On irreducible endospectral graphs[a)]

Jan V. Knop, Wolfgang R. Müller, and Klaus Szymanski
Computer Centre, The University of Düsseldorf, 4000 Düsseldorf, Federal Republic of Germany

Nenad Trinajstić
The Fugjer Bošković Institute, 41001 Zagreb, P. O. Box 1016, Croatia, Yugoslavia

Alexander F. Kleiner
Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311

Milan Randić
Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311 and Ames Laboratory–Department of Energy, Iowa State University, Ames, Iowa 50011

(Received 10 April 1985; accepted for publication 28 May 1986)

The paper considers *endospectral trees*, a special class of graphs associated with the production of numerous isospectral graphs. Endospectral graphs have been considered in the literature sporadically (the name was suggested very recently [M. Randić, SIAM J. Algebraic Discrete Meth. **6**, 145 (1985)]). They are characterized by the presence of a pair of special vertices that, if replaced by any fragment, produce an isospectral pair of graphs. Recently Jiang [Y. Jiang, Sci. Sin. **27**, 236 (1984)] and Randić and Kleiner (M. Randić and A. F. Kleiner, "On the construction of endospectral trees," submitted to Ann. NY Acad. Sci.) considered alternative *constructive* approaches to endospectral trees and listed numerous such graphs. The listing of *all* such trees having $n = 16$ or fewer vertices has been undertaken here. It has been found that relatively few endospectral trees have novel structural features and cannot be reduced to some already known endospectral tree. These few have been named "irreducible endospectral trees." They are responsible for the occurrence of a large number of isospectral trees, leading to, when one considers trees of increasing size, the situation that led Schwenk [A. J. Schwenk, in *New Directions in the Theory of Graphs*, edited by F. Harary (Academic, New York, 1973), pp. 275–307] to conclude that "almost all trees are isospectral."

I. INTRODUCTION

Collatz and Sinogowitz[1] were the first to report that different graphs can have all graph eigenvalues (spectra) the same. Their finding remained dormant for a number of years until the late 1960s and the beginning of the 1970s when there was a rediscovery of isospectral graphs in physics[2] and chemistry.[3] At first there was considerable interest in "hunting down" isospectral graphs,[4] followed by observation of intriguing properties of selected isospectral graphs, which allowed several constructive approaches to be developed.[5] Herndon[6] and Živković, Trinajstić, and Randić[7] independently observed the presence of special vertices in some isospectral graphs, called by Herndon[6] "isospectral points," and called by Živković *et al.*[7] "active sites." The term of Herndon appears a better term, being more specific, and will be adopted here. These vertices, appearing in pairs (in more general situations several vertices or several pairs may appear), have a property that when they are used for adding an arbitrary fragment the exchange of the sites (which produces different graphs) will leave the characteristic polynomial the same. In Fig. 1 we illustrate one such endospectral tree, a graph having $n = 9$ vertices, studied in some detail by Schwenk.[8] The positions 2 and 5 are the isospectral points, or as we prefer to call them here (*vide infra*) endospectral points. *Any* fragment (another graph) attached to either vertex 2 or vertex 5 will produce a pair of isospectral graphs. The simplest is the case of attaching a single edge to either vertex 2 or vertex 5, also illustrated in Fig. 1. The isospectral points thus can produce numerous isospectral graphs. Isospectral points also can occur in different (already isospectral) graphs as illustrated in Fig. 2. Again adding the same fragment at the given sites (each time to a different graph) will produce isospectral graphs as shown in Fig. 2. We find it useful to differentiate the two cases, that of Fig. 1, where a *single* graph has been used to give two isospectral graphs and that of Fig. 2 where two different graphs have been used, in a somewhat analogous way, to produce two isospectral

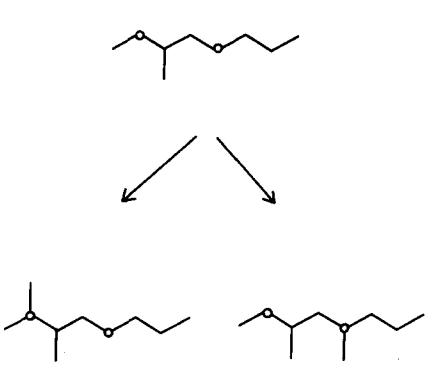


FIG. 1. Schwenk's tree and a pair of isospectral graphs produced by attaching a single edge to endospectral points.

[a)] This work is dedicated to Professor Allen J. Schwenk as an acknowledgment of his many contributions to spectral graph theory.

 0022-2488/86/112601-12$02.50

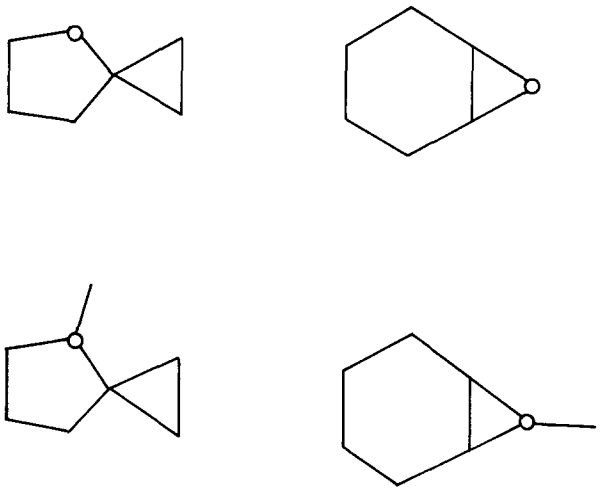


FIG. 2. A pair of isospectral graphs having isospectral points (shown as circles) and the property that when a fragment (here a single edge) is attached to either of the two points another isospectral pair is produced.

graphs. Hence, we refer to the former as endospectral graphs and its "active" points are termed "endospectral points," while those of the two isospectral graphs capable of generating additional isospectral graphs by additions to the selected "active" points, we call "isospectral points." This narrows the term originated with Herndon to cases of the spectral points appearing in *different* (isospectral mates) graphs, while the term "endospectral points" is reserved to such points occurring *within* a single graph. The distinction is particularly useful when both types of graphs are discussed; if no confusion is possible, the term isospectral points can be used, if desired. Finally the reader should be reminded that there are isospectral pairs of graphs that have no isospectral points at all, just as there are graphs that are not isospectral and yet have special points (termed "unusual")[9] characterized by the property common with endospectral and isospectral points: having the same count of self-returning walks and same coefficients for associated eigenvectors. In Table I we give the count of occurrence of these unusual points in trees, as found in another study of Knop and collaborators.[10] Table I gives the percentage of trees having these special structural features, showing a rapid increase in the number of such cases with the increase in the size of trees as measured by n, the number of vertices in a tree.

TABLE I. The number of trees (N), number of trees having vertices of the same code (D), percent N/D, number of pairs of endospectral vertices (EV), and the number of endospectral trees (ET) for $n = 9$ to $n = 16$.

n	N	D	%	EV	ET
9	47	1	2.1	1	1
10	106	3	2.8	0	0
11	235	13	5.5	3	2
12	551	44	8.0	4	4
13	1 301	133	10.2	11	7
14	3 159	364	11.5	18	13
15	7 741	1107	14.3	42	25
16	19 320	3012	15.6	72	45

Legend: n—number of vertices,
N—number of trees (acyclic graphs, restricting the maximal valency to 4),
D—number of isocodal vertices,
%—percent of trees having isocodal vertices,
EV—number of pairs of endospectral vertices,
ET—number of endospectral trees.

Endospectral graphs are of interest for construction and possibly characterization of isospectral graphs. If structural conditions for isospectrality can be fully understood one could revive (conditionally) use of the characteristic polynomials for representation of graphs. Equally, study of endospectral graphs should help to resolve some problems in graph's spectra (e.g., occurrence of common eigenvalues), which in turn are of interest for the problem of graph recognition. In physical chemistry, in the study of structure-property correlations, isospectral graphs may signify molecules with same (or similar) properties, while the chemical physics isospectral networks necessarily point to lattices that will have some statistical constants (e.g., the dimer converings) the same.

II. ENDOSPECTRAL TREES

Besides the tree of Fig. 1, studied by Schwenk, several additional trees have been reported in the literature, see Fig. 3 (Randić,[11] Jiang,[12] and Randić and Kleiner[13]). Jiang was first to report on a single constructional approach using algebraic properties of trees that produced more than a dozen endospectral trees having less than 20 vertices. Thus for the first time one could examine a *collection* of such graphs, instead of individual cases. Among trees, the only known cases are the tree of Schwenk,[8] the tree of Godsil and McKay,[14] and that of Randić.[11] Work of Jiang stimulated a search for alternative *constructional* approaches to generate additional endospectral trees. Although endospectral trees appear to be *rare*, success in constructing additional trees raises the possibility of finding still others. One cannot be sure, unless the constructional approach involves an exhaustive search, which the schemes of Jiang and Randić–Kleiner do not, that all the endospectral trees of a particular size have been found. It is highly desirable to arrive at the complete list of endospectral trees of a particular size. The only sure approach at this moment is an exhaustive search for the endospectral graphs, preferably with the use of a computer. In this paper we report such findings. We have undertaken a systematic computer-assisted search for all endospectral trees having $n = 16$ or fewer vertices. In all, we screened 618 050 trees to report some hundred endospectral cases. Subsequently we have analyzed the derived endospectral trees and found that only about a dozen of them can be considered as "essential," i.e., nontrivially related. These structurally interesting trees we call *irreducible endospectral trees*.

III. COMPUTER SEARCH FOR ENDOSPECTRAL TREES

The basis for the computer-assisted search for endospectral trees is the available program designed for enumeration and construction of all trees.[15,16] The program and the algorithm have been described fully elsewhere[17] so we will mention only the main features of the approach. The first step in generating trees is to devise a unique code for trees. There are

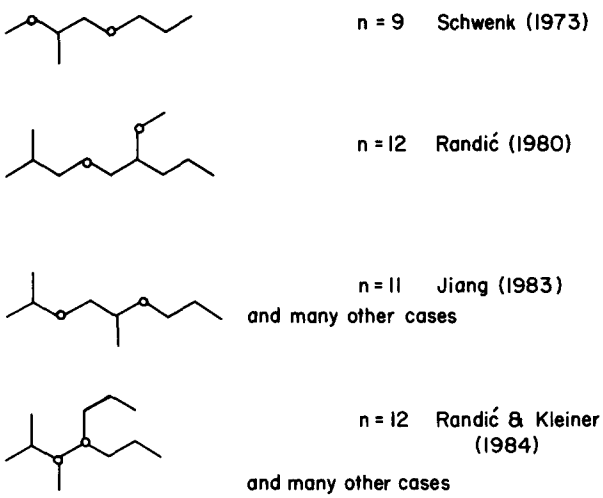


FIG. 3. Some of endospectral trees previously reported in the literature.

several alternative schemes that generate graphs.[18–21] The program used here has been developed and implemented at the Computer Centre at the University of Düsseldorf.[17] Let us illustrate the code used on Schwenk's tree (Fig. 1). The code consists of a string of digits separated occasionally with a colon or a slash. For the graph of Schwenk one obtains

9 43/3:7/311110100 .

The first entry is n, the number of vertices, and the second entry is the ordinal number for the graph (among all graphs having a same n). For example, if one desires to construct all trees having $n = 9$ vertices, use of the particular algorithm will give Schwenk's tree as the 43rd output. The numbers between the slashes /3:7/ are *labels* of the vertices that are endospectral. The labels are implied by the last string of digits, each position in which corresponds to a single vertex. One starts with a vertex having the highest valency, which assumes label 1, and its valency is the first digit of the string (n-tuple). For every vertex but the first, one records not its valency v but $v - 1$, valency decreased by 1. To proceed to the selection of the second vertex one erases vertex 1, which partitions the tree into three disjoint components, each of which has to be examined. We will assume here that the codes for these smaller components are *known* (if not they can always be constructed as a separate task). Having codes for the fragments one concatenates these partial strings into a single code. In order to have a unique code one selects the combinations that make the final code lexicographically maximal. Each vertex of the tree is thus mapped onto one component of the n-tuple. The approach can be extended to rooted trees and even polycyclic graphs built of regular hexagonal cells.[22,23] The listing of the program and discussion of its logical functions has been published.[17] An important component of the program is the subroutine PLTREE, which *plots trees*. Hence the output is graphical as well as digital, which is an important advantage, not yet common to graph generating programs. The subroutine PLTREE transforms the representation of a tree as an n-tuple graphically by drawing the Ariadne thread described by the n-tuple. It spreads the thermal edges uniformly in all directions. Other edges take the directions given by the mean value for the angles of the terminal edges reachable through the chain of intermediate edges from the branching site. In the full drawing, vertices are shown as full circles (in fact they are drawn as small octagons, but these are too small to be recognized necessarily as such). In the present work a minor modification is made by enlarging the circles indicating the endospectral points. The output is shown as Fig. 4, which shows all endospectral trees having from 9 to 13 vertices.

IV. RESULTS

Most of our results are shown graphically starting with Fig. 4, which shows all endospectral trees having $n = 9$ (the graph of Schwenk), $n = 11$, and $n = 12$, as plotted by a computer. Observe that there are no endospectral trees having $n = 10$ vertices. Also observe that among the three outputs for $n = 11$, we have only two distinctive trees, labeled as #185 and #218. Moreover, the tree #185 is related to the Schwenk's graph #43: it can be viewed, as derived from the former, by adding a single edge at *both* endospectral points of Schwenk's graph. It follows from the analysis of Schwenk that such a process will necessarily leave an endospectral graph endospectral. Moreover, both the old endospectral points as well as the two new points will have the endospectral property. We conclude therefore that such augmentation of a graph does not introduce an important novelty and will be referred to as trivial, and will be of no further interest. Hence, we have only one endospectral tree with $n = 11$ vertices, the tree #218, reported by Godsil and McKay.[14] Observed, that although the tree of Schwenk and the tree of Godsil and McKay both have the endospectral points at the same distance they are not simply related. Even though they have some common structural features, they both have the same central parts (the part between the endospectral points) and one of the end parts the same, but they differ in the other thermal group.

A. Endospectral trees having $n = 12$ vertices

The lower part of Fig. 4 shows endospectral trees on 12 vertices, in all, four different trees. None of these trees can be related to either Schwenk's tree or the tree of Godsil and McKay, because they have an odd number of vertices! The reduction and the augmentation process previously mentioned increases or decreases the number of vertices by an even number. The trees cannot be reduced, because there are no endospectral trees on $n = 10$ vertices. Except for the tree #389, the endospectral trees on $n = 12$ vertices have been reported previously in the literature: #435 by Randić and Kleiner[13], #533 by Randić[14], and #539 by Jiang.[12] Graph #435 is particularly interesting; the two endospectral vertices are *adjacent* to one another. Hence the two "halves" of the graph obtained by disconnecting the endospectral vertices balance one another when it comes to the count of self-returning walks. Hence, the two "halves" are "exchangeable," if considered as components of larger graphs. Let us illustrate one such case. We show in Fig. 5 the tree of Godsil and McKay on 11 vertices, which we augmented by adding a single edge at both endospectral vertices. The derived tree has as one of its end groups the terminal portion of the endospectral tree #435 with the adjacent endospectral points.

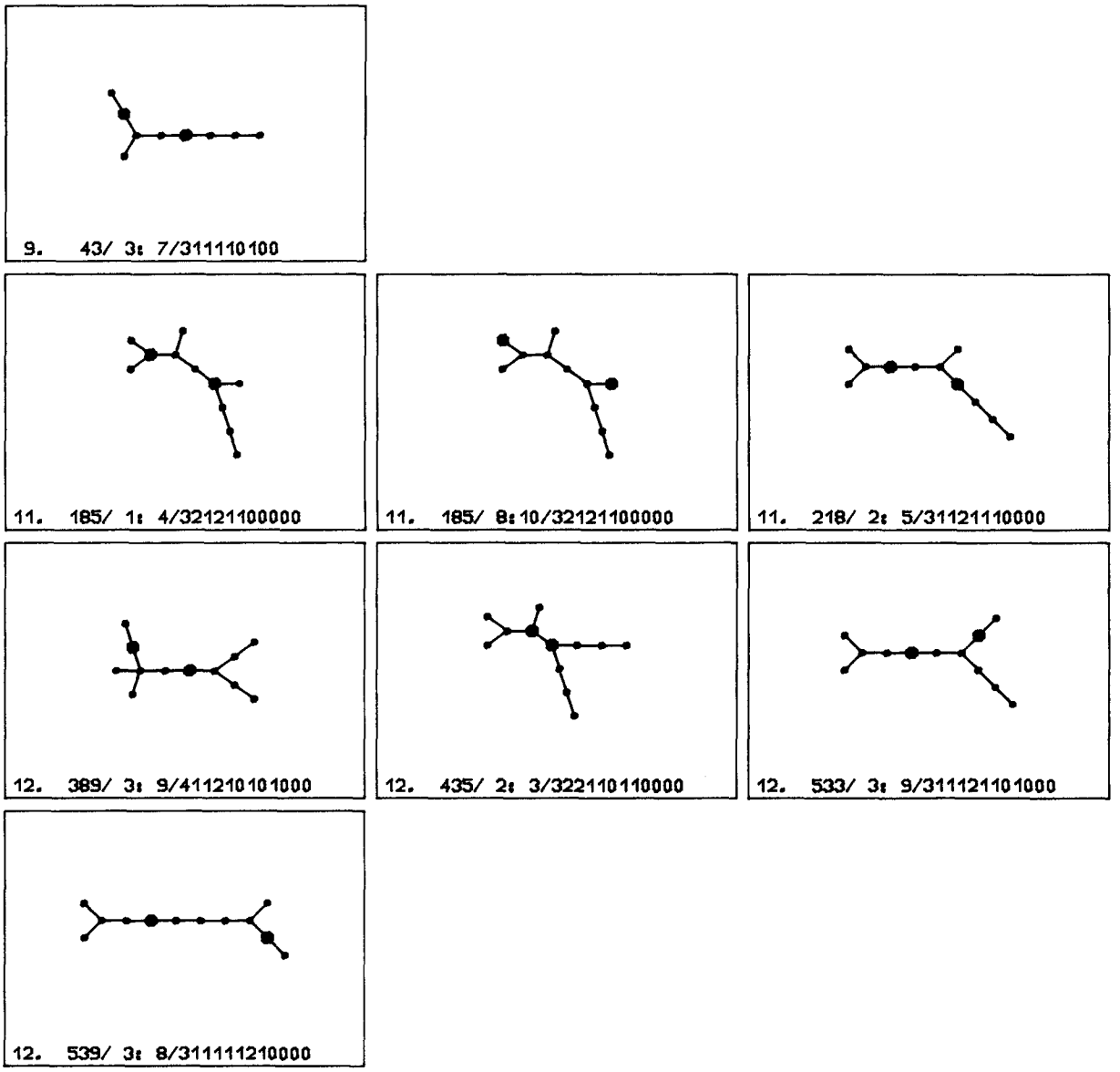
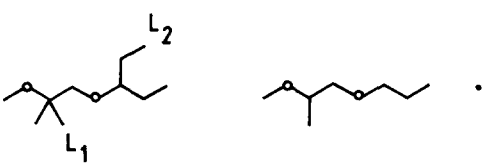


FIG. 4. Computer printout of all endospectral trees having $n = 9$ (one case); $n = 11$ (two trees, one having trivial extension); and $n = 12$ (four cases) vertices.

We now perform a "surgery" by replacing the end of the augmented graph in the middle of Fig. 5 by the *other* endospectral "half" of the tree #435. This produces a tree on $n = 15$ vertices, which we expect to be endospectral, and indeed the graph is found in the computer output under the number #7194. Another more apparent procedure of producing novel endospectral trees from graph #435 is that of the insertion of any symmetrical fragment between the two endospectral points. In Fig. 6 we show several so-derived endospectral trees. Tree #435 is the smallest endospectral tree having adjacent endospectral points, and at the same time the smallest nontrivial endospectral tree having endospectral points with valency greater than 2.

The previously unreported endospectral tree #389 has its endospectral points at a distance 3:



A comparison with Schwenk's graph is instructive: #389 can be viewed as structurally related to the Schwenk's tree by having additional branches L_1 and L_2. It thus appears that adding L_1 near one of the endospectral points is "compensated" by adding the other (larger) fragment L_2 near the other endospectral point. Observe that the fragments L_1 and L_2 are not in a general way equivalent, but just in the particular environment dictated by the other points in #389. However, if the above "balancing" act of the two fragments L_1

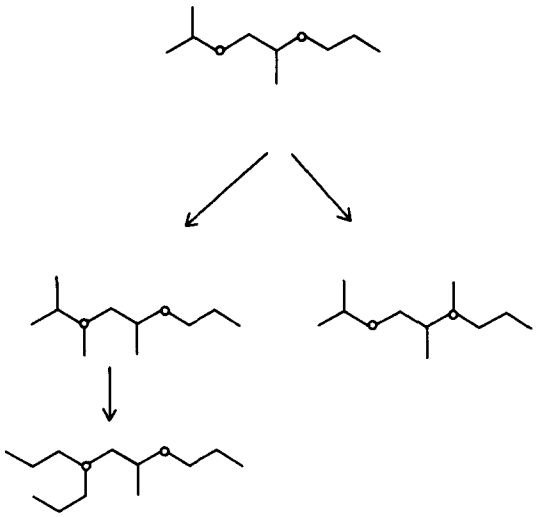
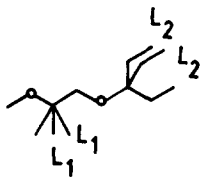


FIG. 5. Construction of endospectral graphs from a parent structure.

and L_2 is valid, it *ought to hold* also if we double the number of L segments. That is, we anticipate that the tree



also will be endospectral. Here instead of L_1 we have $2\times L_1$ and instead of L_2 we have $2\times L_2$, resulting in a tree on $n = 15$ points. By examining our computer output we find the above

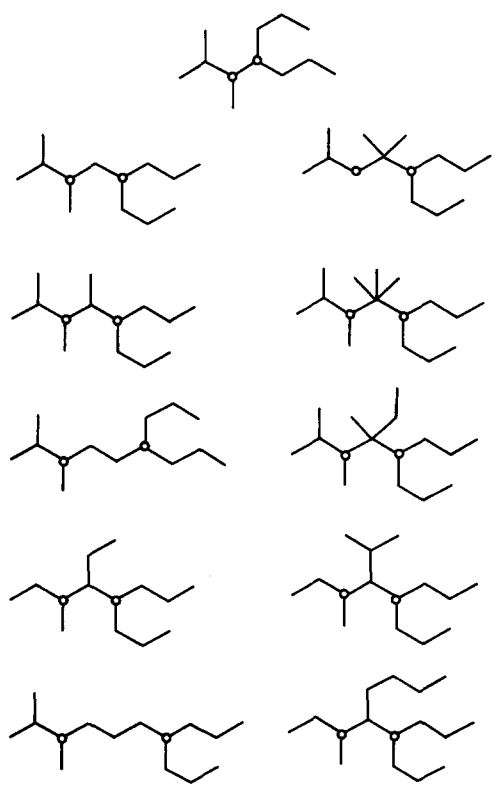


FIG. 6. Simple constructions of numerous endospectral trees all derived by insertion of a fragment between two *adjacent* endospectral vertices.

tree is endospectral tree #3190, thus verifying the validity of our speculation. We may refer to the two vertices used for the substitution as *restricted sites*, by which we want to indicate that *qualified* fragments (like L_1 and L_2) may be attached here without upsetting the endospectrality of the parent (unsubstituted) graph. It is of interest to compute the n-tuples for the three graphs #43, #389, and #3190:

Schwenk's tree #43: 311110100,

this work #389: 411210101000,

this work #3190: 511310101010000.

The above codes can be written in a condensed form as

#43: 3(1)2(1)(10)2(0),

#389: 4(1)2(2)(10)3(0)2,

#3190: 5(1)2(3)(10)4(0)3,

which clearly suggests the existence of a family of endospectral trees having codes

$$(k+2)(1)2(k)(10)(k+1)(0)k, \quad k = 1,2,3,\ldots .$$

B. Endospectral trees having $n = 13$ vertices

In Fig. 7 we show all endospectral trees having 13 vertices. In all seven different trees, some appear two or three times in view of the presence of trivial endospectral points. It is not difficult to identify the trivial cases: #684 and #1138 can be derived from Schwenk's graph #43, while #1120 is similarly related to Jiang's tree #218. Finally #1136 can be derived by insertion of a (symmetric) vertex between the adjacent endospectral points of #435. This leaves only three nontrivial cases, #658, #1191, and #1264, the last case already reported by Jiang.[24]

C. Endospectral trees having $n = 14$ vertices

Five of the 13 endospectral trees (Fig. 8) immediately can be identified as trivial by simply observing that they permit more than one pair of sites for the endospectral vertices. Of these #1423, #2165, #2835, and #2962 have a pair of bonds attached to an existing endospectral point in smaller endospectral trees (#435, #389, #533, and #539, respectively), while #2932 is derived from #435 by *insertion* of a symmetric fragment. Inserted pair of vertices, can also act as isospectral points. In the case of #2632 we also have an *insertion* of two vertices in a symmetrical fashion, thus this case also can be discarded as having no structural novelty. We are left with seven nontrivially related graphs: #2450, #2890, #3004, #3080, #3105, #3120, and #3126. The pair #3105 and #3120 are intriguing, they differ in a single detail, i.e., the linkage of one of the endospectral terminal groups to the central portion of four vertices in a chain. We find tree #3105 is particularly interesting; if one erases either of the endospectral points one obtains as one fragment Schwenk's tree on nine vertices.

V. LARGER ENDOSPECTRAL TREES

We have seen from examining cases $n = 13$ and $n = 14$ that the number of trivially related endospectral trees grows

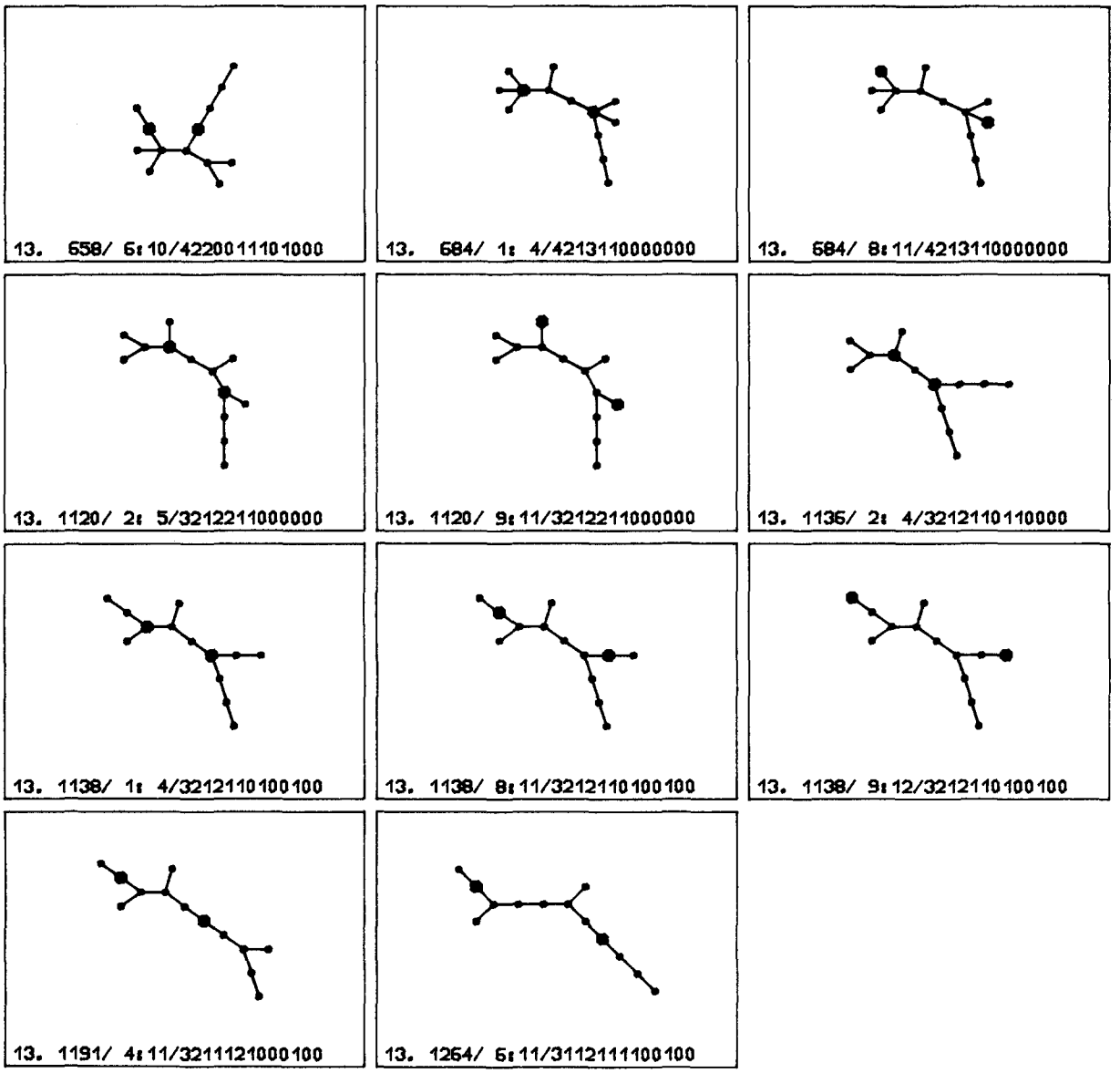


FIG. 7. All endospectral trees having $n = 13$ vertices (seven different trees, several having trivial extensions).

fast. In the remainder of the paper we will therefore no longer consider such cases. Besides the graphs derived by *adding* a fragment at endospectral points of a smaller endospectral tree, which forms one class of trivial extensions, we also have the cases of *insertion* of symmetrical fragments between initially adjacent endospectral points, which form the other class of simple augmentation of endospectral trees. When all the above are discarded, out of 25 endospectral trees with $n = 15$ vertices we are left with 11 trees (shown in Fig. 9): #4557, #4598, #5094, #5233, #6355, #6470, #6533, #7252, #7583, #7607, and #7638. The question is whether any of these 11 can be derived, perhaps by some new structural procedure, from smaller endospectral trees. We have already seen how trees #43, #389, and #3190 with 9, 12, and 15 points, respectively, form a *family* of endospectral trees; each next member in the family to have to increase the size by 3. Are there other such (nontrivially) related endospectral trees? By examining the trees reported in this paper, we find that comparison of the following may be of interest:

#1264 with 13 vertices and the code: 3112111100100,

#2890 with 14 vertices and the code: 32121000111100,

#5233 with 15 vertices and the code: 421111001210000,

#6811 with 16 vertices and the code: 5211110012100000.

The regularity in the codes starts to become apparent, and we are anticipating these endospectral trees to be members of another *family* of endospectral trees. The above graphs are shown in Fig. 10, which suggests that the centrally located vertex *allows* certain substitutions. The above codes appear somewhat scrambled: we see that the correct digits are there, but for the first two members of the family they are in somewhat permuted order. We will see that this is not an error in the programming and the coding, but follows from the fact that endospectral trees #1264 and #2890 are also the leading members of additional families of endospectral trees. The

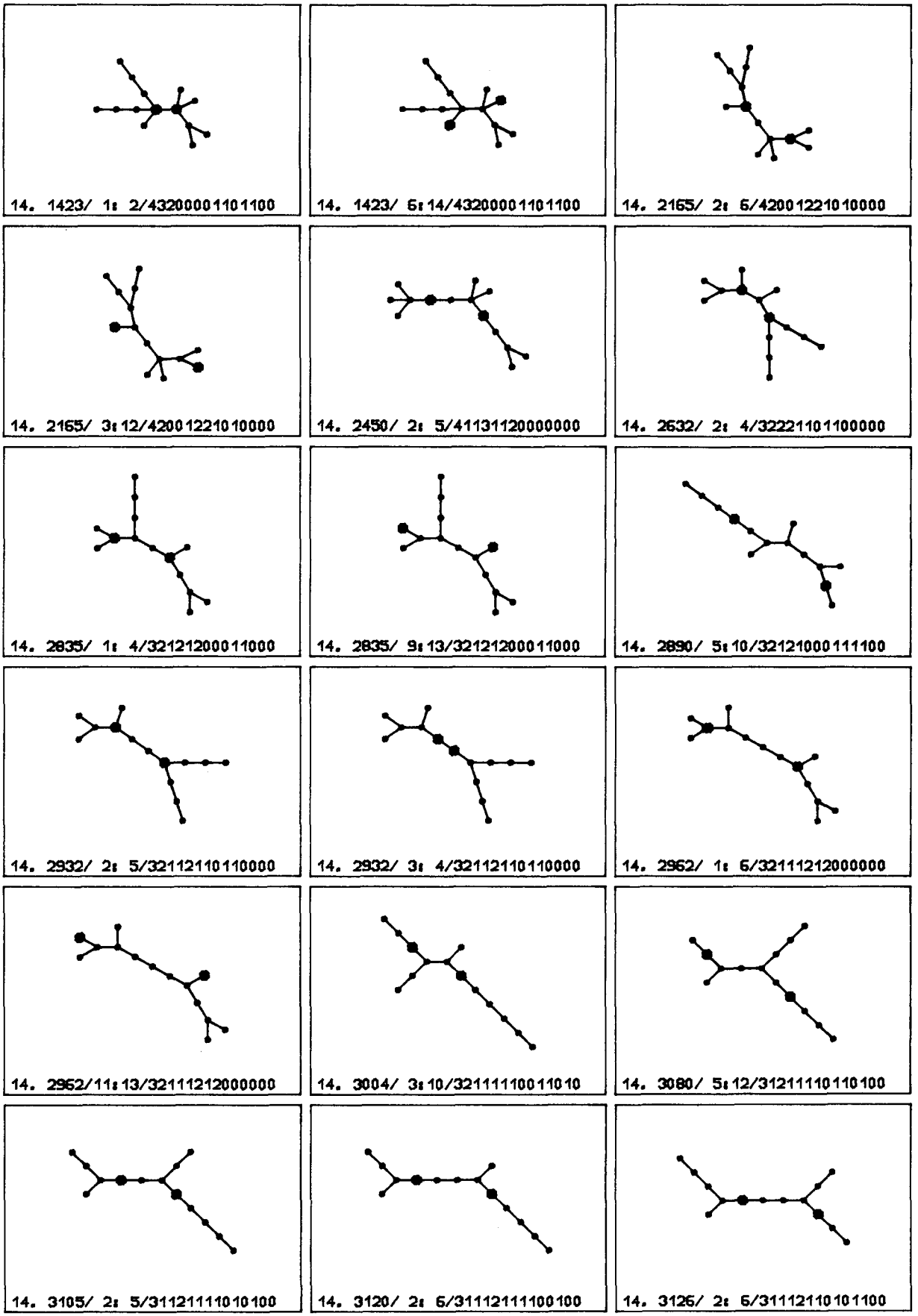


FIG. 8. All endospectral trees having $n = 14$ vertices (13 different cases).

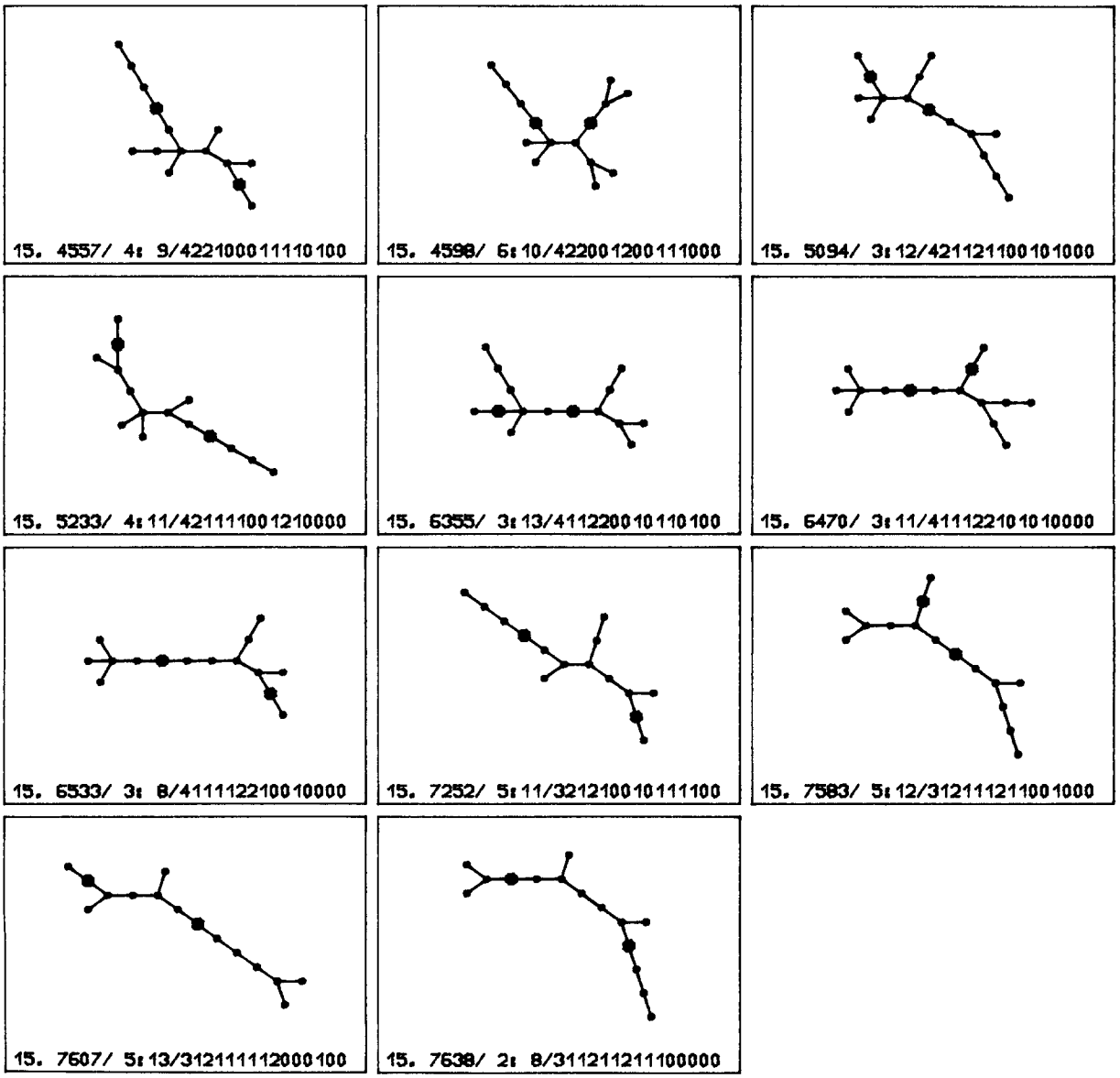
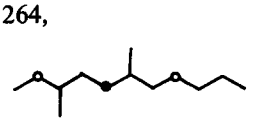


FIG. 9. Nontrivial endospectral trees having $n = 15$ vertices.

problem to consider is as follows: Which fragments *qualify* for the substitution at the central vertex of the parent #1264? It appears that *any* fragment substituted at the central vertex will not upset the inherent endospectrality of the parent skeleton. In Table II and Fig. 11 we illustrate the above for all possible families that can be constructed restricting the size of graphs to $n = 16$. The two portions of the graph #1264 are, however, not equivalent (i.e., exchangeable) because they have a different number of vertices. We anticipate that the above illustration is but a special case of a more general situation that we will state as a proposition.

Proposition 1: Endospectral trees that have a vertex located in the middle of the two endospectral vertices at which a branching occurs will remain endospectral, irrespective of the fragment at the branching site.

The proof of this proposition can be argued along the following lines. The two "halves" of the midsection are identical, except for the relative orientation with respect to the endospectral points. We assume two items: endospectral tree with a midpoint and endospectral tree, the same as above, except for an additional branch (an edge) at the midpoint. The second endospectral tree ensures that at the midpoint there is a 1:1 correspondence between any walk starting at one endospectral point and ending at the midpoint. Reaching the midpoint in this way, one can ignore the "history" of the walk, and clearly from there on the two walks starting at different end points will have counts identical to those of any longer walks extending into the substituted fragment. In searching for additional structural families of endospectral trees, one has to exercise some care and not rush to suggesting novel relationships without carefully examining *all* contributing terms. For example, if we consider the parent tree #1264,



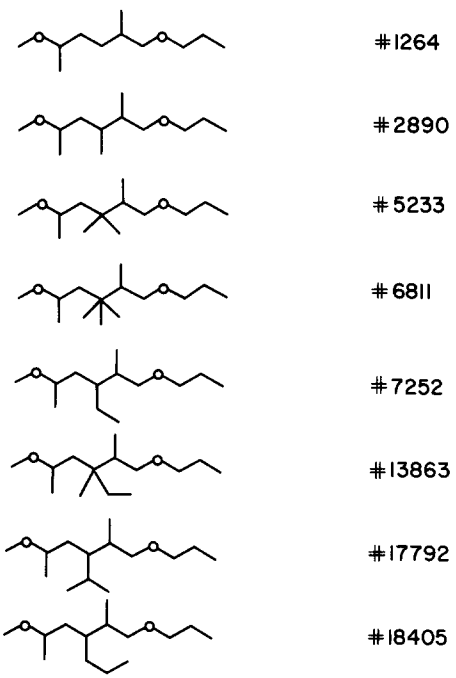
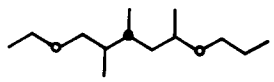


FIG. 10. A sequence of endospectral trees having the same "end" groups and suggesting a presence of a "central" (symmetrical) vertex at which an arbitrary fragment can be attached.

and recognize the special role of the central vertex, we ought not to forget that this property of the central vertex is tied to the presence of the particular ending segments. One may speculate that the tree derived from a *symmetrical* insertion of two vertices adjacent to the central vertex also will be endospectral, and a parent of yet another family of endospectral trees. However, the derived tree *is not* endospectral, because now the central part of the tree requires different end groups (if such exist and can be found). This is the reason why we underlined *endospectral* in our proposition. Hence an endospectral graph has to be taken as a whole. Sometimes one can replace one end group with another, but this does not necessarily mean that analogous constructions valid for one family will hold for the other. An illustration is provided by comparison of #1264 having 13 vertices and #7638 having 15 vertices. The endospectral tree #7638 can be viewed as derived from #1264 by changing one of the end groups. The graph has a midvertex, but is not the source of another family of endospectral graphs because by inserting a single branching edge at the central site we produce a tree,



that *is not* endospectral. This need not be surprising because in #1264 the end groups balance each other creating the special central site, but clearly, such balance is going to be upset by changing only one of the end groups.

Finally in Fig. 11 we illustrate a number of nontrivial endospectral trees having $n = 16$ vertices. They reveal additional interesting structural features. Two trees have *adjacent* endospectral points (of valency 3) and therefore can produce families of endospectral trees analogous to those previously discussed for 12-vertex graph #435.

TABLE II. Regularity in the codes for a family of structurally related endospectral trees.

n	Graph #	Code
13	1264	3112111100100
14	2890	32121000111100
15	7252	321210010111100
16	18165	3212120010110100

VI. SOME OPEN QUESTIONS

We have not exhausted all useful comparisons. It is possible that additional structural relationships will emerge, becoming more visible when the search of construction for larger endospectral trees is continued. The main finding of our work is that endospectral trees are rather rare, if one discards trivial cases, and limits oneself to "sporadic" rather than "families" of such trees. It is conceivable that few structural requirements, if not a single one, may typify numerous endospectral trees. We have not succeeded in pinpointing such a structural element, but with continuing interest in endospectral and isospectral graphs we may have a more complete picture of the structural characteristic of these graphs. There are still a number of open and not yet considered questions, which eventually may help in resolving other such questions. For example: Is there a *single* constructional approach that would cover *all* cases? Can we write *any* "central" portion of a graph (tree) and then find "end" groups that would make the tree endospectral? Are there trees with more than a single (nontrivial) pair of endospectral vertices? In the case of cyclic graphs (Fig. 2), we have cases of two (and more) isospectral points. Can endospectral points be terminal points (of valency 1)? Can endospectral points have different valency? In all the cases considered here (including trivial cases also), the endospectral points always have the same valency: In nontrivial cases valency 2 or 3 (in trivial augmentations higher valencies) can occur. Do endospectral (and isospectral) points come in *pairs*? Or is it possible to have a triplet of such points? Are there endospectral trees (and for that matter isospectral trees) in which there are no vertices of valency 2? Are endospectral trees responsible for *all* cases of isospectral trees, "sporadic" cases merely being the intersection of different families if isospectral graphs? Some of these questions may be easier to answer than others and not all are of equal merit. Schwenk[25] answered the first two questions: endospectral points necessarily have a same valency. Since u and v are endospectral this implies that $Ch(G-u;x) = Ch(G-v;x)$, it must be that $G-u$ and $G-v$ have equally many edges. Therefore, u and v must have the same degree. Concerning terminal endospectral points, when they occur in a tree T, their neighbors u and v are necessarily endospectral in the tree

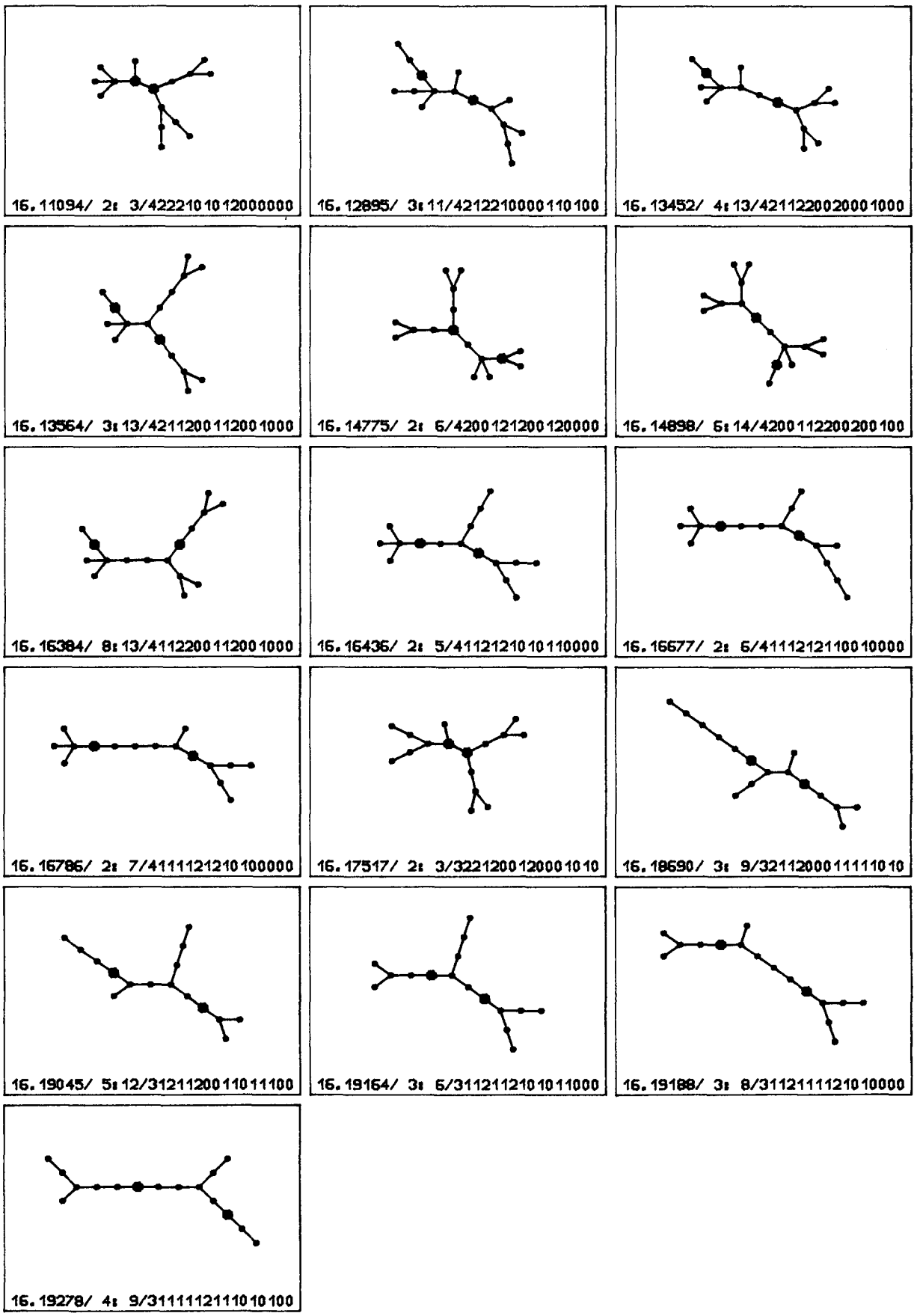


FIG. 11. Nontrivial endospectral trees having $n = 16$ vertices.

$T' = T - u - v$. Hence, the terminal endospectral points can be viewed as trivial. This case is already illustrated in Fig. 4 on #185, Fig. 7: #684, #1120, #1138; and Fig. 8: #2163, #2835, and #2962.

One of the reasons for a study of endospectral trees is that they *possibly* can give some insight into the properties of endospectral cyclic and polycyclic structures, the topic that is of importance in attempts to characterize isospectral graphs in general. If we succeed in characterizing isospectral graphs, then the role of the characteristic polynomial in the comparative study of graphs may be resurrected. Among the numerous (endless?) list of graph invariants, the characteristic polynomial, graph spectra, moments, random walks, and in particular, self-returning random walks are very important. They are all intimately connected, some being expressed in integers (a result of a counting process), others are analytical. A complete understanding of these fundamental invariants is clearly prerequisite for the study of any other composite quantities. Considerable progress was made in the last few years in this area: methods for fast (computer-assisted) construction of the characteristic polynomials are available,[26–30] alternative representations (via Chebyshev polynomials[31–33] have been explored, factoring of the characteristic polynomial (for trees only[34] but some extensions are considered for graphs in general[35]) have been studied, which factoring is not the result of the symmetry of the graphs. Finally the concept of "characteristic equations" has been introduced. These are the equations that determine the coefficients of the characteristic polynomial. It was found that in most cases even isospectral graphs have a different set of characteristic equations,[24] although the existence of some pathological (highly symmetrical polycyclic) structures, to which Schwenk[36] drew attention, make the characteristic equations nonunique. Revival of interest in the characteristic polynomial and related topics may provide some answers to the above questions as well as to other closely related questions. For example, a number of questions concerning graph spectra, such as the occurrence of common eigenvalues,[37] inclusion of spectra of a smaller graph in a larger one,[38,39] the excessive "degeneracy" (i.e., multiplicity of roots), all tied to nodal characteristics of graphs remain for the most part unresolved, although some partial results have been offered.[40] We hope that study of endospectral graphs may directly or indirectly help to resolve some of the problems mentioned.

[1]L. Collatz and U. Sinogowitz, "Spektren endlicher Grafen," Abh. Math. Sem. Univ. Hamburg **21**, 63 (1957). (According to a footnote the published work was based on preliminary results of U. Sinogowitz derived just before his death in 1944.)

[2]C. A. Baker, "Drum shapes and isospectral graphs," J. Math. Phys. 7, 2238 (1966); M. E. Fisher, "On hearing the shape of a drum," J. Combin. Theory **1**, 105 (1966).

[3]A. T. Balaban and F. Harary, "The characteristic polynomial does not uniquely determine the topology of a molecule," J. Chem. Docum. **11**, 258 (1971); K. Kawasaki, K. Mizutani, and H. Hosoya, "Tables of non-adjacent numbers, characteristic polynomials and topological indices. II. Mono- and bicyclic graphs," Nat. Sci. Found. Ann. Rep. Ochanomizu Univ. **22**, 181 (1971).

[4]F. Harary, C. King, A. Mowshowitz, and R. C. Read, "Cospectral graphs and digraphs," Bull. London Math. Soc. **3**, 321 (1975).

[5]W. C. Herndon and M. L. Ellzey, Jr., "Isospectral graphs and molecules," Tetrahedron **31**, 99 (1975); M. Randić, N. Trinajstić, and T. Živković, "Molecular graphs having identical spectra," J. Chem. Soc. Faraday Trans. II **72**, 244 (1976); A. J. Schwenk, W. C. Herndon, and M. L. Ellzey, Jr., "The construction of cospectral composite graphs," Ann. NY Acad. Sci. **319**, 490 (1979); E. Heilbronner, "Some comments on cospectral graphs," Match **5**, 105 (1979).

[6]W. C. Herndon, "Isospectral molecules," Tetrahedron Lett. **1974**, 671.

[7]T. Živković, N. Trinajstić, and M. Randić, "On conjugated molecules with identical topological spectra," Mol. Phys. **30**, 517 (1975).

[8]A. J. Schwenk, "Almost all trees are cospectral," in *New Directions in the Theory of Graphs*, edited by F. Harary (Academic, New York, 1973), pp. 275–307.

[9]M. Randić, A. Graovac, and W. L. Woodworth, "Unusual random walks," Int. J. Quant. Chem. **24**, 435 (1983).

[10]J. V. Knop, W. R. Müller, K. Szymanski, M. Randić, and N. Trinajstić, "Note an acyclic structures and their self-returning walks," Croat. Chem. Acta **56**, 405 (1983).

[11]M. Randić, "Random walks and their diagnostic value for characterization of atomic environment," J. Comput. Chem. **1**, 386 (1980).

[12]Y. Jiang, "Problems on isospectral molecules," Sci. Sin. **27**, 236 (1984).

[13]M. Randić and A. F. Kleiner, "On the construction of endospectral trees," submitted to Ann. NY Acad. Sci.

[14]C. Godsil and B. McKay, "Some computational results on the spectra of graphs," Lect. Notes Math. **560**, 72 (1976).

[15]J. V. Knop, W. R. Müller, Ž. Jeričević, and N. Trinajstić, "Computer enumeration and generation of trees and rooted trees," J. Chem. Inf. Comput. Sci. **21**, 91 (1981).

[16]J. V. Knop, K. Szymanski, G. Jashari, and N. Trinajstić, "The connection between the adjacency matrix and the boundary code of benzenoid hydrocarbons," Croat. Chem. Acta **56**, 443 (1983); J. V. Knop, K. Szymanski, and N. Trinajstić, "A note on trees with polyhexagonal supervertices," Match **16**, 103 (1984).

[17]J. V. Knop, W. R. Müller, K. Szymanski, and N. Trinajstić, *Computer Generation of Certain Classes of Molecules* (SKTH, Zagreb, Croatia, 1985).

[18]Y. Kudo and S-i. Sasaki, "Principle for exhaustive enumeration of unique structures consistent with structural information," J. Chem. Inf. Comput. Sci. **16**, 43 (1976).

[19]V. V. Serov, M. E. Elyashberg, and L. A. Gribov, "Mathematical synthesis and analysis of molecular structures," J. Mol. Struct. **31**, 381 (1976).

[20]C. A. Shelly, T. R. Hays, M. E. Munk, and R. V. Roman, "An approach to automated partial structure expansion," Anal. Chim. Acta **103**, 121 (1978); M. E. Munk, C. A. Shelley, H. B. Woodruff, and M. O. Trulson, "Computer-assisted structure elucidation," Fresenius Z. Anal. Chem. **313**, 473 (1982); C. A. Shelly and M. E. Munk, "CASE, a computer model of the structure elucidation process," Anal. Chim. Acta **133**, 5 (1981).

[21]L. M. Masinter, N. S. Sridharan, J. Lederberg, and D. H. Smith, "Applications of artificial intelligence for chemical interference. XIII. Labeling objects having symmetry," J. Am. Chem. Soc. **96**, 7714 (1974); A. J. Thakkar, "The coming of the computer age to organic chemistry: Recent approaches to systematic synthesis analysis," Topics Curr. Chem. **39**, 3 (1973).

[22]N. Trinajstić, J. V. Knop, W. R. Müller, K. Szymanski, "Computer generation of isomeric structures," Pure Appl. Chem. **55**, 379 (1983).

[23]J. V. Knop, K. Szymanski, Ž. Jeričević, and N. Trinajstić, "Computer enumeration and generation of benzenoid hydrocarbons and identification of bay region," J. Comput. Chem. **4**, 23 (1983).

[24]M. Randić, "On the characteristic equations of the characteristic polynomial," SIAM J. Algebraic Discrete Meth. **6**, 145 (1985).

[25]A. J. Schwenk (private correspondence, February 1985).

[26]K. Balasubramanian, "Computer generation of the characteristic polynomials of chemical graphs," J. Comput. Chem. **5**, 387 (1984).

[27]K. Balasubramanian, "The use of Frame's method for the characteristic polynomials of the chemical graphs," Theor. Chim. Acta **65**, 49 (1984).

[28]K. Balasubramanian, "Spectra of chemical trees," Int. J. Quant. Chem. **21**, 581 (1982).

[29]K. Balasubramanian and M. Randić, "The characteristic polynomials of structures with pending bonds," Theor. Chim. Acta **61**, 307 (1982); K. Balasubramanian and M.Randić, "Spectral polynomials of systems with

general interactions," Int. J. Quant. Chem. **28**, 481 (1985).
[30]P. Křivka, Ž. Jeričević, and N. Trinajstić, "On the computation of the characteristic polynomial of a chemical graph," Int. J. Quant. Chem: Quant. Chem. Symp. (in press).
[31]M. Randić, "On alternative form of the characteristic polynomial and the problem of graph recognition," Theor. Chim. Acta **62**, 485 (1983).
[32]H. Hosoya and M. Randić, "Analysis of the topological dependency of the characteristic polynomial in its Chebyshev expansion," Theor. Chem. Acta **63**, 473 (1983).
[33]N. Ohkami and H. Hosoya, "Topological dependency of the aromatic sextets in polycyclic benzenoid hydrocarbons. Recursive relations of the sextet polynomial," Theor. Chim. Acta **64**, 153 (1983); H. Hosoya and N. Ohkami, "Operator technique for obtaining the recursion formulas of characteristic and matching polynomials as applied to polyhex graphs," J. Comput. Chem. **4**, 585 (1983).
[34]M. Randić, B. Baker, and A. F. Kleiner, "On factoring the characteristic polynomial," Int. J. Quant. Chem: Quant. Chem. Symp. **19**, 107 (1986).
[35]M. Randić, work in progress.
[36]A. J. Schwenk (private correspondence, 1984).
[37]M. Randić (unpublished, preprint available).
[38]T. Živković, N. Trinajstić, and M. Randić, "On topological spectra of composite molecular systems," Croat. Chem. Acta **49**, 89 (1977).
[39]S. S. D'Amato, B. M. Gimarc, and N. Trinajstić, "Isospectral and subspectral molecules," Croat. Chem. Acta **34**, 1 (1981).
[40]U. Wild, J. Keller, and H. H. Günthard, "Symmetry properties of the hückel matrix," Theor. Chem. Acta **14**, 383 (1969).

A complete determination of the zeros of weight-1 6*j* coefficients

Andrew Bremner
Department of Mathematics, Arizona State University, Tempe, Arizona 85287
Simcha Brudno
Chicago College of Osteopathic Medicine, Chicago, Illinois 60615

(Received 28 February 1986; accepted for publication 18 June 1986)

It is shown how all zeros of weight-1 6*j* coefficients arise as particular cases of a four-parameter family of such zeros. The parametrization is given explicitly.

I. INTRODUCTION

A Racah operator is a linear operator acting on a particular abstract Hilbert space, and gives rise to the Racah coefficients. See Biedenharn and Louck[1] for a full discussion, together with motivation for the importance of their study. Considerable interest has been shown in the nontrivial zeros of the Racah coefficients because these determine vector spaces belonging to the null space of a Racah operator, and accordingly give structural information concerning the operator itself.

One method of classifying the zeros of the Racah coefficients has been given by Brudno.[2] Here, it is observed that the explicit expression for each of the coefficients, as given by Racah,[3] is an alternating sum, and the author bases his classification on the number of nonzero terms occurring in this sum. This is shown to be equivalent to a classification by weights of the corresponding Racah operator by Brudno and Louck.[4]

We introduce notation for the 6*j* coefficient $\{{a\ b\ e \atop d\ c\ f}\}$, which up to sign is equal to a Racah coefficient. The coefficient is given by a polynomial function in the arguments a,b,c,d,e,f, which represent angular momentum quantum numbers; an explicit realization of this polynomial is given in Biedenharn and Louck,[1] p. 142. The domain of definition of a,b,c,d,e,f is that they must be non-negative integers or non-negative half-integers satisfying the triangle condition on (a,b,e), (a,c,f), (b,d,f), (c,d,e) [where the triangle condition on (p,q,r) is that $-p+q+r, p-q+r, p+q-r$ are all non-negative integers]. An alternative notation for the 6*j* coefficient is the 4×3 array of Bargmann[5]

$$\begin{Bmatrix} a & b & e \\ d & c & f \end{Bmatrix} = \begin{Bmatrix} d+f-b & c+f-a & c+d-e \\ a+f-c & b+f-d & a+b-e \\ d+e-c & b+e-a & b+d-f \\ a+e-b & c+e-d & a+c-f \end{Bmatrix}.$$

The smallest entry in the Bargmann array is called the weight of the corresponding 6*j* coefficient and is equal to the number one less than the number of terms in the alternating sum, as mentioned above.

A nontrivial zero of a 6*j* coefficient is now defined to be a sextuple (a,b,c,d,e,f) of non-negative integers or non-negative half-integers, such that all entries in the corresponding Bargmann array are non-negative integers. Since coefficients of weight-0 possess no nontrivial zeros, then nontrivial zeros of 6*j* coefficients have corresponding Bargmann arrays with every entry a positive integer.

The first interesting case is that of weight-1 coefficients having two terms in the alternating sum expression for the coefficient. This has been studied by Brudno and Louck[6] (see also Lindner[7,8] for a more general statement) with the following results. If $\{{a\ b\ e \atop d\ c\ f}\} = 0$ and is of weight 1, then by applying a symmetry if necessary, it follows that there exist integers X,Y,Z,U,V,W satisfying

$$\begin{aligned} X^3+Y^3+Z^3 &= U^3+V^3+W^3, \\ X+Y+Z &= U+V+W, \end{aligned} \tag{1}$$

with

$$\begin{Bmatrix} a & b & e \\ d & c & f \end{Bmatrix} = \begin{Bmatrix} \tfrac{1}{4}(X+U-2) & \tfrac{1}{4}(Y+V-2) & \tfrac{1}{4}(W-Z-4) \\ \tfrac{1}{4}(X-U) & \tfrac{1}{4}(Y-V) & \tfrac{1}{4}(W+Z-2) \end{Bmatrix}. \tag{2}$$

They give a one-parameter solution to Eqs. (1) (actually in the form given it is represented as a parametrization homogeneous in two parameters), due to Gérardin (see Dickson,[9] pp. 565 and 713) in 1916. Bremner[10] studies the Diophantine system (1) further, and produces some two-parameter solutions.

It is the object of this paper to show how all solutions of the system (1) may be described in terms of a three-parameter solution. In particular, it follows that all nontrivial zeros of weight-1 6*j* coefficients arise as particular cases of a four-parameter family, which we give explicitly.

II. A FIRST PARAMETRIZATION

In Bremner,[10] it is stated that for a nonrational variety V, the problem of determining all rational points on V is in general a very difficult problem. However, using an observation of the second author of this paper, it turns out that the system (1), representing geometrically a cubic threefold, is actually a rational variety and, accordingly, a complete description of its rational points is readily forthcoming.

In (1), substitute

$$\begin{aligned} Y &= X+\alpha, \quad U = X+\gamma, \\ Z &= X+\beta, \quad V = X+\delta, \\ W &= X+(\alpha+\beta-\gamma-\delta). \end{aligned} \tag{3}$$

The equations reduce to

 0022-2488/86/112613-03$02.50

$$3X^3+3X^2(\alpha+\beta)+3X(\alpha^2+\beta^2)+(\alpha^3+\beta^3)$$
$$=3X^3+3X^2(\alpha+\beta)$$
$$+3X(\gamma^2+\delta^2+(\alpha+\beta-\gamma-\delta)^2)$$
$$+(\gamma^3+\delta^3+(\alpha+\beta-\gamma-\delta)^3),$$

simplifying to

$$2X[\gamma^2+\delta^2+\alpha\beta-\alpha\gamma-\alpha\delta-\beta\gamma-\beta\delta+\gamma\delta]$$
$$=(\alpha+\beta)^2(\gamma+\delta)-(\alpha+\beta)(\gamma+\delta)^2$$
$$-\alpha\beta(\alpha+\beta)+\gamma\delta(\gamma+\delta). \tag{4}$$

We solve Eq. (4) for X, and then backsubstitute into Eqs. (3).

Multiplying throughout by the denominator to ensure polynomial expressions [the system (1) is homogeneous], there results

$$\begin{aligned}
&X:Y:Z:U:V:W\\
&=\alpha^2(-\beta+\gamma+\delta)-\alpha(-\beta+\gamma+\delta)^2+(\gamma+\delta)(\beta-\gamma)(\beta-\delta):\\
&\alpha^2(\beta-\gamma-\delta)+\alpha(-\beta^2+\gamma^2+\delta^2)+(\gamma+\delta)(\beta-\gamma)(\beta-\delta):\\
&\alpha^2(-\beta+\gamma+\delta)+\alpha(\beta^2-(\gamma+\delta)^2)+(-\beta^2(\gamma+\delta)+\beta(\gamma^2+\delta^2)+\gamma\delta(\gamma+\delta):\\
&\alpha^2(-\beta+\gamma+\delta)+\alpha(-\beta^2+2\beta(2\gamma+\delta)-(3\gamma+\delta)(\gamma+\delta))\\
&+\beta^2(\gamma+\delta)-\beta(\gamma+\delta)(3\gamma+\delta)+\gamma(2\gamma^2+3\gamma\delta+3\delta^2):\\
&\alpha^2(-\beta+\gamma+\delta)+\alpha(-\beta^2+2\beta(\gamma+2\delta)-(\gamma+\delta)(\gamma+3\delta))\\
&+\beta^2(\gamma+\delta)-\beta(\gamma+\delta)(\gamma+3\delta)+\delta(3\gamma^2+3\gamma\delta+2\delta^2):\\
&\alpha^2(\beta-\gamma-\delta)+\alpha(\beta^2-4\beta(\gamma+\delta)+(3\gamma^2+4\gamma\delta+3\delta^2))\\
&+(-\beta^2(\gamma+\delta)+\beta(3\gamma^2+4\gamma\delta+3\delta^2)-(\gamma+\delta)(2\gamma^2+\gamma\delta+2\delta^2)).
\end{aligned} \tag{5}$$

This furnishes a parametric solution to (1), ostensibly in four parameters. However, by homogeneity, one can divide throughout by, say, α^3 to see that (5) is really a three-parameter solution in β/α, γ/α, and δ/α. Nonetheless, it is preferable to leave the parametrization in homogeneous form.

We now claim that any solution of (1) occurs as a special case of the parametrization (5). This is obvious, because the inverse transformation to (3) gives

$$\alpha:\beta:\gamma:\delta=Y-X:Z-X:U-X:V-X,$$

and hence a solution $(X_0,Y_0,Z_0,U_0,V_0,W_0)$ to (1) arises from the parametrization (5) using the parameters

$$(\alpha,\beta,\gamma,\delta)=(Y_0-X_0,\ Z_0-X_0,\ U_0-X_0,\ V_0-X_0).$$

To sum up, the parametrization (5) provides a complete description of all the solutions to Eqs. (1).

III. A SIMPLIFIED PARAMETRIZATION

By means of the transformation (2), one now recovers a complete description of all nontrivial zeros of weight-1 6j coefficients, as follows.

From (2) and (5), we obtain the parametrization of a zero coefficient as $\left\{\begin{smallmatrix}a&b&e\\d&c&f\end{smallmatrix}\right\}$, where

$$\begin{aligned}
a&=\tfrac{1}{2}[\alpha^2(-\beta+\gamma+\delta)+\alpha(-\beta^2+\beta(3\gamma+2\delta)-(2\gamma+\delta)(\gamma+\delta))\\
&\quad+\beta^2(\gamma+\delta)-\beta(2\gamma+\delta)(\gamma+\delta)+\gamma(\gamma^2+2\gamma\delta+\delta^2)-1],\\
b&=\tfrac{1}{2}[\alpha(-\beta^2+\beta(\gamma+2\delta)-\delta(2\gamma+\delta))+\beta^2(\gamma+\delta)-\beta(\gamma+\delta)(\gamma+2\delta)+\delta(2\gamma^2+2\gamma\delta+\delta^2)-1],\\
c&=\tfrac{1}{2}[\alpha^2(\beta-\gamma-\delta)+\alpha(-\beta(\gamma+2\delta)+(\gamma^2+2\gamma\delta+2\delta^2))+(\beta\delta(\gamma+\delta)-\delta(\gamma^2+\gamma\delta+\delta^2))],\\
d&=\tfrac{1}{2}[\alpha\gamma(-\beta+\gamma+\delta)+\beta\gamma(\gamma+\delta)-\gamma(\gamma^2+\gamma\delta+\delta^2)],\\
e&=\tfrac{1}{2}[\alpha^2(\beta-\gamma-\delta)+\alpha(-2\beta(\gamma+\delta)+(2\gamma^2+3\gamma\delta+2\delta^2))+\beta(\gamma+\delta)^2-(\gamma+\delta)(\gamma^2+\gamma\delta+\delta^2)-2],\\
f&=\tfrac{1}{2}[\alpha(\beta^2-2\beta(\gamma+\delta)+(\gamma^2+\gamma\delta+\delta^2))+(-\beta^2(\gamma+\delta)+2\beta(\gamma^2+\gamma\delta+\delta^2)-(\gamma+\delta)(\gamma^2+\delta^2))-1].
\end{aligned} \tag{6}$$

Now the triangle conditions on (a,b,e), (a,c,f), (b,d,f), (c,d,e) reduce to the inequalities in terms of X,Y,Z,U,V,W given by

$$W-Y>0,\quad U+V>0,\quad W-X>0,\quad U-Z>0,\quad X+Y>0,\quad V-Z>0.$$

These in turn give the inequalities in terms of $\alpha,\beta,\gamma,\delta$:

$$(-\beta+\gamma+\delta)(\alpha(-\beta+\gamma+\delta)+\beta(\gamma+\delta)-(\gamma^2+\gamma\delta+\delta^2))>0, \tag{7}$$
$$(-\alpha-\beta+\gamma+\delta)(\alpha(-\beta+\gamma+\delta)+\beta(\gamma+\delta)-(\gamma^2+\gamma\delta+\delta^2))>0, \tag{8}$$

$$(\beta-\gamma)(\beta-\delta)(-\alpha+\gamma+\delta)>0, \tag{9}$$

$$(-\alpha-\beta+\gamma+\delta)(-\beta+\gamma+\delta)(-\alpha+\gamma+\delta)>0, \tag{10}$$

$$(\beta-\gamma)(\alpha(-\beta+\gamma+\delta)+\beta(\gamma+\delta)-(\gamma^2+\gamma\delta+\delta^2))>0, \tag{11}$$

$$(\beta-\delta)(\alpha(-\beta+\gamma+\delta)+\beta(\gamma+\delta)-(\gamma^2+\gamma\delta+\delta^2))>0. \tag{12}$$

If we suppose

$$\alpha(-\beta+\gamma+\delta)+\beta(\gamma+\delta)-(\gamma^2+\gamma\delta+\delta^2)<0,$$

then it follows from (7), (8), (11), and (12) that

$$-\beta+\gamma+\delta<0, \tag{13}$$

$$-\alpha-\beta+\gamma+\delta<0, \tag{14}$$

$$\beta-\gamma<0, \tag{15}$$

$$\beta-\delta<0, \tag{16}$$

whence from (9) and (10) that

$$\alpha-\gamma-\delta<0. \tag{17}$$

But (14) and (17) imply $\beta>0$ whereas (13), (15), and (16) imply $\beta<0$, a contradiction.

Accordingly, we must have

$$\alpha(-\beta+\gamma+\delta)+\beta(\gamma+\delta)-(\gamma^2+\gamma\delta+\delta^2)>0, \tag{18}$$

with

$$-\beta+\gamma+\delta>0, \tag{19}$$

$$-\alpha-\beta+\gamma+\delta>0, \tag{20}$$

$$\beta-\gamma>0, \tag{21}$$

$$\beta-\delta>0, \tag{22}$$

$$-\alpha+\gamma+\delta>0. \tag{23}$$

Now (19), (21), and (22) imply that

$$\beta=p+q+r,\quad \gamma=q+r,\quad \delta=p+r, \tag{24}$$

for $p, q, r>0$.

Then (20) gives

$$\alpha=r-s,\quad s>0, \tag{25}$$

and (23) is automatically satisfied.

Further, condition (18) becomes, on using (24) and (25),

$$pq>rs. \tag{26}$$

Thus we may rewrite the parametrization (6) in terms of the (positive) parameters p,q,r,s as follows:

$$\begin{aligned} a&=\tfrac{1}{2}[pq(p+r)+rs(q+s)-1],\\ b&=\tfrac{1}{2}[pq(q+s)+rs(p+r)-1],\\ c&=\tfrac{1}{2}(p+s)(pq-rs),\\ d&=\tfrac{1}{2}(q+r)(pq-rs),\\ e&=\tfrac{1}{2}[(p+q+r+s)(pq-rs)-2],\\ f&=\tfrac{1}{2}[pq(r+s)+rs(p+q)-1]. \end{aligned} \tag{27}$$

Then, subject to the inequality (26), the $6j$ coefficient with arguments a,b,c,d,e,f given by (27) is a zero coefficient of weight 1. Conversely, any nontrivial zero of a $6j$ coefficient of weight 1 arises as a particular case of the parametrized form at (27); indeed, the parameters giving rise to a zero weight-1 $6j$ coefficient $\left\{\begin{smallmatrix} a & b & e \\ d & c & f \end{smallmatrix}\right\}$ are given by

$$(p,q,r,s) = (-a-c+f,-b-d+f,b-d-f,a-c-f).$$

We remark that Eqs. (27) admit the relations

$$\begin{aligned} a+c-f&=p(pq-rs),\\ b+d-f&=q(pq-rs),\\ -b+d+f&=r(pq-rs),\\ -a+c+f&=s(pq-rs),\\ a+b+c+d+1&=(p+q+r+s)pq,\\ a+b-e&=(p+r+s)rs, \end{aligned}$$

giving the precise connection between a,b,c,d,e,f and p,q,r,s.

For numerical examples, we may, for instance, specify $q=s=1, p=r+1$. Then

$$\begin{aligned} a&=\tfrac{1}{2}r(2r+5),\quad d=\tfrac{1}{2}(r+1),\\ b&=\tfrac{1}{2}(r+1)(2r+1),\quad e=\tfrac{1}{2}(2r+1),\\ c&=\tfrac{1}{2}(r+2),\quad f=r(r+2), \end{aligned}$$

where r is an arbitrary integer.

Cases $r=1,2,3,4$ give the zero coefficients

$$\begin{Bmatrix} \frac{7}{2} & 3 & \frac{3}{2} \\ 1 & \frac{3}{2} & 3 \end{Bmatrix},\quad \begin{Bmatrix} 9 & \frac{15}{2} & \frac{5}{2} \\ \frac{3}{2} & 2 & 8 \end{Bmatrix},$$

$$\begin{Bmatrix} \frac{33}{2} & 14 & \frac{7}{2} \\ 2 & \frac{5}{2} & 15 \end{Bmatrix},\quad \begin{Bmatrix} 26 & \frac{45}{2} & \frac{9}{2} \\ \frac{5}{2} & 3 & 24 \end{Bmatrix}.$$

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

[2] S. Brudno, "Nontrivial zeros of the Wigner (3-*j*) and Racah (6-*j*) coefficients. I. Linear solutions," J. Math. Phys. **26**, 434 (1985).

[3] G. Racah, "Theory of complex spectra. II," Phys. Rev. **62**, 438 (1942).

[4] S. Brudno and J. D. Louck, "Nontrivial zeros of the Racah quadruple invariant," J. Math. Phys. **26**, 1125 (1985).

[5] V. Bargmann, "On the representation of the rotation group," Rev. Mod. Phys. **34**, 829 (1962).

[6] S. Brudno and J. D. Louck, Nontrivial zeros of weight 1 $3j$ and $6j$ coefficients: Relation to Diophantine equations of equal sums of like powers," J. Math. Phys. **26**, 2092 (1985).

[7] A. Lindner, "Non-trivial zeros of the Wigner ($3j$) and Racah ($6j$) coefficients," J. Phys. A: Math. Gen. **18**, 3071 (1985).

[8] A. Lindner, *Drehimpulse in der Quantenmechanik* (Teubner, Stuttgart, 1984).

[9] L. E. Dickson, *Theory of Numbers* (Chelsea, New York, 1952), Vol. 2, reprint.

[10] A. Bremner, "On Diophantine equations and nontrivial Racah coefficients," J. Math. Phys. **27**, 1181 (1986).

Algebraic expressions for some multiplicity-free 6j symbols and isoscalar factors for G_2

B. R. Judd
Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, Maryland 21218

(Received 19 December 1985; accepted for publication 16 July 1986)

Generalized 6j symbols for G_2 in which all four triangular conditions involve the seven-dimensional irreducible representation (irrep) (10) are multiplicity-free. Algebraic expressions depending on the highest weights of the accompanying irreps are found by using generalizations of the Racah-Wigner algebra. A rule is given for generalizing the SO(3) phase factors. The results are applied to finding algebraic expressions for a class of isoscalar factors for $SO(7) \supset G_2$.

I. INTRODUCTION

It has been realized for many years that the apparatus of the quantum theory of angular momentum, as represented by the Racah–Wigner algebra for the group SO(3), can be very extensively generalized to other groups. An appreciation of this fact is apparent in the early work of Wigner,[1] but it is only within the last 20 years or so that specific examples have been described. The generalizations to finite groups were initiated by Griffith,[2] and a substantial tabulation of the 3j and 6j symbols for the crystallographic point groups has been recently provided by the book of Butler.[3] Hecht[4] developed the Racah–Wigner algebra of SO(5) for applications in nuclear physics, while a few special cases of generalized 6j symbols for SO(5) have been listed in connection with the Jahn–Teller effect.[5,6] The unitary groups have been the subject of many articles because of their relevance to particle physics as well as nuclear physics. References in those areas can be found in the work of Draayer and Akiyama[7] and of Haase and Butler.[8]

A major problem facing any extension of the Racah–Wigner algebra from SO(3) is the appearance of multiplicity difficulties. A given irreducible representation (irrep) Γ of a group G may occur more than once in the decomposition of the Kronecker product $\Gamma' \times \Gamma''$. This enormously complicates the calculation of general expressions for the 6j symbols for the group G. What is more, all isoscalar factors (that is, factored parts of generalized Clebsch–Gordan coefficients) involve irreps of both G and a subgroup H of G, and it can well happen that a given Γ contains a particular irrep of H more than once. The attention that has been paid to coping with these difficulties has tended to obscure the fact that many types of problems are multiplicity-free or almost so. In the limited regions where multiplicity complications are not encountered we might expect to be able to develop formulas for 3j and 6j symbols as well as for their higher nj forms. The absence of multiplicity difficulties enables us to give definitions of the 6j symbols in terms of recoupling coefficients that exactly parallel the corresponding definitions for SO(3). That is, the equation

$$\begin{Bmatrix} j_1 & j_2 & j_3 \\ j_4 & j_5 & j_6 \end{Bmatrix} = (-1)^{j_1+j_2+j_4+j_5} [(2j_3+1)(2j_6+1)]^{-1/2} \times ((j_1 j_2) j_3, j_4, j_5 | j_1 (j_2 j_4) j_6, j_5)$$

is replaced by

$$\begin{Bmatrix} \Gamma_1 & \Gamma_2 & \Gamma_3 \\ \Gamma_4 & \Gamma_5 & \Gamma_6 \end{Bmatrix} = (-1)^{\varphi(\Gamma_1)+\varphi(\Gamma_2)+\varphi(\Gamma_4)+\varphi(\Gamma_5)} [D(\Gamma_3) D(\Gamma_6)]^{-1/2} \times ((\Gamma_1 \Gamma_2)\Gamma_3, \Gamma_4, \Gamma_5 | \Gamma_1 (\Gamma_2 \Gamma_4)\Gamma_6, \Gamma_5),$$

where $D(\Gamma)$ is the dimension of the irrep Γ of G. Only the form of the phase factors $(-1)^{\varphi(\Gamma)}$ remains to be settled. Of course, the recoupling coefficient itself conceals phase choices that are implicit whenever a sequence of coupled irreps is written down. These can be determined only for specific bases, a task that is distinct from the issues facing us at the moment.

Some work in generalizing the SO(3) 6j symbols to other groups G has been recently carried out for the groups SO(n) (where $n > 3$).[9] The stimulus to do so was provided by the Jahn–Teller effect for an electronic state coupled to two vibrational modes, each belonging to the irrep (10) of SO(5). (Here and throughout this article we label an irrep by its highest weight.) Excitations of a particular mode only involve irreps of the type $(w0)$, and these are particularly easy to handle. By evaluating matrix elements of selected tensor operators in a boson basis and relating the results to the standard Racah–Wigner formalism in which a 6j symbol for SO(n) appears, it proved possible to obtain explicit expressions for such multiplicity-free 6j symbols as

$$\begin{Bmatrix} (w0\cdots0) & (20\cdots0) & (w+1,10\cdots0) \\ (10\cdots0) & (w+1,0\cdots0) & (210\cdots0) \end{Bmatrix}$$

as a function of w.

To demonstrate that the method can be generalized to other groups, formulas were given[9] for all G_2 6j symbols of the type

$$\begin{Bmatrix} (w0) & (10) & U' \\ (10) & (w0) & U'' \end{Bmatrix}, \tag{1.1}$$

where U' and U'' of G_2 range over the five acceptable irreps $(w+1,0)$, $(w1)$, $(w0)$, $(w-1,1)$, and $(w-1,0)$. We follow Racah[10] in using an acute-angled coordinate system to specify the highest weight $(w_1 w_2)$; the connection to the labeling scheme $(a_1 a_2)$ of McKay and Patera[11] is given by

 0022-2488/86/112616-07$02.50

TABLE I. Highest weights (w_1w_2), Dynkin labels (a_1a_2), dimensions $D(U)$, eigenvalues $g(U)$ of Casimir's operator G for G_2, and irreps $\mathscr{D}_L$ of SO(3) for the embedding $(10)\rightarrow\mathscr{D}_3$.

$(w_1\omega_2)$	(a_1a_2)	$D(U)$	$12g(U)$	L
(00)	(00)	1	0	*S*
(10)	(01)	7	6	*F*
(11)	(10)	14	12	*PH*
(20)	(02)	27	14	*DGI*
(21)	(11)	64	21	*DFGHKL*
(30)	(03)	77	24	*PFGHIKM*
(22)	(20)	77	30	*SDGHILN*

$w_1 = a_1 + a_2, w_2 = a_1$. In order to make our notation clear, a few examples are given in Table I. Having tabulated, then, the 25 algebraic expressions (1.1), we are led to ask whether other formulas can be found. There are grounds for optimism because, for the general irrep (w_1w_2) of G_2, the Kronecker product $(w_1w_2)\times(10)$ decomposes according to

$$\begin{aligned}(w_1w_2)\times(10) &= (w_1+1,w_2)+(w_1+1,w_2-1)\\ &\quad+(w_1,w_2+1)+(w_1w_2)+(w_1,w_2-1)\\ &\quad+(w_1-1,w_2+1)+(w_1-1,w_2)\,,\end{aligned}\tag{1.2}$$

a direct sum in which no given irrep appears more than once. Thus all G_2 $6j$ symbols of the types

$$\begin{Bmatrix}(w_1w_2) & (10) & (w_5w_6)\\ (10) & (w_3w_4) & (10)\end{Bmatrix},\quad \begin{Bmatrix}(w_1w_2) & (10) & (w_5w_6)\\ (w_3w_4) & (10) & (w_7w_8)\end{Bmatrix}\tag{1.3}$$

are multiplicity-free and should therefore be susceptible of algebraic evaluation. The products $(w_1w_2)\times(11)$ contain (w_1w_2) twice in the general case, and so the replacing of any irrep (10) by (11) (or by any irrep of higher dimensionality for that matter) opens the way for multiplicity ambiguities. In spite of this limitation, it should be of considerable interest to study those $6j$ symbols of the type (1.3), since (10) is the first nontrivial irrep of G_2, as can be seen from Table I. It is the aim of the present paper to make a start on that program and to show how our knowledge of some G_2 $6j$ symbols makes it possible to calculate general expressions for some isoscalar factors for $SO(7)\supset G_2$.

II. PHASE

In the absence of multiplicity complications we can take the formulas of angular-momentum theory for SO(3) as given, for example, by Edmonds[12] and simply replace every $6j$ symbol by the corresponding generalization to G_2. That is, the six j's are replaced by six U's. The dimensional factors $2j+1$ that appear in the formulas are replaced by $D(U)$, where

$$\begin{aligned}D(U) = \tfrac{1}{120}\,&(w_1+w_2+3)\,(w_1+2)\,(2w_1+w_2+5)\\ &\times(w_1+2w_2+4)\,(w_1-w_2+1)\,(w_2+1)\,.\end{aligned}\tag{2.1}$$

Reduced matrix elements of a SO(3) tensor $\mathbf{T}^{(k)}$ become reduced matrix elements of a G_2 tensor of the type $\mathbf{T}^{(U)}$. Only one problem remains: how do we find the analog of $(-1)^j$? We want our $6j$ symbols for G_2 to exhibit all the symmetries of their SO(3) counterparts—a phase-free interchange of columns and an equally phase-free interchange of pairs of arguments in the upper and lower rows of the $6j$ symbol. Studies of the properties of generalized $6j$ symbols with multiplicity labels attached indicate that the general phase problem is highly complex.[3,13–15] It is natural to hope that we can escape the morass of that analysis, or at least evade its more intricate aspects. After all, we are making two specializations: one to G_2 and a second to multiplicity-free Kronecker products.

Wigner's celebrated analysis[16] of simply reducible groups is of value here. Irreps $\mathscr{D}_j$ of SO(3) are classified as even or odd according to whether they occur in the symmetric or antisymmetric parts of Kronecker squares $\mathscr{D}_k\times\mathscr{D}_k$, where k is integral. This evenness or oddness is represented by the phase factor $(-1)^j$. Because G_2 is not simply reducible a given irrep U' sometimes occurs in both the symmetric and the antisymmetric parts of U^2; however, a glance at the table of special cases[17] indicates that every irrep (10) and every irrep (11) occur in the antisymmetric part of U^2, while (00) and every irrep (20) occur in its symmetric part. To see why this might be true in general, consider the embedding $G_2\supset SO(3)$ for which $(10)\rightarrow\mathscr{D}_3$. Suppose that we introduce the commuting SO(3) tensors $\mathbf{T}_A^{(k)}$ and $\mathbf{T}_B^{(k)}$ that act in identical but distinguishable spaces A and B. We can write

$$\begin{aligned}&(\mathbf{T}_A^{(k)}\mathbf{T}_B^{(k)})^{(L)}|(U_AU_B)(00)S\rangle\\ &\quad=\sum\Xi(U',U'',\tau U_i)|(U'_AU''_B)\tau U_iL\rangle\,,\end{aligned}\tag{2.2}$$

where the second ket indicates that U' and U'' are coupled to those various irreps U_i that contain $\mathscr{D}_L$ of SO(3). The sum in Eq. (2.2) runs over U', U'', U_i, and the multiplicity label τ. Now, we have

$$(\mathbf{T}_A^{(k)}\mathbf{T}_B^{(k)})^{(L)} = (-1)^L(\mathbf{T}_B^{(k)}\mathbf{T}_A^{(k)})^{(L)}\,,\tag{2.3}$$

$$|(U'_AU'_B)\tau U_iL\rangle = (-1)^p|(U'_BU'_A)\tau U_iL\rangle\,,\tag{2.4}$$

where $p=0$ or 1 according to whether τU_i occurs in the symmetric or antisymmetric part of $U'\times U'$. Since (00) necessarily occurs in the symmetric part of U^2,

$$|(U_AU_B)(00)S\rangle = |(U_BU_A)(00)S\rangle\,.\tag{2.5}$$

The replacements (2.3)–(2.5) are made in (2.2). Next, we exchange the labels A and B. For the original equation (2.2) to be recovered we must have

$$\Xi(U',U',\tau U_i) = (-1)^{L+p}\Xi(U',U',\tau U_i)\,.\tag{2.6}$$

Thus the only kets $|(U'_AU'_B)\tau U_iL\rangle$ that can be generated by equations of the type (2.2) correspond to even $L+p$. Almost all irreps U_i of G_2 contain irreps $\mathscr{D}_L$ of SO(3) for which L runs over both odd and even values. These U_i cannot be assigned a unique p by the present method and they may be found in either the symmetric or antisymmetric parts of Kronecker squares (or both). However, we see from Table I that L is always even for (00) and (20), and it is always odd for (10) and (11). Thus p follows L in being even for the former and odd for the latter. Although our argument depends on the assumption that a state $|(U'_AU'_B)\tau U_iL\rangle$ can be generated from an S state by means of an SO(3) tensor of rank L, we have complete freedom in choosing U and $\mathbf{T}^{(k)}$.

The latter, for example, may be the superposition of many of the irreps of G_2 that contain $\mathscr{D}_k$. It is thus highly plausible that $p=0$ for (00) and (20), and that $p=1$ for (10) and (11).

If, in addition, we suppose that the phase $(-1)^{\varphi(U)}$ to be associated with the irrep U [in analogy to $(-1)^j$ for $\mathscr{D}_j$] is of such a form that we can write $\varphi(w_1w_2)=aw_1+bw_2$, then our results for (00), (20), (10), and (11) limit our choice for a and b very severely. In fact, a must be an odd integer and b an even one. Without loss of generality we can take $a=1$ and $b=0$. Thus the phase associated with (w_1w_2) becomes simply $(-1)^{w_1}$.

III. FORMULAS FOR G_2 6j SYMBOLS

The starting point for the construction of algebraic expressions for the 6j symbols (1.3) is the initial five-by-five block of entries for (1.1).[9] To match the phase procedure outlined in Sec. II, it is necessary to set the single undefined phase ε of that table equal to -1. The Racah backcoupling relation [Eq. (6.2.11) of Edmonds[12]] immediately provides expressions for

$$\begin{Bmatrix}(w0) & (10) & U\\ (w0) & (10) & U'\end{Bmatrix}.$$

The Biedenharn–Elliott identity [Eq. (6.2.12) of Edmonds[12]] can now be brought into play. Gaps in the tables can be filled by using the orthonormality relation for 6j symbols [Eq. (6.2.9) of Edmonds[12]]. To reduce the complexity of the algebra, we define a U coefficient similar to that of Jahn[18]:

$$\mathrm{U}\begin{pmatrix}U_1 & U_2 & U_3\\ U_4 & U_5 & U_6\end{pmatrix}=[D(U_3)D(U_6)]^{1/2}\begin{Bmatrix}U_1 & U_2 & U_3\\ U_4 & U_5 & U_6\end{Bmatrix}. \tag{3.1}$$

The introduction of a second meaning for the symbol U should not cause any difficulty since it necessarily precedes a large and visually characteristic parenthesis. The third column of the U coefficient can be interchanged with either of the others on a phase-free basis, but dimensional factors need to be included. For example,

$$\mathrm{U}\begin{pmatrix}U_1 & U_3 & U_2\\ U_4 & U_6 & U_5\end{pmatrix}=\left[\frac{D(U_2)D(U_5)}{D(U_3)D(U_6)}\right]^{1/2}\mathrm{U}\begin{pmatrix}U_1 & U_2 & U_3\\ U_4 & U_5 & U_6\end{pmatrix}. \tag{3.2}$$

Formulas for U coefficients corresponding to the forms (1.3) are given in Tables II and III. The replacement $u=2w+5$ is used to simplify the tabulation. In spite of our procedure for interpreting the phase factors in the standard formulas of the Racah–Wigner algebra, the Biedenharn–Elliott identity often gives only the squares of new U coefficients. Phase factors ε_i are introduced for such occasions and retained in the analysis until freedom to make a specific (though frequently arbitrary) choice is clear.

Tables II and III are limited to reasonably elementary examples. As we proceed to more complicated cases the algebra becomes correspondingly intricate. For example, we find (for $w>1$)

$$\mathrm{U}\begin{pmatrix}(w1) & (10) & (w1)\\ (w1) & (10) & (w1)\end{pmatrix}=\frac{u^6+6u^5-102u^4-448u^3-159u^2+594u-2916}{u(u+2)\,(u-3)\,(u+5)\,(u-7)\,(u+9)}. \tag{3.3}$$

Some compression is possible here. Just as the SO(3) U coefficient

$$\mathrm{U}\begin{pmatrix}L & 1 & L\\ L & 1 & L\end{pmatrix}$$

can be expressed as a function of $L(L+1)$, so Eq. (3.3) can be simplified by introducing Casimir's operator G for G_2. Its eigenvalues g for $(w1)$ are given by[10]

$$g=(u^2+2u-15)/48,$$

and the right-hand side of Eq. (3.3) simplifies to

$$(16g^3-23g^2-17g-3)/g(g-1)\,(16g+5).$$

In spite of the infinite sequences of 6j symbols provided by every line of Tables II and III, an accidental vanishing is rare. The only examples that we have noticed are

$$\begin{Bmatrix}(10) & (10) & (11)\\ (10) & (10) & (11)\end{Bmatrix} \quad\text{and}\quad \begin{Bmatrix}(60) & (10) & (51)\\ (60) & (10) & (51)\end{Bmatrix}.$$

TABLE II. Formulas for $\mathrm{U}\begin{pmatrix}U_1 & (10) & U_3\\ (10) & U_2 & (10)\end{pmatrix}$.

$U_1U_2U_3$	U coefficient
$(w0)\ (w0)\ (w+1,0)$	$[(u+2)\,(u-5)/6u(u-3)]^{1/2}$
$(w0)\ (w0)\ (w1)$	$(u-1)\,[(u+7)/3u(u-3)\,(u+5)]^{1/2}$
$(w0)\ (w0)\ (w0)$	$-[6/(u^2-25)]^{1/2}$
$(w0)\ (w0)\ (w-1,1)$	$(u+1)\,[(u-7)/3u(u+3)\,(u-5)]^{1/2}$
$(w0)\ (w0)\ (w-1,0)$	$-[(u-2)\,(u+5)/6u(u+3)]^{1/2}$
$(w0)\ (w1)\ (w1)$	$-[(u+2)\,(u-7)\,(u+9)/6u(u-3)\,(u+5)]^{1/2}$
$(w0)\ (w1)\ (w-1,1)$	$-[(u-3)\,(u-7)/3u(u-5)]^{1/2}$
$(w0)\ (w-1,1)\ (w-1,1)$	$[(u-2)\,(u+7)\,(u-9)/6u(u+3)\,(u-5)]^{1/2}$
$(w0)\ (w-1,1)\ (w-1,0)$	$-[(u-2)/3(u+3)]^{1/2}$
$(w1)\ (w1)\ (w+1,1)$	$(u+5)[u(u-7)/6(u-1)\,(u+2)\,(u-5)\,(u+7)]^{1/2}$
$(w1)\ (w1)\ (w+1,0)$	$-[u(u-7)\,(u+9)/12(u-3)\,(u-5)\,(u+7)]^{1/2}$
$(w1)\ (w1)\ (w2)$	$(u+5)\,[u(u-3)\,(u+11)/4(u+1)\,(u+3)\,(u-5)\,(u+7)\,(u+9)]^{1/2}$
$(w1)\ (w1)\ (w1)$	$-(u^2+2u+9)\,[6/u(u+2)\,(u-3)\,(u+5)\,(u-7)\,(u+9)]^{1/2}$
$(w1)\ (w1)\ (w-1,1)$	$-(u-3)\,[(u+2)\,(u+9)/6u(u+3)\,(u-5)\,(u+7)]^{1/2}$
$(w1)\ (w1)\ (w-1,2)$	$(u-3)\,[(u+2)\,(u+5)\,(u-9)/4(u^2-1)\,(u-5)\,(u^2-49)]^{1/2}$
$(w1)\ (w-1,1)\ (w-1,2)$	$-[(u^2-81)/2(u^2-49)]^{1/2}$

TABLE III. Formulas for $\mathrm{U}\begin{pmatrix} U_1 & (10) & U_2 \\ U_3 & (10) & U_4 \end{pmatrix}$.

$U_1U_3;U_2U_4$	U coefficient
$(w0)\ (w0);\ (w+1,0)\ (w+1,0)$	$10/u(u-3)$
$(w0)\ (w0);\ (w+1,0)\ (w1)$	$[8(u+2)\ (u-5)\ (u+7)/u^2(u-3)^2(u+5)]^{1/2}$
$(w0)\ (w0);\ (w+1,0)\ (w0)$	$[4(u+2)/u(u-3)\ (u+5)]^{1/2}$
$(w0)\ (w0);\ (w+1,0)\ (w-1,1)$	$[8(u+2)\ (u-7)/u^2(u^2-9)]^{1/2}$
$(w0)\ (w0);\ (w+1,0)\ (w-1,0)$	$[(u^2-4)\ (u^2-25)/u^2(u^2-9)]^{1/2}$
$(w0)\ (w0);\ (w1)\ (w1)$	$(u-7)\ (u+17)/u(u-3)\ (u+5)$
$(w0)\ (w0);\ (w1)\ (w0)$	$[32(u-4)^2(u+7)/u(u-3)\ (u-5)\ (u+5)^2]^{1/2}$
$(w0)\ (w0);\ (w1)\ (w-1,1)$	$(u^2-17)\ [(u^2-49)/u^2(u^2-9)\ (u^2-25)]^{1/2}$
$(w0)\ (w0);\ (w1)\ (w-1,0)$	$-[8(u-2)\ (u+7)/u^2(u^2-9)]^{1/2}$
$(w0)\ (w0);\ (w0)\ (w0)$	$(u^2-61)/(u^2-25)$
$(w0)\ (w0);\ (w0)\ (w-1,1)$	$-[32(u+4)^2(u-7)/u(u+3)\ (u+5)\ (u-5)^2]^{1/2}$
$(w0)\ (w0);\ (w0)\ (w-1,0)$	$-[4(u-2)/u(u+3)\ (u-5)]^{1/2}$
$(w0)\ (w0);\ (w-1,1)\ (w-1,1)$	$-(u+7)\ (u-17)/u(u+3)\ (u-5)$
$(w0)\ (w0);\ (w-1,1)\ (w-1,0)$	$-[8(u-2)\ (u+5)\ (u-7)/u^2(u+3)^2(u-5)]^{1/2}$
$(w0)\ (w0);\ (w-1,0)\ (w-1,0)$	$10/u(u+3)$
$(w0)\ (w+1,0);\ (w+1,0)\ (w1)$	$[16(u-5)/(u-3)^2(u+5)]^{1/2}$
$(w0)\ (w+1,0);\ (w+1,0)\ (w0)$	$-[(u-5)\ (u+7)/(u-3)\ (u+5)]^{1/2}$
$(w0)\ (w+1,0);\ (w1)\ (w1)$	$-(u^2+2u-31)/(u-3)\ (u+5)$
$(w0)\ (w+1,0);\ (w1)\ (w0)$	$-[16(u+7)/(u-3)\ (u+5)^2]^{1/2}$
$(w0)\ (w1);\ (w+1,0)\ (w1)$	$[2(u-7)\ (u+9)/u(u-3)^2(u+5)]^{1/2}$
$(w0)\ (w1);\ (w+1,0)\ (w-1,1)$	$-[(u+2)\ (u-7)/u(u-5)]^{1/2}$
$(w0)\ (w1);\ (w1)\ (w1)$	$-4(u^2-4u-9)/u(u-3)\ (u+5)\quad (w>1)$
$(w0)\ (w1);\ (w1)\ (w0)$	$[(u-1)^2(u+2)\ (u-7)\ (u+9)/u(u-3)\ (u+5)^2(u-5)]^{1/2}$
$(w0)\ (w1);\ (w1)\ (w-1,1)$	$-[8(u+2)\ (u+9)/u^2(u^2-25)]^{1/2}$
$(w0)\ (w1);\ (w0)\ (w-1,1)$	$[2(u-3)\ (u-7)/u(u-5)^2(u+5)]^{1/2}$
$(w0)\ (w1);\ (w-1,1)\ (w-1,1)$	$2(u-6)/u(u-5)$
$(w0)\ (w-1,1);\ (w-1,0)\ (w-1,1)$	$[2(u+7)\ (u-9)/u(u+3)^2(u-5)]^{1/2}$
$(w0)\ (w-1,1);\ (w-1,0)\ (w0)$	$[8(u-2)/(u+3)\ (u^2-25)]^{1/2}$
$(w0)\ (w-1,1);\ (w-1,0)\ (w1)$	$[(u-2)\ (u+7)/u(u+5)]^{1/2}$
$(w0)\ (w-1,1);\ (w-1,1)\ (w-1,1)$	$4(u^2+4u-9)/u(u+3)\ (u-5)\quad (w>2)$
$(w0)\ (w-1,1);\ (w-1,1)\ (w0)$	$-[(u+1)^2(u-2)\ (u+7)\ (u-9)/u(u+3)\ (u+5)\ (u-5)^2]^{1/2}$
$(w0)\ (w-1,1);\ (w-1,1)\ (w1)$	$[8(u-2)\ (u-9)/u^2(u^2-25)]^{1/2}$
$(w0)\ (w-1,1);\ (w0)\ (w1)$	$[2(u+3)\ (u+7)u(u-5)\ (u+5)^2]^{1/2}$
$(w0)\ (w-1,1);\ (w1)\ (w1)$	$-2(u+6)/u(u+5)$
$(w1)\ (w-1,1);\ (w1)\ (w1)$	$-2(u^2+2u-27)/u(u+5)\ (u-7)$
$(w1)\ (w-1,1);\ (w-1,2)\ (w1)$	$[12(u+2)\ (u-3)\ (u-9)/u(u+5)\ (u-7)^2(u+7)]^{1/2}$
$(w1)\ (w-1,1);\ (w-1,2)\ (w0)$	$[3(u^2-9)\ (u^2-81)/4(u^2-25)\ (u^2-49)]^{1/2}$
$(w1)\ (w-1,1);\ (w-1,2)\ (w-1,2)$	$-(u^2-73)/2(u^2-49)$
$(w1)\ (w-1,1);\ (w-1,1)\ (w1)$	$-[(u^2-4)\ (u^2-9)\ (u^2-81)/u^2(u^2-25)\ (u^2-49)]^{1/2}$
$(w1)\ (w-1,1);\ (w-1,1)\ (w-1,2)$	$-[12(u-2)\ (u+3)\ (u+9)/u(u-5)(u+7)^2(u-7)]^{1/2}$
$(w1)\ (w-1,1);\ (w-1,1)\ (w-1,1)$	$2(u^2-2u-27)/u(u-5)\ (u+7)$

The zeros implied by such factors as $(u-7)$ and $(u-9)$ in all other formulas in Tables II and III correspond to the violation of a triangular condition. The most common example is

$$\{(10),\ (11),\ (11)\}=0\,.$$

IV. SYMMETRIES OF JUCYS

It is well known that many formulas of angular-momentum theory are invariant (to within a phase factor) with respect to replacements of the type $j\to -j-1$. The characteristic Casimir form $j(j+1)$ is unchanged, while the dimension $2j+1$ of $\mathscr{D}_j$ merely changes sign. In his book with Savukynas, Jucys[19] touches on the analogous property for G_2. Our choice of u rather than w in the algebraic expressions of Tables II and III exposes this symmetry. Under the replacement $u\to -u$, we find, for $(w0)$,

$$g(w0)\to g(w0),\quad D(w0)\to -D(w0)\,.$$

Other irreps (w_1w_2) appear to become interchanged. For example,

$$g(w1)\to g(w-1,1),\quad D(w1)\to -D(w-1,1)\,.$$

These interchanges are summarized by

$$\begin{aligned}&(w0)\leftrightarrow(w0),\quad (w+1,0)\leftrightarrow(w-1,0)\,,\\ &(w1)\leftrightarrow(w-1,1),\quad (w2)\leftrightarrow(w-2,2)\,,\\ &(w+1,1)\leftrightarrow(w-2,1),\quad (w-1,2)\leftrightarrow(w-1,2)\,,\end{aligned}\tag{4.1}$$

and tally with the substitutions (18.14) of Jucys and Savukynas.[19] It is easy to verify that the formulas of Tables II and III are consistent (to within phase factors) with the simultaneous substitutions $u\to -u$ and (4.1). Since all the formulas of Tables II and III were calculated separately, this procedure provides excellent checks on the magnitudes of the U coefficients.

V. ISOSCALAR FACTORS FOR SO(7)⊃G$_2$

As an example of the usefulness of Tables II and III, we show how certain isoscalar factors can be calculated for $\mathrm{SO}(7)\supset \mathrm{G}_2$. Our method is the analog of one that can be used to find the Clebsch–Gordan (CG) coefficients for $\mathrm{SO}(3)\supset\mathrm{SO}(2)$. If **S** and **L** are two commuting angular momenta with resultant **J**, the CG coefficients in the sum

$$\sum_{M_S,M_L} (SM_S,LM_L|JM_J)|SM_S,LM_L\rangle \tag{5.1}$$

corresponding to well-defined J and M_J can be found by insisting that (5.1) be an eigenfunction of $2\mathbf{S}\cdot\mathbf{L}$ with eigenvalue $J(J+1)-S(S+1)-L(L+1)$. The calculation is carried out by taking $2\mathbf{S}\cdot\mathbf{L}$ in the form $2S_zL_z+S_+L_-+S_-L_+$ (where $S_+=S_x+iS_y$, etc.), which enables enough linear equations to be set up to evaluate the CG coefficients (to within a phase) and check them.

In the notation of Racah,[10] a multiplicity-free isoscalar factor for $SO(7)\supset G_2$ is written

$$(W_1U_1+W_2U_2|W_3U_3)\,, \tag{5.2}$$

where W_i stands for an irrep of SO(7). We can think of (5.2) as a CG coefficient for $SO(7)\supset SO(2)$ with the CG coefficient for $G_2\supset SO(2)$, namely

$$(U_1\beta_1L_1M_{L1},\,U_2\beta_2L_2M_{L2}|\tau U_3\beta_3L_3M_{L3})\,, \tag{5.3}$$

factored out. If we can evaluate (5.2) directly we can avoid the problem of coping with the multiplicity labels β_i and τ in (5.3). As a particular example we choose $W_1\equiv(w10)$ and $W_2\equiv(100)$. Our program consists in finding algebraic expressions for all possible W_3, U_1, U_2, and U_3 in (5.2).

The 21 generators of SO(7) belong to the irrep (110).[10] To distinguish the two commuting sets (the analogs of $\mathbf{S}$ and $\mathbf{L}$) we write $\mathbf{T}_1^{(110)}$ and $\mathbf{T}_2^{(110)}$. We have

$$\mathbf{T}_3^{(110)}=\mathbf{T}_1^{(110)}+\mathbf{T}_2^{(110)} \tag{5.4}$$

as the analog of $\mathbf{J}=\mathbf{S}+\mathbf{L}$. In a similar way, the 14 generators of G_2, which belong to the irrep (11), can be written as $\mathbf{T}_i^{(11)}$. The part of $\mathbf{T}_i^{(110)}$ not coincident with $\mathbf{T}_i^{(11)}$ belongs to the irrep (10) of G_2, thus forming the tensor $\mathbf{T}_i^{(10)}$. Our tensors are conveniently normalized by means of the equations

$$\begin{aligned}&((100)\|T^{(110)}\|(100))=[D(110)]^{1/2}=\sqrt{(21)}\,,\\ &((10)\|T^{(11)}\|(10))=[D(11)]^{1/2}=\sqrt{(14)}\,,\end{aligned} \tag{5.5}$$

where the first reduced matrix element implies reduction with respect to SO(7), and the second with respect to G_2. The quadratic scalar operators

$$G'=\tfrac{1}{5}\mathbf{T}^{(110)}\cdot\mathbf{T}^{(110)},\quad G=\tfrac{1}{4}\mathbf{T}^{(11)}\cdot\mathbf{T}^{(11)} \tag{5.6}$$

are Casimir's operators for SO(7) and G_2. Their eigenvalues $g'(W)$ and $g(U)$ for the irreps $W=(w_1w_2w_3)$ and $U=(w_1w_2)$ are given by

$$\begin{aligned}g'(W)&=\tfrac{1}{10}[w_1(w_1+5)+w_2(w_2+3)+w_3(w_3+1)]\,,\\ g(U)&=\tfrac{1}{12}[w_1^2+w_2^2+w_1w_2+5w_1+4w_2]\,.\end{aligned} \tag{5.7}$$

To find the analog of $2\mathbf{S}\cdot\mathbf{L}$, we note that we can write

$$\begin{aligned}&2\mathbf{T}_1^{(10)}\cdot\mathbf{T}_2^{(10)}\\ &\quad=(\mathbf{T}_1^{(110)}+\mathbf{T}_2^{(110)})\cdot(\mathbf{T}_1^{(110)}+\mathbf{T}_2^{(110)})-(\mathbf{T}_1^{(110)})^2\\ &\qquad-(\mathbf{T}_2^{(110)})^2-(\mathbf{T}_1^{(11)}+\mathbf{T}_2^{(11)})\cdot(\mathbf{T}_1^{(11)}+\mathbf{T}_2^{(11)})\\ &\qquad+(\mathbf{T}_1^{(11)})^2+(\mathbf{T}_2^{(11)})^2=5G_3'-\Omega\,,\end{aligned} \tag{5.8}$$

where

$$\Omega=5G_1'+5G_2'-4G_3+4G_1+4G_2\,. \tag{5.9}$$

The analog of the sum (5.1), which represents the ket $|JM_J\rangle$, is

$$\begin{aligned}&|(W_1W_2)W_3U_3\rangle\\ &\quad=\sum_{U_1,U_2}(W_1U_1+W_2U_2|W_3U_3)|W_1U_1,W_2U_2,U_3\rangle\,,\end{aligned} \tag{5.10}$$

where the ket on the right indicates that U_1 and U_2 are coupled to U_3. The corresponding coupling in (5.1) is represented by the trivial SO(2) condition $M_J=M_S+M_L$. All we have to do now is demand that $2\mathbf{T}_1^{(10)}\cdot\mathbf{T}_2^{(10)}+\Omega$, when acting on the right-hand side of Eq. (5.10), yields the eigenvalue $5g'(W_3)$. The part Ω is diagonal in the basis $|W_1U_1,W_2U_2,U_3\rangle$, and the central problem is to evaluate

$$\langle W_1U_1,W_2U_2,U_3|\mathbf{T}_1^{(10)}\cdot\mathbf{T}_2^{(10)}|W_1U_1',W_2U_2',U_3\rangle\,. \tag{5.11}$$

It is here that we turn to the familiar formulas of the Racah–Wigner algebra. For the special case of $W_1\equiv(w10)$, $W_2\equiv(100)$, Eq. (7.1.6) of Edmonds[12] reduces (5.11) to

$$\begin{aligned}&(-1)^y\begin{Bmatrix}U_3 & (10) & U_1\\ (10) & U_1' & (10)\end{Bmatrix}((w10)U_1\|T_1^{(10)}\|(w10)U_1')\\ &\quad\times((100)\ (10)\|T_2^{(10)}\|(100)\ (10))\,,\end{aligned} \tag{5.12}$$

where $y=\varphi(U_1')+\varphi(10)+\varphi(U_3)$. This phase is calculated according to the rule given in Sec. II. Since both $\mathbf{T}_1^{(10)}$ and $\mathbf{T}_2^{(10)}$ are the components of the generators for their respective SO(7) groups, the values of their reduced matrix elements are independent of how the irreps $(w10)$ and (100) are constructed. For the former, it is convenient to think of the irrep $(w10)$ as being produced by the coupling of two parts (a and b, say) belonging to $(w00)$ and (100). The tensor $\mathbf{T}_1^{(10)}$ is regarded as $\mathbf{T}_a^{(10)}+\mathbf{T}_b^{(10)}$, and we apply Edmonds' Eqs. (7.1.7) and (7.1.8), which refer to tensors acting on the first part or the second part of a coupled system. Since $(w00)\to(w0)$ under the reduction $SO(7)\to G_2$, we have only one kind of reduced matrix element for G_2 to evaluate, namely

$$((w00)\ (w0)\|T^{(10)}\|(w00)\ (w0))\,. \tag{5.13}$$

Applying the Wigner–Eckart theorem, we see that (5.13) is equal to

$$\begin{aligned}&((w00)\ (w0)|(w00)\ (w0)+(110)\ (10))\\ &\quad\times((w00)\|T^{(110)}\|(w00))\,,\end{aligned} \tag{5.14}$$

where the reduced matrix element in this product is reduced with respect to SO(7). Since $(\mathbf{T}^{(110)})^2$ has eigenvalues $5g(W)$, it is easy to show that

$$(W\|T^{(110)}\|W)=[5D(W)g(W)]^{1/2}\,. \tag{5.15}$$

As for the isoscalar factor in (5.14), the reciprocity relation of Racah[10] gives its value as

$$[D(10)D(w00)/D(w0)D(110)]^{1/2}=(\tfrac{1}{3})^{1/2}$$

to within an arbitrary phase. This phase disappears when the reduced matrix element of $\mathbf{T}_2^{(10)}$ in (5.12) is combined with that of $\mathbf{T}_1^{(10)}$.

All the pieces are now in place to calculate the isoscalar factors of Eq. (5.10). They are set out in Table IV. Every line is associated with an arbitrary phase; each has been chosen to reproduce Racah's phase for the two columns headed $(w0)$ and $(w1)$ when we set $w=2$, corresponding to the special cases listed in his Table IIIa.[10] However, all the en-

TABLE IV. The isoscalar factors $((w10)U_1 + (100)(10)|W_3U_3)$.

W_3	U_3	U_1 $(w-1,1)$	$(w0)$	$(w1)$
$(w-1,10)$	$(w-2,1)$	1	0	0
	$(w-1,0)$	$-\left(\frac{16}{(u+1)(u-5)}\right)^{1/2}$	$\left(\frac{(u+3)(u-7)}{(u+1)(u-5)}\right)^{1/2}$	0
	$(w-1,1)$	$-\left(\frac{2(u+7)(u-9)}{u(u+1)(u^2-25)}\right)^{1/2}$	$\left(\frac{8(u-2)(u+3)}{(u+1)(u-5)(u+5)^2}\right)^{1/2}$	$\left(\frac{(u-2)(u+3)^2(u+7)}{u(u+1)(u+5)^2}\right)^{1/2}$
$(w00)$	$(w0)$	$-\left(\frac{2(u-3)(u-7)}{5u(u-5)}\right)^{1/2}$	$\left(\frac{u^2-9}{5(u^2-25)}\right)^{1/2}$	$\left(\frac{2(u+3)(u+7)}{5u(u+5)}\right)^{1/2}$
$(w11)$	$(w-1,0)$	$\left(\frac{(u+3)(u-7)}{(u+1)(u-5)}\right)^{1/2}$	$\left(\frac{16}{(u+1)(u-5)}\right)^{1/2}$	0
	$(w-1,1)$	$-\left(\frac{(u-2)(u+3)(u-9)}{2u(u+1)(u-5)}\right)^{1/2}$	$\left(\frac{(u-1)^2(u+7)}{2(u+1)(u^2-25)}\right)^{1/2}$	$-\left(\frac{9(u+3)}{u(u+1)(u+5)}\right)^{1/2}$
	$(w-1,2)$	$\left(\frac{u-1}{2u}\right)^{1/2}$	0	$\left(\frac{u+1}{2u}\right)^{1/2}$
	$(w0)$	$-\left(\frac{(u+3)(u-7)}{2u(u-5)}\right)^{1/2}$	$\left(\frac{16}{u^2-25}\right)^{1/2}$	$-\left(\frac{(u-3)(u+7)}{2u(u+5)}\right)^{1/2}$
	$(w1)$	$-\left(\frac{9(u-3)}{u(u-1)(u-5)}\right)^{1/2}$	$-\left(\frac{(u+1)^2(u-7)}{2(u-1)(u^2-25)}\right)^{1/2}$	$\left(\frac{(u+2)(u-3)(u+9)}{2u(u-1)(u+5)}\right)^{1/2}$
	$(w+1,0)$	0	$\left(\frac{16}{(u-1)(u+5)}\right)^{1/2}$	$\left(\frac{(u-3)(u+7)}{(u-1)(u+5)}\right)^{1/2}$
$(w20)$	$(w-2,2)$	1	0	0
	$(w-1,1)$	$-\left(\frac{(u-2)(u+3)^2}{2u(u^2-25)}\right)^{1/2}$	$-\left(\frac{(u+3)(u+7)(u-9)}{2(u+5)^2(u-5)}\right)^{1/2}$	$\left(\frac{u-9}{u(u+5)^2}\right)^{1/2}$
	$(w-1,2)$	$-\left(\frac{u+1}{2u}\right)^{1/2}$	0	$\left(\frac{u-1}{2u}\right)^{1/2}$
	$(w0)$	$-\left(\frac{(u+3)(u+7)}{10u(u-5)}\right)^{1/2}$	$-\left(\frac{4(u^2-49)}{5(u^2-25)}\right)^{1/2}$	$\left(\frac{(u-3)(u-7)}{10u(u+5)}\right)^{1/2}$
	$(w1)$	$\left(\frac{u+9}{u(u-5)^2}\right)^{1/2}$	$\left(\frac{(u-3)(u-7)(u+9)}{2(u-5)^2(u+5)}\right)^{1/2}$	$\left(\frac{(u+2)(u-3)^2}{2u(u^2-25)}\right)^{1/2}$
	$(w2)$	0	0	1
$(w+1,10)$	$(w1)$	$-\left(\frac{(u+2)(u-3)^2(u-7)}{u(u-1)(u-5)^2}\right)^{1/2}$	$\left(\frac{8(u+2)(u-3)}{(u-1)(u-5)^2(u+5)}\right)^{1/2}$	$-\left(\frac{2(u-7)(u+9)}{u(u-1)(u^2-25)}\right)^{1/2}$
	$(w+1,0)$	0	$\left(\frac{(u-3)(u+7)}{(u-1)(u+5)}\right)^{1/2}$	$-\left(\frac{16}{(u-1)(u+5)}\right)^{1/2}$
	$(w+1,1)$	0	0	1

tries in the column $(w-1,1)$ possess opposite signs to those of Racah. This phase difference can be traced to our decision to use G_2 phases that parallel those of SO(3). For us, $\varphi(w-1,1)$ differs from $\varphi(w0)$ and $\varphi(w1)$, while for his limited applications Racah was able to avoid having to introduce any general convention for the irreps of G_2.

VI. CONCLUDING REMARKS

The procedure for calculating the entries of Table IV has been described in some detail in order to bring out the parallelism to the familiar SO(3) formulas. Like the $6j$ symbols, the isoscalar factors exhibit the symmetries of Jucys. The checks that these symmetries provide are totally missing, of course, if purely numerical calculations are performed. However, the main significance of our method is that it circumvents the need for a detailed basis. This feature could prove extremely useful for groups possessing a larger number of generators than G_2 or SO(7).

Finding explicit closed forms for the general G_2 $6j$ symbols of the types (1.3) remains an intriguing possibility. Any general formula that encompasses the far from trivial entries of Tables II and III would necessarily have to exhibit a fair degree of complexity. The sequences of the factors $(u-a)$ in Tables II and III seldom suggest the ratio of factorial functions, so it would be difficult at this stage to make conjectures for a general formula.

ACKNOWLEDGMENT

Partial support for the work reported above was provided by the National Science Foundation.

[1] E. Wigner, *Gruppentheorie und ihre Anwendung auf die Quantenmechanik der Atomspektren* (Vieweg, Braunschweig, 1931).

[2] J. S. Griffith, *The Irreducible Tensor Method for Molecular Symmetry Groups* (Prentice-Hall, Englewood Cliffs, NJ, 1962).

[3] P. H. Butler, *Point Group Symmetry Applications* (Plenum, New York, 1981).

[4] K. T. Hecht, Nucl. Phys. **63**, 177 (1965); A **102**, 11 (1967); A **145**, 468 (1970).

[5] S. H. Payne and G. E. Stedman, J. Phys. C **16**, 2725 (1983), Table A3.

[6] B. R. Judd, "Group theoretical approaches," in *The Dynamical Jahn–Teller Effect in Localized Systems*, edited by Yu. E. Perlin and M. Wagner (North-Holland, Amsterdam, 1984).

[7] J. P. Draayer and Y. Akiyama, J. Math. Phys. **14**, 1904 (1973).

[8] R. W. Haase and P. H. Butler, J. Math. Phys. **26**, 1493 (1985).

[9] B. R. Judd, G. M. S. Lister, and M. C. M. O'Brien, J. Phys. A **19**, 2473 (1986).

[10] G. Racah, Phys. Rev. **76**, 1352 (1949).

[11] W. G. McKay and J. Patera, *Tables of Dimensions, Indices, and Branching Rules for Representations of Simple Lie Algebras* (Dekker, New York, 1981).

[12]A. R. Edmonds, *Angular Momentum in Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1957).
[13]P. H. Butler and B. G. Wybourne, Int. J. Quantum Chem. **10**, 581 (1976).
[14]R. P. Bickerstaff and B. G. Wybourne, J. Phys. G **7**, 275 (1981).
[15]P. H. Butler, Philos. Trans. R. Soc. London Ser. A **277**, 545 (1975).
[16]E. P. Wigner, "On representations of certain finite groups" and "On the matrices which reduce the Kronecker products of representations of S. R. groups," in *Quantum Theory of Angular Momentum*, edited by L. C. Biedenharn and H. van Dam (Academic, New York, 1965).
[17]B. R. Judd and H. T. Wadzinski, J. Math. Phys. **8**, 2125 (1967).
[18]H. A. Jahn, Proc. R. Soc. London Ser. A **205**, 192 (1951).
[19]A. P. Jucys and A. J. Savukynas, *Mathematical Foundations of the Atomic Theory* (Acad. Sci. Lith. S. S. R., Vilnius, 1973) (in Russian), pp. 171 and 172.

Group analysis of the three-wave resonant system in (2+1) dimensions

R. A. Leo, L. Martina, and G. Soliani
Dipartimento di Fisica dell' Università, 73100 Lecce, Italy and Istituto Nazionale di Fisica Nucleare, Gruppo di Lecce, Sezione di Bari, Italy

(Received 29 January 1986; accepted for publication 11 July 1986)

The three-wave resonant interaction equations (2D-3WR) in two spatial and one temporal dimension within a group framework are analyzed. The symmetry algebra of this system, which turns out to be an infinite-dimensional Lie algebra whose subalgebra is of the Kac–Moody type, is found. The one- and two-dimensional symmetry subalgebras are classified and the corresponding reduction equations are obtained. From these the new invariant and the partially invariant solutions of the original 2D-3WR equations are obtained.

I. INTRODUCTION

The group analysis of nonlinear differential equations has received much attention in the last years.[1–3] The main feature of this approach is the finding of all infinitesimal generators (which constitute a Lie algebra) of those groups of point transformations that leave the equations under consideration invariant.

Several authors have mainly dealt with systems in one spatial and one temporal dimension.[1–4] The group investigation of nonlinear differential equations, in more than two dimensions, is still at the beginning. However, the application of the group method to some special cases[5] succeeded in revealing new features of certain integrable equations of physical significance, such as the Kadomtsev–Petviashvili and the Davey–Stewartson equations.[6,7] Notable properties are (i) the existence of infinite-dimensional Lie algebras of the Kac–Moody type,[8] (ii) the discovery of interesting classes of solutions admitted by reduced versions of the original systems, and (iii) their possible connection with equations of the Painlevé type.

In this article we study the three-wave resonant (2D-3WR) equations in two spatial and one temporal dimension[9] from the group point of view. These equations, which play an important role in plasma physics and in nonlinear optics,[9,10] allow a linear eigenvalue problem[11,12] and Bäcklund transformations and can be solved via the inverse spectral transform.[13]

In Sec. II we scrutinize the symmetry algebra of the 2D-3WR system, which turns out to be an infinite-dimensional Lie algebra whose subalgebra is of the Kac–Moody type. The symmetry group corresponding to the symmetry algebra is treated in Sec. III. Section IV is devoted to a classification of one- and two-dimensional subalgebras of the 2D-3WR algebras, into conjugacy classes under the action of the adjoint group of the symmetry group of the 2D-3WR system. Furthermore, we introduce a finite-dimensional subalgebra of physical meaning, having a scale generator that leads to an interesting reduced system with nonconstant coefficients in two independent variables. This system, discussed in Sec. V, possesses a linear spectral problem derived using a prolongation technique.[14] In Sec. V we examine the reduced equations arising from the low-dimensional symmetry subalgebras considered in Sec. IV. We obtain both classes of the new invariant and the partially invariant solutions of the 2D-3WR equations. From among these a few stand out that are closely related to solutions of the Painlevé VI equation. Finally, in Sec. VI we report some comments and concluding remarks.

II. THE SYMMETRY ALGEBRA OF THE 2D-3WR SYSTEM

For the sake of definiteness, we consider the three-wave resonant process in (2 + 1) dimensions in the case of explosive instability[9] described by the equations

$$\Delta_j \equiv u_{jt} + c_j u_{jx} + d_j u_{jy} - i u_k^* u_l^* = 0, \qquad j,k,l = 1,2,3, \quad j \neq k \neq l, \tag{2.1}$$

where $u_j(x,y,t)$ are the complex amplitudes of the wave packets, c_j, d_j are their group velocities, the asterisk denotes complex conjugate, and subscripts mean partial derivatives.

The quantities Δ_j are functions defined in the space $X \times U^{(1)}$, where $X = (x,y,t)$ and

$$U^{(1)} = (u_j, u_j^*, u_{jx}, u_{jx}^*, u_{jy}, u_{jy}^*, u_{jt}, u_{jt}^*)$$

are, respectively, the manifold of the space-time variables and of the amplitudes, their derivatives, and their complex conjugates [we have used the abbreviation u_j for (u_1,u_2,u_3) and so on].

In order to look for the symmetry algebra for Eqs. (2.1), let us introduce the vector field

$$V = \xi\,\partial_x + \eta\,\partial_y + \tau\,\partial_t + \phi_j\,\partial_{u_j} + \bar{\phi}_j\,\partial_{u_j^*}, \quad j = 1,2,3, \tag{2.2}$$

on $X \times U$, where $U = (u_j, u_j^*)$ is the space of the dependent variables, and $\xi,\eta,\tau,\phi_j,\bar{\phi}_j$ are functions of (x,y,t,u_j,u_j^*) defined in $X \times U$. Here $\partial_x \equiv \partial/\partial x$, $\partial_y \equiv \partial/\partial_y$,... and the convention over repeated indexes is understood.

The prolongation of V to $X \times U^{(1)}$ is given by[3]

$$\mathrm{pr}\,V = V + \phi_j^x\,\partial_{u_{jx}} + \phi_j^y\,\partial_{u_{jy}} + \phi_j^t\,\partial_{u_{jt}} + \bar{\phi}_j^x\,\partial_{u_{jx}^*} + \bar{\phi}_j^y\,\partial_{u_{jy}^*} + \bar{\phi}_j^t\,\partial_{u_{jt}^*}, \tag{2.3}$$

where the fields $\phi_j^x, \phi_j^y, \ldots$ are defined by

$$\phi_j^m = D_m(\phi_j - u_{i,m}\xi_i) + u_{i,m}^j \xi_i, \quad i = 1,2,3, \tag{2.4}$$

where

$$u_{i,m}^j = \frac{\partial^2 u^j}{\partial_{x_i}\partial_{x_m}}, \quad x_1 = x, \quad x_2 = y, \quad x_3 = t,$$

 0022-2488/86/112623-06$02.50

$$\xi_1=\xi\,,\quad \xi_2=\eta,\quad \xi_3=\tau\,,\quad \phi_j^1=\phi_j^x,\ldots,$$

and D_m is the total derivative with respect to x_m.

We recall that[3] a local group of transformations G is a symmetry group for Eq. (2.1) if and only if

$$\text{pr } V[\Delta_j]=0\,, \tag{2.5a}$$

$$\text{pr } V[\Delta_j^*]=0\,, \tag{2.5b}$$

whenever $\Delta_j=\Delta_j^*=0$, for every generator V of G where pr V is expressed by (2.3). Equation (2.5a) yields the relations

$$(c_j\,\partial_x+d_j\,\partial_y+\partial_t)(c_j\tau-\xi)=0\,, \tag{2.6a}$$

$$(c_j\,\partial_x+d_j\,\partial_y+\partial_t)(d_j\tau-\eta)=0\,, \tag{2.6b}$$

$$\begin{aligned}&-i\phi_k^*u_l^*-i\phi_l^*u_k^*+c_j(\phi_{jx}-i\tau_x u_k^*u_l^*)\\&\quad+d_j(\phi_{jy}-i\tau_y u_k^*u_l^*)+i\phi_{ju_j}u_k^*u_l^*-i\phi_{ju_j^*}u_k u_l\\&\quad+\phi_{jt}-i\tau_t u_k^*u_l^*=0\,,\quad j,k,l=1,2,3\,.\end{aligned} \tag{2.6c}$$

Constraints similar to (2.6) can be obtained from (2.5b). Equations (2.6) are satisfied by

$$\begin{aligned}\xi&=\alpha_{23}^{-1}c_1\varphi_1(\zeta_1)+\alpha_{31}^{-1}c_2\varphi_2(\zeta_2)+\alpha_{12}^{-1}c_3\varphi_3(\zeta_3)\,,\\ \eta&=\alpha_{23}^{-1}d_1\varphi_1(\zeta_1)+\alpha_{31}^{-1}d_2\varphi_2(\zeta_2)+\alpha_{12}^{-1}d_3\varphi_3(\zeta_3)\,,\\ \tau&=\alpha_{23}^{-1}\varphi_1(\zeta_1)+\alpha_{31}^{-1}\varphi_2(\zeta_2)+\alpha_{12}^{-1}\varphi_3(\zeta_3)\,,\end{aligned} \tag{2.7}$$

where

$$\alpha_{kl}=\Delta/(d_k-d_l)\,, \tag{2.8}$$

with

$$\Delta=d_1(c_3-c_2)+d_2(c_1-c_3)+d_3(c_2-c_1) \tag{2.9}$$

and $\varphi_j(\zeta_j)$ $(j=1,2,3)$ are arbitrary functions depending on the variables

$$\zeta_j=x-\frac{c_k-c_l}{d_k-d_l}y+\frac{c_kd_l-c_ld_k}{d_k-d_l}\,,$$

$$j,k,l=1,2,3,\quad j\neq k\neq l. \tag{2.10}$$

Furthermore, we have

$$\phi_j=\rho_j u_j\,,\quad \bar\phi_j=\rho_j u_j^*\,,\quad j=1,2,3\,, \tag{2.11}$$

where the quantities ρ_j are given by

$$\rho_j=-\frac{1}{2}\sum_{k\neq j}\dot\varphi_k(\zeta_k)\,, \tag{2.12}$$

where $\dot\varphi_j\equiv\partial\varphi_j/\partial\zeta_j$. The generators of the Lie-point symmetry algebra A associated with the 2D-3WR equations (2.1) can be written as

$$V=X(\varphi_1)+Y(\varphi_2)+Z(\varphi_3)\,, \tag{2.13}$$

where

$$\begin{aligned}X(\varphi_1)&=\varphi_1(\zeta_1)\partial_{\zeta_1}-\tfrac{1}{2}\dot\varphi_1(\zeta_1)(u_2\,\partial_{u_2}\\&\quad+u_2^*\,\partial_{u_2^*}+u_3\,\partial_{u_3}+u_3^*\,\partial_{u_3^*})\,,\end{aligned} \tag{2.14a}$$

$$\begin{aligned}Y(\varphi_2)&=\varphi_2(\zeta_2)\partial_{\zeta_2}-\tfrac{1}{2}\dot\varphi_2(\zeta_2)(u_1\,\partial_{u_1}\\&\quad+u_1^*\,\partial_{u_1^*}+u_3\,\partial_{u_3}+u_3^*\,\partial_{u_3^*})\,,\end{aligned} \tag{2.14b}$$

$$\begin{aligned}Z(\varphi_3)&=\varphi_3(\zeta_3)\partial_{\zeta_3}-\tfrac{1}{2}\dot\varphi_3(\zeta_3)(u_1\,\partial_{u_1}\\&\quad+u_1^*\,\partial_{u_1^*}+u_2\,\partial_{u_2}+u_2^*\,\partial_{u_2^*})\,,\end{aligned} \tag{2.14c}$$

and the operators ∂_{ζ_j} are expressed by [see (2.10)]

$$\partial_{\zeta_j}=\alpha_{kl}^{-1}(c_j\,\partial_x+d_j\,\partial_y+\partial_t)\,, \tag{2.15}$$

where α_{kl} is given by (2.8) and the indexes $j,k,l=1,2,3$ are cyclic.

In what follows, we shall assume that each function $\varphi_j(\zeta_j)$ takes the form of a Laurent series expansion in the argument ζ_j. Then we restrict ourselves to deal with the subalgebra L of A having the basis

$$\begin{aligned}X_n&=\zeta_1^n\,\partial_{\zeta_1}-\tfrac{1}{2}n\zeta_1^{n-1}(u_2\partial_{u_2}\\&\quad+u_2^*\,\partial_{u_2^*}+u_3\,\partial_{u_3}+u_3^*\partial_{u_3^*})\,,\end{aligned} \tag{2.16a}$$

$$\begin{aligned}Y_n&=\zeta_2^n\,\partial_{\zeta_2}-\tfrac{1}{2}n\zeta_2^{n-1}(u_1\partial_{u_1}\\&\quad+u_1^*\,\partial_{u_1^*}+u_3\,\partial_{u_3}+u_3^*\partial_{u_3^*})\,,\end{aligned} \tag{2.16b}$$

$$\begin{aligned}Z_n&=\zeta_3^n\,\partial_{\zeta_3}-\tfrac{1}{2}n\zeta_3^{n-1}(u_1\,\partial_{u_1}\\&\quad+u_1^*\,\partial_{u_1^*}+u_2\,\partial_{u_2}+u_2^*\,\partial_{u_2^*})\,.\end{aligned} \tag{2.16c}$$

The commutator relations

$$\begin{aligned}&[X_n,X_m]=(m-n)X_{n+m-1}\,,\\&[Y_n,Y_m]=(m-n)Y_{n+m-1}\,,\\&[Z_n,Z_m]=(m-n)Z_{n+m-1}\,,\\&[X_n,Y_m]=[X_n,Z_m]=[Y_n,Z_m]=0\end{aligned} \tag{2.17}$$

hold, where $n,m\in\mathbb{Z}$.

From (2.17) we see that L is the direct sum of the ideals generated by $\{X_n\},\{Y_n\},\{Z_n\}$. Each of these ideals is an algebra isomorphic to the $\mathbb{Z}$-graded algebra $\mathbb{R}[t,t^{-1}](d/dt)$ (see Refs. 5 and 8). Since the last is a simple algebra, then L is semisimple.[1,2] We notice that Eqs. (2.17) enable us to obtain commutator relations among the infinitesimal generators (2.14), provided that any function $\varphi_j(\zeta_j)$ may be expressed as a Laurent series. We have, for example,

$$[X(\varphi_1^{(1)}),X(\varphi_1^{(2)})]=X(\varphi_1^{(1)}\dot\varphi_1^{(2)}-\dot\varphi_1^{(1)}\varphi_1^{(2)})\,. \tag{2.18}$$

Equations (2.17) define an affine Lie algebra of the Kac–Moody type.

III. THE SYMMETRY GROUP OF THE 2D-3WR SYSTEM

In order to obtain the symmetry group of Eq. (2.1), we need to integrate the infinitesimal symmetries (2.13) and (2.14). In doing so, let us consider the equations

$$\begin{aligned}&\frac{dx'}{d\lambda}=\xi(x',y',t',u_j',u_j^{*\prime}),\quad \frac{dy'}{d\lambda}=\eta(x',y',t',u_j',u_j^{*\prime})\,,\\&\frac{dt'}{d\lambda}=\tau(x',y',t',u_j',u_j^{*\prime}),\quad \frac{du_j'}{d\lambda}=\phi_j(x',y',t',u_j',u_j^{*\prime})\,,\\&\frac{du_j^{*\prime}}{d\lambda}=\bar\phi(x',y',t',u_j',u_j^{*\prime})\,,\end{aligned} \tag{3.1}$$

where

$$\begin{aligned}&x'(0)=x,\ y'(0)=y,\ t'(0)=t,\\&u_j'(0)=u_j,\ u_j^{*\prime}(0)=u_j^*\,,\end{aligned} \tag{3.2}$$

ξ,η,τ and $\phi,\bar\phi$ are, respectively, given by (2.7) and (2.9). Now looking at (2.8) and (3.1) we introduce a set of functions $\zeta_j'(\lambda)$, such that

$$\frac{d}{d\lambda}\zeta_1' = \varphi_1(\zeta_1'), \quad \frac{d}{d\lambda}\zeta_2' = \varphi_2(\zeta_2'), \quad \frac{d}{d\lambda}\zeta_3' = \varphi_3(\zeta_3'), \tag{3.3}$$

with the conditions

$$\zeta_1'(0) = \zeta_1, \quad \zeta_2'(0) = \zeta_2, \quad \zeta_3'(0) = \zeta_3. \tag{3.4}$$

The system (3.3) can be easily integrated to yield

$$\zeta_1'(\lambda,\zeta_1) = F_1^{-1}(\lambda + F_1(\zeta_1)), \tag{3.5a}$$

$$\zeta_2'(\lambda,\zeta_2) = F_2^{-1}(\lambda + F_2(\zeta_2)), \tag{3.5b}$$

$$\zeta_3'(\lambda,\zeta_3) = F_3^{-1}(\lambda + F_3(\zeta_3)), \tag{3.5c}$$

where F_j^{-1} is the inverse of the function defined by

$$F_j(\zeta_j) = \int_{\zeta_{j,o}}^{\zeta_j} \frac{ds}{\varphi_j(s)} \quad (j = 1,2,3). \tag{3.6}$$

Of course, the variables x',y',t' may be found in terms of ζ_j' from Eq. (2.8), where primes are understood. Resorting to Eq. (3.1) and taking account of (2.10) we get

$$u_j'(x',y',t') = u_j(x,y,t)\left[\frac{\varphi_k(\zeta_k)\varphi_l(\zeta_l)}{\varphi_k(\zeta_k')\varphi_l(\zeta_k')}\right]^{1/2}, \tag{3.7}$$

where $j \neq k \neq l$ and x,y,t are regarded as functions of x',y',t'.

This set of formulas provides a new solution $\{u_j'(x',y',t')\}$ of the 2D-3WR equations in terms of a known solution $\{u_j(x,y,t)\}$.

IV. LOW-DIMENSIONAL SUBALGEBRAS

A. One-dimensional subalgebras

It is known that the group analysis of differential equations leads to the problem of classifying the subgroups (and the corresponding subalgebras) under which certain classes of solutions are invariant.[1-3] To this aim one needs to build up the so-called optimal system θ_s, that is the set of representatives of the classes of s-dimensional subalgebras, L_s, which are pairwise nonconjugates by the inner automorphism group (adjoint group).[1] In doing so, first we construct the system of one-dimensional subalgebras L_1. Looking over the commutator relation (2.18) and dropping the index 1 for simplicity, from the Campbell–Hausdorff formula we deduce that the action of the adjoint subgroup $\exp[\lambda \text{ ad } X(\psi)]$ on $X(\varphi)$ is

$$\exp[\lambda \text{ ad } X(\psi)]\langle X(\varphi)\rangle = X(\varphi'), \tag{4.1}$$

where the function $\varphi' \equiv \varphi'(\zeta')$ is defined by

$$\frac{d}{d\lambda}\varphi'(\zeta') = \dot{\psi}(\zeta')\varphi'(\zeta') - \dot{\varphi}'(\zeta')\psi(\zeta'), \tag{4.2}$$

with $\varphi'(\zeta')|_{\lambda=0} = \varphi(\zeta)$. Equation (4.2) has the solution[5]

$$\varphi'(\zeta') = \varphi(\zeta(\zeta'))\psi(\zeta')/\psi(\zeta(\zeta')), \tag{4.3}$$

where

$$\zeta'(\zeta) = F^{-1}(\lambda + F(\zeta)). \tag{4.4}$$

Concerning Eq. (4.3), Neuman[15] has shown that a function $\psi(\zeta)$ can be chosen in such a way that $\varphi'(\zeta') \equiv 1$. As a consequence, all one-dimensional subalgebras generated by elements of the $X(\varphi)$ type are conjugated to $\partial_{\zeta_1} \equiv X_0$. Similarly, any generator of the form $Y(\varphi_2)$ and $Z(\varphi_3)$ can be conjugated to $\partial_{\zeta_2} \equiv Y_0$ and $\partial_{\zeta_3} \equiv Z_0$, respectively, via suitable elements $\exp[\lambda \text{ ad } Y(\psi_2)]$ and $\exp[\lambda \text{ ad } Z(\psi_3)]$ of the adjoint group. However, the conjugacy classes of X_0, Y_0, Z_0 do not exhaust all elements of θ_1. For example, let us consider the one-dimensional subalgebra spanned by $X(\varphi_1) + Y(\varphi_2)$. Since the generators X and Y commute [see (2.17)], from the Campbell–Hausdorff expansion we obtain

$$\exp[\lambda_2 \text{ ad } Y(\psi_2)]\exp[\lambda_1 \text{ ad } X(\psi_1)]\langle X(\varphi_1) + Y(\varphi_2)\rangle = X_0 + Y_0, \tag{4.5}$$

where ψ_1 and ψ_2 must be taken in such a way that $\varphi_1'(\zeta_1') = \varphi_2'(\zeta_2') \equiv 1$ [see (4.3)].

Furthermore, it is easy to show that $X_0 + Y_0$ cannot be related to the conjugacy classes of X_0, Y_0, Z_0. We can likewise deal with the subalgebra generated by $X(\varphi_1) + Z(\varphi_3)$, $Y(\varphi_2) + Z(\varphi_3)$, and $X(\varphi_1) + Y(\varphi_2) + Z(\varphi_3)$, respectively. We conclude that the optimal system is given by the conjugacy classes of

$$X_0,\ Y_0,\ Z_0,\ X_0 + Y_0,\ X_0 + Z_0,\ Y_0 + Z_0,\ X_0 + Y_0 + Z_0. \tag{4.6}$$

B. Two-dimensional subalgebras

In order to classify the two-dimensional subalgebras L_2 of the symmetry algebra, we have to determine the optimal system θ_2. Concerning this, it can be shown that only two isomorphy classes of two-dimensional Lie algebras exist, namely,

$$[U_1,U_2] = 0 \tag{4.7}$$

and

$$[U_1,U_2] = U_1. \tag{4.8}$$

In both the cases (4.7) and (4.8), U_1 can be singled out without loss of generality and identified with one of the elements in (4.6). Following a scheme analogous to that used to derive (4.6), we obtain all the two-dimensional Abelian representative subalgebras, i.e.,

$$\{X_0,Y_0\},\ \{X_0,Z_0\},\ \{Y_0,Z_0\},\ \{X_0,Y_0+Z_0\},\ \{Y_0,X_0+Z_0\},\ \{Z_0,X_0+Y_0\}, \tag{4.9}$$

which are related of course to the commutation property of the translation generators. The non-Abelian subalgebras of L_2 type can be classified starting from (4.8). They are

$$\{X_0,X_1+\epsilon_1Y_0+\epsilon_2Z_0\}, \tag{4.10a}$$

$$\{Y_0,Y_1+\epsilon_1X_0+\epsilon_2Z_0\}, \tag{4.10b}$$

$$\{Z_0,Z_1+\epsilon_1X_0+\epsilon_2Y_0\}, \tag{4.10c}$$

$$\{X_0+Y_0,X_1+Y_1+\epsilon_1Z_0\}, \tag{4.10d}$$

$$\{X_0+Z_0,X_1+Z_1+\epsilon_1Y_0\}, \tag{4.10e}$$

$$\{Y_0+Z_0,Y_1+Z_1+\epsilon_1X_0\}, \tag{4.10f}$$

$$\{X_0+Y_0+Z_0,X_1+Y_1+Z_1\}, \tag{4.10g}$$

where the parameters ϵ_1,ϵ_2 can take only two mutually nonconjugate values (0, or, say, 1).

It is worth noticing that the algebra (4.10g) is endowed with the generator of the scale transformation

$$S = X_1 + Y_1 + Z_1 = \zeta_1 \partial_{\zeta_1} + \zeta_2 \partial_{\zeta_2} + \zeta_3 \partial_{\zeta_3} - (u_j \partial_{u_j} + u_j^* \partial_{u_j^*}) \tag{4.11}$$

($j = 1,2,3$). We point out also that this algebra can be regarded as a subalgebra of the algebra L_6 formed by the generators $X_0, Y_0, Z_0, X_1, Y_1, Z_1$ [see (2.16)]. Here L_6 is a solvable six-dimensional algebra that admits the semidirect decomposition[1]

$$L_6 = \{X_0, Y_0, Z_0\} \oplus_s \{X_1, Y_1, Z_1\}$$

with the Abelian ideal $N = \{X_0, Y_0, Z_0\}$. It contains all the infinitesimal transformations of straightforward physical meaning, as, for example, translations and dilatations.

V. 2D-3WR EQUATIONS COMING FROM SYMMETRY REDUCTION

Here we study the reduction equations of the 2D-3WR system (2.1), which can be written in the simple form

$$\alpha_{kl} \frac{\partial}{\partial \zeta_j} u_j = i u_k^* u_l^* , \tag{5.1}$$

whose solutions are related to solutions that are invariant and partially invariant under the subgroups of the subalgebras discussed in the preceding section. To achieve this goal we use some mathematical notions with which the reader is supposed to be acquainted. Anyway, one may consult some basic references (see Refs. 1–3). We recall that the procedure of symmetry reduction consists essentially in finding the invariants of a given subgroup of the symmetry group admitted by the differential equations under consideration. In order to apply the reduction technique to the 2D-3WR interaction, let us begin to analyze the invariants of the subgroups of the one-dimensional subalgebras (4.6).

A. Case I: reduction from one-dimensional subalgebras

Let us deal with one of the symmetries (4.6), say Z_0. A basis of invariants of the subgroup $\exp[\lambda Z_0]$ can be determined by the partial differential equation

$$\frac{\partial}{\partial \zeta_3} I = 0 , \tag{5.2}$$

which is fulfilled by an arbitrary function of the eight independent solutions

$$I_1 = \zeta_1, \quad I_2 = \zeta_2, \quad I_{2+j} = u_j, \quad I_{5+j} = u_j^*, \quad j = 1,2,3 . \tag{5.3}$$

We observe that each solution of Eq. (5.1) can be regarded as a six-dimensional manifold U defined by

$$u_j - u_j(\zeta_1, \zeta_2, \zeta_3) = 0, \quad u_j^* - u_j^*(\zeta_1, \zeta_2, \zeta_3) = 0 , \tag{5.4}$$

in the space $\mathbb{Z}$.

Now if we require the invariance of the manifold U under the subgroup $\exp[\lambda Z_0]$, we have that U is given implicitly by six equations involving invariant functions only. These equations can be solved with respect to six new variables, which depend upon two invariants. Since we would express U in the explicit form (5.4), the rank of the matrix $\partial(I_1, \ldots, I_8)/(u_1, \ldots, u_3^*)$ has to be 6. Of course, this condition is verified for the basis of invariants (5.3). Thus we can introduce the new variables $v_j = I_{2+j}, v_j^* = I_{5+j}$ ($j = 1,2,3$) as functions of the independent variables $\zeta_1 = I_1$ and $\zeta_2 = I_2$. Furthermore, the functions u_j, u_j^* can be trivially written in terms of v_j and v_j^*. Hence, recalling that u_j^* is the complex conjugate of u_j, Eqs. (5.1) provide

$$\alpha_{23} \frac{\partial v_1}{\partial \zeta_1} = i v_2^* v_3^* , \tag{5.5a}$$

$$\alpha_{31} \frac{\partial v_2}{\partial \zeta_2} = i v_1^* v_3^* , \tag{5.5b}$$

$$v_1^* v_2^* = 0 . \tag{5.5c}$$

This system admits three classes of solutions, i.e.,

$$\{0,0,v_3(\zeta_1,\zeta_2)\}, \quad \{v_1(\zeta_2),0,0\}, \quad \{0,v_2(\zeta_1),0\} , \tag{5.6}$$

where the v_j's are arbitrary functions.

The symmetries X_0 and Y_0 lead to similar results.

Now let us consider the operator $Y_0 + Z_0$. Following an analogous procedure, we have the basis if invariants

$$I_1 = \zeta_1, \; I_2 = \zeta_2 - \zeta_3, \; I_{2+j} = u_j, \; I_{5+j} = u_j^*, \; j = 1,2,3 . \tag{5.7}$$

Thus, the reduced system reads

$$\alpha_{23} \frac{\partial v_1}{\partial \zeta_1} = i v_2^* v_3^* , \quad \alpha_{31} \frac{\partial v_2}{\partial \zeta_2} = i v_1^* v_3^* , \quad \alpha_{12} \frac{\partial v_3}{\partial \zeta_2} = - i v_1^* v_2^* , \tag{5.8}$$

where $v_j \equiv v_j(\zeta_1, \zeta_2 - \zeta_3)$. Finally, for the symmetry $X_0 + Y_0 + Z_0$ we obtain the set of invariants

$$I_1 = \zeta_1 - \zeta_3, \quad I_2 = \zeta_2 - \zeta_3, \quad I_{2+j} = u_j , \quad I_{5+j} = u_j^*, \quad j = 1,2,3 .$$

The reduced system becomes

$$\alpha_{23} \frac{\partial v_1}{\partial \zeta_1} = i v_2^* v_3^* , \quad \alpha_{31} \frac{\partial v_2}{\partial \zeta_2} = i v_1^* v_3^* , \quad \alpha_{12} \left(\frac{\partial v_3}{\partial \zeta_1} + \frac{\partial v_3}{\partial \zeta_2} \right) = - i v_1^* v_2^* , \tag{5.9}$$

where $v_j \equiv v_j(\zeta_1 - \zeta_3, \zeta_2 - \zeta_3)$.

The systems (5.8) and (5.9), which can be investigated within the prolongation scheme[16–18] and the inverse spectral transform[9,19] may furnish new solutions of the original 2D-3WR equations.

To conclude this section, we write down an interesting reduced system coming from the scale symmetry (4.11), which leads to the invariants

$$I_1 = \zeta_1/\zeta_3 \equiv z_1, \quad I_2 = \zeta_2/\zeta_3 \equiv z_2, \quad I_j = u_j \zeta_3 , \quad I_j^* = u_j^* \zeta_3, \quad j = 1,2,3 . \tag{5.10}$$

From (5.11) we obtain

$$\alpha_{23} \frac{\partial v_1}{\partial z_1} = i v_2^* v_3^* , \quad \alpha_{31} \frac{\partial v_2}{\partial z_2} = i v_1^* v_3^* , \quad \alpha_{12} \left(z_1 \frac{\partial v_3}{\partial z_1} + z_2 \frac{\partial v_3}{\partial z_2} + v_3 \right) = - i v_1^* v_2^* . \tag{5.11}$$

The system (5.11), which has nonconstant coefficients, has been studied in the framework of the prolongation theory.[14] It allows a linear eigenvalue problem and can be investigated in the context of the inverse spectral transform.

B. Case II: reduction from two-dimensional subalgebras

1. Subcase (a): Abelian subalgebras

Let us consider the subalgebra $\{Y_0,Z_0\}$. The invariant solutions related to this algebra must satisfy both Eqs. (5.5) and the reduced system corresponding to the symmetry Y_0, simultaneously. This implies that the resulting system admits the solutions

$$\{0,v_2(\zeta_1),0\} \quad \text{and} \quad \{0,0,v_3(\zeta_1)\}, \tag{5.12}$$

where v_2, v_3 are arbitrary functions. On the other hand, the algebra $\{Y_0+Z_0, X_0\}$ gives rise to a reduced system formed by Eqs. (5.8) and a set of three equations of the type (5.5). This system affords the solution $\{v_1(\zeta_2-\zeta_3),0,0\}$, where v_1 is an arbitrary function.

2. Subcase (b): non-Abelian subalgebras

Let us consider, for example, the subalgebra (4.10c). A basis of invariants of the subgroup of this subalgebra is furnished by the equations

$$\frac{\partial I}{\partial \zeta_3}=0, \tag{5.13a}$$

$$\left[\zeta_3\frac{\partial}{\partial\zeta_3}+\epsilon_1\frac{\partial}{\partial\zeta_1}+\epsilon_2\frac{\partial}{\partial\zeta_2}-\frac{1}{2}\left(u_1\frac{\partial}{\partial u_1}+u_1^*\frac{\partial}{\partial u_1^*}+u_2\frac{\partial}{\partial u_2}+u_2^*\frac{\partial}{\partial u_2^*}\right)\right]I=0. \tag{5.13b}$$

First let us deal with the case $\epsilon_1=\epsilon_2=1$. Then we find

$$I_1=\zeta_2-\zeta_1,\ I_2=u_1e^{\zeta_1/2},\ I_3=u_1^*e^{\zeta_1/2},\ I_4=u_2e^{\zeta_1/2},\quad I_5=u_2^*e^{\zeta_1/2},\ I_6=u_3,\ I_7=u_3^*. \tag{5.14}$$

Taking account of (5.14), Eqs. (5.1) yield

$$\alpha_{23}\left(\frac{\partial v_1}{\partial\zeta_2}+\frac{1}{2}v_1\right)=-iv_2^*v_3^*,\qquad \alpha_{31}\frac{\partial v_2}{\partial\zeta_2}=iv_1^*v_3^*,\quad v_1^*v_2^*=0, \tag{5.15}$$

where $v_j\equiv v_j(\zeta_2-\zeta_1)$.

A simple solution of Eq. (5.15) is

$$\{v_{1,0}\exp[-\tfrac{1}{2}(\zeta_2-\zeta_1)],0,0\},$$

where $v_{1,0}$ is a constant of integration.

The case $\epsilon_1=0$, $\epsilon_2=1$ (or $\epsilon_1=1$, $\epsilon_2=0$) provides analogous results, while for $\epsilon_1=\epsilon_2=0$ we are led to solutions of the 2D-3WR equations that are partially invariant under the subgroup, say H, of the given subalgebra, in the sense that their manifold U is a partially invariant manifold of H (see Ref. 1, Chap. VI).

The algorithm that can be exploited to determine partially invariant solutions of a system of differential equations and the conditions assuring their existence are well established (see, for instance, Ref. 1, Chap. VI, p. 22). In our case, from (5.13) we get the basis of invariants

$$I_1=\zeta_1,\quad I_2=\zeta_2,\quad I_3=u_1/u_1^*,\quad I_4=u_2/u_2^*,\quad I_5=u_2/u_1,\quad I_6=u_3,\quad I_7=u_3^*, \tag{5.16}$$

so that the rank of the matrix $\partial(I_1,\ldots,I_7)/\partial(u_1,\ldots,u_3^*)$ is 5. As a consequence, the defect δ (see Refs. 1 and 2) of any solution $u\in U$ turns out to be 1. This assures that u exists as a partially invariant H-solution (see Ref. 1, p. 283).

Setting $v_1=I_3$, $v_2=I_4$, $v_3=I_5$, $v_4=I_6$, where $v_j=v_j(\zeta_1,\zeta_2)$, from Eq. (5.1) we obtain

$$u_1u_2=0 \tag{5.17}$$

and, according to the choice (i) $u_1\neq0$, $u_2=0$ or (ii) $u_1=0$, $u_2\neq0$, we deduce that $v_1\equiv v_1(\zeta_2)$ and $v_2\equiv v_2(\zeta_1)$, respectively, where v_1 and v_2 are arbitrary functions. We conclude noticing that in both cases (i) and (ii) we have $u_3=0$.

Furthermore, in correspondence with the choice (i) or (ii), we have that $u_1\equiv u_1(\zeta_2,\zeta_3)$ with

$$u_1(\zeta_2,\zeta_3)/u_1^*(\zeta_2,\zeta_3)=v_1(\zeta_2), \tag{5.18}$$

or $u_2\equiv u_2(\zeta_1,\zeta_3)$, with

$$u_2(\zeta_1,\zeta_3)/u_2^*(\zeta_1,\zeta_3)=v_2(\zeta_1). \tag{5.19}$$

The relations (5.18) and (5.19) imply, respectively,

$$u_1=\hat u_1(\zeta_2,\zeta_3)\exp[if_1(\zeta_2)], \tag{5.20}$$

$$u_2=\hat u_2(\zeta_1,\zeta_3)\exp[if_2(\zeta_1)], \tag{5.21}$$

where the (real) functions $\hat u_j$ and f_j are arbitrary. We remark that u_1 and u_2 become invariant under H in the special case in which they are independent from ζ_3.

A basis of invariants I of the subgroup of the subalgebra (4.10f) arises from the equations

$$\left(\frac{\partial}{\partial\zeta_2}+\frac{\partial}{\partial\zeta_3}\right)I=0, \tag{5.22a}$$

$$\left[\epsilon_1\frac{\partial}{\partial\zeta_1}+\zeta_2\frac{\partial}{\partial\zeta_2}+\zeta_3\frac{\partial}{\partial\zeta_3}-\frac{1}{2}\left(2u_1\frac{\partial}{\partial u_1}+2u_1^*\frac{\partial}{\partial u_1^*}+u_2\frac{\partial}{\partial u_2}+u_2^*\frac{\partial}{\partial u_2^*}+u_3\frac{\partial}{\partial u_3}+u_3^*\frac{\partial}{\partial u_3^*}\right)\right]I=0. \tag{5.22b}$$

We get

$$I_1=(\zeta_2-\zeta_3)\exp(-\zeta_1/\epsilon_1)\equiv z,\quad I_2=u_1(\zeta_2-\zeta_3),\quad I_3=u_1^*(\zeta_2-\zeta_3),\quad I_4=u_2^2(\zeta_2-\zeta_3),\quad I_5=(u_2^*)^2(\zeta_2-\zeta_3),\quad I_6=u_3^2(\zeta_2-\zeta_3),\quad I_7=(u_3^*)^2(\zeta_2-\zeta_3). \tag{5.23}$$

If $v_1(z)=I_2$, $v_2(z)=I_4$, and $v_3(z)=I_6$, the reduced 2D-3WR equations coming from the subalgebra (4.10f) read

$$\alpha_{23}z\frac{\partial v_1}{\partial z}=-i\epsilon_1(v_2^*v_3^*)^{1/2}, \tag{5.24a}$$

$$\alpha_{31}\left(z\frac{\partial v_2}{\partial z}-v_2\right)=2iv_1^*(v_2v_3^*)^{1/2}, \tag{5.24b}$$

$$\alpha_{12}\left(z\frac{\partial v_3}{\partial z}-v_3\right)=-2iv_1^*(v_2^*v_3)^{1/2}. \tag{5.24c}$$

We shall discuss this system for the nontrivial case $\epsilon_1\neq0$ elsewhere.[14]

A notable set of reduced equations is found from the solvable subalgebra (4.10g), where the explicit form of the scale generator is furnished by (4.11). The invariants I of the subgroup of (4.10g) can be written solving the equations

$$\left(\frac{\partial}{\partial\zeta_1}+\frac{\partial}{\partial\zeta_2}+\frac{\partial}{\partial\zeta_3}\right)I=0\,, \tag{5.25}$$

$$\left[\zeta_1\frac{\partial}{\partial\zeta_1}+\zeta_2\frac{\partial}{\partial\zeta_2}+\zeta_3\frac{\partial}{\partial\zeta_3}-u_j\frac{\partial}{\partial u_j}-u_j^*\frac{\partial}{\partial u_j^*}\right]I=0\,, \tag{5.26}$$

which yield the basis

$$\begin{aligned} &I_1=(\zeta_1-\zeta_3)/(\zeta_2-\zeta_3)\equiv z\,,\\ &I_2=u_1(\zeta_1-\zeta_3),\quad I_3=u_2(\zeta_1-\zeta_3)\,,\\ &I_4=u_3(\zeta_1-\zeta_3),\quad I_5=u_1^*(\zeta_1-\zeta_3)\,,\\ &I_6=u_2^*(\zeta_1-\zeta_3),\quad I_7=u_3^*(\zeta_1-\zeta_3)\,. \end{aligned} \tag{5.27}$$

Introducing the variables

$$v_1(z)=I_1/z,\quad v_2(z)=I_2,\quad v_3(z)=(1-z)I_3/z\,, \tag{5.28}$$

the 2D-3WR equations take the form

$$\alpha_{32}\frac{\partial v_1}{\partial z}=\frac{iv_2^*v_3^*}{[z(1-z)]},\quad \alpha_{31}\frac{\partial v_2}{\partial z}=-\frac{iv_1^*v_3^*}{(1-z)},\quad \alpha_{12}\frac{\partial v_3}{\partial z}=-\frac{iv_1^*v_2^*}{z}\,. \tag{5.29}$$

We point out that Eqs. (5.29) coincide with the scaling reduction obtained for the 3WR equations in one spatial and one temporal dimension. Furthermore, if we consider imaginary solutions only, the system (5.29) is reducible to an equation related via a one-to-one transformation to the Painlevé VI equation.[4,14]

VI. CONCLUDING REMARKS

In this paper we have carried out a systematic group analysis of the three-wave resonant equations in two spatial and one temporal dimension. The results of our investigation show that the 2D-3WR system shares many features with other nonlinear integrable partial differential equations of physical significance in $(2+1)$ dimensions, such as the Kadomtsev–Petviashvili, Davey–Stewartson, and modified Kadomtsev–Petviashvili equations.[5–7] In fact, all these equations admit infinite-dimensional symmetry groups whose Lie algebras contain arbitrary functions and involve, as a particular case, Kac–Moody type algebras having a crucial role in the theory of integrable systems.

Concerning the 2D-3WR equations, we have found the connection between one- and two-dimensional symmetry subalgebras and reduction equations, which provide new solutions of the original system. Among these, Eqs. (5.11) and (5.29) deserve a special mention. The former is a nonlinear partial differential system in two independent variables and with nonconstant coefficients coming from the scale generator. It allows a linear spectral problem that has been derived using a prolongation procedure.[14] The latter is related to the Painlevé VI equation via a one-to-one transformation. Other reduced equations arising from our analysis, which seem new at the best of our knowledge, are (5.24). These will be considered in a forthcoming paper.[14]

We conclude by noting that a natural continuation of the present work is the search for the generalized symmetries of the 2D-3WR equations and their algebraic properties. This program will be dealt with in the near future.

[1] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).
[2] N. H. Ibragimov, *Transformation Group Applied to Mathematical Physics* (Reidel, Dordrecht, 1984).
[3] P. J. Olver, J. Differential Geom. **14**, 497 (1979).
[4] A. S. Fokas, R. A. Leo, L. Martina, and G. Soliani, "The scaling reduction of the three wave resonant system and the Painlevé VI equation," preprint Università di Lecce, Italy, 1985.
[5] D. David, N. Kamran, D. Levi, and P. Winternitz, "Symmetry reduction for Kadomtsev–Petviashvili equation using a loop algebra," preprint CRM, Université de Montréal, 1985.
[6] V. V. Kadomtsev and V. I. Petviashvili, Sov. Phys. Dokl., Fluid Mech. **15**, 539 (1970).
[7] A. Davey and K. Stewartson, Proc. R. Soc. London Ser. A **338**, 101 (1974).
[8] V. Kac, *Infinite Dimensional Lie Algebras* (Birkhauser, Boston, 1984).
[9] D. J. Kaup, A. H. Reiman, and A. Bers, Rev. Mod. Phys. **51**, 275 (1979).
[10] V. E. Zakharov, G. L. Musher, and A. M. Rubenchik, Phys. Rep. **129**, (5), 285 (1985).
[11] M. J. Ablowitz and R. Haberman, Phys. Rev. Lett. **35**, 1185 (1975).
[12] G. S. Tondo, Lett. Nuovo Cimento **44**, 297 (1985).
[13] D. J. Kaup, Physica D **1**, 5 (1980).
[14] A. R. Leo, L. Martina, G. Soliani, and G. Tondo, "On certain systems obtained from symmetry reduction of the 2D-3WR equations" (in preparation).
[15] F. Neuman, in *23rd International Symposium on Functional Equations*, Grignano, Italy, to be published by the University of Waterloo, Centre for Information Theory.
[16] M. Leo, R. A. Leo, G. Soliani, and L. Martina, Nuovo Cimento B **88**, 81 (1985).
[17] H. D. Wahlquist and F. B. Estabrook, J. Math. Phys. **16**, 1 (1975).
[18] F. B. Estabrook and H. D. Wahlquist, J. Math. Phys **17**, 1293 (1976).
[19] V. E. Zakharov and S. V. Manakov, Sov. Phys. JETP **42**, 842 (1976).

Generalized Green's functions and spectral densities in the complex energy plane

Erik Engdahl and Erkki Brändas
Quantum Chemistry Group for Research in Atomic, Molecular and Solid State Physics, University of Uppsala, Box 518, S-751 20 Uppsala, Sweden

Magnus Rittby and Nils Elander
Research Institute of Physics, Frescativägen 24, S-104 05 Stockholm, Sweden

(Received 7 October 1985; accepted for publication 2 July 1986)

The Titchmarsh–Weyl theory is applied to the Schrödinger equation in the case when the asymptotic form of the solution is not known. It is assumed that the potential belongs to the Weyl's limit-point classification. A rigorous analytical continuation of the Green's function, obtained from the solution regular at the origin and the square integrable Weyl's solution (regular at infinity), to the "unphysical" Riemann energy sheet is carried out. It is demonstrated how the Green's function can be uniquely constructed from the Titchmarsh–Weyl m-function and its Nevanlinna representation. The behavior of the m-function in the neighborhood of poles is investigated. The m-function is decomposed in a, so called, generalized real part (Reg) and a generalized imaginary part (Img). Reg(m) is found to have a significant argument change upon pole passages. Img(m) is found to be a generalized spectral density. From the generalized spectral density, a spectral resolution of the differential operator and its resolvent is derived. In the expansion contributions are obtained from bound states, resonance states (Gamow states), and the "deformed continuum" given by the generalized spectral density. The present expansion theorem is applicable to the general partial differential operator via a decomposition into partial waves. The numerical procedure involves all quantum numbers l and m, but for convenience, and with the inverse problem in mind, this study is focused on the case when the rotational quantum number equals zero. The theory is tested numerically and analyzed for an analytic model potential exhibiting a barrier and decreasing exponentially at infinity. The potential is Weyl's limit point at infinity and allows for an analytical continuation into a sector in the complex plane. An attractive feature of the generalized spectral density of the present potential is that the poles close to the real axis seem to exhaust or deflate the above-mentioned density inside the pole string. Outside this string the density rapidly approaches that of a free particle. This information is used to derive an approximate representation of the m-function in terms of poles and residues as well as free-particle background. In order to display the features mentioned above, the present study is accompanied with several plots of analytically continued quantities related to the Green's function.

I. INTRODUCTION

In a scattering experiment one is interested in the outcome of the interactions between colliding particles as measured, for instance, by the velocity dependence of the appropriate cross sections of the various processes.

An important feature of the theoretical description of the scattering experiment lies in the possibility of relating observed cross-section data with details of the interaction. In this analysis, short-range properties of the potential as well as long-range and background effects are found to be related to bound and quasibound states formed by the partners of the collision. And, conversely, the spectral density related to the physical process defines the potential uniquely.[1]

Even though atomic and molecular scattering theory is in a relatively privileged position from the viewpoint of rigorous foundation and applicative power, there are nevertheless open problems that need attention. One such problem is related to the question of how to decompose a cross section into resonance contributions and background effects. In a wider context this problem concerns a generalization of the Hamilton–Liouville time evolution generators to the complex energy or k-plane. Generalized spectral properties as well as spectral expansion theorems are therefore needed to analyze and classify the dynamics of the colliding system.

In a previous study,[2] we gave a formulation of scattering theory in terms of the classical Titchmarsh–Weyl theory.[3,4] We paid particular attention to the connection between various spectral densities and the scattering cross section. In addition to an asymptotic analysis of the densities, we also demonstrated how the resonance contribution could be uniquely defined and numerically calculated.

In order to extend the formulation, we will here devote attention to the following developments: continuation of the partial wave Green's function to the second Riemann sheet; evaluation of the generalized spectral density on the second sheet; analysis of the various generalized imaginary and real parts of the Titchmarsh–Weyl m-function; the Nevanlinna representation of the Titchmarsh–Weyl m-function and the associated Green's function; analytic extension of the Green's function based on the Nevanlinna formulation; de-

 0022-2488/86/112629-11$02.50

flation of the generalized spectral density and the behavior of the background contribution; and treatment of potentials with laborious tails, i.e., when the asymptotic form of the solution to the associated differential equation is unknown.

Since the formulation of the inverse problem in scattering theory is focused on the spectral density, the present generalization offers several possibilities. In addition to an analysis of the analytic properties of the interaction potential, it is also possible to use the generalized expansion techniques treated here to solve the Gel'fand–Levitan[5,6] and Marchenko[7,8] equations in connection with the inverse problem in a sector of the complex plane.

II. PRELIMINARY DEFINITIONS AND NOTATIONS

A. Definition of the appropriate Green's functions

Since the present investigation aims at the analytical properties of the differential equation and its associated Green's function we will give a rather detailed preliminary account of the actual equations. The starting point is the time-independent Schrödinger equation

$$(E-H_0)\psi = V\psi\,, \tag{2.1}$$

where

$$H_0(r) = -\frac{\hbar^2}{2\mu}\nabla^2 + V_0(r) \quad \text{and} \quad E = \frac{\hbar^2 k^2}{2\mu}\,, \tag{2.2}$$

with the boundary condition of regularity of $\psi(\mathbf{r})$ at the origin. Asymptotically, $\psi(\mathbf{r})$ then behaves as

$$\psi \sim e^{i\mathbf{kr}} + f(\Omega)e^{ikr}/r\,, \quad |\mathbf{r}|\to\infty\,. \tag{2.3}$$

We solve (2.1) formally by introducing the Green's function defined by

$$\begin{aligned}&(E-H_0(\mathbf{r}))G^+(\mathbf{r}-\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}')\,,\\ &G^+(\mathbf{r}-\mathbf{r}') \sim \text{outgoing waves for } \mathbf{r}' \text{ fixed}\,, \quad |\mathbf{r}|\to\infty\,,\\ &G^+(\mathbf{r}-\mathbf{r}')\to 0 \quad \text{for } \mathbf{r}' \text{ fixed}\,, \quad |\mathbf{r}|\to 0\,.\end{aligned} \tag{2.4}$$

Using (2.4), the formal solution is

$$\psi(\mathbf{r}) = \varphi(\mathbf{r}) + \int d\mathbf{r}'\, G^+(\mathbf{r}-\mathbf{r}')V(\mathbf{r}')\psi(\mathbf{r}')\,. \tag{2.5}$$

Equation (2.5) is an integral equation for $\psi(\mathbf{r})$, where $\varphi(\mathbf{r})$ is the plane wave solution of the homogeneous equation

$$(E-H_0)\varphi = 0\,. \tag{2.6}$$

We decompose $G^+(r-r')$ into partial waves, as is usually done, in the form

$$G^+(\mathbf{r}-\mathbf{r}') = \sum_l \frac{2\mu}{\hbar^2} G_l^+(r,r') \sum_{m=-l}^{l} Y_l^m(\Omega)\,\overline{Y_l^m(\Omega')}\,. \tag{2.7}$$

Inserting this in (2.4) we get

$$\sum_l \sum_{m=-l}^{l} (E-H_l(r))\frac{2\mu}{\hbar^2} G(r,r')Y_l^m(\Omega)\,\overline{Y_l^m(\Omega')} = \delta(\mathbf{r}-\mathbf{r}')\,, \tag{2.8}$$

where

$$H_l(r) = \frac{\hbar^2}{2\mu}\left(-\frac{1}{r}\frac{d^2}{dr^2}(r) + \frac{l(l+1)}{r^2}\right) + V_0(r)\,. \tag{2.9}$$

Noting that

$$\sum_l \sum_{m=-l}^{l} Y_l^m(\Omega)\,\overline{Y_l^m(\Omega')} = \sum_l \sum_{m=-l}^{l} \langle\Omega|lm\rangle\langle lm|\Omega'\rangle = \delta(\Omega-\Omega')\,, \tag{2.10}$$

we find

$$(2\mu/\hbar^2)(E-H_l(r))G_l^+(r,r') = \delta(r-r')/r^2\,. \tag{2.11}$$

For convenience we extract some constants from $H_l(r)$ by defining

$$H_l(r) = \frac{\hbar^2}{2\mu}L_l(r) \quad \text{and} \quad V_0(r) = \frac{\hbar^2}{2\mu}U_0(r)\,, \tag{2.12}$$

with

$$L_l(r) = -\frac{1}{r}\frac{d^2}{dr^2}(r) + \frac{l(l+1)}{r^2} + U_0(r)\,. \tag{2.13}$$

The partial wave Green's function in (2.8) now satisfies

$$(\lambda - L_l)G_l^+(r,r') = \delta(r-r')/r^2\,, \tag{2.14}$$

where the energy parameter λ is given by

$$\lambda = k^2 = (2\mu/\hbar^2)E\,. \tag{2.15}$$

For atomic units we instead get the connection $\lambda = 2E$. Equations (2.12) and (2.13) can be simplified to the following expressions:

$$(\lambda - \hat{L}_l)\hat{G}_l^+(r,r') = \delta(r-r') \tag{2.16}$$

and

$$\hat{L}_l(r) = -\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + U_0(r)\,, \tag{2.16'}$$

upon defining

$$\hat{G}_l^+(r,r') = rr'G_l^+(r,r')\,. \tag{2.17}$$

Since $\hat{G}^+(r,r')$ with r' fixed should be regular both for $r\to 0$ and $r\to\infty$, we find that $\hat{G}_l^+(r,r')$ must be proportional to $\psi_l(r_<)\chi_l^+(r_>)$, where $\psi_l(r)$ is regular for $r\to 0$ and $\chi^+(r)$ is regular for $r\to\infty$. Replacing the outgoing χ^+ by the ingoing χ_l^- we trivially get $\hat{G}_l^-$. Here $r_<$ and $r_>$ are the smaller and the larger of r and r', respectively.

Alternatively, the Green's functions $\hat{G}_l^\pm$ can be obtained from the operator resolvent $(\lambda \pm i\epsilon - \hat{L}_l)^{-1}$ associated with the differential operators (2.12) or (2.16), via an appropriate limiting procedure.

The proportionality constant is easily determined by integration of (2.16) over the junction point $r = r'$ and is found to be the reciprocal of the Wronskian between $\psi_l(r)$ and $\chi_l^+(r)$. Thus we get

$$\hat{G}_l^+(r,r') = \psi_l(r_<)\chi_l^+(r_>)/W(\psi_l\chi_l^+)\,. \tag{2.18}$$

In the free-particle case [$U_0(r)\equiv 0$], this reduces to

$$\hat{G}_l^+(r,r') = \hat{j}_l(r_<)\hat{h}_l^+(r_>)/W(\hat{j}_l\hat{h}_l^+)\,, \tag{2.19}$$

where $\hat{j}_l(\rho)$ and $\hat{h}_l^+(\rho)$ are Riccati–Bessel and Riccati–Hankel functions, respectively.

Above we have discussed the free-particle Green's function, i.e., G_0 and its generalization with the presence of a reference potential U_0. In what follows we will let $U_0\equiv 0$ and consider the Green's function incorporating the actual interaction potential $U = (\hbar^2/2\mu)^{-1}\,\mathrm{V}$. This Green's function

will assume the same form as the one in (2.18).

Before we proceed to an investigation of the analytic properties of G^+ and associated generalized expansions, we will first briefly discuss the classical Titchmarsh–Weyl theory.

B. The classical Titchmarsh–Weyl theory

To set up the framework for the analytical extension, we consider the radial Schrödinger equation in the form

$$\left(-\frac{d^2}{dr^2}+\left(U(r)+\frac{l(l+1)}{r^2}\right)-\lambda\right)\chi(r)=0\,. \tag{2.20}$$

We have dropped the subindex l indicating the rotational quantum number. In order to treat bound states and the continuum in a unified way, we will here first assume λ to contain a nonzero imaginary part.

We define two linearly independent solutions φ and ψ by the left boundary conditions at a point $r=a$ in the interval $(0,\infty)$:

$$\begin{pmatrix}\varphi & \psi\\ \varphi' & \psi'\end{pmatrix}_{r=a}=\begin{pmatrix}\sin\alpha & \cos\alpha\\ -\cos\alpha & \sin\alpha\end{pmatrix}. \tag{2.21}$$

If the potential is less singular than the Coulomb potential at the origin, a can be chosen zero without loss of generality. However, in general, and this concerns the Coulomb potential even if the latter is limit circle at $r=0$, a must be different from zero, with the angle α chosen in the interval $(-\pi/2, \pi/2)$ such that $\psi(r)$ is regular at the origin. The logarithmic derivative of $\psi(r)$ at the point $r=a$ is then $\psi'(a)/\psi(a)=\tan\alpha$.

Provided special care is exercised, the limit $a\to 0$ can be taken in the final spectral density.[2] Furthermore, one usually assumes $-\pi/2<\alpha<\pi/2$, but the limit $|\alpha|\to\pi/2$ can also be taken.[2–4,9]

Any solution to (2.20), except ψ, can be written in the form

$$\chi(r)=\varphi(r)+\psi(r)m(E)\,. \tag{2.22}$$

We do not worry about the overall proportionality constant here since it is only the logarithmic derivative that relates to the quantization condition via suitable boundary values. One important thing to observe is the possible existence of a pole in m occurring when χ and ψ satisfy the same boundary conditions in the limit of real λ. The m-coefficient in (2.22), which will be uniquely defined below, will be seen to be intimately connected to the spectral density associated with (2.20).

We now impose the following real right boundary condition on $\chi(r)$ at $r=b>a$:

$$\cos\beta\chi(b)+\sin\beta\chi'(b)=0\,, \tag{2.23}$$

for some β, where $-\pi\leqslant\beta\leqslant\pi$. Using that β is real implies

$$\mathrm{Im}(\chi'/\chi)_{r=b}=0\,. \tag{2.24}$$

Introducing the square bracket notation

$$[uv]=u\,\overline{v'}-u'\overline{v}\,, \tag{2.25}$$

(2.24) can be rewritten as

$$[\chi\chi](b)=0\,. \tag{2.26}$$

Inserting (2.22), this can be transformed to a condition on m:

$$([\varphi\varphi]+m[\psi\varphi]+\overline{m}[\varphi\psi]+m\overline{m}[\psi\psi])_{r=b}=0\,. \tag{2.27}$$

After using Green's formula and the relation $[\psi\varphi]=-\overline{[\varphi\psi]}$, this condition on m can be shown to force m to lie on a circle in the complex plane with center

$$C_b=-[\varphi\psi](b)/[\psi\psi](b) \tag{2.28}$$

and radius

$$R_b=1/|[\psi\psi](b)|\,. \tag{2.29}$$

Upon using Green's formula, it is a simple matter to show that

$$[\psi\psi](b)=2i\,\mathrm{Im}(\lambda)\int_0^b|\psi|^2\,dr \tag{2.30}$$

goes to infinity for $\mathrm{Im}(\lambda)\neq 0$, in the limit-point case, which includes most potentials of physical interest. A sufficient criterion for a real potential to belong to this classification is given, for instance, by Ref. 10 as

$$V(r)>-Kr^2\,, \tag{2.31}$$

for every $r>r_0$ and some $K>0$.

In the numerical applications, the POP ratio (phi over psi)

$$\mathrm{POP}(r)=\varphi(r)/\psi(r) \tag{2.32}$$

is computed. Provided the unique square integrable solution $\chi(r)$ vanishes at infinity, in the limit-point case, the POP ratio converges to $-m(\lambda)$ as $r\to\infty$ if $\mathrm{Im}(\lambda)\neq 0$ (see Ref. 11). We will refer to such a procedure as the use of the POP method.

For pathological cases, the limit-point solution $\chi(r)$ may not go to zero at infinity.[12] However, if χ' is also square integrable, then $\chi(r)$ must necessarily vanish at infinity,[12] and the POP method is applicable. If χ' is not square integrable, more general techniques as previously discussed in Ref. 11 must be applied. In what follows we will assume that both χ and χ' are square integrable on the interval $[a,\infty)$.

The circle procedure, and, in cases when χ' is square integrable, the POP method, can be used to calculate the wave function and the associated spectral information without explicit knowledge of the asymptotic form of the differential equation.

It can be shown that $m(\lambda)$ is an analytic function of Nevanlinna type (see below) whose imaginary part is a spectral density that will occur in the completeness relation

$$\delta(r-r')=\int_{-\infty}^{\infty}\psi(\omega,r)\psi(\omega,r')\,d\rho(\omega)\,, \tag{2.33}$$

where

$$\rho(\omega_2)-\rho(\omega_1)=\lim_{\epsilon\to 0^+}\frac{1}{\pi}\int_{\omega_1}^{\omega_2}\mathrm{Im}(m(\lambda+i\epsilon))\,d\lambda\,. \tag{2.34}$$

The function χ^{POP} obtained is of either outgoing or ingoing character depending on the sign of $\mathrm{Im}(\lambda)$. This can be visualized by the following argument.

Let us for simplicity assume that $rV(r)\to 0$ when $r\to\infty$. Any solution to (2.20) is then asymptotically a linear combi-

nation of outgoing and ingoing Jost solutions. (In the case of the Stark effect, these solutions would instead be Airy functions of exponential type.) We then get

$$\chi(r) \sim A \exp(ikr) + B \exp(-ikr) . \tag{2.35}$$

For energies λ (on the first Riemann sheet) with $\mathrm{Im}(\lambda) > 0$, we have $\mathrm{Im}(k)$ and $\mathrm{Re}(k) \geqslant 0$, and the second exponential blows up for $r \to \infty$. However, χ^{POP} is square integrable on the interval $[a, \infty)$. The POP method therefore inherently chooses the purely outgoing solution for those energies. Conversely, for complex energy values, with $\mathrm{Im}(\lambda) < 0$, the POP method gives the purely ingoing solution. In the general limit-point case, the concepts of in- and outgoing waves must be interpreted in an extended sense. Symbolically we get

$$\chi^{\mathrm{POP}} = \begin{cases} \chi^{+}, & \text{for } \mathrm{Im}(\lambda) > 0, \\ \chi^{-}, & \text{for } \mathrm{Im}(\lambda) < 0. \end{cases} \tag{2.36}$$

For the m-values obtained by the POP method, we similarly have

$$m^{\mathrm{POP}}(\lambda) = \begin{cases} m^{+}(\lambda), & \text{for } \mathrm{Im}(\lambda) > 0, \\ m^{-}(\lambda), & \text{for } \mathrm{Im}(\lambda) < 0. \end{cases} \tag{2.37}$$

It is possible to let the energy λ approach the continuum part of the real energy axis.[13] However, χ^{POP} is then no longer square integrable, albeit finite at infinity. Care must be taken as to which side of the real energy axis the limit is taken from. Compare, for instance, the corresponding problem of approaching the cut of the Green's function.

As can be seen from (2.35), the solution obtained by the Titchmarsh–Weyl–POP method (corresponding to the point limit of Weyl's circle) is square integrable on the interval $[a, \infty)$. For $\mathrm{Im}(\lambda) < 0$ (on the "unphysical" Riemann sheet), one notices that the corresponding outgoing waves are no longer square integrable.

Furthermore, this defines a unique formal solution to (2.20) that corresponds to a diverging outgoing Gamow wave. This boundary condition leads to the occurrence of complex eigenvalues. On the first sheet the spectral density occurring in the completeness relation (2.33) is, of course, confined to the real axis in agreement with the self-adjoint nature of the differential operator. Hence, when treating divergent waves associated with complex energies the computational procedure no longer converges to the desired solution. This problem, however, can be solved by a simple complex scaling trick.

Although the idea here is to find m even if the asymptotic form of the solutions $f^{\pm}$ is not known, it is easy to obtain the connection between m and f ($\pm$ branches not denoted). This connection has been given many times, see, e.g., Ref. 13. For $\alpha = \pi/2$ one gets

$$m = f'(\sqrt{\lambda}, a)/f(\sqrt{\lambda}, a) . \tag{2.38}$$

From this relation the spectral density exhibiting the properties of $m(\lambda)$ can be found.

III. DILATED VERSION OF THE TITCHMARSH–WEYL THEORY

To carry out the appropriate analytic modifications, we first consider Eq. (2.20) for complex radial distances r':

$$\left(-\frac{d^2}{dr'^2} + \left(U(r') + \frac{l(l+1)}{r'^2} \right) - \lambda \right) \chi(r') = 0 . \tag{3.1}$$

In (3.1) we have assumed the real potential $U(r)$ has an analytic extension to some sector in the complex plane, which, according to the Schwartz reflection principle, fulfills

$$U(r') = \overline{U(\bar{r}')} , \tag{3.2}$$

for complex values r'.

At the starting point $r' = a'$, we define two linearly independent initial value solutions φ and ψ by the boundary conditions

$$\begin{pmatrix} \varphi & \psi \\ \varphi' & \psi' \end{pmatrix}_{r' = a'} = \begin{pmatrix} \sin\alpha & \cos\alpha \\ -\cos\alpha & \sin\alpha \end{pmatrix} . \tag{3.3}$$

If the starting point $r' = a'$ is chosen nonreal, the angle α is nonreal in general. This somewhat more complicated situation will be treated in a later article. Here we restrict a' to be real, $a' = a \geqslant 0$, ensuring the initial value matrix in (3.3) to be real. In the present numerical applications, l is chosen to be zero (and $a = 0$) for convenience.

At this point it is suitable to introduce complex scaling. If a is positive, we are led to the utilization of exterior complex scaling.[14–16] Otherwise, if $a = 0$, as in our numerical investigation, we invoke uniform complex scaling.

We now consider (3.1) along the ray $r' = a + \eta(r - a)$, where r is a real radial distance coordinate and $\eta = |\eta| e^{i\theta}$ ($\theta \geqslant 0$), the complex scaling parameter. In what follows, r belongs to the interval $[a, \infty)$. Rewriting (3.1) using the definitions

$$\chi_\eta(r) = \chi(a + \eta(r - a)) \quad \text{and} \quad r_\eta = a + \eta(r - a) , \tag{3.4}$$

we arrive at

$$\left(-\frac{d^2}{dr^2} + \left(\eta^2 U(r_\eta) + \eta^2 \frac{l(l+1)}{r_\eta^2} \right) - \eta^2 \lambda \right) \chi_\eta(r) = 0 , \tag{3.5}$$

where the differentiations from now on refer to the real coordinate r.

Noting that

$$\chi'_\eta(r) = \eta \left(\frac{d\chi(r')}{dr'} \right)_{r' = a + \eta(r-a)} , \tag{3.6}$$

and similarly for $\varphi'_\eta(r)$ and $\psi'_\eta(r)$, the boundary condition (3.3) assumes the following form:

$$\begin{pmatrix} \varphi_\eta & \psi_\eta \\ \varphi'_\eta & \psi'_\eta \end{pmatrix}_{r = a} = \begin{pmatrix} \sin\alpha & \cos\alpha \\ -\eta\cos\alpha & \eta\sin\alpha \end{pmatrix} . \tag{3.7}$$

Imposing now the boundary condition

$$[\chi_\eta \chi_\eta](b) = 0 , \tag{3.8}$$

for $\chi_\eta(r)$ at $r = b$, we find in parallel with the real case that m must lie on a circle in the complex plane with center

$$C_b = -[\varphi_\eta \psi_\eta](b) / [\psi_\eta \psi_\eta](b) \tag{3.9}$$

and radius

$$R_b = 1/|[\psi_\eta \psi_\eta](b)| . \tag{3.10}$$

The formula corresponding to (2.30) is then

$$[\psi_\eta\psi_\eta](b) = [\psi_\eta\psi_\eta](a) + 2i\int_a^b |\psi_\eta|^2\Big(\mathrm{Im}(\eta^2\lambda) - \mathrm{Im}\Big(\eta^2 U(r_\eta) + \eta^2\frac{l(l+1)}{r_\eta^2}\Big)\Big)dr\,. \quad (3.11)$$

We note that $\mathrm{Im}(\eta^2\lambda)$ is the distance (including sign) from the ray $\eta^{-2}R^+$ in the complex energy plane.

We further note that (3.11) differs from (2.30) in the occurrence of the imaginary part of the effective potential in (3.5). This may perturb the convergence properties of Weyl's circle in the complex plane. For the model potential used in our study, the additional potential terms just mentioned will be dominated by the $\mathrm{Im}(\eta^2\lambda)$ term provided $\mathrm{Arg}(\eta) < \pi/2$. We therefore realize that the radius of Weyl's circle will shrink to zero provided the energy does not belong to the rotated cut $\eta^{-2}R^+$.

After complex scaling, the equation corresponding to (2.35) is now

$$\chi_\eta(r) \sim A\exp(ik\eta r) + B\exp(-ik\eta r)\,. \quad (3.12)$$

The sign of $\mathrm{Im}(k\eta)$, or equivalently the sign of $\mathrm{Im}(\eta^2\lambda)$, determines the behavior of the POP method where the modified POP ratio

$$\mathrm{POP}_\eta(r) = \varphi_\eta(r)/\psi_\eta(r) \quad (3.13)$$

is computed. By an argument similar to the real case, we get

$$\chi_\eta^{\mathrm{POP}} = \begin{cases}\chi_\eta^+\,, & \text{for } \mathrm{Im}(\eta^2\lambda) > 0\,,\\ \chi_\eta^-\,, & \text{for } \mathrm{Im}(\eta^2\lambda) < 0\,,\end{cases} \quad (3.14)$$

and for the m-values

$$m_\eta^{\mathrm{POP}}(\lambda) = \begin{cases}m^+(\lambda)\,, & \text{for } \mathrm{Im}(\eta^2\lambda) > 0\,,\\ m^-(\lambda)\,, & \text{for } \mathrm{Im}(\eta^2\lambda) < 0\,.\end{cases} \quad (3.15)$$

Furthermore it follows that χ_η^+ and χ_η^- are square integrable on $[a,\infty)$ as long as λ does not lie on the rotated cut $\eta^{-2}R^+$. A sufficient condition for the radius $R_b \to 0$ as $b \to \infty$ is then that the integral in (3.11) diverges to infinity.

IV. GENERALIZED SPECTRAL EXPANSION OF GREEN'S FUNCTIONS AND THE RESOLUTION OF THE IDENTITY

A. Spectral expansion over real energy states

We will now derive a spectral resolution of the Green's function (2.18). In addition to dropping the subindex l, we also temporarily suppress the coordinate dependence of the Green's function. The resulting equation can also be looked upon as a resolvent operator expression. One can then write

$$\hat{G}^+(\lambda) = \int_{-\infty}^{\infty}\frac{d\hat{\tau}(\omega)}{\lambda + i0 - \omega} = \sum_j\frac{\mathrm{Res}\,\hat{G}^+(\lambda_j)}{\lambda + i0 - \lambda_j} + \int_0^\infty\frac{[d\hat{\tau}(\omega)/d\omega]d\omega}{\lambda + i0 - \omega}\,, \quad \lambda \neq \lambda_j\,. \quad (4.1)$$

In order to interpret (4.1), we will compare this formula with (2.18) of Sec. II A. This is easily done by letting λ approach the real axis. Let us for the moment take λ real and indicate the limit from the upper part of the complex plane by $\lambda + i0$. As will be seen below, the resolvent representation is based on an analytic representation of Nevanlinna type. As a consequence both $\mathrm{Res}\,\hat{G}^+(\lambda)$ and $d\hat{\tau}(\omega)/d\omega$ $(\omega \geqslant 0)$ are real quantities. Now we extract the imaginary part of (4.1) by using the well-known distribution formula

$$\frac{1}{\lambda + i0 - \omega} = P\frac{1}{\lambda - \omega} - i\pi\delta(\lambda - \omega)\,, \quad (4.2)$$

obtaining

$$\mathrm{Im}\,\hat{G}^+(\lambda) = -\pi\Big(\frac{d\hat{\tau}(\omega)}{d\omega}\Big)_{\omega=\lambda}\,. \quad (4.3)$$

From (2.18) we also get, using $\chi^+ = \varphi + \psi m^+$ and $W(\varphi\psi) = 1$, that

$$\mathrm{Im}\,\hat{G}_l^+(r,r') = -\mathrm{Im}(\psi(r_<)\chi^+(r_>)) = -\psi(r)\psi(r')\mathrm{Im}(m^+)\,, \quad (4.4)$$

where the notations $r_<$ and $r_>$ have been dropped since they occur symmetrically. We have used here that φ and ψ are real since the differential operator as well as the energy is real and the initial conditions are real by construction. We now rewrite (4.1) in the coordinate representation, using (4.3) and (4.4) to the form

$$\hat{G}^+(\lambda;r,r') = \sum_j\frac{\psi(\lambda_j,r)\psi(\lambda_j,r')/\langle\psi(\lambda_j)|\psi(\lambda_j)\rangle}{\lambda + i0 - \lambda_j} + \int_0^\infty\frac{\psi(\omega,r)\psi(\omega,r')(1/\pi)\mathrm{Im}(m^+(\omega))d\omega}{\lambda + i0 - \omega} = \int_{-\infty}^{\infty}\frac{\psi(\omega,r)\psi(\omega,r')d\rho(\omega)}{\lambda + i0 - \omega}\,, \quad (4.5)$$

where

$$\frac{d\rho(\omega)}{d\omega} = \begin{cases}\sum_j\dfrac{\delta(\omega - \lambda_j)}{\langle\psi(\lambda_j)|\psi(\lambda_j)\rangle}\,, & \text{for } \omega < 0\,,\\ \dfrac{1}{\pi}\mathrm{Im}(m^+(\omega))\,, & \text{for } \omega \geqslant 0\,.\end{cases} \quad (4.6)$$

Using (2.18) and (4.1) we further find that the residue parts occurring in (4.1) are related to the normalization of the bound states. By letting the operator $(\lambda + i0 - \hat{L})$ work on (4.5), we immediately conclude that the spectral density in (4.6) is exactly the one occurring in (2.33).

As a general remark it is also possible to relate (4.6) to the Kodaira form needed, for instance, in the Gel'fand–Levitan equation for the inverse problem. In this case one finds that

$$\frac{d\rho(\omega)}{d\omega} = \frac{k^{2l+1}}{|f_l(k)|^2}\,, \quad \text{for } \omega \geqslant 0\,, \quad k^2 = \omega\,, \quad (4.7)$$

where $f_l(k)$ is the well-known Jost function. See, for instance, (1).

B. The Nevanlinna representation

It is obvious that the Green's function previously discussed is uniquely defined once the Titchmarsh–Weyl m-function is known for all energies. However, in order to analytically extend the Green's function [or $m(\lambda)$] it is necessary to study the appropriate analytical properties in more detail. From the Titchmarsh–Weyl theory, it can be shown that $m(\lambda)$ belongs to a class of functions said to be of Nevanlinna type. We will here briefly give the definition.

A function $f(z)$ is said to be of Nevanlinna type if it maps the upper(lower) complex half-plane onto itself. It is, of course, assumed that $f(z)$ is Cauchy analytic in each half-plane with possible singularities on the real axis. Furthermore the theory of Nevanlinna functions leads to the existence of a uniquely defined function $\sigma(\omega)$, called the spectral function of $f(z)$. For general properties of $\sigma(\omega)$, see Refs. 9 and 17. For an example see Fig. 1 where a suitable integration contour for the Cauchy representation of $f(z)$ is displayed. From this it is easy to show that $f(z)$ can be written in the form

$$f(z) = -\sum_j \frac{\mathrm{Res} f(z_j)}{z_j - z} + \int_0^\infty \frac{(1/\pi)\mathrm{Im} f(\omega + i0)d\omega}{\omega - z} = \int_{-\infty}^{\infty} \frac{d\sigma(\omega)}{\omega - z}, \tag{4.8}$$

where

$$\frac{d\sigma(\omega)}{d\omega} = \begin{cases} \sum_j -\mathrm{Res} f(z_j)\delta(\omega - z_j), & \text{for } \omega < 0, \\ \frac{1}{\pi}\mathrm{Im} f(\omega + i0), & \text{for } \omega \geqslant 0. \end{cases} \tag{4.9}$$

In addition, the Nevanlinna property defines $\sigma(\omega)$ as a nondecreasing function of ω. It is required that $|f(z)| \to 0$ sufficiently fast for $|z| \to \infty$ so that the contribution from the circle C_R vanishes for $R \to \infty$.

A good candidate for this representation is $f(\lambda) = m(\lambda) - m_{\mathrm{free}}(\lambda)$, where the free-particle m-function is given by $m_{\mathrm{free}}(\lambda) = i\sqrt{\lambda}$. Here we restrict ourselves to the case $l = 0$ and $a = 0$, although the general Green's function formula that we will obtain can be found for any rotational quantum number l. Hence we get

$$m(\lambda) - i\sqrt{\lambda} = \sum_j \frac{-\mathrm{Res}\, m(\lambda_j)}{\lambda_j - \lambda} + \int_0^\infty \frac{(1/\pi)(\mathrm{Im}(m^+(\omega)) - \sqrt{\omega})d\omega}{\omega - \lambda} = \int_{-\infty}^{\infty} \frac{d\sigma(\omega)}{\omega - \lambda}, \tag{4.10}$$

where

$$d\sigma(\omega) = d\rho(\omega) - d\rho_{\mathrm{free}}(\omega) = \begin{cases} d\rho(\omega) - d(2\omega^{3/2}/3\pi), & \text{for } \omega \geqslant 0, \\ d\rho(\omega), & \text{for } \omega < 0, \end{cases} \tag{4.11}$$

with $\rho(\omega)$ given by (4.6).

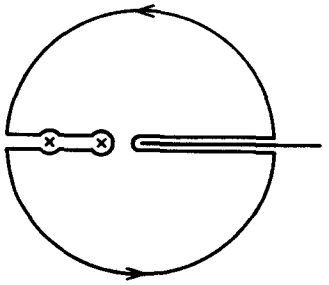


FIG. 1. Integration contour for the Cauchy representation showing the necessary deformation around bound states and the cut along the positive real energy axis.

C. Deformation of the integration contour in the Nevanlinna representation of $m(\lambda)$

We are now going to deform the integration contour in (4.10) from $\omega = 0$ to $\omega = \infty$ and therefore we need the analytical continuation of $\mathrm{Im}(m^+(\omega))$ occurring in the spectral density.

From the analysis of Sec. III we also assume that the relations (3.14) and (3.15) are valid. This means that the pair of functions $m^+(\lambda)$ and $m^-(\lambda)$ have analytic continuations onto a higher-order Riemann sheet. If $m^+(\lambda)$ and $m^-(\lambda)$ are nonreal for real energies, their respective imaginary parts differ only by their signs in agreement with the Nevanlinna character previously discussed. Hence, for real energies we realize that

$$\mathrm{Im}(m^+(\lambda)) = (m^+(\lambda) - m^-(\lambda))/2i. \tag{4.12}$$

The analytic extension $\mathrm{Img}(m^+(\lambda))$, here called the generalized imaginary part of $m^+(\lambda)$, is then immediately given by the left-hand side of (4.12). When we deform the contour in the way shown in Fig. 2, we must take care of the residues of $\mathrm{Img}(m^+(\lambda))$ at the resonance energies corresponding to Gamow waves. The residue contributions corresponding to resonances will appear together with bound states, as can be understood from the figures. Since $m^-(\lambda)$ has no poles in the lower half energy plane [$m^-(\lambda)$ is by definition evaluated on the first Riemann sheet], the formula (4.12) immediately gives

$$\mathrm{Res}(\mathrm{Img}(m^+(\lambda))) = \mathrm{Res}(m^+(\lambda))/2i. \tag{4.13}$$

From the analytic information above the following generalized Nevanlinna representation holds:

$$m(\lambda) - i\sqrt{\lambda} = \sum_j \frac{-\mathrm{Res}\, m(\lambda_j)}{\lambda_j - \lambda} + \int_C \frac{(1/\pi)(\mathrm{Img}(m^+(\omega)) - \sqrt{\omega})d\omega}{\omega - \lambda}, \tag{4.14}$$

where C is the positive real axis rotated downwards twice the argument of η.

D. Spectral expansion of the Green's function with deformed integration contours in the complex energy plane

We are now going to deform the integration contour in (4.5) in the same way as we did in the generalized Nevanlinna representation (4.14). Thus we also need the analytical continuation of the wave function product $\psi(\omega,r)\psi(\omega,r')$.

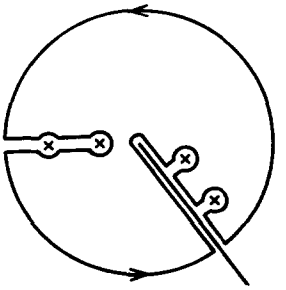


FIG. 2. Integration contour for the Cauchy representation showing the necessary deformation around bound states as well as resonance poles and the rotated cut in the complex energy plane.

We note that since $\psi(\omega,r)\psi(\omega,r')$ occurs in an operator kernel the last factor should be a complex conjugated quantity. This is of no importance here as long as both ω and r are real since then the wave functions are real. The appropriate analytical continuation is then given by $\psi(\omega,r)\overline{\psi(\bar{\omega},\bar{r}')}$. Since $\psi(\omega,r)$ is real for real ω and r, the two conjugations will annihilate each other here.

By defining the residue of the m^+-function at a pole (bound or resonance states) by

$$\mathrm{Res}(m^+(\lambda_j)) = -\langle\psi(\bar{\lambda}_j)|\psi(\lambda_j)\rangle^{-1}, \tag{4.15}$$

where the scalar product takes the general form

$$\langle\psi(\bar{\lambda}_j)|\psi(\lambda_j)\rangle = \int_{\eta R^+}\psi(\lambda_j,r')\psi(\lambda_j,r')dr', \tag{4.16}$$

we arrive at the following generalized spectral resolution of the Green's function:

$$\hat{G}^+(\lambda;r,r') = \sum_j \frac{\psi(\lambda_j,r)\psi(\lambda_j,r')\mathrm{Res}(-m^+(\lambda_j))}{\lambda-\lambda_j} + \int_C \frac{\psi(\omega,r)\psi(\omega,r')(1/\pi)\mathrm{Img}(m^+(\omega))d\omega}{\lambda-\omega}, \quad \lambda\neq\lambda_j, \tag{4.17}$$

where r,r' lie on the ray ηR^+, and C is the rotated continuum $\eta^{-2}R^+$, as before. Note that the discrete sum in (4.17) contains, in addition to bound states, also those resonances that are exposed by the complex rotation, see Fig. 3. This is equivalent to saying that the corresponding Gamow waves are square integrable, i.e., the integral (4.16) converges. All other nonexposed poles are contained in the generalized spectral density occurring in the integral part of (4.17).

As in the conventional case, we now apply the operator $(\lambda-\hat{L}(r))$ on (4.17) obtaining

$$\delta(r-r') = \sum_j \psi(\lambda_j,r)\psi(\lambda_j,r')\mathrm{Res}(-m^+(\lambda_j)) + \int_C \psi(\omega,r)\psi(\omega,r')d\rho(\omega), \tag{4.18}$$

where

$$\frac{d\rho(\omega)}{d\omega} = \frac{1}{\pi}\mathrm{Img}(m^+(\omega)) \tag{4.19}$$

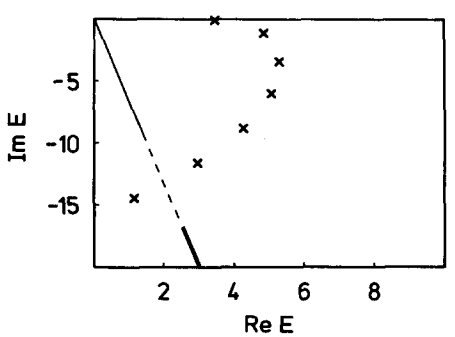


FIG. 3. Pole string for $m^+(\lambda)$ ($\lambda = 2E$). The ray $\eta^{-2}R^+$ displays the integration contour for the Nevanlinna representation (4.14). Along the weakly drawn line the generalized spectral density is vanishingly small. Outside, along the thick solid line the generalized spectral density is that of a free particle. The broken line shows a transition region.

is a generalized spectral density. Since r,r' belong to the ray ηR^+, the delta function in (4.18) is different compared with that occurring in (2.33). It belongs to the space of ultradistributions. See Ref. 18 for more details.

V. NUMERICAL APPLICATION AND RESULTS

A. Numerical integration and the convergence of the POP ratio

The de Vogelaere method[19] for numerical integration of second-order differential equations without a first-derivative term has been implemented in the program POP. The method is suitable since the initial values required for the algorithm consist of the value of the function and its derivative at the starting point. Compare, for instance, with (3.7).

The numerical integration of Eq. (3.5) for the two solutions φ_η and ψ_η continues until convergence of the POP ratio (3.13) is reached. One of the convergence criteria that we have used is based on the change of the POP ratio, which is computed successively for a specified increment of radial distance. If the absolute value of the change is smaller than a certain value supplied by the input, convergence is assumed.

B. Computation and analysis of quantities related to m

The POP method has been tested numerically and analyzed for a model potential,[20] which in atomic units ($\lambda = 2E$) is given by $V(r) = 7.5r^2e^{-r}$. The potential is Weyl's limit point at infinity and allows an analytical continuation into a sector in the complex plane. This potential gives no bound states but it has a barrier that results in the pole string shown in Fig. 3.

For convenience we have only treated the case when the rotational quantum number is zero and the left end point of the interval is $a = 0$.

The program POP gives us the possibility of computing several quantities related to the Titchmarsh–Weyl function $m(\lambda)$. They consist of some special combinations of $m^+(\lambda)$ and $m^-(\lambda)$. Some of these quantities are the analytical continuations of quantities usually studied only for real energies.

The various quantities are as follows.

(1) $m^+(\lambda)$ for energies λ continued to the second Riemann sheet of the lower energy half-plane. $m^+(\lambda)$ is computed using a specified complex rotation angle θ. This allows the energies λ to be situated in the half-plane determined by $\mathrm{Im}(e^{i2\theta}\lambda) > 0$.

First we note the general property $m^+(\lambda) = m^-(\bar{\lambda})$. Therefore, by taking the mirror image of that plot in Fig. 8, see below, with respect to the real energy axis, we get the modulus plot of $m^+(\lambda)$ for our model potential on the first sheet of the upper half-plane.

In Figs. 4 and 5 we show the analytical continuation of $m^+(\lambda)$ onto the second sheet of the lower half energy plane. In addition to the string of poles of m^+, we also find zeros situated between the poles along the same string.

If the potential had been absent, the equimodular contours would have been circles centered at the origin. Depending on the chosen axis setup, this virtually corresponds to ellipses on the plots. With the potential present, the equi-

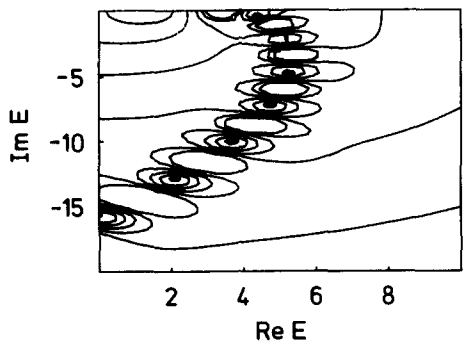


FIG. 4. Abs($m^+(\lambda)$).

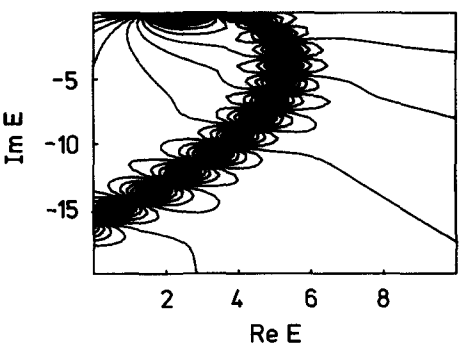


FIG. 6. Arg($m^+(\lambda)$).

modulars curve around one or several of the poles. Figure 6 indicates that the argument of the m^+-function along a ray from the origin is almost constant aside from the pole string region. Looking at the three-dimensional plot in Fig. 7, we see that passing the poles implies an argument shift of around π radians.

On the argument plot we also see some "slip faults" from the poles and the zeros of the m^+-function. This is further described in the section about Reg($m^+(\lambda)$) below.

(2) $m^-(\lambda)$ for energies λ on the first Riemann sheet. $m^-(\lambda)$ is computed using the complex rotation angle $-\theta$ and thus allows the energies to be situated in the half-plane determined by $\mathrm{Im}(e^{-i2\theta}\lambda) < 0$.

Since the solution, regular at the origin, is never square integrable for energies on the first sheet [except for bound states energies where $m(\lambda)$ has a pole] $m^-(\lambda)$ exhibits no pole in the lower half energy plane.

When analytically continued to the second sheet of the upper half-plane, there may be poles in $m^-(\lambda)$. The poles lying close to the real axis may influence the behavior of $m^-(\lambda)$ even on the first sheet.

Figure 8 shows the modulus of $m^-(\lambda)$ for our model potential on the first sheet. The two poles closest to the real axis can be detected from the concentration of contours near it. We see that at larger absolute values of the energy the m^--function rapidly approaches the value corresponding to a free particle, i.e., $m^-(\lambda) = -i\sqrt{\lambda}$. This can also be seen from the argument contour plot in Fig. 9.

(3) The generalized imaginary part of m^+, $\mathrm{Img}(m^+(\lambda)) = (m^+(\lambda) - m^-(\lambda))/2i$. Img($m^+(\lambda)$) is the analytical continuation of Im($m^+(\lambda)$) for real energies. Note that this quantity is complex in general. This quantity is proportional to the generalized spectral density occurring in the spectral resolution of the Green's function and in the Nevanlinna and related representations of the m-function.

For our model potential, Fig. 10 shows that Img($m^+(\lambda)$) is almost zero inside the pole string. This, together with the observed properties of Img($m^+(\lambda) - m^+_{\mathrm{free}}(\lambda)$), described below, is the empirical motivation for the conjecture of deflation, see below.

(4) The generalized real part of m^+, $\mathrm{Reg}(m^+(\lambda) + m^-(\lambda))/2$. Reg($m^+(\lambda)$) is the analytical continuation of Re($m^+(\lambda)$) for real energies. This quantity is also complex in general.

For real energies, Re($m^+(\lambda)$) can be used to localize poles on the second sheet since it shows sign shifts upon passing poles lying close to the real axis. For complex energies, one can therefore assume that Reg($m^+(\lambda)$) exhibits some phase change upon pole passages.

When scanning along lines close to poles that do not lie near to the real axis, it is possible to mimic the above-mentioned behavior either by considering Re($e^{i\alpha} m^+(\lambda)$), where

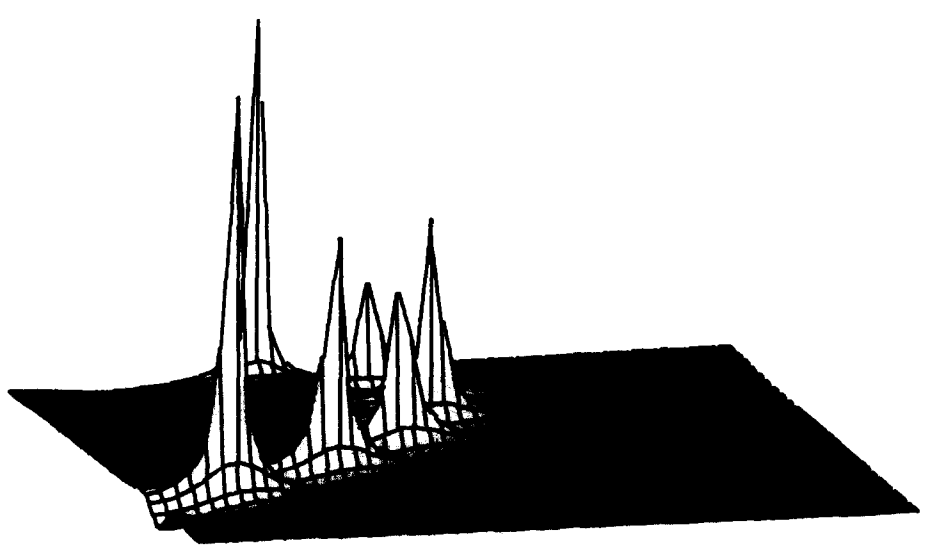


FIG. 5. Abs($m^+(\lambda)$).

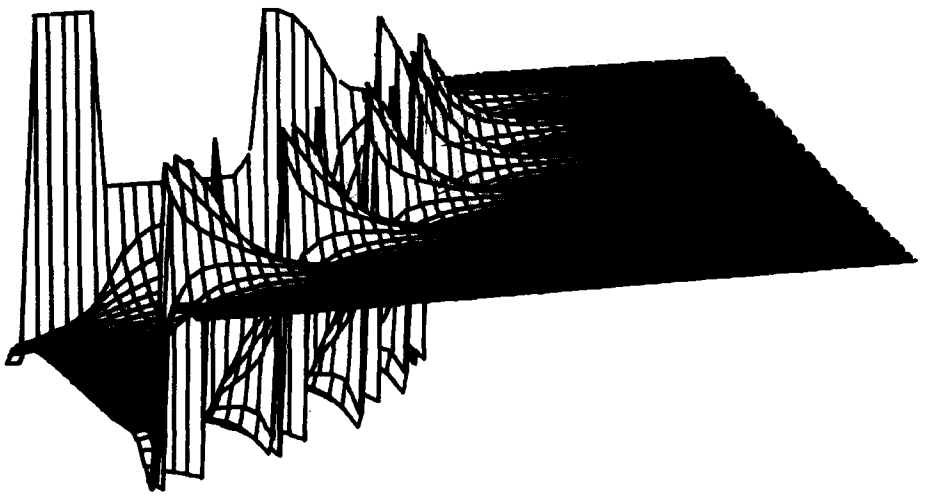


FIG. 7. Arg$(m^+(\lambda))$.

$e^{i\alpha}$ is a suitable chosen phase factor, or by letting the scanning line go in a suitable chosen direction depending on the phase of the residue of the m-function at the resonance pole.

The argument of Reg$(m^+(\lambda))$ for our model potential is shown in Fig. 11. Inside the pole string the argument seems to be almost constant. Outside, where the modulus of Reg$(m^+(\lambda))$ rapidly goes to zero, the argument increases steadily following curves locally parallel to the pole string. The argument increases slowly for trajectories far away, whereas it increases quickly inside the string defined by the zeros of Reg$(m^+(\lambda))$. In the latter case we find "slip faults" in the argument plot since the program POP uses the convention that the argument should be in the interval from $-\pi$ to π. The slip faults go from the zeros of Reg(m^+) to the poles of Reg(m^+) (or m^+ itself). Encircling any slip fault end point shows a complete argument cycling.

(5) Img$(m^+(\lambda) - m^+_{\text{free}}(\lambda))$. This quantity occurs in the spectral density in the Nevanlinna representations for $m^+(\lambda)$. It is also a necessary ingredient in the Gel'fand–Levitan integral equation for the inverse problem. Figure 12 shows the modulus of Img$(m^+(\lambda) - m^+_{\text{free}}(\lambda))$ for our model potential. Our numerical study shows that this quantity is negligible outside the pole string. Inside, where Img$(m^+(\lambda))$ is negligible, it assumes approximately the value $-$ Img$(m^+_{\text{free}}(\lambda))$. Comparing with Fig. 10 we find that only in a narrow transition region do we have a more complicated situation.

We are thus led to a conjecture of a deflation effect. Referring to Figs. 3 and 13–15, we now discuss the following approximation. For the integral appearing in (4.14), we assume a sharp transition point between a negligible and a free-particle spectral density located somewhere on the dashed part of the integration contour. This point will be denoted by Ω. With this approximation we find that the integral occurring in (4.14) is explicitly evaluable. This leads to the following approximate representation of the m-function in terms of poles and residues as well as free-particle background and pole–background interaction:

$$m(\lambda) - i\sqrt{\lambda} = \sum_j \frac{-\operatorname{Res} m(\lambda_j)}{\lambda_j - \lambda} - \frac{2}{\pi}\sqrt{\Omega} + \frac{\sqrt{\lambda}}{\pi} \ln \frac{(\sqrt{\lambda} + \sqrt{\Omega})}{(\sqrt{\lambda} - \sqrt{\Omega})}. \tag{5.1}$$

C. Computation of resonance energies

At a resonance energy $\lambda = \lambda_j$, $m^+(\lambda)$ exhibits a pole. Therefore $1/m^+(\lambda)$ has a zero. By the program POP, these zeros can be searched for using a Newton–Raphson-like method. For such a computation the complex rotation angle must be chosen large enough so as to uncover the pole, i.e., make the corresponding Gamow wave function square integrable.

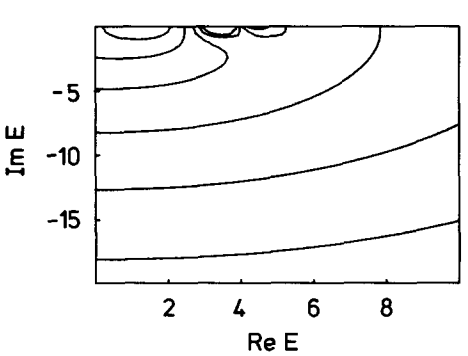


FIG. 8. Abs$(m^-(\lambda))$.

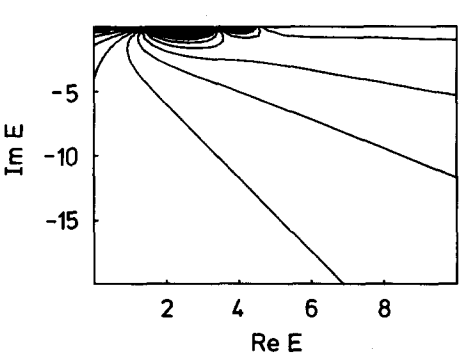


FIG. 9. Arg$(m^-(\lambda))$.

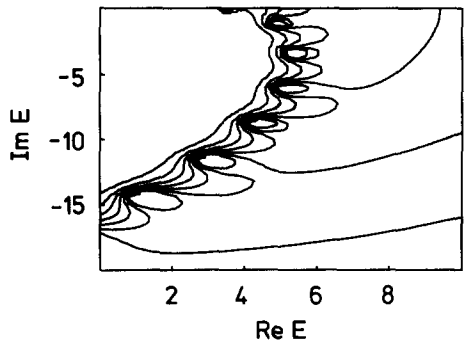


FIG. 10. Abs(Img($m^+(\lambda)$)).

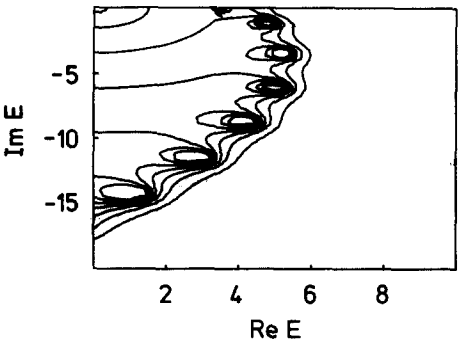


FIG. 12. Abs(Img($m^+(\lambda) - m^+_{\text{free}}(\lambda)$)).

VI. CONCLUSIONS

We have investigated a generalization of the Green's function and the associated completeness relation to a sector in the complex plane defined by the complex rotation method. For bound states, McIntosh[21] has studied the precise effect of the quantization condition for quantum mechanics in Weyl's theory formulation of second-order differential equations. Furthermore he analyzed the associated Green's matrix expression with regard to the corresponding discontinuities. The present study extends this discussion to resonances in the complex plane.

Previously, Berggren[22] has obtained analytically continued Green's functions using Zel'dovich regularization without any complex deformation of the coordinate. Investigations by Gyarmati and Vertse[23] show that the result obtained in the Zel'dovich framework[24] is independent of the convergence factor used. In fact their analysis coincides with the technique of uniform and exterior scaling for the localization of the Gamow wave.

The present study shows that the Berggren continuation, with the Zel'dovich regularization replaced by complex rotations, yields a rigorous formulation of the Green's function and the completeness relation in a sector in the complex plane. This approach has the additional feature of providing a consistent relation between the analytic properties of the potential and the spectral density.[6]

Our numerical results display, in addition to the above-discussed deflation property of the spectral density, that the Green's function can be approximately represented by bound state and resonance poles and a background contribu-

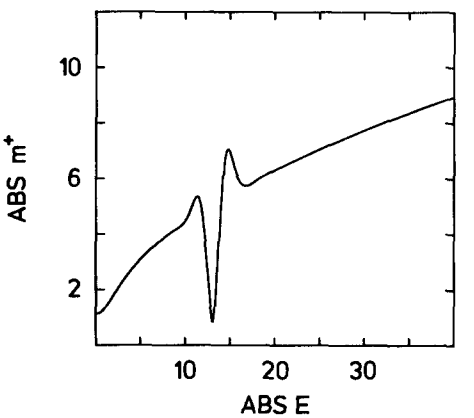


FIG. 13. Abs($m^+(\lambda)$) along the ray displayed in Fig. 3.

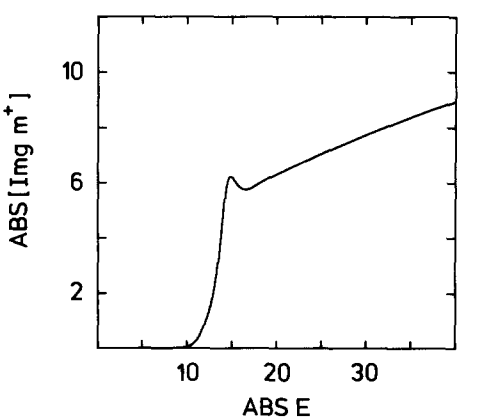


FIG. 14. Abs(Img($m^+(\lambda)$)) along the ray displayed in Fig. 3.

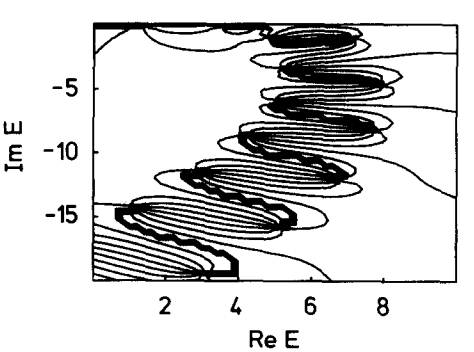


FIG. 11. Arg(Reg($m^+(\lambda)$)).

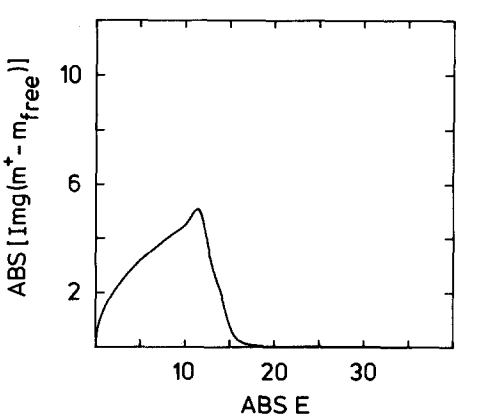


FIG. 15. Abs(Img($m^+(\lambda) - m^+_{\text{free}}(\lambda)$)) along the ray displayed in Fig. 3.

tion. The formula (5.1) has been preliminarily tested and found to contain the relevant information necessary for the construction of the Green's function. The string of corresponding residues used in formula (5.1) shows a smooth trajectory pattern in the complex plane in parallel to the complex pole distribution. For bound states, the residue of m is a negative quantity related to the normalization integral, see (4.15). For resonances with small widths, the real part of $\mathrm{Res}(m)$ is still negative but the imaginary parts are nonzero. When the resonance pole trajectory reaches its maximum real energy and turns, see Fig. 3, we found that for the present potential the real part of $\mathrm{Res}(m)$ changes sign. This seems to represent an alternative way of classifying the resonance poles in a primary and secondary class, see Ref. 25. The primary class, with a negative real part of $\mathrm{Res}(m)$, would then correspond to detectable structures in, for instance, the cross section.

ACKNOWLEDGMENTS

We are indebted to Ms. Lisa Toftemark for linguistic help.

One of us (EE) is grateful for financial support from "Sederhoms fond för utrikes resor" for a research stay at the Quantum Theory Project, Gainesville, Florida.

[1]K. Chadan and P. C. Sabatier, *Inverse Problems in Scattering Theory* (Springer, New York, 1977).
[2]E. Brändas, M. Rittby, and N. Elander, J. Math. Phys. **26**, 2648 (1985).
[3]H. Weyl, Ann. Math. **68**, 220 (1910).
[4]E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second-Order Differential Equations* (Oxford U. P., London, 1946).
[5]I. M. Gel'fand and B. M. Levitan, Izv. Akad. Nauk. SSSR Math. **15**, 309 (1951) [Am. Math. Soc. Trans. Ser. 2 **1**, 253 (1955)].
[6]E. Brändas, E. Engdahl, M. Rittby, and N. Elander, Int. J. Quantum Chem. **19S**, 457 (1985).
[7]V. A. Marchenko, Dokl. Akad. Nauk. SSSR **104**, 659 (1955); Math. Rev. **17**, 740 (1956).
[8]M. Rittby, N. Elander, and E. Brändas, "On the analytical continuation of the partial wave S-matrix using complex scaling techniques," to be published.
[9]W. N. Everitt and C. Bennewitz, in "Tribute to Åke Pleijel," in *Proceedings of the Conference*, September 1979, Department of Mathematics, University of Uppsala, Uppsala, Sweden (U. Uppsala, Uppsala, 1980), p. 49.
[10]E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955).
[11]M. Hehenberger, B. Laskowski, and E. Brändas, J. Chem. Phys. **65**, 4559 (1976).
[12]R. D. Richtmyer, *Principles of Advanced Mathematical Physics* (Springer, New York, 1978).
[13]E. Brändas, M. Hehenberger, and H. V. McIntosh, Int. J. Quantum Chem. **9**, 103 (1975).
[14]B. Simon, Phys. Lett. A **71**, 211 (1979).
[15]J. D. Morgan and B. Simon, J. Phys. B **14**, L167 (1981).
[16]M. Rittby, N. Elander, and E. Brändas, Int. J. Quantum Chem. **23**, 865 (1983).
[17]N. I. Akhiezer, *The Classical Moment Problem* (Oliver and Boyd, Edinburgh, 1965).
[18]I. M. Gel'fand and N. Ya. Vilenkin, *Generalized Functions* (Academic, New York, 1964), Vol. 4.
[19]W. A. Lester, Jr., J. Comp. Phys. **3**, 322 (1968).
[20]R. A. Bain, J. N. Bardsley, B. R. Junker, and C. V. Sukumar, J. Phys. B **7**, 2189 (1974).
[21]H. V. McIntosh, in *Quantum Science, Methods and Structure. A Tribute to Per-Olov Löwdin*, edited by J.-L. Calais, O. Goscinski, J. Linderberg, and Y. Öhrn (Plenum, New York, 1976), p. 277.
[22]T. Berggren, Nucl. Phys. A **109**, 265 (1968).
[23]B. Gyarmati and T. Vertse, Nucl. Phys. A **160**, 523 (1971).
[24]Ya. B. Zel'dovich, Zh. Eksp. Teor. Fiz. **39**, 776 (1960).
[25]M. Rittby, N. Elander, and E. Brändas, Phys. Rev. A **24**, 1636 (1981).

Auto-Bäcklund transformation, Lax pairs, and Painlevé property of a variable coefficient Korteweg–de Vries equation. I

N. Nirmala and M. J. Vedan
Department of Mathematics and Statistics, University of Cochin, Cochin 682 022, India
B. V. Baby[a)]
Department of Mathematics, Bharata Mata College, Cochin 682 021, India

(Received 6 January 1986; accepted for publication 11 July 1986)

Using the Painlevé property of partial differential equations, the auto-Bäcklund transformation and Lax pairs for a Korteweg–de Vries (KdV) equation with time-dependent coefficients are obtained. The Lax pair criterion also makes it possible for some new models of the variable coefficient KdV equation to be found that can represent nonsoliton dynamical systems. This can explain the wave breaking phenomenon in variable depth shallow water.

I. INTRODUCTION

Exciting and important discoveries have been made in the nonlinear dynamics of dissipative and conservative systems. Numerical, analytical, and experimental works in the last two decades show that most of the nonlinear systems exhibit a transformation from "regular" to "chaotic" behavior.[1] Recently,[2,3] the connection between movable singularities and algebraic integrability of dynamical systems is widely studied in different contexts. For an algebraically completely integrable system the independent, single-valued integrals of motions are part of a compact, complex tori on which the motion is linear.

Ward[4] has extended the study of the Painlevé property (PP), well known in the context of ordinary differential equations, to partial differential equations (PDE's). A system of PDE's in n independent variables is considered in the complex domain, the coefficients being analytic on C^n. If S is an analytic noncharacteristic complex hypersurface in C^n, then the PDE that is analytic on S is meromorphic on C^n. A weaker form of the PP was suggested by Weiss *et al.*[5] while studying the Lorentz series expansion of the solutions in the neighborhood of a movable singularity.

It is a well-known conjecture that if a field equation has the PP then it is completely integrable.[6] The limitations of this conjecture, known as the Ablowitz–Ramani–Segur (ARS) conjecture, have been pointed out by many authors.[2,3] The complete integrability is also defined in terms of the existence of the inverse scattering transform (IST) or the auto-Bäcklund transformation (ABT).[7] The existence of an IST solution is assured by that of Lax pairs.

A well-known[7] model for an IST solvable and completely integrable dynamical system is the celebrated constant coefficient Korteweg–de Vries (KdV) equation:

$$u_{,t} + \alpha u u_{,x} + \beta u_{,xxx} = 0, \tag{1.1}$$

the coefficients α and β being constants and the suffix indicating a partial derivative with the respective variables. This equation yields a highly collisionally stable particlelike solution, called a soliton.

Here, we report the results of the PP analysis of a KdV equation with variable coefficients. The PP is used to identify the values of the different parameters for which the system loses its integrability. We have found these parameter values using a property of Lax pairs obtained from the PP. The possible ABT is also developed, when the system is integrable.

Such an equation is particularly significant in the study of the development of a steady solitary wave as it enters a region where the bottom is no longer level.[8–15] It has been found both theoretically and experimentally that when the depth decreases to form a shelf, the solitary wave breaks into a number of "solitons" while if the depth is increasing the solitary wave degenerates into a cnoidal wave.

II. PAINLEVÉ PROPERTY OF VARIABLE COEFFICIENT KdV EQUATION

We introduce a variable coefficient KdV equation:

$$u_{,t} + \alpha t^n u u_{,x} + \beta t^m u_{,xxx} = 0, \tag{2.1}$$

where m and n are real numbers and α and β are constant parameters. The well-known KdV equation (1.1) is obtained when $m = n = 0$. For $\alpha = \frac{3}{2}, \beta = \frac{1}{6}$ and $m = 0$, $n = -\frac{1}{2}$, we can transform (2.1) to the well-known purely concentric KdV equation

$$2v_{,t} + v/t + 3vv_{,x} + \tfrac{1}{3}v_{,xxx} = 0, \tag{2.2}$$

through a nonlinear transformation

$$u = v\sqrt{t}\,. \tag{2.3}$$

Equation (2.2) is studied by several authors,[16–19] and ABT and IST are well known for this system. Some soliton-like solutions of (2.2) in terms of Airy functions are also developed.[18]

Equation (2.1) has the PP when its solutions $u(x,t)$ are "single valued" about the movable, singularity manifolds, determined from the singularity analysis of the Lorentz series expansion

$$u(x,t) = \varphi^n(x,t) \sum_{j=0}^{\infty} u_j(x.t)\varphi^j(x,t), \tag{2.4}$$

where $u_j(x,t)$ and $\varphi(x,t)$ are analytic functions in a neighborhood of the manifold

[a)] Present address: Mathematics Section, International Centre for Theoretical Physics, 34100, Trieste, Italy.

 0022-2488/86/112640-04$02.50

$$\varphi(x,t)=0 \tag{2.5}$$

and η is an integer to be determined. Substituting (2.4) into Eq. (2.1), a leading-order terms analysis uniquely determines the possible values of η. The PP requires that η be a negative integer. The resultant series expansion of (2.1) gives the required ABT and Lax pair for the IST.

The leading-order terms analysis gives the value $\eta=-2$. The recursion relations for $u_j(x,t)$ are found to be

$$\begin{aligned}u_{j-3,t}+(j-4)u_{j-2}\varphi_{,t}+\alpha t^n\sum_{k=0}^{j}u_{j-k}(u_{k-1,x}+(k-2)u_k\varphi_{,x})+\beta t^m\{u_{j-3,xxx}+3(j-4)u_{j-2,xx}\varphi_{,x}\\ +3(j-3)(j-4)u_{j-1,x}\varphi_{,x}{}^2+3(j-4)u_{j-2,x}\varphi_{,xx}+(j-2)(j-3)(j-4)u_j\varphi_{,x}{}^3\\ +3(j-3)(j-4)u_{j-1}\varphi_{,x}\varphi_{,xx}+(j-4)u_{j-2}\varphi_{,xxx}\}=0,\end{aligned} \tag{2.6}$$

where

$$\varphi_{,x}=\frac{\partial\varphi}{\partial x},\quad u_{j,x}=\frac{\partial u_j(x,t)}{\partial x},\quad \text{etc.} \tag{2.7}$$

Collecting terms involving u_j, it is readily found that

$$\beta t^m\varphi_{,x}{}^3(j-6)(j-4)(j+1)u_j=F(u_{j-1},\ldots,u_0,\varphi_{,t},\varphi_{,x},\ldots), \tag{2.8}$$

for $j=0,1,2,\ldots$.

We note that the recursion relations (2.8) are not defined when $j=-1,4$, and 6. These values of j are called the "resonances" of the recursion relation and, corresponding to these values of j, we can insert arbitrary functions of (x,t) instead of $u_j(x,t)$ into the series expansion (2.4). But for $j=-1$, the series expansion (2.4) is not defined and so the admissible values of resonances are $j=4$ and 6 only.

Putting $j=0,1,2,\ldots$ in (2.7), we get

$$j=0,\quad u_0=-(12\beta/\alpha)t^{m-n}\varphi_{,x}{}^2, \tag{2.9}$$

$$j=1,\quad u_1=(12\beta/\alpha)t^{m-n}\varphi_{,xx}, \tag{2.10}$$

$$j=2,\quad (t^{-n}/\alpha)\varphi_{,x}\varphi_{,t}+u_2\varphi_{,x}{}^2-(3\beta/\alpha)t^{m-n}\varphi_{,xx}{}^2+(4\beta/\alpha)t^{m-n}\varphi_{,x}\varphi_{,xxx}=0, \tag{2.11}$$

$$j=3,\quad (t^{-n}/\alpha)\varphi_{,xt}+(m-n)(t^{-n-1}/\alpha)\varphi_{,x}+u_2\varphi_{,xx}-u_3\varphi_{,x}{}^2+(\beta t^{m-n}/\alpha)\varphi_{,xxxx}=0, \tag{2.12}$$

and

$$j=4,\quad \frac{\partial}{\partial x}\left\{\frac{t^{-n}}{\alpha}\varphi_{,xt}+(m-n)\frac{t^{-n-1}}{\alpha}\varphi_{,x}+u_2\varphi_{,xx}-u_3\varphi_{,x}{}^2+\frac{\beta t^{m-n}}{\alpha}\varphi_{,xxxx}\right\}=0, \tag{2.13}$$

which is a compatibility condition. The compatibility condition at $j=6$ involves extensive calculations.

When we assign $u_4=u_6=0$ and for $u_3=0$, we can find

$$u_j=0,\quad \text{for all } j\geqslant 3, \tag{2.14}$$

provided u_2 is a solution of (2.1), which implies that

$$u_{2,t}+\alpha t^n u_2u_{2,x}+\beta t^m u_{2,xxx}=0. \tag{2.15}$$

From Eq. (2.4) and Eqs. (2.9)–(2.15), we get

$$u_0=-(12\beta/\alpha)t^{m-n}\varphi_{,x}{}^2, \tag{2.16}$$

$$u_1=(12\beta/\alpha)t^{m-n}\varphi_{,xx}, \tag{2.17}$$

$$\begin{aligned}&(t^{-n}/\alpha)\varphi_{,x}\varphi_{,t}+u_2\varphi_{,x}{}^2-(3\beta/\alpha)t^{m-n}\varphi_{,xx}{}^2\\ &\quad+(4\beta/\alpha)t^{m-n}\varphi_{,x}\varphi_{,xxx}=0,\end{aligned} \tag{2.18}$$

$$\begin{aligned}&(t^{-n}/\alpha)\varphi_{,xt}+[(m-n)/\alpha]t^{-n-1}\varphi_{,x}+u_2\varphi_{,xx}\\ &\quad+(\beta/\alpha)t^{m-n}\varphi_{,xxxx}=0,\end{aligned} \tag{2.19}$$

$$u_{2,t}+\alpha t^n u_2u_{2,x}+\beta t^m u_{2,xxx}=0, \tag{2.20}$$

and

$$u_j=0,\quad \text{for } j\geqslant 3. \tag{2.21}$$

Substituting (2.16)–(2.21) in Eq. (2.4), we get

$$u(x,t)=-\frac{12\beta}{\alpha}t^{m-n}\frac{\varphi_{,x}{}^2}{\varphi^2}+\frac{12\beta}{\alpha}t^{m-n}\frac{\varphi_{,xx}}{\varphi}+u_2, \tag{2.22}$$

or

$$u(x,t)=\frac{12\beta}{\alpha}t^{m-n}\frac{\partial^2}{\partial x^2}(\log\varphi)+u_2, \tag{2.23}$$

where $u(x,t)$ and u_2 are exact solutions of (2.1) and (2.15), respectively.

Equations (2.16)–(2.23) define the ABT for the variable coefficient KdV equation (2.1) provided (2.18) and (2.19) are consistent. If any one of the solutions $u_2(x,t)$ is known then another solution $u(x,t)$ of Eq. (2.1) can be determined using the ABT. The consistency of Eqs. (2.18) and (2.19) can be verified by using a property of the Lax pairs.

The Lax pairs are obtained from the equations (2.18) and (2.19) by using a transformation

$$\varphi_{,x}=V^2. \tag{2.24}$$

Substituting (2.24) in (2.19) yields

$$\frac{t^{-n}}{\alpha}V_{,t}+\frac{(m-n)}{2\alpha}t^{-n-1}V+u_2V_{,x}+\frac{\beta}{\alpha}t^{m-n}V_{,xxx}$$
$$+\frac{3\beta}{\alpha}t^{m-n}V_{,x}\frac{V_{,xx}}{V}=0. \tag{2.25}$$

Equation (2.19) also transforms to

$$\frac{t^{-n}}{\alpha}V_{,t}+u_2V_{,x}+\frac{1}{2}u_{2,x}V+\frac{4\beta}{\alpha}t^{m-n}V_{,xxx}=0. \tag{2.26}$$

Eliminating $V_{,t}$ from Eqs. (2.25) and (2.26) we get

$$\frac{(m-n)}{2\alpha}t^{-n-1}-\frac{1}{2}u_{2,x}-\frac{3\beta}{\alpha}t^{m-n}\left(\frac{V_{,xx}}{V}\right)_{,x}=0. \tag{2.27}$$

Integrating Eq. (2.27) with respect to x gives

$$\frac{\beta}{\alpha}t^{m-n}\frac{V_{,xx}}{V}+\frac{1}{6}u_2+\frac{(m-n)}{6\alpha}xt^{-n-1}=\lambda(t), \tag{2.28}$$

or

$$f(t)\left\{\frac{\beta}{\alpha}t^{m-n}D^2+\frac{1}{6}u_2-\frac{(m-n)}{6\alpha}xt^{-n-1}\right\}V$$
$$=f(t)\lambda(t)V. \tag{2.29}$$

Thus we get the linear eigenvalue problem

$$LV=\mu V, \tag{2.30}$$

where $\mu=\lambda(t)f(t)$ and L is a linear operator defined by

$$L=f(t)\left\{\frac{\beta}{\alpha}t^{m-n}D^2+\frac{1}{6}u_2-\frac{(m-n)}{6\alpha}xt^{-n-1}\right\}. \tag{2.31}$$

From Eq. (2.26) we get

$$V_t=-\alpha t^n\{(4\beta/\alpha)t^{m-n}D^3+u_2D+\tfrac{1}{2}u_{2,x}\}V, \tag{2.32}$$

or

$$V_t=-BV, \tag{2.33}$$

where the operator B is defined by

$$B=\alpha t^n\{(4\beta/\alpha)t^{m-n}D^3+u_2D+\tfrac{1}{2}u_{2,x}\}. \tag{2.34}$$

Equations (2.30) and (2.34) define the Lax pairs L and B. However, Eq. (2.33) implies that the eigenfunction V is in time evolution so that

$$L_t=LB-BL. \tag{2.35}$$

The L_t in (2.35) denotes the derivative with respect to both the explicit time dependence of L and the implicit dependence through $u_2(x,t)$.

From (2.30) and (2.33) we get the following results for which (2.35) holds:

$$\text{(i) } m=n,\quad f(t)=C, \tag{2.36}$$

$$\text{(ii) } m=2n+1,\quad f(t)=Ct^{n+1}, \tag{2.37}$$

where C is an arbitrary constant. For all other values of m and n the Lax pairs are not consistent and hence the ABT exists only for the values of m and n defined in Eqs. (2.36) and (2.37). Equation (2.36) implies that m and n can be both zero together and then the respective L and B are the well-known Lax pairs of the constant coefficients KdV equation (1.1).

The above study shows that the variable coefficient KdV equation (2.1) is IST solvable and has PP whenever $m=n$ or $m=2n+1$ and these properties are independent of the constant parameters α and β. For all other values of m and n, the system is nonintegrable.

III. DISCUSSION

The variable coefficient KdV equation (2.1) that we have introduced is a new member in the families of integrable as well as nonintegrable PDE's depending on the coefficients. The PP analysis leads to the ABT and Lax pairs when it is integrable. The operator identity (2.35) of the Lax pairs reveals that the system (2.1) can be integrable when $m=n$ and $m=2n+1$ only, whereas for all other values of m and n, the system (2.1) is nonintegrable. The soliton solutions are the products of IST solvable class of nonlinear PDE's.[1,7] The above study shows that the variable coefficient KdV equation (2.1) does not always have a soliton, but only in two special cases. Hence in general a solitary wave solution of (2.1) need not be a soliton and so it need not be collisionally stable always.

The variable coefficient KdV equation (2.1) that we introduced is a model for explaining the observations of soliton-type solution's instability reported earlier in different contexts.[8–15]

The existence of an infinite number of conservation laws is considered as a necessary condition for the existence of soliton solutions of IST solvable equations.[7] Here we are able to give two of these members for general m and n,

$$u_{,t}+((\alpha/2)t^nu^2+\beta t^mu_{,xx})_{,x}=0 \tag{3.1}$$

and

$$\left(\frac{1}{2}u^2\right)_{,t}+\left(\frac{\alpha}{3}t^nu^3-\frac{\beta}{2}t^mu_{,x}{}^2+\beta t^muu_{,xx}\right)_{,xx}=0. \tag{3.2}$$

The higher-order conserved quantities are not so direct and they are now under investigation.

It is interesting to find the soliton solutions of the variable coefficient KdV equation when it is IST solvable and study their time evolutions, etc. Another interesting problem is that of finding the solution of (2.1) for general m and n and then studying its time evolution for various values of m and n. Such a study may shed some light on the possible connection between movable singularity, the Painlevé property, and the soliton stability of particular solutions of a nonlinear PDE.

ACKNOWLEDGMENTS

BVB thanks Professor K. P. Sinha for his warm hospitality during his visit to the Centre for Theoretical Studies, Indian Institute of Science, Bangalore, India.

MJV thanks the Department of Science and Technology Government of India, for financial support to the project [12(10)/82-STP II] under which the work is carried out.

[1] H. G. Helleman, *Fundamental Problems in Statistical Mechanics* (North-Holland, Amsterdam, 1980), Vol. 5.

[2]T. C. Bountis, *Singularities and Dynamical Systems, Mathematics Studies No.* 103, edited by S. N. Penevmatikos (North-Holland, New York, 1985).
[3]W. H. Steeb, M. Kloke, B. M. Spieker, and A. Kunick, Found. Phys. **15**, 677 (1985).
[4]R. S. Ward, Phys. Lett. A **102**, 279 (1984).
[5]J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[6]M. J. Ablowitz, A. Ramani, and H. Segur, J. Math. Phys. **21**, 715 (1980).
[7]R. K. Bullough, P. T. Caudrey, and H. M. Gibbs, *Topics in Current Physics Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer, New York, 1980).
[8]O. S. Madsen and C. C. Mei, J. Fluid Mech. **39**, 781 (1969).
[9]S. Maxon and J. Viecelli, Phys. Fluids **17**, 1614 (1974).
[10]H. Shen and W. F. Ames, Phys. Lett. A **49**, 313 (1974).
[11]R. S. Johnson, Proc. Cambridge Philos. Soc. **73**, 183 (1973).
[12]T. Kakutani and H. Ono, J. Phys. Soc. Jpn. **45**, 333 (1978).
[13]J. W. Miles, J. Fluid. Mech. **84**, 181 (1978).
[14]R. S. Johnson and S. Thompson, Phys. Lett. A **66**, 279 (1978).
[15]R. S. Johnson, J. Fluid. Mech. **97**, 701 (1980).
[16]C. J. Knickerbocker and A. C. Newell, J. Fluid Mech. **153**, 1 (1985).
[17]F. Calogero and A. Degasperis, Lett. Nuovo Cimento **23**, 150, 155 (1978).
[18]A. Nakumara, J. Phys. Soc. Jpn. **49**, 2380 (1980).
[19]W. H. Steeb, M. Kloke, B. M. Spieker, and W. Oevel, J. Phys. A **16**, L447 (1983).

A variable coefficient Korteweg–de Vries equation: Similarity analysis and exact solution. II

N. Nirmala and M. J. Vedan
Department of Mathematics and Statistics, University of Cochin, Cochin 682 022, India
B. V. Baby[a)]
Department of Mathematics, Bharata Mata College, Cochin 682 021, India

(Received 6 January 1986; accepted for publication 11 July 1986)

A Korteweg–de Vries (KdV) equation with time-dependent coefficients is studied in this paper. The similarity transformation for this system is investigated and an exact solution in a particular case is obtained. The Ablowitz–Ramani–Segur (ARS) conjecture is used to identify the integrability of the system. It is found that in some special cases the system may be integrable.

I. INTRODUCTION

One of the most important methods for developing exact solutions of partial differential equations (PDE's) is that of reducing the number of variables, exploiting continuous symmetries of the system. The solutions obtained by this procedure are generally called similarity solutions.[1] This method has been widely used in the past for developing solutions as well as for the test of the Painlevé property (PP) of various systems.[2,3]

In an earlier paper[4] we have analyzed the existence of auto-Bäcklund (ABT), Lax pairs (LP), and the PP of a Korteweg–de Vries (KdV) equation with variable coefficients. In this paper we are reporting some similarity solutions and an exact solution of the equation in a particular case using the standard similarity method.

II. SIMILARITY TRANSFORMATIONS OF A PARTIAL DIFFERENTIAL EQUATION

We shall give the essential details[1] of the Lie continuous point group similarity transformation method to reduce the number of independent variables of a PDE,

$$F(x,t,u,u_t,u_x,u_{xx},\ldots)=0\,, \tag{2.1}$$

under a family of one-parameter infinitesimal continuous point group transformations

$$x=x+\epsilon X(x,t,u)+O(\epsilon^2)\,, \tag{2.2}$$

$$t=t+\epsilon T(x,t,u)+O(\epsilon^2)\,, \tag{2.3}$$

$$u=u+\epsilon U(x,t,u)+O(\epsilon^2)\,. \tag{2.4}$$

Here X, T, and U are the infinitesimals of the variables x, t, and u, respectively, and ϵ is an infinitesimal parameter. The derivatives of u are also transformed according to

$$u_x=u_x+\epsilon[U_x]+O(\epsilon^2)\,, \tag{2.5}$$

$$u_t=u_t+\epsilon[U_t]+O(\epsilon^2)\,, \tag{2.6}$$

$$u_{xxx}=u_{xxx}+\epsilon[U_{xxx}]+O(\epsilon^2)\,, \tag{2.7}$$

where $[U_x]$, $[U_t]$, and $[U_{xxx}]$ are the infinitesimals of the transformations of derivatives u_x, u_t, and u_{xxx}. These are called the first and third extensions depending on the order of the derivative term. These "extensions"[1] are given by

$$[U_x]=U_x+(U_u-X_x)u_x-T_xu_t-X_uu_x{}^2-T_uu_xu_t\,, \tag{2.8}$$

$$[U_t]=U_t+(U_u-T_t)u_t-X_tu_x-T_uu_t{}^2-X_uu_xu_t\,, \tag{2.9}$$

and

$$\begin{aligned}[U_{xxx}]=&U_{xxx}+(3U_{xxu}-X_{xxx})u_x-T_{xxx}u_t+3(U_{xuu}-X_{xxu})u_x{}^2-3T_{xxu}u_xu_t+(U_{uuu}-3X_{xuu})u_x{}^3\\&+3(U_{xu}-X_{xx})u_{xx}-3T_{xx}u_{xt}-3T_{xuu}u_x{}^2u_t+3(U_{uu}-3X_{xu})u_xu_{xx}-3T_{xu}u_tu_{xx}-6T_{xu}u_{xt}u_x\\&-3T_xu_{xxt}+(U_u-3X_x)u_{xxx}-X_{uuu}u_x{}^4-6X_{uu}u_x{}^2u_{xx}-3T_{uu}u_x{}^2u_{xt}\\&-T_{uuu}u_x{}^3u_t-3X_uu_{xx}{}^2-3T_uu_xu_{xxt}-3T_uu_{xx}u_{xt}-3T_{uu}u_xu_tu_{xx}-4X_uu_xu_{xxx}-T_uu_tu_{xxx}\,.\end{aligned} \tag{2.10}$$

The invariance requirement of (2.1) under the set of transformations (2.2)–(2.10) leads to the invariant surface condition

$$T\frac{\partial F}{\partial t}+X\frac{\partial F}{\partial x}+U\frac{\partial F}{\partial u}+[U_x]\frac{\partial F}{\partial u_x}+[U_t]\frac{\partial F}{\partial u_t}+[U_{xxx}]\frac{\partial F}{\partial u_{xxx}}=0\,. \tag{2.11}$$

On solving (2.11), the infinitesimals X, T, and U can be uniquely determined, which give the similarity group under which the system (2.1) is invariant.

By the infinitesimal transformations (2.2)–(2.4) we have

$$u(x+\epsilon X+O(\epsilon^2),\ t+\epsilon T+O(\epsilon^2))=u+\epsilon U+O(\epsilon^2)\,. \tag{2.12}$$

On expanding and equating the $O(\epsilon)$ terms on either side of (2.12) we get

[a)] Present address: Mathematics Section, International Centre for Theoretical Physics, 34100, Trieste, Italy.

 0022-2488/86/112644-03$02.50

$$T\frac{du}{dt}+X\frac{du}{dx}-U=0\,. \tag{2.13}$$

The solutions of (2.13) are obtained by Lagrange's condition

$$dt/T=dx/X=du/U\,. \tag{2.14}$$

Equation (2.14) gives the solution

$$x=x(t,c_1,c_2)\,, \tag{2.15}$$

$$u=u(t,c_1,c_2)\,, \tag{2.16}$$

where c_1, c_2 are arbitrary integration constants. The constant c_1 plays the role of an independent variable called the similarity variable σ and c_2 that of a dependent variable called the similarity solution $f(\sigma)$ such that

$$u(x,t)=f(\sigma)\,. \tag{2.17}$$

Substituting (2.17) in the original equation (2.1) the resultant equation is an ordinary differential equation involving only the derivatives with respect to the similarity variable σ.

III. SIMILARITY TRANSFORMATION AND LIE ALGEBRA OF VARIABLE COEFFICIENT KdV EQUATION

Here we consider a variable coefficient KdV equation

$$u_t+\alpha t^n uu_x+\beta t^m u_{xxx}=0\,, \tag{3.1}$$

where α and β are arbitrary constant parameters and n and m are real numbers. In a special case this equation can be reduced to the well-known cylindrical KdV equation.[4]

Under the family of infinitesimal transformations (2.2)–(2.4) the variable coefficient KdV equation (3.1) yields

$$[U_t]+\alpha t^n(u_xU+u[U_x])+\alpha nt^{n-1}uu_xT +\beta t^m[U_{xxx}]+\beta mt^{m-1}u_{xxx}T=0\,. \tag{3.2}$$

On substituting the expressions for the extensions from (2.8)–(2.10) and solving for the infinitesimals X, T, and U we get the constraint equations

$$-X_t+\alpha t^n(U+u(U_u-X_x))+n\alpha t^{n-1}uT=0\,, \tag{3.3}$$

$$U_t+\alpha t^n uU_x+\beta t^m U_{xxx}=0\,, \tag{3.4}$$

$$tU_u-3tX_x+mT=0\,, \tag{3.5}$$

$$U_u-T_t=0\,,\ U_{xu}-X_{xx}=0\,,\ U_{uu}-3X_{ux}=0\,, \tag{3.6}$$

$$T_x=T_u=X_u=0\,. \tag{3.7}$$

The constraints (3.3)–(3.7) can be uniquely solved. Then we get the following solutions for X, T, and U.

(i) When m and n are arbitrary,

$$T=0\,, \tag{3.8}$$

$$X=a[\alpha t^{n+1}/(n+1)]+b\,, \tag{3.9}$$

$$U=a\,. \tag{3.10}$$

For the Lie algebra,

$$G_1=\frac{\alpha t^{n+1}}{n+1}\frac{\partial}{\partial x}+\frac{\partial}{\partial u}\,, \tag{3.11}$$

$$G_2=\frac{\partial}{\partial x}\,, \tag{3.12}$$

$$[G_1,G_2]=0\,. \tag{3.13}$$

(ii) When $m=3n+5$,

$$T=t\,, \tag{3.14}$$

$$X=(2+n)x+a[\alpha t^{n+1}/(n+1)]+b\,, \tag{3.15}$$

$$U=u+a\,. \tag{3.16}$$

The Lie algebra is the same as in the last case [(3.11)–(3.13)].

(iii) When $m=-2$ and $n=-\frac{3}{2}$,

$$T=t^{1/2}\,, \tag{3.17}$$

$$X=-(xt^{-1/2}/2)-2a\alpha t^{-1/2}+b\,, \tag{3.18}$$

$$U=(ut^{-1/2}/2)+(x/4\alpha)+a\,. \tag{3.19}$$

The Lie algebra is same as in (3.11)–(3.13), with $n=-\frac{3}{2}$.

In all the above cases [(3.8)–(3.19)], a and b are arbitrary integration constants.

IV. SIMILARITY, SELF-SIMILAR AND EXACT SOLUTIONS

Using (2.14) and (2.17) we can find the similarity variables, similarity reduced equations, and similarity solutions for the above three cases [(3.8)–(3.19)].

The set of infinitesimals (3.8)–(3.10) gives the similarity variable

$$\sigma_1=t \tag{4.1}$$

and the similarity reduced equation

$$\frac{df_1}{d\sigma_1}+\frac{(n+1)a\alpha\sigma_1^n}{a\alpha t^{n+1}+(n+1)b}f_1=0\,. \tag{4.2}$$

The corresponding similarity solution is

$$u(x,t)=[(n+1)ax/a\alpha t^{n+1}+(n+1)b)]+f_1\,. \tag{4.3}$$

Equations (4.2) and (4.3) give an exact solution of the variable coefficient KdV equation (2.1).

$$u(x,t)=[a(n+1)x+c]/[a\alpha t^{n+1}+b(n+1)]\,. \tag{4.4}$$

The solution (4.4) is not so useful as the third derivative with respect to the variable x vanishes.

The set of infinitesimals (3.14)–(3.16) yields the similarity variable

$$\sigma_2=\frac{x}{t^{n+2}}+\frac{a\alpha}{(n+1)t}+\frac{b}{(n+2)t^{n+2}}\,. \tag{4.5}$$

The corresponding similarity reduced equation is

$$\beta\frac{d^3f_2}{d\sigma_2^3}+\alpha f_2\frac{df_2}{d\sigma_2}+f_2-(n+2)\sigma_2\frac{df_2}{d\sigma_2}=0 \tag{4.6}$$

and the similarity solution is

$$u(x,t)=tf_2(\sigma_2)-a\,. \tag{4.7}$$

When $n=-3$, Eq. (4.6) can be reduced to a second-order equation by integration with respect to σ_2. This yields

$$\beta\frac{d^2f_2}{d\sigma_2^2}+\frac{\alpha}{2}f_2^2+\sigma_2f_2=\text{const}\,. \tag{4.8}$$

Equation (4.8) is not easily solvable.

From Eq. (3.17)–(3.19) we get the similarity variable

$$\sigma_3 = xt^{1/2} + 4a\alpha t^{1/2} - bt\,. \tag{4.9}$$

The corresponding similarity reduced equation is

$$\beta \frac{d^3 f_3}{d\sigma_3^3} + \alpha f_3 \frac{df_3}{d\sigma_3} + \frac{b}{2\alpha} = 0 \tag{4.10}$$

and the similarity solution is

$$u(x,t) = -\sigma_3/2\alpha + bt/2\alpha + t^{1/2} f_3(\sigma_3)\,. \tag{4.11}$$

Equation (4.10) can be exactly solved for the case $b = 0$. This gives the following solution of the variable coefficient KdV equation (3.1) for $m = -2$, $n = -\frac{3}{2}$:

$$u(x,t) = \frac{-(4\alpha + x)t^{1/2}}{2\alpha} + \frac{4t^{1/2}}{[(\sqrt{-\alpha/3\beta})(x + 4a\alpha)t^{1/2} + c]^2}\,. \tag{4.12}$$

The exact solution (4.12) is real valued only when $\alpha < 0$ or $\beta < 0$ and not both simultaneously negative. The solution (4.12) has no characteristics of a stable configurationlike "soliton."[5]

The self-similar[1] solution can be developed for the variable coefficient KdV equation (3.1) using the dimensional analysis. The self-similar transformation is very much identical to the similarity transformations; nevertheless self-similar solutions are not always obtainable by similarity procedure.

For the variable coefficient KdV equation (3.1) we got the self-similar transformation

$$u(x,t) = t^{(m-3n-2)/3} F(\eta)\,, \tag{4.13}$$

where $\eta(x,t)$ is the self-similar variable

$$\eta(x,t) = xt^{-(m+1)/3}\,. \tag{4.14}$$

Equation (4.13) yields the following self-similarity reduced ordinary differential equation, on substituting in (3.1):

$$\beta \frac{d^3 F}{d\eta^3} + \alpha F \frac{dF}{d\eta} - \left(\frac{m+1}{3}\right)\eta \frac{dF}{d\eta} + \frac{m-3n-2}{3} F = 0\,. \tag{4.15}$$

Unfortunately Eq. (4.15) cannot be solved for any values of m and n.

V. DISCUSSION

The variable coefficient KdV equation has attracted the attention of several authors since 1969.[6–13] The equation is a standard mathematical model[6] for explaining the soliton breaking phenomena observed in variable depth shallow water. So far no exact solution for this model exists in the literature. Our work is an attempt in this direction.

Using the well-known Ablowitz–Ramani–Segur (ARS) conjecture[3] one can study the PP of a PDE by reducing it to an ordinary differential equation (ODE), using similarity or self-similar transformations. Equation (4.2) is linear and so it is clearly a Painlevé-type. For $n = -3$, Eq. (4.8) is not a Painlevé-type equation whereas (4.11) can be integrated once and it will reduce to Painlevé-type. This equation (4.15) can be reduced to a second-order equation for $n = -1$, but not a Painlevé-type.

The exact solution (4.12) that we developed has no smooth property of a soliton solution, which indicates that the system has decaying solutions other than soliton solutions when coefficients of KdV equation are variables.

ACKNOWLEDGMENTS

BVB thanks Professor K. P. Sinha for his warm hospitality during his visit to the Centre for Theoretical Studies, Indian Institute of Science, Bangalore, India.

MJV thanks the Department of Science and Technology Government of India, for the financial support to the Project [12(10)/82-STP II] under which the work is carried out.

[1] G. M. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer, Berlin, 1974).
[2] M. Lakshmanan and P. Kaliappan, J. Math. Phys. **24**, 795 (1983).
[3] M. J. Ablowitz, A. Ramani, and H. Segur, Lett. Nuovo Cimento **23**, 333 (1978); J. Math. Phys. **21**, 715, 1006 (1980).
[4] N. Nirmala, M. J. Vedan, and B. V. Baby, J. Math. Phys. **27**, 2640 (1986).
[5] A. C. Scott, F. Y. F. Chu, and D. W. McLaughlin, Proc. IEEE **61**, 1443 (1973).
[6] O. S. Madsen and C. C. Mei, J. Fluid Mech. **39**, 781 (1969).
[7] R. S. Johnson, Proc. Cambridge Philos. Soc. **73**, 183 (1973).
[8] S. Maxon and J. Viecelli, Phys. Fluids **17**, 1614 (1974).
[9] J. W. Miles, J. Fluid Mech. **84**, 181 (1978).
[10] R. S. Johnson and S. Thompson, Phys. Lett. A **66**, 279 (1978).
[11] T. Kakutani and H. Ono, J. Fluid Mech. **97**, 701 (1980).
[12] R. S. Johnson, J. Fluid Mech. **97**, 701 (1980).
[13] W. H. Steeb, M. Kloke, B. M. Spieker, and W. Oevel, J. Phys. A **16**, L447 (1983).

Periodic fixed points of Bäcklund transformations and the Korteweg–de Vries equation

John Weiss[a)]
Institute for Pure and Applied Physical Sciences, University of California, San Diego, La Jolla, California 92093

(Received 28 October 1985; accepted for publication 11 July 1986)

A new method for studying integrable systems based on the "periodic fixed points" of Bäcklund transformations (BT's) is presented. Normally the BT maps an "old" solution into a "new" solution and requires a known "seed" solution to get started. Besides this limitation, it can also be difficult to qualitatively classify the result of applying the BT several times to a known solution. By studying the periodic fixed points of the BT (regarded as a nonlinear map in a function space), integrable systems of equations of finite degree (equal to the order of the fixed point) and a method for the systematic classification of the solutions of the original system are obtained.

I. INTRODUCTION

A Bäcklund transformation (BT) maps solutions of a nonlinear system into solutions of (we assume the same) nonlinear system. Customarily, the BT is applied iteratively to a known (trivial) solution to generate a sequence of solutions that may be of interest.

For instance, the (Schwarzian KdV) equation[1]

$$\varphi_t/\varphi_x + \{\varphi;x\} = \lambda \tag{1.1}$$

has the Bäcklund transformations[1]

$$\text{(i)}\quad \varphi = (a\psi + b)/(c\psi + d), \quad ad - bc = 1, \tag{1.2}$$

$$\text{(ii)}\quad \varphi_x = \psi_x^{-1}, \tag{1.3}$$

where

$$\psi_t/\psi_x + \{\psi;x\} = \lambda. \tag{1.4}$$

The expression

$$\{\varphi;x\} = \frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) - \frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2 \tag{1.5}$$

is the Schwarzian derivative, which is invariant under the Moebius group (1.2).[2,3]

From Eq. (1.1) and identifying the variables

$$V = \varphi_{xx}/\varphi_x, \tag{1.6}$$

$$U = \{\varphi;x\}, \tag{1.7}$$

it can be shown that[1]

$$V_t + \frac{\partial}{\partial x}\left(V_{xx} - \frac{1}{2}V^3\right) = \lambda V_x, \tag{1.8}$$

$$U_t + \frac{\partial}{\partial x}\left(U_{xx} + \frac{3}{2}U^2\right) = \lambda U_x, \tag{1.9}$$

where Eq. (1.9) is the KdV and (1.8) the modified KdV equation.[1]

Now, from Eqs. (1.1) and (1.4), the BT (1.3) is completed by (either of) the equations

$$\frac{\varphi_t}{\varphi_x} + \frac{\psi_t}{\psi_x} = \left(\frac{\varphi_{xx}}{\varphi_x}\right)^2 + 2\lambda = \left(\frac{\psi_{xx}}{\psi_x}\right)^2 + 2\lambda, \tag{1.10}$$

$$\frac{\varphi_t}{\varphi_x} - \frac{\psi_t}{\psi_x} = 2\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right) = -2\frac{\partial}{\partial x}\left(\frac{\psi_{xx}}{\psi_x}\right). \tag{1.11}$$

That is (1.3) and (say) (1.10) imply, by the integrability conditions,

$$\varphi_{tx} = \varphi_{xt}, \quad \psi_{tx} = \psi_{xt}, \tag{1.12}$$

that (φ,ψ) satisfy Eq. (1.1). Since (1.3) is in involution, the effective BT is the composition of (1.2) and (1.3):

$$\varphi_x = (c\psi + d)^2/\psi_x, \tag{1.13}$$

$$\frac{\varphi_t}{\varphi_x} = -\frac{\psi_t}{\psi_x} + \left(\frac{\psi_{xx}}{\psi_x}\right)^2 - 4\frac{\partial^2}{\partial x^2}\ln\left(\psi + \frac{d}{c}\right) + 2\lambda. \tag{1.14}$$

Through the iterative application of (1.13) and (1.14),

$$\varphi = \varphi_{n+1}, \quad \psi = \varphi_n, \tag{1.15}$$

with $\lambda = 0$,

$$\varphi_0 = x, \tag{1.16}$$

the sequence of rational solutions of the KdV equation can be found.[1] These are

$$\varphi_1 = x^3 + 12t,$$
$$\varphi_2 = (x^6 + 60tx^3 - 720t^2)/x, \tag{1.17}$$

etc.

On the other hand, with the solution

$$\varphi_0 = e^{x+bt}, \quad b = \lambda + \tfrac{1}{2}, \tag{1.18}$$

an application of (1.13) and (1.14) obtains

$$\varphi_1 = c^2e^{x+bt} + 2cd(x - 2t) - d^2e^{-x-bt}. \tag{1.19}$$

The continued iterative application of (1.13) and (1.14) produces solutions that are rational functions of e^{x+bt}, x, and t. The "secular" terms in (1.19) will vanish only if d or c vanishes. Say, $d = 0$. But in this case $\varphi_1 = c^2\varphi_0$ or φ_1 is a fixed point of the BT

$$\varphi_{n+1,x} = \varphi_n^2/\varphi_{n,x}, \tag{1.20}$$

$$\frac{\varphi_{n+1,t}}{\varphi_{n+1,x}} + \frac{\varphi_{n,t}}{\varphi_{n,x}} = \left(\frac{\varphi_{n,xx}}{\varphi_{n,x}}\right)^2 - 4\frac{\partial^2}{\partial x^2}\ln\varphi_n + 2\lambda. \tag{1.21}$$

Since we do not find the usual N-soliton solutions by a straightforward, iterative application of the BT, we propose, instead, to study the periodic fixed points of the BT (1.20),

[a)] Mailing address: Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138.

 0022-2488/86/112647-10$02.50

(1.21). That is, we define Eqs. (1.20) and (1.21) with

$$n = 1,2,3,4,\ldots,\mathrm{mod}(N). \tag{1.22}$$

The periodic fixed points continue to define a strong Bäcklund transformation of Eq. (1.1). That is, the integrability conditions

$$\varphi_{n+1,xt} = \varphi_{n+1,tx} \tag{1.23}$$

continue to imply that φ_n satisfy Eq. (1.1) and, by the periodicity mod(N), the set $\{\varphi_n, n = 1,2,\ldots,\mathrm{mod}(N)\}$ are solutions of (1.1). Therefore, it is enough to solve (1.20) and then fix the time-dependent constants of integration by (1.21).

The period three fixed points satisfy the system of equations:

$$\varphi_{1x} = \varphi_3^2/\varphi_{3x}, \quad \varphi_{2x} = \varphi_1^2/\varphi_{1x}, \quad \varphi_{3x} = \varphi_2^2/\varphi_{2x}. \tag{1.24}$$

It is not difficult to show that Eqs. (1.24) with

$$\epsilon_i = \varphi_{ix}/\varphi_i \tag{1.25}$$

are equivalent to a Hamiltonian system and each ϵ_i will satisfy the equation

$$\epsilon_x^2 = \epsilon^2(\epsilon - b)^2 - 4a\epsilon, \tag{1.26}$$

which defines ϵ as a Jacobi elliptic function. Also, each φ_i will satisfy the equation

$$\varphi_x^{3/2} = b\varphi\varphi_x^{1/2} + e^{-bx/2}\varphi + ae^{bx/2}, \tag{1.27}$$

and, when $b = 0$, Eq. (1.27) is

$$x + c = \int_0^{\phi} \frac{ds}{(s^3 + a)^{2/3}}. \tag{1.28}$$

Equation (1.28) defines a conformal mapping from the interior of a circle of radius $|q|^{1/3}$ in the φ plane into an equilateral triangle in the x plane. On the other hand, when the modulus in Eq. (1.26) is 1, the usual one-soliton solution for the KdV equation is found.

Before considering the periodic fixed points for the KdV equation, we note the following remarks.

Remark 1: In general, it is to be expected that the fixed points of a BT of an integrable system themselves define integrable systems of finite degree. The degree of the system (number of arbitrary constants in the solution) equals the order of the fixed points.

Remark 2: When the order of the fixed point approaches infinity we expect that the solutions of the associated systems are "dense" in the manifold of solutions of the original system.

Remark 3: There may exist analogies between the fixed points of mappings and the fixed points of Bäcklund transformations, especially as regards their stability to perturbation.

II. PERIODIC FIXED POINTS AND THE KdV EQUATION

Consider the periodic fixed point of the Bäcklund transformation (1.20) of order N:

$$\varphi_{j+1,x} = \varphi_j^2/\varphi_{j,x}, \tag{2.1}$$

where

$$j = 1,2,3,\ldots,\mathrm{mod}(N). \tag{2.2}$$

We define the variables

$$\epsilon_j = \varphi_{jx}/\varphi_j, \tag{2.3}$$

$$e^{\theta_j} = \epsilon_j, \tag{2.4}$$

$$V_j = \varphi_{jxx}/\varphi_{jx}, \tag{2.5}$$

and find from Eq. (2.1) the equations

$$\epsilon_{j+1,x}/\epsilon_{j+1} + \epsilon_{jx}/\epsilon_j = \epsilon_j - \epsilon_{j+1}, \tag{2.6}$$

$$\theta_{j+1,x} + \theta_{j,x} = e^{\theta_j} - e^{\theta_{j+1}}, \tag{2.7}$$

$$V_{j+1,x} + V_{j,x} = \tfrac{1}{2}(V_j^2 - V_{j+1}^2), \tag{2.8}$$

where, from (2.1),

$$V_{j+1} + V_j = 2(\varphi_{jx}/\varphi_j) = 2\epsilon_j = 2e^{\theta_j}. \tag{2.9}$$

Next, define the N by N matrices

$$A = \begin{pmatrix} 1 & 1 & 0 & & & \\ 0 & 1 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 1 & & \\ & & 0 & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 & \\ 1 & & & & & 1 \end{pmatrix}, \tag{2.10}$$

$$B = \begin{pmatrix} 1 & -1 & 0 & & & & \\ 0 & 1 & -1 & 0 & 0 & & \\ 0 & 0 & 1 & -1 & & & \\ & & 0 & & \ddots & \ddots & 0 \\ & & & \ddots & & \ddots & -1 \\ -1 & & & & & & 1 \end{pmatrix}. \tag{2.11}$$

It can be shown that

$$\det B = |B| = 0, \tag{2.12}$$

for all N and the one-dimensional null space of B is spanned by the N-vector

$$\hat{b}_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{2.13}$$

Also,

$$|A| = 0 \tag{2.14}$$

for $N = 2k$ and A has a one-dimensional null space spanned by

$$\hat{a}_0 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}. \tag{2.15}$$

Finally, for $N = 2K + 1$,

$$|A| \neq 0, \tag{2.16}$$

and

$$A^{-1} = \tfrac{1}{2}(I + \Omega), \tag{2.17}$$

where

$$\Omega = \begin{pmatrix} 0 & -1 & 1 & -1 & 1 & \cdots & \cdots & -1 & 1 \\ 1 & 0 & -1 & 1 & -1 & \cdots & \cdots & 1 & -1 \\ -1 & 1 & 0 & -1 & 1 & \cdots & \cdots & -1 & 1 \\ 1 & -1 & 1 & 0 & -1 & \cdots & \cdots & 1 & -1 \\ -1 & 1 & -1 & 1 & 0 & -1 & \cdots & -1 & 1 \\ \vdots & & & & & \ddots & \ddots & \vdots & \\ & & & & & & & -1 & \\ -1 & 1 & -1 & 1 & & \cdots & 1 & 0 & \end{pmatrix} \tag{2.18}$$

is a $(2K+1,2K+1)$ antisymmetric matrix

$$\Omega^t = -\Omega, \tag{2.19}$$

with

$$|\Omega| = 0. \tag{2.20}$$

The one-dimensional null space of (2.18) is spanned by the $(2K+1)$ vector

$$\hat{c}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{2.21}$$

and it can be shown that

$$\Omega = A^{-1}B. \tag{2.22}$$

Now, define the N-vectors

$$\hat{V} = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{pmatrix}, \tag{2.23}$$

$$\hat{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}, \tag{2.24}$$

and

$$\hat{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}. \tag{2.25}$$

Then, using (2.2), (2.8), (2.10), and (2.11),

$$A\hat{V}_x = \frac{1}{2} B \begin{pmatrix} V_1^2 \\ V_2^2 \\ \vdots \\ V_N^2 \end{pmatrix}, \tag{2.26}$$

and similar equations for (2.6) and (2.7). For even $N = 2K$, A and B are singular and the contraction of Eq. (2.26) with the null vectors (2.13) and (2.15) obtains the conditions

$$\text{(i)} \quad \frac{\partial}{\partial x} \sum_1^N V_j = 0, \tag{2.27}$$

$$\text{(ii)} \quad \sum_{j=1}^{K} V_{2j-1}^2 = \sum_{j=1}^{K} V_{2j}^2. \tag{2.28}$$

For odd $N = 2K+1$, A is invertible and, using (2.17) and (2.22), we find the equations

$$\hat{V}_x = \frac{1}{2} \Omega \begin{pmatrix} V_1^2 \\ V_2^2 \\ \vdots \\ V_N^2 \end{pmatrix}, \tag{2.29}$$

$$\hat{\theta}_x = \Omega \begin{pmatrix} e^{\theta_1} \\ e^{\theta_2} \\ \vdots \\ e^{\theta_N} \end{pmatrix}, \tag{2.30}$$

and

$$\hat{\epsilon}_x = \begin{pmatrix} \epsilon_1 & & 0 \\ & \epsilon_2 & \\ 0 & & \epsilon_N \end{pmatrix} \Omega \begin{pmatrix} \epsilon_1 & & 0 \\ & \epsilon_2 & \\ 0 & & \epsilon_N \end{pmatrix} \circ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{2.31}$$

From (2.9), (2.10), and (2.7),

$$\hat{\epsilon} = \begin{pmatrix} e^{\theta_1} \\ \vdots \\ e^{\theta_N} \end{pmatrix} = \frac{1}{2} A \hat{V} \tag{2.32}$$

and

$$\hat{V} = (I + \Omega) \begin{pmatrix} e^{\theta_1} \\ \vdots \\ e^{\theta_N} \end{pmatrix} = (I + \Omega)\hat{\epsilon}. \tag{2.33}$$

In this paper, we will require that

$$N = 2K + 1 \tag{2.34}$$

and show that Eqs. (2.29)–(2.31) are completely integrable K-dimensional Hamiltonian systems with K integrals (and one Casimir) in involution.

From (2.29) and (2.30),

$$\hat{V}_x = \Omega \nabla_{\hat{V}} H_3, \tag{2.35}$$

$$\hat{\theta}_x = \Omega \nabla_{\hat{\theta}} J_1, \tag{2.36}$$

where

$$H_3 = \frac{1}{6} \sum_{j=1}^{N} V_j^3, \tag{2.37}$$

$$J_1 = \sum_{j=1}^{N} e^{\theta_j}, \tag{2.38}$$

$$\nabla_{\hat{V}} = \begin{pmatrix} \partial/\partial V_1 \\ \vdots \\ \partial/\partial V_N \end{pmatrix}. \tag{2.39}$$

Using the notation of Ref. 4, Ω, being a constant coefficient, antisymmetric matrix, is cosymplectic and Eqs. (2.35) and (2.36) are Hamiltonian systems. Furthermore, the Hamiltonian systems (2.35) and (2.36) are connected by an invertible "Miura" transformation (2.32), (2.33), which may be written as

$$\hat{V} = (I + \Omega)\nabla_{\hat{\theta}} J_1. \tag{2.40}$$

It is a result of Ref. 4 that if two Hamiltonian systems,

$$\hat{U}_x = \Omega_1 \nabla_{\hat{U}} H_1, \tag{2.41}$$

$$W_x = \Omega_2 \nabla_{\hat{W}} J_1, \tag{2.42}$$

are connected by a Miura transformation

$$\hat{U} = \hat{B}(\hat{W}), \quad \hat{W} = \hat{C}(\hat{U}) \tag{2.43}$$

and

$$\begin{aligned} \nabla_{\hat{W}} \hat{U} &= \nabla_{\hat{W}} \hat{B}, \\ \nabla_{\hat{U}} \hat{W} &= \nabla_{\hat{U}} \hat{C}, \end{aligned} \tag{2.44}$$

then the forms

$$\begin{aligned} \Omega_1' &= (\nabla_W \hat{B})\Omega_2(\nabla_W \hat{B})^*, \\ \Omega_2' &= (\nabla_U \hat{C})\Omega_1(\nabla_U \hat{C})^* \end{aligned} \tag{2.45}$$

are cosymplectic for Eqs. (2.41) and (2.42), respectively. In effect, this obtains the dual-Hamiltonian formulation for Eqs. (2.41) and (2.42) and the recursion operators for the functional gradients of the conserved quantities

$$\begin{aligned} \Omega_1 \nabla_{\hat{U}} H_{j+1} &= \Omega_1' \nabla_{\hat{U}} H_j, \\ \Omega_2 \nabla_{\hat{W}} J_{j+1} &= \Omega_2' \nabla_{\hat{W}} J_j. \end{aligned} \tag{2.46}$$

We refer the reader to Refs. 4–6 for further information.

For Eqs. (2.29)–(2.31), using the formula

$$\nabla_{\hat{\theta}} \hat{V} = (I + \Omega)\nabla_{\hat{\theta}}^2 J_1, \tag{2.47}$$

$$\nabla_{\hat{\theta}}^2 J_1 = \begin{pmatrix} e^{\theta_1} & & 0 \\ & \ddots & \\ 0 & & e^{\theta_N} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} V_1 + V_2 & & 0 \\ & \ddots & \\ 0 & & V_N + V_1 \end{pmatrix}, \tag{2.48}$$

$$\nabla_{\hat{V}} \hat{\theta} = \begin{pmatrix} e^{-\theta_1} & & \\ & \ddots & \\ & & e^{-\theta_N} \end{pmatrix} A, \tag{2.49}$$

$$\nabla_{\hat{V}} \hat{\epsilon} = \tfrac{1}{2} A, \tag{2.50}$$

and the above result, the recursion operators for the gradients of the integrals of Eqs. (2.29)–(2.31) are

$$\Omega \nabla_{\hat{V}} H_{2n+1} = M_{\hat{V}} \nabla_{\hat{V}} H_{2n-1}, \tag{2.51}$$

$$\Omega \nabla_{\hat{\theta}} J_{2n+1} = M_{\hat{\theta}} \nabla_{\hat{\theta}} J_{2n-1}, \tag{2.52}$$

$$\Omega_1 \nabla_{\hat{\epsilon}} H'_{2n+1} = M_{\hat{\epsilon}} \nabla_{\hat{\epsilon}} H'_{2n-1}, \tag{2.53}$$

where

$$M_{\hat{V}} = (I + \Omega) \begin{pmatrix} (V_1 + V_2)/2 & & 0 \\ & \ddots & \\ 0 & & (V_N + V_1)/2 \end{pmatrix} \Omega \begin{pmatrix} (V_1 + V_2)/2 & & 0 \\ & \ddots & \\ 0 & & (V_N + V_1)/2 \end{pmatrix} (I - \Omega), \tag{2.54}$$

$$M_{\hat{\theta}} = \begin{pmatrix} e^{-\theta_1} & & 0 \\ & \ddots & \\ 0 & & e^{-\theta_N} \end{pmatrix} A\Omega A^* \begin{pmatrix} e^{-\theta_1} & & 0 \\ & \ddots & \\ 0 & & e^{-\theta_N} \end{pmatrix}, \tag{2.55}$$

$$\Omega_1 = BA^* = A^*B = \begin{pmatrix} 0 & -1 & 0 & & 0 & 1 \\ 1 & 0 & -1 & & 0 & 0 \\ 0 & 1 & 0 & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & 1 & 0 & \\ -1 & & & & & \end{pmatrix}, \tag{2.56}$$

$$M_{\hat{\epsilon}} = \begin{pmatrix} \epsilon_1 & & 0 \\ & \epsilon_2 & \\ 0 & & \epsilon_N \end{pmatrix} \Omega \begin{pmatrix} \epsilon_1 & & 0 \\ & \epsilon_2 & \\ 0 & & \epsilon_N \end{pmatrix}. \tag{2.57}$$

In general, the forms (2.54)–(2.57) are cosymplectic and singular, since $|\Omega| = 0$. Also, since Ω is singular the systems (2.29)–(2.31) have "Casimir" invariants

$$H_1 = \sum_{j=1}^{N} V_j, \quad J_0 = \sum_{j=1}^{N} \theta_j, \quad H_1' = \sum_{j=1}^{N} \epsilon_j. \tag{2.58}$$

We note the following results:

$$\Omega \nabla_{\hat{V}} H_3 = M_{\hat{V}} \nabla_{\hat{V}} H_1, \tag{2.59}$$

$$\Omega \nabla_{\hat{V}} H_1 = 0, \tag{2.60}$$

$$\Omega \nabla_{\hat{\theta}} J_0 = M_{\hat{\theta}} \nabla_{\hat{\theta}} J_1 = 0, \tag{2.61}$$

$$\Omega_1 \nabla_{\hat{\epsilon}} H_1' = 0. \tag{2.62}$$

It is somewhat nontrivial to solve (2.51)–(2.53) for the integrals since the operators involved are singular. To do so requires a fairly detailed analysis to arrive at the result.

Theorem 1: For the Hamiltonian systems,

$$\hat{V}_x = \Omega \nabla_{\hat{V}} H_3, \tag{2.63a}$$

$$\hat{\epsilon}_x = \Omega_1 \nabla_{\hat{\epsilon}} H_3', \tag{2.63b}$$

there are K integrals

$$H_3, H_5, \ldots, H_{2K+1}, \quad H_3', H_5', \ldots, H_{2K+1}', \tag{2.64}$$

and one Casimir integral

$$H_1 = \sum_{1}^{N} V_j, \quad H_1' = \sum_{1}^{N} \epsilon_j. \tag{2.65}$$

The integrals are

$$H_{2K+1} = \prod_{j=1}^{N} (V_j + V_{j+1}), \tag{2.66}$$

$$H_{2K+1}' = \prod_{j=1}^{N} \epsilon_j, \tag{2.67}$$

$$H_{2K+1-2m} = (1/m!)(-\Delta)^m \circ H_{2K+1}, \tag{2.68}$$

$$H_{2K+1-2m}' = (1/m!) L^m \circ H_{2K+1}', \tag{2.69}$$

for $m = 0,1,2,\ldots,K$, where

$$\Delta = \sum_{j=1}^{N} \frac{\partial^2}{\partial V_j^2}, \tag{2.70}$$

$$L = -\sum_{j=1}^{N} \frac{\partial^2}{\partial \epsilon_j \, \partial \epsilon_{j+1}}. \tag{2.71}$$

Furthermore, the integrals are in involution. That is,

$$(\nabla_{\hat{V}} H_{2j+1})^t \circ \Omega \nabla_{\hat{V}} H_{2l+1} = 0, \tag{2.72}$$

$$(\nabla_{\hat{\epsilon}} H_{2j+1}')^t \circ \Omega_1 \nabla_{\hat{\epsilon}} H_{2l+1}' = 0, \tag{2.73}$$

for $(j,l) \in (0,1,\ldots,K)$. Also, (H_{2j+1}, H_{2j+1}') are homogeneous of degree $2j+1$. That is,

$$\begin{aligned} H_{2j+1}(\lambda \hat{V}) &= \lambda^{2j+1} H_{2j+1}(\hat{V}), \\ H_{2j+1}'(\lambda \hat{\epsilon}) &= \lambda^{2j+1} H_{2j+1}'(\hat{\epsilon}), \end{aligned} \tag{2.74}$$

for $j = 0,1,2,3,\ldots,K$.

Finally,

$$M_{\hat{V}} \nabla_{\hat{V}} H_{2K+1} = 0, \quad M_{\hat{\epsilon}} \nabla_{\hat{\epsilon}} H_{2K+1}' = 0, \tag{2.75}$$

which implies that

$$\begin{aligned} \Omega \nabla_{\hat{V}} H_{2K+3} &= \Omega \nabla_{\hat{V}} H_1 = 0, \\ \Omega_1 \nabla_{\hat{\epsilon}} H_{2K+3}' &= \Omega_1 \nabla_{\hat{\epsilon}} H_1' = 0, \end{aligned} \tag{2.76}$$

and, in terms of the recursion operators (2.51) and (2.53), the integrals $(H_{2j+1}, H_{2j+1}'; j = 0,1,\ldots,K)$ provide a basis for their solutions.

To begin, it is enough to solve one of the sequences (2.51)–(2.53) since the results for each follow by a change of variable. For instance,

$$M_{\hat{V}} = (I + \Omega) M_{\hat{\epsilon}} (I - \Omega), \tag{2.77}$$

$$\nabla_{\hat{\epsilon}} = (I - \Omega) \nabla_{\hat{V}}, \tag{2.78}$$

where

$$\hat{\epsilon} = \tfrac{1}{2} A \hat{V}.$$

Now, consider the sequence

$$\Omega \hat{h}_{2n+1} = M_{\hat{V}} \hat{h}_{2n-1}, \tag{2.79}$$

where the $\hat{h}_{2j+1}$ are N-vectors (not necessarily the gradients of integrals) and

$$\hat{h}_1 = \nabla_{\hat{V}} H_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{2.80}$$

Since Ω is singular with null vector (2.80), the right-hand side of (2.79) must be orthogonal to (2.80), for every n, to be solvable. That is,

$$\hat{h}_1^t \circ \Omega \hat{h}_{2n+1} = \hat{h}_1^t \circ M_V \hat{h}_{2n-1} = 0. \tag{2.81}$$

However, using the antisymmetry of $M_{\hat{V}}$ and induction on n,

$$\begin{aligned} \hat{h}_1^t \circ M_V \hat{h}_{2n-1} &= -\hat{h}_{2n-1} \circ M_V \hat{h}_1 = -\hat{h}_{2n-1} \circ \Omega \hat{h}_3 \\ &= \hat{h}_3 \circ \Omega \hat{h}_{2n-1} = \hat{h}_3 \circ M_V \hat{h}_{2n-3}, \end{aligned} \tag{2.82}$$

and after j steps

$$\hat{h}_1^t \circ M_{\hat{V}}\, \hat{h}_{2n-1} = -\hat{h}_{2n-1-2j} \circ \Omega \hat{h}_{2j+3} = \hat{h}_{2j+3} \circ M_{\hat{V}}\, \hat{h}_{2n-3-2j}, \tag{2.83}$$

which vanishes identically when

$$j = (n+1)/2 \quad \text{or} \quad j = n/2. \tag{2.84}$$

Therefore, the sequence (2.79) beginning from (2.80) exists (although it is not unique). An identical argument, using the identity

$$\hat{h}^t_{2m+1} \circ \Omega \hat{h}_{2n+1} = -\hat{h}^t_{2n-1} \circ \Omega \hat{h}_{2m+3} \tag{2.85}$$

to raise or lower indices, establishes that the "symmetries" $\{\hat{h}_{2j+1}\}$ are in involution:

$$\hat{h}^t_{2m+1} \circ \Omega \hat{h}_{2n+1} = 0. \tag{2.86}$$

Since

$$\hat{V}_x = \Omega \nabla_{\hat{V}} H_3 = M_{\hat{V}} \nabla_{\hat{V}} H_1, \tag{2.87}$$

the symmetries $\{\hat{h}_{2j+1}\}$, which are the gradients of functions, will obtain integrals of (2.87) that are in involution.

To find the integrals it is convenient to consider the sequence (2.53). Let

$$H_{2K+1} = \prod_{j=1}^{N} \epsilon_j. \tag{2.88}$$

Then

$$M_{\hat{\epsilon}} \nabla_{\hat{\epsilon}} H_{2K+1} = 0. \tag{2.89}$$

To see this, note that for every j, $j = 1,2,3,\ldots,N$, where $N = 2K+1$,

$$\epsilon_j \frac{\partial}{\partial \epsilon_j} H_{2K+1} = H_{2K+1}. \tag{2.90}$$

Therefore,

$$\nabla_{\hat{\epsilon}} H_{2K+1} = H_{2K+1} \nabla_{\hat{\epsilon}} H_1 = H_{2K+1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{2.91}$$

and

$$M_{\hat{\epsilon}} \nabla_{\hat{\epsilon}} H_{2K+1} = H_{2K+1} \begin{pmatrix} \epsilon_1 & & \\ & \ddots & \\ & & \epsilon_N \end{pmatrix} \Omega \nabla_{\hat{\epsilon}} H_1 = 0, \tag{2.92}$$

where

$$H_1 = \sum_{j=1}^{N} \epsilon_j. \tag{2.93}$$

Now, define the operator

$$L = -\sum_{j=1}^{N} \frac{\partial^2}{\partial \epsilon_j\, \partial \epsilon_{j+1}}, \tag{2.94}$$

and apply the operator m times to (2.90) to obtain the identity

$$\epsilon_j \frac{\partial}{\partial \epsilon_j} L^m \circ H_{2K+1} = L^m \circ H_{2K+1} + m \frac{\partial}{\partial \epsilon_j} \left(\frac{\partial}{\partial \epsilon_{j-1}} + \frac{\partial}{\partial \epsilon_{j+1}} \right) L^{m-1} \circ H_{2K+1}, \tag{2.95}$$

for each $j = 1,2,\ldots,\mathrm{mod}(N)$. Therefore,

$$\begin{pmatrix} \epsilon_1 & & 0 \\ & \epsilon_2 & \\ & \ddots & \\ 0 & & \epsilon_N \end{pmatrix} \nabla_{\hat{\epsilon}} L_m \circ H_{2K+1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} L^m \circ H_{2K+1} + m \begin{pmatrix} D_1(D_N + D_2) \\ \vdots \\ D_j(D_{j-1} + D_{j+1}) \\ \vdots \\ D_N(D_{N-1} + D_1) \end{pmatrix} L^{m-1} \circ H_{2K+1}, \tag{2.96}$$

where

$$D_j = \frac{\partial}{\partial \epsilon_j}.$$

From the definition of Ω, (2.18), the first term on the rhs of (2.96) vanishes when applied to Ω and the second term obtains

$$M_\epsilon \nabla_\epsilon L^m \circ H_{2K+1} = m \begin{pmatrix} \epsilon_1 & & 0 \\ & \epsilon_2 & \\ & \ddots & \\ 0 & & \epsilon_N \end{pmatrix} \begin{pmatrix} D_1(-D_2 + D_N) \\ \vdots \\ D_j(-D_{j-1} + D_{j+1}) \\ \vdots \\ D_N(-D_{N-1} + D_1) \end{pmatrix} L^{m-1} \circ H_{2K+1}. \tag{2.97}$$

From (2.95),

$$\epsilon_j D_j(-D_{j-1}+D_{j+1})L^{m-1}\circ H_{2K+1} \tag{2.98}$$
$$=(-D_{j-1}+D_{j+1})\epsilon_j D_j L^{m-1}\circ H_{2K+1}$$
$$=(-D_{j-1}+D_{j+1})\{L^{m-1}\circ H_{2K+1}+(m-1)D_j(D_{j-1}+D_{j+1})L^{m-2}\circ H_{2K+1}\}$$
$$=(-D_{j-1}+D_{j+1})L^{m-1}\circ H_{2K+1}+(m-1)D_j(-D_{j-1}^2+D_{j+1}^2)L^{m-2}\circ H_{2K+1}$$
$$=(-D_{j-1}+D_{j+1})L^{m-1}\circ H_{2K+1}, \tag{2.99}$$

since

$$(-D_{j-1}^2+D_{j+1}^2)L^{m-2}\circ H_{2K+1}$$
$$=L^{m-2}(-D_{j-1}^2+D_{j+1}^2)H_{2K+1}=0. \tag{2.100}$$

Combining (2.99) and (2.97),

$$\Omega_1\nabla_{\hat\epsilon}L^{m-1}\circ H_{2K+1}=(1/m)M_{\hat\epsilon}\nabla_{\hat\epsilon}L^m\circ H_{2K+1}, \tag{2.101}$$

where Ω_1 is defined by (2.56). Therefore, by induction on m,

$$H'_{2K+1-2m}=(1/m!)L^m\circ H_{2K+1}, \tag{2.102}$$

for $m=0,1,2,\ldots,K$, will satisfy the recursion relation

$$\Omega_1\nabla_{\hat\epsilon}H'_{2K+3-2m}=M_{\hat\epsilon}\nabla_{\hat\epsilon}H'_{2K+1-2m}, \tag{2.103}$$

for $m=1,2,3,\ldots,K$. Furthermore

$$H'_1=\frac{1}{K!}L^K\circ H_{2K+1}=(-1)^K\sum_{j=1}^N\epsilon_j \tag{2.104}$$

and

$$\Omega_1\nabla_{\hat\epsilon}H_1=M_{\hat\epsilon}\nabla_{\hat\epsilon}H_{2K+1}=0. \tag{2.105}$$

Equations (2.105) and (2.103) show that, in spite of the nonuniqueness of solutions of the recursion (2.103), the integrals (2.102) provide a basis for the solution of (2.103) in that this set of integrals is closed under the recursion (2.103). The general solution of (2.103) can be found by adding an arbitrary function of H_1 to a H'_{2j+1} at each step of (2.103) (since the gradient of H_1 is in the null space of Ω_1) and allowing for the effect of this addition under the recursion (2.103).

It is immediate that

$$H'_{2j+1}(\lambda\hat\epsilon)=\lambda^{2j+1}H'_{2j+1}(\hat\epsilon), \tag{2.106}$$

for $j=1,2,3,\ldots,K$.

By the previous remarks the $\{H'_{2j+1};\ j=0,1,2,\ldots,K\}$ define $(K+1)$ integrals in involution and, by construction, these integrals are independent. Thus, (2.63b) is a K-dimensional completely integrable, Hamiltonian system with one Casimir integral.

After the change of variable [(2.77) and (2.78)] and a convenient scaling, we find that

$$H_{2K+1}=\prod_1^N(V_j+V_{j+1}), \tag{2.107}$$

$$H_{2K+1-2m}=(1/m!)(-\Delta)^m\circ H_{2K+1}, \tag{2.108}$$

where $m=0,1,2,\ldots,K$ and

$$\Delta=\sum_1^N\frac{\partial^2}{\partial V_j^2}, \tag{2.109}$$

are $(K+1)$ independent integrals of (2.63a) in involution, satisfying the recursion (2.51). The properties of (2.63a) follow directly from those of (2.63b) and will not be repeated.

We note that, as defined by (2.108),

$$H_3=(-1)^{K-1}\left\{\frac{1}{6}\left(\sum_{j=1}^N V_j\right)^3-\frac{1}{6}\left(\sum V_j^3\right)\right\}, \tag{2.110}$$

which is not identical to (2.37), where

$$H_3''=\frac{1}{6}\sum V_j^3. \tag{2.111}$$

However, except for a possible difference in sign,

$$\Omega\nabla_{\hat V}H_3=\Omega\nabla_{\hat V}H_3'' \tag{2.112}$$

and insofar as the dynamics of (2.35) and (2.63b) is concerned there is no difference.

We turn next to the time dependence of the fixed points.

Theorem 2: For the modified KdV equation

$$V_t+\frac{\partial}{\partial x}\left(V_{xx}-\frac{1}{2}V^3\right)=0, \tag{2.113}$$

the order $N=2K+1$ periodic fixed point of the Bäcklund transformation is determined by the commuting (completely integrable) Hamiltonian systems

$$\hat V_x=\Omega\nabla_{\hat V}H_3''=\frac{1}{2}\Omega\begin{pmatrix}V_1^2\\ \vdots\\ V_N^2\end{pmatrix}, \tag{2.114}$$

$$\hat V_t=\Omega\nabla_{\hat V}H_5''=M_{\hat V}\nabla_{\hat V}H_3'', \tag{2.115}$$

where

$$H_3''=\frac{1}{6}\sum_{j=1}^N V_j^3. \tag{2.116}$$

Remark 1: The H_5'' of (2.115) is not the H_5 of (2.108) but differs by a constant multiple of $H_1^2H_3$ [as defined by (2.108)]. The time dynamics determined by H_5 is equivalent to a partial Galilean transformation $(t\to t, x\to x+ct)$ of Eq. (2.113).

To see (2.115) we can begin from the BT

$$\frac{\varphi_{jt}}{\varphi_{jx}}+\frac{\varphi_{j+1x}}{\varphi_{j+1x}}+4\frac{\partial}{\partial x}\left(\frac{\varphi_{jx}}{\varphi_j}\right)=\left(\frac{\varphi_{jxx}}{\varphi_{jx}}\right)^2 \tag{2.117}$$

or directly from

$$\hat V_t+\frac{\partial}{\partial x}\begin{pmatrix}V_{1xx}-\frac{1}{2}V_1^3\\ \vdots\\ V_{Nxx}-\frac{1}{2}V_N^3\end{pmatrix}=0 \tag{2.118}$$

and use (2.114) to find that

$$V_t=\frac{1}{2}\left\{\frac{1}{2}\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}\Omega+\frac{1}{2}\Omega\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}-\Omega\begin{pmatrix}V_1 & & \\ & \ddots & \\ & & V_N\end{pmatrix}\Omega\begin{pmatrix}V_1 & & \\ & \ddots & \\ & & V_N\end{pmatrix}\Omega\right\}\begin{pmatrix}V_1^2\\ \vdots\\ V_N^2\end{pmatrix}, \tag{2.119}$$

where we use the identity

$$\begin{pmatrix}V_{1x} & & 0\\ & \ddots & \\ 0 & & V_{Nx}\end{pmatrix}-\Omega\begin{pmatrix}V_{1x} & & 0\\ & \ddots & \\ 0 & & V_{Nx}\end{pmatrix}\Omega=\frac{1}{2}\left\{\Omega\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}-\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}\Omega\right\}. \tag{2.120}$$

Let

$$\Omega_2=\frac{1}{2}\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}\Omega+\frac{1}{2}\Omega\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}-\Omega\begin{pmatrix}V_1 & & 0\\ & \ddots & \\ 0 & & V_N\end{pmatrix}\Omega\begin{pmatrix}V_1 & & 0\\ & \ddots & \\ 0 & & V_N\end{pmatrix}\Omega. \tag{2.121}$$

We show that

$$M_{\hat{V}}=\Omega_2. \tag{2.122}$$

Using

$$A^{-1}=\tfrac{1}{2}(I+\Omega),\quad A^{*-1}=\tfrac{1}{2}(I-\Omega), \tag{2.123}$$

$$\Omega=A^{-1}B,\quad \Omega A^*=-B^*,\quad \Omega_2=\tfrac{1}{4}(I+\Omega)\Omega_3(I-\Omega), \tag{2.124}$$

where

$$\Omega_3=\frac{1}{2}B\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}A^*-\frac{1}{2}A\begin{pmatrix}V_1^2 & & 0\\ & \ddots & \\ 0 & & V_N^2\end{pmatrix}B^*+B\begin{pmatrix}V_1 & & 0\\ & \ddots & \\ 0 & & V_N\end{pmatrix}\Omega\begin{pmatrix}V_1 & & 0\\ & \ddots & \\ 0 & & V_N\end{pmatrix}B^*. \tag{2.125}$$

It is straightforward to find

$$\Omega_3=\begin{pmatrix}V_1+V_2 & & & 0\\ & V_2+V_3 & & \\ & & \ddots & \\ 0 & & & V_N+V_1\end{pmatrix}\Omega\begin{pmatrix}V_1+V_2 & & & 0\\ & V_2+V_3 & & \\ & & \ddots & \\ 0 & & & V_N+V_1\end{pmatrix} \tag{2.126}$$

and

$$\Omega_2=M_V,$$

which establishes (2.115).

Remark 2: The system (2.114) is a scale-invariant Hamiltonian system of the type studied by Yoshida.[7] That is, (2.114) is invariant under the scaling

$$X\to a^{-1}X,\quad \hat{V}\to a\hat{V}. \tag{2.127}$$

By a leading order analysis

$$\hat{V}\sim X^{-1}\hat{C}, \tag{2.128}$$

it can be shown that there is an N-vector $\hat{c}$,

$$\hat{c}=(2,4,6,8,\ldots,2K,-2K,-2K+2,\ldots,-2,0)^t, \tag{2.129}$$

and resonances

$$r=-2K+1,-2K+3,\ldots,-1,1,3,5,\ldots,2K+1. \tag{2.130}$$

By Ref. 7 the resonances $\{1,3,\ldots,2K+1\}$ correspond to the homogeneous invariants $\{H_{2j+1};j=0,1,\ldots,K\}$.

In a somewhat different direction, the general system (2.26) (for arbitrary N) has a commutator representation

that is related to the factorization method of Infeld and Hull.[8] Define the "raising" and "lowering" operators

$$L_j^+ = D + \tfrac{1}{2}V_j, \tag{2.131}$$

$$L_j^- = D - \tfrac{1}{2}V_j, \tag{2.132}$$

where $D = \partial/\partial x$ and the N by N matrix operators

$$L_\lambda = \begin{pmatrix} \lambda & L_1^+ & & & \\ 0 & \lambda & \ddots & L_2^+ & 0 \\ & & \ddots & \ddots & \\ & & 0 & \ddots & L_{N-1}^+ \\ L_N^+ & & & & \lambda \end{pmatrix}, \tag{2.133}$$

$$M_\sigma = \begin{pmatrix} \sigma & & & & L_{N-1}^- \\ L_1^- & \sigma & & 0 & \\ & L_2^- & \ddots & & \\ 0 & & \ddots & \ddots & \\ & & & L_{N-1}^- & \sigma \end{pmatrix}. \tag{2.134}$$

Then

$$L_\lambda M_\sigma = M_\sigma L_\lambda \tag{2.135}$$

if and only if

$$L_{j+1}^+ L_{j+1}^- = L_j^- L_j^+, \tag{2.136}$$

for $j = 1,2,3,\ldots,\mathrm{mod}(N)$. Condition (2.136) is

$$V_{jx} + V_{j+1\,x} = \tfrac{1}{2}(V_j^2 - V_{j+1}^2), \tag{2.137}$$

which is

$$A\hat{V}_x = \frac{1}{2} B \begin{pmatrix} V_1^2 \\ \vdots \\ V_N^2 \end{pmatrix}, \tag{2.138}$$

or, Eqs. (2.26). When $N = 2K+1$, (2.138) is the system (2.114). With

$$L = L_\lambda - \lambda I, \quad M = M_\sigma - \sigma I, \tag{2.139}$$

it can be shown that

$$L^N = \begin{pmatrix} L_1^+ L_2^+ \cdots L_N^+ & & & 0 \\ & L_2^+ L_3^+ \cdots L_1^+ & & \\ 0 & & \ddots & \\ & & & L_N^+ L_1^+ \cdots L_{N-1}^+ \end{pmatrix}, \tag{2.140}$$

$$M^N = \begin{pmatrix} L_N^- L_{N-1}^- \cdots L_1^- & & & 0 \\ & L_1^- L_N^- \cdots L_2^- & & \\ 0 & & \ddots & \\ & & & L_{N-1}^- L_{N-2}^- \cdots L_N^- \end{pmatrix}. \tag{2.141}$$

From (2.135), or directly using (2.136), the pair of scalar operators

$$P = L_1^+ L_2^+ \cdots L_N^+, \quad Q = L_N^- L_{N-1}^- \cdots L_1^-, \tag{2.142}$$

and other associated pairs, commute:

$$PQ = QP. \tag{2.143}$$

Although $\deg(P) = \det(Q) = N$ (not relatively prime), the operators (P,Q) will, according to Burchnall and Chaundy,[9] satisfy an algebraic identity of degree N, or less, in (P,Q). We note that

$$L^* = -M, \quad Q^* = (-1)^N P. \tag{2.144}$$

For $N = 3,5$, we have found that the matrix operators (L,M) satisfy

$$L^3 - M^3 = h_1(LM) + h_3, \tag{2.145}$$

$$L^5 - M^5 = h_1(LM)^2 + h_3(LM) + h_5, \tag{2.146}$$

where the h_{2j+1} are integrals of (2.137), homogeneous of degree $(2j+1)$.

The modified KdV sequence is defined to be[1]

$$V_t + L_V^n \circ V_X = 0, \tag{2.147}$$

where $n = 0,1,2,\ldots$ and

$$L_V = D(D+V)D^{-1}(D-V). \tag{2.148}$$

We conjecture that the time dependence of the periodic fixed points evolves according to the system

$$\hat{V}_t = \Omega \nabla_V H'_{2n+3}, \tag{2.149}$$

for $n = 0,1,2,3,\ldots$, and where

$$\hat{V}_x = \Omega \nabla_V H_3. \tag{2.150}$$

The H'_{2n+3} are suitable integrals of (2.150), homogeneous of degree $2n+3$. With $N = 2K+1$, we have found previously (2.105) that

$$\Omega \nabla_{\hat{V}} H_{2K+3} = \Omega \nabla_{\hat{V}} H_1 = M_{\hat{V}} \nabla_{\hat{V}} H_{2K+1} = 0, \tag{2.151}$$

or, for the $(2K+3)$ MKdV flow associated with H_{2K+3},

$$\hat{V}_t = 0. \tag{2.152}$$

Therefore, if the conjecture is verified, the evolution of the fixed points occurs on a manifold of steady state solutions of a higher-order equation (a Lax–Novikov equation) and the theory of periodic fixed points of Bäcklund transformation is, in this sense, equivalent to the theory of the finite-zone potentials.[10,11]

ACKNOWLEDGMENTS

This work was supported by the Air Force Office of Scientific Research Grant No. AFOSR 84-0128.

[1]J. Weiss, J. Math. Phys. **25**, 13 (1984).
[2]E. Hille, *Ordinary Differential Equations in the Complex Plane* (Wiley, New York, 1976).
[3]J. Weiss, J. Math. Phys. **24**, 1405 (1983).
[4]B. Fuchssteiner and A. S. Fokas, Physica D **4**, 47 (1981).
[5]A. S. Fokas and R. L. Anderson, J. Math. Phys. **23**, 1066 (1982).
[6]B. A. Kupershmidt and G. Wilson, Invent. Math. **62**, 403 (1981).
[7]H. Yoshida, Celestial Mech. **81**, 363, 381 (1983).
[8]L. Infeld and T. Hull, Rev. Mod. Phys. **23**, 21 (1951).
[9]J. L. Burchnall and T. W. Chaundy, Proc. R. Soc. London A Ser. **118**, 557 (1928).
[10]P. D. Lax, Commun. Pure Appl. Math. **28**, 141 (1975).
[11]I. M. Gelfand and L. A. Dikii, *Integrable Systems*, edited by S. P. Novikov (Cambridge U. P., London, 1981), p. 13; B. A. Dubrovin, V. B. Matveer, and S. P. Novikov, *ibid.* p. 53.

On the stable analytic continuation with a condition of uniform boundedness

I. Sabba Stefanescu
Institut für Theoretische Kernphysik, Universität Karlsruhe, Karlsruhe, West Germany

(Received 11 November 1985; accepted for publication 2 July 1986)

It is shown that, if $h(x)$ is any continuous function defined on some interval $[-a,b] \subset (-1,1)$ of the real axis, then, in general, its best L^2 approximant, in the class of functions holomorphic and bounded by unity in the unit disk of the complex plane, is a finite Blaschke product. An upper bound is placed on the number of factors of the latter and a method for its construction is given. The paper contains a discussion of the use of these results in performing a stable analytic continuation of a set of data points under a condition of uniform boundedness, as well as some numerical examples.

I. INTRODUCTION

In this paper, we consider the following problem: (A) let $h(x)$ be a continuous function defined on the segment $[-a,b] \subset (-1,1)$, and $\rho(x)$ an increasing function with normalized bounded variation defined on the same interval. The distance from $h(x)$ to another continuous function $f(x)$ on $[-a,b]$ is measured by

$$\chi^2(\rho;h-f) = \int_{-a}^{b} (h(x)-f(x))^2 \, d\rho(x) . \tag{1.1}$$

Let further $H_R^\infty(D)$ be the set of functions real holomorphic in the open unit disk D: $|z|<1$ $(z=x+iy)$ and uniformly bounded there, with the norm

$$\|f\|_\infty \equiv \sup_{|z|<1} |f(z)| . \tag{1.2}$$

The question is to show the existence and to describe that element of $H_R^\infty(D)$ which realizes

$$\chi^2_{\min}(\rho;h) \equiv \inf\{\chi^2(\rho;h-f):\ f \in H_R^\infty(D),\ \|f\|_\infty \leqslant 1\} , \tag{1.3}$$

i.e., the best $L^2(\rho)$-approximant of $h(x)$ in terms of functions of $H_R^\infty(D)$, bounded by unity inside D.

In Sec. II, we shall argue that the situation $\chi^2_{\min} > 0$ is in a certain sense the usual one. Then, if a number of χ_1^2 exists so that $\chi^2_{\min} > \chi_1^2 > 0$, we show that (i) the unique function on which the infimum is attained is a finite Blaschke product and (ii) an upper bound may be placed on the number of its factors, dependent on $h(x)$ and χ_1^2. An estimate of χ_1^2 may be obtained from the solution of the related H_R^2 problem: (A') find

$$\chi^2_{\min,2}(\rho;h) \equiv \inf\{\chi^2(\rho;h-f):\ f \in H_R^2(D),\ \|f\|_2 \leqslant 1\} , \tag{1.4}$$

where $H_R^2(D)$ is the analog of $H_R^\infty(D)$ with the norm

$$\|f\|_2^2 \equiv \frac{1}{2\pi} \oint |f(e^{i\theta})|^2 \, d\theta . \tag{1.5}$$

The problem (A') is explicitly solvable (see Sec. IV for a reminder) and, since the infimum in (1.4) is taken on a larger set, it is true that

$$\chi^2_{\min,2}(\rho;h) \leqslant \chi^2_{\min}(\rho;h) . \tag{1.6}$$

Therefore, if $\chi^2_{\min,2}(\rho;h) > 0$ (which is the usual situation), we do obtain an upper bound on the number of factors of the Blaschke product in the solution of problem (A), depending only on the "data function" $h(x)$. In Sec. III we shall show, however, even more, i.e., there exists a nice connection between, roughly speaking, the dual extremal problem that arises in relation to (A') and the dual problem related to (A). This allows us to set up an algorithm for the determination of the extremal element of (A), which avoids resorting to an opaque and cumbersome minimization in a space of Blaschke parameters with dimension equal to the bound mentioned above. We reduce, namely, the problem to that of solving a certain nonlinear equation for which the convergence of an algorithm related to the Newton–Kantorowich procedure may be completely analyzed (see Secs. III and VI and Appendix C).

The fact that the function realizing the extremum (1.4) is a finite Blaschke product if $\chi^2_{\min} > 0$ was derived by different means (using the Schur–Pick–Nevanlinna interpolation theory) in Ref. 1, for the case of a piecewise constant weight function $\rho(x)$ with N jumps. It was shown there that the extremal product consists of at most N Blaschke factors. The present paper provides a refinement in that, for large N, it turns out that the number of factors does not increase indefinitely, as long as $\chi^2_{\min}(\rho;h)$ stays larger than any given $\chi_1^2 > 0$.

The argument we present concerning the existence and uniqueness of the finite Blaschke product that realizes (1.3) rests upon (a slight generalization of) a theorem of Rogosinski and Shapiro (Ref. 2), concerning the extrema of linear functionals defined over $H^\infty(D)$ (see also Ref. 3 for a related treatment). In Sec. II we repeat the reasoning of Ref. 2 in the present setting, both for the convenience of the reader and because several of its intermediate steps are of further relevance to this paper.

The interest in the solution of problem (A) is in the first place of a mathematical nature; it provides an example of a rather complicated looking extremal problem (of a mixed H^2–H^∞ type) whose solution always lies in a finite-dimensional class of functions. In fact, the results stated above may presumably look more surprising if they are formulated as follows: Consider any real analytic function $f_0(z)$, holomorphic and bounded by unity in the unit disk and let it be affected by errors on some interval $[-a,b]$ of the real diameter of the disk, in such a way that one obtains a (continuous) func-

 0022-2488/86/112657-30$02.50

tion $h(x)$, which admits of no analytic extension to $|z|<1$. Then, look for the best fit to $h(x)$ in the class $\|f\|_\infty<1$: one always obtains a Blaschke product as a result and never the original f_0, regardless of the manner in which the errors have been distributed.

There are also other reasons for studying question (A): its solution allows an answer to the following problem (B): let $S(\chi_0;h)$ be the set of functions f in $H_R^\infty(D)$ obeying

$$\chi^2(\rho;h-f)\leqslant\chi_0^2\,, \tag{1.7}$$

for some $\chi_0^2>0$. One is required to show the existence and describe that element of $H_R^\infty(D)$ which realizes

$$M_0(\chi_0;h)\equiv\inf\|f\|_\infty\equiv\inf\sup_{|z|<1}|f(z)| \tag{1.8}$$

over all f in $S(\chi_0;h)$.

The solution $f_0(z)$ of problem (B) provides a stable analytic continuation of the data $h(x)$ to the whole interior of the unit disk (see Refs. 4–6 and also Sec. VI of this paper). Usually, for such applications, $\rho(x)$ is a piecewise constant function, with jumps equal to $1/N\epsilon_i^2$ at the points $x_i\in[-a,b]$, where ϵ_i is the estimated error of the measurement $h(x_i)$ at x_i of a certain function $f_0(z)$, holomorphic in D. The most well-known instance in high energy physics where such an extrapolation is required is presumably the Chew–Low–Goebel extrapolation of the $\pi N\to\pi\pi N$ differential cross section to the (second-order) pion pole.[7,8] To be sure, there exist other, simpler, methods to achieve the same end: one of them is to map the unit disk onto the natural domain of convergence of a series of polynomials (Refs. 9 and 10), orthogonal on the image of $[-a,b]$ and truncate the series suitably (see also Ref. 11). Also, methods that are numerically very successful have been developed for the situation when the stabilizing condition $\|f\|_\infty\leqslant1$ is replaced by the condition $\|f\|_2\leqslant1$, or variants of it (Refs. 12–14).

It is of interest to consider also modifications (A_c), (B_c) of problems (A) and (B): one looks for the extrema (1.3) and (1.8) under the n_c supplementary conditions

$$f(x_{0,i})=f_{0;i}\,, \tag{1.9}$$

for some points $\{x_{0,i}\}_{i=1}^{n_c}$, lying (in general) outside $[-a,b]$, and fixed, given values $f_{0,i}$.

The solutions of problems (A_c) and (B_c), for $n_c=1$ may be used to give a numerical answer to the following question (C): given a point x outside $[-a,b]$, find all possible values assumed at x by functions $f(z)$ in $S(\chi_0;h)\cap\{f: f\in H_R^\infty(D),\ \|f\|_\infty\leqslant1\}$. This problem [with $\rho(x)$ piecewise constant] has received attention over the past decade in relation to the interpolation of the spacelike data on the pion form factor $F_\pi(t)$.[1,15,16] Measurements in the timelike region provide a bound $M(t)$ to $F_\pi(t)$ and the problem of finding the allowed values of $F_\pi(t)$ at some point $t_1<0$ outside the (spacelike) data region may be reduced to (C), after dividing off an outer function

$$E(z)=\exp\frac{1}{2\pi}\oint\frac{e^{i\theta}+z}{e^{i\theta}-z}\ln M(\theta(t))d\theta\,, \tag{1.10}$$

where $z=z(t)$ is a well-known conformal mapping.[1] In fact, the form factor is constrained by $F_\pi(t=0)=1$ and then problem (C) may be reduced to (A_c) and (B_c), with $n_c=2$.

A problem similar to (A) and (B) above was considered in Ref. 17. There, the set of functions $F(z)$ was considered, with $F(z)$ represented as

$$F(z)=\frac{1}{\pi}\int_\alpha^\beta\frac{f(t)}{t-z}\,dt\,, \tag{1.11}$$

where $f(t)$ is real, and we looked for the smallest possible value of $\chi^2(\rho;h-F)$, defined over a data region outside $[\alpha,\beta]$, among all $F(x)$ satisfying (1.11) with

$$\|f\|_1=\int_\alpha^\beta|f(t)|dt<A\,, \tag{1.12}$$

for a given A. It turned out that, if $\chi^2_{\min}(\rho;h)>\chi_1^2>0$, then the extremal function is made up of a finite number of δ-functions and a bound, depending on $h(x)$ and χ_1^2, was derived for the number of the latter. The proofs in Sec. IV of this paper bear some similarity to those of Ref. 17 but are more involved.

A recent study exists[18] concerning problem (C) above, with the distance (1.1) between functions on $[-a,b]$ replaced by the uniform norm. For arbitrary $h(x)$, this form of the problem is considerably harder than the one of this paper. The authors solve it completely in the special case $h(x)\equiv0$ and show that, as in problem (C) of this paper, Blaschke products realize the extreme allowed values at a given x outside $[-a,b]$. They are also able to make definite statements concerning the location of the zeros of these particular Blaschke products, which turn out to be a natural (and surprising) generalization of the Chebyshev polynomials.

The present paper is organized as follows. In Sec. II, after a short geometrical discussion, we derive the announced result concerning problem (A): the extremal function is a finite Blaschke product. In Sec. III, we discuss the determination of the sets $\|f\|_\infty\leqslant1$ and $\|f\|_2\leqslant1$ in terms of their supporting hyperplanes and obtain, essentially, a description of the former set by means of the latter. This allows us to reduce the solution of (A) to that of a nonlinear integral equation. Further, using geometrical and function theoretical arguments, we show that the operators appearing in this equation have pleasant differential properties (in particular, their Fréchet derivative is, in general, invertible). In Sec. IV, we derive the bound announced above concerning the number of zeros of the extremal Blaschke product. Section V extends the results of Secs. II–IV to the case when the values of the extremal function are prescribed at some points [problems (A_c) and (B_c)]. In Sec. VI, we discuss the basis of the numerical solution of problems (B) and (C) and, finally, in Sec. VII, we present some numerical examples and conclusions.

Appendix A extends the geometrical discussion of Sec. II and establishes a relation between a solution of problem (C) with $\chi^2_{\min}=0$ and the methods of this paper where $\chi^2_{\min}>0$. Appendices B and C are complements to Sec. III and V and Appendix D discusses the convergence of the algorithms used for the solution of the integral equations of Secs. III and V. Appendix E removes an assumption made for simplicity throughout the text.

II. THE DESCRIPTION OF THE EXTREMAL FUNCTIONS

We shall describe the functions realizing the extremum (1.3) in the family

$$\mathscr{F}_{a,\infty} \equiv \{f: f \in H_R^{\infty}(D), \|f\|_{\infty} \leqslant a\} (\equiv \mathscr{F}_a) \tag{2.1}$$

with $a = 1$. The generalization to arbitrary a is obvious and will be used in Sec. VI. It is advantageous to distinguish between $\mathscr{F}_{a,\infty}$ and $\tilde{\mathscr{F}}_{a,\infty}$ ($\equiv \tilde{\mathscr{F}}_a$), the image of $\mathscr{F}_{a,\infty}$ in $L^2(\rho)$ by the inclusion $H_R(D) \subset L^2(\rho)$. If $\rho(x)$ is of finite type (i.e., has only N points of increase), two distinct elements of $\mathscr{F}_a$ need not be distinct in $\tilde{\mathscr{F}}_a$. We suppose that

$$\chi^2_{\min}(\rho;h) \neq 0. \tag{2.2}$$

This condition requires some comments.

Assume first that $\rho(x)$ has on some interval $(a', b') \subset (-a,b)$ a strictly positive derivative. Then, condition (2.2) means that no extension of $h(x)$ outside the points of increase of $\rho(x)$ leads to a function in $\mathscr{F}_1$. The question is how likely is this situation. To this end, we notice that every $f(x)$ in $\mathscr{F}_1$ may be approximated arbitrarily well in the sense of the norm (1.1) by functions continuous on $[-a,b]$ and lying outside $\tilde{\mathscr{F}}_1$, e.g., by modifying $f(x)$ and (a',b') to broken lines joining points $(x_i, f(x_i))$, $(x_{i+1}, f(x_{i+1}))$, $a' < x_i < b'$. Thus, $\tilde{\mathscr{F}}_1$ does not contain any ball $S(\chi_0; f)$ centered at f, no matter how small χ_0. On the other hand, we shall show below that $\tilde{\mathscr{F}}_1$ is closed with respect to the convergence generated by (1.1); therefore, any point outside it has a neighborhood disjoint from $\tilde{\mathscr{F}}_1$. It is in this sense that we state that condition (2.2) is fulfilled with a large chance (generically). Intuitively, we may state that the noise in the data, although it respects the continuity of the measured function in our case, destroys the fine correlations implied by analyticity and boundedness.

If (2.2) is violated, i.e., $\chi^2_{\min}(\rho;h) = 0$, then, by the uniqueness of analytic continuation, the minimal value of $\chi^2(\rho; h-f)$ is attained on one function in $\mathscr{F}_1$, about which, of course, nothing more may be said.

If $\rho(x)$ is of finite type, let $\mathscr{H}_N$ be the set of continuous $g(x)$, $x \in [-a,b]$, such that $\chi^2_{\min}(g) = 0$. We shall argue that condition (2.2), i.e., $h(x) \notin \mathscr{H}_N$ is, in some sense, increasingly likely as N increases. We show, namely, the following (which is a more detailed form of the argument used in Ref. 17).

Let $\tilde{\rho}(x)$ have an infinite number of points of increase on $[-a,b]$, and let $\rho_N(x)$ be a sequence of approximants of finite type to $\rho(x)$, such that (a) the points of increase $x_1, x_2, \ldots, x_N$ of $\rho_N(x)$ are included in those of $\rho_{N+1}(x)$; (b) $\rho_N(x_i + 0) = \tilde{\rho}(x_i + 0)$, $i = 1,2,\ldots,N$; and (c) as $N \to \infty$,

$$\sup_{x \in [-a,b]} |\tilde{\rho}(x) - \rho_N(x)| \to 0. \tag{2.3}$$

Further, define, for any $f \in \mathscr{F}_1$,

$$r_{f,N} \equiv \sup\left\{r: \int_{-a}^{b} (g(x) - f(x))^2 \, d\rho_N(x) \leqslant r \Rightarrow g(x) \in \mathscr{H}_N\right\}. \tag{2.4}$$

Then, as $N \to \infty$, $r_{f,N} \to 0$, uniformly with respect to $f \in \mathscr{F}_1$.

To understand the meaning of this statement, consider the set

$$\mathscr{S}_N = \{w \in R^N,\ w = (w_1, w_2, \ldots, w_N):$$

$$f(x_i) = w_i,\ i = 1,\ldots,N, \text{ for some } f \in \mathscr{F}_1\}, \tag{2.5}$$

i.e., the set of possible values assumed at $x_1, x_2, \ldots, x_N$ by functions $f \in \mathscr{F}_1$. We may identify $\mathscr{S}_N$ with $\tilde{\mathscr{F}}_1$ if we use the isomorphism of $L^2(\rho_N)$ with R^N. Equation (2.4) is the radius of the largest ball contained in $\mathscr{S}_N$, and centered at the point $\{f(x_i)\}_{i=1}^N$.

To prove the statement, we shall show that, for any $\epsilon > 0$, we can find $N_0(\epsilon)$ such that, for $N \geqslant N_0(\epsilon)$ and for any $f \in \tilde{\mathscr{F}}_1$, the ball of radius ϵ centered at f contains points lying outside $\mathscr{S}_N$. Indeed, we choose first $N_1(\epsilon)$ so that the left-hand side of (2.3) is less than $\epsilon/16$ for $N \geqslant N_1(\epsilon)$. Assume then that $\rho_{N_1+1}(x)$ has a point of increase $\bar{x}$ between x_k and x_{k+1}. Let then $N_0(\epsilon) = N_1(\epsilon) + 1$ and define

$$\begin{aligned} g(x) &= f(x), \text{ for } x \leqslant x_k \text{ or } x \geqslant x_{k+1} \\ &= f(x) + (f(\bar{x}) + 3)(x - x_k)/(\bar{x} - x_k),\ x_k < x < \bar{x}, \\ &= f(x) + (f(\bar{x}) + 3)(x - x_{k+1})/(\bar{x} - x_{k+1}), \\ &\quad x < x < x_{k+1}. \end{aligned} \tag{2.6}$$

Clearly, by the maximum modulus theorem, $|f(\bar{x})| < 1$, so that $g(\bar{x}) > 1$ and, consequently, $g(x) \notin \mathscr{H}_N$, for all $N \geqslant N_0(\epsilon)$. On the other hand, for any $N \geqslant N_0(\epsilon)$,

$$\begin{aligned} &\int_{-a}^{b} (g(x) - f(x))^2 \, d\rho_N(x) \\ &\leqslant 16(\rho_N(x_{k+1} - 0) - \rho_N(x_k + 0)) \\ &\leqslant 16(\tilde{\rho}(x_{k+1} - 0) - \rho_{N_1}(x_k + 0)) \\ &\leqslant 16 \sup_{-a < x < b} (\tilde{\rho}(x) - \rho_{N_1}(x)) \leqslant \epsilon \end{aligned} \tag{2.7}$$

and this proves our point.

Thus, we can state that the body $\mathscr{S}_N$ gets increasingly flattened as N grows and, for high N, the effect of the noise is that the experimental values $\{h_i\}$ yield in general a point in R^N lying outside $\mathscr{S}_N$, so that condition (2.2) is fulfilled. This remark was made for the first time in Ref. 19 and, in the present setting, in Ref. 15; it was used in the work of Refs. 1, 16, and 20 and was further discussed in Ref. 21 and in Ref. 17.

One should mention that the improbable situation $\{h(x_i)\}_{i=1}^N \in \mathscr{S}_N$ allows, under special circumstances, an elegant treatment of problem (C) of the Introduction. As a consequence of Theorem 2.2 of Sec. VI, Ref. 22, if both points $\underline{h}, \bar{h} \in R^N$, with coordinates

$$\underline{h}_i = h(x_i) + (-1)^i \epsilon_i,\ \bar{h}_i = h(x_i) + (-1)^{i+1} \epsilon_i, \tag{2.8}$$

belong to $\mathscr{S}_N$, and $x_1 > x_2 > x_3 > \cdots > x_N$, then one may obtain exact upper and lower bounds for the values assumed at a point $x \neq x_i$ by all functions in $\mathscr{F}_1$ obeying

$$h(x_i) - \epsilon_i < f(x_i) < h(x_i) + \epsilon_i,\ i = 1,2,\ldots,N \tag{2.9}$$

($\epsilon_i > 0$). Namely, if $x < x_N$, the upper bound is obtained by performing a Pick–Nevanlinna interpolation (explained, e.g., in Ref. 15 and 23) of the points $\bar{h}_i$ and letting the final free function be equal to $+1$ if N is even and -1 if N is odd. The lower bound is obtained from $\underline{h}_i$ with the free function equal to ∓ 1 if N is even/odd. Similar statements are possible if x is situated in a different manner [staying in $(-1,1)$]. Apart from the statements of Ref. 22, the only

information one needs in order to derive these results is the positivity of the determinant

$$\Delta_N(x_i;\theta_k)\equiv\det|P(x_i;\theta_k)|_{i,k=1}^N , \tag{2.10}$$

where $P(x,\theta)$ is the Poisson kernel

$$P(x,\theta)=\frac{1}{2\pi}\frac{1-x^2}{1-2x\cos\theta+x^2}, \tag{2.11}$$

and $1>x_1>x_2>\cdots>x_N>-1$, $0<\theta_1<\theta_2<\cdots<\theta_N<\pi$. Imitating the method of Ref. 24, one obtains the recurrence relation

$$\Delta_N(x_i;\theta_k)=2^{N-1}\frac{1-x_N^2}{2\pi}\Delta_{N-1}(x_i;\theta_k)\frac{\Pi_{i=1}^N(\cos\theta_i-\cos\theta_N)\Pi_{i=1}^{N-1}(x_i-x_N)\Pi_{i=1}^N(1-x_ix_N)}{\Pi_{i=1}^N(1-2x_i\cos\theta_N+x_i^2)\Pi_{i=1}^{N-1}(1-2x_N\cos\theta_i+x_N^2)}, \quad \Delta_1\equiv P(x_1;\theta_1) , \tag{2.12}$$

which makes positivity explicit.[25]

In a recent paper (Ref. 26), under the assumptions that the point with coordinates $\{h(x_i)\}_{i=1}^N$ belongs to $\mathscr{S}_N$, and that both points $\bar{\epsilon},\underline{\epsilon}$, with coordinates $\{(-1)^i\epsilon_i\}_{i=1}^N$, $\{(-1)^{i+1}\epsilon_i\}_{i=1}^N$ in turn, also belong to $\mathscr{S}_N$, the author derives upper bounds for the departure of a linear extrapolation formula

$$f(x)=\sum_{i=1}^N C_i(x;x_i;\epsilon_i)h(x_i) , \tag{2.13}$$

where the coefficients C_i are specified functions of x,x_i,ϵ_i, from the set of values assumed at x by those functions of $\mathscr{F}_1$, which obey (2.9) (see Ref. 27).

Unfortunately, the arguments of Refs. 1,15–17 and 19–21 as well as the foregoing one show that the "chance" (understood as above) for the conditions of validity of these results [i.e., $(\bar{h},\underline{h})\in\mathscr{S}_N$ or $h\in\mathscr{S}_N$, $(\underline{\epsilon},\bar{\epsilon})\in\mathscr{S}_N$] to be met is vanishingly small, as soon as N assumes realistic values (e.g., $\gtrsim 5$ in the example of Ref. 17). It may be, however, of interest to understand the transition between these results and those obtained in Ref. 1, when the experimental point $\{h(x_i)\}_{i=1}^N$ lies outside $\mathscr{S}_N$. We give a discussion of this point in Appendix A.

Clearly, the arguments for the generality of the condition $\chi^2_{\min}>0$ are not restricted to the special type of norm (1.2) used to define $\mathscr{F}_1$. In fact, such a question arises in connection with the construction of a function $C(x)$, which reduces the errors ϵ_i of the data points h_i to a constant value, as proposed in Ref. 28. For two data sets at points $\{x_i\}_{i=1}^{N/2}$, $\{x_i'\}_{i=1}^{N/2}$, such that $x_{i-1}'<x_i<x_i'<x_{i+1}$, with errors ϵ and 2ϵ in turn, the minimal L^2 norm of $\mathrm{Re}\ln C(e^{i\theta})$ is a rather large number [of the order of 10^{14} for a total of 15 points distributed equidistantly on $(-0.5,0.5)$].

Before proceeding, we note that, if $\rho(x)$ has an infinite number of points of increase, then $\bar{\mathscr{F}}_1$ is convex and closed with respect to the distance (1.1). Indeed, if a sequence $f_n(x)\in\bar{\mathscr{F}}_1$ converges in the sense of (1.1) to $f_0(x)\in L^2(\rho)$, it contains a subsequence $f_{n_k}(x)$ that converges pointwise, a.e.-ρ to $f_0(x)$. By a form of the principle of uniform boundedness (due to Vitali), the sequence $f_{n_k}(x)$ converges even uniformly on compact subsets of D. The limit function $f_0(x)$ is thus real holomorphic in D and bounded on any compact subset of D by unity. Thus, it is in $\bar{\mathscr{F}}_1$.

If $\rho(x)$ has only a finite number of points of increase, then the set $\mathscr{S}_N\in R^N$, defined by Eq. (2.5), is convex and closed (see Ref. 1).

In order to have easy reference to results available in textbooks of analysis, we take the completion of the space of continuous functions with the form (1.1) and obtain thus the Hilbert space $L^2(\rho)$. Then, using, e.g., Theorem 4.10 of Ref. 29 (p. 83), we conclude that there exists a unique $f_0(x)\in\bar{\mathscr{F}}_1$ [a unique point $P_0\in\mathscr{S}_N$ if $P(x)$ is of finite type] on which $\chi^2_{\min}(\rho;h)$ is actually attained.

The characterization of the extremal elements of problem (A) is obtained by transforming it to a problem of maximization in the space dual to $L^2(\rho)$, which, by a well-known result (Ref. 29, p. 89, Theorem 4.17) may be identified with $L^2(\rho)$ itself. We shall use the notation $\|\cdot\|_\rho$ and $(\cdot,\cdot)_\rho$ for the norm and scalar products obtained from (1.1). The result we need is a corollary of the Hahn–Banach theorem (see Ref. 30, p. 58, Theorem 3.4 and Ref. 31, p. 136, Theorem 1) and states for our problem

$$\chi_{\min}(\rho;h)=\sup_{\|n\|_\rho\le 1}[(n,h)_\rho-\sup_{f\in\bar{\mathscr{F}}_1}(n,f)_\rho] , \tag{2.14}$$

where the supremum on the right is attained by some n_0 in $L^2(\rho)$. The essential element in (2.14) is the convexity of $\bar{\mathscr{F}}_1$. Further, we have seen that $\chi_{\min}(\rho;h)$ is realized by a unique $f_0\in\bar{\mathscr{F}}_1$; then, Eq. (2.14) implies

$$\chi_{\min}=\|h-f_0\|_\rho\le(n_0,h)_\rho-(n_0,f_0)_\rho\le\|h-f_0\|_\rho , \tag{2.15}$$

where we have used Schwarz's inequality in the last step. We conclude that

$$\sup_{f\in\bar{\mathscr{F}}_1}(n_0,f)_\rho=(n_0,f_0)_\rho \tag{2.16}$$

and

$$n_0(x)=(h(x)-f_0(x))/\|h-f_0\|_\rho . \tag{2.17}$$

This shows that the extremal $n_0(x)\in L^2(\rho)$ on the right-hand side of (2.14) is also unique.

We are now able to sketch the main direction of our argument (presented in Sec. IV) concerning the bound on the number of parameters of the solution of problem (A). In the rest of this section, we shall show that, independently of the special form of $\rho(x)$ and for any $n(x)$ in $L^2(\rho)$, the functions $f(n;x)$ realizing the supremum in (2.16) are uniquely determined in $\mathscr{F}_1$ and are finite Blaschke products. Consequently, so is $f_0(z)$, which realizes $\chi^2_{\min}$. In Sec. IV, we prove essentially that, if a sequence $\{n_k(x)\}_{k=1}^\infty$ is such that the corresponding extremals $\{f_0(x;n_k)\}$ have a number of zeros that increases indefinitely, then $n_k(x)$ tends weakly to zero. Consequently, if $\chi_{\min}(\rho;h)$ is realized on a function that has a too large number of zeros, Eq. (2.14)

implies that it may be unable to stay larger than some number χ_1, given in advance. This will set an upper bound on the number of zeros of $f_0(x)$, depending on $h(x)$ and χ_1, as announced. Now, following (to some extent) Ref. 2, we state the following theorem.

Theorem 2.1: Let $n(x)\in L^2(\rho)$. There exists a unique $f(n;z)$ in $\mathscr{F}_1$ such that $\sup\{(n,f)_\rho : f\in\mathscr{F}_1\}$ is attained on $f(n;x)$ and $f(n;z)$ is a finite Blaschke product.

Proof: We rewrite, using Cauchy's theorem

$$(n,f)_\rho = \left(n,\frac{1}{2\pi}\oint \frac{f(e^{i\theta})}{e^{i\theta}-x}e^{i\theta}\,d\theta\right)_\rho$$

$$= \frac{1}{2\pi}\oint k(n;\theta)f(\theta)e^{i\theta}\,d\theta\,, \tag{2.18}$$

with

$$k(n;\theta) = k(n;z=e^{i\theta}) = \int_{-a}^{b}\frac{n(x)}{e^{i\theta}-x}\,d\rho(x)\,. \tag{2.19}$$

Notice, $k(n;z)$ is holomorphic in the whole z plane, except for the line $[-a,b]$. The problem is to find the supremum of the right-hand side of (2.18), for fixed $k(n;\theta)$ among all $f(\theta)$, which are boundary values of $f\in\mathscr{F}_1$. (These are defined almost everywhere, see Ref. 32, p. 6, Theorem 1.3.)

We derive the existence and properties of the extremal $f(n;z)$ using again duality relations (see, e.g., Ref. 32, Chaps. 7 and 8 and Ref. 31, Sec. 5.8). To this end, we recall we can identify isometrically the space $L^\infty(T)$ of measurable complex functions absolutely bounded on the unit circle T: $|z|=1$ with the space of continuous linear functionals defined on $L^1(T)$ (Ref. 29, p. 136, Theorem 6.16). The action of such a functional, denoted by h, on $L^1(T)$ is then given by

$$\langle h,g\rangle = \oint h(\theta)g(\theta)d\theta\,, \tag{2.20}$$

with $h(\theta)\in L^\infty(T)$, $g(\theta)\in L^1(T)$.

Consider now the real subspace $L^\infty_R(T)$ of $L^\infty(T)$ consisting of functions $f(\theta)$ with the symmetry property (a.e. on $-\pi\leqslant\theta<\pi$)

$$f(\theta) = f^*(-\theta)\,. \tag{2.21}$$

One can verify that $L^\infty_R(T)$ is isomorphic (isometrically) with the real vector space of real continuous linear functionals defined on the subspace $L^1_R(T)$ of $L^1(T)$, of functions $g(\theta)$ obeying the symmetry (2.21). Further, it follows from Cauchy's theorem that the subspace $H^\infty_R(D)$ of $L^\infty_R(T)$ made up of functions $f(\theta)$, which admit of a bounded holomorphic extension to D: $|z|<1$, generates linear functionals that vanish on the subspace $H^1_R(D)$ of $L^1_R(T)$, of functions $g(z)$, holomorphic in D and such that $\oint|g(re^{i\theta})|d\theta$ is uniformly bounded in $r<1$. It can also be shown (e.g., using Ref. 32, Theorem 3.7, p. 40) that $H^1_R(D)$ represents in fact the whole annihilator of $H^\infty_R(D)$.

With this the following (duality) relation holds (Refs. 30 and 31):

$$\sup_{\substack{f\in H^\infty_R(D)\\ \|f\|_\infty<1}}\frac{1}{2\pi}\oint k(n;\theta)f(\theta)e^{i\theta}\,d\theta$$

$$= \inf_{g\in H^1_R(D)}\frac{1}{2\pi}\oint|k(n;\theta)-g(\theta)|d\theta\,. \tag{2.22}$$

The supremum on the left is achieved by some $f(n;z)\in H^\infty_R(D)$. Thus, the existence of the extremal element of Theorem 2.1 is guaranteed. Further, using the argument of Ref. 32, p. 132, one can show that, since $k(n;\theta)$ is continuous on T, the infimum on the right is also realized by a function $g(n;z)\in H^1_R(D)$. With this, we may obtain part of the statement of Theorem 2.1 from the conditions under which the following chain of inequalities is saturated:

$$\frac{1}{2\pi}\oint k(n;\theta)f(n;\theta)e^{i\theta}\,d\theta$$

$$= \frac{1}{2\pi}\oint(k(n;\theta)-g(n;\theta))f(n;\theta)e^{i\theta}\,d\theta$$

(a)

$$\leqslant\frac{1}{2\pi}\oint|k(n;\theta)-g(n;\theta)|\,|f(n;\theta)|d\theta\,,$$

(b)

$$\leqslant\frac{1}{2\pi}\oint|k(n;\theta)-g(n;\theta)|d\theta\,. \tag{2.23}$$

In step (b), equality can occur only if $|f(n;\theta)|=1$ almost everywhere on T, where $k(n;\theta)-g(n;\theta)\neq 0$. However, $k(n;\theta)-g(n;\theta)$ cannot vanish on a set of positive measure, without vanishing completely [Ref. 32, p. 17; $k(n;\theta)$ is holomorphic in a domain around T]. Thus, almost everywhere on $|z|=1$,

$$|f(n;\theta)| = 1\,. \tag{2.24}$$

We next show that $f(n;z)$ contains only a finite number of zeros in $|z|<1$. The condition for equality in step (a) of (2.23) is that the function

$$L(n;\theta) = (k(n;\theta)-g(n;\theta))f(n;\theta)e^{i\theta} \tag{2.25}$$

be of constant phase on the unit circle. From the symmetry condition (2.21), we conclude that, in fact, a.e. on $(-\pi,\pi)$,

$$L(n;\theta)\geqslant 0\,. \tag{2.26}$$

Equation (2.26) suggests an application of Schwarz's reflection principle to $L(n;z)$ across the unit circle, and thus the conclusion that $L(n;z)$ is in fact holomorphic in the whole z plane, except for the segments $[-a,b]\cup[1/b,\infty]\cup[-\infty,-1/a]$ $(a>0)$. A repetition of the reasoning on which Schwarz's principle is based (see, e.g., Ref. 33, p. 309) shows that the only requirement for its validity is that, for any arc (θ_1,θ_2),

$$\lim_{r\to 1}\int_{\theta_1}^{\theta_2}(rL(n;re^{i\theta})-L(n;e^{i\theta}))d\theta = 0\,. \tag{2.27}$$

This is, however, guaranteed by the fact that $g(n;z)$ belongs to $H^1_R(D)$ through Theorem 2.6 of Ref. 32 (p. 21). It follows that $L(n;z)$ and, therefore, $f(n;z)$ can vanish only a finite number of times inside a crown $1/r<|z|<r$, $1>1/r>\max[a,b]$. Further, $f(n;z)$ may vanish only a finite number of times in $|z|<1$ and may be written as $(|\alpha_i|<1)$

$$f(n;z) = \prod_{i=1}^{p}\frac{z-\alpha_i}{1-\alpha_i^* z}\psi(z)\,, \tag{2.28}$$

with $\psi(z)$ nonvanishing in $|z|<1$, $|\psi(e^{i\theta})|=1$ a.e. on $|z|=1$, $\psi\in H^\infty_R(D)$.

We now show that, in fact, $\psi(z)\equiv\pm 1$. To this end, we

write, for any $r<1$ and $|z|<r$, $z=|z|e^{i\psi}$,

$$\ln|\psi(z)| = \frac{1}{2\pi}\oint \frac{r^2-|z|^2}{r^2-2r|z|\cos(\theta-\psi)+|z|^2} \times \ln|\psi(re^{i\theta})|d\theta . \tag{2.29}$$

If we show that the positive quantities

$$I(r) = -\oint \ln|\psi(re^{i\theta})|d\theta \tag{2.30}$$

tend to zero as $r\to 1$, we conclude that $|\ln|\psi(z)|\,|$ is bounded by arbitrarily small quantities, for any z in $|z|<1$, and it follows that $|\psi(z)|\equiv 1$. To show that $I(r)\to 0$ as $r\to 1$, we notice first that a number $0<\alpha\leqslant\frac{1}{2}$ exists so that both $|L(n;e^{i\theta})|^{-2\alpha}$ and $|g(n;e^{i\theta})|^{\alpha}\,|L(n;e^{i\theta})|^{-\alpha}$ are integrable on $-\pi\leqslant\theta\leqslant\pi$. This follows from the fact that $L(n;e^{i\theta})$ has only zeros of finite multiplicity (isolated) on $|z|=1$ and from Schwarz's inequality

$$\oint |g(n;e^{i\theta})|^{\alpha}\,|L(n;e^{i\theta})|^{-\alpha}\,d\theta \leqslant \left[\oint |g(n;e^{i\theta})|^{2\alpha}\,d\theta\right]^{1/2}\left[\oint |L(n;e^{i\theta})|^{-2\alpha}d\theta\right]^{1/2}, \tag{2.31}$$

if we take into account that $|g(n;e^{i\theta})|\in L^1(T)$. Using the inequality

$$-\ln x \leqslant \frac{1}{\alpha}\frac{1-x^{\alpha}}{x^{\alpha}} \quad (0<x<1), \tag{2.32}$$

we write

$$\begin{aligned} I(r) &\leqslant \frac{1}{\alpha}\oint \frac{1-|\psi(re^{i\theta})|^{\alpha}}{|\psi(re^{i\theta})|^{\alpha}}d\theta \\ &\leqslant \frac{1}{\alpha}\oint (1-|\psi(re^{i\theta})|^{\alpha})\beta(e^{i\theta}) \\ &\quad + \frac{2}{\alpha}\oint (\beta(re^{i\theta})-\beta(e^{i\theta}))d\theta , \end{aligned} \tag{2.33}$$

with

$$\beta(e^{i\theta}) = (|g(n;e^{i\theta})|^{\alpha}+|k(n;e^{i\theta})|^{\alpha})/|L(n;e^{i\theta})|^{\alpha} . \tag{2.34}$$

In (2.33) we have used the inequality

$$(x+y)^{\alpha} < x^{\alpha}+y^{\alpha} , \tag{2.35}$$

valid for $x,y>0$, $0<\alpha<1$. The first term on the right-hand side of (2.33) goes to zero as $r\to 1$, by the dominated convergence theorem. The second term also vanishes as $r\to 1$, as a consequence of the fact that $g(n;z)$ belongs to $H^1(D)$ (see Ref. 32, Theorem 2.6, p. 21).

From this argument and knowing the holomorphy domain of $L(n;z)$ it follows that $g(n;z)$ is in fact holomorphic in the whole z plane, except for a cut along $[-\infty,-1/a]\cup[1/b,\infty]$.

The uniqueness of the extremal functions $f(n;\theta)$ and $g(n;\theta)$ is a consequence of the positivity condition (2.26). The latter is true for $L(n;\theta)$ constructed with any combination of an extremal $g(n;\theta)$ and an extremal $f(n;\theta)$. We pick then out a definite $g(n;\theta)$ and infer from (2.26) that all extremal $f(n;z)$ must have the same phase (mod 2π) along $|z|=1$. Since $|f(n;\theta)|=1$, this implies that they must be identical. However, if $f(n;\theta)$ is given, the combination $zg(n;z)f(n;z)$ [and thus $g(n;z)$] is completely determined from its boundary values by means of the (complexified) Poisson formula:

$$\operatorname{Im} k(n;\theta)f(n;\theta)e^{i\theta} = \operatorname{Im} g(n;\theta)f(n;\theta)e^{i\theta} . \tag{2.36}$$

This ends the proof of Theorem 2.1.

Returning now to problem (A), we see that, with the remarks surrounding Eq. (2.14), we have proved the following theorem.

Theorem 2.2: There exists a unique $f_0(z)\in\mathscr{F}_1$ on which $\chi_{\min}(\rho;h)$ is attained. The function $f_0(z)$ is a finite Blaschke product.

Notice, in both Theorems 2.1 and 2.2 we assert the uniqueness of the extremal element in $\mathscr{F}_1$, rather than $\bar{\mathscr{F}}_1$. This reproduces the results of Ref. 1 where, using Schur–Pick–Nevanlinna interpolation theory, the authors show that the unique point P_0 in $\mathscr{S}_N$ on which $\chi_{\min}(\rho;h)$ is attained is indeed generated by a unique Blaschke product. In the next section, we turn to the problem of actually constructing $f_0(z)$.

III. AN INTEGRAL EQUATION FOR PROBLEM (A)

The theorems of the preceding sections provide no means of computing the extremal $f_0(x)$. In fact, given $n(x)\in L^2(\rho)$, the argument of Theorem 2.1 does not show how to find the Blaschke product $f(n;x)$, which maximizes $(n,f)_{\rho}$ over $\mathscr{F}_1$. If such a method (expected to be nonlinear) were known, Eqs. (2.16) and (2.17) would provide an equation for the unique extremal function $\tilde{n}_0(x) = \|h-f_0\|_{\rho}n_0(x)$ associated to $f_0(x)$. Since $f(n;x)$ is unaffected by a change from $n(x)$ to $\lambda n(x)$ for any positive λ, this equation reads simply

$$\tilde{n}_0(x) = h(x) - f(\tilde{n}_0;x) . \tag{3.1}$$

Equation (3.1) has no other solutions in $L^2(\rho)$ apart from $n_0(x)$. Indeed, any $n_1(x)$ obeying (3.1) leads, when normalized, to equalities in (2.15) and is thus identical to $\tilde{n}_0(x)$.

Although it does not give the explicit dependence on $n(x)$ of $f(n;x)$, the argument of Theorem 2.1 does provide a characterization of $f(n;x)$, for given $n(x)$, by means of the special properties of the function $L(n;z)$, Eq. (2.25). To see how this is done, we recall the function $L(n;z)$ is made up of $n(x)$, the extremal $g(n;x)$ and the extremal $f(n;x)$, is positive along $|z|=1$ [Eq. (2.26)] and is holomorphic in the z plane cut along $(-\infty,-1/a]\cup[-a,b]\cup[1/b,\infty]$. The latter property is expressed more accurately through the following lemma.

Lemma 3.1: $L(n;z)$ obeys the representation

$$L(n;z) = 2\pi\int_{-a}^{b} n(x)f(n;x)P(x;z)d\rho(x) , \tag{3.2}$$

where $P(x;z)$ is the Poisson kernel [generalizing (2.11) to $|z|\neq 1$].

To prove this lemma, we denote by $\bar{L}(z)$ the right-hand side of (3.2) and verify first that it has the correct analyticity properties of $L(n;z)$. This is evident from the decomposition

$$\begin{aligned} -\frac{1}{x}P(x;z) &= -\frac{1}{2\pi x}\frac{1-x^2}{x^2-x(z+1/z)+1} \\ &= \frac{1}{2\pi}\left[\frac{1}{x-z}+\frac{1}{x-1/z}-\frac{1}{x}\right]. \end{aligned} \tag{3.3}$$

Further, $\bar{L}(z)$ is clearly real on the unit circle. As a consequence, the difference

$$\Delta(z) \equiv L(n;z) - \bar{L}(z) \tag{3.4}$$

obeys the same reflection symmetry as $L(n;z)$ and $\bar{L}(z)$:

$$\Delta(z) = \Delta^*(1/z^*) . \tag{3.5}$$

Now, using Eq. (2.25) and the expression (2.19) for $k(n;z)$ we verify easily that, in fact, $\Delta(z)$ is holomorphic in the unit disk. It follows then from (3.5) that it is everywhere holomorphic and thus a constant. To show that the constant vanishes, we compute $\lim_{z\to 0} \Delta(z)$. On one hand, from the definition of $L(n;z)$, Eq. (2.25), we see that the possible nonzero term of $L(n;z)$ near $z=0$ comes from $zk(n;z)f(n;z)$. On the other hand, from (3.2), we verify that, near $z=0$,

$$\bar{L}(z) = z\int_{-a}^{b} \frac{n(x)f(n;x)}{z-x}\,d\rho(x) + O(z) = zk(z)f(n;z) + O(z) . \tag{3.6}$$

Thus, $\Delta(z)\to 0$ as $z\to 0$, and this proves Lemma 1.

Notice that the right-hand side of Eq. (3.2) contains only $n(x)$ and $f(n;x)$, whereas $g(n;x)$ has dropped out. We may now imagine that we are given a function $n(x)$ and a Blaschke product $B(x;\alpha)$, with zeros at the points α_i, $i=1,2,...,p$, and that we construct with them a function $\tilde{L}(z;n;B)$ by means of the right-hand side of Eq. (3.2) [with $f(n;x)$ replaced by $B(x;\alpha)$]. The question is to decide whether $\tilde{L}(z;n;B)$ is indeed the function $L(n;z)$ associated to $n(x)$ [and the extremal $f(n;x)$]. From Eq. (2.25) we expect this to be so only if

$$\tilde{L}(z=\alpha_i;n;B) = 0, \quad i=1,2,...,p , \tag{3.7}$$

whereas from Eq. (2.26) we obtain the (necessary) condition

$$\tilde{L}(z=e^{i\theta};n;B) \geqslant 0, \quad 0\leqslant\theta<2\pi . \tag{3.8}$$

A more precise inspection shows, however, that Eq. (3.7) is not necessary if one of the α_i's happens to coincide with a point of discontinuity of $\rho(x)$. Nevertheless, it turns out that conditions that are both necessary and sufficient and are related to (3.7) and (3.8) may be formulated, essentially from geometrical considerations. For simplicity, we shall confine ourselves in the following to the situation when the extremal Blaschke product has no multiple factors. The changes that have to be made to allow for this are described in Appendix E. With this, we formulate the following lemma.

Lemma 3.2: The Blaschke product $B(x;\alpha)$ is the extremal function $f(n;x)$ associated to $n(x)$ by Theorem 2.1 if and only if both

$$(n(x),B(x;\alpha)P(x;\alpha_i)/\alpha_i)_\rho = 0, \quad i=1,2,...,p , \tag{3.9}$$

and

$$(n(x),B(x;\alpha)P(x;e^{i\theta}))_\rho \geqslant 0, \quad 0\leqslant\theta<2\pi . \tag{3.10}$$

Notice, Eq. (3.10) is, by (3.2), identical with (3.8); however, (3.9) reduces to (3.7) if $\alpha\notin[-a,b]$, but not necessarily otherwise.

Proof: We show first the "only if" part. Let $B(x;\alpha)$ be the extremal Blaschke product associated to $n(x)$; then it is true that, for any change $\Delta\alpha_i$ of the positions of the zeros,

$$(n(x),B(x;\alpha_i) - B(x;\alpha_i+\Delta\alpha_i))_\rho \geqslant 0 . \tag{3.11}$$

Consider then first only a change $\Delta\alpha_k = \lambda e^{i\varphi}$, $\Delta\alpha_{k+1} = \lambda e^{-i\varphi} = (\Delta\alpha_k)^*$ in the positions of a pair of complex conjugate zeros α_k, $\alpha_{k+1} = \alpha_k^*$, with a fixed angle φ. We may apply Taylor's first-order formula with respect to λ to write, using (3.11),

$$\left(n(x),\frac{\partial B}{\partial\alpha_k}(x;\tilde{\alpha}_k(x))\,\lambda e^{i\varphi} + \frac{\partial B}{\partial\alpha_k^*}(x;\tilde{\alpha}_k^*(x))\,\lambda e^{i\varphi}\right)_\rho \leqslant 0 , \tag{3.12}$$

with $0\leqslant(\tilde{\alpha}_k(x)-\alpha_k)/\lambda e^{i\varphi}\leqslant 1$. In writing (3.12) we have allowed formally a Blaschke factor $(x-\alpha)/(1-\alpha^* x)$ to depend upon two variables, α and α^*. We divide now Eq. (3.12) by λ and let λ tend to zero. Using the dominated convergence theorem, we obtain the condition that, for all φ,

$$\mathrm{Re}\left[\left(n(x),\frac{\partial B}{\partial\alpha_k}(x;\alpha)\right)_\rho e^{i\varphi}\right] \leqslant 0 . \tag{3.13}$$

Letting in turn $\varphi=0$ and $\varphi=\pi$ in Eq. (3.13), and then $\varphi=\pi/2$ and $\varphi=-\pi/2$, we see that (3.13) implies that, as a complex quantity

$$\left(n(x),\frac{\partial B}{\partial\alpha_k}(x;\alpha)\right)_\rho = 0 . \tag{3.14}$$

A simple calculation verifies that

$$\frac{\partial B}{\partial\alpha_k}(x;\alpha) = \frac{2\pi B(x;\alpha)P(x;\alpha_k)}{\alpha_k} , \tag{3.15}$$

so that we obtain Eq. (3.9) for α_i complex.

If α_i is a real zero, we need only consider real variations $\Delta\alpha_i$, repeat the reasoning above, and allow $\Delta\alpha_i$ to have both signs. Taking limits $\Delta\alpha_i\to 0$ with $\Delta\alpha_i>0$ and $\Delta\alpha_i<0$, we obtain immediately (3.9) also for real α_i. Notice the expression $P(x;\alpha)/\alpha$ is well behaved even if $\alpha=0$, except at $x=0$, where it diverges like $1/x$. However, in this case, $B(x;\alpha)$ contains a factor x, so that (3.9) is always meaningful.

Finally, consider the variations of $B(x;\alpha)$ obtained by the addition of two complex conjugate zeros $\alpha_{p+1},\alpha_{p+2}$ near the points $z=\exp(\pm i\theta)$. Let $\Delta\alpha_{p+1} = -\lambda e^{i\theta}$, $\Delta\alpha_{p+2} = -\lambda e^{-i\theta}$, $\lambda>0$. We obtain an equation analogous to (3.13), however, with an opposite sign. We can now let λ tend to zero only with a positive sign, so that we can only conclude that [using Eq. (3.15)],

$$\mathrm{Re}(n(x),B(x;\alpha)P(x;e^{i\theta}))_\rho \geqslant 0 . \tag{3.16}$$

The brackets in Eq. (3.16) are manifestly real [$P(x;e^{i\theta})$ is real] so that we have obtained (3.10). Thus, conditions (3.9) and (3.10) are indeed necessary.

To show that they are sufficient, we write (z complex)

$$\tilde{L}(z;n;B) = (n(x),B(x;\alpha)P(x;z))_\rho \tag{3.17}$$

and construct, by analogy with (2.25),

$$\bar{g}(z) = k(n;z) - \tilde{L}(z;n;B)/zB(z;\alpha) . \tag{3.18}$$

We claim that $\bar{g}(z)$ is holomorphic in the whole z plane minus the intervals $(-\infty,-1/a)\cup(1/b,\infty)$. It is in fact enough to show that it has no singularities on $[-a,b]$, since Eqs. (3.9) show that it has no singularities if z lies outside $[-a,b]$, $|z|<1$. Using Eq. (3.3) and (3.9), we see further that it is enough to show that the function

$$\Phi(z) = (n(x),\psi(z;x))_\rho \frac{1}{B(z;\alpha)} \equiv \frac{\gamma(z)}{B(z;\alpha)}, \tag{3.19}$$

with

$$\psi(z;x) = \frac{B(z;\alpha) - B(x;\alpha)}{z-x} - B(x;\alpha)\frac{x}{1-xz}, \tag{3.20}$$

is holomorphic at all points of $[-a,b]$. It follows from (3.19) and (3.20) that $\Phi(z)$ has in fact at most poles on $[-a,b]$ at the possible zeros of $B(z;\alpha)$. Let x_0 be such a zero; we compute the residuum at such a point (apart from a factor) by letting z approach x_0 in $\gamma(z)$, Eq. (3.19). If $x \neq x_0$, we obtain, using $B(x_0;\alpha) = 0$,

$$\psi(x_0;x) = -2\pi B(x;\alpha)P(x;x_0)/x_0. \tag{3.21}$$

If $x = x_0$, the last term in (3.20) vanishes, but the first term gives [using (3.15)]

$$\psi(x_0;x_0) = -2\pi \lim_{x\to x_0} B(x;\alpha)P(x;x_0)/x_0. \tag{3.22}$$

Then (using the dominated convergence theorem),

$$\gamma(x_0) = -2\pi(n(x),B(x;\alpha)P(x;x_0)/x_0)_\rho \tag{3.23}$$

and it vanishes, by virtue of (3.9). Thus, $g(z)$ is indeed holomorphic at all points of $[-a,b]$ and thus in the z plane minus $(-\infty,-1/a]\cup[1/b,\infty)$. In particular, $\bar{g}(z)\in H_R^1(D)$. Notice, we could have used the analyticity of $g(n;z)$ in $|z|<1$ to show directly, using Eq. (3.18), that conditions (3.9) are necessary. Our proof has the advantage of giving some meaning to the function $L(n;z)$.

We shall now show that, choosing $f(\theta) = B(e^{i\theta};\alpha)$ and $g(\theta) = \bar{g}(e^{i\theta})$ in Eq. (2.22), we obtain equality of the quantities under the sup and inf signs on the left- and right-hand sides; this will imply that $B(z;\alpha)$ and $\bar{g}(z)$ are indeed extremal functions associated to $k(n;z)$ and, thus, to $n(x)$. Using Eq. (3.18), (3.17), and (3.10) in turn, we verify

$$\begin{aligned}
&\oint k(n;\theta)B(e^{i\theta};\alpha)e^{i\theta}\,d\theta \\
&= \oint (k(n;\theta) - \bar{g}(e^{i\theta}))B(e^{i\theta};\alpha)e^{i\theta}\,d\theta \\
&= \oint \tilde{L}(e^{i\theta};n;B)d\theta \\
&= \oint |k(n;\theta) - \bar{g}(e^{i\theta})|d\theta.
\end{aligned} \tag{3.24}$$

This ends the proof of Lemma 3.2.

It may appear that the characterization of the extremal Blaschke product by means of Eqs. (3.9) and (3.10) falls short of providing the explicit dependence $f(n;x)$ needed in Eq. (3.1) in order to compute the extremal function $f_0(x)$ of problem (A). However, we shall now show that, by solving a problem analogous to the one of Theorem 2.1 in the Hardy space $H_R^2(D)$, we are able to obtain a large supply of pairs $(n(x),B(x;\alpha))$, which satisfy identically Eqs. (3.9) and (3.10).

Thus, we consider the problem of determining, for a given $N(x)\in L^2(\rho)$,

$$\|k_N\|_2 \equiv \sup\{(N,f)_\rho : f\in H_R^2(D),\ \|f\|_2^2 \leqslant 1\}, \tag{3.25}$$

and of describing those functions on which $\|k_N\|_2$ is attained. The solution is a straightforward generalization of the expansion in reproducing kernels of Refs. 12, 19, 34, and 35 and is expressed in the following lemma.

Lemma 3.3: The function $F(N;z)$ on which $\|k_N\|_2$ is attained may be written as

$$F(N;z) = \text{const} \int_{-a}^{b} \frac{N(x)d\rho(x)}{1-xz}. \tag{3.26}$$

The (positive) constant is determined from the condition $\|F(N;z)\|_2^2 = 1\ (= 1/\|k_N\|_2)$.

Proof: We write as in (2.18) [explaining also the notation $\|k_N\|_2$ in Eq. (3.25)]

$$\begin{aligned}
(N,f)_\rho &= \frac{1}{2\pi}\oint k_N(e^{i\theta})f(\theta)e^{i\theta}\,d\theta \\
&= \frac{1}{2\pi}\oint (\tilde{k}_N(z))^* f(z)z|dz|,
\end{aligned} \tag{3.27}$$

where we have denoted by $\tilde{k}_N(z)$ the function equal to $(k_N(z))^*$ on $|z| = 1$ and holomorphic in $|z|<1$,

$$\tilde{k}_N(z) = z\int_{-a}^{b} \frac{N(x)d\rho(x)}{1-xz}. \tag{3.28}$$

The last expression in (3.27) is the scalar product of $\tilde{k}_N(z)$ with $zf(z)$ in the Hilbert space $H_R^2(D)$. Its maximum is attained if $f(z)$ is proportional to $\tilde{k}_N(z)/z$. This proves Lemma 3.3. The analyticity domain of $F(N;z)$ contains the z plane without the interval $(-\infty,-1/a)\cup(1/b,\infty)$.

We next prove identities similar to Eqs. (3.9) and (3.10) in the following lemma.

Lemma 3.4: If α_i, $i = 1,2,...,p,p+1,...,q$ are the zeros of $F(N;z)$ in the z plane minus $(-\infty,-1/a]\cup[1/b,\infty)$, then

$$(N(x),F(N;x)P(x;\alpha_i)/\alpha_i)_\rho = 0,\ i = 1,2,...,q. \tag{3.29}$$

Also, for $0\leqslant\theta\leqslant 2\pi$,

$$(N(x),F(N;x)P(x;e^{i\theta}))_\rho \geqslant 0. \tag{3.30}$$

Proof: We write the canonical factorization (Ref. 32, p. 24, Theorem 2.8)

$$F(N;x) = B(N;x;\alpha)E(N;x), \tag{3.31}$$

where $B(N;x;\alpha)$ is a finite Blaschke product containing the zeros α_i, $i = 1,2,...,p$ of $F(N;x)$, lying in $|z|<1$ and $E(N;x)$ is the corresponding outer function [cf. Eq. (1.10)]. [Since $F(N;z)$ is holomorphic in $|z|<r$, $r>1$, the singular part of the factorization (3.31) is absent.] We may now repeat the proof of Lemma 3.2 and consider variations of $F(N;x)$ obtained by displacing the zeros α_i, $i = 1,2,...,p$, i.e., replacing $B(N;x;\alpha)$ by $B(N;x;\alpha+\Delta\alpha)$. This leaves the condition $\|f\|_2^2 \leqslant 1$ unchanged. With the argument of Lemma 3.2, we obtain Eq. (3.29) for the p zeros lying in $|z|<1$. Further, the same reasoning as in Lemma 3.1 shows that the function

$$\mathcal{L}(N;z) \equiv F(N;z)k_N(z) = F(N;z)(F(N;1/z^*))^*, \tag{3.32}$$

which is clearly holomorphic in the z plane minus $(-\infty,-1/a)\cup(-a,b)\cup(1/b,\infty)$, obeys the representation

$$\mathcal{L}(N;z) = 2\pi(N(x),F(N;x)P(x;z))_\rho/\|k_N\|_2. \tag{3.33}$$

The positivity of $\mathcal{L}(N;z)$ on $|z| = 1$ [manifest in (3.32)] leads to Eq. (3.30). Also, from Eq. (3.32), we see that, if a zero $\alpha_i \notin [-a,b]$, then $\mathcal{L}(N;\alpha_i) = 0$, which is Eq. (3.29). This ends the proof of Lemma 3.4.

We now notice that we may rewrite Eq. (3.29) by means of the factorization (3.31) as

$$(N(x)E(N;x),B(N;x;)P(x;\alpha_i)/\alpha_i)_\rho = 0,\ i = 1,2,...,q. \tag{3.34}$$

Since the same can be done with Eq. (3.30), a comparison with Eqs. (3.9) and (3.10) yields by Lemma 3.2 the following statement, which is essential for the construction of the extremal $f_0(x)$ of problem (A).

Theorem 3.1: For any $n(x)\in L^2(\rho)$, the Blaschke product $B(N;x;\alpha)$ of Eq. (3.31) realizes $\sup\{(n,f)_\rho : f\in H_R^\infty(D), \|f\|_\infty \leqslant 1\}$, if

$$n(x) = N(x)E(N;x)\|k_N\|_2 \equiv N(x)E_0(N;x). \tag{3.35}$$

In Eq. (3.35), we have introduced a factor $\|k_N\|_2$, to remove the normalization constant in Eq.(3.26). We are clearly free to do this, since Eqs. (3.9) and (3.10) and Eqs. (3.29 and 3.30) are invariant if $n(x)$, or $N(x)$ are multiplied by positive, x-independent, quantities. Therefore, in Eq. (3.35) $E_0(N;x)$ is the outer function obtained from

$$F_0(N;z) \equiv \int_{-a}^{b} \frac{N(x)d\rho(x)}{1-xz} \equiv B(N;z;\alpha)E_0(N;z). \tag{3.36}$$

We may now regard Eq. (3.35) as a nonlinear mapping $\mathscr{E}: L^2(\rho)\to L^2(\rho)$ (the left-hand side is always in $L^2(\rho)$ since $F_0(N;x)$ is continuous on $[-a,b]$). Theorem 3.1 may then be understood as follows: suppose $n(x)$ is given and we are required to find the associated extremal $f(n;x)$ in the set $\|f\|_\infty \leqslant 1, f\in H_R^\infty(D)$; we may regard then Eq. (3.35) as a nonlinear integral equation for the function $N(x)$. If we can solve this equation (compute $\mathscr{E}^{-1}n$), the solution to the original question is given by $B(N;x;\alpha)$. Let us show that this is, in principle at least, always possible.

Lemma 3.5: The mapping $\mathscr{E}: L^2(\rho)\to L^2(\rho)$ given by Eq. (3.35) is onto and has an inverse $\mathscr{E}^{-1}$.

Proof: We show first that $\mathscr{E}$ is one-to-one on its domain of values. Assume to this end that two functions $N_1(x)$ and $N_2(x)$ exist, both satisfying (3.35) for a given $n(x)$. Since there exists a unique Blaschke product $B(n;x)$ realizing the extremum of $(n,f)_\rho$ over $\|f\|_\infty \leqslant 1$, the two functions

$$F_0(N_i;x) = \int_{-a}^{b} \frac{N_i(x)}{1-xz}\,d\rho(x),\quad i = 1,2, \tag{3.37}$$

must lead to the same Blaschke factor $B(n;x)$. Thus

$$F_0(N_i;x) = B(n;x)E_{i;0}(x),\quad i = 1,2, \tag{3.38}$$

and we have to show that the two functions $E_{i;0}(x)$ are in fact identical. It is enough to show that they have the same modulus on $|z| = 1$. But

$$|E_{i;0}(e^{i\theta})|^2 = F_0(N_i;z)F_0^*(N_i;1/z^*)|_{z=e^{i\theta}} = \mathscr{L}_0(N_i;z)|_{z=e^{i\theta}} \tag{3.39}$$

with $\mathscr{L}_0(N_i;z) \equiv \|k_N\|_2^2 \mathscr{L}(N_i;z)$, Eq. (3.32). However, by the representation (3.33), both functions $\mathscr{L}_0(N_i;z)$ may be written as

$$\mathscr{L}_0(N_i;z) = 2\pi(N_i(x),B(n;x)E_{i;0}(x)P(x;z))_\rho = 2\pi(n(x),B(n;x)P(x;z))_\rho = L(n;z), \tag{3.40}$$

with $L(n;z)$ the function associated to $n(x)$ by (2.25), and use was made of Lemma 3.1. Therefore, the two functions $\mathscr{L}_0(N_i;z)$ are identical and it follows that $F_0(N_1;z) = F_0(N_2;z)$, i.e., $N_1(x) = N_2(x)$ (a.e.-ρ).

To show that $\mathscr{E}$ is onto, we pick out an $n\in L^2(\rho)$ and let $B(n;x)$ be its associated Blaschke product by Theorem 2.1. We define then, using property (2.26) of $L(n;e^{i\theta})$

$$e(\theta) \equiv [L(n;e^{i\theta})]^{1/2} \tag{3.41}$$

[we take $e(\theta)\geqslant 0$] and construct the function $\tilde{E}(z)$, holomorphic and free of zeros in $|z| < 1$, with modulus $e(\theta)$ on $|z| = 1$. Clearly, for $|z| = 1$,

$$L(n;z) = \tilde{E}(z)\tilde{E}^*(1/z^*), \tag{3.42}$$

where $\tilde{E}^*(1/z^*)$ is holomorphic and nonvanishing in $|z| > 1$. Since $L(n;z)$ is holomorphic in the z plane minus $(-\infty,-1/a]\cup[-a,b]\cup[1/b,\infty)$, Eq. (3.42) can serve to extend $\tilde{E}(z)$ to a holomorphic function in the z plane minus $(-\infty,-1/a]\cup[1/b,\infty)$; namely, we declare for $|z| > 1$,

$$\tilde{E}(z) = L(n;z)/\tilde{E}^*(1/z^*). \tag{3.43}$$

Notice that $\tilde{E}(z)$ vanishes at all points in $|z| > 1$, where $L(n;z)$ vanishes. A reasoning similar to that of Lemma 3.1 and use of Eq. (3.2) show that $\tilde{E}(z)$ may be represented as

$$\tilde{E}(z) = \int_{-a}^{b} \frac{n(x)B(n;x)}{\tilde{E}(x)} \frac{1}{1-xz}\,d\rho(x). \tag{3.44}$$

Consider now the function

$$\tilde{F}(z) = B(n;z)\tilde{E}(z). \tag{3.45}$$

Use of Eq. (3.44), of the identity $B(n;z)B^*(n;1/z^*) = 1$, and again a reasoning similar to that of Lemma 3.1 show that, in fact, $\tilde{F}(z)$ is obtained through the formula

$$\tilde{F}(z) = \int_{-a}^{b} \frac{n(x)}{\tilde{E}(x)} \frac{1}{1-xz}\,d\rho(x). \tag{3.46}$$

In Eq. (3.46) we see that, denoting $N(x) = n(x)/\tilde{E}(x)$, it follows that $\tilde{F}(z) \equiv F_0(N;x)$ and, from (3.45), $\tilde{E}(z) = E_0(N;x)$. Thus, an $N(x)$ obeying (3.35) has been found. This ends the proof of Lemma 3.5.

Next follow some comments.

(a) If $B(n;x)$ is the extremal Blaschke product associated to $n(x)$, we see from Eq. (3.35) that it is in fact the extremal function for many other $n(x)$, namely, all those obtained by varying $N(x)$ in such a manner that the zeros of $F_0(N;x)$, Eq. (3.36), in $|z| < 1$, stay unchanged; only $E(N;x)$ varies. Thus, we may regard the Blaschke products as "corners" of the set $\mathscr{F}_1$ in $L^2(\rho)$.

(b) With the help of Eq. (3.35) we may now cast Eq. (3.1) in the form of a nonlinear integral equation for the unknown function $N_0(x)$,

$$n_0(x) = N_0(x)E_0(N_0;x) = h(x) - B(N_0;x). \tag{3.47}$$

This is a considerable improvement over (3.1). In fact, it is easy to see that the integral equation (3.47) has a unique solution. Indeed, we have seen that there exists a unique $n_0(x)$ satisfying (3.1) and the statement follows from Lemma 3.5. The problem is, of course, whether the solution of (3.47) is or is not an awful task. It turns out that the solution is not completely out of reach, in view of the following lemma.

Lemma 3.6: If $N(x)$ is such that $F_0(N;z)$ does not van-

ish anywhere on $|z|=1$, the Fréchet derivative of the nonlinear operator $\mathscr{E}: L^2(\rho)\to L^2(\rho)$ given by (3.35) exists, is continuous, and has a bounded inverse.

Thus, under the conditions of Lemma 3.6, the operator $\mathscr{E}(N)$ is a local diffeomorphism of $L^2(\rho)$; the inverse mapping $\mathscr{E}^{-1}(n)$ also has a continuous Fréchet derivative (see Ref. 36, p. 56, Theorem 4.2.1). Also, as a consequence of Lemma 3.6, we shall show in Sec. VII that we can give a systematic method of solving Eq. (3.47).

Proof: The fact that $\mathscr{E}$ has a continuous Fréchet derivative follows immediately from the known representation of $E_0(N;x)$:

$$E_0(N;x)=\exp\left[\frac{1}{4\pi}\oint\frac{e^{i\theta}+x}{e^{i\theta}-x}\ln F_0(N;e^{i\theta})F_0(N;e^{-i\theta})d\theta\right]. \tag{3.48}$$

Straightforward calculations show that the action of the Fréchet derivative $\partial\mathscr{E}/\partial N$ on a vector $\delta N(x)\in L^2(\rho)$ is given by

$$\left(\frac{\partial\mathscr{E}(N)}{\partial N}\right)(\delta N)(x) = \delta N(x)E_0(N;x)+N(x)E_0(N;x)\times\oint P(x;e^{i\theta})\left(\frac{\delta F_0(e^{i\theta})}{F_0(N;e^{i\theta})}+\frac{\delta F_0(e^{-i\theta})}{F_0(N;e^{-i\theta})}\right)d\theta, \tag{3.49}$$

with

$$\delta F_0(z)=\int_{-a}^{b}\frac{\delta N(x)d\rho(x)}{1-xz}\equiv\left(\frac{\partial F_0}{\partial N}\right)(\delta N)(z). \tag{3.50}$$

The integrand of Eq. (3.49) indicates why we needed the restrictions on $N(x)$ in the statement of the Lemma. The situation when $F_0(N;e^{i\theta})$ vanishes for some θ is discussed after the end of the proof. Equation (3.49) may be transformed by means of the residuum theorem to yield

$$\left(\frac{\partial\mathscr{E}(N)}{\partial N}\right)(\delta N)(x)=\delta N(x)E_0(N;x)+N(x)E_0(N;x)\times\left(\frac{\delta F_0(x)}{F_0(N;x)}-\frac{\delta B(x)}{B(N;x)}\right), \tag{3.51}$$

where

$$\delta B(x)=-2\pi B(N;x)\sum_{i=1}^{p}P(x;\alpha_i)\frac{1}{\alpha_i}\frac{\delta F_0(\alpha_i)}{F_0'(N;\alpha_i)}, \tag{3.52}$$

and α_i, $i=1,2,\ldots,p$ are the zeros of $B(N;z)$ in $|z|<1$.

For any given left-hand side $\delta n(x)$, Eq. (3.51) is a Fredholm equation of the second kind for $\delta N(x)$. It can be solved for any $\delta n(x)$ if the equation

$$\frac{\partial\mathscr{E}(N)}{\partial N}(\delta N)=0, \tag{3.53}$$

does not admit of nonzero solutions. To show this, we see that Eq. (3.2) defines [via (3.35)] a mapping from $L^2(\rho)$ into the set of holomorphic functions representable in the form $(\gamma(x),P(x;z))_\rho$ with $\gamma(x)\in L^2(\rho)$ and the norm of $L^2(\rho)$. The action of its Fréchet derivative on $\delta N(x)$ is given [using the notation (3.53)]

$$\delta L(z)\equiv\left(\frac{\partial L(N)}{\partial N}\right)(\delta N)(z) = (\delta n(x),B(x)P(x;z))_\rho+(n(x),\delta B(x)P(x;z))_\rho. \tag{3.54}$$

Now, if $\delta n(x)=0$ [from (3.53)] and $\delta N(x)$ is such that $\delta B(x)=0$ [notice that the image of the action of $(\partial B/\partial N)(N)$ on $L^2(\rho)$ is a p-dimensional subspace of $L^2(\rho)$], then $\delta L(z)=0$. Further, from Eq. (3.41) we see that

$$0=\delta L(N;e^{i\theta})=2|E_0(N;e^{i\theta})|\delta|E_0(N;e^{i\theta})|, \tag{3.55}$$

so that $\delta|E_0(N;e^{i\theta})|=0$. This leads, however, to ($x\in[-a,b]$)

$$\delta E_0(N;x)=0 \tag{3.56}$$

[from (3.48)]. Since $E_0(N;x)\neq 0$, it follows from

$$\delta n(x)=E_0(N;x)\delta N(x)+N(x)\delta E_0(N;x) \tag{3.57}$$

that $\delta N(x)=0$. Thus, if there are nontrivial solutions to (3.53), they are such that $\delta B(x)\neq 0$.

To handle this situation, we consider the mapping $L^2(\rho)\to 0\in R^p$ given by Eqs. (3.9). Writting (3.9) in the form (3.14), we conclude that, if $\delta n(x)=0$,

$$\delta(L(\alpha_i))=\left(n(x),\sum_{k=1}^{p}\frac{\partial^2 B}{\partial\alpha_i\partial\alpha_k}\delta\alpha_k\right)_\rho=0, \quad i=1,2,\ldots,p, \tag{3.58}$$

with

$$\delta\alpha_k=-\frac{\delta F_0(\alpha_k)}{F_0'(N;\alpha_k)}. \tag{3.59}$$

Multiplying equation "i" in (3.58) by $\delta\alpha_i$ and adding them together, we obtain that (3.53) implies the vanishing of the quadratic form

$$Q(\delta\alpha)\equiv\sum_{i,k}\left(n(x),\frac{\partial^2 B}{\partial\alpha_i\,\partial\alpha_k}\right)_\rho\delta\alpha_i\delta\alpha_k=0. \tag{3.60}$$

If there are nontrivial solutions $\delta N(x)$ to (3.53), then $\delta\alpha_i\neq 0$, for at least some i, and (3.60) shows that $Q(\delta\alpha)$ is not definite. However, we shall show that $Q(\delta\alpha)$ is in fact strictly negative definite, if $\delta B(x)\neq 0$. This shows that, in fact, we must have $\delta B(x)=0$, if $\delta n(x)=0$. This leads in turn to $\delta N(x)\equiv 0$ and proves Lemma 3.6.

To show that (3.60) is negative definite, we resort for simplicity to the function $F(N;z)$ of Lemmas 3.3 and 3.4, which realizes the extremum of $(N,f)_\rho$ under the condition $\|f\|_2\leqslant 1$. [$F(N;z)$ differs from $F_0(N;z)$ by a constant factor).] We notice first that, at a change

$$F(N;z)\to F(N;z)+\Delta F(N;z)\equiv B(z;\alpha+\Delta\alpha)E(N;z),$$

the functional $(N,f)_\rho$ decreases strictly. Indeed, using the fact that (from Lemma 3.3)

$$F(N;e^{i\theta})=(k_N(e^{i\theta}))^*/\|k_N\|_2, \tag{3.61}$$

we see that the constraint $\|F(N;z)+\Delta F(N;z)\|_2^2=1$ implies

$$2(k_N^*(e^{i\theta}),\Delta F)_2/\|k_N\|_2+\|\Delta F\|_2^2=0. \tag{3.62}$$

[The index 2 on $(\cdot,\cdot)$ shows the scalar product in $H_R^2(D)$ leading to the norm (1.5).] Then, as claimed,

$$(N,\Delta F)_\rho=(k_N^*,\Delta F)_2=-\|\Delta F\|_2^2\|k_N\|_2/2<0. \tag{3.63}$$

On the other hand, we write this change by means of a Taylor expansion as in Lemmas 3.2 and 3.4 and conclude that, using Eqs. (3.29)

$$\sum_{i,j}\left(N(x),E(N;x)\frac{\partial^2 B}{\partial\alpha_i\,\partial\alpha_j}(x;\alpha_i+\theta(x)\Delta\alpha_i)\right)_\rho \Delta\alpha_i\Delta\alpha_j$$
$$= -\|k_N\|_2\frac{1}{2\pi}\oint|E(N;e^{i\theta})|^2$$
$$\times\left|\sum_i\frac{\partial B}{\partial\alpha_i}(\theta;\alpha_i+\psi(\theta)\Delta\alpha_i)\Delta\alpha_i\right|^2<0, \tag{3.64}$$

with $0\leqslant\theta(x)\leqslant 1$, $0\leqslant\psi(\theta)\leqslant 1$. Letting $\Delta\alpha_i=\lambda\beta_i$, $\lambda\in R$, and allowing λ to approach zero, we obtain, for arbitrary directions β_i [and using Eq. (3.35)],

$$\sum_{i,k}\left(n(x),\frac{\partial^2 B}{\partial\alpha_i\,\partial\alpha_j}(x;\alpha_i)\right)_\rho\beta_i\beta_j$$
$$= -\|k_N\|_2^2 2\pi\oint|E(N;e^{i\theta})|^2$$
$$\times\left|\sum_{i=1}^{p}\beta_i\frac{P(e^{i\theta};\alpha_i)^2}{\alpha_i}\right|^2 d\theta. \tag{3.65}$$

The right-hand side of (3.65) is, however, strictly negative definite, if $\delta B(x)\neq 0$. Indeed, a comparison with (3.52) and (3.59) with $\delta\alpha_i=\beta_i$ shows that the last factor in (3.65) is the analytic continuation of $\delta B(z)$ to $|z|=1$, and it cannot vanish identically there. This ends the proof of Lemma 3.6.

We now discuss the situation when $F(N;z)$ vanishes on $|z|=1$. We restrict ourselves to elements $\tilde{N}(x)\in L^2(\rho)$ for which the following property holds.

(H) The function $F(\tilde{N};z)$ has either one simple zero or two complex conjugate zeros on $|z|=1$.

This restriction is made only partly for simplicity; it is discussed at the end of Appendix E. It is enough to study the case when $F(\tilde{N};\alpha_{p+1})=0, |\alpha_{p+1}|=1, \alpha_{p+2}=\alpha_{p+1}^*$, $\alpha_{p+2}\neq\alpha_{p+1}$; the situation of one real zero is analogous. We consider two types of approximants $\{N_k(x)\}_{k=1}^\infty$ to $N(x)$: the "exterior" ones, for which $F(N_k^e;x)$ vanishes at two points $\alpha_{p+1,k}^e$, $(\alpha_{p+1,k}^e)^*$ with $|\alpha_{p+1,k}^e|>1$, $\alpha_{p+1,k}^e\to\alpha_{p+1}$ as $k\to\infty$, and "interior" ones, $F(N_k^i;x)$ vanishing at $\alpha_{p+1,k}^i$, $(\alpha_{p+1,k}^i)^*$, $|\alpha_{p+1,k}^i|<1$. From Eq. (3.51) it is easy to see that

$$\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_e\equiv\lim_{k\to\infty}\frac{\partial\mathscr{E}}{\partial N}(N_k^e)$$
$$\neq\lim_{k\to\infty}\frac{\partial\mathscr{E}}{\partial N}(N_k^i)\equiv\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_i, \tag{3.66}$$

although both limits exist (strongly). In fact, Eq. (3.52) gives the difference of the limits [its action on $\delta N(x)\in L^2(\rho)$]

$$\left[\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_e-\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_i\right](\delta N(x))$$
$$= -2\pi N(x)E_0(N;x)\sum_{k=p+1}^{p+2}P(x;\alpha_k)\frac{\delta F_0(\alpha_k)}{\alpha_k F_0'(N;\alpha_k)}. \tag{3.67}$$

In Appendix B we shall show that, in fact, both $((\partial\mathscr{E}/\partial N)(\tilde{N}))_e$ and $((\partial\mathscr{E}/\partial N)(\tilde{N}))_i$ have bounded inverses.

At such a point $\tilde{N}(x)\in L^2(\rho)$, we may clearly define the Gateaux differentials $\delta\mathscr{E}(\tilde{N};\delta N)$ in all directions $\delta N(x)$; if $\delta N(x)$ is such that $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})<0$ [$\delta\alpha_{p+1}$ is given in Eq. (3.59)], then

$$\delta\mathscr{E}(\tilde{N};\delta N)=\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_i(\delta N), \tag{3.68}$$

whereas, for the other half-space $\delta\mathscr{E}(\tilde{N};\delta N)$ is obtained from $(\partial\mathscr{E}/\partial N)_e$. The two definitions coincide if $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})=0$, since the Poisson kernel is real for $|\alpha_k|=1$ and real analytic. This implies the following generalization of Lemma 3.6 to include the boundary points $\tilde{N}(x)$, which is of relevance for the algorithm described in Sec. VII and amusing by itself.

Lemma 3.7: If condition (H) is satisfied, the mapping $L^2(\rho)\to L^2(\rho)$ given by $\delta\mathscr{E}(N;\delta N)$ is onto and invertible.

Proof: The image of the hyperplane $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})=0$, which is the same under $((\partial\mathscr{E}/\partial N)(\tilde{N}))_i$ and $((\partial\mathscr{E}/\partial N)(\tilde{N}))_e$ is a hyperplane containing the origin, since, say, $((\partial\mathscr{E}/\partial N)(\tilde{N}))_i$ is invertible. There exists then an element $k(x)\in L^2(\rho)$, determined up to multiplication by a scalar, so that this hyperplane is given by

$$(k(x),\delta n(x))=0. \tag{3.69}$$

If we denote by $p(x)$ the vector determining the linear functional of δN given by $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})$ [obtainable from (3.59)], then the solutions $k_1(x)$, $k_2(x)$ of the two equations

$$\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_i^\dagger(k_1(x))=\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_e^\dagger(k_2(x))=p(x) \tag{3.70}$$

are parallel to $k(x)$ and differ among themselves by a scalar factor λ. Now, the sets of vectors $\delta N(x)$ given by (a) $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})<0$ and (b) $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})>0$ are mapped in turn by $((\partial\mathscr{E}/\partial N)(\tilde{N}))_i$, $((\partial\mathscr{E}/\partial N)(\tilde{N}))_e$ one-to-one onto the half-spaces of elements $\delta n(x)$ satisfying

$$(k_1(x),\delta n(x))_\rho<0,\quad (k_2(x),\delta n(x))_\rho>0. \tag{3.71}$$

The two half-spaces are distinct if $\lambda>0$. We show in Appendix B that this is indeed the case. This ends the proof of Lemma 3.7.

We are now able to formulate a statement concerning Eq. (3.47) that determines the extremal function $B(N_0;x)$ of problem (A). We denote by $\mathscr{A}$ the operator from $L^2(\rho)$ into $L^2(\rho)$ whose action is given by

$$\mathscr{A}(N(x))\equiv N(x)E_0(N;x)+B(N;x). \tag{3.72}$$

Then we obtain the following theorem.

Theorem 3.2: For any $h(x)$ outside $\mathscr{F}_1$, the equation $\mathscr{A}(N(x))=h(x)$ has a unique solution. The operator $\mathscr{A}$ has a continuous Fréchet derivative with a bounded inverse at all points $N(x)$ such that $F_0(N;z)$, Eq. (3.36), does not vanish on $|z|=1$. At those remaining points, for which (H) is true, the operator whose action on $\delta N(x)\in L^2(\rho)$ is given by the Gateaux differential $\delta\mathscr{A}(N;\delta N)$ also has a bounded inverse.

Proof: The first part of the theorem has been already proved in comment (b) following Lemma 3.5. The only statement to be settled in the second part is the existence of a bounded inverse of the Fréchet derivative of $\mathscr{A}(N(x))$ at points $N(x)$ such that $F_0(N;z)$, Eq. (3.36), does not vanish on $|z|=1$.

To this end, consider the action of $(\partial\mathscr{A}/\partial N)(N)$ on $\delta N(x)$,

$$\frac{\partial \mathscr{A}}{\partial N}(N)(\delta N(x)) = \delta n(x) + \delta B(x). \tag{3.73}$$

We shall show that a constant $c>0$ exists so that, for any $\delta N(x)$ with $\|\delta N(x)\|_\rho = 1$,

$$\left(\delta N(x), \left(\frac{\partial \mathscr{E}}{\partial N}\right)^\dagger \left(\frac{\partial \mathscr{A}}{\partial N}\right) \delta N(x)\right)_\rho \geqslant c. \tag{3.74}$$

Since, by Lemma 3.6, $((\partial \mathscr{E}/\partial N)(N))$ has a bounded inverse, Eq. (3.74) implies that $((\partial \mathscr{A}/\partial N)(N))$ also has one. Equation (3.74) is obtained from (3.73) by ρ-scalar multiplication with $\delta n(x)$. It is clearly enough to show that

$$(\delta n(x), \delta B(x))_\rho \geqslant 0. \tag{3.75}$$

This is, however, a consequence of the convexity of $\mathscr{F}_1$. Indeed, consider two functions $n(x), n(x)+\Delta n(x)$ in $L^2(\rho)$ [corresponding by (3.35) to $N(x)$, $N(x)+\Delta N(x)$] and their associated extremal Blaschke products in $\mathscr{F}_1$: $B(x)$, $(B+\Delta B)(x)$ Then, the convexity of $\mathscr{F}_1$ implies both

$$(n(x), B(x))_\rho \geqslant (n(x), B+\Delta B)(x))_\rho, \tag{3.76}$$

$$(n(x)+\Delta n(x), B(x))_\rho \leqslant (n(x)+\Delta n(x), (B+\Delta B)(x))_\rho. \tag{3.77}$$

We subtract them, use the definition of the Fréchet derivative, and divide through by $\|\Delta N(x)\|_\rho^2$; we obtain that, for any $\delta N(x) \in L^2(\rho)$, $\|\delta N(x)\|_\rho = 1$, $\delta N(x) \equiv \Delta N(x)/\|\Delta N(x)\|_\rho$:

$$\left(\frac{\partial n}{\partial N}(N)\delta N(x), \frac{\partial B}{\partial N}(N)\delta N(x)\right)_\rho + \left(\frac{\kappa_n(\Delta N)}{\|\Delta N\|_\rho}, \frac{\partial B}{\partial N}(N)\delta N\right)_\rho + \left(\frac{\partial n}{\partial N}\delta N, \frac{\kappa_B(\Delta N)}{\|\Delta N\|_\rho}\right)_\rho + \left(\frac{\kappa_n(\Delta N)}{\|\Delta N\|_\rho}, \frac{\kappa_B(\Delta N)}{\|\Delta N\|_\rho}\right)_\rho \geqslant 0, \tag{3.78}$$

with $\|\kappa_n(\Delta N)\|_\rho, \|\kappa_B(\Delta N)\|_\rho = o(\|\Delta N\|_\rho)$. By letting $\|\Delta N\|_\rho \to 0$, (3.75) follows.

To prove the last statement of the theorem, we have to show first that $((\partial \mathscr{A}/\partial N)(\tilde{N}))_i, ((\partial \mathscr{A}/\partial N)(\tilde{N}))_e$ are invertible operators. To this end, we consider again sequences $N_k^i, N_k^e \to \tilde{N}$, and notice that, for any fixed $\delta N(x) \in L^2(\rho)$, and for all N_k^i, N_k^e, the quadratic form in $\delta N(x)$ given by (3.75) is positive. Consequently, this will be so also for $k \to \infty$. With the reasoning above, we conclude then that $((\partial \mathscr{A}/\partial N)(\tilde{N}))_i, ((\partial \mathscr{A}/\partial N)(\tilde{N}))_e$ are invertible. Further, the latter operators coincide in $L^2(\rho)$ on the hyperplane $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1}) = 0$ (if $\alpha_{p+1} \neq \alpha_{p+1}^*$), as one explicitly verifies. Consequently, the statement of Lemma 3.7 holds also for $\mathscr{A}(\tilde{N})$ [see also comment (b) of Appendix B] and this proves the last part of Theorem 3.2.

In Sec. VII, we shall use Theorem 3.2 to set up an algorithm for finding the solution $N_0(x)$ of Eq. (3.49) and thus the extremal element of problem (A).

IV. BOUNDS ON THE NUMBER OF ZEROS OF THE SOLUTION OF PROBLEM (A)

In the previous section, we have shown how the extremal element of problem (A) may be obtained by solving a certain nonlinear integral equation. The equation has some properties that make it amenable to numerical treatment. However, the methods that have been used so far fail to show how we may estimate, or find an upper bound on the number of zeros of the extremal Blaschke product $f_0(x)$ [$\equiv B(N_0;x)$ of Eq. (3.49)], in an *a priori* manner. They do give a hint, however; the analyticity of the function $F_0(N;z)$, Eq. (3.35), in a domain that strictly includes the unit disk makes it intuitively difficult for too many of its zeros to accumulate in $|z|<1$. In this section, we make this argument precise and derive an upper bound on the number of zeros of $f_0(x)$ as a functionof $h(x)$ and of a lower bound χ_1^2 to $\chi_{\min}^2(\rho;h)$. The main idea of the argument is sketched in Sec. II, following Eq. (2.17).

We shall use for our purposes a function related to $L(n;z)$ of Eq. (2.25), namely,

$$R(n;z) \equiv L(n;z)/(zf(n;z)) \equiv k(n;z) - g(n;z). \tag{4.1}$$

One can show easily that $R(n;z)$ satisfies the representation

$$R(n;z) = \int_{-a}^{b} \frac{n(x)d\rho(x)}{z-x} + \int_{-a}^{b} \frac{xn(x)(f(n;x))^2\, d\rho(x)}{1-xz}, \tag{4.2}$$

which is analogous to that of Lemma 2.1 for $L(n;z)$. As a consequence of Eq. (2.26), one notices that the total variation of the phase of the function $R(n;z)$ along the unit circle is $-2\pi(p+1)$, where p is the number of factors of the Blaschke product $f(n;x)$. [The counting is done by setting equal to zero the phase variation at those, possibly existing, points on $|z|=1$ where $L(n;z)$, and thus $R(n;z)$ have double zeros.]

Therefore, each of the functions $R_R(n;\theta) \equiv \mathrm{Re}\, R(n;e^{i\theta})$, $R_I(n;\theta) \equiv \mathrm{Im}\, R(n;e^{i\theta})$ must vanish $2(p+1)$ times at least on $0 \leqslant \theta \leqslant 2\pi$. These functions read

$$R_R(n;\theta) = \int_{-a}^{b} \frac{n(x)[\cos\theta(1-x^2f(n;x)^2) - x(1-f(n;x)^2)]}{x^2 - 2x\cos\theta + 1} \times d\rho(x), \tag{4.3}$$

$$R_I(n;\theta) = \int_{-a}^{b} \frac{n(x)\sin\theta(1-x^2f(n;x)^2)}{x^2-2x\cos\theta+1}\, d\rho(x). \tag{4.4}$$

Instead of $R_I(n;\theta)$, it is of more interest to consider

$$\bar{R}_I(n;\theta) = R_I(n;\theta)/\sin\theta, \tag{4.5}$$

which vanishes at least $2p$ times in $0 \leqslant \theta \leqslant 2\pi$. We now prove the following lemma.

Lemma 4.1: There exist constants $C_R>0$, $C_I>0$, $0<\gamma<1$, independent of $n(x)$, such that, if $f(n;z)$ consists of p Blaschke factors,

$$|R_R(n;\theta)| < C_R\gamma^p, \quad |\bar{R}_I(n;\theta)| < C_I\gamma^p. \tag{4.6}$$

Proof: We consider $R_R(n;\theta)$ only, since the inequality for $\bar{R}_I(n;\theta)$ is obtained in a clearly similar manner. Let $R_R(n;z)$ be the analytic extension of $R_R(n;\theta)$ to the complex z plane, through $\cos\theta = (z+1/z)/2$. It is a holomorphic function in the whole z plane except for the "cuts" $(-\infty, -1/a] \cup [-a,b] \cup [1/b, \infty)$. Consider now a simple closed curve $\mathscr{C}$ surrounding $[-a,b]$ and strictly contained in $|z|<1$ and its image $\mathscr{C}'$, obtained by reflection across the unit circle (thus lying strictly in $|z|>1$). We can bound

$|R_R(n;z)|$ along $\mathscr{C}$, independently of $n(x)$ [and of $f(n;x)$], using the Schwarz inequality, the condition $\|n\|_\rho \leqslant 1$, and the fact that $|f(n;x)| \leqslant 1$, for $x \in [-a,b]$:

$$|R_R(n;z| < \left[\int_{-a}^{b} \frac{(|z^2+1|+2|z|\,|x|)^2}{|(x^2+1)z - x(z^2+1)|^2}\, d\rho(x)\right]^{1/2} \equiv \overline{C}_R(z), \quad z \in \mathscr{C}. \tag{4.7}$$

Since $R_R(n;z)$ is real along $|z|=1$, (4.7) holds also for $z \in \mathscr{C}'$. Let $C_{1R}(z)$ be that function, harmonic in the domain $\mathscr{D}$ bounded by the curves $\mathscr{C}$ and $\mathscr{C}'$, which assumes the values $\ln \overline{C}_R(z)$, for z on $\mathscr{C}$ and $\mathscr{C}'$. Further, we introduce the Green's function $\mathscr{G}(z;z_0)$ of the domain $\mathscr{D}$ with a pole at z_0: $\mathscr{G}(z;z_0)=0$, for z on $\mathscr{C}$ and $\mathscr{C}'$ and $(z_0 \in \mathscr{D})$

$$\mathscr{G}(z;z_0) = \log(1/|z-z_0|) + \varphi_{z_0}(z), \tag{4.8}$$

with $\varphi_{z_0}(z)$ harmonic in $\mathscr{D}$. With this, we construct the difference (the following few lines are actually a derivation of "Lindelöf's principle"[37])

$$D(z) = \ln|R_R(n;z)| - C_{1R}(z) + \sum_{i=1}^{2p+2} \mathscr{G}(z;z_i). \tag{4.9}$$

Writing $(z_k = e^{i\theta_k})$

$$R_R(n;z) = \prod_{k=1}^{2p+2} (z - e^{i\theta_k}) \widetilde{R}_R(z) \tag{4.10}$$

and using the definition (4.8) of $\mathscr{G}(z;z_0)$, we can write

$$D(z) = \ln|\widetilde{R}_R(z)| - C_{1R}(z) + \sum_{k=1}^{2p+2} \varphi_{z_k}(z) \tag{4.11}$$

and see explicitly that $D(z)$ is subharmonic in $\mathscr{D}$, since $\ln|\widetilde{R}_R(z)|$ is so. Now, on the boundary $\mathscr{C} \cup \mathscr{C}'$ of $\mathscr{D}$, using the definition of $\mathscr{G}(z;z_0)$, $D(z) \leqslant 0$. Thus $D(z) \leqslant 0$ in all of $\mathscr{D}$ and we conclude that

$$|R_R(n;z)| \leqslant \exp[C_{1R}(z)] \prod_{k=1}^{2p+2} \exp(-\mathscr{G}(z;z_i)). \tag{4.12}$$

Now, $\mathscr{G}(z;z_i) > 0$ for $z \in \mathscr{D}$; in particular

$$\gamma_1 \equiv \min_{\theta,\varphi} \mathscr{G}(e^{i\theta},e^{i\varphi}) > 0. \tag{4.13}$$

Then, Eq. (4.12) implies the first inequality in (4.6), if $z = e^{i\theta}$, $\widetilde{C} = \max C_{1R}(e^{i\theta})$, $\gamma = e^{-2\gamma_1}$, and $C_R = e^{\widetilde{C}}\gamma^2$. This ends the proof of Lemma 4.1.

Defining now

$$s_1(n;x) = n(x)(1 - x^2 f(n;x)^2), \tag{4.14}$$

$$s_2(n;x) = n(x)(1 - f(n;x)^2), \tag{4.15}$$

we conclude from the inequalities (4.6) that

$$(s_i(n;x), P(x;\theta)\sigma_i(x))_\rho < C_i \gamma^p, \quad i=1,2, \tag{4.16}$$

with $C_1 = C_I/(2\pi)$, $C_2 = (C_R + C_I)/(2\pi)$, $\sigma_1(x) = 1/(1-x^2)$, $\sigma_2(x) = x/(1-x^2)$, and $P(x;\theta)$ given in Eq. (2.11). It is important that we can recover $n(x)$ linearly from the $s_i(n;x)$ and independently of $f(n;x)$:

$$n(x) = [s_1(n;x) - x^2 s_2(n;x)]/(1-x^2). \tag{4.17}$$

Clearly, $s_i(n;x) \in L^2(\rho)$, $i=1,2$, and, since $|f(n;x)| \leqslant 1$, $x \in [-a,b]$,

$$\|s_1(n,x)\|_\rho \leqslant 1, \quad \|s_2(n;x)\|_\rho \leqslant 1. \tag{4.18}$$

We shall now show that the inequalities (4.16) and (4.18) imply an upper bound on the scalar product of the $s_i(n;x)$ with any function $g(x) \in L^2(\rho)$; this upper bound tends to zero as p increases. Then, with

$$h_1(x) = h(x)/(1-x^2), \quad h_2(x) = x^2 h(x)/(1-x^2), \tag{4.19}$$

and using

$$|(n,h)_\rho| \leqslant |(s_1,h_1)_\rho| + |(s_2,h_2)_\rho|, \tag{4.20}$$

we conclude from Eq. (2.14) that $\chi^2_{\min}(\rho;h)$ may become arbitrarily small if p is large. A comparison with the lower bound χ_1^2 to $\chi^2_{\min}(\rho;h)$ will yield then the desired upper bound on the number of zeros.

The derivation of the bound on $(s_i(n;\cdot),g)_\rho$ for some $g \in L^2(\rho)$ runs analogously to Ref. 17, but we give here more accurate estimates. We regard the scalar product in (4.16) as a compact operator A_i mapping $L^2(\rho)$ into $L^2(T)$ [with the norm (1.5)] and denote by B_i the compact positive operator $A_i^\dagger A_i$. We can then write a weaker form of (4.16) as

$$(s_i,B_i s_i)_\rho = \|A_i s_i(n;\cdot)\|_2^2 \leqslant C_i^2 \gamma^{2p} \equiv k_i^2, \quad i=1,2. \tag{4.21}$$

In the following, we drop the subscript i on the various quantities, as the consideration of one index is obviously sufficient. With this, the subset of $L^2(\rho)$ delimited by (4.18) and (4.21) is weakly compact and the linear functional $\Phi(s) \equiv (g,s)_\rho$, determined by the given $g(x)$, attains its maximum on it. Then, application of a known statement about Lagrange multipliers (Theorem 1 of Ref. 31, p. 217, Sec. 8.4) shows that, as a consequence of the convexity of the function to be minimized and of the constraints (4.18) and (4.21), there exist positive numbers λ, μ so that the constrained minimum of $-\Phi(s)$ is equal to the unconstrained minimum of

$$\mathscr{L}(s;\mu;\nu) = -\Phi(s) + \mu((s,s)_\rho - 1) + \nu((s,Bs)_\rho - k^2). \tag{4.22}$$

The minimum of $\mathscr{L}(s;\mu;\nu)$ is achieved at the same point as the unconstrained minimum s_0 and the Lagrange multipliers are such that

$$\mu((s_0,s_0)_\rho - 1) + \nu((s_0,Bs_0)_\rho - k^2) = 0. \tag{4.23}$$

With this, the following is verified by straightforward calculations.

Lemma 4.2: Assume $k^2 < (g,Bg)_\rho/(g,g)_\rho$. Define

$$\alpha(g) = \lim_{\substack{x \to 0 \\ x>0}} \sigma(x;g) \tag{4.24}$$

with

$$\sigma(x;g) = \frac{(g,B(x+B)^{-2}g)_\rho}{(g,(x+B)^{-2}g)_\rho}, \tag{4.25}$$

then

$$\Phi_{\max}(k) = k \frac{(g,(\bar{x}+B)^{-1}g)_\rho}{(g,B(\bar{x}+B)^{-2}g)_\rho^{1/2}}, \tag{4.26}$$

where $\bar{x}=0$ if $k^2 < \alpha(g)$ and is the unique positive root of the equation

$$\sigma(x;g) = k^2, \tag{4.27}$$

if $k^2 > \alpha(g)$.

The possibly unclear points in this lemma concern the existence of the limit in Eq. (4.24) and the uniqueness of the root in Eq. (4.27). Both points are settled by verifying directly that $\sigma(x;g)$ is a strictly monotonically increasing function of x. The relation with Eqs. (4.22) and (4.23) is established through $x=\mu/\nu$. For most situations, $\alpha(g)$ in Eq. (4.24) vanishes. However, if $\rho(x)$ is of finite type and B does not have the eigenvalue zero, then $\alpha(g)>0$, although it is, in general, very small.

We make now a statement concerning the behavior of $\Phi_{\max}(k)$, Eq. (4.26), for small k [i.e., as we increase the number p of zeros of the extremal function, cf. Eq. (4.21)].

Lemma 4.3: For any $\epsilon>0$, there exists $k_0(g;\epsilon)$ such that $\Phi_{\max}(k)<\epsilon$, for any $k<k_0$ $(k>0)$.

Proof: If $\alpha(g)>0$, this is obvious from (4.26). If $\alpha(g)=0$, this is not evident, since x depends on k through Eq. (4.27). Let $P(N_0)$ be the projector in $L^2(\rho)$ onto the subspace spanned by the eigenfunctions $\{\psi_j\}_{j=N_0+1}^{\infty}$ of B, for some N_0 to be specified $(BP(N_0)\neq 0)$. Using Eq. (4.25) in (4.26) and applying Schwarz's inequality, we get

$$\Phi_{\max}(k)\leqslant \frac{((1-P(N_0))g,(\bar{x}+B)^{-1}(1-P(N_0))g)_\rho}{(g,(\bar{x}+B)^{-2}g)_\rho^{1/2}} + \|P(N_0)g\|_\rho. \tag{4.28}$$

Now, if $\psi(N_0;x;g)$ denotes the first expression in (4.27) for variable $x>0$, one verifies that for all $x>0$,

$$0<\psi^2(N_0;x;g)<\sigma(x;g)\|B^{-1/2}(1-P(N_0))g\|_\rho^2.$$

Therefore, $\psi(N_0;x;g)\to 0$, as $x\to 0$. Let $x_0(\epsilon/2)$ be the smallest value of x for which $\psi(N_0;x;g)=\epsilon/2$ and let $k_0(g;\epsilon)$ be computed from Eq. (4.27),

$$k_0^2(g;\epsilon)\equiv\sigma(x_0(\epsilon/2);g). \tag{4.29}$$

The statement of the lemma follows then from the strict monotonicity of $\sigma(x;g)$ and choosing N_0 so that $\|P(N_0)g\|_\rho\leqslant\epsilon/2$.

It is worth noticing that, even if $\alpha(g)\neq 0$ [as is generally the case if $\rho(x)$ is of finite type], the estimate obtained by the method used for $\alpha(g)=0$ yields usually a larger value of $k_0(g;\epsilon)$ than the one obtained directly from (4.26).

We can now state the main result of this section and reinstate to this end the index i [e.g., in Eq. (4.16)].

Theorem 4.1: Assume $\chi_{\min}(\rho;h)>\chi_1>0$. Then the Blaschke product, which is the solution of problem (A), cannot contain more than

$$N_0(\chi_1;h)\equiv\max_{i=1,2}[(\ln(k_{i0}/C_i)/\ln\gamma)] \tag{4.30}$$

factors, where $k_{i0}=k_0(h_i;\chi_1/2)$, $i=1,2$, is obtained from Lemma 4.3, C_i is given following Eq. (4.16), and γ following Eq. (4.13).

Proof: We choose in Lemma 4.3, $g(x)=h_1(x)$, Eq. (4.19) and $\epsilon=\chi_1/2$, and obtain in this way a number $k_{10}=k_0(h_1,\chi_1/2)$, such that, if $C_1^2\gamma^{2p}<k_{10}$, $|(s_1,h_1)|_\rho<\chi_1/2$. The reasoning can be repeated for $i=2$ to yield k_{20}. [Notice that the expression $k_0(g;\epsilon)$ depends itself on the index i through the operator B.] Since the second term in Eq. (2.14) is positive, $\chi_{\min}(\rho;h)<|(n,h)_\rho|$. With Eq. (4.20), this ends the proof.

As announced in the Introduction, we now recall shortly the computation of $\chi^2_{\min,2}(\rho;h)$, Eq. (1.4), which gives immediately a possible value of χ_1^2 in Theorem 4.1. Namely, from the duality relation (2.14) and Lemma 3.3 it is clear that the extremal function must be of the form (3.36) for some $N(x)\in L^2(\rho)$. Let $\tilde{A}$ be the compact operator mapping $L^2(\rho)$ into $L^2(\rho)$ given by the integral in (3.36), with z restricted to $[-a,b]$. One has then the problem of determining

$$\chi^2_{\min,2}(\rho;h)=\inf\{\|h-\tilde{A}N\|_\rho^2;\ N\in L^2(\rho)\} \tag{4.31}$$

under the additional constraint (meaning $\|F(N;z)\|_2^2\leqslant 1$)

$$(N,\tilde{A}N)_\rho\leqslant 1. \tag{4.32}$$

With the same discussion concerning Lagrange multipliers as the one preceding Lemma 4.2, one verifies easily that the following is true.

Lemma 4.4: Let μ be the unique positive root of the equation

$$(h,\tilde{A}(\tilde{A}+\mu)^{-2}h)_\rho=1, \tag{4.33}$$

then

$$\chi^2_{\min,2}(\rho;h)=\|I-\tilde{A}(\tilde{A}+\mu)^{-1}h\|_\rho^2 \tag{4.34}$$

and is attained on

$$f_0^{(2)}(x;h)=\tilde{A}(\tilde{A}+\mu)^{-1}h(x). \tag{4.35}$$

Numerical estimates (see Sec. VII) show, however, that the bound (4.30) is usually rather weak, when compared to the observed number of zeros of the extremal function $f_0(x)$. Its virtue is mainly that it has a weak ρ dependence, so that it stays finite when the number N of jumps of $\rho(x)$ increases indefinitely [with the total variation of $\rho(x)$ staying finite].

V. MINIMIZATION OF χ^2 WITH A FINITE NUMBER OF LINEAR CONSTRAINTS

In this section, we consider a subset $\mathcal{F}_{1;c}(\xi;w)$ of $\mathcal{F}_1$ [Eq. (2.1)] consisting of those function that assume preassigned values at m given points $\xi_1,\xi_2,\ldots,\xi_m$, $\xi_i\in\mathbb{R}$, $|\xi_i|<1$,

$$f(\xi_i)=w_i,\quad i=1,2,\ldots m. \tag{5.1}$$

The set of points $\{w_i\}_{i=1}^m\in R^m$ for which there exist interpolating functions in $\mathcal{F}_1$, satisfying (5.1), is denoted by $\mathcal{S}_m$ [cf. Eq. (2.5), with $N=m$]. We wish to discuss problem (A_c): find the minimal value of $\chi^2(\rho;h-f)$ over $\mathcal{F}_{1;c}(\xi;w)$ and describe the extremal function. This discussion is a prerequisite to the (numerical) solution of problem (C) of the Introduction.

We start with two preparatory steps. The first step is a more accurate description of the set $\mathcal{S}_m$, obtained from the Schur–Pick–Nevanlinna interpolation theory. We state this as the following lemma.

Lemma 5.1: Consider for every $\{w_i\}_{i=1}^m\in\mathcal{S}_m$ the quadratic form in α,

$$F(\alpha)=\sum_{i,j}\frac{1-w_iw_j}{1-\xi_i\xi_j}\alpha_i\alpha_j. \tag{5.2}$$

Then $F(\alpha)$ is strictly positive definite if $w\in\operatorname{int}\mathcal{S}_m\equiv\mathring{\mathcal{S}}_m$ and vanishes for some set of values $\{\alpha_i\}_{i=1}^m\neq 0$ if $w\in\partial\mathcal{S}_m(=\mathcal{S}_m\setminus\mathring{\mathcal{S}}_m)$. If $w\in\partial\mathcal{S}_m$ there exists just one function $f\in\mathcal{F}_1$, obeying (5.1). This function is a Blaschke

product with at most $m-1$ zeros. If $w\in\mathring{\mathcal{S}}_m$, there exists an infinite set of functions in $\mathcal{F}_1$ satisfying (5.1).

We do not give a proof of this statement, since it may be easily abstracted from Ref. 22, in conjunction with the algorithms of Refs. 1, 15, and 23. As a consequence of it, we see that, if $\{w_i\}_{i=1}^m\in\partial\mathcal{S}_m$, the problem of minimization is trivial and $f_c(n;x)$ is uniquely determined by the set $(\xi_i;w_i)_{i=1}^m$ with the interpolation method of Refs. 1, 15, and 23.

The second preparation is formal and necessary for an easy application of the results of Sec. II–IV. The first step is to enlarge the function $\rho(x)$ defined on $[-a,b]$ to a new function $\rho_c(x)$ of bounded variation over an interval $[-\alpha,\beta]\subset(-1,1)$, containing $[-a,b]\cup\{\xi_i\}_{i=1}^m$ by declaring all the points ξ_i to have unit measure and the set of points in $[-\alpha,\beta]$ outside $[-a,b]\cup\{\xi_i\}_{i=1}^m$ to have zero measure. [We assume that, if some of the ξ_i's lie inside $[-a,b]$, they are of measure zero with respect to $\rho(x)$.] We define further a new data function $h_c(x)$ by $h_c(\xi_i)=w_i$ and $h_c(x)=h(x)$, for all $x\in[-a,b]$, $x\neq\xi_i$. Naturally, $L^2(\rho_c)$ is the Hilbert space of ρ_c-measurable functions $g(x)$ for which

$$\chi^2(\rho_c;g)\equiv\|g\|_c^2=\int_{-\alpha}^{\beta}g(x)^2d\rho_c(x). \tag{5.3}$$

We identify $L^2(\rho)$ with that subspace of $L^2(\rho_c)$ made up of functions $g(x)$ with $g(\xi_i)=0$. We may clearly identify $L^2(\rho_c)$ with $L^2(\rho)\otimes R^m$ and write, e.g., $h_c\equiv(h,w)$, $h\in L^2(\rho)$, $w=\{w_i\}_{i=1}^m$. Further, we denote by $\omega_i(x)$ the characteristic functions of the points $\xi_i(\omega_i(\xi_i)=1,\omega_i(x)=0,x\neq\xi_i)$. Clearly, $\omega_i\in L^2(\rho_c)$ and, if $f\in L^2(\rho_c)$,

$$(\omega_i,f)_c=f(\xi_i), \tag{5.4}$$

and $(\omega_i,g)_c=0$, if $g\in L^2(\rho)$. In fact, one reason for introducing $L^2(\rho_c)$ is that the value of f at ξ_i is a continuous linear functional in this space. With this, we rewrite the duality relation (2.14) in a form valid in $L^2(\rho_c)$; in the rest of this section we suppress the explicit ξ and w dependence of $\mathcal{F}_{1;c}$; the symbol $\tilde{\mathcal{F}}_{1;c}$ is clear from Sec. II [following (2.1)],

$$\inf_{f\in\tilde{\mathcal{F}}_{1;c}}\|h-f\|_\rho=\inf_{f\in\tilde{\mathcal{F}}_{1;c}}\|h_c-f\|_c$$
$$=\sup_{\substack{(\omega_\rho n)_c=0\\ \|n\|_c^2\leq 1}}\left[(n,h_c)_c-\sup_{f\in\tilde{\mathcal{F}}_{1;c}}(n,f)_c\right]. \tag{5.5}$$

Now, as in Sec. II, one sees that, since $\tilde{\mathcal{F}}_{1;c}$ is convex and closed, there exists a unique $f_{0;c}(x)\in\tilde{\mathcal{F}}_{1;c}$ on which the infimum in Eq. (5.5) is attained; also from general arguments (see Ref. 31, p. 136, Theorem 1), an extremal $n_{0;c}(x)$ on the right-hand side exists and is in fact unique, in view of the equalities [cf. Eqs. (2.16) and (2.17)]

$$\sup\{(n_{0;c},f)_\rho:\ f\in\tilde{\mathcal{F}}_{1;c}\}$$
$$=(n_{0;c},f_{0;c})_\rho,\quad n_{0;c}(x)=\frac{h(x)-f_{0;c}(x)}{\|h(x)-f_{0;c}(x)\|_\rho}. \tag{5.6}$$

One of the important difference with respect to Secs. II–IV is that the set $\tilde{\mathcal{F}}_{1;c}$ is not symmetrical with respect to the origin [it does not contain $f(x)$ and $-f(x)$]. As a consequence, the second term on the right-hand side of (5.5) is not necessarily positive. However, Theorems 2.1 and 2.2 can be easily generalized to the set $\tilde{\mathcal{F}}_{1;c}$, using Lagrange multipliers.

Theorem 5.1: Let $n(x)\in L^2(\rho)$. There exists a unique $f_c(n;z)\in\tilde{\mathcal{F}}_{1;c}$, such that $\sup\{(n,f)_\rho:f\in\tilde{\mathcal{F}}_{1;c}\}$ is attained on it; $f_c(n;z)$ is a finite Blaschke product.

The existence of $f_c(n;x)$ follows from general arguments. The set $\tilde{\mathcal{F}}_{1;c}$ in $L^2(\rho)$ is closed in the norm (2.1) and convex; hence it is weakly closed and, since it is contained in the set $\|f\|_\rho\leq A$ for some $A>0$, it is weakly compact. Thus, any linear (continuous) functional attains its maximum on it. If $\{w\}_{i=1}^m\in\partial\mathcal{S}_m$, the statement of the theorem follows from Lemma 5.1. If $\{w_i\}\in\mathring{\mathcal{S}}_m$, it is convenient to switch to $L^2(\rho_c)$ and recall that, as a consequence of a general theorem concerning Lagrange multipliers (see Ref. 31 and Appendix C), there exist numbers λ_i (*not* necessarily positive) such that [using Eq. (5.4)]

$$(n,f_c(n;x))_c=\sup_{f\in\tilde{\mathcal{F}}_{1;c}}(n,f)_c$$
$$=\sup_{f\in\tilde{\mathcal{F}}_1}\left[(n,f)_c+\sum_i\lambda_i((\omega_i,f)_c-w_i)\right], \tag{5.7}$$

where the supremum on the right is achieved at the same $f_c(n;x)$. The problem of maximizing the right-hand side of (5.7) over $\mathcal{F}_1$, for any choice of λ_i, is the same as the one solved by Theorem 2.1, with the only replacement $n(x)\to n_\lambda(x)$, where

$$n_\lambda(x)=\left[\left(1+\sum_{i=1}^m\lambda_i^2\right)^{1/2}\right]^{-1}\left(n(x)+\sum_i\lambda_i\omega_i(x)\right). \tag{5.8}$$

We conclude from Theorem 2.1 that $f_c(n;z)$ is indeed a finite Blaschke product. The uniqueness of $f_c(n;z)$ is obtained through the following consideration: if two different Blaschke products f_{c1},f_{c2} lead to the same extremal value of $(n,f)_\rho$ over $\tilde{\mathcal{F}}_{1;c}$, then their convex combination $f_{c\mu}=\mu f_{c1}+(1-\mu)f_{c2}$ for any μ, $0<\mu<1$, leads to the same value. However, for no value of $\mu\neq 0,1$ can $f_{c\mu}(\theta)$ have unit modulus for all θ, unless $f_{c1}=f_{c2}$, as is easy to verify. This contradicts the fact that the extremals are Blaschke products only and proves Theorem 5.1.

We may now obviously state, using the first of Eqs. (5.6), the following theorem.

Theorem 5.2.: The infimum of $\chi^2(\rho;h-f)$ over $\tilde{\mathcal{F}}_{1;c}$ is realized by a finite Blaschke product $f_{0;c}(z)$.

In the following, we shall assume, unless otherwise stated, that $w\in\mathring{\mathcal{S}}_m$. Then, as a consequence of Lemma 5.1, the extremal Blaschke products in Theorem 5.1 (and thus the one in Theorem 5.2) contain $p\geq m$ zeros. Further, Lagrange multipliers λ_i exist, so that (5.7) holds. This latter allows a generalization of Theorems 3.1 and 3.2 of Sec. III to problem (A_c). We notice namely that many arguments of Secs. II–IV may be used without change, except for the replacements $n(x)\to n_\lambda(x)$, $[-a,b]\to[-\alpha,\beta]$, $k(n;z)\to k_\lambda(n;z)$, $L(n;z)\to L_\lambda(n;z)$, for example:

$$k_\lambda(n;z)=\frac{k(n;z)+\sum_{i=1}^m\lambda_i[1/(\xi_i-z)]}{(1+\sum_{i=1}^m\lambda_i^2)^{1/2}}$$
$$=(n_\lambda(x),1/(x-z))_c. \tag{5.9}$$

In particular, $L_\lambda(n;z)$ satisfies

$$L_\lambda(n;z) = 2\pi(n_\lambda(x), P(x;z) f_c(n;x))_c. \tag{5.10}$$

Expressions (5.9) and (5.10) may be written for any set $\{\lambda_i\}_{i=1}^m$; we shall, however, need in the following only the "correct," but unknown so far, λ_i's, for which Eq. (5.7) holds. In fact, we may characterize these λ_i's by the following lemma.

Lemma 5.2: If the Blaschke product that maximizes $(n_\lambda, f)_c$ over $\mathscr{F}_1$ assumes the values w_i at the points ξ_i ($w \in \mathring{\mathscr{S}}_m$), it also maximizes $(n, f)_\rho$ over $\mathscr{F}_{1;c}$.

The proof is given in Appendix C. With this, we see that the generalization of Lemma 3.2 to the present case reads as the following lemma.

Lemma 5.3: Assume $w = \{w_i\}_{i=1}^m \in \mathring{\mathscr{S}}_m$. The Blaschke product $B(x;\alpha)$ is the extremal function $f_c(n;x)$ associated to $n(x)$ by Theorem 5.1 if and only if m real numbers λ_i exist, so that

$$B(\xi_i;\alpha) = w_i, \tag{5.11}$$

$$(n_\lambda(x), B(x;\alpha)P(x;\alpha_i)/\alpha_i)_c = 0, \quad i = 1,2,...,p, \tag{5.12}$$

$$(n_\lambda(x), B(x;\theta)P(x;e^{i\theta}))_c \geqslant 0, \quad 0 \leqslant \theta < 2\pi. \tag{5.13}$$

Notice, at this stage we do not yet know whether the Lagrange multipliers associated to a given $n(x)$ and to a set $\{\xi_i;w_i\}_{i=1}^m$, $w \in \mathring{\mathscr{S}}_m$, are unique. This is settled by the following lemma.

Lemma 5.4: If $\{w_i\}_{i=1}^m \in \mathring{\mathscr{S}}_m$, there exists a unique set of m Lagrange multipliers λ_i, for which (5.7) holds.

Proof: Assume the opposite were true and write out conditions (5.12) for two such different sets $\{\lambda_i\}_{i=1}^m$, $\{\lambda_i'\}_{i=1}^m$. By Theorem 5.1, $f_c(n;x) = B(x;\alpha)$ is the same in both cases and thus the p zeros α_i are the same. By Lemma 5.1, $p \geqslant m$. We multiply each of the equations (5.12) by the normalization factor $(1 + \Sigma_i \lambda_i^2)^{1/2}$ and subtract the equations corresponding to the same α_i from each other. Using Eq. (5.8) and Eq. (3.15), we obtain a set of p homogeneous equations for the m quantities $\Delta\lambda_k = \lambda_k - \lambda_k'$:

$$\sum_{k=1}^m \Delta\lambda_k \frac{\partial B}{\partial \alpha_i}(\xi_k;\alpha) = 0, \quad i = 1,2,...,p. \tag{5.14}$$

However, the rank of the matrix $\| (\partial B/\partial\alpha_i)(\xi_k;\alpha) \|_{i=1,p}^{k=1,m}$ is precisely m. Indeed, if $B(\xi_k;\alpha) \neq 0$, for all k, this follows directly from Eq. (3.15) and the identity

$$\det \left| \frac{\partial B}{\partial \alpha_i}(\xi_k;\alpha) \right|_{i,k=1}^m \equiv \left(\prod_{i=1}^m 2\pi B(\xi_i;\alpha) \right) \det \left(\frac{P(\xi_k;\alpha_i)}{\alpha_i} \right)_1^m, \tag{5.15}$$

where the last determinant is nonvanishing, by Eq. (2.12) evaluated for complex θ. If, however, say, $B(\xi_m;\alpha) = 0$, $\xi_m = \alpha_m$, one verifies using Eq. (2.12) that the limit $\alpha_m \to \xi_m$ of the right-hand side of (5.15) exists and is nonzero. This shows that $\Delta\lambda_k = 0$, for all k, and ends the proof of Lemma 5.5.

We summarize next our knowledge in more geometrical language: by Theorem 2.1, for any element (n,λ) of $L^2(\rho_c) \simeq L^2(\rho) \otimes R^m$, there exists a Blaschke product $B(n;\lambda;x)$ that realizes[38] $\sup\{(n,\lambda;f)_c : f \in \mathscr{F}_1\}$ (see Ref. 39). It may happen that $B(n;\lambda;z)$ has $p \leqslant m-1$ zeros [as it does, e.g., if $(n;\lambda)$ is of the special form $(0;\lambda)$] in which case the values $\{w_i\}_{i=1}^m$ lie on $\partial\mathscr{S}_m$. The set of elements (n,λ) whose associated $B(n;\lambda;x)$ have this property make up a closed cone $\mathscr{K}$ in $L^2(\rho_c)$. If $(n;\lambda)$ belongs to $\mathscr{N} \equiv \mathbf{C}\mathscr{K}$, the values $w_i = B(n;\lambda;\xi_i)$ lie in $\mathring{\mathscr{S}}_m$. At fixed n, there is a certain allowed region for λ, such that $(n,\lambda) \in \mathscr{N}$; as λ moves through this region, the corresponding $B(n;\lambda;\xi_i)$ moves throughout $\mathring{\mathscr{S}}_m$, by Lemma 5.4. By Theorem 3.1, we can generate pairs $((n,\lambda);B(n;\lambda;x))$ by means of the solutions of an $H^2(D)$ problem [with respect to the scalar product in $L^2(\rho_c)$]. Out of these pairs, we can pick the solution of our problem, i.e., the correct Lagrange set $(\lambda_1,\lambda_2,...,\lambda_m)$ by solving for λ the equation

$$w_i = B(n;\lambda;\xi_i), \quad i = 1,2,...,m. \tag{5.16}$$

This we can do, by Lemma 5.4.

Following Sec. III, we consider, for given pairs $(N;\Lambda) \in L^2(\rho_c)$, the functions

$$F_{0;c}(N;\Lambda;z) = \int_{-\alpha}^{\beta} \frac{N(x) + \Sigma_i \Lambda_i \omega_i(x)}{1 - xz} d\rho_c(x) \tag{5.17}$$

and their canonical decomposition

$$F_{0;c}(N;\Lambda;z) = B(N;\Lambda;z) E_0(N;\Lambda;z). \tag{5.18}$$

By Theorem 3.1, $B(N;\Lambda;z)$ realizes $\sup\{(n,\lambda;f)_c : f \in \mathscr{F}_1\}$ with

$$n(x) = N(x) E_0(N;\Lambda;x), \tag{5.19}$$

$$\lambda_i = \Lambda_i E_0(N;\Lambda;\xi_i), \quad i = 1,...,m. \tag{5.20}$$

If it happens that the point with coordinates $\{w_i\}_{i=1}^m$ given by (5.16) lies in $\mathring{\mathscr{S}}_m$, then, by Lemma 5.2, $B(N;\Lambda;z)$ realizes even $\sup\{(n,f)_\rho : f \in \mathscr{F}_{1;c}(\xi;w)\}$.

We generalize next the integral equation (3.47) to the situation of this section. The role of the unknown is played by a pair $(N_{0;c};\Lambda_0)$ in $L^2(\rho_c)$. From the second of Eqs. (5.6) and from (5.16), we obtain, by means of (5.18)–(5.20) the set

$$n_{0;c}(x) = N_{0;c}(x) E_0(N_{0;c};\Lambda_0;x) = h(x) - B(N_{0;c};\Lambda_0;x), \tag{5.21}$$

$$w_i = B(N_{0;c};\Lambda_0;\xi_i), \quad i = 1,...,m. \tag{5.22}$$

It is easy to show that the set of equations (5.21) and (5.22) has a unique solution if $w \in \mathring{\mathscr{S}}_m$. Since, given $\{w_i\}_{i=1}^m$, there exists a unique $n_{0;c}(x)$ satisfying

$$n_{0;c}(x) = h(x) - f(n_{0;c};x) \tag{5.23}$$

[cf. Eq. (5.6)], it is enough to show that the set of equations given by (5.19) and (5.22) has a solution for any pair $(n;w)$ in $L^2(\rho) \otimes \mathring{\mathscr{S}}_m$. Now, the mapping $\mathscr{C}_c$: $L^2(\rho_c) \to L^2(\rho) \otimes \mathscr{S}_m$ given by Eqs. (5.19) and (5.22) may be written as

$$\mathscr{C}_c = \mathscr{B} \circ \mathscr{C}_c', \tag{5.24}$$

with $\mathscr{C}_c'$ the mapping $L^2(\rho_c) \to L^2(\rho_c)$ given by (5.19) and (5.20) and $\mathscr{B}$ the mapping $R^m \to \mathscr{S}_m$ given, at fixed $n \in L^2(\rho)$, by (5.16). But Lemma 3.5 applied to $L^2(\rho_c)$ shows that $\mathscr{C}_c'$ is one-to-one and onto and Lemma 5.4 shows that $\mathscr{B}$ is invertible, if $w \in \mathring{\mathscr{S}}_m$. Clearly, the solution $(N_{0;c};\Lambda_0)$ lies in the set $\mathscr{N}_2 \equiv (\mathscr{C}_c')^{-1}\mathscr{N}$. We conclude that the set of equations (5.21) and (5.22) has indeed a unique solution.

We now state directly the analog of Theorem 3.2. Let $\mathscr{A}_c(N;\Lambda)$ be the mapping from

$$\mathscr{N}_2 \subset L^2(\rho_c) \to L^2(\rho) \otimes \mathscr{S}_m$$

given by

$$\mathscr{A}_c(N;\Lambda;x) = (N(x)E_0(N;\Lambda;x) + B(N;\Lambda;x), B(N;\Lambda;\xi_i)). \tag{5.25}$$

Recalling the definition of $h_c(x)$ preceding Eq. (5.3) and assuming $w \in \mathring{\mathscr{S}}_m$, we state the following theorem.

Theorem 5.3: For any $h(x)$ outside $\mathscr{F}_{1;c}$, the equation

$$\mathscr{A}_c(N;\Lambda;x) = h_c(x) \tag{5.26}$$

has a unique solution in $\mathscr{N}_2$. The operator $\mathscr{A}_c$ has a continuous Fréchet derivative with a bounded inverse at all points of $\mathscr{N}_2$ such that $\mathscr{F}_{c;0}(N;\Lambda;z)$, Eq. (5.18), does not vanish on $|z| = 1$. At those points of $\mathscr{N}_2$, which do not have this latter property, but for which conditon (H), Sec. III, is true, the operator whose action $L^2(\rho_c)$ is given by the Gateaux differential $\delta\mathscr{A}_c(N;\Lambda;\delta N;\delta\Lambda)$ also has a bounded inverse.

It may appear unfortunate that we have to restrict the statement to the set $\mathscr{N}_2$, whose description is complicated. It turns out, however, that, in a numerical search, at least if m is small (= 1 or 2), we never get outside $\mathscr{N}_2$. [That is, the functions $F(N;\Lambda;z)$, Eq. (5.17) have $p \geqslant m$ zeros in $|z| < 1$.]

Proof: Consider first the situation when $F_c(N;\Lambda;z)$ does not vanish on $|z| = 1$. We write $\mathscr{A}_c$ as the composition $\mathscr{A}_c = \mathscr{A}_c' \circ \mathscr{B} \circ \mathscr{C}_c'$ with $\mathscr{A}_c'$ the mapping from $L^2(\rho) \otimes \mathscr{S}_m$ into itself given by $n(x) \to n(x) + B(n;x;w)$, $w \to w$; $B(n;x;w)$ is the Blaschke product realizing the extremum in Theorem 5.1, with the w dependence explicitly shown. We shall show that $\mathscr{A}_c'$ has a continuous Fréchet derivative; then, by the convexity argument of Theorem 3.2 applied to $\mathscr{F}_{1;c}$ [in $L^2(\rho)$], it follows easily that this derivative has a bounded inverse. Further, by Lemma 2.6 applied in $L^2(\rho_c)$, the derivative $\partial\mathscr{C}_c'/\partial(N;\Lambda)$ exists, is continuous, and has a bounded inverse. Thus, we have to show that $\partial\mathscr{B}/\partial\lambda$ exists and is invertible at points of $\mathscr{N}$, i.e., that the Jacobian $(\partial w_i/\partial\lambda_k)_{i,k=1}^m$ is nonvanishing.

To this end, we use Lemma 3.3 in $L^2(\rho_c)$ and consider the decompositions (5.18) corresponding to two functions $F_{c,0}(N;\Lambda;z)$, $F_{c;0}(n+\Delta n;\Lambda+\Delta\Lambda;z)$, where N, Λ, ΔN, $\Delta\Lambda$ are so chosen that $\mathscr{C}_c(N;\Lambda) = (n;\lambda)$, $\mathscr{C}_c(N+\Delta N;\Lambda+\Delta\Lambda) = (n;\lambda+\Delta\lambda)$ for a given $n(x)$ and given sets $\{\lambda_i\}_{i=1}^m$, $\{\Delta\lambda_i\}_{i=1}^m$. Such a choice is possible by Lemma 3.5 applied in $L^2(\rho_c)$. We consider now the variation of the functional $(N,\Lambda;F)_c$ as the Blaschke product $B(N;\Lambda;z) \equiv B(n;\lambda;z)$ changes to $B(n;\lambda+\Delta\lambda;z)$; the functional decreases, by Lemma 3.3. Equation (3.63) gives a quantitative expression for this change:

$$\begin{aligned}
&(N,\Lambda;E(N;\Lambda)B(n;\lambda))_c - (N,\Lambda;E(N;\Lambda)B(n;\lambda+\Delta\lambda))_c \\
&\quad = -(n,\Delta_\lambda B(n;\lambda))_\rho - \sum_{i=1}^m \lambda_i \Delta_\lambda w_i \\
&\quad = \frac{1}{2}\oint |E^2(N;\Lambda;\theta)|\,|\Delta_\lambda B(n;\lambda;\theta)|^2\,d\theta \geqslant 0,
\end{aligned} \tag{5.27}$$

where the symbols $\Delta_\lambda B(n;\lambda)$, $\Delta_\lambda w_i$ denote finite differences. On the other hand,

$$\begin{aligned}
&(N+\Delta N,\Lambda+\Delta\Lambda;E(N+\Delta N;\Lambda+\Delta\Lambda)B(n;\lambda+\Delta\lambda))_c \\
&\quad - (N+\Delta N,\Lambda+\Delta\Lambda;E(N+\Delta N;\Lambda+\Delta\Lambda)B(n;\lambda))_c \\
&\quad = (n,\Delta_\lambda B(n;\lambda))_\rho + \sum_{i=1}^m (\lambda_i+\Delta\lambda_i)\Delta_\lambda w_i \\
&\quad = \frac{1}{2}\oint |E^2(N+\Delta N;\Lambda+\Delta\Lambda;\theta)|\,|\Delta_\lambda B(n;\lambda;\theta)|^2\,d\theta \geqslant 0.
\end{aligned} \tag{5.28}$$

We now add Eqs. (5.27) and (5.28), write $\Delta\lambda_i = \beta_i\epsilon$ for some fixed nonzero vector $\{\beta_i\}_{i=1}^m$, divide through by ϵ^2, and let $\epsilon \to 0$. The finite differences go over into derivatives, which exist by virtue of the fact that the mapping $\mathscr{C}_c'$ has a bounded inverse Fréchet derivative. We obtain

$$\sum_{i,j=1}^m \beta_i \frac{\partial w_i}{\partial\lambda_j}\beta_j = \oint |E^2(N;\Lambda)| \left|\sum_{i=1}^m \beta_i \frac{\partial B(n;\lambda;\theta)}{\partial\lambda_i}\right|^2 d\theta. \tag{5.29}$$

The last expression is strictly positive, for any choice of the $\{\beta_i\}_{i=1}^m \neq 0$, if the functions $(\partial B/\partial\lambda_i)(n;\lambda;\theta)$ are linearly independent. We show briefly that this is the case indeed. To this end, we notice that, under the assumption that the zeros of $B(n;\lambda;z)$ are simple, it follows from (3.52) that $(\partial/\partial\lambda_i)(B(n;\lambda;z))$ is a meromorphic function of z; thus if the functions $(\partial B/\partial\lambda_i)(n;\lambda;\theta)$ were linearly dependent, nonzero β_i's would exist so that, for all z ($p \geqslant m$, since $w \in \mathring{\mathscr{S}}_m$),

$$\sum_{i=1}^m \beta_i \sum_{j=1}^p B(z;\alpha)\frac{P(z;\alpha_j)}{\alpha_j}\frac{\partial\alpha_j}{\partial\lambda_i} = 0. \tag{5.30}$$

This implies, by the linear independence of the $P(z;\alpha_j)$ that, for all p,

$$\sum_{i=1}^m \beta_i \frac{\partial\alpha_j}{\partial\lambda_i} = 0, \quad j = 1,2,\ldots,p. \tag{5.31}$$

However, differentiation of any subset of m identities (5.12) with respect to λ_i, multiplication of the results by β_i, and addition over i yields, using (5.31),

$$\sum_{i=1}^m \beta_i \frac{\partial B}{\partial\alpha_k}(\xi_i) = 0, \quad k = 1,2,\ldots,m, \tag{5.32}$$

and this implies, with the reasoning of Lemma 5.4 that all $\beta_i = 0$. Thus, the right-hand side of (5.29) is positive for any choice of the β_i and consequently the Jacobian $(\partial w_i/\partial\lambda_j)_{i,j=1}^m$ is nonzero. Thus, $\mathscr{C}_c = \mathscr{B} \circ \mathscr{C}_c'$ has an invertible Fréchet derivative. It also follows that we can solve (5.16) at fixed left-hand side and obtain a function $\lambda = \lambda(n;w;\xi)$ with a continuous Fréchet derivative with respect to $n(x)$. Substituting it in $B(n;\lambda;x)$, $x \in [-a,b]$, we obtain a Fréchet differentiable mapping with respect to $n(x)$ from $L^2(\rho)$ into $L^2(\rho)$. It follows that $\mathscr{A}_c'$ has a continuous Fréchet derivative at $(n;w)$, which, as argued at the outset, must be invertible. This disposes of the situation $F_c(N;\Lambda;z) \neq 0$, for $|z| = 1$.

Now, as in Sec. III, if $F_c(\tilde{N};\tilde{\Lambda};z_0) = 0$, for z_0 on $|z| = 1$ and the zero is simple [cf. condition (H), Sec. III], we may build distinct "exterior" and "interior" limits for $\partial\mathscr{C}_c/\partial(N;\Lambda)$ and these, by Lemma 3.6, have bounded inverses. It follows that the derivative $\partial\mathscr{B}/\partial(n;\lambda)$ has two distinct limits at points $(\tilde{n};\tilde{\lambda}) = \mathscr{C}_c'(\tilde{N};\tilde{\Lambda})$. One must show that both these limits have nonvanishing Jacobians. This is done by

taking the "exterior" and "interior" limits, in turn, $(n_k,\lambda_k)\to(\tilde{n},\tilde{\lambda})$, at fixed directions β_i, for both sides of the identity (5.29) and verifying that the right-hand side remains strictly positive in this process. The limit of the left- and right-hand sides, as $(n,\lambda)\to(\tilde{n},\tilde{\lambda})$ exist in both cases, since the corresponding limits of $\partial\mathscr{E}_c(N;\Lambda)/\partial(N;\Lambda)$ as $(N,\Lambda)\to(\tilde{N},\tilde{\Lambda})$ exist. As above, we may then look for consequences of the hypothesis that, for some $\{\beta_i\}_{i=1}^m$, the right-hand side of (5.29) would vanish. For the "exterior" limit, there is no change in the argument and we can derive again Eq. (5.32) and thus conclude $(\partial w_i/\partial\lambda_j)_e(\tilde{n};\tilde{\lambda})\neq 0$. In the "interior" process, Eq. (5.31) is true only for those zeros staying in $|z|<1$; it has to be modified for those zeros that approach $|z|=1$. However, one can verify that conclusion (5.32) is still correct, so that $(\partial w_i/\partial\lambda_j)_i(\tilde{n},\tilde{\lambda})\neq 0$.

Finally, it follows that

$$\frac{\partial\mathscr{A}_c'}{\partial n}=\left(1+\frac{\partial\mathscr{B}}{\partial n}+\frac{\partial\mathscr{B}}{\partial\lambda}\frac{\partial\lambda}{\partial n}\right)(n;w)$$

has two distinct limits as $(n,w)\to(\tilde{n},\tilde{w})$ according to the position of (n,w) with respect to the image through $\partial\mathscr{E}_c/\partial(N,w)$ of the plane $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})=0$. Taking the limit $n,w\to\tilde{n},\tilde{w}$ in Eq. (3.75) [referring to $\tilde{\mathscr{F}}_{1;c}(\xi;w)$], we conclude that, as in Theorem 3.2, both $(\partial\mathscr{A}_c'/\partial n)_e$, $(\partial\mathscr{A}_c'/\partial n)_i$ are invertible. Now, both $(\partial\mathscr{A}_c/\partial(N;\Lambda))_{e,i}$ have the same action on the plane $\mathrm{Re}(\delta\alpha_{p+1}/\alpha_{p+1})=0$ and $\mathscr{A}_c$ is one-to-one in a neighborhood of h_c. The argument of Lemma 3.7 leads then to the statement of the theorem. This ends the proof.

Theorem 5.3 justifies the performance of an iteration of the Newton type for the solution of problem (A_c).

We now present briefly the mechanism that limits the number of zeros of the extremal function of problem (A_c), leading to a statement similar to Theorem 4.1. With the change $L^2(\rho)\to L^2(\rho_c)$, we may take over many of the arguments of Sec. IV. There are, however, two difficulties, with one common root.

(a) The argument of Lemma 4.1 provides a bound like (4.16) [or (4.21)] in terms of the number p of zeros of the extremal function $f_{0;c}(n;x)$,

$$|(s_{i;\lambda},P(x;\theta)\sigma_i(x))_c|\leqslant C_{i;c}\gamma_c^p\equiv k_{i;c},\quad i=1,2, \tag{5.33}$$

with $C_{i;c},\gamma_c$ obtained similarly to C_i,γ of Eq. (4.16) with the geometry appropriate to the replacement $[-a,b]\to[-\alpha,\beta]$. In Eq. (5.33), $s_{i;\lambda}(n;x)$ is related to $n_\lambda(x)$, Eq. (5.8), by Eqs. (4.14) and (4.15). On the other hand, as one sees from Eqs. (5.5) and (4.30), one needs to show that, as p increases, the quantities $s_i(n;x)$ related to $n(x)$ by (4.14) and (4.15) and which do *not* contain the Lagrange multipliers λ_i tend weakly to zero, as p increases.

(b) The set $\mathscr{F}_{1;c}$ is asymmetrical, so that, as observed following Eqs. (5.6), we have to place a bound also on the second term in the duality relation (5.5). With the notation of Eq. (5.10), it follows from the saturation of inequalities analogous to (2.23) that

$$(n,f_c(n;x))_\rho=\frac{1}{2\pi}\oint L_\lambda(n;\theta)\,d\theta\left(1+\sum_i\lambda_i^2\right)^{1/2}+\sum_{i=1}^m\lambda_i w_i. \tag{5.34}$$

The integral is bounded by the estimates (5.33) and vanishes exponentially with p. Thus, we have again the task of showing that, for a sequence $\{n_k(x)\}_{k=1}^\infty$ such that $f_c(n_k;x)$ have an increasing number of zeros, the corresponding $\lambda_{i,k}$ tend to zero. Thus we need the following lemma.

Lemma 5.7: For any $\epsilon_c>0$, there exists $\kappa_\lambda(\epsilon_c)$ so that $|\lambda_l|<\epsilon_c$, $l=1,2,...,m$, if $k_{i;c}<\kappa_\lambda(\epsilon_c)$, $i=1,2$.

Proof: We apply Lemma 4.3 in $L^2(\rho_c)$ choosing in turn $g(x)=\omega_l(x)$, $l=1,2,...,m$ [$\omega_l(x)$ is the characteristic function of ξ_l]. Let the corresponding functional be $\Phi(\omega_l;\cdot)$. Then, with $i=1$,

$$\Phi(\omega_l;s_{1;\lambda})=-(\omega_l,s_{1\lambda}(n;\cdot))_c=\frac{\lambda_l(1-\xi_l^2w_l^2)}{(1+\Sigma_{j=1}^m\lambda_j^2)^{1/2}}. \tag{5.35}$$

The statement of the lemma is then obvious, if we choose $\kappa_\lambda(\epsilon_c)=\min_{i,l}k_{0;i}(\omega_l;\epsilon_{i;l})$, where $k_{0;i}$ is the function of Lemma 4.3 and $\epsilon_{i;l}$ is an appropriate function of ϵ_c, obtained from (5.35) [e.g., if $\epsilon_c\leqslant(1/m)^{1/2}$, $\epsilon_{1;l}=\epsilon_c(1-\xi_l^2w_l^2)/2$]. This ends the proof. This lemma shows again the advantages of introducing $L^2(\rho_c)$.

With this, the problem of obtaining a bound on the number of zeros of $f_c(n;x)$ is solved with precisely the same argument and qualitatively with the same result as in Theorem 4.1. We skip the details, to be obtained from Ref. 40.

VI. THE STABLE ANALYTIC CONTINUATION OFF A CERTAIN SET OF INTERIOR POINTS

In this section, we show how the developments of Secs. II–V allow us to give a numerical solution to problems (B) [and (B_c)] and (C) of the Introduction.

In relation to problem (B), we recall the definition of the stable extrapolate: to each $\epsilon>0$, we associate a data function $h_\epsilon(x)\in L^2(\rho)$, so that

$$\chi^2(\rho;h_\epsilon-f_t)\leqslant\chi_0^2\epsilon, \tag{6.1}$$

where $f_t(x)$ is the true, unknown, analytic function, $f_t\in H_R^\infty(D)$, that is being measured. Consider the family $\Phi(\epsilon;h_\epsilon)$ of functions $f\in H_R^\infty(D)$, for which

$$\chi^2(\rho;h_\epsilon-f)\leqslant\chi_0^2\epsilon. \tag{6.2}$$

Any $f\in\Phi(\epsilon;h_\epsilon)$ is a valid extrapolation of h_ϵ to $|z|<1$. A procedure of analytic continuation is a prescription for choosing a unique f_e out of $\Phi(\epsilon;h_\epsilon)$, given ϵ and h_ϵ. The procedure is called stable (in $|z|<1$) if, as $\epsilon\to 0$, for any choice of h_ϵ in (6.1), $f_e(\epsilon;h_\epsilon;z)\to f_t(z)$, for all z in $|z|<1$. The function $f_e(\epsilon;h_\epsilon;z)$ chosen by a stable procedure is called a stable extrapolate to $|z|<1$.

Consider now the following prescription: for each ϵ and $h_\epsilon(x)$, define

$$M_0(\epsilon;h_\epsilon)\equiv\inf\{\|f\|_\infty:\ \chi^2(\rho;h_\epsilon-f)\leqslant\chi_0^2\epsilon,\ f\in H_R^\infty(D)\} \tag{6.3}$$

[cf. Eq. (1.8)]. We show below that there exists a unique $f_e(\epsilon;h_\epsilon)\in H_R^\infty(D)$ which realizes the extremum in (6.3). The claim is that $f_e(\epsilon;h_\epsilon)$ is a stable extrapolate to $|z|<1$ (cf. the related problem of Ref. 41).

This is a consequence of Tykhonov's criterion for stabil-

ity (Ref. 4, p. 28, §11), which we may state in our case as follows: If the set of extrapolates $f_e(\epsilon)$, for $0<\epsilon<\epsilon_0$ is contained within a set $\mathscr{S}$, compact with respect to the uniform convergence on compact subsets of $|z|<1$, then $f_e(\epsilon)$ provides a stable extrapolation of the data to $|z|<1$. Now, the set of functions uniformly bounded in $|z|<1$ makes up such a compact set (see, e.g., Ref. 29, Theorem 14.6) and one has to show that all functions defined by (6.3) obey such a bound, i.e., that the sequence $M_0(\epsilon_k;h_{\epsilon_k})$ is bounded as $\epsilon_k\to 0$. But $f_t(z)\in\Phi(\epsilon;h_\epsilon)$, for all ϵ, so that $M_0(\epsilon;h_\epsilon)\leqslant\|f_t\|_\infty<\infty$, which shows that Tykhonov's criterion is satisfied. We now prove the following lemma.

Lemma 6.1: There exists a unique $f_e(\epsilon;h_\epsilon)$ which realizes the infimum in Eq. (6.3). It is of the form $M_0B_{M_0}(z)$, with $B_{M_0}(z)$ a finite Blaschke product.

Proof: One considers the function of M [recall definition (2.1)]

$$\tilde{\chi}_{\min}(M)=\inf\{\chi(\rho;h-f),\ f\in\mathscr{F}_M\}. \tag{6.4}$$

It is easy to verify that $\tilde{\chi}_{\min}(M)$ is in fact a strictly monotonically decreasing convex function of M on some interval $[O,M_{\max}]$ ($M_{\max}$ may be infinite). Strict monotonicity follows from the uniqueness of the function in $\mathscr{F}_M$ that realizes the infimum. From Theorem 2.2, it follows it is of the form $MB_M(z)$, with $B_M(z)$ a finite Blaschke product. The two inequalities $[\rho(b)-\rho(-a)=1]$,

$$\tilde{\chi}^2_{\min}(0)-\tilde{\chi}^2_{\min}(M)\leqslant M(M+2\|h_\epsilon\|_\rho) \tag{6.5}$$

and

$$\begin{aligned}\tilde{\chi}_{\min}(M)&<M\chi\left(\rho;\frac{h_\epsilon}{M+\delta}-B_{M+\delta}\right)+\frac{\delta}{M+\delta}\|h_\epsilon\|_\rho\\&\leqslant\tilde{\chi}_{\min}(M+\delta)\\&\quad+\frac{\delta}{M+\delta}(\|h_\epsilon\|_\rho-\tilde{\chi}_{\min}(M+\delta)),\quad M>0,\end{aligned} \tag{6.6}$$

show that $\tilde{\chi}_{\min}(M)$ is a Lipschitz continuous function of M. Therefore,it assumes any value between $\tilde{\chi}_{\min}(0)=\|h_\epsilon\|_\rho$ and $\tilde{\chi}_{\min}(\|f_t\|_\infty)$ exactly once. In particular, M_0 is the unique root of the equation

$$\tilde{\chi}_{\min}(M_0)=\chi_0\epsilon. \tag{6.7}$$

The function $f_e(\epsilon;h_\epsilon)$ realizes $\chi_{\min}(M_0)$. This ends the proof of Lemma 6.1.

We now prove a similar statement for problem (C) of the Introduction. Let

$$\alpha(M;\epsilon)=\sup\{f(x_0):\ f\in\mathscr{S}(M;\epsilon)\}, \tag{6.8}$$

$$\beta(M;\epsilon)=\inf\{f(x_0):\ f\in\mathscr{S}(M;\epsilon)\}, \tag{6.9}$$

where

$$\mathscr{S}(M;\epsilon)=\mathscr{F}_M\cap\{f:\ \chi(\rho;h-f)\leqslant\chi_0\epsilon,\ f\in L^2(\rho)\}. \tag{6.10}$$

The functions $\alpha(M;\epsilon),\beta(M;\epsilon)$ are, at fixed M, defined only for $\chi_0\epsilon\geqslant\tilde{\chi}_{\min}(M)$. We prove, namely, the following lemma.

Lemma 6.2: There exists a unique function in $\mathscr{S}(M;\epsilon)$ assuming the value $\alpha(M;\epsilon)$ at x_0. It is of the form $MB_M(z)$, with $B_M(z)$ a finite Blaschke product. The same is true for $\beta(M;\epsilon)$.

Proof: We consider the function

$$\bar{\chi}_{\min}(f_0;M)=\inf\{\chi(\rho;h-f):\ f\in\mathscr{F}_M;\ f(x_0)=f_0\}, \tag{6.11}$$

and shall show, similarly to Lemma 6.1, that it is a strictly monotonically increasing, Lipschitz continuous function of f_0, on the interval $[f_{0M}(x_0),M]$, where $f_{0M}(z)$ is the unique function of $\mathscr{F}_M$ on which $\tilde{\chi}_{\min}(M)$ is achieved. It is convenient to introduce

$$\hat{\chi}_{\min}(f_0;M)=\inf\{\chi(\rho;h-f):\ f\in\mathscr{F}_M;\ f(x_0)\geqslant f_0\} \tag{6.12}$$

and show first that it is attained on the same function as $\bar{\chi}_{\min}(f_0;M)$ (and thus has the same value). To this end, it is of advantage to switch to the space $L^2(\rho_c)$ (see Sec. V), where $f(x_0)=(\omega_0,f)_c$ and recall that (Ref. 31, p. 217, §8.3, Theorem 1) in view of the convexity of $\mathscr{F}_M$, of the functional $\chi(\rho;h-f)$ and of the constraint $f_0-(\omega_0,f)_c\leqslant 0$ (which is fulfilled at least by $f=M$) a Lagrange multipler $\lambda\leqslant 0$ exists so that

$$\begin{aligned}&\hat{\chi}_{\min}(f_0;M)\\&\quad=\inf\{\chi(\rho;h-f)+\lambda[(\omega_0,f)_c-f_0]:\ f\in\mathscr{F}_M\}.\end{aligned} \tag{6.13}$$

Further, for the extremal function $\bar{f}$, $\lambda[(\omega_0,\bar{f})_c-f_0]=0$. If $(\omega_0,\bar{f})_c\neq f_0$, we conclude that $\lambda=0$ which means $\hat{\chi}_{\min}(f_0;M)=\tilde{\chi}_{\min}(M)$. This is, however, impossible, by Theorem 2.2, if $f_0>f_{0M}(x_0)$, since $\tilde{\chi}_{\min}(M)$ is attained on a unique function $f_{0M}(z)$.

It is thus evident that $\bar{\chi}_{\min}(f_0;M)=\hat{\chi}_{\min}(f_0;M)$ and, in view of Theorem 5.2, that the unique extremal functions are the same. As in Lemma 6.1, it follows now that $\bar{\chi}_{\min}(f_0;M)$ is a strictly monotonically increasing, convex function of f_0, on $[f_{0M}(x_0),M]$. [Strict monotonicity follows from the uniqueness in $\mathscr{F}_M$ of the element achieving $\bar{\chi}_{\min}(f_0;M)$.] To show the Lipschitz continuity of $\bar{\chi}_{\min}(f_0;M)$, we consider two values $f_{01},f_{02},f_{0M}(x_0)\leqslant f_{01}<f_{02}\leqslant M$ and their corresponding extremal elements in $\mathscr{F}_{M;c}(x_0;f_{0i})$, $MB(z;f_{0i})$, $i=1,2$. Then, using the Schur–Pick–Nevanlinna algebra (Refs. 1 and 23), we may write ($i=1,2$)

$$B(z;f_{0i})=\frac{f_{0i}/M+\tilde{\beta}(z)B_i(z)}{1+f_{0i}\tilde{\beta}(z)B_i(z)/M} \tag{6.14}$$

with $\tilde{\beta}(z)=(z-x_0)/(1-zx_0)$ and $B_i(z)$ are Blaschke products. Then

$$\begin{aligned}\bar{\chi}_{\min}(f_{02};M)&<\chi\left(\rho;h-M\frac{f_{02}/M+\tilde{\beta}(x)B_1(x)}{1+f_{02}\tilde{\beta}(x)B_1(x)/M}\right)\\&\leqslant\chi(\rho;h-MB(x;f_{01}))\\&\quad+M\chi\left(\rho;\frac{f_{02}/M+\tilde{\beta}(x)B_1(x)}{1+f_{02}\tilde{\beta}(x)B_1(x)/M}\right.\\&\quad\left.-\frac{f_{01}/M+\tilde{\beta}(x)B_1(x)}{1+f_{01}\tilde{\beta}(x)B_1(x)/M}\right)\\&<\bar{\chi}_{\min}(f_{01};M)\\&\quad+|f_{02}-f_{01}|\frac{1+1/M}{[\min_{x\in[-a,b]}(1-|\tilde{\beta}(x)|)]^2},\end{aligned} \tag{6.15}$$

which shows the Lipschitz continuity of $\bar{\chi}_{\min}(f_0;M)$. With

the same reasoning, we establish that $\bar{\chi}_{\min}(f_0;M)$ is a strictly monotonically decreasing, Lipschitz continuous, convex function of f_0 on the interval $[-M, f_{0M}(x_0)]$. It follows that the equation for f_0,

$$\chi_{\min}(f_0;M) = \chi_0\epsilon \tag{6.16}$$

has, for each $\epsilon>0$, at most two distinct roots μ_1, μ_2, $-M\leqslant\mu_1\leqslant f_{0M}(x_0)\leqslant\mu_2\leqslant M$; we agree to let $\mu_1=-M$, if $\chi_0\epsilon\geqslant\chi(\rho;h(x)+M)$ and $\mu_2=M$, if $\chi_0\epsilon\geqslant\chi(\rho;h(x)-M)$. Now, if $f_{0M}(x_0)\leqslant f_{01}\leqslant\mu_2$, there exist functions $f\in\mathcal{F}_M$, $\chi(\rho;h-f)\leqslant\chi_0\epsilon$ assuming the value f_{01} at x_0, e.g., $MB(z;f_{01})$. However, if $f_{01}>\mu_2$, in view of the equality $\bar{\chi}_{\min}(M;f_0)=\hat{\chi}_{\min}(M;f_0)$, there exist no such functions. Thus, $\mu_2=\alpha(M;\epsilon)$ and, similarly, $\mu_1=\beta(M;\epsilon)$. This ends the proof of Lemma 6.2.

With this, problems (B) and (C) have been reduced to the solution of two equations, (6.7) and (6.16), with known right-hand sides. The evaluation of the left-hand sides for given M and f_0 require the numerical solution of problems (A) and (A_c). The Lipschitz continuity of $\bar{\chi}_{\min}(M)$ and $\bar{\chi}_{\min}(M;f_0)$, Eqs. (6.5) and (6.6) and (6.15), show that we can achieve any desired precision in the determination of the roots of (6.7) and (6.16) by solving problem (A) or (A_c) only for a finite set of values of M or of f_0. Thus, we have to describe in detail the numerical computation of $\tilde{\chi}_{\min}(M)$, $\bar{\chi}_{\min}(M;f_0)$, for given M and f_0.

Before turning to this, we make two remarks.

(a) There is no difficulty to allow for further constraints in problems (B) and (C), e.g., require their solution under the conditions $f(\xi_i)=w_i$, for several fixed points ξ_i and values w_i. One can always reduce the problem to a sequence of numerical solutions of problems (A_c).

(b) In a series of papers (see, e.g., Refs. 42–45), an equivalent method for the solution of problem (C) was used. One computes, for every fixed ϵ, the value

$$M_0(\epsilon;f_0)=\inf\{\|f\|_\infty : f\in H_R^\infty(D), \quad \chi(\rho;h-f)\leqslant\chi_0\epsilon;\ f(x_0)=f_0\}. \tag{6.17}$$

The set $[\beta(M;\epsilon),\alpha(M;\epsilon)]$ of possible values at x_0 is given by the set of values f_0 for which $M_0(\epsilon;f_0)\leqslant M$ (see, also Ref. 40). The curves $M_0(\epsilon;f_0)$ are sometimes quite spectacular, and may be used to indicate the quality of the extrapolation even if M is unknown. However, their computation requires more effort in our case than the solution of Eq. (6.16).

In principle, according to Theorem 4.1 and to the discussion of Sec. V, following Lemma 5.7, the solutions of problems (A) and (A_c) are equivalent to minimizations in a finite-dimensional space of parameters. It is shown in Ref. 40 that the dimension of this space may be chosen independently of the value of M. However, as shown on examples in Sec. VII, the bounds on the dimension of this space are considerably larger than the number of factors of the extremal function, which is observed in actual calculations. The method based on the solution of the nonlinear integral equation, Eq. (3.47) [or of the set (5.21) and (5.22)], turns out, however, to be quite efficient.

Before considering this, we discuss an apparently straightforward method of numerical solution, which uses Lemma 3.5. One tries, namely, to minimize directly the nonlinear functional of $N(x)\in L^2(\rho)$,

$$\chi^2(\rho;h-B)=(h(x)-B(N;x),h(x)-B(N;x))_\rho. \tag{6.18}$$

This functional is Fréchet differentiable at all $N(x)\in L^2(\rho)$, with the exception of those $N(x)$ for which zeros of $F_0(N;x)$, Eq. (3.36), cross $|z|=1$; there is no difficulty to account for the latter (see Lemma 3.7). One may then attempt to minimize (6.18) by the method of steepest descent (see Ref. 31, §10.5, Ref. 46). Since the action of $(\partial B(N;x)/\partial N)$ on $\delta N\in L^2(\rho)$ is obtained through the finite set of linear functionals $(\partial\alpha_i/\partial N)$ [Eq. (3.59)], the procedure is equivalent to a minimization with respect to the parameters α_i, whose number is varied in a controlled manner, when zeros of $F(N;x)$ cross $|z|=1$. Lemma 3.5 is invoked to make sure that we scan in fact the whole space of normals $n(x)\in L^2(\rho)$. This is an improvement of principle over the direct minimization described above. However, it has the drawback that it is not well defined everywhere in $L^2(\rho)$. In other words, if the Fréchet derivative of (6.18), which is the element of $L^2(\rho)$ given by

$$\frac{\partial\chi^2}{\partial N}(N;x) = -2\left(\frac{\partial B}{\partial N}(N;x)\right)^\dagger (h(\cdot)-B(N;\cdot)), \tag{6.19}$$

vanishes at some $N(x)$, it does not follow the corresponding $B(N;x)$ is the extremal Blaschke product. To understand this, we return to Eqs. (3.9) and (3.10) and remark that, replacing there $n(x)$ by $h(x)-B(N;x)$, problem (A) may be regarded as that of finding the unique number p and the unique set of p values for the α_i's so that (3.9) and (3.10) are satisfied. There are, in general, may solutions to (3.9) alone; all of them lead to vanishing $\partial\chi^2/\partial N(N;x)$, as is easily verified. The remarkable fact is that only one of them—the extremal function—satisfies (3.10), i.e., with the replacement above:

$$(h(x)-B(N;x),B(N;x)P(x;e^{i\theta}))_\rho\geqslant 0. \tag{6.20}$$

No similar problems of false extrema appear if one attempts a solution of the integral equation (3.47). This may be achieved by minimizing:

$$\Phi(N)=\|\mathscr{A}(N)(x)-h(x)\|_\rho^2. \tag{6.21}$$

The functional $\Phi(N)$ has only one finite local minimum, at the solution $N_0(x)$ of Eq. (3.47). Indeed, at any point $N_1(x)\neq N_0(x)$, the (Gateaux) derivative of $\Phi(N)$ in the direction δN is given by

$$\frac{\partial\Phi}{\partial N}(N_1)\delta N=2(\mathscr{A}(N_1)-h,\delta\mathscr{A}(N_1;\delta N)). \tag{6.22}$$

By Theorem 3.2, choosing either

$$\delta N_1=-\left(\frac{\partial\mathscr{A}}{\partial N}(N_1)\right)^\dagger(\mathscr{A}(N_1)-h) \tag{6.23}$$

or

$$\delta N_2=-\left(\frac{\partial\mathscr{A}}{\partial N}(N_1)\right)^{-1}(\mathscr{A}(N_1)-h), \tag{6.24}$$

the expression (6.22) is strictly negative. (We have used the symbol $\partial\mathscr{A}/\partial N$ even at points where it is not defined; however, there is no ambiguity, in view of Lemma 3.7.) Consequently, there exist points $N(x)$ near $N_1(x)$, with $\Phi(N)$

$<\Phi(N_1)$, as asserted. The choice (6.23) corresponds to the method of steepest descent and (6.24) to Newton's method (see Ref. 31, Chap. X, Ref. 47, Chap. XV and XVIII, Ref. 48). In an actual calculation, for a given $N_1(x)$ and a given choice of δN_i, we locate numerically the minimum of the function of one variable $f(\lambda) = \Phi(N_1 + \lambda \delta N_i)$, obtain this way a new point $N_2(x)$, and continue in this manner. If all steps of the minimization lie inside a domain in which a uniform lower bound on $\|(\partial \mathscr{A}/\partial N)(N)\delta N\|_\rho$, for $\|\delta N\|_\rho \geqslant 1$ or a uniform upper bound on $\|((\partial \mathscr{A}/\partial N)(N))^{-1}\|_\rho$ is known, and if the distance of the points N_k from the separating surface $F(N;e^{i\theta}) = 0$ can be controlled, we can show that this procedure converges to the solution of Eq. (3.47). Indeed (see Ref. 31, p. 289, for similar reasoning), if we asume that $\|((\partial \mathscr{A}/\partial N)(N_k))^{-1}\|_\rho < K$, independently of k, and that Taylor's second-order formula (see Ref. 36, p. 77, Theorem 5.6.1) may be applied [which ignores the discontinuities of $(\partial \mathscr{A}/\partial N)$] we would have at the $(k+1)$st step for the choice (6.24) the majorization

$$\Phi\left(N_k - \lambda\left[\left(\frac{\partial \mathscr{A}}{\partial N}\right)(N_k)\right]^{-1}(\mathscr{A}(N_k) - h)\right)$$
$$\leqslant \Phi(N_k) - \lambda\Phi(N_k) + \lambda^2 BK^2\Phi(N_k), \qquad (6.25)$$

where B is an upper bound on the second derivative. Choosing $\lambda = 1/2BK^2$, Eq. (6.25) implies

$$\Phi(N_{k+1}) - \Phi(N_k) \leqslant -\Phi(N_k)/(2BK^2). \qquad (6.26)$$

Now, the condition $\Phi(N_k) > 0$, for all k and Eq. (6.26) are consistent only if $\lim_{k\to\infty}\Phi(N_k) = 0$. Further, this fact and the existence of the inverse Fréchet derivative $((\partial \mathscr{A}/\partial N)(N))^{-1}$ show that the corresponding sequence $N_k(x)$ tends to $N_0(x)$ in $L^2(\rho)$ norm, if $F(N_0;z) \neq 0$, for $|z| = 1$. In Appendix B [comments (a) and (b)], we show that, even if the latter is not the case, the convergence $\Phi(N_k) \to \Phi(N_0)$ implies $N_k(x) \to N_0(x)$ in $L^2(\rho)$. The calculations we present in Sec. VII are based on the algorithm described above. Unfortunately, as we have pointed out, the reasoning leading to (6.26) is not complete. We can turn nevertheless the arguments above into rigorous convergence statements provided we restrict $h(x)$ to a sufficiently small neighborhood of $\mathscr{A}(N_1)$ and we slightly modify the prescription (6.24) for the minimization step. This is done in Appendix D and suffices in fact to justify the systematic use of the integral equation (3.47) for the solution of problem (A). Indeed, it is enough to find solutions, within a sufficiently good approximation, for functions $h_{\mu_i}(x) = (1-\mu_i)\mathscr{A}(N_1(x)) + \mu_i h(x)$, for a finite number of μ_i, lying close enough to each other. This allows us to formulate the conclusion of this section in the following theorem.

Theorem 6.1: The algorithm based on the minimization of $\Phi(N)$ with the choice (6.4) for the direction of minimization and the modifications of Appendix D provides a sequence of Blaschke factors $B(N_k;x)$, uniformly convergent in $|z| \leqslant 1$ to $f_0(x)$, the solution of problem (A), provided $\Phi(N_1)$ is sufficiently small.

It is possible, although somewhat laborious, to obtain a quantitative characterization of this convergence and we skip this point. In view of Theorem 5.3, a statement analogous to Theorem 6.1 may obviously be formulated for problem (A_c).

VII. NUMERICAL EXAMPLES AND CONCLUSIONS

We consider 15 points $\{x_i\}_{i=1}^{15}$, distributed equidistantly on $[-a,a]$, $a>0$. Usually, we shall take $a = 0.5$. At these points, we prescribe a data function $h(x)$ with constant errors, of magnitude σ. The data function is obtained by perturbing the values $f(x_i)$ of a certain $f(z) \in \mathscr{F}_1$ with random numbers obeying a Gaussian distribution with standard deviation σ. We wish to obtain a numerical feeling for the procedures described in Secs. III–VI.

The bounds on the numbers of zeros of the best Blaschke product (Sec. IV) involve quantities C_R, C_I, and γ (Lemma 4.1) which are purely geometrical (except for the ρ-dependence in C_R and C_I). Their computation requires knowledge of $\overline{C}_R(z)$, Eq. (4.7), along a closed curve $\mathscr{C}$, surrounding the data domain and lying in $|z|<1$ and on a closed curve $\mathscr{C}'$, situated in $|z|>1$ and avoiding $(-\infty, -1/a] \cup [1/a, \infty)$. It is convenient to choose $\mathscr{C}$ and $\mathscr{C}'$ as level lines $|\zeta(z)| = \text{const}$ of the mapping $\zeta(z)$ leading from the plane cut along $(-\infty, -1/a] \cup [-a,a] \cup [1/a, \infty)$ to an annulus with radii 1, R in the ζ plane (as obtained by means of the incomplete elliptic integral[9,10]). As is well known, R is a monotonically decreasing function of a, $R(a) \to 1$, as $a \to 1 (a>1)$. The computation of the Green's function and thus of the bounds in Lemma 4.1 for a circular crown with radii R_1, R_2, $1 < R_1 < R_2 < R$, is straightforward. On one hand, the bound γ_1, Eq. (4.13), decreases quickly towards zero as $R_2/R_1 \to 1$, but on the other hand, the functions $\overline{C}_R(z)$, $\overline{C}_I(z)$, Eq. (4.7), for z on $\mathscr{C}$, $\mathscr{C}'$, increase indefinitely as R_1 approaches unity or R_2 approaches R. Thus, a balance has to be achieved and, as a rule, it is profitable to compute the bounds for $R_1 = 1.1$, $R_2 = R/1.1$. The constants C_R, C_I are then of the order of unity $(1 < C_R, C_I < 10)$. The dependence on $\rho(x)$ (in particular on the number of points on $[-a,a]$) is negligible (a few percent for $N \gtrsim 10$).

From Eq. (4.30), one sees that, if $\gamma \ll 1$, the bound on the number of zeros is likely to be good. From Table I we see that γ is depressingly close to unity if $a \gtrsim 0.5$ so that the bound gets very weak for increasing a. The values of k_{i0} depend on $h(x)$ but are, in general, much smaller than unity (see Table I). The operators B_i, Eq. (4.21), are again determined (apart from a weak ρ-dependence) on geometrical grounds; they have eigenvalues which decrease rapidly to zero; as a consequence, $\alpha(g)$, Eq. (4.24), is, for "usual" g's, very small [$\sim 10^{-7}$ for (7.1)] and only the estimate of k_{i0} obtained by the method used in Lemma 4.2 for $\alpha(g) = 0$ is of relevance.

In Table I, we show the values of the upper bound Eq. (4.30) on the number of zeros of the best fit to the data function obtained by perturbing

$$f_1(z) = 1/(z^2+2), \qquad (7.1)$$

for $z \in [-a,a]$ with noise. The interesting feature is the rapid increase of the bound with a. For small a, the bound (4.30) turns out to be tight. Its rapid deterioration may be seen as an expression of the "thickening" of the body $\mathscr{S}_N$ as a increases; the correlation of the values of $f(z) \in \mathscr{F}_1$ at two different points becomes weaker as the distance between the points increases. We recall that the derivation of the bound (4.30) was done neglecting the second term in the duality

TABLE I. Numerical values for the quantities appearing in Sec. IV and the upper bound $N_{\max}$ for the number of zeros of the best fit; the data function was obtained by perturbing example (7.1) with 5% errors [of $f(0)$]. The data are given at 15 points distributed equidistantly on $(-\alpha,\alpha)$; the constants C_i are defined following Eq. (4.16), γ following (4.13), and k_{01}, k_{02} are defined in Lemma 4.3. The bound $N_{\max}$ is essentially controlled by γ, which is of purely geometrical origin.

α	χ_1	C_1	C_2	$\ln k_{10}/C_1$	$\ln k_{20}/C_2$	$\ln \gamma$	$N_{\max}$
0.1	0.0199	0.21	4.74	−8.47	−8.74	−1.48	6
0.2	0.0192	0.215	1.93	−7.91	−7.94	−0.82	10
0.3	0.0180	0.240	1.42	−11.03	−13.42	−0.51	27
0.5	0.0162	0.302	1.13	−10.69	−11.73	−0.18	63
0.7	0.0176	0.353	0.70	−10.60	−11.77	−0.044	260

relation (2.14); the latter describes the "thickness" of $\mathscr{S}_N$ [in the direction $n(x)$].

Numerical experience shows, however (Refs. 1, 16, and 20), that even for larger a, the number of zeros of the best fit does not increase, in general, as badly as indicated in Table I. The method of determining the best fit by means of the nonlinear integral equation (3.47) [or of the minimization of $\Phi(N)$, Eq. (6.21)] allows a control over the number of zeros at each step of the iteration. A possible starting point for the minimization of $\Phi(N)$, Eq. (6.21), is the function $N_0(x)$ with

$$N_0(x) = h(x) - F_0(x) \tag{7.2}$$

and $F_0(x)$ the best fit to $h(x)$ under the L^2 condition $\|f\|_2 \leqslant 1$, Eq. (1.5). If the direction of $N(x)$ is fixed, one can still minimize easily $\Phi(N)$ as a function of the magnitude of $N(x)$. If $(N(x), h(x) - B(N;x))_\rho > 0$, the minimal value of $\Phi(N)$ is obtained for $N' = \cos\theta_0 \times N$ and is equal to

$$\Phi(N') = \|h(x) - B(N;x)\|^2 \sin^2\theta_0 \tag{7.3}$$

with θ_0 the angle between N and $h(x) - B(N;x)$. We start the minimization from this value of N (after having verified that $\cos\theta_0 > 0$). If $\|N(x)\|_\rho$ is small and thus also $E_0(N;x)$, $-a < x < a$, the first term in the Fréchet derivative $\partial\mathscr{A}/\partial N$, Eq. (3.73) is small numerically compared to the second one. The operator $(\partial B/\partial N)(N)$ is from $L^2(\rho)$ into R^p and has no inverse. Thus, although $(\partial\mathscr{A}/\partial N)^{-1}$ exists, its conditioning may be not very good (although manageable). We start from the function $f_1(z)$, Eq. (7.1), and generate $h(x)$ as explained above, choosing $\sigma = 0.05\, f_1(0)$. It follows that $d\rho(x)/dx = 1/(15\sigma^2)\sum_{i=1}^{15} \delta(x - x_i)$. We wish to find the set of allowed values of the functions $f \in \mathscr{F}_1$ and obeying $\chi(\rho; h - f) \leqslant 1$, at the points $x_1 = 0.6$, $x_2 = 0.7$, and $x_3 = 0.8$. The procedure described in Sec. VII was applied and the results are shown in Fig. 1 as curves $\chi_{\min}(1; f_{0i})$ vs f_{0i}. The precision can be made arbitrarily high. The function $f_1(z)$ has no zeros in $|z| < 1$; its best approximants in $\mathscr{F}_1$, with fixed values f_{0i} at x_i, are Blaschke products with four to six factors. The construction of the curves is quickly convergent, once a run has led us from the L^2 solution to a solution of problem (A_c), for a certain f_{0i}.

To sum up, we have presented a systematic method to solve the problem of determining the best analytic approximant in the usual least squares sense, within the class of functions uniformly bounded and holomorphic in the unit disk. The essential point of the method is the solution of a certain nonlinear integral equation, which leads itself to an approximation scheme, related to the Newton–Kantorowich procedure.[47]

The geometrical meaning of the integral equation is clearest in the form (3.1). The equation may be obviously written for many other problems which require the determination of the minimal distance to a convex set $S \in L^2(\rho)$. It gives the solution in all cases when the "linear" problem of finding the element $f(n;x)$ which realizes $\sup\{(n,f), f \in S\}$ may be solved for all $n(x)$. In fact, we have shown that, if the nonlinear dependence on $n(x)$ of $f(n;x)$ is known, and if its Fréchet derivative with respect to $n(x)$ is a compact operator, then the operator

$$\mathscr{A}(n(x)) = n(x) + f(n;x), \tag{7.4}$$

has a Fréchet derivative with a bounded inverse. Indeed, this follows from the convexity argument of Theorem 3.2. In problems of analytic continuation and interpolation, the set S is also compact in $L^2(\rho)$, so that $f(n;x)$ is a compact

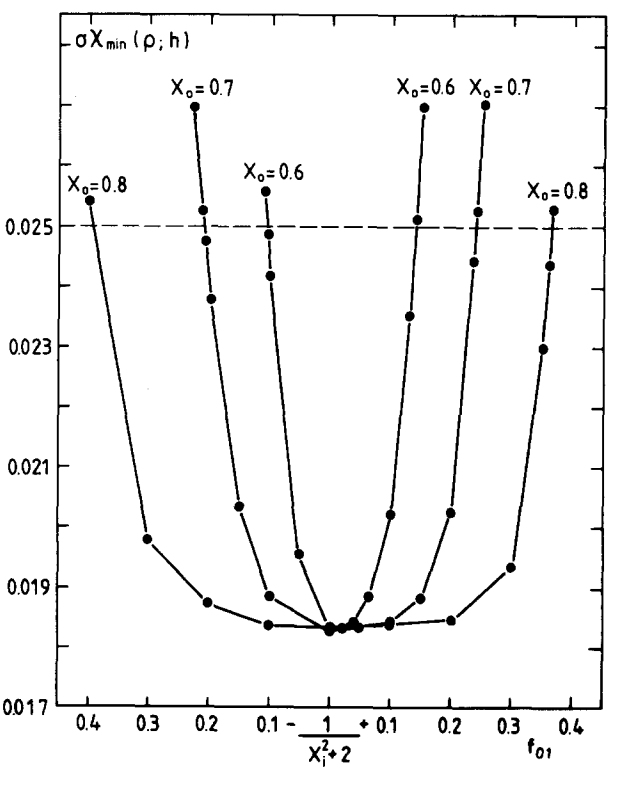


FIG. 1. Curves $\chi_{\min}(1; f_{01})$ as a function of f_{01} for the data function obtained from (7.1). The curves are obtained by solving the set of integral equations (5.24) and (5.25) with the procedure described in Sec. VI. The set of possible values at $x_{01} = 0.6, 0.7$, and 0.8 in turn is given by the segment delimited by these curves on the line $\chi_{\min} = 0.025(f(0) \times 0.05)$.

operator and, crudely speaking, we expect also its Fréchet derivative to be so. As an example, consider the H^2 problem, treated in Lemma 4.4. The analog of Eq. (3.1) is (using Lemma 3.3)

$$N(x)+\frac{1}{\|k_N\|_2}\int_{-a}^{b}\frac{N(x')}{1-x'x}\,d\rho(x')=h(x). \tag{7.5}$$

If we assign a certain value λ to $\|k_N\|_2$, Eq. (7.5) is Fredholm of the second kind, without positive eigenvalues. The difficulty is, clearly, that, given $N_\lambda(x)$ for a certain choice of λ, we are not sure that $\|k_{N_\lambda}\|_2=\lambda$. Thus, in principle, we have to solve the equation for all $\lambda>0$ and find that value λ for which $\|k_{N_\lambda}\|_2=\lambda$. This is implicitly done in Lemma 4.4. On the other hand, the Fréchet derivative of the left-hand side of Eq. (7.5) is [its action on $\delta N(x)\in L^2(\rho)$]

$$\frac{\partial\mathscr{A}}{\partial N}(N)\delta N=\delta N(x)+\frac{1}{\|k_N\|_2}\int_{-a}^{b}\frac{d\rho(x')}{1-x'x}-\frac{(\delta N,F(N;x))_\rho}{\|k_N\|_2^3} \tag{7.6}$$

and is obviously the identity plus a compact operator from $L^2(\rho)$ into $L^2(\rho)$. Thus, we can achieve the same ends as in Lemma 4.4 by solving (7.5) as a nonlinear integral equation by the Newton method. This is too complicated in this case, but is of general applicability and may be used in other situations when $f(n;x)$ is known (see below).

In the problem treated in this paper, the dependence $f(n;x)$ was, however, not known; most of the work in Sec. III and V was then dedicated to showing that we can indeed "scan" the whole space $L^2(\rho)$ of functions $n(x)$ and generate the corresponding extremal functions $f(n;x)$ (Blaschke products) by varying another function $N(x)$, related to $n(x)$ by Eq. (3.35). The function $N(x)$ gives also rise, via Eq. (3.36) to the Blaschke product associated to $n(x)$. The fact that the correspondence $n(x)\leftrightarrow N(x)$ is invertible and with an invertible Fréchet derivative (with the exceptions of Lemma 3.6) appeared to the author as remarkable.

Further, the generalization of these results to the case when several values $f(\xi_i)=w_i$ are given in advance—as is the case in any application—is not obvious; the most intricate part is contained in the proof of Theorem 5.3.

One should also mention that a pleasant feature of the integral equation (3.47) is the virtual independence of its method of solution on the number N of experimental points. Indeed, for continuous data distributions, the only limitation is given by the precision of the solution of a Fredholm equation of the second kind [i.e., the construction of $((\partial\mathscr{A}/\partial N)(N))^{-1}(\mathscr{A}(N)-h)$].

Condition (H), Sec. III, restricts the set of vectors $N(x)$ through which the iteration may proceed. The set of excluded points may be regarded as exceptional; it never occurred in our numerical experiments. Further (see Appendix E), part of (H) may presumably be relaxed, without any change of the results. It is of interest to understand how the solution of $\mathscr{A}(N)=h$ may be recovered, if $F_0(N;z)$ has a multiple zero on $|z|=1$.

Clearly, the interest in the results of this paper lies mostly in their exact character. Also, one may presumably state that problems (B) and (C) of the Introduction represent analytic continuation problems in a "pure" form: if a model for the stabilizing condition is available, then it is likely to take the form $|f(e^{i\theta})|\leqslant M(\theta)$, which, by means of (1.10), is the same as the one treated here. Unfortunately, there are few problems in high energy physics (the one treated in Ref. 1 is an exception) where this stabilizing condition is known with sufficient confidence.

However, a problem closely related to the present one is that of the inversion of a large number of moments of a function $f(x)$, positive definite on some interval (a,b), if the moments are affected by errors. The usual methods of moment inversion rely on the assumption that the moments are consistent with the positivity constraint on $f(x)$. A very small amount of noise invalidates this assumption, however, in the same manner as discussed in Sec. II for the body $\mathscr{S}_N$. The best fit to error affected moments is then, in analogy to Sec. II, given by the moments generated by a function $f_0(x)$ consisting of a few positive δ-functions placed on (a,b). The problem arises to find a systematic method for the determination of the position of their support. An equation similar to (3.1) seems to be appropriate. The author hopes to return to these questions in the future.

ACKNOWLEDGMENTS

The author became aware of the special role played by Blaschke products in problems (A)–(C) of this paper during his collaboration with H. Raszillier and W. Schmidt several years ago. He also profited very much from the interest D. Atkinson had in the problem of analytic continuation at that time and from his encouragement.

APPENDIX A: ON THE NUMBER OF ZEROS OF THE EXTREMAL BLASCHKE PRODUCT $f(n;z)$

In this Appendix, we are concerned with a measure $\rho(x)$ of finite type, with N jumps at the points $x_i, x_1>x_2>x_3>\cdots>x_N$. The possible values $f_1,f_2,\ldots,f_N$ assumed by functions in $\mathscr{F}_{1;\infty}$ at $x_1,x_2,\ldots,x_N$ make up the convex and closed set $\mathscr{S}_N$, Eq. (2.5), in R^N. By means of the Cayley transformation

$$w_i=(1+f_i)/(1-f_i), \tag{A1}$$

the points of $\mathscr{S}_N$ are placed in one-to-one correspondence with those of the cone K_N of points with coordinates given by the possible values assumed at $\{x_i\}_{i=1}^N$ by the functions holomorphic and with positive real part in $|z|<1$. Using the results of Ref. 22 (Chap. VI), the positivity of the determinant (2.10) and of its minors and the fact that the transformation (A1) is monotonical, i.e., leads a set $f_{i-}<f_i<f_{i+}$, $i=1,\ldots,N$, into a set $w_i(f_{i-})<w_i<w_i(f_{i+})$, one verifies that (a) the boundary of $\mathscr{S}_N$ consists of two $(N-1)$-dimensional manifolds ("surfaces") in R^N, generated by all Blaschke products with precisely $N-1$ zeros and their common closure, generated by products with less than $(N-1)$ zeros; (b) if two points with coordinates $(f_1,\ldots,f_N)$, $(f_1',\ldots,f_N')$ belong to $\mathscr{S}_N$ and

$$(-1)^{N-k}f_k<(-1)^{N-k}f_k',\quad k=1,\ldots,N,$$

then the whole set of points with coordinates $\bar{f}_k$,

$(-1)^{N-k}f_k \leqslant (-1)^{N-k}\bar{f}_k \leqslant (-1)^{N-k}f'_k$ also belongs to $\mathscr{S}_N$; and (c) the planes tangent to $\mathscr{S}_N$ at the points of the two surfaces are given by

$$\sum_{i=1}^{N} n_i (f_i - f_{i0}) = 0, \tag{A2}$$

where the $\{n_i\}_{i=1}^N$ have alternating signs (see Ref. 22, §1.5, Theorem 5.3). There are several planes tangent to $\mathscr{S}_N$ at the points of the common closure of the two surfaces ("edges," see, also, Ref. 17).

As we have seen in Sec. II, for large N, $\mathscr{S}_N$ gets very flattened. Consider then a point $\{h_i\}_{i=1}^N \in \text{int}\,\mathscr{S}_N$; we may construct the largest parallelepiped $(\underline{h},\bar{h})$ in R^N that is contained in $\mathscr{S}_N$, so that $\underline{h}_k(-1)^{N-k} \leqslant h_k(-1)^{N-k} \leqslant \bar{h}_k(-1)^{N-k}$. If we wish to extrapolate the "data" $\{h_k\}_{k=1}^N$ with "errors" $(\underline{h}_k,\bar{h}_k)_{k=1}^N$ to a point $x \in (-1,1)$, $x \neq x_i$, $i = 1,2,...,N$, we may apply the method described in Sec. II, in relation to Eq. (2.8). Assume now $\{h_i\}_{i=1}^N$ is affected by errors $(-1)^{N-i}\lambda\epsilon_i, \lambda, \epsilon_i > 0$, with increasing λ. Already for very small λ, the point $\{h_i\}_{i=1}^N$ crosses the surface of $\mathscr{S}_N$ and the problem of extrapolation is reduced to that of finding $\chi^2_{\min}(\rho;h)$, Eq. (1.3). For small λ, the latter is realized by a point on the surface that has been crossed, i.e., by a Blaschke product with $N-1$ zeros. The normal $\{n_{i,0}\}_{i=1}^N$ of the plane tangent to $\mathscr{S}_N$ at $f(n_0;z)$ has, as mentioned above, components with alternating signs. Thus, $f(n_0;z)$ has as many factors as there are sign variations of $\{n_{i,0}\}_{i=1}^N$. Now, if λ increases or if we choose a different direction of displacement, the point $\{h_i\}_{i=1}^N$ will in general move outside of $\mathscr{S}_N$ in such a way that its minimal distance to $\mathscr{S}_N$ is attained on an "edge," i.e., on a Blaschke product with less than $N-1$ factors. It is amusing that some relation persists between the number of factors of $f(n_0;z)$ and that of the sign variations of the corresponding $\{n_{i,0}\}_{i=1}^N = \{h_i - f(n_0;x_i)\}_{i=1}^N$. Namely, for *any* normal $\{n_i\}_{i=1}^N$.

Theorem A1: The number of zeros of the extremal Blaschke product $f(n;z)$ is at most equal to the number of sign variations of the normal $\{n_i\}_{i=1}^N$.

Proof: We show first a similar statement for the Cayley transform (A1). We recall (see Ref. 22, Chap. IV) that the boundary of K_N is made up of points with components w_i^b, given by ($k \leqslant N$, Ref. 22)

$$w_i^b = \sum_{j=1}^{k} P(x_i;e^{i\theta}j)\lambda_j \equiv \oint P(X;e^{i\theta})d\sigma_b(\theta) \tag{A3}$$

with $\lambda_j > 0$; the angles $\theta_j \in [0,2\pi)$ are the zeros of a polynomial

$$\mathscr{P}(\bar{n};e^{i\theta}) = \sum_{i=1}^{N} \bar{n}_i P(x_i;e^{i\theta}), \tag{A4}$$

which is positive on $[0,\pi]$. Such a polynomial is, by reflection symmetry, in fact positive on $[0,2\pi)$, so that all its zeros are at least double. In Eq. (A4) the $\bar{n}_i$'s are the components of the normal of a plane tangent to K_N at $\{w_i^b\}_{i=1}^N$. With this, we claim that the number of distinct zeros of $\mathscr{P}(\bar{n};e^{i\theta})$ on $0 \leqslant \theta < 2\pi$ cannot exceed the number of sign variations of the set $\{\bar{n}_i\}_{i=1}^N$.

To see this, we notice that the function $\mathscr{P}(\bar{n};z)/z$ has $2N$ poles at x_i and $1/x_i$, with residua equal to $\bar{n}_i$; it vanishes twice at infinity and has $2N-2$ further zeros either falling in pairs $(z_i, 1/z_i^*)$ or lying on $|z| = 1$. If $\bar{n}_i\bar{n}_{i+1} > 0$, there must be a zero lying between x_i and x_{i+1} and therefore a zero between $1/x_{i+1}$ and $1/x_i$. Consequently, if $V(\bar{n})$ is the number of sign variations of $\{\bar{n}_i\}_{i=1}^N$, the number of zeros lying on $[x_N,x_1]$ and its reflection across $|z| = 1$ is $2N-2-2V$. Thus, at most $V(n)$ double zeros lie on $|z| = 1$, which proves our claim.

Let the coordinates of a point in R^N be $(f_1,...,f_N)$. Then a plane tangent to $\mathscr{S}_N$ at $(f_1^0, f_2^0,...,f_N^0)$ may be written in the form (A2). If $\{f_i\}_{i=1}^N \in \mathscr{S}_N$, $f_i \neq f_i^0$, then

$$\sum_{i=1}^{N} n_i f_i < \sum_{i=1}^{N} n_i f_i^0. \tag{A5}$$

From (A1) it follows that all points of K_N with coordinates $\{w_i\}_{i=1}^N$ obey $[w_i^0 = w_i(f_i^0), w \neq w^0]$

$$S(w,w^0) \equiv \sum_{i=1}^{N} n_i \left(\frac{w_i - 1}{w_i + 1} - \frac{w_i^0 - 1}{w_i^0 + 1}\right) < 0. \tag{A6}$$

Consider now the plane tangent to $S(w,w^0) = 0$ at w^0,

$$T(w,w^0) \equiv \sum_{i=1}^{N} \frac{2n_i}{(1+w_i^0)^2}(w_i - w_i^0) = 0. \tag{A7}$$

We claim that, in fact, $T(w,w^0) = 0$ is tangent to K_N at w_0. To this end, consider any point $\{w_i\}_{i=1}^N \in K_N$ and join it to $\{w_i^0\}_{i=1}^N$ by a line segment, which is contained in K_N. All points of the line obey

$$\tilde{S}(\lambda) \equiv S(\lambda w + (1-\lambda)w^0, w^0) < 0. \tag{A8}$$

This implies that, for any $w \in K_N$ $[\tilde{S}(0) = 0]$

$$0 \geqslant \frac{d\tilde{S}}{d\lambda} = T(w;w^0). \tag{A9}$$

Thus, $T(w;w^0) = 0$ leaves K_N on one side and touches it at least at w^0. In fact, considering the ray $\lambda w^0 \in K_N$, $\lambda > 0$, it is easy to see that

$$\sum_{i=1}^{N} n_i' w_i^0 \equiv \sum_{i=1}^{N} \frac{2n_i}{(1+w_i^0)^2} w_i^0 = 0, \tag{A10}$$

so that $T(w;w_0) = 0$ passes through the origin in w-space. It follows that the polynomial $\mathscr{P}(n_i';z)$ has at most $V(n)$ double zeros on $|z| = 1$ and that the function $\sigma_b^0(\theta)$ generating $w_i^0 \in K_N$ by means of (A3) has at most $V(n)$ positive jumps on $[0,2\pi)$ and is otherwise constant. The Cayley transform (A1) shows that the phase variation of the Blaschke product $f(n;z)$ with $f(n;x_i) = f_i^0$ is then at most $2\pi V(n)$, and this ends the proof of Theorem A1. For the extremal $f(n_0;x)$ this means the number of its zeros is at most equal to the number of its oscillations around $h(x)$.

APPENDIX B: COMPLEMENT TO SEC. III

In this Appendix, we wish to show that $((\partial\mathscr{E}/\partial N)(\tilde{N}))_e$, $((\partial\mathscr{E}/\partial N)(\tilde{N}))_i$, Eq. (3.66), have bounded inverses. It is easy to see that both $(\partial\mathscr{E}/\partial N)_e$, $(\partial\mathscr{E}/\partial N)_i$ are the sum of an operator with bounded inverse $[E(\tilde{N}) \times \text{identity}]$ and a compact operator, i.e., they are of Fredholm type. For our statement, it is thus enough to argue that each of the equations

$$\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_e (\delta N) = 0, \quad \left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_i (\delta N) = 0 \tag{B1}$$

implies $\delta N = 0$. We proceed by analogy to the proof of Lemma 3.6 and show first that (B1) imply, for their possible solutions $\delta N \in L^2(\rho)$, in turn,

$$\left(\frac{\partial B}{\partial N}\right)_e (\delta N) = 0, \quad \left(\frac{\partial B}{\partial N}\right)_i (\delta N) = 0. \tag{B2}$$

Equations (B2) are obtained by considering in turn the limits $N_k^e \to \tilde{N}$, $N_k^i \to \tilde{N}$ of the identities

$$\frac{\partial L}{\partial N}(\mathscr{E}(N_k^{e,i}); \alpha_j(N_k^{e,i}))(\delta N) = 0, \tag{B3}$$

valid for every $\delta N \in L^2(\rho)$ [cf. Eq. (3.58)]. If δN satisfies, say, the second of the equations (B1), then (B3) implies, similarly to Lemma 3.5, the vanishing of the quadratic form

$$Q_i(\tilde{N};\delta\alpha) = \sum_{i,j}^{p+2} \left(n(\tilde{N}), \left(\frac{\partial^2 B}{\partial\alpha_i \partial\alpha_j}\right)_i (\tilde{N})\right)_\rho \delta\alpha_i\, \delta\alpha_j = 0. \tag{B4}$$

In (B4), we have considered the case of two complex zeros α_{p+1}, $\alpha_{p+2} = \alpha_{p+1}^*$ approaching $|z| = 1$, and have used the fact that the zeros α_i have a continuous Fréchet derivative with respect to $N(x)$. [This follows from the assumption that $F(N;z)$ has only simple zeros in $|z| \leqslant 1$.]

To evaluate (B4), we consider the limit $N_k^i \to \tilde{N}$, at fixed directions β_i, of the identity (3.65). The left-hand side obviously has a limit [as in (B4)]. On the right-hand side, we obtain again the modulus squared of $E(N;e^{i\theta})$ times the analytic continuation to $|z| = 1$ of $[(\partial B/\partial N)(\tilde{N})]_i(\delta N)(x)$. If δN exists so that $[(\partial\mathscr{E}/\partial N)(\tilde{N})]_i(\delta N) = 0$, but $[(\partial B/\partial N)(\tilde{N})]_i(\delta N) \neq 0$, we obtain a contradiction with (B4). Thus, the second equation (B2) must be valid, if (B1) is true. The same argument shows that the first equation (B1) implies the first of (B2).

Now, the condition $(\delta B)_i(x) = 0$ does not imply that, for all k, $1 \leqslant k \leqslant p+2$, $\delta\alpha_k = 0$. Indeed, the functions $P(x;\alpha_{p+1})$, $P(x;\alpha_{p+2})$ are identical if $\alpha_{p+1} = \alpha_{p+2}^*, |\alpha_{p+1}| = 1$. Since the functions $\{P(x;\alpha_i)\}_{i=1}^{p+1}$ are, however, linearly independent [see Eq. (2.12)], we conclude that $(\delta B)_i(x) = 0$ means $\delta\alpha_i = 0$ only for $i = 1,2,\ldots,p$ but still allows $\delta\alpha_{p+1} \neq 0$ if $\mathrm{Re}(\delta\alpha_{p+1}^* \alpha_{p+1}) = 0$. No such possibility is allowed if α_{p+1} is real.

We consider next the representation (3.39) and (3.40) for $L(n;z)$, and take at fixed z, the Fréchet derivative with respect to $N(x)$; it is a continuous function of $N(x)$, even at points like $\tilde{N}(x)$, with $F(\tilde{N};e^{i\theta})$ vanishing for some θ. Taking limits $N_k^i \to \tilde{N}$, $N_k^e \to \tilde{N}$, one easily verifies that

$$\begin{aligned}\frac{\partial L}{\partial N}(\tilde{N};z)(\delta N) &= \left(\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_i (\delta N)(x), B(x;\alpha)P(x;z)\right)_\rho \\ &\quad + \left(n(x), \left(\frac{\partial B(\tilde{N})}{\partial N}\right)_i (\delta N)(x) P(x;z)\right)_\rho \\ &= \left(\left(\frac{\partial\mathscr{E}}{\partial N}(\tilde{N})\right)_e (\delta N)(x), B(x)P(x;z)\right)_\rho \\ &\quad + \left(n(x), \left(\frac{\partial B}{\partial N}(\tilde{N})\right)_e (\delta N)(x) P(x;z)\right)_\rho .\end{aligned} \tag{B5}$$

In view of (B1) and (B2), it follows that the possible solutions $\delta N \in L^2(\rho)$ of any of the equations (B1) satisfy, for all z,

$$\frac{\partial L}{\partial N}(\tilde{N};z)(\delta N) = 0. \tag{B6}$$

We now show that (B6) implies, however, for all z

$$\frac{\partial F_0}{\partial N}(\delta N)(z) = 0. \tag{B7}$$

The latter, in view of the representation (3.50) means $\delta N(x) = 0$, a.e.-ρ and proves our point.

To show (B7), we consider Eq. (3.39) for $L(n;z)$ [cf. Eq. (3.40)] and write, for $z = e^{i\theta}(|\alpha_{p+1}(\tilde{N})| = 1)$,

$$\begin{aligned}L(n(\tilde{N});z) \equiv{}& e^{-2i\theta}(e^{i\theta} - \alpha_{p+1})^2 \\ &\times (e^{i\theta} - (1/\alpha_{p+1}))^2 L_1(n(\tilde{N});e^{i\theta}).\end{aligned} \tag{B8}$$

We also write

$$F_0(\tilde{N};z) = B(\tilde{N};z)(z - \alpha_{p+1})(z - (1/\alpha_{p+1}))E_1(\tilde{N};z), \tag{B9}$$

with $E_1(\tilde{N};z) \neq 0$ in $|z| \leqslant 1$. From (B8) and (B9), we deduce that [$\tilde{n} = n(\tilde{N})$]

$$|E_1(\tilde{N};e^{i\theta})|^2 = L_1(\tilde{N};e^{i\theta}). \tag{B10}$$

We now compute the variation of the double zeros at $\tilde{N}$, if δN is constrained by (B6). A double zero is generated by the two coincident zeros $\alpha_{p+1}(\tilde{N})$, $\alpha'_{p+1}(\tilde{N}) = 1/\alpha_{p+1}^*(\tilde{N})$ of $F_0(\tilde{N};z)$, $F_0^*(\tilde{N};1/z^*)$, each of which is Fréchet differentiable with respect to N. We obtain a constraint on the individual variations of these zeros by considering those of their symmetric combinations, e.g., their sum

$$\delta\alpha_{p+1} + \delta\alpha'_{p+1} = \frac{1}{2\pi i}\oint z\delta\left(\frac{\partial L/\partial z}{L}(\tilde{N})\right)(z)dz = 0, \tag{B11}$$

where the integration is carried out on a contour enclosing α_{p+1} and no other zero of L, and we have used (B6). However, $\delta\alpha'_{p+1} = -\delta\alpha_{p+1}^*/(\alpha_{p+1}^*)^2$ and $\delta\alpha_{p+1} = \lambda\beta_{p+1}$ is constrained by $\mathrm{Re}(\delta\alpha_{p+1}\alpha_{p+1}^*) = 0$. It is easy to verify that these two constraints imply $\delta\alpha_{p+1} = 0$. If $\alpha_{p+1} = \alpha_{p+1}^*$, then, we have seen that $\delta\alpha_{p+1} = 0$ anyway.

Now, using (B6), we conclude that, for all θ,

$$e^{-2i\theta}(e^{i\theta} - \alpha_{p+1})^2(e^{i\theta} - (1/\alpha_{p+1}))^2 \delta L_1(n(\tilde{N});e^{i\theta}) = 0, \tag{B12}$$

which, in view of the analyticity of $\delta L_1(n(\tilde{N});z)$ around $|z| = 1$ implies $\delta L_1(n(\tilde{N});e^{i\theta}) = 0$. Through (B10), we obtain $\delta|E_1(N;e^{i\theta})| = 0$ and, using the representation of functions nonvanishing in $|z| < 1$ [cf. Eq. (3.48)], we deduce that, for those δN obeying (B1), for all z,

$$\delta E_1(N;z) = 0. \tag{B13}$$

Now, taking the variation of (B9) and using also (B2), we obtain (B7), as announced.

We next give an interpretation of the results of this appendix and also settle the proof of Lemma 3.7. We assume condition (H) is true. The position of the zero $\alpha(N)$, which tends to α_{p+1} as $N(x)$ approaches $\tilde{N}(x)$, is a Fréchet differentiable functional in a certain neighborhood $\mathscr{U}$ of $\tilde{N}$; the critical surface Σ on which the discontinuity of $(\partial\mathscr{E}/\partial N)(N)$ occurs is a level surface of the functional $\mathscr{L}(N) = \alpha(N)\alpha^*(N)$, namely, $\mathscr{L}(N) = 1$. The operator

$\mathscr{E}(N)$ of (3.35) is Fréchet differentiable in $\mathscr{U} \cap (\mathscr{Z}(N) < 1)$. We can define now an "analytic continuation" of $\mathscr{E}$ to $\mathscr{Z}(N) > 1$ by deforming slightly the contour of integration ($|z| = 1$) of $E_0(N;z)$, Eq. (3.48). We allow, namely, for two semicircles around $\alpha_{p+1}, \alpha_{p+1}^*$ leaving these points in the interior of the contour. For all $N(x)$ for which $\mathscr{Z}(N) < 1$ ($N \in \mathscr{U}$) the operator $\mathscr{E}_i$ defined this way through (3.35) coincides with $\mathscr{E}$. It does not do so for those N in a sufficiently small neighborhood $\mathscr{U}_1 \subset \mathscr{U}$ of N, for which $\alpha(N)$ still lies inside the modified contour, but $\mathscr{Z}(N) > 1$. However, $\mathscr{E}_i(N)$ has a continuous Fréchet derivative throughout $\mathscr{U}_1$ and $(\partial \mathscr{E}_i/\partial N)(\tilde{N}) = ((\partial \mathscr{E}/\partial N)(\tilde{N}))_i$. By the results of this appendix and the implicit function theorem, $\mathscr{E}_i$ realizes a diffeomorphism of a neighborhood $\mathscr{V} \subset \mathscr{U}_1$ of $\tilde{N}$ onto a neighborhood $\mathscr{W}$ of $\tilde{n} = \mathscr{E}(\tilde{N})$. Let Σ' be the image of Σ under $\mathscr{E}_i$; it is a level surface of $\mathscr{Z}_i(n) = \mathscr{Z}(\mathscr{E}_i^{-1}(n))$. It is easy to see that its tangent plane at $\mathscr{E}(\tilde{N})$ is given by

$$(k_1(x), n(x) - \tilde{n}(x))_\rho = 0, \tag{B14}$$

with $k_1(x)$ defined in Eq. (3.70). Here $\mathscr{E}_i(N)$ maps $\mathscr{V} \cap (\mathscr{Z}(N) < 1)$ onto $\mathscr{W} \cap (\mathscr{Z}_i(n) < 1)$ in a one-to-one manner.

Similarly, we may define an analytic continuation of $\mathscr{E}(N)$ from $\mathscr{Z}(N) > 1$ to $\mathscr{Z}(N) < 1$. For this, we take semicircles around $\alpha_{p+1}, \alpha_{p+1}^*$ leaving these points outside the contour. We obtain this way another diffeomorphism $\mathscr{E}_e(N)$ of the neighorhood $\mathscr{V}$ of $\tilde{N}$ onto a neighborhood $\mathscr{W}_1$ of $\tilde{n}$. Also, we may define $\mathscr{Z}_e(n) = \mathscr{Z}(\mathscr{E}_e^{-1}(n))$. The sets $\mathscr{V} \cap (\mathscr{Z}(N) \gtrless 1)$ are mapped one-to-one and onto $\mathscr{W}_1 \cap (\mathscr{Z}_e(n) \lesseqgtr 1)$. Also, in $\mathscr{W}_0 \equiv \mathscr{W} \cap \mathscr{W}_1$, the sets $\mathscr{Z}_e(n) = 1$ and $\mathscr{Z}_i(n) = 1$ coincide. As a consequence, one may verify the following: the set $\mathscr{W}_0 \cap (\mathscr{Z}_e(n) > 1)$ either coincides with $\mathscr{W}_0 \cap (\mathscr{Z}_i(n) < 1)$ or is disjoint from it. In the latter situation, it coincides with $\mathscr{W}_0 \cap (\mathscr{Z}_i(n) > 1)$.

To settle this ambiguity, we use the fact that the operator $\mathscr{E}(N)$ is one-to-one in $L^2(\rho)$ (Lemma 3.5). Consequently, the images of the sets $\mathscr{V} \cap (\mathscr{Z}(N) < 1)$ and $\mathscr{V} \cap (\mathscr{Z}(N) > 1)$ are disjoint. These images contain in turn the sets $\mathscr{W}_0 \cap (\mathscr{Z}_i(n) < 1)$ and $\mathscr{W}_0 \cap (\mathscr{Z}_e(n) > 1)$; thus, these latter sets must be disjoint. We conclude that $\mathscr{W}_0 \cap (\mathscr{Z}_e(n) > 1) \equiv \mathscr{W}_0 \cap (\mathscr{Z}_i(n) > 1)$. The latter set contains points $n(x)$ such that (a) the whole segment $\mu\tilde{n} + (1-\mu)n, 0 < \mu < 1$ is contained in it [cf. Appendix D, following Eq. (D3)] and (b) $(k_1(x), n(x) - \tilde{n}(x))_\rho > 0$. For such an $n(x)$, it is true that

$$\begin{aligned} 0 &< \mathscr{Z}_e(n) - \mathscr{Z}_e(\tilde{n}) \\ &= \mathscr{Z}(\mathscr{E}_e^{-1}(n)) - \mathscr{Z}(\mathscr{E}_e^{-1}(\tilde{n})) \\ &= \left(p(x), \left(\frac{\partial \mathscr{E}}{\partial N}(\tilde{N}) \right)_e^{-1} (n - \tilde{n}) \right)_\rho + o(\|n - \tilde{n}\|_\rho). \end{aligned} \tag{B15}$$

Dividing by $\|n - \tilde{n}\|_\rho$ and letting then $\|n - \tilde{n}\|_\rho$ tend to zero, we obtain that, for $\delta n = (n - \tilde{n})/\|n - \tilde{n}\|_\rho$, both

$$(k_1(x), \delta n(x)) > 0, \quad (k_2(x), \delta n(x)) \geqslant 0, \tag{B16}$$

where we have used Eq. (3.70). Since $k_1(x) = \lambda k_2(x)$, $\lambda \neq 0$, (B16) implies $\lambda > 0$. This ends the proof of Lemma 3.7.

Next follow some comments.

(a) The reasoning above shows that, in fact, $\mathscr{E}^{-1}(n)$ is a continuous operator from $L^2(\rho)$ into $L^2(\rho)$. Indeed, given a neighborhood $\mathscr{V}$ of $\tilde{N}$, it is enough to choose $\mathscr{W} \subset \mathscr{E}_e(\mathscr{V}) \cap \mathscr{E}_i(\mathscr{V})$ to make sure that $\mathscr{E}^{-1}(\mathscr{W}) \subset \mathscr{V}$.

(b) The same reasoning may be repeated for the operator $\mathscr{A}(N)$, Eq. (3.72). We may write, by analogy,

$$\mathscr{A}_i(N) = N(x) E_{0;i}(N;x) + F(N;x)/E_{0;i}(N;X), \tag{B17}$$

where $E_{0;i}(N;x)$ is the outer function defined on the modified contour described above, and similarly for $\mathscr{A}_e(N)$. In the course of the proof of Theorem 3.2 we have shown that $(\partial \mathscr{A}/\partial N)_e$, $(\partial \mathscr{A}/\partial N)_i$ have bounded inverses. This, together with the fact that $\mathscr{A}(N)$ is one-to-one in neighborhoods of N and h (see Theorem 3.2) shows that Lemma 3.7 is true also for $\mathscr{A}(N)$. Similarly, $\mathscr{A}^{-1}(h)$ is continuous.

(c) The argument of this Appendix concerning the existence of $(\partial \mathscr{E}/\partial N)_e^{-1}$ may be applied even if several simple zeros approach $|z| = 1$, staying away from each other. [This goes beyond hypothesis (H).] There are in this case several relevant approximating sequences $\{N_k\}$ to N, corresponding to some of the zeros lying outside $|z| = 1$ and the others inside. Associated limiting Fréchet derivatives can be defined and they have bounded inverses.

APPENDIX C: SOME STATEMENTS ON LAGRANGE MULTIPLIERS

We prove two theorems concerning Lagrange multipliers.

Theorem C.1: (See also Problem 7, Chap. 8, Ref. 31.) If $F(\alpha)$, Eq. (5.2), is strictly positive definite, there exist numbers λ_i such that Eq. (5.7) is true.

Proof: By Lemma 5.1, the point with coordinates $\{w_i\}_{i=1}^m$ is contained in the open, convex set $\mathring{\mathscr{S}}_m$. We define, for $y \in \mathring{\mathscr{S}}_m$ the function

$$\begin{aligned} \Psi(y) = \sup\{(n,f)_\rho :\ & f \in \mathscr{F}_1;\ f(\xi_i) = y_i, \\ & \{y_i\}_{i=1}^m \in \mathring{\mathscr{S}}_m\}. \end{aligned} \tag{C1}$$

Let $f_c(y)$ be a function on which the supremum is attained. Then, if $y_1, y_2 \in \mathring{\mathscr{S}}_m$,

$$\begin{aligned} &\lambda\Psi(y_1) + (1-\lambda)\Psi(y_2) \\ &\quad = (n, \lambda f_c(y_1) + (1-\lambda) f_c(y_2))_\rho \\ &\quad \leqslant (n, f_c(\lambda y_1 + (1-\lambda) y_2))_\rho = \Psi(\lambda y_1 + (1-\lambda) y_2), \end{aligned} \tag{C2}$$

so that $\Psi(y)$ is a concave function. Consider then in $R^m \times R$ the convex set $\mathscr{K}$ given by the points (y,g), with $y \in \mathring{\mathscr{S}}_m$, $g \leqslant \Psi(y)$ and let $g + \Sigma \lambda_i y_i = \text{const}$ be a supporting hyperplane (Ref. 49, Theorem 11.6) through the point $(w, \Psi(w))$. The fact that the coefficient of g may be chosen equal to 1 is a consequence of the condition $w \in \mathring{\mathscr{S}}_m$. It is obvious that the coefficients λ_i of this hyperplane are the desired Lagrange multipliers.

Theorem C.2: Assume $\sup\{(n,f)_\rho + \Sigma_i \lambda_i (f(\xi_i) - w_i);\ f \in \mathscr{F}_1\}$ is attained on a function $f_{c;\lambda}(z)$ with $f_{c;\lambda}(\xi_i) = w_i$. Then, if $n \neq 0$, $\sup\{(n,f)_\rho;\ f \in \mathscr{F}_1, f(\xi_i) = w_i\}$ is attained on the same $f_{c;\lambda}(z)$.

Proof: If the second supremum is attained on a different

function f_c, it is true that $(n,f_c)_\rho > (n,f_{c;\lambda})_\rho$, since we have seen (Theorem 5.1) that f_c is unique. However, if this is so, $f_{c;\lambda}$ cannot be the function on which the first supremum is attained, since f_c gives a larger value to the quantity in brackets. This ends the proof (and gives Lemma 5.2).

APPENDIX D: A CONVERGENT ALGORITHM FOR THE SOLUTION OF EQ. (3.47)

In this Appendix, we describe an algorithm for the minimization of $\Phi(N)$, Eq. (6.21), for which convergence can be proved. More precisely, we show that if $\Phi(N_1)$ is small enough, we can construct a Cauchy sequence $\{N_k\}_{k=1}^{\infty}$ converging to the solution N_0 of $\Phi(N_0)=0$ by modifying suitably the Newton–Kantorowich iteration. It is profitable to make explicit the dependence of $\Phi(N)$ on h and write in the following $\Phi(N;h)$. If N_1 does not lie on a separating surface $\Sigma: F(N;e^{i\theta})=0$, for some θ on $[0,\pi]$, then, by Theorem 3.2, the usual form of the implicit function theorem applies to the equation $\mathscr{A}(N(x))=h(x)$, and one knows that there exist neighborhoods $\mathscr{U}$ of N_1 and $\mathscr{V}$ of $\mathscr{A}(N_1)$ and a continuous differentiable operator $\mathscr{A}^{-1}$ defined on $\mathscr{V}$ with values in $\mathscr{U}$, so that $\mathscr{A}^{-1}(\mathscr{A}(N_1))=N_1$ and for all x in $\mathscr{V}$, $\mathscr{A}(\mathscr{A}^{-1}(x))=x$. The construction of $\mathscr{A}^{-1}$ may be done by the usual Newton iteration (see, e.g., Ref. 36, p. 57ff) so that our statement requires no further comment. As N_1 approaches Σ, the neighborhoods $\mathscr{U},\mathscr{V}$ of $N_1,\mathscr{A}(N_1)$ shrink indefinitely, since the proof of the theorem requires the continuous differentiability of $\mathscr{A}$ in a certain neighborhood of N_1. However, we have shown in Appendix B that $\mathscr{A}(N)$ still provides a bicontinuous correspondence between neighborhoods $\mathscr{U}$ of N_1 and $\mathscr{V}$ of $\mathscr{A}(N_1)$ even if N_1 lies on Σ. We shall now describe a slight modification of the Newton procedure, which allows the explicit construction of $\mathscr{A}^{-1}(N)$ also in this case.

From the outset, we point out that the difficulty in proving the convergence of the iteration near points of Σ is that one cannot apply Taylor's second-order formula to estimate the variation of $\Phi(N;h)$. To circumvent this, we part the step of Newton's method into two: one part up to Σ, the other following it. However, we need to make sure that we no longer meet Σ while performing the second step, after a too short distance. This is achieved by moving in a direction that is "sufficiently normal" to Σ.

In the following, if $N_1(x)\in\Sigma$, the symbol $\|((\partial\mathscr{A}/\partial N)(N_1))^{-1}\|_\rho$ means $\max(\|(\partial\mathscr{A}/\partial N)(N_1))_e^{-1}\|_\rho, \|((\partial\mathscr{A}/\partial N)(N_1))_i^{-1}\|_\rho)$. Now if $\|((\partial\mathscr{A}/\partial N)(N_1))^{-1}\|_\rho = K_1$, then it is easy to show that, if $K>K_1$, there exists a neighborhood $\mathscr{U}(N_1;r)$ of $N_1(x)$: $\|N-N_1\|<r(K)$ such that, for all N in $\mathscr{U}(N_1;r(K))$, $\|((\partial\mathscr{A}/\partial N)(N))^{-1}\|<K$. We skip the proof of this statement.

We shall use restriction (H), Sec. III, and assume thus: (a) $F(N_1e^{i\theta})=0$ only for two real values of θ, $-\pi<\theta<\pi$, $\theta_1,-\theta_1$, $z_1=e^{i\theta_1}$; and (b) $F'(N_1;e^{i\theta_1})\neq 0$. With this, we restrict $\mathscr{U}(N_1;r)$ even further as follows: (i) given $\tau,\sigma>0$, sufficiently small, there exists $r_1(\tau,\sigma)\leqslant r(K)$ such that, for all N in $\mathscr{U}(N_1;r_1(\tau,\sigma))$, $F(N;z)$ vanishes only once in $|z-z_1|<\tau$ and $|F'(N_1;z)|>\sigma$ if $|z-z_1|<\tau$; and (ii) we choose $\hat{r}\leqslant r_1(\tau,\sigma)$ such that for all $N\in\mathscr{U}(N_1;\hat{r})$, $F(N;z)\neq 0$ for z on $|z|=1$, except for points contained in disks of radius τ around z_1,z_1^*.

We now discuss how to control the direction of the step of minimization as we cross the surface Σ. To this end, we notice that Σ may be viewed as a level surface of the (nonlinear) functional $\mathscr{Z}(N)\equiv z_0(N)z_0^*(N)$, with $z_0(N)$ the root of $F(N;z)$ in $|z-z_1|<\tau$. Since $F_0'(N;z)\neq 0$, for $|z-z_1|<\tau$, $\mathscr{Z}(N)$ is in fact of class C^∞ with respect to N in $\mathscr{U}(N_1;\hat{r})$; its Fréchet derivative at $N_1(x)$ is determined by the vector in $L^2(\rho)$,

$$g(N_1;x)\equiv \mathrm{Re}\left[e^{-i\theta_1}\frac{1}{1-xe^{i\theta_1}}\frac{1}{F'(N_1;e^{i\theta_1})}\right]. \tag{D1}$$

Let now $\tilde{B}$ be an upper bound on the norm of the Fréchet derivative of $g(N;x)$ for N in $\mathscr{U}(N_1;\hat{r})$. Then, if N', $N''\in\mathscr{U}(N_1;\hat{r})$, it is true that

$$|z_0(N')z_0^*(N')-z_0(N'')z_0^*(N'')-(g(N''),N''-N')_\rho| \leqslant \tfrac{1}{2}\tilde{B}\,\|N''-N'\|_\rho^2. \tag{D2}$$

With this, if we confine ourselves to the cone of directions ΔN given by, say,

$$|(g(N_1),\Delta N)_\rho| \geqslant 3\tilde{B}\hat{r}\|\Delta N\|_\rho, \tag{D3}$$

then any line $\tilde{N}+\lambda\Delta N$, originating at $\tilde{N}\in\mathscr{U}(N_1;\hat{r})$ with $z_0(\tilde{N})z_0^*(\tilde{N})=1$ intersects the surface Σ only once in $\mathscr{U}(N_1;\hat{r})$, namely, for $\lambda=0$. This follows from (D2) and a mean value theorem for $\|g(N_1)-g(\tilde{N})\|_\rho$. Condition (D3) is meaningful only if

$$\mu = 3\tilde{B}\hat{r}/\|g(N_1)\|_\rho < 1. \tag{D4}$$

This may be achieved by simply choosing a smaller $\hat{r}$, if necessary.

We next give a prescription for choosing the general step (the k th) of the algorithm for minimization, ΔN_k. The prescription depends on the magnitude of the component of the standard Newton step

$$\Delta N_{0,k}\equiv\left(\frac{\partial\mathscr{A}}{\partial N}(N_k)\right)^{-1}(\mathscr{A}(N_k)-h) \tag{D5}$$

on the direction of $g(N_1)$. Namely, if

$$|(\Delta N_{0,k},g(N_1))_\rho| \geqslant \mu\|g(N_1)\|_\rho\|\Delta N_{0,k}\|_\rho, \tag{D6}$$

then we choose

$$\Delta N_k = \Delta N_{0,k}. \tag{D7}$$

However, if (D6) is violated, we let

$$\Delta N_k = \Delta N_{0k} + g(N_1)\,[a_k\Phi^{1/2}(N_k;h) - (g(N_1),\Delta N_{0k})_\rho/\|g(N_1)\|_\rho^2], \tag{D8}$$

where a_k is a constant, with modulus independent of k, but

$$\operatorname{sgn} a_k = \operatorname{sgn}(g(N_1),\Delta N_{0k})_\rho \tag{D9}$$

and

$$|a_k| = a > \mu K/\|g(N_1)\|_\rho. \tag{D10}$$

This choice of step ensures that the modulus of the component of ΔN_k on $g(N_1)$ either satisfies (D6) or is equal to $a\Phi^{1/2}(N_k;h)$. We choose further a in such a manner that, besides (D10), condition (D3) is fulfilled. If (D6) is false, then, according to (D8),

$$\|\Delta N_k\|_\rho \leqslant \|\Delta N_{0k}\|_\rho + a\|g(N_1)\|_\rho \Phi^{1/2}(N_k;h)$$
$$\leqslant (K + a\|g(N_1)\|_\rho)\Phi^{1/2}(N_k;h) \quad \text{(D11)}$$

and one sees that (D3) is fulfilled if, e.g.,

$$a = \mu K / \|g(N_1)\|_\rho (1-\mu) \quad \text{(D12)}$$

[consistent with (D10)]. It follows that, in case (D6) does not hold,

$$\|\Delta N_k\|_\rho \leqslant K\Phi^{1/2}(N_k;h)/(1-\mu). \quad \text{(D13)}$$

The estimates (D11) and (D13) are meaningful only if the whole sequence $\{N_k\}_{k=1}^{\infty}$ belongs to $\mathscr{U}(N_1;\hat{r})$. We shall show below that, if $\|\mathscr{A}(N_1) - h(x)\|_\rho^2 \equiv \Phi(N_1;h)$ is sufficiently small, this will be the case.

Assuming for a while that this is true, if $N_k \in \Sigma$, then the line $N_k + \lambda \Delta N_k$, $0 \leqslant \lambda \leqslant 1$ no longer intersects Σ_1. However, for arbitrary $N_k \in \mathscr{U}(N_1;\hat{r})$, it is possible that we meet Σ at $N_k' = N_k + \lambda \Delta N_k$, $0 < \lambda \leqslant 1$. We wish to make sure that $\Phi(N_k';h) < \Phi(N_k;h)$. To this end, we apply Taylor's second-order formula along the segment joining N_k, N_k'. If $\hat{r}$ is further chosen so that, for some $C, B > 0$ and any $N \in \mathscr{U}(N_1;\hat{r})$, both

$$\left\| \frac{\partial \mathscr{A}}{\partial N}(N) g(N_1) \right\|_\rho < C,$$
$$\left\| \frac{\partial^2 \mathscr{A}}{\partial N^2}(N) \right\|_\rho < B \quad (N \notin \Sigma), \quad \text{(D14)}$$

and assuming first that (D6) is violated, we obtain [using (D13)]

$$\Phi^{1/2}(N_k';h) \leqslant \left[1 - \lambda(1 - aC) + \lambda^2 \frac{K^2}{(1-\mu)^2} \times B \frac{\Phi^{1/2}(N_k;h)}{2}\right] \Phi^{1/2}(N_k;h). \quad \text{(D15)}$$

The function of λ in the brackets is less than unity for all $0 < \lambda < 1$ if it is less than unity for $\lambda = 1$, i.e., if

$$aC + \frac{K^2}{(1-\mu)^2} B \frac{\Phi^{1/2}(N_1;h)}{2} = p < 1, \quad \text{(D16)}$$

where we have used the induction hypothesis $\Phi(N_k;h) < \Phi(N_1;h)$. With the expression (D12) for a and Eq. (D4), we see that, if both $\hat{r}$ and $\Phi^{1/2}(N_1;h)$ are sufficiently small, we can satisfy (D16). In case (D6) is fulfilled, then only the second term occurs in (D16), which is thus automatically obeyed.

If the line $N_k + \lambda \Delta N_k$ meets Σ at N_k', we let N_k' be the new starting point and continue. We obtain this way a sequence of points $N_1,\ldots,N_k,N_k',N_{k+1},\ldots$, where some of the indices may appear at most twice, and a corresponding decreasing sequence $\Phi(N_1;h),\ldots,\Phi(N_k;h),\Phi(N_k';h),\ldots$ such that $\Phi(N_{k+1};h)/\Phi(N_k;h) \leqslant p^2 < 1$. It follows that the sequence $\Phi(N_k;h)$ tends to zero. Also

$$\|N_{k+1} - N_k\|_\rho \leqslant \|N_{k+1} - N_k'\|_\rho + \|N_k' - N_k\|_\rho$$
$$< [2K/(1-\mu)]\Phi^{1/2}(N_k;h)$$
$$\leqslant [(2K/(1-\mu)]p^k \Phi^{1/2}(N_1;h), \quad \text{(D17)}$$

which shows that $\{N_k\}_{k=1}$ is a Cauchy sequence. All terms N_k, N_k' of the sequence stay in a ball around N_1 of radius R given by

$$\|N_k - N_1\|_\rho \leqslant \sum_{i=1}^{k-1} \|N_{i+1} - N_i\|_\rho$$
$$< \left(\frac{2K}{1-\mu}\right)\Phi^{1/2}(N_1;h) \frac{1}{1-p} \equiv R. \quad \text{(D18)}$$

Clearly, R may be made smaller than $\hat{r}$ if $\Phi^{1/2}(N_1;h)$ is sufficiently small. If this is so, then $N_0' = \lim_{k\to\infty} N_k$ belongs to $\mathscr{U}(N_1;\hat{r})$ and, since $\Phi(N;h)$ is continuous,

$$\lim_{k\to\infty} \Phi(N_k;h) = 0 = \Phi(N_0';h),$$

it follows that $N_0' = N_0$, the solution of the equation $\Phi(N_0;h) = 0$.

To sum up, we have shown that, if $\hat{r}$ is chosen so that (D4) and (D14) are true, the quantity $aC < 1$ [cf. (D16)] and the conditions preceding Eq. (D1) are met, then for any $h(x)$ such that $\Phi(N_1;h)$ satisfies (D16) and gives rise in (D18) to $R \leqslant \hat{r}$, the equation $\Phi(N;h) = 0$ has a solution in $\mathscr{U}(N_1;\hat{r})$ and this solution may be found by the modified Newton iteration described above. This solution is unique by Theorem 3.2. This establishes the claim made in this Appendix.

It is clearly possible to replace the Newton method with the minimization procedure of Sec. VI, provided we confine ourselves to a search over a (uniformly with respect to k) finite number of steps (D8) and choose the constants in (D18) appropriately. This justifies Theorem 6.1.

APPENDIX E: MULTIPLE ZEROS

The theorems presented in this paper are true regardless of the multiplicity of the zeros of the extremal Blaschke product $B(N;z)$ in $|z| < 1$. However, several steps of the proofs [cf., e.g., Eqs. (3.52), (3.59), and (3.60)] and even the statements of some of the lemmas (3.2, 3.4, and 5.3) make use of the assumption (introduced preceding Lemma 3.2) that the zeros of $B(N;z)$ are simple. In this appendix, we describe a general procedure to treat multiple zeros and complete this way the justification of the theorems of Secs. III and V.

First, expressions (2.25) and (3.2) for $L(n;z)$ suggest that, if, say, α_1 is a zero of the rth order of $B(n;z)$, then $r-1$ of the equations (3.9) must be replaced by

$$\left(n(x), B(x;\alpha)) \frac{\partial^k (P(\alpha_1)/\alpha_1)}{\partial \alpha_1^k}\right)_\rho = 0, \quad k = 1,2,\ldots,r-1. \quad \text{(E1)}$$

The first part of the proof of Lemma 3.2 under this condition is obtained from the requirement that the expression on the right-hand side of Eq. (3.18) is holomorphic in $|z| < 1$ [cf. comment following Eq. (3.23)]. In particular, all negative Laurent coefficients in the expansion of (3.19) around $z = \alpha_1$ (possibly $\alpha_1 \in [-a,b]$) should vanish. Using (3.2) this shows that (E1) are necessary conditions. The second part of the proof stays unchanged.

Similarly, under the conditions above, $r-1$ of Eq. (3.29) of Lemma 3.4 must be replaced by

$$\left(N(x), F(N;x) \frac{\partial^k (P(\alpha_1)/\alpha_1)}{\partial \alpha_1^k}\right)_\rho = 0, \quad k = 1,2,\ldots,r-1. \quad \text{(E2)}$$

The proof that Eqs. (3.2) are necessary conditions for $F(N;x)$ to be extremal for (3.25) (even if $\alpha_1 \in [-a,b]$) is obtained by differentiating with respect to z the identity [which uses (3.26)]

$$\left(N(x), \frac{F(x)-F(z)}{x-z}\right)_\rho = \left(N(x), F(x)\frac{x}{1-xz}\right)_\rho \quad \text{(E3)}$$

and letting $z=\alpha_1$. The integrands are holomorphic functions of z, for $z\in[-a,b]$, so that differentiation poses no problems. Thus, Theorem 3.1 is justified for all $N(x)\in L^2(\rho)$ indeed.

The difficulty that appears in the arguments of Lemma 3.5 if we allow for multiple zeros is that the functionals $\alpha_i(N)$ describing the dependence of the position of a zero of $F(N;z)$ [and thus of $B(N;z)$] on $N(x)$, are no longer Fréchet differentiable. Indeed, the condition for application of the implicit function theorem ($F_0'(N;z)\neq 0$) fails. However, the symmetric combinations

$$\begin{aligned}
\gamma_1(N) &= -\sum_{i=1}^{p} \alpha_i(N), \\
\gamma_2(N) &= \sum_{i<j}^{p} \alpha_i(N)\alpha_j(N), \\
&\vdots \\
\gamma_p(N) &= (-1)^p \prod_{i=1}^{p} \alpha_i(N)
\end{aligned} \quad \text{(E4)}$$

are always Fréchet differentiable. This is obtained by expressing them as polynomials in the symmetric sums

$$\sigma_k(N) = \sum_{i=1}^{p} \alpha_i(N)^k \quad \text{(E5)}$$

(by means of Newton's formulas, see Ref. 50) and writing the latter as

$$\sigma_k(N) = \frac{1}{2\pi i}\oint_{\mathscr{C}} \frac{z^k(\partial F(N;z)/\partial z)}{F(N;z)dz}, \quad \text{(E6)}$$

where $\mathscr{C}$ is a contour surrounding the p zeros of $F(N;z)$ and no others and $F(N;z\in\mathscr{C})\geqslant\mu>0$. From (E6) it is clear that all σ_k's and thus all γ_i's are Fréchet differentiable functionals of N.

Now, the Blaschke products can be expressed conveniently in terms of $\gamma_i(N)$:

$$B(N;z) = \frac{\Sigma_{i=0}^{p}\gamma_i(N)z^{N-i}}{\Sigma_{i=0}^{p}\gamma_i(N)z^i} \quad (\gamma_0(N)=1), \quad \text{(E7)}$$

and, instead of the α_i's, we may take the γ_i's as parameters. As in Lemma 3.2 we obtain another form of the p conditions for $B(n;z)$ to be the extremal function associated to $n(z)$,

$$\left(n(x), \frac{\partial B(N;x)}{\partial \gamma_i}\right) = 0, \quad i=1,2,\ldots,p. \quad \text{(E8)}$$

The reasoning of Lemma 3.6 following Eq. (3.57) may be done again with no other change but the replacement of the $\delta\alpha_i$ with the $\delta\gamma_i$, which are always well-defined quantities.

The changes in Lemma 5.3 are the same as in Lemma 3.2, whereas in Lemma 5.4 it is sufficient to verify that the determinant (2.10) stays nonvanishing even if $P(\xi_k;\alpha_i)/\alpha_i$ is replaced by $\partial(P(\xi_k;\alpha_i)/\alpha_i)/\partial\alpha_i$, for some i. The statement of Theorem 5.3 is obtained by taking the γ_j, Eq. (E4), as parameters and resorting to (E8) instead of (5.12). The functions $(\partial B/\partial\gamma_j)(z)$ are indeed linearly independent; this follows from the fact that, if constants μ_i existed so that

$$\sum_{k=1}^{p} \mu_i \frac{\partial B(N;z)}{\partial\gamma_i} \equiv 0, \quad \text{(E9)}$$

then they are also such that, for $|z|=1$ ($\gamma_0=1$),

$$\mathrm{Im}\left(\sum_{k=0}^{p}\gamma_k z^k\right)\left(\sum_{k=1}^{p}\frac{\mu_k}{z^k}\right) = 0. \quad \text{(E10)}$$

The first factor has only zeros in $|z|>1$ and the second factor cannot produce sufficiently many zeros conjugate to these with respect to the unit circle, since its degree is less than p. Thus $\mu_i=0$, for all i.

We may now inquire whether it is possible by this means to get rid of the restriction (H) of Sec. III. e.g., to allow for multiple zeros to cross the circle $|z|=1$. In any neighborhood of a point $\tilde{\tilde{N}}$, so that $F(\tilde{\tilde{N}};z)$ has a double zero on $|z|=1$, there exist points N, so that $F(N;z)$ has a simple zero at $\alpha(N)$ in $|z|<1$ and another simple zero at $\alpha_1(N)$ in $|z|>1$. For such points N, we can again write the Fréchet derivatives of $(\partial B/\partial N)(N)$ by means of Eq. (3.52). However, contrary to the situation of Appendix B, as N approaches $\tilde{\tilde{N}}$, the Fréchet derivative of $\alpha(N)$ with respect to N diverges, and the limit of $(\partial B/\partial N)(N)$ does not exist for all $\delta N\in L^2(\rho)$. This cannot be cured by the method above and a different treatment of these points is required. The author has not pursued this item further.

On the other hand, it may be possible to relax that part of (H) referring to several simple zeros lying on $|z|=1$ for some N. As pointed out in Appendix B, comment (c), the various inverse Fréchet derivatives exist in this case. However, it appears difficult to give a proof of Lemma 3.7 for this situation in sufficient generality. The author presumes its statement stays nevertheless true.

[1] G. Nenciu, I. Raszillier, W. Schmidt, and H. Schneider, Nucl. Phys. B **63**, 285 (1973).
[2] W. W. Rogosinski and H. S. Shapiro, Acta Math. **90**, 287 (1952).
[3] P. D. Lax, Commun. Pure Appl. Math. **VIII**, 437 (1955).
[4] A. N. Tykhonov and V. Arsenin, *Methodes de Resolution des problemes mal poses* (Mir, Moscow, 1973).
[5] M. M. Lavrentiev, *Some Improperly Posed Problems of Mathematical Physics* (Springer, Berlin, 1967).
[6] M. Bertero and G. A. Viano, Nuovo Cimento **38**, 1915 (1965).
[7] C. Goebel, Phys. Rev. Lett. **1**, 337 (1958).
[8] G. F. Chew and F. E. Low, Phys. Rev. **113**, 1640 (1959).
[9] R. E. Cutkosky and B. B. Deo, Phys. Rev. Lett. **22**, 1272;1859 (1968).
[10] S. Ciulli, Nuovo Cimento A **61**, 787 (1969); **62**, 301 (1969).
[11] I. S. Stefanescu, Institut für Theoretische Kernphysik preprint TKP 81/10.
[12] R. E. Cutkosky, Ann. Phys. (NY) **54**, 110 (1969).
[13] E. Pietarinen, Nuovo Cimento A **12**, 522 (1972); Nucl. Phys. B **55**, 541 (1973).
[14] R. E. Cutkosky, J. Math. Phys. **14**, 1231 (1973).
[15] I. Raszillier and W. Schmidt, Nucl. Phys. B **55**, 106 (1973).
[16] H. D. Kiehlmann and W. Schmidt, Nucl. Phys. B **94**, 67 (1975).
[17] I. S. Stefanescu, J. Math. Phys. **21**, 175 (1980).
[18] G. Auberson and S. Ciulli, preprint Montpellier, PM84/10.
[19] G. Nenciu and I. Raszillier, Nuovo Cimento A **11**, 319 (1972).
[20] I. Raszillier, W. Schmidt, and I. S. Stefanescu, Nucl. Phys. B **109**, 452 (1976); Z. Phys. A **277**, 211 (1976).

[21]D. Atkinson, in *Applied Inverse Problems, Lecture Notes in Physics*, Vol. 85 (Springer, Berlin, 1978), p. 111.
[22]M. G. Krein and A. A. Nudelman, *The Theory of Markov Moments and Extremal Problems*, (Nauka, Moscow, 1973) (in Russian).
[23]G. Nenciu, Nucl. Phys. B **53**, 584 (1973).
[24]G. Polya and G. Szegö, *Sätze aus der Analysis* (Springer, Berlin, 1964), Bd. II, Kap. VII, Problem 3.
[25]This method of finding the allowed values of $f(x)$, if $f\in\mathcal{F}_1$ and obeys (2.9), was communicated to the author by S. Ciulli in 1977 (with a different proof).
[26]P. Dită, J. Phys. A **17**, 957 (1984).
[27]No new definition of χ^2 is introduced in Ref. 17, as asserted in Ref. 26.
[28]S. Ciulli and T. D. Spearman, J. Math. Phys. **23**, 1752 (1982).
[29]W. Rudin, *Real and Complex Analysis*, (McGraw–Hill, New York, 1973), 2nd ed.
[30]W. Rudin, *Functional Analysis* (McGraw–Hill, New York, 1973).
[31]D. G. Luenberger, *Optimization by Vector Space Methods* (Wiley, New York, 1969).
[32]P. L. Duren, *Theory of H^p Spaces* (Academic, New York, 1970).
[33]J. E. Marsden, *Basic Complex Analysis* (Freeman, San Francisco, 1973).
[34]H. Meschkowski, *Hilbertsche Raume mit Kernfunktion* (Springer, Berlin 1962).
[35]S. Okubo, J. Math. Phys. **15**, 963 (1974).
[36]H. Cartan, *Calcul differentiel* (Hermann, Paris, 1967).
[37]M. Heins, Ann. Math. **61**, 440 (1955); known to the author from S. Stoilov, *Theory of Functions of a Complex Variable* (Editura Academiei, Bucharest, 1958) (in Romanian).
[38]If an element of $L^2(\rho_c)$ is represented by a pair (n,λ) with $n\in L^2(\rho)$, $\lambda\in R^m$, the scalar product is written as $(\cdot;\cdot)_c$.
[39]We consider now the set $\tilde{\mathcal{F}}_1$ in $L^2(\rho_c)$.
[40]I. S. Stefanescu, Institut für Theoretische Kernphysik preprint, TKP 85/2.
[41]S. Ciulli and G. Nenciu, Nuovo Cimento **8**, 735 (1972).
[42]I. Caprini, S. Ciulli, C. Pomponiu, and I. S. Stefanescu, Phys. Rev. D **5**, 1658 (1972).
[43]I. Caprini, Nuovo Cimento Lett. **5**, 443 (1972); Nucl. Phys. B **68**, 561 (1974).
[44]I. S. Stefanescu, Nucl. Phys. B **56**, 287 (1973).
[45]S. Ciulli, C. Pomponiu, and I. S. Stefanescu, Phys. Rep. C **17**, 134 (1975).
[46]M. Z. Nashed, J. Math. Mech. **13**, 765 (1964).
[47]L. W. Kantorowich and G. P. Akilow, *Funktionalanalysis in Normierten Räumen* (Akademic, Berlin, 1964).
[48]M. Altman, Bull. Acad. Sci. Pol. Cl. III, **5**, 1031 (1957).
[49]R. T. Rockafellar, *Convex Analysis* (Princeton U.P., Princeton, 1970).
[50]H. W. Turnbull, *Theory of Equations* (Oliver and Boyd, Edinburgh, 1957), p. 73.

Inverse scattering for geophysical problems when the background is variable

A. G. Ramm
Mathematics Department, Kansas State University, Manhattan, Kansas 66506

(Received 19 December 1985; accepted for publication 2 July 1986)

An exact method for finding an inhomogeneity for a variable background from the knowledge of the scattered field on some manifolds is given.

I. INTRODUCTION

Let

$$[\nabla^2 + \omega^2 n(x) + \omega^2 v(x)]u(x,y,\omega) = -\delta(x-y),\ x \in R^3. \tag{1}$$

Here u is the acoustic field generated by a point source situated at the point y in an inhomogeneous medium described by the refraction coefficient $n(x) + v(x)$, where $n(x)$ is the known variable background and $v(x)$ is a compactly supported inhomogeneity. We want to find $v(x)$ from the measurements of $u(x,y,\omega)$ for all x and y running through some manifold M and $0 < \omega < \omega_0$, where ω_0 is a small fixed frequency. If the manifold is the plane $P = \{x: x_3 = 0\}$, the background $n(x) = 1$, or

$$n(x) = \begin{cases} 1, & x_3 > 0, \\ n_0 = \text{const}, & x_3 \leqslant 0, \end{cases}$$

the inverse problem was solved in Refs. 1 and 2. In Ref. 2 the cases when $M = S_{R_1} = \{x: |x| = R_1\}$ or $M = \overset{\circ}{C}_{R_1} = \{x: x_1^2 + x_2^2 = R_1^2\}$ were studied, $R_1 > R$ and supp $v \subset B_R = \{x: |x| \leqslant R\}$, where supp v is the support of v. We assume (without loss of generality) that $v = 0$ if $x_3 \geqslant 0$.

There are two basic results in this paper. The first is a linear integral equation for finding $v(x)$ for a fairly general background $n(x)$. The second is an analytic exact method of finding $v(x)$ in the case when $v = v(x_1, x_3)$, and $M = l_1 \cup l_2$, where $l_1 = \{x: x_1 = a,\ x_2 = 0\}$, $l_2 = \{x: x_1 = -a,\ x_2 = 0\}$, $a > R$, $x \in l_1$, $y \in l_2$. In geophysics this is the well-to-well exploration scheme.

The method we use was developed in Refs. 1 and 2, but here some new ideas are added.

In Ref. 3 the exact inversion theory originated in Ref. 1 is presented in a systematic way. Many authors used a reference background for inversion assuming that the difference between the reference and actual backgrounds is small in some sense (e.g., Ref. 4). Here this "smallness" assumption is dropped and the inversion problem is treated globally, without using perturbation theory.

II. BASIC EQUATIONS

Let us assume that the Green's function

$$\Delta G + \omega^2 n(x) G = -\delta(x-y) \tag{2}$$

is uniquely defined for $0 < \omega < \omega_0$, and

$$\lim_{\omega \to 0} G = g_0 = (4\pi|x-y|)^{-1}. \tag{3}$$

Since $n(x)$ is assumed to be known, the function G can be considered known.

Equation (1) can be written as the integral equation

$$u(x,y,\omega) = G(x,y,\omega) + \omega^2 \int G(x,z,\omega) v(z) u(z,y,\omega) dz, \tag{4}$$

where the integral is taken over the support of v. As in Ref. 1 one can prove that (4) is solvable by iterations if ω is small enough, e.g., if

$$\omega^2 \max_x \int |G(x,z,\omega) v(z)| dz < 1,$$

and the following limit exists:

$$f(x,y) \equiv 16\pi^2 \lim_{\omega \to 0} (u - G)\omega^{-2} = \int \frac{v(z)dz}{|x-z|\,|z-y|}. \tag{5}$$

Here we use (3). The left side of (5) can be measured. The function $f(x,y)$ is our datum. Equation (5) is formally the same as Eq. (2.9) in Ref. 2. The basic new point is that we use $u - G$ as the scattered field now while in Ref. 2 the scattered field was

$$u - \exp(i\omega|x-y|)[4\pi|x-y|]^{-1},$$

which corresponds to the constant background $n(x) = 1$. Here the scattered field corresponds to the variable background $n(x)$. The important fact, which makes the method work, is that for fairly general backgrounds $n(x)$ the limit (3) does not depend on the background, although for $\omega \neq 0$ the Green's function G depends on the background.

In Ref. 2 Eq. (5) was solved analytically for the case when $x, y \in P$, and, under additional assumptions on $v(z)$, for the case when $x, y \in l = \{x: x_1 = x_2 = 0\}$, which corresponds to the model logging problem. Here we study the basic equation (5) in the case when $x \in l_1$ and $y \in l_2$, which corresponds to well to well exploration. We assume that

$$n = n(z_3). \tag{6}$$

The assumption $n = n(z_3)$ is convenient in geophysical applications. Equation (5) is a linear integral equation of the first kind for v and can be solved numerically by means of a regularization method.

III. INVERSION FOR A VARIABLE BACKGROUND

(i) First, consider the problem when the manifold, on which one measures the data, is $M = P = \{x: x_3 = 0\}$.

There are two basic steps in solving the inverse well-to-well problem. First, one computes the data $f(x,y)$. This computation needs the measured field u and the computed (or

 0022-2488/86/112687-03$02.50

calculated analytically) field G. Second, one solves numerically Eq. (5) with the data found in the first step. In this paper we discuss the first step under the assumption that the background is

$$n(x)=n(x_3)=\begin{cases}1, & x_3\geqslant 0,\\ 1-[(1-n_0)/d\,]x_3, & -d<x_3\leqslant 0,\\ n_0, & x_3<-d,\quad n_0<1.\end{cases}\tag{7}$$

One can compute G, the solution to (2), for the background (7), using the Fourier transform in variables x_1 and x_2:

$$G=\frac{1}{(2\pi)^2}\int \tilde{G}\exp(i\lambda\cdot x^1)d\lambda,\quad d\lambda=d\lambda_1\,d\lambda_2,$$
$$x^1=(x_1,x_2,0)\,,\tag{8}$$
$$\tilde{G}=\int G\exp(-i\lambda\cdot x^1)dx^1\,.\tag{9}$$

Note that $G(x,y)=G(x_1-y_1;x_2-y_2;x_3,y_3)$. Taking this Fourier transform of Eq. (2) one obtains

$$\frac{d^2\tilde{G}}{dx_3^2}+(\omega^2 n(x_3)-\lambda^2)\tilde{G}=-\delta(x_3-y_3)\exp(-i\lambda\cdot y^1),\qquad \lambda^2=\lambda_1^2+\lambda_2^2\,.\tag{10}$$

Thus, $\tilde{G}$ is the Green's function of the one-dimensional equation (10). It can be written as

$$\tilde{G}=\begin{cases}\dfrac{h_+(x_3)h_-(y_3)}{-2ik_1}\exp(-i\lambda\cdot y^1), & x_3\geqslant y_3,\\ \dfrac{h_+(y_3)h_-(y_3)}{-2ik_1}\exp(-i\lambda\cdot y^1), & y_3\geqslant x_3,\quad k_1=(\omega^2 n_0-\lambda^2)^{1/2}\end{cases}\tag{11}$$

where $k_1>0$ if $\omega^2 n_0>\lambda^2$, $k_1=i(\lambda^2-\omega^2 n_0)^{1/2}$ if $\omega^2 n_0<\lambda^2$, and $h_\pm$ are the solutions to the equation

$$h''+[\omega^2 n(x_3)-\lambda^2]h=0\,,\tag{12}$$

which are given for $x_3>0$ by the formula

$$h_\pm=\exp(\pm ik_0x_3),\quad x_3>0,\quad k_0=(\omega^2-\lambda^2)^{1/2}\,.\tag{13}$$

If $d=0$, then

$$h_+=\frac{1}{2}\left(1+\frac{k_1}{k_0}\right)\exp(ik_0x_3)+\frac{1}{2}\left(1-\frac{k_1}{k_0}\right)\times\exp(-ik_0x_3),\quad x_3<0,\quad k_0=\sqrt{\omega^2-\lambda^2}\,,\tag{14}$$

$$h_-=\frac{1}{2}\left(1-\frac{k_1}{k_0}\right)\exp(ik_0x_3)+\frac{1}{2}\left(1+\frac{k_1}{k_0}\right)\exp(-ik_0x_3),\quad x_3<0\,.\tag{15}$$

If $d>0$, then expressions for $h_\pm$ are more complicated. In the interval $0>x_3>-d$, Eq. (12) is

$$h''+(\omega^2-\lambda^2+[\omega^2(1-n_0)/d\,]x_3)h=0\,,\tag{16}$$

and its two linearly independent solutions can be written by means of Bessel functions. Indeed, let $\alpha=\omega^2-\lambda^2$, $\beta=\omega^2(1-n_0)^{-1}$, $\xi=\alpha+\beta x_3$. Then (16) takes the form

$$h_{\xi\xi}+\beta^2\xi h=0.\tag{17}$$

This equation is solvable in Bessel functions (Ref. 5, formula 8.491.7)

$$h(\xi)=A\xi^{1/2}J_{1/3}\left(\frac{2\beta}{3}\xi^{3/2}\right)+B\xi^{1/2}J_{-1/3}\left(\frac{2\beta}{3}\xi^{3/2}\right),\qquad \xi=\alpha+\beta x_3\,,\tag{18}$$

where A and B are arbitrary constants and $J_p(x)$ is the Bessel function. Therefore one can find $h_\pm(x_3)$ explicitly: $h_\pm$ are given by (13) in the region $x_3>0$, by

$$h_\pm=A_\pm\xi^{1/2}J_{1/3}\left(\frac{2\beta}{3}\xi^{3/2}\right)+B_\pm\xi^{1/2}J_{-1/3}\left(\frac{2\beta}{3}\xi^{3/2}\right),\qquad -d<x_3<0,\quad \xi=\alpha+\beta x_3\,,\tag{19}$$

and by

$$h_\pm=C_\pm\exp(ik_1x_3)+D_\pm\exp(-ik_1x_3),\quad x_3<-d\,.\tag{20}$$

The coefficients $A_\pm, B_\pm, C_\pm, D_\pm$ are to be found from the requirement that the functions $h_\pm(x_3)$ and $h'_\pm(x_3)=dh_\pm/dx_3$ are continuous at $x_3=0$ and at $x_3=-d$. This requirement leads to a linear system of four equations for each of the sets of four coefficients (A_+, B_+, C_+, D_+) and (A_-, B_-, C_-, D_-). The coefficients can be written explicitly. The formulas for the coefficients are complicated and therefore omitted here. Our point is that the Green's function $\tilde{G}$ can be written analytically for $n(x_3)$ given by (7). If $\tilde{G}$ is known then G is given explicitly by (8). One can check that (3) holds. Therefore one can compute $f(x,y)$ in (5) and obtain the linear integral equation (5) for $v(z)$. If $x, y\in P$ this equation is solved analytically in Ref. 2.

(ii) Let us now consider the case when $v=v(x_1,x_3)$. This means that v is not compactly supported: it is constant as a function of x_2. Nevertheless one can prove that in this case Eq. (5) is still valid, and the limit on the left side of (5) exists. Let us assume that $x\in l_1=\{x:\ x_1=-a,\ x_2=0,\ x_3<0\}$ and $y\in l_+=\{x:\ x_1=a,\ x_2=0,\ x_3<0\}$. Equation (5) takes the form

$$\int dz_1\,dz_3 v(z_1,z_3)T(x,y,z_1,z_3)=f(x,y),\quad x\in l_-,\quad y\in l_+\,,\tag{5'}$$

where

$$T(x,y,z_1,z_3)=\int_{-\infty}^{\infty}\frac{dz_2}{[(x+a)^2+z_3^2+z_2^2]^{1/2}[(y-a)^2+z_3^2+z_2^2]^{1/2}}$$
$$=\frac{\pi}{\sqrt{(y-a)^2+z_3^2}}K\left(\frac{\sqrt{(y-a)^2-(x+a)^2}}{\sqrt{(y-a)^2+z_3^2}}\right)\quad\text{if } (y-a)^2>(x+a)^2, \tag{21}$$

where formula 3,152 from Ref. 5 is used and $K(q)=F(\pi/2,q)$, and where

$$F(\phi,k)=\int_0^{\phi}\frac{d\phi}{\sqrt{1-q^2\sin^2\phi}}$$

is the elliptic integral of the first kind. If $(x+a)^2>(y-a)^2$, then the right side of (21) is

$$\frac{\pi}{[(x+a)^2+z_3^{2\,1/2}]}K\left(\left[\frac{(x+a)^2-(y-a)^2}{(x+a)^2+z_3^2}\right]^{1/2}\right).$$

Equation (5′) is a linear equation of the first kind for v. The integral in (5′) is taken over the support of v in the plane z_1 and z_3. This equation can be solved numerically by a regularization method.

IV. INVERSION FOR $n(x_3)$ OF THE SEISMIC DATA

Assume that $v=0$, $y=0$, and $n(x)=n(x_3)$ in (1), $x_3=z$ in what follows. Take the Fourier transform of (1) in x^1 to get

$$\frac{d^2\tilde{G}}{dz^2}+(\omega^2n(z)-\lambda^2)\tilde{G}=-\delta(z)\,. \tag{22}$$

The data are the values $u(x^1,z=0,y=0,\omega)$, $0<\omega<\omega_0$, where ω_0 is a (small) fixed number. The Fourier transform $\tilde{u}(\lambda,z=0,y=0,\omega)=\tilde{G}(\lambda,0,\omega)$ is known. Let us assume that $n(z)=1$ if $z>0$, $n(z)=n_0$ if $z<-d$. We wish to find $n(z)$ in the region $-d<z<0$ from the knowledge of $\tilde{G}(\lambda,0,\omega)$. It follows from (22) that

$$\tilde{G}(z)=\Gamma(z)+\omega^2\int_{-\infty}^{\infty}\Gamma(z-z')n(z')\tilde{G}\,dz',$$
$$\Gamma(z)=\frac{\exp(-\lambda|z|)}{2\lambda}. \tag{23}$$

Equation (23) is uniquely solvable by iteration if $\lambda^2>\omega^2\max n(z)$. We have

$$f(z,\lambda)\equiv 4\lambda^2\lim_{\omega\to 0}(\tilde{G}-\Gamma)\omega^{-2}$$
$$=\int_{-\infty}^{\infty}\exp(-\lambda|z-z'|)n(z')\exp(-\lambda|z'|)dz'\,. \tag{24}$$

Set $z=0$ in (24) to get

$$f(0,\lambda)\equiv\phi(\lambda)$$
$$=\int_0^{\infty}\exp(-2\lambda|z'|)dz'+n_0$$
$$\times\int_{-\infty}^{-d}\exp(-2\lambda|z'|)dz'$$
$$+\int_{-d}^{0}\exp(-2\lambda|z'|)n(z')dz'\,. \tag{25}$$

Denote

$$F(p)=\phi(\lambda)-\frac{1}{2\lambda}-\frac{\exp(-2\lambda d)}{2\lambda},\quad \lambda=\frac{p}{2},$$

and $n(-t)=f(t)$. Then (25) can be written as

$$\int_0^d\exp(-pt)f(t)dt=F(p)\,.$$

The function $F(p)$ is defined for $p>0$. We can compute $f_N(t)$, which approximates $f(t)$ with a prescribed accuracy in $L^2[0,d]$ if $f\in L^2$ [in $C[0,d]$ if $f\in C([0,d])$] using the data $F(p)$ on the interval $0<p<b$, where $b>0$ is a fixed number, by the formulas[3,6]

$$f_N(t)=\int_0^b dp\,F(p)H_N\left(p-\frac{b}{2}\right)\exp(pt)\,,$$

where

$$H_N(p)=(2\pi)^{-1}\int_{-\infty}^{\infty}\delta_N(iy)\exp(-iyp)dy\,,$$
$$H_N(p)=0,\ \text{if } p>b/2 \text{ or } p<-b/2\,,$$

and

$$\delta_N(iy)=\left(\frac{N}{4\pi d^2}\right)^{1/2}\left(1+\frac{y^2}{4d^2}\right)^N$$
$$\times\left(\frac{\sin(yb/2(2N+q))}{yb/2(2N+q)}\right)^{2N+q},\quad q\geqslant 1\,,$$

where $q\geqslant 1$ is an arbitrarily fixed number. The following estimate holds:

$$\|f_N(t)-f(t)\|\leqslant\epsilon(N)\to 0 \text{ as } N\to\infty\,,$$

where the norm is $L^2([0,d])(C([0,d]))$ if $f\in L^2([0,d])(C([0,d]))$.

[1] A. G. Ramm, Phys. Lett. A **99**, 258 (1983).
[2] A. G. Ramm, Inverse Problems **1**, 133 (1985).
[3] A. G. Ramm, *Scattering by Obstacles* (Reidel, Dordrecht, 1986).
[4] V. H. Weston, J. Math. Phys. **22**, 2523 (1981).
[5] I. Gradshteyn and I. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1965).
[6] A. G. Ramm, "Signal estimation from incomplete data," J. Math. Anal. Appl. (to appear).

A note on topology of supermanifolds

Shuichi Matsumoto
Department of Mathematics, University of the Ryukyus, Okinawa, Japan
Kiyotaka Kakazu
Department of Physics, University of the Ryukyus, Okinawa, Japan

(Received 2 April 1986; accepted for publication 18 June 1986)

A topological structure on the space of supernumbers is introduced, with which this space becomes a Fréchet space. The definition of supermanifold and "superprojective space" are given. The superprojective space is one of paracompact Hausdorff supermanifolds.

I. INTRODUCTION

There are some different ways of defining the topological structure on the space of supernumbers.[1–4] However, it seems, at least to the authors, that each way has its merits and demerits. The topological structure of Rogers[3] is richer in that it is nontrivial in the soul part of supernumbers. On the other hand, the coarser topologies used by DeWitt[1] and Batchelor[2] are much easier to study, because there is no necessity for considering the convergence of norms. In this article we attempt to define a topology that has both merits mentioned above.

In Sec. II we define our topology; Sec. III contains a discussion of the differentiability of functions of supernumbers. In Sec. IV, we give the definition of supermanifold and consider "superprojective space" as an example of a paracompact Hausdorff supermanifold.

II. SUPERNUMBERS

Let Λ_∞ be an infinite-dimensional Grassmann algebra over complex number field $\mathbf{C}$. Let ζ^a, $a = 1,2,...$, be a set of generators for Λ_∞, which anticommute; $\zeta^a\zeta^b = -\zeta^b\zeta^a$ for all a and b. The elements of Λ_∞ are called supernumbers. Every supernumber can be expressed in the form

$$u = \sum_{n=0}^{\infty} \sum_{a_1<\cdots<a_n} u_{a_1\cdots a_n}\zeta^{a_1}\cdots\zeta^{a_n} = u_B + u_S \quad (u_{a_1\cdots a_n}\in\mathbf{C}),$$

where u_B is an ordinary complex number and

$$u_S = \sum_{n=1}^{\infty} \sum_{a_1<\cdots<a_n} u_{a_1\cdots a_n}\zeta^{a_1}\cdots\zeta^{a_n}.$$

The numbers u_B and u_S are called the body and the soul of the supernumber u, respectively.

If a supernumber u has the form

$$u = \sum_{n=0}^{\infty} \sum_{a_1<\cdots<a_{2n}} u_{a_1\cdots a_{2n}}\zeta^{a_1}\cdots\zeta^{a_{2n}},$$

then it is called a c-number. If u has the form

$$u = \sum_{n=0}^{\infty} \sum_{a_1<\cdots<a_{2n+1}} u_{a_1\cdots a_{2n+1}}\zeta^{a_1}\cdots\zeta^{a_{2n+1}},$$

then it is called an a-number. The set of all c-numbers is a commutative subalgebra of Λ_∞, which is denoted by C_c. The set of all a-numbers is denoted by C_a.

Let N_∞ denote the set of sequences $\mathbf{a} = (a_1,...,a_n)$ where the a's are natural numbers and $a_1 < \cdots < a_n$, then the elements $\zeta^{a_1}\cdots\zeta^{a_n}$, $(a_1,...,a_n)\in N_\infty$, form a basis for Λ_∞. For an element $\mathbf{a} = (a_1,...,a_n)$ in N_∞, we define $\zeta^{\mathbf{a}}$ by $\zeta^{a_1}\cdots\zeta^{a_n}$. With this convention every supernumber can be expressed in the form

$$u = \sum_{\mathbf{a}\in N_\infty} u_{\mathbf{a}}\zeta^{\mathbf{a}}.$$

Let N_m $(m = 1,2,...)$ be subsets of N_∞ so that (1) N_m is a finite set for all m, and (2) ϕ is in N_1, $N_m \subset N_n$ for $m < n$, and

$$N_\infty = \bigcup_{m=1}^{\infty} N_m.$$

(Here ϕ represents the empty sequence in N_∞.) Let p_m: $\Lambda_\infty \to \mathbf{R}$ $(m = 1,2,...)$ be defined by

$$p_m(u) = \sum_{\mathbf{a}\in N_m} |u_{\mathbf{a}}|,$$

where

$$u = \sum_{\mathbf{a}\in N_\infty} u_{\mathbf{a}}\zeta^{\mathbf{a}}.$$

Then the p's are seminorms on Λ_∞ and we have $p_1(u) \leqslant p_2(u) \leqslant \cdots$ for all u in Λ_∞. Set

$$V_{m,l}(u) = \{v\in\Lambda_\infty \mid p_m(v-u) < 1/l\}.$$

Here u is in Λ_∞ and m, l are natural numbers. Then Λ_∞ has a topology such that $\{V_{m,l}(u)\}_{m,l}$ is a fundamental neighborhood system of u. As is easily seen, this topology admits the second countability axiom. Furthermore we can make Λ_∞ into a metric space by defining a metric

$$d(u,v) = \sum_{m=1}^{\infty} \frac{(1/2^m)p_m(u-v)}{1+p_m(u-v)}.$$

A sequence $\{u^j\}_{j=1,2,...}$ of supernumbers converges to a supernumber u iff $\lim_{j\to\infty} |u^j_{\mathbf{a}} - u_{\mathbf{a}}| = 0$ for all $\mathbf{a}$ in N_∞. Hence the metric d is complete. Thus with our topology Λ_∞ becomes a Fréchet space.[5] It is the purpose of this article to demonstrate that our topology has desirable properties for analysis over supernumbers.

III. SUPERANALYTIC FUNCTIONS

In this section we discuss an analytic theory of functions of c-numbers and a-numbers.

Consider first C_c. Let U be an open neighborhood of an element u in C_c and let f be a mapping from U into Λ_∞.

 0022-2488/86/112690-03$02.50

Then f is said to be superanalytic at u if there exists a function $\epsilon(h)$ $(h\in C_c)$ and an element α in Λ_∞ such that

(1) $f(u+h)-f(u)=\alpha h+\epsilon(h)h \quad (h\in C_c)$,

(2) $\epsilon(h)\to 0 \quad (h\to 0)$ with respect to our topology.

The supernumber α is called the derivative of f with respect to u and denoted by $(d/du)f(u)$. If f is superanalytic at every point u in U then f is said to be superanalytic on U.

Our definition of a superanalytic function may be regarded as a mathematically rigorous form of that of DeWitt.[1] Actually we can prove exactly the following theorems, which are described by DeWitt without proofs.

Theorem 1: Let f be any ordinary complex holomorphic function on an open set V in $\mathbf{C}$. Then we can extend f uniquely to a superanalytic function

$$f(u)=\sum_{n=0}^{\infty}\frac{1}{n!}f^{(n)}(u_B)u_S{}^n, \tag{1}$$

where $f^{(n)}$ denotes the nth derivative of f and the definition is valid for all c-numbers u such that u_B is in V.

Theorem 2: The general form of a superanalytic function on a connected open set of C_c is

$$f(u)=\sum_{\mathbf{a}\in N_\infty} f_{\mathbf{a}}(u)\zeta^{\mathbf{a}},$$

where the $f_{\mathbf{a}}(u)$ are functions like (1).

Here we give a proof only of the first theorem.

Proof of Theorem 1: We first note that the right-hand side of (1) converges with respect to our topology. We have

$$\begin{aligned} f(u+h)-f(u) &= \sum_{n=0}^{\infty}\frac{1}{n!}\{f^{(n)}(u_B+h_B)(u_S+h_S)^n-f^{(n)}(u_B)u_S{}^n\} \\ &= \sum_{l=1}^{\infty}\left\{\sum_{m=0}^{\infty}\frac{1}{m!}f^{(l+m)}(u_B)u_S{}^m\right\}\left\{\sum_{j+k=l}\frac{1}{j!k!}h_B{}^j h_S{}^k\right\} \\ &= \sum_{l=1}^{\infty}\frac{1}{l!}\left\{\sum_{m=0}^{\infty}\frac{1}{m!}f^{(l+m)}(u_B)u_S{}^m\right\}h^l. \end{aligned} \tag{2}$$

Therefore, f is a superanalytic function on $U=\{u \text{ in } C_c \mid u_B \text{ is in } V\}$. To prove the uniqueness of the extension, we let g be a superanalytic function on U such that $g(z)=0$ for all z in V. Expand g in the form

$$g(u)=\sum_{\mathbf{a}\in N_\infty} g_{\mathbf{a}}(u)\zeta^{\mathbf{a}} \quad (g_{\mathbf{a}}:\ U\to\mathbf{C}).$$

For a natural number m, let $\Lambda_{\infty,m}$ be the finite-dimensional subspace of Λ_∞ spanned by the set $\{\zeta^{\mathbf{b}}\mid \mathbf{b} \text{ in } N_m\}$. Then we can see that $\mathbf{C}\subset\Lambda_{\infty,1}\subset\Lambda_{\infty,2}\subset\cdots$ and $\bigcup_{m=1}^{\infty}\Lambda_{\infty,m}$ is a dense subset of Λ_∞. Now, for fixed $\mathbf{a}$ and m, we restrict the domain of $g_{\mathbf{a}}(u)$ to $U\cap\Lambda_{\infty,m}$. Then we can regard the function $g_{\mathbf{a}}(u)$ as an ordinary holomorphic function of the complex variables $u_{\mathbf{b}}$, $\mathbf{b}\in N_m$. Since $(d/du)g(u)=0$ for all $u\in V$, the ordinary derivatives of $g_{\mathbf{a}}(u)$ at u in V with respect to $u_{\mathbf{b}}$, $\mathbf{b}\in N_m$, are all zero. Hence $g_{\mathbf{a}}(u)=0$ for all u in $U\cap\Lambda_{\infty,m}$. Thus we obtained that $g(u)=0$ for all u in $U\cap(\bigcup_{m=1}^{\infty}\Lambda_{\infty,m})$. Since $\bigcup_{m=1}^{\infty}\Delta_{\infty,m}$ is dense in Λ_∞, $g(u)=0$ for all u in U. Q.E.D.

Rogers[3] uses a l_1 norm to define her topology on the space of supernumbers. Therefore, there are frequent occasions when it is imperative to verify convergence of norms. On the other hand we use seminorms $\{p_m(u)\}_m$ to define our topology, consequently there is no such problem. As a result we may deal much more easily with an infinite series of supernumbers. [For example, see (2).]

Using Theorem 2, we can show the following theorem.

Theorem 3: Let f be superanalytic on a connected open set U of C_c. Set $U_B=\{z\in\mathbf{C}\mid z=u_B \text{ for some } u \text{ in } U\}$, and set $\tilde{U}=\{u\in C_c\mid u_B \text{ is in } U_B\}$. Then f has the unique extension to a superanalytic function on $\tilde{U}$.

A superanalytic function of a-numbers is defined similarly:

$$f(u+h)-f(u)=\alpha h+\epsilon(h)h=h\beta+h\rho(h) \quad (h\in C_a),$$

$$\epsilon(h)\to 0 \text{ and } \rho(h)\to 0 \quad (h\to 0).$$

The supernumbers α and β are called, respectively, the right and left derivatives of f and are denoted by $f(u)(\overleftarrow{d}/du)$ and $(\vec{d}/du)f(u)$.

This definition is also a mathematically rigorous form of DeWitt's one. We can show the following theorem.

Theorem 4: The general form of a superanalytic function on a connected open set U in C_a is

$$f(u)=a+bu \quad (u\in U),$$

where a and b are arbitrary supernumbers. Therefore, any superanalytic function on a connected open set in C_a is extensible to a superanalytic function on C_a.

Let ψ be a mapping from an open set of ${C_c}^m\times{C_a}^n$ to ${C_c}^{m'}\times{C_a}^{n'}$. This mapping is said to be superanalytic if the coordinates u'^j $(j=1,2,\ldots,m'+n')$ of the image point $\psi(u)$ are superanalytic functions of the coordinates u^i $(i=1,2,\ldots,m+n)$ of the point u.

IV. SUPERMANIFOLD AND SUPERPROJECTIVE SPACE

Since DeWitt and Batchelor use a coarser topology on the space of supernumbers, their supermanifolds are not even Hausdorff. However our supermanifold may be assumed to be Hausdorff. Our definition of supermanifolds is as follows.

A superanalytic manifold M of dimension (m,n) is a Hausdorff topological space with a collection of pairs $(U_i,\ \psi_i)$, where

(1) each U_i is an open set of M and its associated ψ_i is a homeomorphism of U_i onto an open set in ${C_c}^m\times{C_a}^n$,

(2) $\cup_i U_i=M$,

(3) $\psi_i\cdot\psi_j^{-1}$ is superanalytic on $\psi_j(U_i\cap U_j)$.

In the rest of this section we will consider a generalization of the complex projective space $P^m(\mathbf{C})$.

For non-negative integers m and n, we set

$$\Lambda^{m+1,n}=\{u_1,\ldots,u_{m+1},v_1,\ldots,v_n) \text{ in } {C_c}^{m+1}\times{C_a}^n \mid (u_{1B},\ldots,u_{m+1\,B})\neq 0\}.$$

Let

$$C_c{}^*=\{\lambda \text{ in } C_c\mid \lambda_B\neq 0\},$$

i.e., $C_c{}^*$ is the set of all invertible c-numbers. We define an equivalence relation on $\Lambda^{m+1,n}$ as follows: $(u,v)\equiv(u'v')$ if there exists an element $\lambda\in C_c{}^*$ such that $u'=\lambda u$ and $v'=\lambda v$.

Let $P^{m,n}(\Lambda_\infty)$ be the set of all equivalence classes of $\Lambda^{m+1,n}$. For (u,v) in $\Lambda^{m+1,n}$, let $[u,v]$ be the equivalence class of (u,v). Give $P^{m,n}(\Lambda_\infty)$ the topology that makes the mapping π: $(u,v)\to[u,v]$ continuous. Then it is seen that π is an open mapping and $P^{m,n}(\Lambda_\infty)$ is a Hausdorff topological space that admits the second countability axiom. Let

$$U_i=\{[u,v]|(u,v) \text{ in } \Lambda^{m+1,n},\ u_{iB}\neq 0\}$$

$(i=1,2,...,m+1)$ and let ψ_i be the mapping from U_i to $C_c{}^m\times C_a{}^n$ defined by

$$\psi_i([u,v])=\left(\frac{u_1}{u_i},...,\frac{u_{i-1}}{u_i},\ \frac{u_{i+1}}{u_i},...,\frac{u_{m+1}}{u_i},\ \frac{v_1}{u_i},...,\frac{v_n}{u_i}\right).$$

Then with the collection $\{(U_i,\ \psi_i)\}_i$, we obtain the supermanifold $P^{m,n}(\Lambda_\infty)$ of dimension (m,n).

When we consider integration over $P^{m,n}(\Lambda_\infty)$ (see Ref. 6), the following theorem is useful for us.

Theorem 5: $P^{m,n}(\Lambda_\infty)$ is paracompact.

Proof: Since $P^{m,n}(\Lambda_\infty)$ is a Hausdorff space that admits the second countability axiom, it is enough to show that the space is regular. Let p_0 and A be a point and a closed subset of $P^{m,n}(\Lambda_\infty)$, respectively, such that p_0 is not in A. Let q_0 be a point in $\Lambda^{m+1,n}$ such that $\pi(q_0)=p_0$. Then q_0 is in the open set $\pi^{-1}(A^c)$. On the other hand, $\Lambda^{m+1,n}$ has a metric ρ. There is a positive number δ such that $\{q \text{ in } \Lambda^{m+1,n}|\rho(q,q_0)<\delta\}$ does not have the intersection with $\pi^{-1}(A)$. Let

$$U=\{\lambda q|\lambda \text{ in } C_c{}^*,\ \rho(q,q_0)<\delta/2\}.$$

Then

$$\overline{U}=\{\lambda q|\lambda\in C_c{}^*,\ \rho(q,q_0)\leqslant\delta/2\}$$

and $\pi^{-1}(A)\subset\overline{U}^c$. Therefore, we obtain that p_0 is in the open set $\pi(U)$, A is contained in the open set $\pi(\overline{U}^c)$, and the open sets $\pi(U)$, and $\pi(\overline{U}^c)$ have no intersection. Thus $P^{m,n}(\Lambda_\infty)$ is regular. Q.E.D.

[1] B. DeWitt, *Supermanifolds* (Cambridge U.P., Cambridge, 1984).
[2] M. Batchelor, Trans. Am. Math. Soc. **253**, 329 (1979); **258**, 257 (1980).
[3] A. Rogers, J. Math. Phys. **21**, 1352 (1980).
[4] Y. Manin, in *Lecture Note in Mathematics*, Vol. 1111 (Springer, New York, 1985).
[5] J. Dieudonné and L. Schwartz, Ann. Inst. Fourier (Grenoble) **1**, 61 (1949).
[6] S. Matsumoto and K. Kakazu, in preparation.

Killing tensors in spaces of constant curvature

G. Thompson
Department of Pure Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

(Received 7 November 1984; accepted for publication 2 July 1986)

A Killing tensor is one possible way of generalizing the notion of a Killing vector on a Riemannian or pseudo-Riemannian manifold. It is explained how Killing tensors may be identified with functions that are homogeneous polynomials in the fibers on the associated cotangent bundle. As such, Killing tensors may be identified with first integrals of the Hamiltonian geodesic flow, which are homogeneous polynomials in the momenta. Again using this identification, it is shown that in flat spaces the dimension of the vector space of Killing tensors is maximal and that the Killing tensors are generated by the Killing vectors. Finally, using Riemann's model for the metric in spaces of constant curvature, a comparison argument is used to show that similar results are valid in that more general context.

I. INTRODUCTION

A Killing tensor field on a Riemannian or pseudo-Riemannian manifold is one possible way of generalizing the familiar concept of a Killing vector. Killing tensors have been studied extensively by relativists, since they are essentially identical with first integrals of the dynamical field that are polynomial in the momentum variables.[1–4] More recently, it has been pointed out that Killing tensors also play a role in the search for first integrals in the context of classical mechanics.[5] Indeed, many authors have in effect solved Killing's equations, without explicitly isolating the notion of a Killing tensor.[6–9]

In this paper I shall primarily be concerned with the structure of Killing tensors in spaces of constant curvature. Specifically, in Sec. IV I shall show that in such a space the number of independent Killing tensors is maximal and that every Killing tensor consists of a sum of symmetrized products of Killing vectors. The proof of this result is somewhat indirect; I establish the result first for flat spaces and then use a comparison argument based on Riemann's model for the metric in spaces of constant curvature, to obtain the general result. In Secs. II and III I give some background material on Killing tensors, some of which is not well known. In particular, in Sec. II I shall describe how $S(M)$, the graded algebra of symmetric, contravariant tensor fields on a smooth manifold M, carries naturally the structure of a graded Lie algebra that extends the Lie structure of the vector fields on M. When M is a Riemannian or pseudo-Riemannian manifold, this Lie structure enables one to give an elegant characterization of Killing tensors as those elements of $S(M)$ that commute with the (contravariant) metric tensor G. I shall also describe how $S(M)$ with its two algebraic structures may be identified with a subspace of $F(T^*M)$, the ring of smooth functions on the cotangent bundle of M. This isomorphism gives a very convenient description of Killing tensors as functions on T^*M that are homogeneous polynomials in the fibers, which is used throughout the paper. In Sec. III I explain the relationship between Killing tensors and first integrals in mechanics, in terms of the machinery developed in Sec. II.

Before giving my notational conventions I should like to mention two points that are potential sources of confusion in the paper. First of all, I shall have occasion several times to deal with Riemannian or pseudo-Riemannian manifolds (M,g) that are flat; by this I mean that the Riemann tensor determined by g is zero. It follows that (M,g) is *locally* isometric to $\mathbb{R}^n$ with an inner product whose signature is the same as that of g.[10–12] However, such an isometry need only be locally defined, the obstruction to extending it to a global isometry being the fundamental group $\pi_1(M)$ of M. In the sequel, several results involving flat spaces are to be interpreted as global theorems in case $\pi_1(M)$ is trivial, or local theorems otherwise. In particular, it is only in the former case that it is valid to speak of global, linear coordinates on M.

The second possible point of confusion concerns the Lie algebra of Killing tensors on a Riemannian or pseudo-Riemannian manifold (M,g). $S(M)$ with its commutative algebra structure may be considered to be either a graded or filtered object.[13] In terms of the isomorphism mentioned above with the subspace of $F(T^*M)$, this amounts to the distinction between functions on T^*M that are homogeneous or inhomogeneous polynomials in the fibers, respectively. In fact it is only in Sec. III that I shall briefly wish to allow for inhomogeneous polynomials because there I shall consider mechanical Hamiltonians in which there is a potential function in addition to the kinetic energy term. Otherwise, I shall be concerned with Killing tensors that correspond to first integrals which are homogeneous in the momentum variables. I shall, however, show in Sec. IV that the components of a Killing tensor in flat space are polynomial and, moreover, that each homogeneous part is also a Killing tensor. Thus the term "homogeneous Killing tensor" will be used only in flat space and will correspond to a function on T^*M that is a homogeneous polynomial in both the position and momentum variables.

As regards notation, M will denote a smooth, that is C^∞, manifold of dimension m. Further $\mathcal{L}$ and ∇ will denote, respectively, Lie and covariant differential operators, ∇ corresponding to the Riemannian or pseudo-Riemannian metric g. The graded algebra of symmetric and skew-symmetric contravariant tensor fields on M will be represented by $S(M)$ and $\Lambda(M)$, respectively. I shall use $\pi: T^*M \to M$ to denote the cotangent fibration of M and $\{f,g\}$ for the Pois-

 0022-2488/86/112693-07$02.50

son bracket of two elements f and g of $F(T^*M)$, the ring of smooth functions on T^*M. In addition, I shall employ the notation of the classical tensor calculus with range and summation conventions in operation unless otherwise stated (in particular in Proposition 4.6) and $A_{(i_1\cdots i_n)}$ will denote the symmetric part of the valence n covariant tensor field $A_{i_1\cdots i_n}$.

II. KILLING TENSORS

Let (M,g) be a Riemannian or pseudo-Riemannian manifold, with M of dimension m. A symmetric, covariant tensor field K on M of valence n is said to be a Killing tensor if

$$(\nabla_{X_1}K)(X_2,\ldots,X_{n+1}) + (\nabla_{X_2}K)(X_3,\ldots,X_{n+1},X_1) + \cdots + (\nabla_{X_{n+1}}K)(X_1,\ldots,X_n) = 0\,, \tag{2.1}$$

for any collection of $n+1$ vector fields $X_1,X_2,\ldots,X_{n+1}$ on M. Equivalently, introducing a local coordinate system (x^i), K satisfies the index condition

$$K_{(i_1\cdots i_n;i_{n+1})} = 0\,. \tag{2.2}$$

For $n=1$, (2.1) or (2.2) reduces to the usual definition of a (covariant) Killing vector (or one-form) on M.[10] It is well known that for $n=1$, (2.1) is equivalent to the following condition[11,14]

$$\mathcal{L}_{\tilde{K}}g = 0\,, \tag{2.3}$$

where now $\tilde{K}$ is the contravariant vector field dual to K via the metric g. For future reference I should also recall that in Euclidean n-space with the standard metric whose components relative to the natural coordinates are δ_{ij}, the solution of Killing's equations [(2.2) with $n=1$ or (2.3) thought of as a first-order system of partial differential equations for the unknown functions K^i] leads to the following basis for the Lie algebra of Killing vectors:

$$T_i = \frac{\partial}{\partial x^i} \quad (1\leqslant i\leqslant m)\,, \tag{2.4}$$

$$R_{ij} = x^j\frac{\partial}{\partial x^i} - x^i\frac{\partial}{\partial x^j} \quad (1\leqslant i<j\leqslant m)\,. \tag{2.5}$$

T_i and R_{ij} are the well-known symmetries which correspond to the conservation laws of linear and angular momentum via Noether's theorem.[15]

I shall next describe how (2.3) may be generalized and thereby give an alternative, intrinsic definition of Killing tensors.[4] To do this, it is necessary to consider Schouten's bracket on the algebra of symmetric contravariant tensor fields $S(M)$ on M. The collection of (contravariant) vector fields on M has, of course, the structure of a real Lie algebra under the Lie bracket of vector fields. This bracket structure can be extended as a biderivation (of degree zero) to $S(M)$. Specifically, one extends the bracket to decomposable, symmetric contravariant tensor fields and then extends by linearity to the whole of $S(M)$. One easily checks that this endows $S(M)$ with a well-defined structure of a real Lie algebra. If $A,B\in S(M)$ with valence p and q, respectively, the bracket of A with B, which I denote by $[A,B]$ and which is of valence $p+q-1$, is given in component form by

$$[A,B]^{j_1\cdots j_{p+q-1}} = pA^{i(j_1\cdots j_{p-1}}B^{j_p\cdots j_{p+q-1})}_{\ \ ,i} - qB^{i(j_1\cdots j_{q-1}}A^{j_q\cdots j_{p+q-1})}_{\ \ ,i}\,. \tag{2.6}$$

[The Lie structure on $S(M)$ just described, which is known as the Schouten bracket, should be distinguished from a structure on $\Lambda(M)$ (the algebra of skew-symmetric contravariant tensor fields on M), which goes by the same name. This structure is also of current geometric interest and is obtained by extending the Lie product to $\Lambda(M)$ as a biderivation of degree 1 (see Ref. 16).] In terms of the Schouten bracket, a (contravariant) Killing tensor is simply one that commutes with g.

In addition to its Lie structure, $S(M)$ of course enjoys another algebraic structure, namely, it is a commutative algebra with respect to the symmetrized tensor product, which I shall denote by $A\odot B$ [$A,B\in S(M)$]. These two algebraic structures are related by

$$[A\odot B,C] = [A,C]\odot B + A\odot[B,C]\,, \tag{2.7}$$

where $A,B,C\in S(M)$; that is to say, the Schouten bracket acts as a derivation with respect to the commutative algebraic structure on $S(M)$; indeed that is how it is defined.

I shall now consider an apparently very different way of viewing the algebraic structures on $S(M)$ just described, in terms of the geometry of the cotangent bundle T^*M associated to M. The starting point for this is the observation that an element A of $S(M)$ defines naturally a real-valued function on T^*M (a homogeneous polynomial in the momenta), which I shall denote by a. Specifically, if A has valence n,

$$a(p) = A(p,\ldots,p)\,, \tag{2.8}$$

where $p\in T^*M$ and there are n arguments on the right-hand side of (2.8). If (x^i,p_i) is a coordinate system adapted to T^*M, then (2.8) may be written in component form as

$$a(x^i,p_i) = A^{i_1\cdots i_n}(x^i)p_{i_1}\cdots p_{i_n}\,. \tag{2.9}$$

Consider now the various algebraic structures of $F(T^*M)$ (the set of smooth, real-valued functions on T^*M). Then $F(T^*M)$ is naturally a commutative algebra, with multiplication simply the pointwise product of functions. Moreover $F(T^*M)$ is a Lie algebra with respect to the Poisson bracket $\{\ ,\ \}$ arising from the canonical symplectic structure on T^*M. These two algebraic structures are related by

$$\{ab,c\} = \{a,c\}b + a\{b,c\}\,, \tag{2.10}$$

where $a,b,c\in F(T^*M)$.

The similarity between (2.7) and (2.10) is not coincidental: indeed one has the following result.

Theorem 2.1: The map described above from $S(M)$ to $F(T^*M)$ by $A\mapsto a$ is an isomorphism (into) of both the Lie and commutative algebra structures. □

The proof of Theorem 2.1 may be effected either by using local coordinates, or, more elegantly, by establishing the result on tensors of valence 0 and 1, respectively, and then using (2.7) to deduce the general case by induction.[4]

III. KILLING TENSORS AND FIRST INTEGRALS

In this section I shall explain how Killing tensors figure into the calculation of first integrals in mechanics. This provides an application of the formalism developed in Sec. II and, in particular, of Theorem 2.1. I shall consider systems of "standard" mechanical type, that is to say, the dynamics is that of a particle on a Riemannian or pseudo-Riemannian

manifold moving under the influence of a force term which is the gradient of a potential function. As such, the Hamiltonian formulation is available.

The Hamiltonian h is a function on T^*M, say

$$h = \tfrac{1}{2} G^{ij}(x^k) p_i p_j + V(x^k) , \tag{3.1}$$

where (x^i,p_i) is an adapted system on T^*M, G^{ij} the inverse of the metric g_{ij} and V the potential function. Exploiting the isomorphism described in Sec. II, I may write

$$H = \tfrac{1}{2} G + \pi^* V , \tag{3.2}$$

where $\pi: T^*M \to M$ is the canonical submersion. The Hamiltonian vector field X_h associated to h is given by

$$X_h = G^{ij} p_j \frac{\partial}{\partial x^i} - \frac{\partial V}{\partial x^i}\frac{\partial}{\partial p_i} . \tag{3.3}$$

Suppose that now $f: T^*M \to \mathbb{R}$ is a first integral of X_h that is polynomial of degree n in the momentum variables. Then once again, by the isomorphism of $S(M)$ with the subspace of $F(T^*M)$, I may write

$$F = K_n + K_{n-1} + \cdots + K_0 , \tag{3.4}$$

where $K_0, K_1, \ldots, K_n$ are symmetric, contravariant tensors of valence $0,1,2,\ldots,n$, respectively. Now by Theorem 2.1, f is a first integral of X_h, that is $X_h f = 0$ iff the Schouten bracket $[H,F]$ vanishes. On equating grades, the latter condition is clearly equivalent to the following

$$\begin{aligned}
[G,K_n] &= 0 , \\
[G,K_{n-1}] &= 0 , \\
[G,K_{n-2}] + 2[\pi^* V, K_n] &= 0 , \\
[G,K_{n-3}] + 2[\pi^* V, K_{n-1}] &= 0 , \\
&\vdots \\
[G,K_0] + 2[\pi^* V, K_2] &= 0 , \\
[\pi^* V, K_1] &= 0 .
\end{aligned} \tag{3.5}$$

From (3.1) it is apparent that K_n and K_{n-1} are Killing tensors. Moreover, the equations decouple into two sets in such a way that it suffices to consider first integrals of purely odd and purely even degrees. Of course when the function V is zero and one is considering simply the geodesic flow of g, every K_i is Killing and corresponds to a homogeneous polynomial first integral. The conditions (3.5) have been given before in local coordinates and used to determine several new systems with polynomial first integrals.[5]

IV. KILLING TENSORS IN SPACES OF CONSTANT CURVATURE

It is immediately apparent from the intrinsic characterization of Killing tensors given in Sec. II, that the symmetrized product of two Killing tensors is also Killing. In particular this also applies to Killing vectors. The principal result in this section asserts that in a space of constant curvature, any Killing tensor is generated by Killing vectors; that is to say, a Killing tensor consists of sums of symmetrized products of Killing vectors. Before dealing with spaces of constant curvature, I shall consider flat spaces and then deduce the more general result in the shape of Proposition 4.6 and Theorem 4.7.

Proposition 4.1: Let (M,g) be a flat Riemannian or pseudo-Riemannian manifold. Then a (covariant) Killing tensor field K of degree n, is a polynomial of degree n relative to any system of linear coordinates (x^i) on M.

Proof: Let (x^i) be a system of global, linear coordinates on M [this of course is possible, if and only if (M,g) is an inner-product space, but there is an obvious corresponding local result]. Also, let differentiation relative to the Levi-Civita connection of g be denoted by a comma. I shall show by induction that for $0 \leqslant k \leqslant n$,

$$K_{i_1 \cdots i_k (j_1 \cdots j_{n-k} j_{n-k+1} \cdots j_{n+1})} = 0 . \tag{4.1}$$

Then (4.1) for $k = n$ yields the required result and for $k = 0$ the first step in the induction, which is valid because it is precisely the Killing condition (2.2).

Suppose then (4.1) holds where $0 \leqslant k \leqslant n-1$. Then

$$\begin{aligned}
0 &= K_{i_1 \cdots i_k (j_1 \cdots j_{n-k} j_{n-k+1} \cdots j_n) i_{k+1})} \\
&= (n-k/n+k) K_{i_1 \cdots i_{k+1} (j_1 \cdots j_{n-k-1} j_{n-k} \cdots j_{n+1})} \\
&\quad + (k+2/n+2) K_{i_1 \cdots i_k (j_1 \cdots j_{n-k} j_{n-k+1} \cdots j_{n+1}) i_{k+1}} ,
\end{aligned} \tag{4.2}$$

since K is totally symmetric and the order in which partial derivatives are computed is immaterial. But now by the induction hypothesis, the second term on the right-hand side of (4.2) is zero, whence so is the first and the proof is complete. □

Proposition 4.2: let (M,g) be a flat Riemannian or pseudo-Riemannian manifold. Let K be a covariant Killing tensor field of valence n (corresponding to a first integral of the geodesic flow which is homogeneous in momenta). Let $K = K_0 + K_1 + \cdots + K_n$, where the components of each K_r relative to a linear coordinate system (x^i) are homogeneous polynomials [in the position variables, such a decomposition being possible by Proposition (4.1)]. Then each K_r is also a Killing tensor.

Proof: In a system of linear coordinates, (2.2) assumes the form

$$K_{(i_1 \cdots i_n i_{n+1})} = 0 . \tag{4.3}$$

The result follows immediately by equating coefficients in (4.3). □

Suppose now that (M,g) is a Riemannian or pseudo-Riemannian manifold that is not necessarily flat. The technique used in the proof of Proposition 4.1 may be adapted to show the following: if K is a covariant Killing tensor field of valence n and (x^i) a local coordinate system on M, all derivatives of K of order greater than n are expressible in terms of derivatives of K of order less than or equal to n, together with the metric g and its derivatives. I shall illustrate this explicitly for the case $n = 2$, which is perhaps the most interesting case from a physical point of view,[1-4] and it should then be clear how one deals with Killing tensors of arbitrary valence.

Let the components of K be K_{ij} in the coordinate system (x^i). Then notice first of all that (2.2) is equivalent to

$$K_{(ij,k)} = 2K_{\alpha(i}\Gamma^{\alpha}_{jk)} , \tag{4.4}$$

where Γ^i_{jk} are the Christoffel symbols of the Levi-Civita connection associated to g. Next, it is straightforward to check the following identities:

$$K_{l(i,jk)} = \tfrac{1}{2} K_{(ij,kl)} - K_{(ij,k)l} , \tag{4.5}$$

$$K_{lm,ijk} = K_{l(i,jkm)} - 3K_{l(i,jk)m} \,. \tag{4.6}$$

From (4.5) and (4.6) one obtains, with one additional symmetrization,

$$K_{lm,ijk} = \tfrac{1}{2} K_{(lm,ijk)} - \tfrac{5}{4} \left[K_{(ij,kl)m} + K_{(ij,km)l} \right] + 3K_{(ij,k)lm} \,. \tag{4.7}$$

Hence, from (4.4) and (4.7)

$$K_{lm,ijk} = \left[K_{\alpha(i} \Gamma^{\alpha}_{jk} \right]_{,lm)} - \tfrac{5}{2} \left[K_{\alpha(i} \Gamma^{\alpha}_{jk} \right]_{,l)m} - \tfrac{5}{2} \left[K_{\alpha(i} \Gamma^{\alpha}_{jk} \right]_{,m)l} + 6 \left[K_{\alpha(i} \Gamma^{\alpha}_{jk)} \right]_{,lm} \,. \tag{4.8}$$

Since the Γ^{i}_{jk}'s depend on g and its first-order derivatives, (4.8) gives $K_{lm,ijk}$ in terms of derivatives up to second order of K_{lm}, and derivatives up to third order of g. In particular, if g is flat, coordinates can be chosen so that the right-hand side of (4.8) vanishes, which reiterates Proposition 4.1 in the case $n = 2$. To obtain formulas analogous to (4.8) for Killing tensors of valence n, one would have n identities instead of just (4.5) and (4.6); these would enable one to obtain a general formula, for which (4.7) corresponds to the case $n = 2$. In conjunction with (4.4), one could then derive a formula for $K_{i_1 \cdots i_m j_1 \cdots j_{n+1}}$ in terms of derivatives of K and g up through orders n and $n + 1$, respectively.

The preceding considerations are useful in the proof of the following theorem. In the theorem there occur two positive integer parameters: m (the dimension of the ambient manifold) and n the degree of the Killing tensor being considered. The case m arbitrary, $n = 1$ is classical[10,17] and the case m arbitrary, $n = 2$ has been proved much more recently by Kalnins and Miller.[18]

Theorem 4.3: The collection of analytic (covariant) Killing tensors of valence n on M is a finite-dimensional vector space of dimension, say K^{m}_{n}. Moreover,

$$K^{m}_{n} \leqslant \frac{(m+n-1)!(m+n)!}{(m-1)!m!n!(n+1)!}$$

and equality holds if M is flat.

Proof: Let K be an analytic covariant Killing tensor on M of valence n. (In terms of the isomorphism given by Theorem 2.1, K corresponds to a first integral of the Hamiltonian geodesic flow which is a homogeneous polynomial of degree n in the fibers of $T^{*}M$.) Now consider K together with its derivatives of all orders evaluated at some point p in M. I have already indicated how the derivatives of K of order $n + 1$ may be expressed in terms of derivatives of lower order (together with g and its derivatives which are known). Thus all derivatives of K of order greater than n may be expressed in terms of derivatives of order less than or equal to n.

Now (2.2) is equivalent to the following, of which (4.4) corresponds to the case $n = 2$,

$$K_{(i_1 \cdots i_m i_{n+1})} = n K_{\alpha(i_1 \cdots i_n} \Gamma^{\alpha}_{i_n i_{n+1})} \,. \tag{4.9}$$

Next consider (4.9) and the totality of equations obtained from it by differentiating at most n times. This may be thought of as a homogeneous system of linear equations in which the unknowns are

$$K_{i_1 \cdots i_m j_1}, K_{i_1 \cdots i_m j_1 j_2}, \ldots, K_{i_1 \cdots i_m j_1 \cdots j_{n+1}} \,.$$

There are

$$\sum_{r=0}^{n+1} \binom{m+r-1}{r} \binom{m+n-1}{n}$$

such unknowns. On the other hand, it is not difficult to check that the number of independent linear equations is

$$\sum_{r=1}^{n+1} \binom{m+r-2}{r-1} \binom{m+n}{n+1}$$

[one obtains all such independent equations by differentiating (4.9) first of all with respect to x^{j_1} $(1 \leqslant j_1 \leqslant n)$, then with respect to x^{j_1} and x^{j_2} $(1 \leqslant j_1 \leqslant j_2 \leqslant n)$, then with respect to x^{j_1}, x^{j_2}, and x^{j_3} $(1 \leqslant j_1 \leqslant j_2 \leqslant n)$ etc.]. Thus the excess of the number of unknowns over the number of equations is

$$\sum_{r=0}^{n+1} \binom{m+r-1}{r} \binom{m+n-1}{n} - \sum_{r=1}^{n+1} \binom{m+r-2}{r-1} \binom{m+n}{n+1} = \frac{(m+n-1)!(m+n)!}{(m-1)!m!n!(n+1)!} \,. \tag{4.10}$$

The functions $K_{i_1 \cdots i_n}$ and their derivatives are all being considered at a fixed point p in M. If one confines one's attention to Killing tensors that are real analytic, it should be clear that the integer appearing on the right-hand side of (4.10) gives an upper bound on the dimension of the vector space of Killing tensors; that is to say, it represents the maximum number of free parameters if one attempts a power series solution about p of (4.9), considered as a system of first-order partial differential equations. Thus

$$K^{m}_{n} \leqslant \frac{(m+n-1)!(m+n)!}{(m-1)!m!n!(n+1)!} \,.$$

In general one will have a strict inequality, because there will be integrability conditions constraining the derivatives of the $K_{i_1 \cdots i_n}$ of order $n + 2$ and higher. In a flat space these conditions are satisfied identically, that is to say, the system of partial differential equations obtained by differentiating (4.9) no more than n times, is completely integrable in the sense of the Frobenius theorem. □

The integrability conditions alluded to in Theorem 4.3 can be obtained in the case $n = 2$ as follows. Consider (4.8) and differentiate $K_{lm,ijk}$ with respect to x_n, say. The resulting expression for $K_{lm,ijkn}$ will contain derivatives of the K_{lm}'s of order three which can be eliminated using (4.8). If one insists that two such expressions $K_{lm,ijkn}$ and $K_{lm,ijnk}$ are equal, one obtains conditions relating the K_{lm}'s and their first and second derivatives. If M is flat, then (4.8) reduces simply to

$$K_{lm,ijk} = 0 \tag{4.11}$$

and hence the integrability conditions are satisfied identically. We shall also see presently, by an indirect argument, that they are also satisfied identically in spaces of constant curvature.

We have already seen from Propositions 4.1 and 4.2 that in flat spaces, Killing tensors are polynomials and that each homogeneous part of such a Killing tensor is also Killing. We may therefore ask for the dimension of the space of Killing tensors of valence n whose components are homogeneous polynomials of degree r, where $0 \leqslant r \leqslant n$; denote this number by $K^{m}_{n|r}$. Such homogeneous Killing tensors are of

considerable significance if one has in mind applications of Killing tensors to classical mechanics.[5,18]

Corollary 4.4:

$$K^m_{n|r} = (n-r-1)\,\frac{(m+r-2)!(m+n-1)!}{r!(n+1)!(m-1)!(m-2)!}.$$

Proof: Referring to the proof of Theorem 4.3, the number of unknowns of degree r is $\binom{m+r-1}{r}\binom{m+n-1}{n}$ and the number of independent linear equations they satisfy is $\binom{m+r-2}{r-1}\binom{m+n}{n-1}$ (where $r=0$, this latter quantity should be interpreted as zero). Subtraction of $\binom{m+r-2}{r-1}\binom{m+n}{n-1}$ from $\binom{m+r-1}{r}\binom{m+n-1}{n}$ gives the value for $K^m_{n|r}$ stated. □

I show next that in a flat space (M,g), the Killing tensors are generated by the Killing vectors. I shall exploit the isomorphism given by Theorem 2.1 and prove the result by lifting to T^*M. To facilitate the proof, it is convenient to introduce the following four vector fields on T^*M. Letting (x^i,p_i) be an adapted coordinate system on T^*M induced from a linear coordinate system (x^i) on M let Γ be the vector field on T^*M induced by the position vector field $x^i(\partial/\partial x^i)$ on M, in virtue of T^*M being (locally) isomorphic with $M\oplus M^*$. Let $\Delta = p_i(\partial/\partial p_i)$ be the Liouville vector field, $X = p_i(\partial/\partial x^i)$ the dynamical field (the Hamiltonian flow determined by any flat metric on M) and $Y = x^i(\partial/\partial p_i)$. One may readily check that Γ,Δ,X, and Y satisfy the relations of a four-dimensional algebra, but the only Lie bracket relations I shall need are

$$[\Gamma,\Delta] = 0, \tag{4.12}$$

$$[X,Y] = \Delta - \Gamma. \tag{4.13}$$

Notice also that a function $f: T^*M \to \mathbf{R}$ is homogeneous in the position or momentum variables of degree n iff $\Gamma f = nf$ or $\Delta f = nf$, respectively.

I have shown quite generally in Sec. II that when one looks for first integrals of a geodesic flow that are polynomial in the momenta, it suffices to look for homogeneous polynomials. It also follows from Proposition 4.2 that, in a flat space, it is even sufficient to consider first integrals which are homogeneous in the position variables.

Theorem 4.5: In a flat Riemannian or pseudo-Riemannian manifold (M,g) the Killing tensors are generated by the Killing vectors, that is to say, any Killing tensor consists of sums of symmetrized Killing vectors.

Proof: Again I shall make use of the isomorphism given by Theorem 2.1 and show that a first integral f of the geodesic flow X, which is a homogeneous polynomial of degree n in the momenta, consists of sums of products of first integrals of degree one. I may also assume that f is homogeneous in the position variables of degree r, where $0 \leqslant r \leqslant n$. Thus f satisfies

$$\Gamma f = rf, \tag{4.14}$$

$$\Delta f = nf, \tag{4.15}$$

$$Xf = 0. \tag{4.16}$$

The idea of the proof is to apply the second-order differential operator

$$(x^ip_j - x^jp_i)\left(\frac{\partial^2}{\partial x^i\partial p_j} - \frac{\partial^2}{\partial x^i\partial p_i}\right)$$

to f. This has the effect of decomposing f into a sum of terms, each of which has as a factor the angular momentum-type integral $x^ip_j - x^jp_i$, which corresponds to the Killing vector R_{ij} [see (2.5)]. The result will then follow by induction. Thus

$$\begin{aligned}
&(x^ip_j - x^jp_i)\left(\frac{\partial^2 f}{\partial x^i\partial p_j} - \frac{\partial^2 f}{\partial x^j\partial p_i}\right)\\
&\quad = 2\Gamma(\Delta f) - 2x^ip_j\frac{\partial^2 f}{\partial x^i\partial p_j}\\
&\quad = 2\Gamma(\Delta f) - (X(Yf) + Y(Xf) - \Gamma f - \Delta f)\\
&\quad = 2\Gamma(\Delta f) - ([X,Y]f + 2Y(Xf) - \Gamma f - \Delta f)\\
&\quad = 2\Gamma(\Delta f) - (\Delta f - \Gamma f + 2Y(Xf) - \Gamma f - \Delta f)\\
&\quad = 2\Gamma(\Delta f) - 2Y(Xf) + 2\Gamma f\\
&\quad = 2r(n+1)f.
\end{aligned} \tag{4.17}$$

The penultimate equality here is an identity and the last equality follows from (4.14)–(4.16).

Now it is straightforward to check that each of the quantities

$$\frac{\partial^2 f}{\partial x^i\partial p_j} - \frac{\partial^2 f}{\partial x^j\partial p_i}$$

is a first integral whenever f is a first integral of X. Moreover, if f is homogeneous of degree n and r in the p_i's and x^i's, respectively, each

$$\frac{\partial^2 f}{\partial x^i\partial p_j} - \frac{\partial^2 f}{\partial x^j\partial p_i}$$

will be homogeneous of degree $n-1$ and $r-1$, respectively, provided i is different from j. The result now follows by induction on n: it is evidently true when $n=1$ and if true for $n-1$, then (4.17) demonstrates that it is true for n provided that $r \neq 0$; however, if $r=0$, then f is a polynomial in the p_i's alone and thus corresponds to a Killing tensor that consists of a sum of symmetrized products of the Killing vector T_i. Thus the result is also valid when $r=0$. □

It is important to observe that although Theorem 4.5 shows that in flat space the Killing tensors are generated by the $\frac{1}{2}m(m+1)$ Killing vectors T_i and R_{jk}, it is not necessarily true that a basis for, say, the valence n Killing tensors consists of all n-fold products

$$T_{i_1\cdots i_{n-r}}R_{j_{n-r+1}k_{n-r+1}}\cdots R_{j_nk_n} \qquad (0 \leqslant r \leqslant n). \tag{4.18}$$

For example, if $m \geqslant 3$ and $i<j<k$,

$$T_iR_{jk} + T_kR_{ij} + T_jR_{ki} = 0 \tag{4.19}$$

is a nontrivial relation involving the T_i's and R_{jk}'s. Despite this, one may choose as a basis for the Killing tensors of valence n whose components are homogeneous polynomials of degree r, a certain number of products of the form given by (4.18) (in fact,

$$(n-r+1)\,\frac{(m+r-2)!(m+n-1)!}{r!(n+1)!(m-1)!(m-2)!}$$

such products according to Corollary 4.4).

It is a well-known classical result that an m-dimensional Riemannian or pseudo-Riemannian manifold possesses a full complement, that is to say, $\frac{1}{2}m(m+1)$ linearly indepen-

dent Killing vectors iff it is a space of constant curvature.[10,17] It is also a well-known result, originally due to Riemann, that such a space of constant curvature K, say, has a local coordinate system (x^i) such that the metric g_{ij} assumes the form

$$g_{ij}\,dx^i\,dx^j = \frac{e_1(dx^1)^2 + e_2(dx^2) + \cdots + e_m(dx_m)^2}{[1 + (K/4)(e_1(x^1)^2 + e_2(x^2)^2 + \cdots + e_m(x_m)^2)]^2}, \tag{4.20}$$

where each e_i is plus or minus one, and the total number of negative e_i's determine the signature of g.[10–12,17] I shall now give the explicit local form of the Killing vectors in a space of constant curvature, relative to which the metric takes the form given by (4.20); these do not appear to have been stated in the classical references.[10,17]

Proposition 4.6: In a Riemannian or pseudo-Riemannian manifold (M,g) of constant, nonzero curvature K, there are $\frac{1}{2}m(m+1)$ linearly independent Killing vectors. A basis for the space of Killing vectors is given by

$$\overline{R}_{ij} = e^i x^j \frac{\partial}{\partial x^i} - e^j x^i \frac{\partial}{\partial x^j} \quad (1 \leqslant i < j \leqslant m), \tag{4.21}$$

$$\overline{T}_i = \left(2e_i(x_i)^2 - \sum_{j=1}^{m} e_j(x^j)^2 + \frac{K}{4}\right)\frac{\partial}{\partial x^i} + 2e_i \sum_{\substack{j=1 \\ j\neq i}}^{m} x^i x^j \frac{\partial}{\partial x^j} \quad (1 \leqslant i \leqslant m), \tag{4.22}$$

where (x^i) is a coordinate system in which the metric assumes the form given by (4.20) and it is to be understood in (4.21) and (4.22) that the summation convention does not apply.

Proof: Once more it is convenient to identify a Killing vector with a real-valued function on T^*M. As such, if the Killing vector is written locally as $K^i(\partial/\partial x^i)$, one must demand that the K^i's satisfy the following condition:

$$\left\{\left(1 + \frac{K}{4}\sum_{i=1}^{m} e_i(x^i)^2\right)^2 \sum_{j=1}^{m} e_j(p_j)^2, \sum_{k=1}^{m} K^k p_k\right\} = 0, \tag{4.23}$$

where again the summation convention is suspended.

When (4.23) is expanded, one finds that the K^i's must satisfy

$$\frac{\partial K^i}{\partial x^i} - \frac{\partial K^j}{\partial x^j} = 0 \quad (1 \leqslant i, j \leqslant n), \tag{4.24}$$

$$e_i \frac{\partial K^j}{\partial x^i} + e_j \frac{\partial K^i}{\partial x^j} = 0 \quad (1 \leqslant i < j \leqslant n), \tag{4.25}$$

$$\sum_{k=1}^{m} e_k K x^k K^k - 2\left(1 + \frac{K}{4}\sum_{k=1}^{m} e_k(x^k)^2\right)\frac{\partial K^i}{\partial x^i} = 0 \quad (1 \leqslant i \leqslant n), \tag{4.26}$$

and again the summation convention does not apply in (4.24)–(4.26). The system of partial-differential equations consisting of (4.24)–(4.26) is linear in the sense that the solutions form a linear space. It is straightforward now to check that the components of the $\frac{1}{2}m(m+1)$ linearly independent Killing vectors in (4.21) and (4.22) constitute a basis for the solution space. □

The following and final theorem subsumes all the results of this section.

Theorem 4.7: Let (M,g) be a Riemannian or pseudo-Riemannian manifold of constant curvature K. Then the dimension of the space of (covariant) Killing tensors K_n^m is equal to

$$\frac{(m+n-1)!(m+n)!}{(m-1)!m!n!(n+1)!}.$$

Moreover, any Killing tensor on (M,g) consists of sums of symmetrized products of Killing vectors.

Proof: The idea of the proof is to show that, roughly speaking, there are as many linearly independent Killing tensors in a space of constant curvature, as there are in flat space; since a flat space admits the maximum number of linearly independent Killing tensors, then so too must a space of constant curvature.

Now let $(K^0_1,\ldots,K^0_{i_0}),(K^1_1,\ldots,K^1_{i_1}),\ldots,(K^n_1,\ldots,K^n_{i_n})$ be bases for the Killing tensors in flat space of valence n whose components are homogeneous polynomials of degree $0,1,\ldots,n$, respectively. The set of these $(n+1)$ bases collectively form a basis for the space of Killing tensors of valence n, in view of Propositions 4.1 and 4.2. Moreover, as was explained in the remarks following Theorem 4.5, each of the tensors K_i $(0 \leqslant r \leqslant n,\ 1 \leqslant i \leqslant i_r)$ consists of sums of symmetrized products of the Killing vectors T_k and R_{lm}. Thus I shall write $K^r_i = K^r_i(T_k,R_{lm})$ and remark that each summand in k^r_i contains $n-r$ T_k's and r R_{lm}'s.

The considerations of the previous paragraph concerned a linear space (or at any rate, an open submanifold of a flat space) and as such did not require the introduction of local coordinates. Suppose now, however, that coordinates (x^i) are introduced on a space of constant curvature relative to which the metric assumes the form given by (4.20). I shall then write $\overline{K}^j_i = \overline{K}^j_i(\overline{T}_k,\overline{R}_{lm})$ to signify that $\overline{K}^j_i$ is the same polynomial function of $\overline{T}_k$ and $\overline{R}_{lm}$ [See (4.21) and (4.22)] as K^j_i is of T_k and R_{lm}. Next, let $A^0_n,A^1_n,\ldots,A^n_n$, respectively, be linear combinations of $(\overline{K}^0_1,\ldots,\overline{K}^0_{i_0})$, $(\overline{K}^1_1,\ldots,\overline{K}^1_{i_1}),\ldots,(\overline{K}^n_1,\ldots,\overline{K}^n_{i_n})$, respectively, suppose that

$$\lambda_0 A^0_n + \lambda_1 A^1_n + \cdots + \lambda_n A^n_n = 0, \tag{4.27}$$

where each $\lambda_i \in \mathbb{R}$. Then it is apparent from (4.21) and (4.22) that each A^r_n contains no term of degree lower than r in the x^i's; but then (4.27) implies either $\lambda_i = 0$ or $A_i = 0$. Thus the subspaces spanned, respectively, by $(\overline{K}^0_1,\ldots,\overline{K}^0_{i_0})$, $(\overline{K}^1_1,\ldots,\overline{K}^1_{i_1}),\ldots,(\overline{K}^n_1,\ldots,\overline{K}^n_{i_n})$ intersect mutually in a trivial fashion. It follows that the dimension of the space of Killing tensors of valence n in a space of constant curvature is at least the sum of the dimensions of the subspaces spanned by $(\overline{K}^0_1,\ldots,\overline{K}^0_{i_0})$, $(\overline{K}^1_1,\ldots,\overline{K}^1_{i_1}),\ldots,(\overline{K}^n_1,\ldots,\overline{K}^n_{i_n})$, respectively.

Now consider the subspace spanned by $(\overline{K}^r_1,\ldots,\overline{K}^r_{i_r})$. It is clear from (4.21) and (4.22) that the lowest-order terms of this subspace of polynomials, that is, those of degree r, agree with those of the subspace spanned by $(K^r_1,\ldots,K^r_{i_r})$ considered as homogeneous polynomials in the x^i's of degree r. It follows from Corollary 4.4 that the dimension of the subspace spanned by $(\overline{K}^r_1,\ldots,\overline{K}^r_{i_r})$ is at least

$$(n-r+1)\frac{(m+r-2)!(m+m-1)}{r!(n+1)!(m-1)!(m-2)!}.$$

All in all, one may conclude that the dimension of the space of Killing tensors in a space of constant curvature is at least as great as that in the corresponding flat space. Since this latter dimension is maximal, these two dimensions are actually equal. It is clear from the comparison argument just given that in a space of constant curvature the Killing tensors are generated by the Killing vectors. □

[1]P. J. Sommers, J. Math. Phys. **14**, 787 (1973).
[2]I. Hauser and R. J. Malhiot, J. Math. Phys. **15**, 816 (1974).
[3]I. Hauser and R. J. Malhiot, J. Math. Phys. **16**, 1626 (1975).
[4]N. M. J. Woodhouse, Commun. Math. Phys. **44**, 9 (1975).
[5]G. Thompson, J. Math. Phys. **25**, 3474 (1984).
[6]G. Darboux, Arch. Neerl. Sci. Exactes Nat. **6**, 371 (1901).
[7]G. Thompson, J. Phys. A **17**, 955 (1984).
[8]G. Thompson, J. Phys. A **17**, L411 (1984).
[9]B. Dorizzi, B. Grammaticos, and A. Ramani, J. Math. Phys. **24**, 2282 (1983).
[10]L. P. Eisenhart, *Riemannian Geometry* (Princeton U. P., Princeton, NJ, 1949).
[11]M. Spivak, *A Comprehensive Introduction to Differential Geometry* (Publish or Perish, Berkeley, CA, 1979), Vol. 2.
[12]J. Wolf, *Spaces of Constant Curvature* (Publish or Perish, Boston, 1974).
[13]N. Jacobson, *Lie Algebras* (Interscience, New York, 1962).
[14]B. O'Neill, *Semi-Riemannian Geometry* (Academic, New York, 1983).
[15]V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).
[16]A. Weinstein, J. Differential Geometry, **18**, 523 (1983).
[17]K. Yano, *The Theory of the Lie Derivative and Its Applications* (North-Holland, Amsterdam, 1956).
[18]E. Kalnins and W. Miller, SIAM J. Math. Anal. **11**, 1011 (1980).

Some problems of spinor and algebraic spinor structures

Krystyna Bugajska
Institut für Theoretische Physik der Technischen Universität, 3392 Clausthal Zellerfeld, West Germany

(Received 6 January 1986; accepted for publication 28 May 1986)

A concrete realization of the Milnor–Lichnerowicz spinor bundle by algebraic spinors is considered in the case when the holonomy group of the Levi-Civita connection is equal to the Crumeyrolle group. Some relationships between the existence of parallel spinor fields on a space-time manifold $\mathscr{M}$ and its topological invariants are given.

I. INTRODUCTION

In the physical as well as in the mathematical literature the most popular approach to the problem of spinor structure is given by the Milnor–Lichnerowicz one. However, starting from the Sauter, Edington, and Sommerfeld investigations algebraic spinors are also studied intensely. Algebraic spinors were defined as elements of a minimal left ideal of an appropriate Clifford algebra.

At the beginning of this paper we show how we can introduce the structure of the left Clifford modules on the Milnor–Lichnerowicz (ML for short) bundle $\xi_s[\mathbb{C}^4]$. It allows us to consider, for example, the Clifford multiplication of vectors and ML spinors. But it does not mean that this bundle can be identified with a bundle of minimal left ideals; quite the opposite is the case. For example, for a Riemannian manifold S^4 we can introduce ML spinors together with their Clifford left module structure but we cannot construct any algebraic spinor structure (given by a global field of primitive idempotents) on it.

In this paper we are interested in a Lorentz space-time manifold. Fortunately for this case the topological obstructions for the ML spin structure and for an algebraic spin structure are exactly the same. However, we restrict our considerations to the case when the holonomy group of the Levi-Civita connection is equal to the Crumeyrolle group. In this case the parallel transport of algebraic spinors given by the Levi-Civita connection lifted to the Clifford bundle is exactly the same as given by the spin connection. Equivalently we can say that the global field $f(x)$ of primitive idempotents is parallel as well as a section of the Clifford bundle and also as a section of the section of the spinor bundle.

We will see that in this case we have a totally geodesic codimension-1 foliation of $\mathscr{M}$. However, because the normal bundle of this foliation is isomorphic to the isotropic tangent line bundle we meet many difficulties. To avoid them we construct a concrete Riemannian metric that seems to be quite natural in our case. Then we prove a lemma that allows us to know the holonomy group of this Riemannian metric as well as to see that this is a bundlelike metric compatible with our foliation. Now we can, using the result of Johnson and Naveira, relate the existence of parallel spinor fields on a Lorentz manifold $\mathscr{M}$ with the vanishing of the Pontryagin ring in the top dimension.

We will denote by $R^{s,t}$ and $(s+t=n)$-dimensional vector space equipped with a quadratic form of signature (s,t) and by $R_{s,t}$ its universal Clifford algebra.

II. DIRAC OPERATOR AND ITS SQUARE

Let $\mathscr{M}$ be a semi-Riemannian oriented four-manifold of signature (s,t). Let us suppose that the bundle ξ_0 of oriented, orthonormal frames over $\mathscr{M}$ lifts to give a principle $\mathrm{Spin}_+(s,t)$ bundle ξ_s. Now we can define via the spin representation a vector bundle

$$\mathscr{S} = \xi_s[\mathbb{C}^4], \tag{2.1}$$

called the bundle of Dirac spinors.

However, owing to the metric structure g on $\mathscr{M}$ we have a semidefinite quadratic form on the tangent bundle $T\mathscr{M}$. Thus we can associate to $\mathscr{M}$ a Clifford bundle $\mathscr{C}^{\mathbb{C}}(T^*\mathscr{M})$ in a natural way. Its fiber at a point $x\in\mathscr{M}$ is the complexified Clifford algebra of the tangent space $(T_x\mathscr{M}, g(x))$. So we can study bundles of modules over these bundles of algebras.

As a matter of fact the group $\mathrm{Spin}_+(s,t)$ can be defined as a subgroup of the group of units of the Clifford algebra $R^{\mathbb{C}}_{s,t}$ and a complex representation of the $\mathrm{Spin}_+(s,t)$ can be given by a minimal left ideal of $R^{\mathbb{C}}_{s,t}$. This is also a module for the Clifford algebra $R^{\mathbb{C}}_{s,t}$ and the action of $\mathrm{Spin}_+(s,t)$ is induced by the Clifford multiplication.

It is known[1] that any minimal left ideal Δ is given by a primitive idempotent $f^2=f$ of $R^{\mathbb{C}}_{s,t}$,

$$\Delta = R^{\mathbb{C}}_{s,t} f. \tag{2.2}$$

Any global field of primitive idempotents on a spin manifold $\mathscr{M}$ defines a bundle Ψ of left modules over the bundle of algebras $\mathscr{C}^{\mathbb{C}}(M,g)$. We will call Ψ the bundle of algebraic spinors.

However, we can give to $\mathscr{S}=\xi_s[\mathbb{C}^4]$ also the structure of a left module bundle.

Let $\{\epsilon_\alpha\}_{\alpha\in\Lambda}$ be local trivializations of ξ_0 related to some contractible covering $\{\mathscr{U}_\alpha\}_{\alpha\in\Lambda}$ of $\mathscr{M}$ such that

$$\epsilon_\alpha(x) = \epsilon_\beta(x) g_{\beta\alpha}(x), \quad \forall x\in\mathscr{U}_\alpha\cap\mathscr{U}_\beta, \tag{2.3}$$

and

$$g_{\alpha\beta}\colon\ \mathscr{U}_\alpha\cap\mathscr{U}_\beta\to \mathrm{SO}_+(s,t).$$

Now we can use the canonical maps[2]

$$\theta_\alpha\colon\ \epsilon_\alpha(x)\to\theta(e)\hookrightarrow R^{\mathbb{C}}_{s,t},$$
$$\theta_\beta\colon\ \epsilon_\beta(x)\to\theta(e')\hookrightarrow R^{\mathbb{C}}_{s,t}. \tag{2.4}$$

Here $e=\{e_i\}$ as well as $e'=\{e'_i\}$ is an orthonormal basis in $R^{s,t}$ related by the transformation $g_{\beta\alpha}(x)$ given by (2.3),

$$e = e' g_{\beta\alpha}(x), \tag{2.5}$$

 0022-2488/86/112700-06$02.50

and θ: $R^{s,t} \hookrightarrow R^{\mathbb{C}}_{s,t}$ is the canonical embedding. Now using (2.4) we can write

$$e' = \gamma_{\beta\alpha} e'_i \gamma_{\beta\alpha}^{-1}, \quad i = 1,...,4, \quad \gamma_{\beta\alpha} \in R^{\mathbb{C}}_{s,t}. \tag{2.6}$$

Let $\{\mathfrak{z}_\alpha\}_{\alpha\in\Lambda}$ be local trivializations of ξ_s related with $\{\mathcal{U}_\alpha, \epsilon_\alpha\}_{\alpha\in\Lambda}$. From (2.6) we obtain that ξ_s can be seen as a principal bundle whose structural group is given by elements of $\mathrm{Spin}_+ \subset R^{\mathbb{C}}_{s,t}$ [we will denote the group $\mathrm{Spin}_+(s,t)$ by Spin_+ for short]

$$\mathfrak{z}_\alpha(x) = \mathfrak{z}_\beta(x)\gamma_{\beta\alpha}(x), \quad \gamma_{\beta\alpha}(x) \in R^{\mathbb{C}}_{s,t}. \tag{2.7}$$

Now for any element $\omega\in\Delta$ we can construct the element

$$[\mathfrak{z}_\alpha(x),\omega] := (\mathfrak{z}_\alpha(x),\omega)\mathrm{Spin}_+ = (\mathfrak{z}_\beta(x)\gamma_{\beta\alpha}(x),\omega)\mathrm{Spin}_+ = [\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\omega]. \tag{2.8}$$

In other words we can construct the bundle $\xi_s[\Delta]$ of left modules over the bundle of algebras $\mathscr{C}^{\mathbb{C}}(M,g)$. It is nothing more than $\xi_s(\mathbb{C}^4)$ with the identification of $\mathbb{C}^4$ with Δ. We will see later [(2.17)] that in a general case $\xi_s[\Delta]$ cannot be identified with a bundle of minimal left ideals of the Clifford bundle. Nevertheless the isomorphism between Δ and $\mathbb{C}^4$ together with (2.4)–(2.8) allows us to consider the Clifford multiplication of, for example, vectors and spinors.

Let us take any vector $\nu\in R^{s,t}$. We define

$$[\mathfrak{z}_\alpha(x),\nu] := (\mathfrak{z}_\alpha(x),\nu)\mathrm{Spin}_+ = (\mathfrak{z}_\beta(x)\gamma_{\beta\alpha}(x),\nu)\mathrm{Spin}_+ = [\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\nu\gamma_{\beta\alpha}^{-1}(x)]. \tag{2.9}$$

It means that when we understand the structure group of ξ_s as given by some elements of the Clifford algebra $R^{\mathbb{C}}_{s,t}$ according to (2.6) and (2.7), then

$$\xi_s[R^4] = T\mathcal{M}. \tag{2.10}$$

Moreover from (2.4), (2.6), and (2.9) we obtain immediately that

$$\xi_s[R^{\mathbb{C}}_{s,t}] = \mathscr{C}^{\mathbb{C}}(\mathcal{M},g) \tag{2.11}$$

and

$$[\mathfrak{z}_\alpha(x),\omega] := (\mathfrak{z}_\alpha(x),\omega)\mathrm{Spin}_+ = (\mathfrak{z}_\beta(x)\gamma_{\beta\alpha}(x),\omega)\mathrm{Spin}_+ = [\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\omega\gamma_{\beta\alpha}^{-1}(x)], \quad \forall\omega\in R^{\mathbb{C}}_{s,t}. \tag{2.12}$$

Now let us take some section $\tilde{\psi}(x)$ of the bundle

$$\tilde{\Psi} := \xi_s[\Delta]$$

and some section $\omega(x)$ of the bundle $\mathscr{C}^{\mathbb{C}}(\mathcal{M},g)$. According to the above formulas we have

$$\tilde{\psi}(x) = [\mathfrak{z}_\alpha(x),\omega] = [\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}\omega], \quad \omega\in\Delta, \tag{2.13}$$

and

$$\omega(x) = [\mathfrak{z}_\alpha(x),\omega] = [\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\omega\gamma_{\beta\alpha}^{-1}(x)], \quad \omega\in R^{\mathbb{C}}_{s,t}, \tag{2.14}$$

for every $x\in\mathcal{U}_\alpha\cap\mathcal{U}_\beta$. We can define their Clifford multiplication in the following way:

$$\omega(x)\tilde{\psi}(x) = [\mathfrak{z}_\alpha(x),\omega][\mathfrak{z}_\alpha(x),\omega\omega] = [\mathfrak{z}_\alpha(x),\omega\omega] = [\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\omega\omega] \quad \text{by (2.8)}. \tag{2.15}$$

It is easy to see that by multiplication of $[\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\omega\gamma_{\beta\alpha}^{-1}(x)]$ by $[\mathfrak{z}_\beta(x),\gamma_{\beta\alpha}(x)\omega]$ we obtain the same result. Thus although in a general case $\xi_s[\Delta]$ cannot be given as a bundle of minimal left ideals of the Clifford bundle (2.11) we can always consider $\tilde{\Psi}$ as a bundle of Clifford modules.

But why can we not always consider $\xi_s[\Delta]$ as a subbundle of $\mathscr{C}^{\mathbb{C}}(M,g)$?

According to (2.2) Δ is a minimal left ideal of $R^{\mathbb{C}}_{s,t}$ defined by some primitive idempotent $f\in R^{\mathbb{C}}_{s,t}$. Now for every $x\in\mathcal{U}_\alpha$ elements $[\mathfrak{z}_\alpha(x),f]$ define a local field, say $f_\alpha(x)$ of primitive idempotents of

$$\mathscr{C}^{\mathbb{C}}(\mathcal{M},g)|_{\mathcal{U}_\alpha}.$$

Similarly $[\mathfrak{z}_\beta(x),f]$ defines a field $f_\beta(x)$. However, in a general case

$$f_\alpha(x) = [\mathfrak{z}_\alpha(x),f] \neq [\mathfrak{z}_\beta(x),f] = f_\beta(x), \quad x\in\mathcal{U}_\alpha\cap\mathcal{U}_\beta. \tag{2.16}$$

If the structural group of ξ_s can be reduced to an appropriate subgroup we can still have the possibility that

$$\mathscr{C}^{\mathbb{C}}(\mathcal{M},g)|_{\mathcal{U}_\alpha\cap\mathcal{U}_\beta} f_\alpha(x) = \mathscr{C}^{\mathbb{C}}(\mathcal{M},g)|_{\mathcal{U}_\alpha\cap\mathcal{U}_\beta} f_\beta(x). \tag{2.17}$$

But in a general case this condition is not satisfied (for example, for some Riemannian structures on S^4).

From now on we will define the bundle Ψ of algebraic spinors on $\mathcal{M}$ as a concrete realization of the ML bundle $\xi_s[\mathbb{C}^4]$ given by some concrete global field f of primitive idempotents of $\mathscr{C}^{\mathbb{C}}(\mathcal{M},g)$.

It is known that the necessary and sufficient condition for the existence of a global field $f(x)$ is the existence of the reduction of the bundle ξ_s to the subbundle $\xi_{\mathscr{C}}$. The structural group $\mathscr{C}$ is given by the following conditions:

$$\mathscr{C} := \{\gamma\in\mathrm{Spin}_+ \subset R^{\mathbb{C}}_{s,t};\ \gamma f\gamma^{-1} = \gamma f = f\}. \tag{2.18}$$

This group was investigated by Crumeyrolle[3] and for this reason we will call it the Crumeyrolle group.

The existence of a global field f of primitive idempotents is not equivalent, in a general case, to the existence of a bundle ξ_s. For example, the Riemannian spin manifold S^4 does not admit a global field f for any metric structure on it. To see this, let

$$\Delta = \Delta^+ \oplus \Delta^- \tag{2.19}$$

be the decomposition of Δ into the direct sum of two two-complex-dimensional irreducible representations of Spin_+. This decomposition implies the decomposition of the module bundle $\xi_s[\Delta]$ into subbundles

$$\tilde{\Psi} \equiv \xi_s[\Delta] = \tilde{\Psi}^+ \oplus \tilde{\Psi}^-, \tag{2.20}$$

where $\tilde{\Psi}^\pm := \xi_s[\Delta^\pm]$. The decomposition (2.19) is interchanged by the Clifford multiplication with vectors, i.e., we have to deal with the following homomorphisms:

$$R^{s,t} \underset{R}{\otimes} \Delta^+ \to \Delta^-, \tag{2.21}$$

$$R^{s,t} \underset{R}{\otimes} \Delta^- \to \Delta^+. \tag{2.21'}$$

This implies the vector bundle homomorphisms

$$T\mathcal{M} \underset{R}{\otimes} \tilde{\Psi}^+ \to \tilde{\Psi}^-, \tag{2.22}$$

$$T\mathcal{M} \underset{R}{\otimes} \tilde{\Psi}^- \to \tilde{\Psi}^+, \tag{2.22'}$$

or

$$\mu\colon T\mathcal{M} \underset{R}{\otimes} \tilde{\Psi} \to \tilde{\Psi}. \tag{2.23}$$

In other words any element of Δ^+ determines a real isomorhpism of $R^{s,t}$ with Δ^- given by the Clifford multiplication. Thus any global field of primitive idempotents $f \in \Gamma(\mathcal{C}^{\mathbb{C}}(\mathcal{M},g))$ (or isotropic bivectors of $\Lambda T^{\mathbb{C}}\mathcal{M}$) determines a real isomorphism

$$T\mathcal{M} \cong \tilde{\Psi}^-, \tag{2.24}$$

i.e., an almost complex structure on $\mathcal{M}$.

But it is known that S^4 does not admit any almost complex structure, so we cannot construct any global field f on S^4.

Fortunately for the most interesting physical case, namely for a Lorentzian space-time manifold $\mathcal{M}$, the following two conditions are equivalent[4]: (i) the existence of a lifting ξ_s of ξ_0, and (ii) the existence of a global field of primitive idempotents f. Nevertheless, the possibility of giving to $\xi_s[\Delta] = :\tilde{\Psi}$ the structure of a module bundle over the algebras $\mathcal{C}^{\mathbb{C}}(M,g)$ allows us to introduce a first-order differential operator $\mathcal{D}\colon \Gamma(\tilde{\Psi}) \to \Gamma(\tilde{\Psi})$ for any signature of a metric structure of $\mathcal{M}$. We define this operator $\mathcal{D}$, called the Dirac operator, by the following composition:

$$\Gamma(\tilde{\Psi}) \xrightarrow{\nabla^s} \Gamma(T^*M \otimes \tilde{\Psi}) \xrightarrow{*} \Gamma(T\mathcal{M} \otimes \tilde{\Psi}) \xrightarrow{\mu} \Gamma(\tilde{\Psi}), \tag{2.25}$$

where ∇^s is the covariant derivative (relative to the connection on $\tilde{\Psi}$ induced by a metric g), $*$ means the identification of $T^*\mathcal{M}$ with $T\mathcal{M}$ given by a metric g, and μ denotes the Clifford multiplication (2.23).

Locally in the basis $\{\mathcal{U}_\alpha, \epsilon_\alpha, \mathcal{F}_\alpha\}_{\alpha\in\Lambda}$,

$$\mathcal{D} = \sum_{i=1}^{4} \tau(e_i) e_i \nabla^s_{e_i}, \tag{2.26}$$

where $\tau(i) = \pm 1$ according to $g(e_i,e_j) = \tau(e_i)\delta_{ij}$. Let us recall that

$$\nabla e_j = \sum_{i=1}^{4} \omega_{ji}\tau(i) e_i, \tag{2.27}$$

with

$$\omega_{ij} = g(\nabla e_i, e_j)\tau(e_i)\tau(e_j)$$

and

$$\nabla^s \tilde{\Psi} = d\tilde{\psi} + \frac{1}{2}\sum_{i<j} \tau(i)\tau(j)\omega_{ij}\, e_i \cdot e_j\, \tilde{\psi}. \tag{2.28}$$

Now we can write locally

$$\mathcal{D}\tilde{\psi} = \sum_{i=1}^{4} \Big[\tau(i) e_i \cdot e_i\, \tilde{\psi} + \frac{1}{2}\sum_{j<k} \tau(i)\tau(j)\tau(k)\omega_{jk} e_i \cdot e_j \cdot e_k\, \tilde{\psi}\Big]. \tag{2.29}$$

The square $\mathcal{D}^2$ locally has the form

$$\mathcal{D}^2 = \frac{s}{4} - \sum_j \tau(j)\nabla^s_{e_i}\nabla^s_{e_i} - \sum_{j,k} \tau(j)\tau(k) g(\nabla_{e_k} e_j, e_k)\nabla^s_{e_j}, \tag{2.30}$$

where s is the scalar curvature

$$s = \sum_{ij} \tau(i)\tau(j) R_{ijji}.$$

Using the formula (2.30) Lichnerowicz has shown[5] that if the scalar curvature is non-negative but not identically zero on a compact spin Riemannian manifold, then there are no harmonic spinors, i.e.,

$$\operatorname{Ker} \mathcal{D} = 0.$$

This fact together with the Atiyah–Singer index theorem gives that the Hirzebruch genus $\hat{A}$ of such a manifold must be zero. In other words one cannot have a metric with non-negative scalar curvature (except identically zero) on a compact, spin Riemannian manifold whose $\hat{A}$ genus is not zero.

This does not mean that there is an analog of Hodge's theorem, i.e., that we can express the dimension of the space of harmonic spinors in terms of topological invariants of the manifold. On the contrary, Hitchin[6] has shown that the dimension of the null space of the Dirac operator depends on the metric used to define this operator. Besides, Eliasson[7] and Aubin[8] have shown that every compact manifold of dimension $\geqslant 3$ possesses a Riemannian metric whose total curvature is negative. Further, using Trudinger's result[9] it can be seen that every compact manifold of dimension $\geqslant 3$ admits a Riemannian metric with constant negative scalar curvature. Moreover, it is known[10] that there are no topological obstructions to scalar curvatures that may change sign as long as they are negative somewhere.

It appears that such a relation between a positive scalar curvature and the absence of harmonic spinors is valid only in the Riemannian case. We cannot obtain a similar result for a semi-Riemannian space. Quite the opposite, there are known[11] examples of a compact semi-Riemannian manifold with positive scalar curvature for which the space of harmonic spinors is nonzero. We can meet such a situation for $SU(2) \cong S^3$ equipped with its natural metric of signature $(-,-,+)$ or for $T^3 = S^1 \times S^1 \times S^1$ equipped with a metric of signature $(+,+,-)$. Besides, until now no relation between the existence of a nontrivial harmonic spinor space $\operatorname{Ker} \mathcal{D}$ and some topological invariants of a Lorentzian manifold $\mathcal{M}$ was known. Later we will show that such a relationship does exist.

III. PENROSE FOLIATION

Let $\mathcal{M}$ be a Lorentzian space-time manifold[12] and $\mathcal{S} = \xi_s[\mathbb{C}^4]$ its Milnor–Lichnerowicz bundle of Dirac spinors. Let us consider the case when the holonomy group of Levi-Civita connection is equal to

$$\mathcal{C} = \rho(\tilde{\mathcal{C}}) \tag{3.1}$$

with the covering map $\rho\colon \mathrm{Spin}_+ \to \mathcal{L}_0$. Then we can realize $\mathcal{S}$ by the bundle Ψ of minimal left ideals of $\mathcal{C}^{\mathbb{C}}(\mathcal{M},g)$ determined by a parallel field f of primitive idempotents. By (2.18), and by a vector bundle isomorphism between $\mathcal{S}$ and

Ψ, we also have a nonvanishing field of parallel spinors $\tilde{\psi}(x)\in\Gamma(\mathscr{S})$.

A spinor field $\tilde{\psi}(x)$ is parallel iff

$$\nabla^s\tilde{\psi}\equiv 0, \tag{3.2}$$

i.e., by (2.28) if

$$d\tilde{\psi} = -\frac{1}{2}\sum_{i<j}\tau(i)\tau(j)\omega_{ij}e_i\cdot e_j\,\tilde{\psi}. \tag{3.3}$$

Now from (2.30) we obtain immediately that if a manifold $(\mathscr{M},g)$ admits a parallel spinor field then its scalar curvature has to vanish for any possible signature of g. Moreover, for a Riemannian manifold $(\mathscr{M},g)$ it also has to be Ricci flat.[13] For a Lorentzian manifold this condition is replaced[11] by the following: (i) the isotropy of the Ricci tensor, and (ii) the existence of a parallel nonvanishing light vector field. The example of the Schwarzschild metric shows that the condition of the isotropy of the Ricci tensor is not sufficient.

Let us consider an orthonormal frame ϵ_x at a point $x\in\mathscr{M}$. Let $\xi_{\mathscr{C}}$ be the holonomy bundle through ϵ_x and let $\epsilon(x)$ be some global section of $\xi_{\mathscr{C}}$,

$$\begin{aligned}&\epsilon(x) = (e_1(x),e_2(x),e_3(x),e_4(x)),\\ &e_1^2 = e_2^2 = e_3^2 = -e_4^2 = 1.\end{aligned} \tag{3.4}$$

Now we can see that the structure group of $\xi_{\mathscr{C}}$ can be generated by

$$A_1 = e_{14} - e_{31},\quad A_2 = e_{24} + e_{23}. \tag{3.5}$$

Now we can ask the following question: given a holonomy bundle $\xi_{\mathscr{C}}$ through ϵ_x, can we find a parallel field of primitive idempotents.

It is known[14] that any primitive idempotent of $R^{\mathbb{C}}_{3,1}$ has a form

$$f = \tfrac{1}{4}(1+\omega_1)(1+\omega_2),$$

where $\omega_i^2 = 1$, $i = 1,2$, and $\omega_1\omega_2 = \omega_2\omega_1$. It is also known that for any minimal left ideal there are many primitive idempotents that define it. We will be interested only in such primitive idempotents that can be written as

$$f(x) = \tfrac{1}{4}(1+e_{I_1})(1+e_{I_2}) \tag{3.6}$$

in the basis $\epsilon(x)$. Here the I_i are multi-indices, $i = 1,2$.

It is obvious that if our field $f(x)$ has a form (3.6) in the frame $\epsilon(x)$ then it has exactly the same form with respect to any orthonormal basis at the point $x\in\mathscr{M}$ that belongs to $\xi_{\mathscr{C}}$.

A general element f of the form (3.6) can be written in the following ways:

$$\begin{aligned}&\text{I.}\quad e_{I_1} = e_i,\quad e_{I_2} = e_{j4},\\ &\text{II.}\quad e_{I_1} = e_{i4},\quad e_{I_2} = e_{ijk},\quad i\neq j\neq k\neq 4.\\ &\text{III.}\quad e_{I_1} = e_{ik4},\quad e_{I_2} = e_{ij4},\end{aligned} \tag{3.7}$$

However, only

$$f^1(x) = \tfrac{1}{4}(1+e_1(x))(1-e_{34}(x)) \tag{3.8}$$

and

$$f^2(x) = \tfrac{1}{4}(1+e_2(x))(1-e_{34}(x)) \tag{3.9}$$

have their isotropy group generated by A_1 and A_2 of (3.5). In an appropriate basis we have

$$A_1 = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad A_2 = \begin{pmatrix}0 & i\\ 0 & 0\end{pmatrix}, \tag{3.10}$$

i.e., $\tilde{\mathscr{C}}$ contains the elements of the form

$$s = \begin{pmatrix}1 & z\\ 0 & 1\end{pmatrix},\quad z\in\mathbb{C}. \tag{3.11}$$

Now we can easily see that $\mathscr{C} = \rho(\tilde{\mathscr{C}})$ transforms the elements $\epsilon(x)$ as follows:

$$\begin{aligned}&e_1 \overset{\mathscr{C}}{\to} e_1 + 2\,\mathrm{Re}\,z(e_3+e_4),\\ &e_2 \overset{\mathscr{C}}{\to} e_2 + 2\,\mathrm{Im}\,z(e_3+e_4),\\ &e_3 \overset{\mathscr{C}}{\to} \tfrac{1}{2}(-\mathrm{Re}\,ze_1 - \mathrm{Im}\,ze_2\\ &\qquad + (1-z\bar{z})(e_3+e_4) + (e_3-e_4)),\\ &e_4 \overset{\mathscr{C}}{\to} \tfrac{1}{2}(\mathrm{Re}\,ze_1 - \mathrm{Im}\,ze_2 + (1+z\bar{z})(e_3+e_4) - (e_3-e_4),\end{aligned} \tag{3.12}$$

and

$$\begin{aligned}&e_3+e_4 \overset{\mathscr{C}}{\to} e_3+e_4,\\ &e_3-e_4 \overset{\mathscr{C}}{\to} -\mathrm{Re}\,ze_1 - \mathrm{Im}\,ze_2 - z\bar{z}(e_3+e_4) + (e_3-e_4).\end{aligned} \tag{3.13}$$

This fact explains the possibility of the construction of two inequivalent parallel primitive idempotents (3.8) and (3.9) for a given holonomy bundle $\xi_{\mathscr{C}}$. The interpretation of this fact in the Penrose picture is given in Ref. 15.

The transformation laws (3.12) imply that the codimension-1 distribution P spanned by

$$P = \{X(x),e_1(x),e_2(x)\},$$

with

$$X(x) = (1/\sqrt{2})(e_3+e_4), \tag{3.14}$$

is parallel. Further, from the fact that our connection is torsionless, i.e., for every vector field $A,B\in\Gamma(T\mathscr{M})$, we have

$$[A,B] = \nabla_A B - \nabla_B A, \tag{3.15}$$

one obtains that the distribution P is involutive. This means that we have a transversally oriented codimension-1 foliation $\mathscr{F}$ of our space-time manifold $(\mathscr{M},g)$.

However, it is easy to see that a metric g induces on each leaf of $\mathscr{F}$ a degenerate metric of signature $(0^1,+^2,-^0)$, i.e., determines a Galilean structure on any leaf of $\mathscr{F}$.

Moreover, any leaf is autoparallel by (3.15). Using again the torsionless property of the connection we see that this is equivalent to the fact that any leaf of $\mathscr{F}$ is a totally geodesic submanifold of $(\mathscr{M},g)$.[16]

Let us notice that the situation is considerably different from that of a Riemannian manifold. First, for a Riemannian metric to any parallel distribution there always exists a complementary distribution that is also parallel.[17] For a semi-Riemannian manifold this need not be the case. For example, a linear bundle complementary to P can be spanned by the field $Y(x) = (1/\sqrt{2})(e_3(x) - e_4(x))$, which is not parallel. Thus it is not possible even to try to give a result analo-

gous to the de Rham decomposition theorem.

Second, in the Riemannian case the integrability of the complementary distribution implies that the notion of totally geodesic foliation is interchangeable with the notion of a bundlelike metric.[18] In our case a vector bundle complementary to P has to be integrable (because any one-dimensional distribution spanned by a nonvanishing vector field is involutive[19]). However, by the isotropy of the field the notion of a bundlelike metric foliation cannot even be introduced. [By definition $(\mathscr{M},g)$ has a bundlelike metric compatible with a foliation $\mathscr{F}$ if the local submersions defining $\mathscr{F}$ may be chosen to be metric submersions.] Besides we have also to examine the Gauss and Weingarten formulas as well as the Gauss and Codazzi equations.

These remarks suggest that we could try to introduce some Riemannian metric on $\mathscr{M}$ and investigate our codimension-1 foliation $\mathscr{F}$ by means of it.

IV. BUNDLELIKE METRIC

We have already seen that if the bundle $\xi_{\mathscr{C}}$ can be taken as the holonomy bundle of a space-time manifold (M,g) then we have a codimension-1 foliation of $\mathscr{M}$ determined by the involutive Penrose distribution P.

The following general problem is still open: given a foliation $\mathscr{F}$ on a manifold $\mathscr{M}$, is there a bundlelike Riemannian metric on $\mathscr{M}$ compatible with $\mathscr{F}$? However, Johnson and Naveira[20] have made significant progress with this question. Namely, they have shown that if a codimension-1 foliation admits such a metric then the Pontryagin ring of the manifold has to vanish in the top dimension. In this section we will construct such a Riemannian metric $\hat{g}$. In this way we obtain a relationship between the existence of parallel spinors on $(\mathscr{M},g)$ and topological invariants of $\mathscr{M}$.

Now let us consider the possible Riemannian metrics on $\mathscr{M}$ that can be constructed from our fixed Lorentzian structure on $\mathscr{M}$. It is known that for any concrete Lorentzian metric g any nonvanishing timelike vector field, say V, defines a Riemannian metric $\hat{g}$ on $\mathscr{M}$ given by the following formula:

$$\hat{g}(A,B) = g(A,B) - 2g(A,V)g(B,V)/g(V,V), \quad \forall A,B\in\Gamma(T\mathscr{M}). \tag{4.1}$$

(In other words any section of ξ_0 mod SO(3) determines a concrete Riemannian structure.) Let us take the following vector field $V(x)\in\Gamma(T\mathscr{M})$:

$$V(x) = (1/\sqrt{2})(X(x) - Y(x)) = e_4(x). \tag{4.2}$$

Now

$$\hat{g}(A,B) = g(A,B) + 2g(A,V)g(B,V) \tag{4.3}$$

and we see immediately that the field of orthonormal (with respect to g) frames $\epsilon(x)$ is also orthonormal with respect to $\hat{g}$. Thus the SO(4,R) principal bundle ξ_R, which is equivalent to $\hat{g}$, is given by

$$\xi_R = \hat{\epsilon}(x)\mathrm{SO}(4,R). \tag{4.4}$$

Here $\hat{\epsilon}(x)$ denotes the same set of vectors as $\epsilon(x)$ but with Euclidean metric properties. For this reason we shall also distinguish the components $\{\hat{e}_i(x)\}$ of $\hat{\epsilon}(x)$ from the $\{e_i\}$ of $\epsilon(x)$ although they form the same elements of $T_x\mathscr{M}$. We see that

$$\xi_V := \epsilon(x)\mathrm{SO}(3,R) = \hat{\epsilon}(x)\cdot\mathrm{SO}(3,R) = \xi_0\cap\xi_R \tag{4.5}$$

is a subbundle of the two bundles ξ_0 and ξ_R and corresponds to the isotropy bundle of our vector field $V(x)$.

Although we can show which linear frames of $\mathscr{M}$ form the orthonormal bundle ξ_R we still have no information about the holonomy group related with $\hat{g}$.

Lemma: Let $\xi_{\hat{\mathscr{C}}}$ be the holonomy bundle of the Riemannian structure $(M,\hat{g})$ through $\hat{\epsilon}_x$. Then its structural group $\hat{\mathscr{C}}$ can be generated by

$$\hat{A}_1 = \tfrac{1}{2}(\hat{e}_{14} - \hat{e}_{31}), \quad \hat{A}_2 = \tfrac{1}{2}(\hat{e}_{24} + \hat{e}_{23}), \quad \hat{A}_3 = \hat{e}_{12}. \tag{4.6}$$

Proof: The elements $\hat{e}_{ij} = \hat{e}_i\cdot\hat{e}_j$ satisfy the commutation relations determined by the Clifford multiplication of the algebra $R_{4,0}\cong\mathscr{C}(T_xM,\hat{g})$. Hence we obtain

$$[\hat{A}_1,\hat{A}_2] = -\hat{A}_3, \quad [\hat{A}_2,\hat{A}_3] = -\hat{A}_1, \quad [\hat{A}_3,\hat{A}_1] = -\hat{A}_2.$$

Let us denote the Lie group generated by the elements (4.6) by $\hat{\mathscr{C}}$. Let us construct the principal bundle

$$\xi_{\hat{\mathscr{C}}} := \hat{\epsilon}(x)\cdot\hat{\mathscr{C}}\subset\xi_R. \tag{4.7}$$

Any connection on the bundles $\xi_0,\xi_{\mathscr{C}},\xi_R$, or $\xi_{\hat{\mathscr{C}}}$ determines a connection in the bundle of linear frames $\xi_{\mathrm{GL}(4,R)}$, i.e., a linear connection of M (see Ref. 17). On the other side it is known that a linear connection is a metric one only if it is induced from a connection in the principal bundle of orthonormal frames with respect to this metric. Moreover every semi-Riemannian or Riemannian manifold admits a unique metric connection with vanishing torsion.

In our case all bundles $\xi_0,\xi_{\mathscr{C}},\xi_R$, and $\xi_{\hat{\mathscr{C}}}$ are equivalent to the trivial bundles, and the section $\epsilon(x)$ of $\xi_{\mathrm{GL}(4,R)}$ determines these global trivializations. For this reason it is enough to define the horizontal subspaces of appropriate connections only at points given by $\epsilon(x)$. Let us consider a bundle $\xi_{\mathscr{C}}$. Let

$$(x,e)\equiv\epsilon(x), \quad \forall x\in\mathscr{M},$$

where e is the unit of $\mathscr{C}$. Let the horizontal space at (x,e) be spanned by

$$H_{(x,e)} = \{e_i(x) + X_i(x)\}, \tag{4.8}$$

where $X_i\in$Lie algebra of $\mathscr{C}$ given by (3.5). Now, for every point (x,a), $a\in\mathscr{C}$,

$$H_{(x,a)} = \{e_i(x) + \mathrm{ad}\,a^{-1}X_i(x)\}. \tag{4.8'}$$

Because $\xi_{\mathscr{C}}$ is the holonomy bundle of our Lorentzian structure g, the horizontal subspaces of its Levi-Civita connection are determined also by (4.8) and are given by (4.8'), but now with $a\in\mathscr{L}_0$. Of course it determines a connection on $\xi_{\mathrm{GL}(4,R)}$ whose horizontal subspaces at any point (x,a) are given again by formula (4.8') with $a\in\mathrm{GL}(4,R)$.

Now let us consider a horizontal distribution on $\xi_{\hat{\mathscr{C}}}$ defined by

$$H_{(x,\hat{e})} = \{e_i(x) + \hat{X}_i(x)\}, \quad \forall x\in\mathscr{M}. \tag{4.9}$$

Here $\hat{e}$ is the unit of $\hat{\mathscr{C}}$, and $\hat{X}_i$ has exactly the same form as X_i but with obviously different multiplication law. By the

formulas analogous to (4.8') we obtain horizontal distributions on ξ_R and $\xi_{GL(4,R)}$, which satisfy all propertiesof horizontal distributions of $\hat{g}$-metric and linear connections, respectively.

Instead of considering the horizontal subspaces we can introduce the dual notion of the one-form of the connections. Now, from the global trivializations of the considered principal bundles it is enough to determine the Lie-algebra-valued one-forms on $\mathscr{M}$.

We see immediately that the form

$$\alpha(e_i(x)) = -X_i(x) \in \text{Lie algebra of } \mathscr{C} \tag{4.10}$$

defines the one-form of the Levi-Civita connection of our Lorentzian structure g.

Similarly

$$\hat{\alpha}(e_i(x)) = -\hat{X}_i(x) \in \text{Lie algebra of } \hat{\mathscr{C}} \tag{4.11}$$

defines some $\hat{g}$-metric connection on ξ_R. Now we have to check if this connection is torsionless.

Let us consider the two-jet extension $\mathscr{J}^2\mathscr{M}$ of $\mathscr{M}$ (see Ref. 18) and the first-order differential prolongation $K\mathscr{M}$ (see Ref. 21) of the bundle $\xi_{GL(4,R)} = :L\mathscr{M}$ of linear frames. This last bundle can be identified with the one-jet extension of $\mathscr{M}$. We have the following diagram:

$$\begin{array}{ccc} \mathscr{J}^2\mathscr{M} & \overset{i}{\hookrightarrow} & K\mathscr{M} \\ {\scriptstyle \omega'}\uparrow\downarrow & & \downarrow\vdots\ {\scriptstyle \omega} \\ \mathscr{J}^1\mathscr{M} & = & L\mathscr{M} \end{array} \tag{4.12}$$

The fiber of $K\mathscr{M} \to L\mathscr{M}$ is given by the set of all jets of local sections of $L\mathscr{M}$, i.e., by the set

$$\text{Hom}(R^4, \text{gl}(4)) \cong \text{gl}(4) \otimes (R^4)^* \equiv R^4 \otimes (R^4)^* \otimes (R^4)^*. \tag{4.13}$$

[Here $(R^4)^*$ is the dual space to R^4.]

The fiber of $\mathscr{J}^2\mathscr{M} \to \mathscr{J}^1\mathscr{M}$ is given by the set of all equivalence classes of C^2 embeddings of $(R^4,0)$ into $(\mathscr{M},x)$. Two such embeddings φ and φ' are equivalent if the composition $\varphi^{-1} \circ \varphi'$ has the same derivatives up to order 2 at 0 as the identity map. It means that the fiber of $\mathscr{J}^2\mathscr{M} \to \mathscr{J}^1\mathscr{M}$ is given by $R^4 \otimes S^2(R^4)^*$, where S^2 denotes the symmetrized tensor power.

We can tell that $\mathscr{J}^2\mathscr{M}$ is isomorphic to the subbundle of $K\mathscr{M}$ given by the first-order jets of local holonomic sections of $L\mathscr{M}$.

Now any linear connection is given by a section ω of the bundle $K\mathscr{M} \to L\mathscr{M}$. This connection has no torsion iff its one-form ω factors through i (see Ref. 21) [see (4.12)], i.e.,

$$\omega = i \circ \omega'. \tag{4.14}$$

Now let ω and $\hat{\omega}$ denote a one-form of linear connections on $\xi_{GL(4,R)} = L\mathscr{M}$ determined by the horizontal distributions (4.8) and (4.9) [or equivalently by one-forms (4.10) and (4.11), respectively]. The difference between ω and $\hat{\omega}$ treated as maps from $L\mathscr{M}$ to $K\mathscr{M}$ can be determined by elements $p_x \in \text{Hom}(R^4, \text{gl}(4))$ such that

$$p_x(e_i(x)) = \hat{X}_i(x) - X_i(x) = j_x^*(X_i(x) - X_i(x)). \tag{4.15}$$

Here j is the local diffeomorphism

$$R^4 \supset \mathscr{U} \overset{j}{\to} \mathscr{U}' \subset R^4,$$

which induces a transformation of linear frames of R^4 [i.e., $GL(4,R)$] in such a way that

$$j_* A_i = \hat{A}_i, \quad i = 1,2.$$

In other words the horizontal distribution $\hat{H}_{(x,\hat{e})}$ can be obtained from $H_{(x,e)}$ by some map of local linear frames. Now because ω factors through i we see immediately that $\hat{\omega}$ has the same property. In this way we have obtained that the $\hat{g}$-metric torsionless connection on $\xi_{GL(4,R)}$ can be induced from the connection on $\xi_{\hat{\mathscr{C}}}$ given by (4.9). By uniqueness of the torsionless, metric connection we obtain that $\xi_{\hat{\mathscr{C}}}$ is the holonomy bundle of $\hat{g}$ through ϵ_x. ■

Now let us return to our problem of Penrose foliation. Let us recall that a linear bundle complementary to the Penrose distribution can be spanned by the vector field $Y(x)$. We can see that $Y(x)$ is parallel with respect to our Riemannian structure $\hat{g}$. It implies that the orthogonal distribution is also parallel. But this is just the distribution P tangent to the foliation $\mathscr{F}$. From a general theory of Riemannian structures[17] we obtain that $\mathscr{F}$ is totally geodesic with respect to $\hat{g}$. Besides we have that $\hat{g}$ is a bundlelike metric compatible with $\mathscr{F}$. Thus we can use the Johnson and Naveira result and relate the existence of parallel spinors on (M,g) with the topological invariants described by the Pontryagin ring of M.

ACKNOWLEDGMENTS

I wish to thank Professor H. Doebner for his hospitality at Clausthal University as well as for his help and numerous discussions.

This research was supported by the Humboldt Foundation.

[1] C. Chevalley, *The Algebraic Theory of Spinors* (Columbia U.P., New York, 1954).
[2] D. Husemoller, *Fibre Bundles* (McGraw–Hill, New York, 1966).
[3] A. Crumeyrolle, *Algebres de Clifford et Spineurs* (Toulouse, 1979).
[4] K. Bugajska, Int. J. Theor. Phys. **18**, 77 (1979).
[5] A. Lichnerowicz, C. R. Acad. Sci. Paris Ser A-B **1963**, 257.
[6] N. Hitchin, Adv. Math. **14**, 1 (1974).
[7] H. Eliasson, Math. Scand. **29**, 317 (1971).
[8] T. Aubin, J. Diff. Geom. **4**, 383 (1970).
[9] N. Trudinger, Ann. Scuola Norm. Sup. Pisa **22**, 265 (1968).
[10] J. Kazdan and F. Warner, J. Diff. Geom. **10**, 113 (1975).
[11] H. Baum, *Spin-Structuren und Dirac-Operatoren über pseudoriemannschen Mannigfaltigkeiten* (Teubner, Leipzig, 1981), Band 41.
[12] R. Geroch, J. Math. Phys. **9**, 1739 (1968).
[13] Th. Friedrich, Dissertation, Humboldt University, Berlin, 1978.
[14] P. Lounesto "On primitive idempotents of Clifford algebras," Helsinki Report-HTKK-MAT-A113, 1977.
[15] K. Bugajska, "Internal structure of space-time," submitted for publication.
[16] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (New York, Interscience, 1969), Vol. II.
[17] See Ref. 16, Vol. I (1963).
[18] B. Reinhart, *Differential Geometry of Foliations* (Springer, Berlin, 1983).
[19] F. Kamber and P. Tondeur, "Foliated bundles and characteristic classes," in *Lecture Notes in Mathematics*, Vol. 493 (Springer, Berlin, 1975).
[20] D. Johnson and A. Naveira, Geom. Dedicata **11**, 347 (1981).
[21] S. Sternberg, *Lectures on Differential Geometry* (Chelsea, New York, 1983), 2nd ed.

Hydrodynamics of a topologically nontrivial metric

Tina A. Harriott and J. G. Williams
Department of Mathematics, Mount Saint Vincent University, 166 Bedford Highway, Halifax, Nova Scotia B3M 2J6, Canada

(Received 24 April 1986; accepted for publication 2 July 1986)

The curvature and Einstein tensor are computed for a metric having one or more kinks (solitons) present. It is pointed out that the components of the fluid velocity four-vector can be identified in a natural way with certain parameters present in the metric. Making this identification, a number of hydrodynamical quantities are computed.

I. INTRODUCTION

Finkelstein and Misner[1] were the first to draw attention to the existence of an integral counting number N that could be used to classify the metrics of general relativity up to a homotopy. The classification is obtained by considering the set of homotopy classes of cross sections of the metric tensor bundle. If space-time is assumed to be $R^3 \times R^1$ then the relevant set of homotopy classes can be shown to be $[R^3 \times R^1, S_{4,1}]$, where, in the notation of Steenrod (see Ref. 2), $S_{4,1}$ denotes the set of 4×4 real symmetric matrices of signature $(-+++)$. At any instant of time, under the assumption of asymptotic flatness (or equivalently by closing R^3 with a one-point compactification), computation of this set of homotopy classes yields the group of integers Z. The number $N \in Z$ is called the kink number or soliton number of the metric. Metrics whose kink numbers differ cannot be continuously deformed into one another.

In what follows, Greek indices run over 0,1,2,3 and Latin indices run over 1,2,3. In particular, $\{x^i\}$ denote spatial coordinates and x^0 refers to time. The symbol $\eta = ||\eta_{\mu\nu}|| = \text{diag}(-1,1,1,1)$ is used to denote the Minkowski metric. Its kink number is $N = 0$. We shall use $||\delta_{\mu\nu}|| = \text{diag}(1,1,1,1)$ to denote the Kronecker delta. The four-covariant derivative of any tensor $S_{\mu\nu}$ is denoted by $S_{\mu\nu;\lambda}$ and the usual derivative is denoted by $S_{\mu\nu,\lambda}$.

Consider the metric

$$g_{\mu\nu} = \delta_{\mu\nu} - 2\phi_\mu \phi_\nu. \tag{1a}$$

The functions ϕ_μ are functions of the space-time coordinates and take values on the three-sphere: $\Sigma\phi_\mu \phi_\mu = 1$. The ϕ_μ hence control the value of N. This metric was introduced by Williams and Zia,[3] and was discussed by Shastri, Williams, and Zvengrowski.[4] Specific examples of metrics of this type have been studied by Finkelstein and McCollum,[5] Clément,[6] and Wiliams.[7] A number of related metrics (with nonzero time–space terms) have been studied by Harriott.[8]

The metric $g_{\mu\nu}$ of Eq. (1) is its own inverse, so that $g^{\mu\nu} = g_{\mu\nu}$. Defining the fields ϕ^μ to be $\phi^\mu = g^{\mu\nu}\phi_\nu$ it follows that

$$g^{\mu\nu} = \delta^{\mu\nu} - 2\phi^\mu\phi^\nu. \tag{1b}$$

The contravariant and covariant components of ϕ are related through a sign change:

$$\phi^\mu = -\phi_\mu. \tag{2}$$

The $\{\phi^\mu\}$ also define a mapping into the three-sphere: $\Sigma\phi^\mu\phi^\mu = 1$, and our preference will be to work with the ϕ^μ rather than the ϕ_μ.

In order to understand how the kink metric of Eq. (1) arises, it is helpful to consider the polar representation of $||g_{\mu\nu}||$. *Any* real nonsingular matrix M has a polar representation in which M can be written *uniquely* as the product of an orthogonal matrix Q and a positive definite symmetric matrix S:

$$M = QS.$$

Assume M to be a general relativistic metric so that M is a 4×4 real symmetric matrix of signature $(-+++)$. It follows (according to Steenrod[2]) that S is a 4×4 positive definite symmetric matrix that commutes with Q: $QS = SQ$, and that Q itself is a 4×4 symmetric orthogonal matrix of signature $(-+++)$. Furthermore, it can be shown that Q can be decomposed according to $Q = P^T\eta P$, where η is $\text{diag}(-1,1,1,1)$ and P is an orthogonal matrix.

Select any row or column of P. To be specific, suppose that the first row is selected and that its elements are denoted by $\phi^0,\phi^1,\phi^2,\phi^3$. By performing this operation, we are projecting into the base space of the $S_{4,1}$ fiber bundle. Since P is orthogonal, its rows and columns are normalized to unity so that $\Sigma\phi^\mu\phi^\mu = 1$. Thus the $\{\phi^\mu\}$ define a mapping into S^3. (More correctly, because of the $\pm$ sign ambiguity, we have a mapping into $SO_3 = RP^3$. This space is homeomorphic to a three-sphere with antipodal points identified.) If the $\{\phi^\mu\}$ define a nontrivial mapping, kinks will be present in the metric. In this way, any metric $M = QS$ can be split into a "kink part" Q and a "nonkink part" S. Almost all of the commonly studied metrics of general relativity have a trivial $Q = \text{diag}(-1,1,1,1)$ and a nontrivial S. In this paper, we have taken exactly the opposite viewpoint, namely that the nonkink part is trivial, $S = \text{diag}(1,1,1,1)$, and that the kink part Q has interesting structure. With $Q = P^T\eta P$, our metric of Eq. (1) corresponds to

$$P = \begin{bmatrix} \phi^0 & \phi^1 & \phi^2 & \phi^3 \\ -\phi^1 & \phi^0 & \phi^3 & -\phi^2 \\ -\phi_2 & -\phi^3 & \phi^0 & \phi^1 \\ -\phi^3 & \phi^2 & -\phi^1 & \phi^0 \end{bmatrix}.$$

It is usual to allow only those $\{\phi^\mu\}$ that map the infinite boundary of R^3 into some particular fixed point of S^3, say $(1,0,0,0,)$. Such a restriction leads to asymptotic flatness,

$$\lim_{|\mathbf{x}|\to\infty} g_{\mu\nu} = \eta_{\mu\nu}$$

 0022-2488/86/112706-03$02.50

and prevents any kinks present from "escaping at infinity." However, kink conservation is best considered in terms of the kink current N^μ. For the metric of Eq. (1), N^μ is identical to the usual skyrmionic current of strong interaction theory[9]:

$$N^\mu = (12\pi^2)^{-1}\epsilon^{\mu\nu\lambda\rho}\epsilon_{\alpha\beta\gamma\delta}\phi^\alpha\,\partial_\nu\phi^\beta\,\partial_\lambda\phi^\gamma\,\partial_\rho\phi^\delta. \tag{3}$$

Integrating the N^0 component over three-space yields the kink number N. Its conservation will be demonstrated in Sec. II where it will be shown that $N^\mu{}_{;\mu} = 0$, in the frame being considered.

In this paper, we study the metric of Eq. (1) in its general form. Since $\delta^{\mu\nu}$ is not a tensor, it is clear from Eq. (1b) that ϕ^μ cannot be a four-vector. Indeed, since ϕ^μ is obtained from the metric by a complicated projection procedure, it is understandable that ϕ^μ should not be as simple an object as a vector. Consequently, Eq. (3) for N^μ is only valid in the frame in which we have chosen to work. If we transform to a different frame, the expression for N^μ will change and, of course, we shall lose the simple form of the metric, as given by Eq. (1).

II. CURVATURE PROPERTIES

Our convention for the signature of the metric and the definitions of the Christoffel symbols and the Ricci and Einstein tensors is in agreement with Misner, Thorne, and Wheeler.[10] We shall work with units in which $c = 8\pi G = 1$. Since, in the frame being considered, the metric of Eq. (1) has a constant determinant (equal to -1), it follows that $\Gamma^\nu_{\mu\nu} = 0$. This has implications for the N^μ of Eq. (3). It is clear from the antisymmetric nature of Eq. (3) that $N^\mu{}_{,\mu} = 0$. Hence

$$N^\mu_{;\mu} = N^\mu_{,\mu} + \Gamma^\nu_{\mu\nu}N^\mu = 0,$$

so that N^μ is a conserved current.

The Christoffel symbols are given by

$$\Gamma^\lambda_{\mu\nu} = 2^{-1}g^{\lambda\eta}(g_{\mu\eta,\nu} + g_{\nu\eta,\mu} - g_{\mu\nu,\eta}).$$

Using the metric quantities ϕ^μ, the $\Gamma^\lambda_{\mu\nu}$ can be written

$$\Gamma^\lambda_{\mu\nu} = 2\phi_{(\mu}\partial_{\nu)}\phi^\lambda - 2\phi^\lambda\,\partial_{(\mu}\phi_{\nu)} + (\partial_\lambda - 2\phi^\lambda\phi^\eta\,\partial_\eta)(\phi_\mu\phi_\nu),$$

where ( ) indicates symmetrization.

Since $\Gamma^\nu_{\mu\nu} = 0$, the Ricci tensor simplifies to

$$R_{\mu\nu} = \Gamma^\lambda_{\mu\nu,\lambda} - \Gamma^\eta_{\mu\xi}\Gamma^\xi_{\eta\nu},$$

and hence in terms of the parameters ϕ^μ this becomes

$$\begin{aligned}R_{\mu\nu} = 2[&\phi^\lambda{}_{,\lambda}\phi^{(\mu}{}_{,\nu)} - \phi^{(\mu}{}_{,\lambda}\phi^{\lambda)}{}_{,\nu} - \phi^\lambda\phi^\gamma\phi^\mu{}_{,\lambda}\phi^\nu{}_{,\gamma}\\ &- 2\phi^\lambda\phi^\gamma{}_{,\gamma}\phi^{(\mu}\phi^{\nu)}{}_{,\lambda} + \phi^\lambda_{,\mu}\phi^\lambda_{,\nu}\\ &- \phi^\mu\phi^\nu\{\phi^\lambda_{,\gamma}\phi^\gamma_{,\lambda} - (\delta^{\lambda\gamma} - 2\phi^\lambda\phi^\gamma)\phi^\rho_{,\lambda}\phi^\rho_{,\gamma}\}\\ &+ \phi^\lambda\phi^{(\mu}{}_{,\nu)\gamma} - \phi^{(\mu}\phi^{\lambda}{}_{,\nu)\lambda}\\ &+ \phi^{(\mu}(\delta^{\lambda\gamma} - 2\phi^\lambda\phi^\gamma)\phi^{\nu)}{}_{,\lambda\gamma}].\end{aligned}$$

The curvature scalar is

$$\begin{aligned}R = g^{\mu\nu}R_{\mu\nu} = 2\{&\phi^\mu{}_{,\nu}\phi^\mu{}_{,\nu} + \phi^\mu{}_{,\mu}\phi^\nu{}_{,\nu}\\ &- \phi^\mu\phi^\nu\phi^\lambda{}_{,\mu}\phi^\lambda{}_{,\nu} + 2\phi^\mu\phi^\nu{}_{,\mu\nu}\}.\end{aligned}$$

The Einstein tensor is

$$G^\lambda_\nu = R^\lambda_\nu - \tfrac{1}{2}\delta^\lambda_\nu R,$$

where R^λ_ν splits conveniently into a symmetric part S^λ_ν, an antisymmetric part A^λ_ν, and a part V^λ_ν that vanishes on contraction:

$$\begin{aligned}R^\lambda_\nu &= S^\lambda_\nu + A^\lambda_\nu + V^\lambda_\nu,\\ S^\lambda_\nu &= 2\{\phi^\gamma{}_{,\gamma}\phi^{(\nu}{}_{,\lambda)} - \phi^{(\lambda}{}_{,\gamma}\phi^{\gamma}{}_{,\nu)} - \phi^\gamma\phi^\eta\phi^\lambda{}_{,\gamma}\phi^\nu{}_{,\eta}\\ &\quad + \phi^\lambda\phi^\nu(\phi^\gamma{}_{,\eta}\phi^\eta{}_{,\gamma} + \phi^\gamma{}_{,\gamma\eta}\phi^\eta)\\ &\quad + \phi^\gamma{}_{,\lambda}\phi^\gamma{}_{,\nu} + \phi^\gamma\phi^{(\lambda}{}_{,\nu)\gamma}\},\\ A^\lambda_\nu &= 2\{\phi^\gamma{}_{,\gamma[\nu}\phi^{\lambda]} + g^{\gamma\eta}\phi^{[\nu}\phi^{\lambda]}{}_{,\gamma\eta}\},\\ V^\lambda_\eta &= 2\phi^\gamma\{\phi^\lambda\phi^\nu{}_{,\eta}\phi^\eta_{,\gamma} - \phi^\nu\phi^\eta{}_{,\eta}\phi^\lambda{}_{,\gamma}\}.\end{aligned}$$

The symbol [ ] denotes antisymmetrization.

III. FLUID VELOCITY IDENTIFICATION AND DETERMINATION OF OTHER KINEMATICAL QUANTITIES

In a fluid model, one assumes the existence of a unique velocity field u^μ representing the average velocity of matter.

In terms of general coordinates x^μ, $u^\mu = (dx^\mu/d\tau)|_{y^i}$, where (y^i,τ) are local comoving coordinates. The y^i label the fluid particles in an arbitrarily chosen space section of spacetime, and τ labels proper time measured from this space section along the fluid flow lines $y^i = \text{const}$. The following normalization condition is a direct consequence of the definition of u^μ:

$$u^\mu u_\mu = -1. \tag{4}$$

Since Eq. (2) implies $\phi^\mu\phi_\mu = g_{\mu\nu}\phi^\mu\phi^\nu = -1$, Eq. (4) strongly suggests that the velocity be identified with the kink parameter:

$$u^\mu = \phi^\mu. \tag{5}$$

This identification will be assumed from now on. Since ϕ^μ is not a vector, Eq. (5) is not covariant. We are simply choosing to work in a frame in which the relationship between u^μ and ϕ^μ is postulated to take a particularly simple form.

Following Ellis,[11] we note that the general relativistic stress-energy tensor $T_{\mu\nu}$ can be written

$$T_{\mu\nu} = \rho u_\mu u_\nu + 2q_{(\mu}u_{\nu)} + ph_{\mu\nu} + \pi_{\mu\nu}.$$

The projection tensor $h_{\mu\nu}$ is defined by

$$h_{\mu\nu} = g_{\mu\nu} + u_\mu u_\nu.$$

The function ρ is the energy density, p the kinetic pressure, q_μ the energy flux (due to diffusion, heat conduction, etc.), and $\pi_{\mu\nu}$ the anisotropic pressure (viscosity) term. Note that

$$q_\mu u^\mu = \pi^\mu_\mu = \pi_{\mu\nu}u^\nu = 0.$$

These functions occur in or are related to functions that occur in the following decomposition[11] of the covariant derivative of the velocity,

$$u_{\mu;\nu} = \omega_{\mu\nu} + \sigma_{\mu\nu} + \tfrac{1}{3}\theta h_{\mu\nu} - \dot{u}_\mu u_\nu.$$

The vorticity tensor $\omega_{\mu\nu}$ is defined by $\omega_{\mu\nu} = u_{[\mu;\lambda}h^\lambda_{\nu]}$, where [ ] denotes antisymmetrization. The shear tensor $\sigma_{\mu\nu}$ is defined by $\sigma_{\mu\nu} = u_{(\mu;\lambda}h^\lambda_{\nu)} - \theta h_{\mu\nu}/3$. The function θ denotes the isotropic (volume) expansion $\theta = u^\mu_{;\mu}$ and $\dot{u}_\mu$ is

the acceleration vector defined by $\dot{u}^\mu = u^\mu_{;\nu}u^\nu$ with $\dot{u}^\mu u_\mu = 0$. The vorticity and shear tensors satisfy $\omega_{\mu\nu}u^\nu = \sigma_{\mu\nu}u^\nu = 0$.

A scalar vorticity and a scalar shear can be defined according to $\omega = (\omega^{\mu\nu}\omega_{\mu\nu}/2)^{1/2}$ and $\sigma = (\sigma^{\mu\nu}\sigma_{\mu\nu}/2)^{1/2}$. The above tensors can be interrelated by phenomenological equations which are usually[11] postulated to take the form

$$\pi_{\mu\nu} = -\lambda\sigma_{\mu\nu},$$
$$q_\mu = -\kappa h^\nu_\mu (T_{;\nu} + T\dot{u}_\nu),$$

where T denotes temperature and λ,κ are positive constants.

Using the identification of Eq. (5) and the metric of Eq. (1), the various hydrodynamical quantities simplify as follows:

$$h_{\mu\nu} = \delta_{\mu\nu} - u_\mu u_\nu, \quad u^\mu{}_{;\nu} = -\{\partial_\mu\phi_\nu - \phi^\gamma\partial_\gamma(\phi^\mu\phi_\nu)\},$$
$$u_{\mu;\nu} = -\{\partial_\mu\phi_\nu + \phi^\gamma\partial_\gamma(\phi^\mu\phi_\nu)\}, \quad \dot{u}_\mu = \phi^\nu\partial_\nu\phi_\mu,$$
$$\theta = \partial_\mu\phi^\mu, \quad \omega_{\mu\nu} = \partial_{[\nu}\phi_{\mu]} + \phi^\eta\phi_{[\nu}\partial_\eta\phi_{\mu]},$$
$$\sigma_{\mu\nu} = -\{\partial_{(\nu}\phi_{\mu)} + \phi^\eta\phi_{(\nu}\partial_\eta\phi_{\mu)} + \partial_\eta\phi^\eta(\delta_{\mu\nu} - \phi_\mu\phi_\nu)/3\},$$
$$\omega^2 = \tfrac{1}{2}\partial_\nu\phi_\mu\partial_{[\nu}\phi_{\mu]}, \quad \sigma^2 = \tfrac{1}{2}\partial_\nu\phi_\mu\partial_{(\nu}\phi_{\mu)}.$$

IV. SUMMARY AND CONCLUSIONS

In this paper we have studied a metric whose form is sufficiently simple to allow the straightforward computation of the usual tensors of general relativity and yet whose form has sufficient structure to allow for the existence of kinks. The relation $\phi^\mu\phi_\mu = -1$ justified our equating the kink parameter ϕ^μ with the fluid velocity. This may allow inhomogeneous and anisotropic space-time solutions. The ultimate justification, of course, must lie with demonstrating the consistency of the Einstein equations $G_{\mu\nu} = T_{\mu\nu}$, using the $T_{\mu\nu}$ appropriate for a realistic fluid. This is the next stage of the work. Since "tumbling" light cones[5] are a feature of space-times with kink metrics, there will be regions (perhaps the size of elementary particles or perhaps within black holes) where causality is violated. Their interpretation is also an outstanding problem.

ACKNOWLEDGMENTS

We would like to acknowledge a number of helpful conversations with Dr. Alan Coley, Dr. Ken Dunn, Dr. Werner Israel, and Dr. C. W. Misner. Thanks are due to Mount Saint Vincent University for supporting this research through Grant No. A04-006. One of us (JGW) would like to acknowledge the award of a research grant from the Natural Sciences and Engineering Research Council of Canada.

[1] D. Finkelstein and C. W. Misner, Ann. Phys. (NY) **6**, 230 (1959); "Further results in topological relativity," in *Les Théories Relativistes de la Gravitation, Actes du Colloque International*, organisé à Royaumont, 21–27 June 1959, edited by M. A. Lichnerowicz and M. A. Tonnelat (C.N.R.S., Paris, 1962).

[2] N. Steenrod, *The Topology of Fibre Bundles* (Princeton U.P., Princeton, NJ, 1951), Chap. 40.

[3] J. G. Williams and R. K. P. Zia, J. Phys. A **6**, 1 (1973).

[4] A. R. Shastri, J. G. Williams, and P. Zvengrowski, Int. J. Theor. Phys. **19**, 1 (1980).

[5] D. Finkelstein and G. McCollum, J. Math. Phys. **16**, 2250 (1975).

[6] G. Clément, Gen. Relativ. Gravit. **16**, 131, 477 (1984).

[7] J. G. Williams, Lett. Nuovo Cimento **43**, 282 (1985).

[8] T. A. Harriott, Ph.D. Thesis, Dalhousie University, 1986 (unpublished).

[9] T. H. R. Skyrme, Proc. R. Soc. London Ser. A **260**, 127 (1961); A. P. Balachandran, "The Skyrme soliton–A review," in *Solitons in Nuclear and Elementary Particle Physics, Proceedings of the Lewes Workshop*, 2–16 June 1984, edited by A. Chodos, E.Hadjimichael, and C. Tze (World Scientific, Singapore, 1984).

[10] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[11] G. F. R. Ellis, "Relativistic cosmology," in *Proceedings of the International School of Physics "Enrico Fermi," Course XLVII–General Relativity and Cosmology*, held at Varenna on Lake Como, 30 June–12 July 1969, edited by B. K. Sachs (Academic, New York, 1971).

Quantum causal structures

L. Szabó
Institute for Theoretical Physics, Eötvös University, Budapest, Hungary

(Received 24 April 1986; accepted for publication 18 June 1986)

The axioms of a causal structure are reformulated. A natural generalization is suggested for the case when the subset lattice of space-time events is replaced by a lattice coming from the quantum theory.

I. INTRODUCTION

One of the most important consequences of the quantum theory is that our basic concepts of the probability theory have drastically changed. The classic Boole-lattice-based probability theory has been succeeded by a new one based on a more general lattice structure of physical events.[1] In mathematical physics one usually starts to work with the "natural" Boolean lattice of physical events. Therefore it was a very important step in the "quantization" of basic ideas when the subset structure in the definition of topology had been replaced by the quantum lattice.[2]

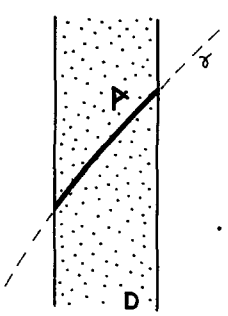
In this paper following this "quantization" program, axioms are suggested describing the causal structure of physical events. In this new causal structure the role of the lattice of space-time subsets is played by the dual of the quantum lattice of events.

II. EVENTS IN QUANTUM THEORY AND EVENTS IN SPACE-TIME

The usual definition of an event in (quantum) physics is the following: An event means a possible result of a possible observation performed on a physical object.

At first sight this notion seems to be very far from the notion of a space-time point. However each space-time point—being an event as well—can be formulated in the language of physical observations. This fact suggests that there must be some relation between these two notions.

There are physical events that do not correspond to a single space-time point, but they do correspond to a collection of space-time points, i.e., to a subset of the space-time. For example the event "the cloud camera D has detected the particle γ" corresponds to the subset A:



Thus the physical events can be identified with the subsets (the single points included) of space-time.

A subset in space-time means a complex of events contained is the subset. Therefore the union of subsets A and B corresponds to the conjunction of events identified with A and B. Inversely, the common part of A and B means the disjunction of the corresponding physical events. This means that the lattice of physical events and the lattice of space-time subsets are dual to each other. In classical physics this correspondence can be correct since the lattice of physical events is Boolean. But it cannot be correct in quantum theory, because quantum logic is not Boolean, consequently, the dual lattice is not Boolean either. A possible resolution of this inadequacy is if the Boolean subset lattice is exchanged for the dual of the quantum lattice of events, and if the whole space-time structure is built on this ground up. As an initial effort one can reformulate the axioms of causality according to the conception above.

III. THE ROLE OF THE SUBSET LATTICE IN THE CAUSAL STRUCTURE

Let us recall the Kronheimer–Penrose axioms of a causal structure. Let X be the underlying set. Two relations $<_c$ and $\ll$ are given with the following properties[3]:

$$\begin{aligned}
&\text{(i) } x <_c x,\\
&\text{(ii) if } x <_c y \text{ and } y <_c z, \text{ then } x <_c z,\\
&\text{(iii) from } x <_c y \text{ and } y <_c x \text{ follows that } x = y,\\
&\text{(iv) not } x \ll x,\\
&\text{(v) if } x \ll y, \text{ then } x <_c y,\\
&\text{(vi) if } x <_c y \text{ and } y \ll z, \text{ then } x \ll z,\\
&\text{(vii) if } x \ll y \text{ and } y <_c z, \text{ then } x \ll z,
\end{aligned} \tag{1}$$

where $x, y, z \in X$.

The causal future set and the chronological future set are defined as

$$J^+(A) =: \{x \in X \mid \text{there exists } a \in A \text{ such that } a <_c x\},$$

$$I^+(A) =: \{x \in X \mid \text{there exists } a \in A \text{ such that } a \ll x\}.$$

The causal and chronological future sets have the following properties:

$$\begin{aligned}
&\text{(i) } J^+(A) \supset A,\\
&\text{(ii) if } J^+(\{x\}) \supset \{y\} \text{ and } J^+(\{y\}) \supset \{x\}, \text{ then } x = y,\\
&\text{(iii) if } A \subset J^+(B) \text{ and } B \subset J^+(C), \text{ then } A \subset J^+(C),\\
&\text{(iv) } J^+(A \cup B) = J^+(A) \cup J^+(B),\\
&\text{(v) } J^+(A \cap B) \subset J^+(A) \cap J^+(B),\\
&\text{(vi) if } A \subset J^+(B) \text{ and } B \subset I^+(C), \text{ then } A \subset I^+(C),\\
&\text{(vii) if } A \subset I^+(B) \text{ and } B \subset J^+(C), \text{ then } A \subset I^+(C),\\
&\text{(viii) not } \{x\} \subset I^+\{x\},
\end{aligned} \tag{2}$$

where A, B, and C are subsets of X.

 0022-2488/86/112709-02$02.50

It is obvious that a pair of maps on the subset lattice

$$J^+: \mathscr{P}(X) \to \mathscr{P}(X), \quad I^+: \mathscr{P}(X) \to \mathscr{P}(X),$$

satisfying (2) define a causal structure in the sense of (1) via the following definition of causal and chronological relations:

$$x <_c y \quad \text{iff} \quad y \in J^+(\{x\}), \quad x \ll y \quad \text{iff} \quad y \in I^+(\{x\}).$$

One can introduce many different topologies on X. The most reasonable of them is the Alexandrov topology, i.e., the coarsest topology on X in which each $I^+(A)$ is open.

IV. QUANTUM CAUSAL STRUCTURES

Let us now replace the subset lattice $\mathscr{P}(X)$ by the dual to the quantum lattice. Denote Q the quantum lattice of events and let $(S, \wedge, \vee) = Q^*$ be the dual of Q.

Definition: A causal structure is a pair of maps

$$J^+: S \to S, \quad I^+: S \to S,$$

with the following properties:

(i) $J^+(A) > A$ and $I^+(A) < J^+(A)$,

(ii) if $A < J^+(B)$ and $B < J^+(C)$, then $A < J^+(C)$,

(iii) for any $x, y \in \mathscr{A}(S)$ from $x < J^+(y)$ and $y < J^+(x)$ follows that $x = y$,

(iv) $J^+(A \vee B) = J^+(A) \vee J^+(B)$, (3)

(v) $J^+(A \wedge B) < J^+(A) \wedge J^+(B)$,

(vi) $I^+(A \vee B) = I^+(A) \vee I^+(B)$,

(vii) $I^+(A \wedge B) < I^+(A) \wedge I^+(B)$,

(viii) if $A < J^+(B)$ and $B < I^+(C)$, then $A < I^+(C)$,

(ix) if $A < I^+(B)$ and $B < J^+(C)$, then $A < I^+(C)$,

(x) not $x < I^+(x)$, for $x \in \mathscr{A}(S)$,

where $A, B, C \in S$ and $\mathscr{A}(S)$ denotes the set of atoms in S. If S is a Boolean lattice it can be represented by a suitable subset lattice and the causality defined above leads to the usual Kronheimer–Penrose causality.

V. ALEXANDROV *T*-STRUCTURE

In case S is not Boolean it cannot be equivalent to any subset lattice, therefore one cannot define a point set topology on an "underlying set of causal structure." Fortunately there is a nice generalization of the topology for a non-Boolean lattice.[2]

Denote $\mathscr{F}(S)$ the set of filters of S and $\Omega(S)$ the set of maximal filters. A T-structure on S is a map

$$T: \Delta \subset \Omega(S) \to \mathscr{F}(S)$$

such that (i) $T(A) \subset A$, and (ii) for any $B \in T(A)$ there exists $C \in T(A)$ such that $C < B$ and

$$C \in \bigcap_{C \subset D} T(D).$$

An element $B \in S$ is said to be open if

$$B \in \bigcap_{B \subset A} T(A).$$

One can define a T-structure on S (let us call it Alexandrov T-structure) associated to the causal structure as the coarsest T-structure in which for any $A \in S$ the chronological future $I^+(A)$ is open.

VI. CONCLUSIONS

In a quantum causal structure one can define the causal and chronological relations as

$$A <_c B \quad \text{iff} \quad B < J^+(A), \quad A \ll B \quad \text{iff} \quad B < I^+(A). \quad (4)$$

It is reasonable to regard the set of atoms $\mathscr{A}(S)$ as the "space-time set." One can restrict the relations (3) for the atoms. These restricted relations on $\mathscr{A}(S)$ satisfy the Kronheimer–Penrose axioms. However there can be many relations on $\mathscr{A}(S)$ satisfying the axioms which are not generated by any quantum causal structure on the whole S. It means that we have a possibly strong physical restriction for the possible causal relations.

A possible relevancy of the quantum causal structures to the analysis of the "delayed-choice experiments"[4] will be discussed later.

ACKNOWLEDGMENT

I am indebted to Dr. L. B. Szabados for the continuous fruitful discussions.

[1] J. M. Jauch, *Foundation of Quantum Mechanics* (Addison–Wesley, Reading, MA, 1968).

[2] A. R. Marlow, Int. J. Theor. Phys. **19**, 516 (1980).

[3] E. H. Kronheimer and R. Penrose, Proc. Cambridge Philos. Soc. **63**, 481 (1967).

[4] J. A. Wheeler, in *Mathematical Foundations of Quantum Theory*, edited by A. R. Marlow (Academic, New York, 1978).

Special solutions of singular nonlinear Schrödinger equations with polynomial nonlinearities

L. Brüll and H.-J. Kapellen
Mathematisches Institut der Universität zu Köln, Weyertal 86-90, 5000 Köln 41, West Germany

(Received 13 March 1986; accepted for publication 18 June 1986)

The singular nonlinear Schrödinger equation $iu_t = -u_{xx} + f(|u|^2)u + \chi h(|u|^2)h'(|u|^2)u$, where f has the form $f(s) = as^n$, $n \geqslant 1$, $a \in \mathbb{R}$, is investigated. A classification is given of those nonlinearities f and h that allow the existence of solitary waves and kink solutions. Further, in several cases the solutions are given in explicit form.

I. INTRODUCTION

In this short paper we consider singular nonlinear Schrödinger equations of the form

$$iu_t = -u_{xx} + f(|u|^2)u + \chi h(|u|^2)_{xx}h'(|u|^2)u, \tag{1}$$

where f has the form $f(s) = as^n$, $n \geqslant 1$, $a \in \mathbb{R}$. For the case $\chi = 0$ we have the usual nonlinear Schrödinger equation, while for the case $\chi \neq 0$ this equation appears frequently in recent physical literature concerning, for example, plasma physics,[1] superfluid films,[2] or the Heisenberg ferromagnetic spin chain.[3] More examples in both cases $\chi = 0$ and $\chi \neq 0$ can be given.[4]

For $\chi = 0$ much work has been done on the investigation of special solutions of (1) such as solitary waves or kink solutions (see, for example, Ablowitz/Segur[5] or Berestycki/Lions[6]). In this paper we classify the nonlinearities $f(s) = a \cdot s^n$ and h that allow special solutions of the described form and give explicit solitary waves, kink solutions, and spatial periodic traveling waves to some equations of the form (1).

II. SOLITARY WAVES AND KINK SOLUTIONS

Let us consider solutions of (1) of the form

$$\begin{aligned} u(x,t) &= r(x+ct)e^{i\phi(x+dt)}, \\ \phi(x+dt) &= -(c/2)(x+dt). \end{aligned} \tag{2}$$

It is possible to show[4,7] that in the case of solitary waves (for arbitrary space dimension n) and in the case of traveling waves—where r changes sign—the phase has to take the form (2).

Substituting (2) into (1) yields the nonlinear singular scalar field equation

$$\begin{aligned} r'' &= r[\lambda + f(r^2) + \chi(h(r^2))_{xx}h'(r^2)], \\ \lambda &= (c/2)(c/2 - d). \end{aligned} \tag{3}$$

Here the derivative is taken with respect to the argument $x + ct$, which we abbreviate in the following by x. Hence (3) can be rewritten as

$$\begin{aligned} r''(1 - 2\chi h'^2(r^2)r) = r[\lambda + f(r^2) + 4\chi h''(r^2)h'(r^2)r'^2r] \\ + 2\chi h'^2(r^2)r'^2. \end{aligned} \tag{4}$$

Obviously, the existence and characterization of critical points of Eq. (4) is decisive in the study of solitary waves and kink solutions. Putting $r'' = r' = 0$ in (4) yields

$$r(\lambda + f(r^2)) = 0. \tag{5}$$

Hence the critical points of (4) are determined only by the nonlinearity $f(s)$.

Most applications deal with polynomial nonlinearities $f(s) = a \cdot s^n$, $n \geqslant 1$. For that reason and for simplicity we study only this case. The case of arbitrary f can be handled by analogous phase plane arguments in nearly the same way.

Obviously, $r = 0$ is a critical point of (4). The others are given by

$$r_0 = \sqrt[2n]{-\lambda/a}, \tag{6}$$

provided $\lambda/a < 0$.

Now it is easily seen that a solitary wave type solution [that means a solution with $\lim_{|x|\to\infty} r(x) = 0$] corresponds to a homoclinic orbit of the critical point 0, whereas a kink solution [that means a solution with $\lim_{x\to\infty} r(x) = a \neq b = \lim_{x\to-\infty} r(x)$] corresponds to a heteroclinic orbit in the (r,r')-phase plane. Since a homoclinic orbit always includes a critical point, $\lambda/a < 0$ is necessary to get a solitary wave or kink solution of (1). We show in the following that this condition is sufficient in some sense.

Theorem 1: Assume

$$a > 0, \quad \lambda < 0, \quad r_0 := \sqrt[2n]{-\lambda/a}\,.$$

Then (1) has a kink solution if

$$\chi h'^2(r^2) \cdot r \neq \tfrac{1}{2}, \quad \forall r \in [0, r_0] \tag{7a}$$

holds.

Theorem 2: Assume

$$a < 0, \quad \lambda > 0, \quad r_1 := \sqrt[2n]{-\lambda(n+1)/a}.$$

Then (1) has a solitary wave type solution if

$$\chi h'^2(r^2) \cdot r \neq \tfrac{1}{2}, \quad \forall r \in [0, r_1] \tag{7b}$$

holds.

Remarks: (a) The conditions (7a) and (7b) are necessary to insure that Eq. (4) does not become singular in the region of interest.

(b) Since we can change the parameter $\lambda = (c/2) \times (c/2 - d)$, we get infinitely many kink solutions for $a > 0$ and infinitely many solitary waves for $a < 0$.

Proof of Theorem 1: It is easy to see that the (r,r') phase plane of (4) is symmetric with respect to the r and r' axes. This implies that a critical point of (4) is either a center or a hyperbolic saddle point. Further one can prove that between two saddle points there must be a center and between two centers there must be a saddle. Because of $\lambda < 0$, (0,0) is a center and hence $(-r_0,0)$, $(r_0,0)$ are hyperbolic saddle

 0022-2488/86/112711-03$02.50

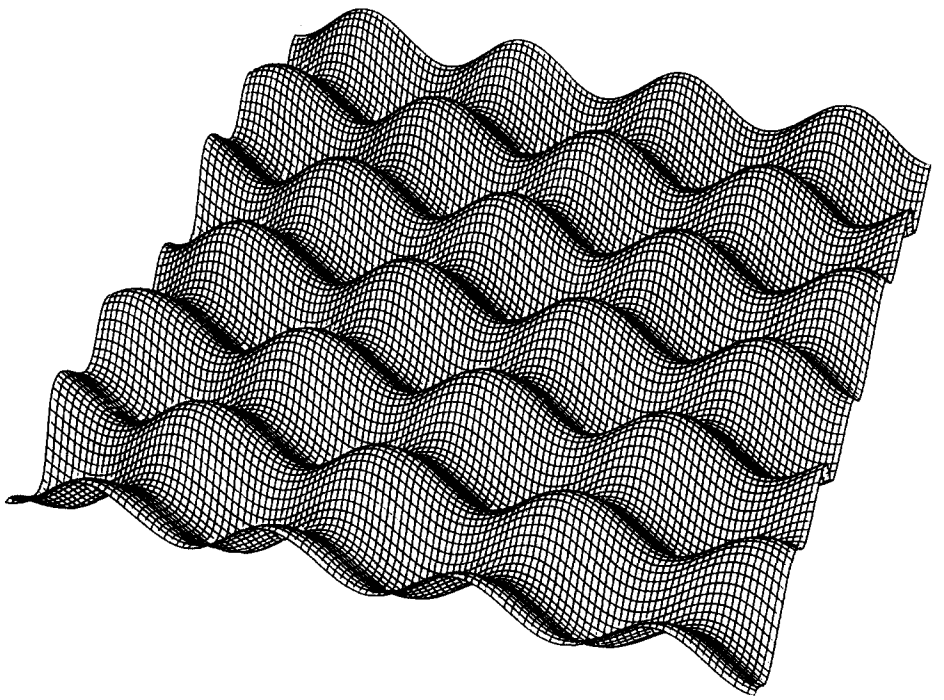


FIG. 1. The real part (egg crate) of the solution of the periodic traveling wave with $a = -6$, $\chi = -\frac{3}{2}$, $c = -2$, $d = 1$, and $c_3 = 0$.

points. Since there are no more critical points, the right part of the unstable manifold of $(-r_0,0)$ is identical with the left part of the stable manifold of $(r_0,0)$. Hence, a heteroclinic orbit exists.

Remarks: (a) The proof shows that for the described kink solution

$$\lim_{x\to\infty} r(x) = -\lim_{x\to-\infty} r(x)$$

holds.

(b) Theorem 2 can be proved by phase space arguments similar to those of the proof of Theorem 1.

III. SPECIAL SOLUTIONS

We consider in the following the equation

$$iu_t = -u_{xx} + a|u|^2u + b|u|^4u + \chi\partial_{xx}|u|^2u, \tag{8}$$

and look for solutions of the form (2).

The analogous equation to (3) now reads

$$r'' = r(\lambda + ar^2 + br^4 + 2\chi(rr'' + r'^2)). \tag{9}$$

Multiplying by $2r'$ and integrating yields

$$r'^2 = \lambda r^2 + (a/2)r^4 + (b/3)r^6 + 2\chi r^2r'^2 + D, \tag{10}$$

which can be rewritten as

$$r'^2 = -\frac{b}{6\chi}r^4 - \frac{1}{2\chi}\left(\frac{b}{6\chi} + \frac{a}{2}\right)r^2 - \xi + \frac{D+\xi}{1-2\chi r^2}, \tag{11}$$

where

$$\xi = \xi(a,b,c,d,\chi) = \frac{1}{2\chi}\left[\frac{1}{2\chi}\left(\frac{b}{6\chi} + \frac{a}{2}\right) + \lambda\right].$$

Now we choose d such that

$$\xi(a,b,c,d,\chi) = -D,$$

and hence (11) yields

$$r'^2 = -\frac{b}{6\chi}r^4 - \frac{1}{2\chi}\left(\frac{b}{6\chi} + \frac{a}{2}\right)r^2 + D. \tag{12}$$

(a) Solitary waves: In order to get solitary waves we need $D = 0$ in (12). Assuming $b/6\chi \geqslant 0$ and

$$-\frac{1}{2\chi}\left(\frac{b}{6\chi} + \frac{a}{2}\right) > 0 \quad \text{and} \quad \lambda = -\frac{1}{2\chi}\left(\frac{b}{6\chi} + \frac{a}{2}\right)$$

we find by simple integration the solitary wave type solution

$$r(x+ct) = \pm\sqrt{-\frac{3}{b}\left(\frac{b}{6\chi} + \frac{a}{2}\right)} \times \operatorname{sech}\left(\sqrt{-\frac{1}{2\chi}\left(\frac{b}{6\chi} + \frac{a}{2}\right)}(x+ct) + c_1\right), \tag{13}$$

where c_1 is an arbitrary real constant.

(b) Kink-type solutions: Assuming

$$\chi < 0, \quad b > 0, \quad (a/2 + b/6\chi) < 0,$$

and

$$\lambda = \frac{3}{2}a\left(\frac{b}{6\chi} + \frac{a}{2}\right) + \frac{9}{4}b\left(\frac{b}{6\chi} + \frac{a}{2}\right)^2$$

we find by integration the solution

$$r(x+ct) = \pm\sqrt{-\frac{3}{2}\left(\frac{a}{2} + \frac{b}{6\chi}\right)} \times \tanh\left(\sqrt{\frac{b}{4\chi}\left(\frac{a}{2} + \frac{b}{6\chi}\right)}(x+ct) + c_2\right),$$

with an arbitrary real constant c_2.

(c) Periodic traveling waves: Assume $b = 0$, $a/4\chi > 0$, and choose λ in such a manner that

$$(2/a)(\lambda + a/4\chi) < 0.$$

Then integration yields

$$r(x+ct)=\sqrt{-\frac{2}{a}\left(\lambda+\frac{a}{4\chi}\right)}\times\sin\left(\sqrt{\frac{a}{4\chi}}\,(x+ct)+c_3\right),$$

where c_3 is an arbitrary real constant.

Figure 1 shows the real part (egg crate) of the solution of the periodic traveling wave with $a=-6$, $\chi=-\frac{3}{2}$, $c=-2$, $d=1$, and $c_3=0$.

ACKNOWLEDGMENT

One of us (H.-J. K.) was supported by the Deutsche Forschungsgemeinschaft under Grant No. BA 735/3-2.

[1] M. Goldman, "Strong turbulence of plasma waves," Rev. Mod. Phys. **56**, 709 (1984).

[2] S. Kurihara, "Large-amplitude quasi-solitons in superfluid films," J. Phys. Soc. Jpn. **50**, 3262 (1981).

[3] S. Takeno and S. Homma, "Classical planar Heisenberg ferromagnet, complex scalar field and nonlinear excitations," Prog. Theor. Phys. **65**, 172 (1981).

[4] L. Brüll and H. Lange, "Stationary, oscillatory and solitary wave type solution of singular nonlinear Schrödinger equations," Expositiones Math. (to appear).

[5] M. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).

[6] H. Berestycki and P.-L. Lions, "Nonlinear scalar field equations. I. Existence of a ground state," Arch. Rat. Mech. Anal. **82**, 313 (1983).

[7] L. Brüll, "On solitary waves for nonlinear Schrödinger equations in higher dimensions," Appl. Anal. (to appear).

On renormalization in nonrelativistic interaction models with infinities

Helmut Kröger, Anais Smailagic, and Ralph Girard
Département de Physique, Université Laval, Québec, Province of Quebec, Canada G1K 7P4

(Received 21 November 1985; accepted for publication 11 July 1986)

The Green's function of two models with nonrelativistic separable interaction giving rise to infinities in the perturbation expansion is studied. These infinities do not arise from the $E + i0$ limit, but come from the slow falloff behavior of the vertices, modeled after the infinities in Feynman graphs of field theory. Both models are analytically solvable. It is found that the Green's function obtained from summing the renormalized perturbation series is identical to the direct solution of the Green's function, which requires only an intermediate regularization. In the first model the interaction is split in a singular part giving raise to infinities and a regular part. It is shown that the Green's function is the same as the Green's function derived from only the regular part. This effect is similar to the effect occurring in ϕ^4 field theory in $3 + 1$ dimensions, where the ϕ^4 interaction vanishes after renormalization and the S matrix is trivial. The second model is constructed such that parts of the singular interaction survive in the Green's function.

I. INTRODUCTION

It is well known that perturbation theory has been successful in describing quantum electrodynamics (QED) and to a certain extent the short distance behavior of quantum chromodynamics (QCD). On the other hand nonperturbative effects are essential for describing the confinement region in QCD. Recently new evidence has been found for the need of nonperturbative contributions in p–p scattering.[1]

One approach to overcome the limitations of perturbation theory and extend results to the large coupling regime relies on the Borel summation technique.[2] Another very successful technique is the lattice approach.[3] Several other methods have been discussed in the literature.[4–7] The authors have proposed recently another approach,[8,9] which is similar to the Hamiltonian formulation of lattice theory.

In these nonperturbative methods one is faced with the question of renormalization. In Ref. 10, Wilson, applying standard cutoff regularization, has found no need for wave-function and coupling-constant renormalization, contrary to standard renormalization of perturbation theory. Hence one can ask if there is a need for renormalization inherent in physics or if it is an artifact to overcome difficulties with a mathematically ill-defined theory.

In order to investigate these questions, we study in this paper analytically solvable nonrelativistic models. We choose interactions that generate infinities in the perturbative expansion of the Green's function, with the intention to model the infinities of Feynman graphs in relativistic field theories. However, these infinities do not arise due to the $E + i0$ limit, but come from the falloff behavior of the interaction. We have chosen the interaction to be separable, which leads to an analytically solvable Green's function. Separable potentials were introduced in nuclear physics as early as 1954 by Yamaguchi[11] and were used by Mitra[12] in 1962 to facilitate the solution of the Schrödinger equation.

In this paper we have compared two methods for calculating the full Green's function: the standard perturbative approach and the direct analytical solution. In both methods we find the same Green's function. Although both methods require intermediate regularization, only the perturbative expansion requires renormalization, which turns out to be finite.

II. MODEL I

We consider in three dimensions a nonrelativistic Hamiltonian $H = H^0 + H^{\text{int}}$, with H^0 being the free Hamiltonian and H^{int} being the interaction. Let $G(z) = (z - H)^{-1}$ and $G^0(z) = (z - H^0)^{-1}$ denote the corresponding full and free Green's function, respectively. We consider a separable interaction

$$H^{\text{int}} = \sum_{i,j=1}^{3} |\chi_i\rangle \lambda_{ij} \langle \chi_j|, \tag{2.1}$$

because it allows for an analytical solution of the Green's function given by

$$G^{\text{anl}}(z) = G^0(z) + \sum_{i,j=1}^{3} G^0(z)|\chi_i\rangle g_{ij}(z) \langle \chi_j| G^0(z), \tag{2.2}$$

where

$$g(z) = \lambda (1 - g^0(z)\lambda)^{-1}, \tag{2.3}$$

$$g^0(z)_{ij} = \langle \chi_i | G^0(z) | \chi_j \rangle. \tag{2.4}$$

This expression for G is only meaningful when the matrix g^0 exists and where g has no poles. G can be expanded in a perturbation series as

$$\begin{aligned} G^{\text{pert}}(z) &= G^0(z) + G^0(z) H^{\text{int}} G^0(z) \\ &\quad + G^0(z) H^{\text{int}} G^0(z) H^{\text{int}} G^0(z) + \cdots \\ &= G^0(z) + \sum_{ij} G^0(z)|\chi_i\rangle \lambda_{ij} \langle \chi_j| G^0(z) \\ &\quad + \sum_{ijkl} G^0(z)|\chi_i\rangle \lambda_{ij} \langle \chi_j| G^0(z)|\chi_k\rangle \\ &\quad \times \lambda_{kl} \langle \chi_l| G^0(z) + \cdots . \end{aligned} \tag{2.5}$$

We want to have the property that perturbation theory gives infinite contributions in any order, which happens if some of

 0022-2488/86/112714-06$02.50

the matrix elements g^0_{ij} are infinite. We want to model the infinities occurring in Feynman graphs of field theories, which we do by assuming a slow falloff behavior in momentum space of the vertices. At present we take $z = E + i\epsilon$, $\epsilon \neq 0$, while the physical limit $\epsilon \to +0$ will be discussed later. In the first model we take

$$|\chi_i\rangle = \int d\mathbf{q}_1\, d\mathbf{q}_2\, \chi_i(\mathbf{q}_1,\mathbf{q}_2)|\mathbf{q}_1,\mathbf{q}_2\rangle, \quad i = 1,2,3, \tag{2.6}$$

with

$$\chi_1(\mathbf{q}_1,\mathbf{q}_2) = 1/\sqrt{q_1 q_2}. \tag{2.7}$$

This gives an infinite matrix element g^0_{11}.

By introducing a cutoff Λ in momentum space we regularize it and obtain

$$g^0_{11}(z,\Lambda) = 4\pi^2\left[2\Lambda^2 \ln\left(\frac{z-2\Lambda^2}{z-\Lambda^2}\right) + z\ln(z-2\Lambda^2) - 2z\ln(z-\Lambda^2) + z\ln(z)\right], \tag{2.8}$$

which behaves asymptotically as

$$g^0_{11} \underset{\Lambda\to\infty}{\sim} 4\pi^2[2\ln(2)\Lambda^2 - 2z\ln(\Lambda)]. \tag{2.9}$$

The other vertices χ_2, χ_3 we choose with a sufficiently fast falloff behavior such that

$\langle\chi_i|G_0|\chi_j\rangle$ finite if $(i,j) \neq (1,1)$.

For our purposes there is no need to specify them explicitly. We denote by $\chi_i(\Lambda)$ the cutoff vertex and by $H^{\text{int}}(\Lambda)$, $G(z,\Lambda)$ the corresponding interaction Hamiltonian and the full Green's function, respectively. One has, in analogy to Eqs. (2.2)–(2.4),

$$G(z,\Lambda) = G^0(z) + \sum_{i,j=1}^{3} G^0(z)|\chi_i(\Lambda)\rangle \times g_{ij}(z,\Lambda)\langle\chi_j(\Lambda)|G^0(z), \tag{2.10}$$

where g obeys the matrix equation

$$g(z,\Lambda) = \lambda(1 - g^0(z,\Lambda)\lambda)^{-1}, \tag{2.11}$$

with

$$g^0(z,\Lambda)_{ij} = \langle\chi_i(\Lambda)|G^0(z)|\chi_j(\Lambda)\rangle. \tag{2.12}$$

Now let us consider for simplicity the particular case

$$\lambda_{ij} = \lambda\delta_{ij} \tag{2.13}$$

and let us calculate the behavior of g for large Λ. One finds a stable limit given by

$$\lim_{\Lambda\to\infty} g(z,\Lambda) = \lambda \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1-\lambda g^0_{33}(z) & \lambda g^0_{23}(z) \\ 0 & \lambda g^0_{32}(z) & 1-\lambda g^0_{22}(z) \end{pmatrix} \times \left[\det\begin{pmatrix} 1-\lambda g^0_{22}(z) & -\lambda g^0_{23}(z) \\ -\lambda g^0_{32}(z) & 1-\lambda g^0_{33}(z) \end{pmatrix}\right]^{-1}. \tag{2.14}$$

If we define

$$g^0_{(23)}(z) = \begin{pmatrix} g^0_{22}(z) & g^0_{23}(z) \\ g^0_{32}(z) & g^0_{33}(z) \end{pmatrix}, \tag{2.15}$$

and

$$g_{(23)}(z) = \lambda(1 - \lambda g^0_{(23)}(z))^{-1}, \tag{2.16}$$

one can express

$$\lim_{\Lambda\to\infty} g(z,\Lambda) = 0_{(1)} \otimes g_{(23)}(z). \tag{2.17}$$

Hence $G(z,\Lambda)$ has a stable limit, given by

$$\lim_{\Lambda\to\infty} G(z,\Lambda) = G^0(z) + \sum_{i,j=2}^{3} G^0(z)|\chi_i\rangle g_{(23)}(z)_{ij}\langle\chi_j|G^0(z). \tag{2.18}$$

One can verify that this Green's function is identical to the Green's function obtained from the Hamiltonian

$$H^{\text{ren}} = H^0 + \sum_{i,j=2}^{3} |\chi_i\rangle\lambda_{ij}\langle\chi_j|, \tag{2.19}$$

which we call the renormalized Hamiltonian. It differs from the original H by the absence of the singular vertex part $|\chi_1\rangle\lambda\langle\chi_1|$.

Now we want to calculate the Green's function from the perturbation theory. Because of the infinities arising from g^0_{11} in each order of the perturbation expansion, we consider the regularized expansion

$$\begin{aligned} G^{\text{pert}}(z,\Lambda) &= G^0(z) + \sum_{i=1}^{3} G^0(z)|\chi_i(\Lambda)\rangle\lambda\langle\chi_i(\Lambda)|G^0(z) \\ &\quad + \sum_{i,j=1}^{3} G^0(z)|\chi_i(\Lambda)\rangle\lambda\langle\chi_i(\Lambda)| \\ &\quad \times G^0(z)|\chi_j(\Lambda)\rangle\lambda\langle\chi_j(\Lambda)|G^0(z) + \cdots \\ &= G^0(z) + \sum_{i,j=1}^{3} G^0(z)|\chi_i(\Lambda)\rangle\lambda[1 + \lambda g^0(z,\Lambda) \\ &\quad + (\lambda g^0(z,\Lambda))^2 + \cdots]_{ij}\langle\chi_j(\Lambda)|G^0(z). \end{aligned} \tag{2.20}$$

Renormalization means here to sum up the infinite parts thus defining a new propagator and a new interaction that are both finite and independent of Λ in the limit $\Lambda\to\infty$. We split g^0 in a singular and a regular part

$$g^0(z,\Lambda) = g^{0s}(z,\Lambda) + g^{0r}(z,\Lambda), \tag{2.21}$$

$$g^{0s}(z,\Lambda) = \begin{pmatrix} g^0_{11}(z,\Lambda) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.22}$$

We define as a new propagator $(1 - \lambda g^{0s}(z,\Lambda))^{-1}$, obtained from summing up the singular part g^{0s}

$$(1 - \lambda g^{0s}(z,\Lambda))^{-1} = 1 + \lambda g^{0s}(z,\Lambda) + (\lambda g^{0s}(z,\Lambda))^2 + \cdots. \tag{2.23}$$

We define as a new interaction the regular part $g^{0r}(z,\Lambda)$. We claim that both the propagator and the interaction have a finite limit when Λ tends to infinity. From the definition of $g^{0s}(z,\Lambda)$ one obtains

$$\lim_{\Lambda\to\infty}(1 - \lambda g^{0s}(z,\Lambda))^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = 0_{(1)} \otimes 1_{(23)}, \tag{2.24}$$

which is also independent of λ. The limit of g^{0r} is simply

$$\lim_{\Lambda\to\infty} g^{0r}(z,\Lambda) = \begin{pmatrix} 0 & g^0_{12}(z) & g^0_{13}(z) \\ g^0_{21}(z) & g^0_{22}(z) & g^0_{23}(z) \\ g^0_{31}(z) & g^0_{32}(z) & g^0_{33}(z) \end{pmatrix}. \quad (2.25)$$

The perturbation series in the square brackets of Eq. (2.20) can be rewritten in terms of the new propagator and new interaction:

$$\begin{aligned} &1 + \lambda g^0(z,\Lambda) + (\lambda g^0(z,\Lambda))^2 + \cdots \\ &\quad = (1 - \lambda g^{0s}(z,\Lambda))^{-1} + (1 - \lambda g^{0s}(z,\Lambda))^{-1} \\ &\qquad \times \lambda g^{0r}(z,\Lambda)(1 - \lambda g^{0s}(z,\Lambda))^{-1} \\ &\qquad + (1 - \lambda g^{0s}(z,\Lambda))^{-1}\lambda g^{0r}(z,\Lambda)(1 - \lambda g^{0s}(z,\Lambda))^{-1} \\ &\qquad \times \lambda g^{0r}(z,\Lambda)(1 - \lambda g^{0s}(z,\Lambda))^{-1} + \cdots . \end{aligned} \quad (2.26)$$

We call the rhs of Eq. (2.26) the renormalized perturbation series, which corresponds to summing the skeleton graphs in field theory. We claim that taking the limit $\Lambda\to\infty$ and summing the renormalized series one obtains the same Green's function as given by Eq. (2.18).

One calculates

$$\begin{aligned} &\lim_{\Lambda\to\infty} (1 - \lambda g^{0s}(z,\Lambda))^{-1}\lambda g^{0r}(z,\Lambda)(1 - \lambda g^{0s}(z,\Lambda))^{-1} \\ &\quad = 0_{(1)} \otimes \lambda g^0_{(23)}(z). \end{aligned} \quad (2.27)$$

Thus the rhs of Eq. (2.26) can be expressed in the limit $\Lambda\to\infty$ as

$$\begin{aligned} &0_{(1)} \otimes 1_{(23)} + 0_{(1)} \otimes \lambda g^0_{(23)}(z) + (0_{(1)} \otimes \lambda g^0_{(23)}(z))^2 + \cdots \\ &\quad = 0_{(1)} \otimes (1_{(23)} - \lambda g^0_{(23)}(z))^{-1}. \end{aligned} \quad (2.28)$$

One should note that the rhs is a meromorphic function in λ of degree 1 over 2. Its Taylor series has a finite radius of convergence and allows an unique analytic continuation beyond except at the two poles.

Substituting this result in Eq. (2.20) and taking the limit $\Lambda\to\infty$ of the regular vertices $\chi_i(\Lambda)$, $i = 2,3$, the result is

$$\begin{aligned} \lim_{\Lambda\to\infty} G^{\text{ren pert}}(z,\Lambda) = G^0(z) + \sum_{i,j=2}^{3} G^0(z)|\chi_i\rangle \\ \times g_{(23)}(z)_{ij}\langle\chi_j|G^0(z), \end{aligned} \quad (2.29)$$

in agreement with the Green's function obtained by the direct analytical solution given by Eq. (2.18).

III. DISCUSSION OF MODEL I

The model is a nonrelativistic model. The interaction has been chosen such that the perturbation expansion of the Green's function gives rise to infinities. The interaction was also chosen to be separable for the reason to give a closed analytical solution of the Green's function for each finite cutoff and in the cutoff limit. One should note, however, that the separability property also appears in some self-interacting field theories; e.g., for the nonlinear Schrödinger model or the ϕ^4 model, the interaction matrix element between two-particle Fock space states is separable, after the total energy-momentum delta function is split off. We start from a Hamiltonian H given by Eqs. (2.1), (2.6), and (2.7). It can be written in the form

$$H = H^0 + H^{\text{int}}_{\text{reg}} + H^{\text{int}}_{\text{sing}}, \quad (3.1)$$

$$H^{\text{int}}_{\text{reg}} = \sum_{i=2}^{3} |\chi_i\rangle\lambda\langle\chi_i|, \quad H^{\text{int}}_{\text{sing}} = |\chi_1\rangle\lambda\langle\chi_1|, \quad (3.2)$$

where the last part of the interaction is called singular, because the matrix element g^0_{11} becomes infinite. One of the main results is

$$\lim_{\Lambda\to\infty} G(z,\Lambda) = G^{\text{ren}}(z), \quad (3.3)$$

where $G^{\text{ren}}(z) = (z - H^{\text{ren}})^{-1}$ and $H^{\text{ren}} = H^0 + H^{\text{int}}_{\text{reg}}$.

Inspection of the derivation shows that this is always true when $\lim_{\Lambda\to\infty} g^0_{ij}(z,\Lambda)$ is finite for $ij \neq 11$, but $\lim_{\Lambda\to\infty} g^0_{11}(z,\Lambda)$ is infinite. The property of g^0_{11} being infinite comes from the large momentum behavior of $\chi_1(\mathbf{q}_1,\mathbf{q}_2)$. The particular choice of χ_1 given by Eq. (2.7) leads to a quadratic divergence of g^0_{11}. A different choice, e.g., $\chi_1(\mathbf{q}_1,\mathbf{q}_2) = 1/(\mathbf{q}_1\mathbf{q}_2)$ would lead to a logarithmic divergence. Hence there is a class of examples for χ_1, which leads to the same G^{ren} and H^{ren}. The class consists of those functions $\chi_1(\mathbf{q}_1,\mathbf{q}_2)$, which are smooth functions of the variables $\mathbf{q}_1,\mathbf{q}_2$ and fall off for large q_1q_2 like $(q_1q_2)^{-1}$ or slower. The class may be even larger. We can formulate this in terms of an invariance property under a transformation group. Let us consider the following class of functions:

$$\begin{aligned} K = \{\chi_1 | \chi_1(\mathbf{q}_1,\mathbf{q}_2) = \chi_1^{\text{supp}}(\mathbf{q}_1,\mathbf{q}_2) \\ + (q_1q_2)^{-\alpha},\ \alpha\in[0,1]\}, \end{aligned} \quad (3.4)$$

where χ_1^{supp} is a continuous function with compact support. Let T denote the one-to-one transformations of $[0,1]$ onto $[0,1]$, which forms a group. Corresponding to the group T we define a group of transformations $\mathcal{T}$ on the class K. Corresponding to each $t\in T$ we define a $\tau\in\mathcal{T}$ by

$$\begin{aligned} &\tau\chi_1(\mathbf{q}_1,\mathbf{q}_2) = \chi_1(\mathbf{q}_1',\mathbf{q}_2'), \quad \hat{q}_1' = \hat{q}_1, \quad q_1' = q_1^{t(\alpha)/\alpha}, \\ &\hat{q}_2' = \hat{q}_2, \quad q_2' = q_2^{t(\alpha)/\alpha}, \end{aligned} \quad (3.5)$$

where α is the falloff exponent of χ_1.

One easily checks that $\chi_1^{\text{supp}}(\mathbf{q}_1',\mathbf{q}_2')$ considered as a function of $\mathbf{q}_1,\mathbf{q}_2$ is continuous and has a compact support, and $(q_1',q_2')^{-\alpha} = (q_1q_2)^{-\beta}$, $\beta = t(\alpha)\in[0,1]$, which shows that $\mathcal{T}$ maps K on K. Here $\mathcal{T}$ is a group that follows from T being a group. Hence we can consider $\mathcal{T}$ as a group of symmetry transformations that leaves G^{ren} and the renormalized Hamiltonian H^{ren} invariant and hence describes the same physics. The group $\mathcal{T}$ should be seen in contrast to the usual renormalization group, which describes transformations between regularized cutoff dependent Hamiltonians, giving the same physics.

Another important feature of this model is its parallel with ϕ^4 theory. It is generally believed[13–19] that the renormalized ϕ^4 theory has in $3+1$ dimensions a unity S matrix, while in one or two space dimensions the S matrix differs from unity. This feature is inherent in our model, too. Let us consider the Hamiltonian

$$H = H^0 + |\chi_1\rangle\lambda\langle\chi_1|, \quad (3.6)$$

where $\chi_1(\mathbf{q}_1,\mathbf{q}_2)$ behaves asymptotically for large q_1q_2 like $(q_1q_2)^{-1}$, but is regular at the origin. It gives in three dimen-

sions a logarithmically divergent matrix element g_{11}^0. Then we have

$$\lim_{\Lambda\to\infty} G(z,\Lambda) = G^0(z). \tag{3.7}$$

Hence the full Green's function tends towards the free Green's function, which leads to a unity S matrix. If, however, we consider the Hamiltonian (3.6) only in one or two dimensions, then g_{11}^0 is no longer divergent, hence

$$\lim_{\Lambda\to\infty} G(z,\Lambda) \neq G^0(z), \tag{3.8}$$

and the S matrix is different from unity. Thus this model might serve to give a better understanding of the mechanism of triviality of ϕ_{3+1}^4 theory.

Finally let us discuss the relation between the analytical solution and the perturbative solution. By construction of the model there are infinities in the perturbation series of the Green's function. The reason is that the interaction Hamiltonian is an ill-defined operator in Hilbert space. Therefore the perturbation expansion, being a polynomial in H^{int}, requires renormalization in order to give finite results. Summing the renormalized perturbation series we find the same Green's function as obtained by the direct analytical solution. It is interesting to note that for the latter no renormalization is needed, but only an intermediate regularization. Thus we find that $(z-H)^{-1}$ behaves more regularly than H or polynomials in H. One expects that also $\exp(iH)$, which is closely related to the S matrix and which can be expressed as a contour integral of the resolvant, behaves more regularly than H itself. One reason behind this is that for $\text{Im}(z)\neq 0$, $(z-H)^{-1}$ and $\exp(iH)$ are bounded operators even if H is unbounded. Hence as a conclusion from our model investigation we suggest for field theories, which cannot be solved analytically, to search for approximate solutions in the form of the resolvent $(z-H_n)^{-1}$ or in the form of the exponent $\exp(iH_n)$ (see also Refs. 8 and 9).

IV. MODEL II

From our investigation of model I we know that the Hamiltonian

$$H = H^0 + \sum_{i=1}^{2} |\chi_i\rangle\lambda\langle\chi_i|,$$

with χ_1 being a singular vertex but χ_2 being a regular vertex, i.e., g_{11}^0 is infinite, but g_{ij}^0 is finite for $ij\neq 11$, has the renormalized Hamiltonian

$$H^{\text{ren}} = H^0 + |\chi_2\rangle\lambda\langle\chi_2|;$$

i.e., the singular vertex drops out of the renormalized Hamiltonian and the Green's function $\lim_{\Lambda\to\infty} G(z,\Lambda)$. From that one might suspect that if both χ_1 and χ_2 are singular vertices then the renormalized Hamiltonian should be H^0 only. Model II shall serve as a warning that is not in general the case. We choose now

$$H^{\text{int}} = \sum_{i=1}^{2} |\chi_i\rangle\lambda\langle\chi_i|. \tag{4.1}$$

We choose the singular vertices identical to the singular vertex of model I given by Eq. (2.7), apart from some overall factors. We take χ_i of the form given by Eq. (2.6). We write

$$|\chi_i\rangle = |\chi_i^s\rangle + |\chi_i^r\rangle, \quad i=1,2, \tag{4.2}$$

and put

$$\chi_i^s(\mathbf{q}_1,\mathbf{q}_2) = c_i/\sqrt{q_1 q_2}, \quad i=1,2. \tag{4.3}$$

From model I we know that $\langle\chi_i^s|G_0|\chi_j^s\rangle$ are infinite matrix elements. The vertices χ_i^r are chosen such that the matrix elements $\langle\chi_i^r|G_0|\chi_j^r\rangle$, $\langle\chi_i^s|G_0|\chi_j^r\rangle$, and $\langle\chi_i^r|G_0|\chi_j^s\rangle$ are finite for all i,j. As in the first model we introduce a cutoff Λ in the vertices. Then the cutoff Green's function $G(z,\Lambda)$ as well as the matrices $g^0(z,\Lambda)$ and $g(z,\Lambda)$ are given by Eqs. (2.10)–(2.12) (g^0 and g are now 2×2 matrices). As in Eq. (2.21) we split g^0 in a singular and a regular part, the singular part now being defined by

$$g_{ij}^{0s}(z,\Lambda) = \langle\chi_i^s(\Lambda)|G^0(z)|\chi_j^s(\Lambda)\rangle. \tag{4.4}$$

Denoting the rhs of Eq. (2.8) by $\sigma(z,\Lambda)$, Eq. (4.4) can be expressed as

$$g_{ij}^{0s}(z,\Lambda) = c_i^* c_j \sigma(z,\Lambda). \tag{4.5}$$

Let us calculate $\lim_{\Lambda\to\infty} g(z,\Lambda)$. Model I yields for this limit zeros in the first row and the first column [Eq. (2.14)]. It occurred for these matrix elements because the denominator $\det(1-\lambda g^0(z,\Lambda))$ was of first order in $\sigma(z,\Lambda)$ but the numerator was only of zeroth order in $\sigma(z,\Lambda)$. Hence one would expect for model II the denominator $\det(1-\lambda g^0(z,\Lambda))$ to be of second order in $\sigma(z,\Lambda)$, the numerator to be of first order in $\sigma(z,\Lambda)$ and hence to obtain zero for the $\lim_{\Lambda\to\infty} g(z,\Lambda)$. However, this is not the case, because in $\det(1-\lambda g^0(z,\Lambda))$ the leading order cancels. One obtains

$$g(z) = \lim_{\Lambda\to\infty} g(z,\Lambda) = \lambda \begin{pmatrix} -|c_2|^2 & c_1^* c_2 \\ c_1 c_2^* & -|c_1|^2 \end{pmatrix} D^{-1},$$

$$D = -|c_1|^2 - |c_2|^2 + \lambda\left(|c_1|^2 g_{22}^{0r}(z) + |c_2|^2 g_{11}^{0r}(z) - c_1^* c_2 g_{21}^{0r}(z) - c_1 c_2^* g_{12}^{0r}(z)\right) \tag{4.6}$$

and

$$\lim_{\Lambda\to\infty} G(z,\Lambda) = G^0(z) + \sum_{i,j=1}^{2} G^0(z)|\chi_i\rangle g_{ij}(z)\langle\chi_j|G^0(z). \tag{4.7}$$

In the case $c_1\neq c_2$ the singular vertices χ_i^s are still present in the limit of the Green's function. Now let us consider in the following the special case $c_1 = c_2$. Then Eqs. (4.6) and (4.7) simplify to give

$$\lim_{\Lambda\to\infty} g(z,\Lambda) = \lambda \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} D^{-1},$$

$$D = -2 + \lambda\left(g_{11}^{0r}(z) + g_{22}^{0r}(z) - g_{12}^{0r}(z) - g_{21}^{0r}(z)\right). \tag{4.8}$$

Defining

$$|\chi^r\rangle = |\chi_1^r\rangle - |\chi_2^r\rangle, \tag{4.9}$$

we can express

$$\lim_{\Lambda\to\infty} G(z,\Lambda) = G^0(z) + G^0(z)|\chi^r\rangle \times \frac{\lambda}{2-\lambda\langle\chi^r|G^0(z)|\chi^r\rangle}\langle\chi^r|G^0(z). \tag{4.10}$$

One can verify that this Green's function is identical to the Green's function obtained from the Hamiltonian

$$H^{\text{ren}} = H^0 + |\chi^r\rangle(\lambda/2)\langle\chi^r|, \tag{4.11}$$

which we call a renormalized Hamiltonian. It differs from the original H in three ways: (i) absence of singular vertices, (ii) shift of the coupling constant

$$\lambda \to \lambda' = \lambda/2,$$

and (iii) addition of cross terms of the regular vertices

$$-|\chi_1^r\rangle\lambda'\langle\chi_2^r| - |\chi_2^r\rangle\lambda'\langle\chi_1^r|.$$

Now let us calculate the Green's function by summing up the renormalized perturbation series. The regularized perturbation expansion is given by Eq. (2.20) (except that we have only 2×2 matrices now). We take the definition of the new propagator and new interaction from model I, but $g^{0s}(z,\Lambda)$ being now given by Eq. (4.4). We calculate the limits

$$\lim_{\Lambda\to\infty}(1-\lambda g^{0s}(z,\Lambda))^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \tag{4.12}$$

which is independent from λ too, and

$$\lim_{\Lambda\to\infty}\lambda g^{0r}(z,\Lambda) = \lambda\begin{pmatrix} g_{11}^{0r}(z) & g_{12}^{0r}(z) \\ g_{21}^{0r}(z) & g_{22}^{0r}(z) \end{pmatrix}. \tag{4.13}$$

We define the renormalized perturbation series by the rhs of Eq. (2.26). Taking the limit $\Lambda\to\infty$ the nth term is

$$\begin{aligned}&\lim_{\Lambda\to\infty}(1-\lambda g^{0s}(z,\Lambda))^{-1}\lambda g^{0r}(z,\Lambda)\\ &\quad\times(1-\lambda g^{0s}(z,\Lambda))^{-1}\cdots(1-\lambda g^{0s}(z,\Lambda))^{-1}\\ &\quad\times\lambda g^{0r}(z,\Lambda)(1-g^{0s}(z,\Lambda))^{-1}\\ &\quad=\frac{1}{2}\left[\frac{\lambda}{2}\gamma^{0r}(z)\right]^n\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},\end{aligned} \tag{4.14}$$

with

$$\gamma^{0r}(z) = g_{11}^{0r}(z) + g_{22}^{0r}(z) - g_{12}^{0r}(z) - g_{21}^{0r}(z). \tag{4.15}$$

Summing it up yields

$$\frac{1}{2-\lambda\gamma^{0r}(z)}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

After multiplying this expression with λ, the expression agrees with $\lim_{\Lambda\to\infty} g(z,\Lambda)$ given by Eq. (4.8). Hence substituting it in Eq. (2.20), one obtains also in this model

$$\lim_{\Lambda\to\infty} G^{\text{ren pert}}(z,\Lambda) = \lim_{\Lambda\to\infty} G(z,\Lambda). \tag{4.16}$$

V. S MATRIX

We have obtained for model I, and model II in the case $c_1 = c_2$, Green's functions $\lim_{\Lambda\to\infty} G(z,\Lambda)$ and Hamiltonians H^{ren}, where all singular vertices are absent [Eqs. (2.18), (2.19), (4.10), and (4.11)]. So far we have considered only a complex energy $z = E + i\epsilon$, $\epsilon \neq 0$. In order to obtain the physical transition amplitude, one has to perform the limit $\epsilon \to +0$. Also in the coefficient matrix g only regular vertices appear [Eqs. (2.16) and (4.8)]. Hence the limit $g(E+i\epsilon)$ can be performed in a standard way, provided that the vertices also have a regular behavior in the vicinity of E such that $\lim_{\epsilon\to+0} g^0(E+i\epsilon)$ exists, which holds, e.g., for the Yamaguchi form factor. Hence one has a well-defined physical transition amplitude $T(E+i0)$ related to the Green's function in the standard way by

$$G(z) = G^0(z) + G^0(z)T(z)G^0(z). \tag{5.1}$$

Then the physical S matrix is related to $T(E+i0)$ also in the standard way.

VI. CONCLUSION

In this paper we have studied nonrelativistic models with separable interactions, such that the interaction generates infinities in the perturbation series of the Green's function. We have compared the direct analytical solution with the perturbative solution. Both methods led to the same Green's function. However, in the perturbative approach renormalization was needed, while in the direct analytical calculation only an intermediate regularization was necessary. We have chosen the interaction to contain vertices that produce infinities in the perturbation expansion in order to model infinities that appear in field theory. In the first model we have one singular vertex giving rise to one infinite matrix element g_{11}^0. This vertex drops out of the cutoff limit of the Green's function and the renormalized Hamiltonian. There is a class of Hamiltonians differing in the singular vertex but leading to the same renormalized Hamiltonian, which can be expressed in terms of a symmetry under a group of transformations. In the second model we have two singular vertices. This model shows that in general the renormalized Hamiltonian is not simply obtained by subtracting the singular vertex part from the original Hamiltonian, as was the case in the first model.

Although we have studied nonrelativistic models, we have a pattern of infinities similar to a field theory. Particularly model I resembles ϕ^4 theory, which, after renormalization, is a free theory in $3+1$ dimensions, but is an interacting theory in one or two space dimensions. Choosing a suitable vertex yields the same feature in our model and hence may serve for a better understanding of ϕ^4 theory.

From the study of both models we find that the resolvent $(z-H)^{-1}$ behaves more regularly than H itself. One expects that also $\exp(iH)$ behaves more regularly than H because first the time evolution $\exp(iHt)$ is closely related to the S matrix, which was found to be well defined, and second $\exp(iHt)$ can be expressed as a contour integral of the resolvent $(z-H)^{-1}$. Hence we suggest to look for approximate solutions of not analytically solvable field theories in the form of $(z-H_n)^{-1}$ or $\exp(iH_n t)$, where H_n is some approximate Hamiltonian.

One might speculate if the results of this model study, namely no need for renormalization if the Green's function can be calculated directly, have parallels in field theories. In general for those an analytical solution is not known, but nonperturbative approximation methods are available.[4–9] Numerical investigations based on the method of Refs. 8 and 9 are underway.

ACKNOWLEDGMENT

One of the authors (H.K.) is grateful for support from the Natural Sciences and Engineering Research Council of Canada.

[1]Phys. Today **38** (8), 17 (1985).
[2]J. Zinn-Justin, Phys. Rep. **70**, 109 (1981).
[3]K. G. Wilson, Phys. Rev. D **10**, 2445 (1974).
[4]For a review see, e.g., "Non-perturbative aspects in quantum field theory," proceedings of Les Houches Winter Advanced Study Institute, 1978, edited by E. Brézin and J. L. Gervais, in Phys. Rep. **49**, 93 (1979).
[5]C. S. Hsue, H. Kümmel, and P. Ueberholz, Phys. Rev. D **32**, 1435 (1985).
[6]H. C. Pauli and S. J. Brodsky, Phys. Rev. D **32**, 1993 (1985).
[7]A. C. Davis and W. Nahm, Phys. Lett. B **159**, 294 (1985).
[8]H. Kröger, A. Smailagic, and R. Girard, Phys. Rev. D **32**, 3221 (1985).
[9]R. Girard and H. Kröger, Phys. Lett. B **164**, 117 (1985).
[10]K. G. Wilson, Phys. Rev. D **6**, 419 (1972).
[11]Y. Yamaguchi, Phys. Rev. **95**, 1628 (1954).
[12]For a review see A. N. Mitra, Adv. Nucl. Phys. **3**, 1 (1969).
[13]K. G. Wilson and J. Kogut, Phys. Rep. C **12**, 75 (1974).
[14]G. A. Baker and J. M. Kincaid, Phys. Rev. Lett. **42**, 1431 (1979).
[15]M. Aizenmann, Phys. Rev. Lett. **47**, 1 (1981); Commun. Math. Phys. **86**, 1 (1982).
[16]J. Fröhlich, Nucl. Phys. B **200**, 281 (1982).
[17]C. Aaragao de Carvalho, C. S. Carragiolo, and J. Fröhlich, Nucl. Phys. B **215**, 209 (1983).
[18]B. Freedman, P. Smolensky, and D. Weingarten, Phys. Lett. B **113**, 481 (1982).
[19]I. G. Halliday, Rep. Prog. Phys. **47**, 987 (1984).

Low-energy scattering for medium-range potentials

Roger G. Newton
Physics Department, Indiana University, Bloomington, Indiana 47405

(Received 15 May 1986; accepted for publication 16 July 1986)

The low-energy behavior of the transmission coefficient in one dimension and of the phase shifts in two and three dimensions is studied for the Schrödinger equation with central potentials that have finite absolute moments of order between 1 and 2. Resulting modifications of Levinson's theorem are also derived.

I. INTRODUCTION

The standard assumption on which most of the known properties of scattering amplitudes for the Schrödinger equation in one, two, or three dimensions with a central potential have been proved[1-3] is that $V\in L^1_2(\mathbb{R})$ or $V(|x|)\in L^1_2(\mathbb{R}_+)$, respectively, where

$$L^1_\sigma(\mathbb{R}) = \left\{V(x) \;\middle|\; \int_{\mathbb{R}} dx|V(x)|(1+|x|^\sigma) < \infty\right\}$$

and analogously for $\mathbb{R}_+$.[4] In this paper we will investigate what properties of the scattering amplitude will be changed if $V\notin L^1_2$. We will particularly study the cases when $V\in L^1_1$ and remark only parenthetically on those instances when $V\notin L^1_1$.

The technique used will generally be based on specific assumptions such as $V\in L^1_\sigma$, $1\leqslant\sigma\leqslant 2$. The modifications that arise in one dimension when $\sigma<2$ have to do primarily with the way in which the transmission coefficient can approach zero as the wave number $k\to 0$ (see Sec. II). In two and three dimensions, with central potentials, we examine the behavior of the phase shifts as $k\to 0$, and the effect on the differential scattering length or cross section (see Sec. III for two dimensions and Sec. IV for three dimensions). In each case we also determine the changes in Levinsen's theorem that arise from modifications in the small-k behavior of the transmission amplitude or the Jost function. Their analyticity and large-k behavior are, of course, unchanged by our weakened assumptions, since those require only that $V\in L^1$.

For ease of reading all detailed proofs are given in three appendices.

II. ONE DIMENSION

We begin by considering solutions of the equation

$$f'' = V(x)f, \quad x,V\in\mathbb{R}. \tag{2.1}$$

If $V\in L^1_2$ one easily proves these well-known facts: (a) continuous solutions f_1 and f_2 exist that satisfy the boundary conditions

$$\lim_{x\to+\infty} f_1(x) = \lim_{x\to-\infty} f_2(x) = 1,$$
$$\lim_{x\to+\infty} f_1'(x) = \lim_{x\to-\infty} f_2'(x) = 0; \tag{2.2}$$

(b) in general f_1 grows linearly as $x\to-\infty$, and so does f_2 as $x\to\infty$; (c) for an exceptional set of potentials, the functions f_1 and f_2 are linearly dependent and hence, uniformly bounded; (d) for no potential in L^1_2 is there a solution of (2.1) that tends to zero as $x\to\infty$ or as $x\to-\infty$.

If $V\notin L^1_1$ the situation is radically different, as can be seen explicitly when $V\sim cx^{-2}$ as $x\to\infty$. Solutions of (1) may grow more rapidly than linearly, and for some V there are solutions that tend to zero; in fact, there may be bound states; solutions that satisfy (2.2) generally do not exist. The question is, what happens if $V\in L^1_1$ but $V\notin L^1_2$?

The answer is given by the following lemma.

Lemma 2.1: If $V\in L^1_\sigma$, $1\leqslant\sigma<2$, then (2.1) has unique continuous solutions that satisfy the boundary conditions (2.2). We have

$$f_1(x) = \begin{cases} 1+o(x^{1-\sigma}), & \text{as } x\to+\infty, \\ -\gamma x + O(x^{2-\sigma}), & \text{as } x\to-\infty, \end{cases}$$
$$f_2(x) = \begin{cases} \gamma x + O(x^{2-\sigma}), & \text{as } x\to+\infty, \\ 1+o(x^{1-\sigma}), & \text{as } x\to-\infty, \end{cases}$$

where

$$\gamma \doteqdot f_1f_2' - f_1'f_2 = \int_{-\infty}^{\infty} dx\, Vf_1 = \int_{-\infty}^{\infty} dx\, Vf_2.$$

In the exceptional case when $\gamma=0$, we have $f_1(x) = af_2(x) = a + o(x^{1-\sigma})$ as $x\to-\infty$, $a\neq 0$. There also exists a solution $g(x)$ that is linearly independent of $f_1(x)$ and which is such that as $x\to\infty$, $g(x) = x + O(x^{2-\sigma})$. (For a proof see Appendix A.)

Corollary: If $V\in L^1_1$ there is no solution of (2.1) that tends to zero as $x\to+\infty$ or as $x\to-\infty$. Hence (2.1) has no L^2-solutions.

Thus the situation for $V\in L^1_\sigma$, $1\leqslant\sigma<2$, is very similar to that when $V\in L^1_2$, except for the size of the error term. An explicit example is given by $V(x) = \theta(x-1)\frac{1}{4}(x^{-3} - 3x^{-5/2})$ [where $\theta(x)$ is the Heaviside function], for which

$$f_1(x) = \begin{cases} \exp(-x^{-1/2}), & x>1, \\ (x+1)/2e, & x<1, \end{cases}$$
$$f_2(x) = \begin{cases} 1, & x<1, \\ [e - \int_1^x dy \exp(2y^{-1/2})/2e]\exp(-x^{-1/2}), & x>1. \end{cases}$$

As $x\to\infty$, $f_2 = -x/2e - 3x^{1/2}/2e + O(1)$.

The solutions f_1 and f_2 of (2.1) that satisfy the boundary conditions (2.2) are the unique solutions of the Volterra equations (see Appendix A)

$$f_1(x) = 1 - \int_x^\infty dy(x-y)V(y)f_1(y) \tag{2.3a}$$

and

 0022-2488/86/112720-11$02.50

$$f_2(x) = 1 + \int_{-\infty}^{x} dy(x-y)V(y)f_2(y). \tag{2.3b}$$

If $V \in L^1_{\mu\eta} \doteq L^1_\mu(\mathbb{R}_-) \cap L^1_\eta(\mathbb{R}_+)$, where $\mu,\eta \geqslant 1$, then Lemma 2.1 holds with σ replaced by μ for the limits as $x \to -\infty$ and by η for the limits as $x \to +\infty$.

We now consider solutions $f_1(k,x)$ and $f_2(k,x)$ of the Schrödinger equation

$$f'' + k^2 f = Vf \tag{2.4}$$

defined by the integral equations

$$f_1(k,x) = e^{ikx} - \int_x^{\infty} dy\, k^{-1} \sin[k(x-y)]V(y)f_1(k,y), \tag{2.5a}$$

$$f_2(k,x) = e^{-ikx} + \int_{-\infty}^{x} dy\, k^{-1} \times \sin[k(x-y)]V(y)f_2(k,y). \tag{2.5b}$$

When $k \neq 0$, Im $k \geqslant 0$, these Volterra equations can be solved by iteration if $V \in L^1$, and they lead to the well-known analyticity and continuity properties[1] of the Jost solutions f_1 and f_2. We are now interested in their behavior near $k = 0$. If $V \in L^1_2$ it is well known that as $k \to 0$, $f_1(k,x) \to f_1(x)$, $f_2(k,x) \to f_2(x)$, where $f_1(x)$ and $f_2(x)$ are the solutions of (2.3a) and (2.3b), respectively, and the remainders are linear in k. More generally we have

Lemma 2.2: If $V \in L^1_\sigma$, $1 \leqslant \sigma \leqslant 2$, then

$$e^{-ikx}f_1(k,x) = f_1(x) + g_1(k,x),$$
$$e^{ikx}f_2(k,x) = f_2(x) + g_2(k,x),$$

where

$$|g_1(k,x)| \leqslant \begin{cases} C\nu(k), & x \geqslant 0, \\ C\left[\nu(k) - \dfrac{2|k|x}{1-2|k|x}\right](1-x), & x < 0, \end{cases}$$

$$|g_2(k,x)| \leqslant \begin{cases} C\nu(k), & x \leqslant 0, \\ C\left[\nu(k) + \dfrac{2|k|x}{1+2|k|x}\right](1+x), & x > 0. \end{cases}$$

Here $\nu(k)$ is bounded, independent of x, and $o(|k|^{\sigma-1})$ as $|k| \to 0$ if $\sigma < 2$; if $\sigma = 2$, then $\nu = O(k)$. (For a proof, see Appendix A.)

This tells us how $f_1(k,x)$ and $f_2(k,x)$ approach $f_1(x)$ and $f_2(x)$, respectively, as $|k| \to 0$. Now, the transmission and reflection amplitudes are given by[1]

$$T(k) = \frac{2ik}{2ik - I_1(k)},$$

$$R_l(k) = \frac{I_2(k)}{2ik - I_1(k)}, \quad R_r(k) = \frac{I_3(k)}{2ik - I_1(k)},$$

where

$$I_1(k) = \int_{-\infty}^{\infty} dx\, V(x)e^{-ikx}f_1(k,x) = \int_{-\infty}^{\infty} dx\, V(x)e^{ikx}f_2(k,x),$$

$$I_2(k) = \int_{-\infty}^{\infty} dx\, V(x)e^{ikx}f_1(k,x),$$

$$I_3(k) = \int_{-\infty}^{\infty} dx\, V(x)e^{-ikx}f_2(k,x).$$

The behavior of these integrals is as follows.

Lemma 2.3: If $V \in L^1_\sigma$, $1 \leqslant \sigma < 2$, then for $j = 1,2,3$ as $k \to 0$

$$I_j(k) = \gamma + o(|k|^{\sigma-1}),$$

where γ is the constant defined in Lemma 2.1. If $\sigma = 2$ the remainder is $O(k)$. (For a proof see Appendix A.)

Note that this lemma does not imply that there exists a $\mu > \sigma - 1$ such that $I_j - \gamma$ goes exactly like k^μ, i.e., that

$$\lim_{k \to 0} k^{-\mu}[I_j(k) - \gamma]$$

exists and differs from zero.

For $\sigma = 2$ one can show that $I_j(k) = \gamma + ic_j k + o(k)$ as $k \to 0$, where c_j is real. It then follows that T, R_l, and R_r are real and continuous at $k = 0$ even if $\gamma = 0$. When $\sigma < 2$ they may be complex and hence discontinuous as $k \to 0 \pm$. [Since[5] $T(-k) = \overline{T}(k)$, T will not be continuous unless it is real in the limit as $k \to 0$; similarly for R_l and R_r.]

As a result of Lemma 2.3 we have the following theorem.

Theorem 1: If $V \in L^1_1$ and $\gamma \neq 0$ (where γ is defined in Lemma 2.1), then as $k \to 0$,

$$T(k) = -2ik/\gamma + o(k)$$

and $R_l(0) = R_r(0) = -1$. In the exceptional case when $\gamma = 0$,

if $V \in L^1_\sigma$, $1 \leqslant \sigma < 2$, then $1/T(k) = o(|k|^{\sigma-2})$,

while for $\sigma = 2$,

$$1/T(k) = O(1)$$

and

$$T(k) = O(1).$$

Remarks: (1) If $V \in L^1_{\mu\eta}$ [defined below Eq. (2.3b)] where $\mu,\eta \geqslant 1$, then Lemma 2.2 holds with $\sigma = \eta$ for g_1 and $\sigma = \mu$ for g_2. Lemma 2.3 and Theorem 1 hold with $\sigma = \min(\mu,\eta)$.

(2) If $V(x) = V_1(x) + \lambda_1 \delta(x - x_0)$, then all the results hold, provided that V_1 satisfies the hypotheses of the lemmas and the theorem.

The breakdown of these results when $V \notin L^1_1$ may be explicitly seen for the potential

$$V(x) = a\theta(x)(1+x)^{-2} + b\theta(-x)(1-x)^{-2},$$

for which the Schrödinger equation is solvable in terms of Hankel functions of orders $\rho = (a + \frac{1}{4})^{1/2}$ and $\sigma = (b + \frac{1}{4})^{1/2}$. One then finds that as $k \to 0$, $T(k) = \text{const}\, k^{\rho+\sigma}$. If $a < -\frac{1}{4}$ and $b < -\frac{1}{4}$, then ρ and σ become $\rho = i|\rho|$ and $\sigma = i|\sigma|$, and

$$k^{\rho+\sigma} = \exp[i(|\rho| + |\sigma|)\log k\,],$$

which has no limit as $k \to 0$. If a or b are large enough, T vanishes faster than k, which is impossible if $V \in L^1_1$.

Theorem 1 directly leads to the Levinson theorem. If $V \in L^1$, T is known to be the boundary value of an analytic

function meromorphic in $\mathbb{C}^+$, with simple poles at $k=i\kappa$ if κ^{-2} is an eigenvalue (bound state), and such that $\lim_{|k|\to\infty} T(k)=1$. If T has n simple poles in $\mathbb{C}^+$ and tends to zero exactly like k^α as $k\to 0$ then by the "argument principle" its phase δ, continuously defined by

$$T=|T|e^{i\delta},$$

satisfies the relation

$$\delta(0)-\delta(\infty)=\pi n-\tfrac{1}{2}\pi\alpha.$$

Since

$$\det\begin{pmatrix} T & R_r \\ R_l & T\end{pmatrix}=\det S=\frac{T}{\overline{T}}=e^{2i\delta}$$

and each pole of T in $\mathbb{C}^+$ signifies a bound state, the Levinson theorem for $V\in L^1_1$ has to be modified in the exceptional case. According to Theorem 1 we have the following.

Theorem 2 (Levinson theorem): If $V\in L^1_1(\mathbb{R})$, generically

$$\delta(0)-\delta(\infty)=\pi(n-\tfrac{1}{2}),$$

where n is the number of bound states. In the exceptional case of $\gamma=0$ (see Lemma 2.1) if T goes exactly like k^β near $k=0$ (which is consistent with Theorem 1 if $V\in L^1_\sigma$, $1\leqslant\sigma<2$, for $\beta<2-\sigma\leqslant 1$, but not implied by it), then

$$\delta(0)-\delta(\infty)=\pi(n-\tfrac{1}{2}\beta).$$

III. TWO DIMENSIONS

We now consider the Schrödinger equation in two dimensions with a central potential. The equation is separable and the radial equations are

$$f''+[(\tfrac{1}{4}-\lambda^2)/r^2]f+k^2f=Vf,\quad \lambda=0,1,2,\ldots. \tag{3.1}$$

A "regular solution," defined by the boundary condition

$$\lim_{r\to 0} r^{-(1/2)-\lambda}\phi_\lambda(k,r)=1$$

satisfies the Volterra equation

$$\phi_\lambda(k,r)=\phi_{\lambda 0}(k,r)-\int_0^r dr'\,g_\lambda(k,r,r')V(r')\phi_\lambda(k,r'), \tag{3.2}$$

where

$$\phi_{\lambda 0}(k,r)=r^{1/2}J_\lambda(kr)(k/2)^{-\lambda}\lambda!, \tag{3.3}$$

$$g_\lambda(k,r,r')=\tfrac{1}{2}\pi(rr')^{1/2}[J_\lambda(kr)Y_\lambda(kr')-J_\lambda(kr')Y_\lambda(kr)], \tag{3.4}$$

and J_λ and Y_λ are the usual Bessel and Neumann functions. The Jost solution, defined by

$$\lim_{r\to\infty} e^{-ikr}f_\lambda(k,r)=1$$

is the solution of

$$f_\lambda(k,r)=f_{\lambda 0}(k,r)+\int_r^\infty dr'\,g_\lambda(k,r,r')V(r')f_\lambda(k,r'), \tag{3.5}$$

where, in terms of the usual Hankel function

$$f_{\lambda 0}(k,r)=e^{i(1/2)\pi(\lambda+1/2)}(\pi kr/2)^{1/2}H^{(1)}_\lambda(kr). \tag{3.6}$$

We define the Jost function by

$$\mathcal{F}_\lambda(k)=-(\pi^{1/2}/2\lambda!)(k/2)^{\lambda-1/2}e^{i(1/2)\pi((1/2)-\lambda)}\times W(\phi_\lambda,f_\lambda), \tag{3.7}$$

where $W(\phi,f)\doteq\phi f'-\phi' f$. From (3.2) and (3.5) we then find that

$$\begin{aligned}\mathcal{F}_\lambda(k)&=1-i\pi\frac{(k/2)^\lambda}{2\lambda!}\int_0^\infty dr\,r^{1/2}V(r)\\&\quad\times H^{(1)}_\lambda(kr)\phi_\lambda(k,r)\\&=1+e^{i(1/2)\pi((1/2)-\lambda)}\left(\frac{\pi}{2k}\right)^{1/2}\int_0^\infty dr\,r^{1/2}\\&\quad\times V(r)J_\lambda(kr)f_\lambda(k,r).\end{aligned} \tag{3.8}$$

Here $\mathcal{F}_\lambda$ is defined so that $\lim_{k\to\infty}\mathcal{F}_\lambda(k)=1$ and

$$\phi_\lambda=\lambda!\pi^{-1/2}(k/2)^{(1/2)-\lambda}e^{i(1/2)\pi(\lambda+1/2)}\times(\mathcal{F}_\lambda\bar{f}_\lambda-e^{i\pi((1/2)-\lambda)}\bar{\mathcal{F}}_\lambda f_\lambda),$$

which implies that the S matrix is given by

$$S_\lambda=\bar{\mathcal{F}}_\lambda/\mathcal{F}_\lambda=e^{2i\delta_\lambda}. \tag{3.9}$$

Use of the first form of (3.8) then leads to the representation

$$S_\lambda(k)=1+i\pi\int_0^\infty dr\,r^{1/2}V(r)J_\lambda(kr)\phi_\lambda(k,r)\frac{(k/2)^\lambda}{\lambda!\mathcal{F}_\lambda(k)}. \tag{3.10}$$

The scattering amplitude in two dimensions is given by[6]

$$f(\theta)=\left(\frac{1}{2\pi ik}\right)^{1/2}\sum_{\lambda=0}^\infty\epsilon_\lambda(\cos\lambda\theta)(S_\lambda-1) \tag{3.11}$$

and the scattering length

$$L=\int_0^{2\pi}d\theta\,|f(\theta)|^2=\left(\frac{4}{k}\right)\sum_{\lambda=0}^\infty\epsilon_\lambda\sin^2\delta_\lambda, \tag{3.12}$$

where $\epsilon_\lambda=2$, $\lambda=1,2,\ldots$, $\epsilon_0=1$.

(a) The case of $\lambda=0$.

Lemma 3.1: If $V\in L^1_\sigma(\mathbb{R}_+)$, $\sigma>1$, the equation

$$f''+(1/4r^2)f=Vf \tag{3.13}$$

has a unique solution that satisfies the boundary condition

$$\lim_{r\to 0}\phi_0(r)r^{-1/2}=1.$$

This solution is continuous and generally grows as $r^{1/2}\log r$ when $r\to\infty$; in the exceptional case in which

$$\Gamma\doteq\int_0^\infty dr\,r^{1/2}V(r)\phi_0(r)=0 \tag{3.14}$$

it grows as $r^{1/2}$. There is a linearly independent solution $g_0(r)$, which is such that

$$\lim_{r\to 0} r^{-1/2}g_0(r)/\log r=1.$$

(For a proof see Appendix B. It will be noted from the proof that the weaker hypothesis

$$\int_0^\infty dr\,r|V|(1+|\log r|)^2<\infty \tag{3.15}$$

suffices.)

Corollary: If $V\in L^1_\sigma(\mathbb{R}_+)$, $\sigma>1$, (3.13) has no solution in $L^2(\mathbb{R}_+)$.

For $k \neq 0$, the solutions ϕ_0 and f_0 and the Jost function $\mathcal{F}_0$ are defined by (3.2), (3.5), and (3.7) setting $\lambda = 0$.

Lemma 3.2: If $V \in L^1_\sigma(\mathbb{R}_+)$, $1 < \sigma < 3$, then

$$\mathcal{F}_0(k) = A \log k + B + E(k), \tag{3.16}$$

where

$$|E(k)| \leqslant Ck^{\alpha - 1}, \tag{3.16'}$$

for all $\alpha \in (1, \sigma)$, and $A = 0$ if and only if $\Gamma = 0$ [as defined in (3.14)], i.e., we have the exceptional case. (For a proof see Appendix B.)

The constants A and B are given by

$$A = -\frac{2}{\pi} \int_0^\infty dr\, r^{1/2} V(r) g(r), \tag{3.17}$$

$$B = 1 + i \int_0^\infty dr\, r^{1/2} V(r) h(r), \tag{3.18}$$

where g and h are solutions of (3.13) defined by the equations

$$g(r) = r^{1/2} + \int_r^\infty dr' (rr')^{1/2} \log\left(\frac{r'}{r}\right) V(r') g(r'), \tag{3.19}$$

$$h(r) = [1 + iW_0(0)] g(r) + iH(r), \tag{3.20}$$

$$H(r) = r^{1/2} \log r + \int_r^\infty dr' (rr')^{1/2} \log\left(\frac{r'}{r}\right) V(r') H(r'). \tag{3.21}$$

The function $W_\lambda(x)$ is defined by the Bessel and Neumann functions[7]

$$W_\lambda(x) = Y_\lambda(x) - (2/\pi) \log x J_\lambda(x). \tag{3.22}$$

Lemma 3.2 allows us to conclude by (3.9) how S_0 behaves near $k = 0$. If $A \neq 0$, then

$$S_0 = \frac{A \log k + \overline{B} + \overline{E}(k)}{A \log k + B + E(k)} = 1 + \frac{\overline{B} - B}{A \log k} + O\left(\frac{1}{(\log k)^2}\right)$$

and since $\operatorname{Re} h = g$ implies $\operatorname{Im} B = -(\pi/2)A$, we have

$$S_0(k) = 1 + \frac{i\pi}{\log k} + O\left(\frac{1}{(\log k)^2}\right).$$

On the other hand if $A = 0$, then B is real and cannot vanish. This is because $A = 0$ implies that

$$B = 1 - \int_0^\infty dr\, r^{1/2} V(r) H(r)$$

and hence as $r \to 0$

$$H(r) = Br^{1/2} \log r + o(r^{1/2} \log r),$$

$$g(r) = o(r^{1/2} \log r).$$

Thus if $B = 0$, $H(r) = cg(r)$ because both are solutions of (3.13). But as $r \to \infty$, $g(r)$ goes as $r^{1/2}$ and $H(r)$ goes as $r^{1/2} \log r$, so B cannot vanish. Therefore in the exceptional case

$$S_0 = 1 + O(k^{\alpha - 1}),$$

and we have the following theorem for the phase shift defined in (3.9).

Theorem 3: If $V \in L^1_\sigma(\mathbb{R}_+)$, $1 < \sigma < 3$, in the generic case when $\Gamma \neq 0$ [see (3.14)], as $k \to 0$,

$$\delta_0(k) = (\pi/2)/\log k + o(1/\log k) \pmod{\pi},$$

but when $\Gamma = 0$,

$$\delta_0(k) = O(k^{\alpha - 1}) \pmod{\pi},$$

for all $\alpha < \sigma$.

Thus in all cases δ_0 tends to zero (mod π) at the origin. However, the approach to zero is faster in the exceptional case. In fact, if $\sigma > 2$ one may expect (though this is not implied by Theorem 3) that the derivative of δ_0 at $k = 0$ vanishes in the exceptional case, and tends to $-\infty$ generically.

By (3.11) the partial-wave amplitude for $\lambda = 0$ in two dimensions is equal to

$$a_0 = (e^{2i\delta_0} - 1)/(2i\pi k)^{1/2}.$$

Thus, by Theorem 3, in the generic case

$$\lim_{k \to 0} a_0(k) k^{1/2} \log k = (i\pi/2)^{1/2},$$

while in the exceptional case

$$a_0 = O(k^{\alpha - 3/2}), \quad \alpha < \sigma.$$

It also follows from Lemma 3.2 that the form of Levinson's theorem does not depend on whether the case is generic or exceptional.[8]

Theorem 4 (Levinson theorem): In all cases for $V \in L^1_\sigma(\mathbb{R}_+)$, $\sigma > 1$,

$$\delta_0(0) - \delta_0(n) = \pi n_0,$$

where n_0 is the number of bound states for $\lambda = 0$.

This follows directly from (3.16), the "argument principle," and the fact that for C a semicircle of radius ϵ in the upper half-plane,

$$\int_C d \log \mathcal{F}_0(k) = i \int_0^\pi d\phi \frac{1}{\log \epsilon} \underset{\epsilon \to 0}{\to} 0.$$

(b) The cases of $\lambda \geqslant 1$.

Lemma 3.3: If $V \in L^1_1(\mathbb{R}_+)$ the equation

$$f'' + [(\tfrac{1}{4} - \lambda^2)/r^2] f = Vf \tag{3.23}$$

has a unique solution that satisfies the boundary condition

$$\lim_{r \to 0} \phi_\lambda(r) r^{-1/2 - \lambda} = 1.$$

This solution is continuous and bounded by $Cr^{1/2 + \lambda}$. In the exceptional case when

$$\Gamma_\lambda \doteqdot 1 + \frac{1}{2\lambda} \int_0^\infty dr\, r^{(1/2) - \lambda} V(r) \phi_\lambda(r) = 0, \tag{3.24}$$

ϕ_λ goes as $r^{(1/2) - \lambda}$ when $r \to \infty$ rather than as $r^{(1/2) + \lambda}$. (For a proof, see Appendix B.)

Corollary: If $\lambda = 1$, there are no L^2-solutions of (3.23). If $\lambda > 1$, $\Gamma_\lambda = 0$ implies that $\phi_\lambda \in L^2$ (a bound state of zero energy).

The behavior of $S_l(k)$ near $k = 0$ is obtainable from (3.10) and Lemma 3.3. We have the following theorem.

Theorem 5: If $\Gamma_\lambda = \mathcal{F}_\lambda(0) \neq 0$ and $V \in L^1_\sigma(\mathbb{R}_+)$, $1 \leqslant \sigma < 1 + 2\lambda$, $\lambda \geqslant 1$, then as $k \to 0$,

$$\delta_\lambda(k) = o(k^{\sigma - 1}) \pmod{\pi},$$

and for $\sigma = 1 + 2\lambda$,

$$\delta_\lambda(k) = O(k^{2\lambda}).$$

(For a proof see Appendix B.)

We also have the following lemma.

Lemma 3.4: If $V \in L^1_\sigma(\mathbb{R}_+)$, $1 \leqslant \sigma < 3$, then

$$\mathscr{F}_\lambda(k) = \mathscr{F}_\lambda(0) + o(k^{\sigma-1}), \quad \lambda \geqslant 1.$$

(For a proof, see Appendix B.)

This lemma implies that in the exceptional case the phase shift need not approach a multiple of π as $k \to 0$. In fact, we have, by the same argument as that given at the end of Sec. II.

Theorem 6 (Levinson theorem): If $V \in L^1_\sigma(\mathbb{R}_+)$ and $\mathscr{F}_\lambda(0) \neq 0, \lambda \geqslant 1$,

$$\delta_\lambda(0) - \delta_\lambda(\infty) = \pi n_\lambda,$$

where n_λ is the number of bound states. If $\mathscr{F}_\lambda(0) = 0$ and $\mathscr{F}_\lambda(k) = Ck^\alpha + o(k^\alpha), \alpha > \sigma - 1$, as $k \to 0$ (which is compatible with Lemma 3.4 but not implied by it), then

$$\delta_\lambda(0) - \delta_\lambda(\infty) = \pi(n_\lambda + \tfrac{1}{2}\alpha).$$

Theorem 5 implies that if $V \in L^1_\sigma$, $1 \leqslant \sigma < 3$, then all partial-wave scattering amplitudes for $\lambda \geqslant 1$ generically are $o(k^{\sigma-3/2})$ as $k \to 0$. Thus, by Theorem 3 they vanish relative to that for $\lambda = 0$ if $\sigma > 1$, and they are bounded if $\sigma \geqslant \frac{3}{2}$. If $\mathscr{F}_0(0) = 0$, then $a_0 = O(k^{\alpha-3/2} \log k)$, so that we have the following theorem.

Theorem 7: If $V \in L^1_\sigma(\mathbb{R}_+)$, $\sigma > 1$, then in the generic case, i.e., if $\mathscr{F}_\lambda(0) \neq 0$ for all integers λ, the differential scattering length in two dimensions becomes independent of the scattering angle as $k \to 0$ and

$$\lim_{k \to 0} L(k) k (\log k)^2 = \pi^2.$$

If $\mathscr{F}_0(0) = 0$, then

$$L(k) = O[k^{2\alpha-3}(\log k)^2],$$

for all $\alpha < \sigma < 3$.

The universal value of the limit of $L(k)\ k(\log k)^2$ is noteworthy.[2] Also note that if $\sigma > \frac{3}{2}$, then in the exceptional case for $\lambda = 0$, L is bounded as $k \to 0$.

IV. THREE DIMENSIONS

We next consider the Schrödinger equation in three dimensions with a central potential. It is separable, and the radial equation is the same as (3.1), but with $\lambda = l + \frac{1}{2}$, $l = 0,1,2,\ldots$. Subscripts in this section will refer to l rather than λ.

(a) The case $l=0$. If $f_0(k,r)$ is the Jost solution defined by the boundary condition

$$\lim_{r \to 0} e^{ikr} f_0(k,r) = 1$$

and $\phi_0(k,r)$ is the regular solution defined by

$$\phi_0(k,0) = 0, \quad \phi_0'(k,0) = 1,$$

then the Jost function $\mathscr{F}_0$ is given by

$$\mathscr{F}_0 = W(f_0,\phi_0) \doteq f_0\phi_0' - f_0'\phi_0. \tag{4.1}$$

It is well known (but since there does not appear to be a detailed published proof, a proof is given in Appendix C) that if $V \in L^1_2(\mathbb{R}_+)$ and if $\mathscr{F}_0(0) = 0$, then near $k = 0$,

$$\mathscr{F}_0(k) = -iak + o(k), \quad a \neq 0. \tag{4.2}$$

If $V \notin L^1_2$, this need no longer hold and we have the following lemma.

Lemma 4.1: If $V \in L^1_\sigma(\mathbb{R}_+)$, $1 \leqslant \sigma < 2$, then as $k \to 0$,

$$\mathscr{F}_0(k) = \mathscr{F}_0(0) + o(k^{\sigma-1}),$$

while for $\sigma = 2$,

$$\mathscr{F}_0(k) = \mathscr{F}_0(0) + O(k).$$

(For a proof see Appendix C.)

The S matrix is given by

$$S_0 = e^{2i\delta_0} = \mathscr{F}_0(-k)/\mathscr{F}_0(k) \tag{4.3}$$

and hence we have by Lemma 4.1 the following theorem.

Theorem 8: If $V \in L^1_\sigma(\mathbb{R}_+)$, $1 \leqslant \sigma < 2$, then generically, i.e., if $\mathscr{F}_0(0) \neq 0$, in three dimensions

$$\delta_0(k) = o(k^{\sigma-1}) \pmod{\pi}$$

as $k \to 0$, and if $\sigma = 2$

$$\delta_0(k) = O(k) \pmod{\pi}.$$

If $\mathscr{F}_0(0) = 0$ and $\mathscr{F}_0(k) = ak^\alpha + o(k^\alpha), a \neq 0, \alpha > \sigma - 1$ (which is consistent with but not implied by Lemma 4.1), then

$$\delta_0(0) = \tfrac{1}{2}\pi\alpha \pmod{\pi}.$$

The implications for the partial-wave amplitude are easy to see. Since the partial-wave amplitude for $l = 0$ is given by

$$a_0 = \frac{1}{2ik}(e^{2i\delta_0} - 1),$$

we have generically [i.e., when $\mathscr{F}_0(0) \neq 0$] for $V \in L^1_\sigma$, $1 \leqslant \sigma < 2$

$$a_0(k) = o(k^{\sigma-2}), \tag{4.4}$$

while for $\sigma = 2$

$$a_0(k) = O(1). \tag{4.4'}$$

Therefore there is assurance of a finite cross section at $k = 0$ only for $\sigma = 2$.

Lemma 4.1 and Eq. (4.3) also immediately lead to Levinson's theorem by the same argument as at the end of Sec. II.

Theorem 9 (Levinson theorem): If $V \in L^1_\sigma(\mathbb{R}_+)$, $1 \leqslant \sigma$, then in the generic case, i.e., if $\mathscr{F}_0(0) = 0$,

$$\delta_0(0) - \delta_0(\infty) \neq \pi n_0,$$

where n_0 is the number of bound states of $l = 0$. If $\mathscr{F}_0(0) = 0$ and $\mathscr{F}_0(k) = ak^\alpha + o(k^\alpha)$, $a \neq 0$, for $\alpha > \sigma - 1$ (which is consistent with but not implied by Lemma 4.1), then

$$\delta_0(0) - \delta_0(\infty) = \pi(n_0 + \tfrac{1}{2}\alpha).$$

(b) The case $l \geqslant 1$. All the formulas and representations (3.2) to (3.10) now hold, with $\lambda = l + \frac{1}{2}$, $l = 1,2,\ldots$. [Of course, $\lambda!$ is to be replaced by $\Gamma(\lambda + 1)$.] We then obtain the following lemma.

Lemma 4.2: The same as Lemma 3.3, with $\lambda = l + \frac{1}{2}$.

Corollary: If

$$\Gamma_l \doteq 1 + \int_0^\infty dr\, r^{-l} V(r) \phi_l(r) = 0,$$

then for $l \geqslant 1$, $\phi_l \in L^2$ and there is a bound state of zero energy.

We also have the analog of Lemma 3.4, with the same proof. (The logarithmic terms in the estimates used are now absent, of course.)

Lemma 4.3: If $V\in L^1_\sigma(\mathbb{R}_+)$, $1\leqslant\sigma<3$, then

$$\mathscr{F}_l(k)=\mathscr{F}_l(0)+o(k^{\sigma-1}).$$

Also, from (3.10) and Lemma 4.2 we derive the following theorem.

Theorem 10: If $\mathscr{F}_l(0)\neq 0$, $l\geqslant 1$, and $V\in L^1_\sigma(\mathbb{R}_+)$, $1\leqslant\sigma<2l+2$, then as $k\to 0$,

$$\delta_l(k)=o(k^{\sigma-1})\ (\mathrm{mod}\ \pi)$$

and for $\sigma=2l+2$

$$\delta_l(k)=O(k^{2l+1})\ (\mathrm{mod}\ \pi).$$

If $\mathscr{F}_l(0)=0$ and $\mathscr{F}_l(k)=ak^\alpha+o(k^\alpha)$, $a\neq 0$, $\alpha>\sigma-1$ (which is consistent with Lemma 4.2 but not implied by it), then

$$\delta_l(0)=\tfrac{1}{2}\pi\alpha\quad(\mathrm{mod}\ \pi).$$

The Levinson theorem has the same form as in the two-dimensional case for $\lambda\geqslant 1$.

Theorem 11 (Levinson theorem): The same as Theorem 6 with $\lambda=l+\frac{1}{2}$.

Since in the exceptional case (which by the corollary to Lemma 4.2 for $l\geqslant 1$ is a zero energy bound state) the phase shift need not approach an integal multiple of π, the cross section may grow to infinity as k^{-2}. By Theorems 8 and 10 in the generic case if $\sigma<2$, all the partial waves may vanish equally as $k\to 0$ and hence the differential cross section may be both unbounded and angle-dependent.

Theorem 12: If $V\in L^1_\sigma(\mathbb{R}_+)$, $1\leqslant\sigma<2$, then in the generic case, i.e., if $\mathscr{F}_0(0)\neq 0$ and there is no zero-energy bound state, the differential cross section may tend to infinity as $o(k^{2\sigma-4})$ and it may be angle dependent. If there is a zero-energy bound state of angular momentum $l\geqslant 1$ [or $\mathscr{F}_0(0)=0$], then the cross section may grow as k^{-2}, with the angle dependence of the square of the Legendre polynomial of order l.

We note that if $\sigma\geqslant 2$, the zero-energy cross section is isotropic and finite, except if $\mathscr{F}_0(0)=0$.

ACKNOWLEDGMENT

This work was supported in part by a grant from the National Science Foundation.

APPENDIX A: PROOFS FROM SEC. II

We shall use C for an arbitrary constant that need not have the same value everywhere.

Proof of Lemma 2.1: If $V\in L^1_1$, the Volterra equations (2.3a) and (2.3b) can be solved by iteration. That gives uniqueness, continuity, and for $x>0$,

$$|f_1(x)|\leqslant C\exp\left[2\int_x^\infty dy\,y|V(y)|\right]\leqslant C.$$

For $x<0$, therefore,

$$\left|\int_0^\infty dy(x-y)V(y)f_1(y)\right|\leqslant C(1+|x|),$$

and hence by (2.3a)

$$|f_1(x)|=\left|1-\int_x^0-\int_0^\infty\right|$$

$$\leqslant C(1+|x|)+2|x|\int_x^0 dy|V|\,|f_1|,$$

$$\frac{|f_1(x)|}{1+|x|}\leqslant C+2\int_x^0 dy(1+|y|)|V|\frac{|f_1(y)|}{1+|y|},$$

and by iteration $|f_1(x)|/(1+|x|)\leqslant C$. So

$$|f_1(x)|\leqslant\begin{cases}C(1+|x|), & x\leqslant 0,\\ C, & x\geqslant 0.\end{cases}$$

We now use this result in (2.3a) for $x<0$:

$$f_1(x)=1-x\int_{-\infty}^\infty dy\,Vf_1+x\int_{-\infty}^x dy\,Vf_1+\int_x^\infty dy\,yVf_1,$$

$$\left|\int_0^\infty dy\,yVf_1\right|\leqslant C,$$

$$\left|\int_x^0 dy\,yVf_1\right|\leqslant C|x|(1+|x|^{1-\sigma})\int_x^0 dy(1+|y|)^\sigma|V|$$

$$\leqslant C(1+|x|^{2-\sigma}),$$

$$\left|x\int_{-\infty}^x dy\,Vf_1\right|\leqslant C(1+|x|^{2-\sigma})\int_{-\infty}^x dy(1+|y|)^\sigma|V|.$$

For $x>0$ by (2.3a) and the boundedness of f_1,

$$|f_1(x)-1|\leqslant C\int_x^\infty dy|V(y)|y^\sigma x^{1-\sigma}=o(x^{1-\sigma})$$

as $x\to+\infty$. Similarly for f_2. If $\gamma=0$, f_1 and f_2 must be multiples of one another since their Wronskian vanishes.

Let $g(x)$ be a solution of (2.1) such that $W(g,f_1)=gf_1'-g'f_1=1$. Then define $h=g/f_1$ so that $h'=1/f_1^2$. Therefore $\lim_{x\to\infty}h'(x)=1$ and hence

$$h(x)=\int_b^x dy(f_1^{-2}-1)+x$$

$$=x+\int_b^x dy\,o(y^{1-\sigma})=x+O(x^{2-\sigma}).\qquad\square$$

The remark below (2.3b) follows from the above proof. Before proving Lemma 2.2 we define

$$h(k,x)\doteq\int_x^\infty dy\,\alpha(k(y-x))(x-y)V(y)f_1(y),\quad(\mathrm{A1})$$

where

$$\alpha(x)=1-[(\sin x)/x]e^{ix}.$$

Lemma A.1: If $V\in L^1_\sigma$, $1\leqslant\sigma\leqslant 2$, then

$$|h(k,x)|\leqslant\begin{cases}C\nu(k), & \text{for } x\geqslant 0,\\ C\left[\nu(k)+\dfrac{2k|x|}{1+2k|x|}\right](1+|x|), & \text{for } x<0,\end{cases}$$

where $\nu(k)$ is bounded, independent of x, $\nu(k)=o(k^{\sigma-1})$ as $k\to 0$ if $\sigma<2$, and $\nu(k)=O(k)$ if $\sigma=2$.

Proof: We use

$$|\alpha(x)|\leqslant C[|x|/(1+|x|)].$$

Now, for $x\geqslant 0$ by Lemma 2.1,

$$|h|\leqslant C\int_x^\infty dy\,\frac{k(x-y)^2}{1+k(y-x)}\,|V(y)|\leqslant C\int_0^\infty dy\,\frac{ky^2}{1+ky}\,|V|$$
$$\leqslant Ck^{\sigma-1}\int_0^\infty dy\,y^\sigma|V|\left(\frac{ky}{1+ky}\right)^{2-\sigma}\doteq\nu(k).$$

One easily proves that the integral tends to zero as $k\to 0$ if $\sigma<2$, by splitting it into

$$\int_0^\infty=\int_0^a+\int_a^\infty,$$

where $a=k^{-1/2}$.

For $x<0$,

$$\left|\int_0^{|x|}\cdots\right|\leqslant C\int_0^{|x|}dy|V|\,\frac{k(|x|+|y|)^2}{1+k(|x|+|y|)}\leqslant C\frac{2k|x|^2}{1+2k|x|},$$
$$\left|\int_{|x|}^\infty\cdots\right|\leqslant Ck^{\sigma-1}\int_{|x|}^\infty dy|V|y^\sigma\left(\frac{ky}{1+ky}\right)^{2-\sigma}\leqslant\nu(k),$$
$$\left|\int_x^0\cdots\right|\leqslant C\int_x^0 dy|V|\,\frac{2k|x|^2}{1+2k|x|}(1+|y|)$$
$$\leqslant C\frac{2k|x|^2}{1+2k|x|}.\qquad\square$$

Next we consider the Volterra equation

$$g(k,x)=h(k,x)-\int_x^\infty dy\,k^{-1}\sin k(x-y)\times e^{ik(y-x)}V(y)g(k,y),\qquad(A2)$$

where h is the function defined by (A1).

Lemma A.2: If $V\in L^1_\sigma$, $1\leqslant\sigma\leqslant 2$, then (A2) has a unique solution g and this solution satisfies the same inequalities as stated in Lemma A.1 for h.

Proof: Equation (A2) is solvable by iteration, using

$$|\sin x|\leqslant C\,[|x|/(1+|x|)].$$

From $x\geqslant 0$ this gives directly $|g(k,x)|\leqslant C\nu(k)$. For $x<0$ use this result in

$$\left|\int_0^\infty dy\,k^{-1}\sin k(x-y)e^{ik(y-x)}V(y)g(k,y)\right|$$
$$\leqslant C\nu(k)\int_0^\infty dy|V|(|x|+y)$$
$$\leqslant C\nu(k)(1+|x|).$$

Therefore by (A2) and Lemma A1 for $x<0$,

$$|g|\leqslant\Sigma(k,x)+\left|\int_x^0 dy\,k^{-1}\sin k(y-x)e^{ik(y-x)}V(y)g(k,y)\right|,$$

where

$$\Sigma(k,x)=C\left[\nu(k)+\frac{2k|x|}{1+2k|x|}\right](1+|x|).$$

Since $\Sigma(k,x)/(1+|x|)$ increases with x,

$$\left|\frac{g(k,x)}{\Sigma(k,x)}\right|$$
$$\leqslant 1+C\int_x^0 dy(1+|y|)|V|\left|\frac{g(k,y)}{\Sigma(k,y)}\right|\leqslant C$$

by iteration. Thus $|g(k,x)|\leqslant C\Sigma(k,x)$. $\square$

We now consider the integral equation (2.5a), which can be solved by iteration if $V\in L^1$ for $k\neq 0$, Im $k\geqslant 0$. Multiplication by e^{-ikx} and subtraction of (2.1) shows that the function

$$g(k,x)=f_1(k,x)e^{-ikx}-f_1(x)\qquad(A3)$$

satisfies (A2). Thus we obtain Lemma 2.2 from Lemma A2 for f_1, and similar for f_2. $\square$

Proof of Lemma 2.3: We have, with γ defined in Lemma 2.1 and g by (A3),

$$I_1(k)=\gamma+\int_{-\infty}^\infty dx\,V(x)g(k,x).$$

Now use Lemma (A2) for g, and

$$\frac{2k|x|}{1+2k|x|}\leqslant 2\left(\frac{2k|x|}{1+2k|x|}\right)^{2-\sigma}k^{\sigma-1}|x|^{\sigma-1}.$$

Furthermore, by the same argument used in the proof of Lemma A1,

$$\int_{-\infty}^\infty dx|x|^\sigma|V|\left(\frac{2k|x|}{1+2k|x|}\right)^{2-\sigma}=o(1)$$

as $k\to 0$ for $\sigma<2$; for $\sigma=2$, of course, it is $O(1)$.

For I_2 we write

$$I_2-\gamma=\int_{-\infty}^\infty dx\,V(x)[e^{ikx}f_1(k,x)-f_1(x)]$$
$$=\int_{-\infty}^\infty dx\,V(x)g(k,x)e^{2ikx}+2i\int_{-\infty}^\infty dx\,V(x)f(x)\sin kx\,e^{ikx},$$

use Lemma A2, Lemma 2.1, and

$$|\sin kx|\leqslant C\frac{k|x|}{1+k|x|}\leqslant C(k|x|)^{\sigma-1}\left(\frac{k|x|}{1+k|x|}\right)^{2-\sigma}.$$

Similar arguments are used for I_3. $\square$

APPENDIX B: PROOFS FROM SEC. III

Proof of Lemma 3.1: The solution ϕ_0 of (3.13) satisfies the Volterra equation

$$\phi_0(r)=r^{1/2}+\int_0^r dr'(rr')^{1/2}\log\left(\frac{r}{r'}\right)V(r')\phi_0(r').$$

Using the inequality

$$|\log(r/r')|\leqslant(1+|\log r|)(1+|\log r'|),$$

we iterate

$$\frac{|\phi_0(r)|}{r^{1/2}(1+|\log r|)}$$
$$\leqslant 1+\int_0^r dr'\,r'(1+|\log r'|)^2|V(r')|\frac{|\phi_0(r')|}{r'^{1/2}(1+|\log r'|)}$$
$$\leqslant\exp\left[\int_0^r dr'\,r'(1+|\log r'|)^2|V(r')|\right]\leqslant C.$$

Hence

$$|\phi_0(r)|\leqslant Cr^{1/2}(1+|\log r|).$$

It follows from the Volterra equation that

$$\phi_0(r) = r^{1/2}\log r\int_0^\infty dr'\, r'^{1/2}V(r')\phi_0(r')$$
$$- r^{1/2}\log r\int_r^\infty dr'\, r'^{1/2}V\phi_0$$
$$+ r^{1/2}\left[1 - \int_0^r dr'\, r'^{1/2}\log r' V(r')\phi_0(r')\right].$$

But

$$\left|\int_r^\infty dr'\, r'^{1/2}V\phi_0\right|$$
$$\leqslant C\int_r^\infty dr'\, r'|V|\frac{(1+|\log r'|)^2}{1+|\log r|} = o(1/|\log r|)$$

and hence

$$\phi_0(r) = \Gamma r^{1/2}\log r + r^{1/2}$$
$$\times\left(1 - \int_0^\infty dr'\, r'^{1/2}\log r' V\phi_0\right) + o(r^{1/2}),$$

where Γ is defined by (3.14). Note that we needed only

$$\int_0^\infty dr\, r|V|(1+|\log r|)^2 < \infty. \tag{B1}$$

It also follows that as $r\to 0$, $\phi_0(r) = r^{1/2} + o(r^{3/2}(\log r)^2)$, if $V\in L^1$.

Let g_0 be such that $W(\phi_0,g_0) = \phi_0 g_0' - \phi_0' g_0 = 1$, and define $h(r) = g_0/\phi_0$. Then

$$h' = 1/\phi_0^2 = r^{-1} + o(\log r)^2, \quad r\to 0.$$

Therefore $h = \log r + O(r(\log r)^2)$ and $g_0 = r^{1/2}\log r + O(r^{3/2}(\log r)^2)$ as $r\to 0$. □

Proof of Lemma 3.2: Since both $J_0(x)$ and $Y_0(x)$ decrease like $x^{-1/2}$ as $x\to\infty$, while $J_0\to 1$ as $x\to 0$ and Y_0 goes as $\log x$,

$$|W_0(x)| \leqslant C\frac{1+|\log x|}{1+|\log x| + x^{1/2}}, \tag{B2}$$

$$|J_0(x)| \leqslant \frac{C}{(1+x)^{1/2}}, \tag{B3}$$

where W_0 is defined by (3.22). Replacing Y_0 in (3.4) by (3.22) therefore leads to

$$|g_0(k,r,r')| \leqslant CZ(k,r)Z(k,r'),$$

where

$$Z(k,r) = \frac{r^{1/2}(1+|\log r|)(1+|\log kr|)}{1+|\log kr| + (kr)^{1/2}}.$$

Thus from (3.5)

$$|f_0(k,r)| \leqslant Ck^{1/2}Z(k,r) + CZ(k,r)$$
$$\times\int_r^\infty dr'\, Z(k,r')|V(r')|\,|f_0(k,r')|$$

and by iteration

$$|f_0(k,r)| \leqslant Ck^{1/2}Z(k,r)$$
$$\times\exp\left[C\int_r^\infty dr'\, Z^2(k,r')|V(r')|\right].$$

Since

$$Z(k,r) \leqslant r^{1/2}(1+|\log r|),$$

it follows that if $V\in L^1_\sigma$, $\sigma > 1$, then

$$|f_0(k,r)| \leqslant CZ(k,r)k^{1/2}. \tag{B4}$$

Near $r = 0$ the function $f_0(k,r)$ goes as

$$f_0(k,r) = ar^{1/2} + br^{1/2}\log r + o(r^{1/2})$$

and we find from (3.5)

$$b = (\pi k/2)^{1/2}e^{3i\pi/4}$$
$$- \int_0^\infty dr\, r^{1/2}J_0(kr)V(r)f_0(k,r)$$

and by (3.8), $b = (\pi k/2)^{1/2}e^{3i\pi/4}\mathcal{F}_0$. One then readily finds from (B3) and (B4) that

$$\lim_{k\to\infty}\mathcal{F}_0(k) = 1.$$

We next need more detailed estimates of J_0 and W_0:

$$|J_0(x) - 1| \leqslant C[x/(1+x)]^2 \tag{B5}$$

and

$$|W_0(x) - W_0(0)| \leqslant C[x/(1+x)]^2, \tag{B6}$$

which follow from their boundedness and approach to zero as x^2. Therefore, from (3.6) and (3.22)

$$f_{00}(k,r) = k^{1/2}\log k\, e^{3i\pi/4}(2r/\pi)^{1/2}$$
$$+ k^{1/2}e^{i\pi/4}(\pi r/2)^{1/2}$$
$$\times[1 + (2i/\pi)\log r + iW_0(0)] + R_0, \tag{B7}$$

$$|R_0| \leqslant C\frac{(kr)^{5/2}}{(1+kr)^2}(1+|\log kr|)$$
$$\leqslant C(kr)^{\alpha-(1/2)}, \quad 1<\alpha<3. \tag{B8}$$

Furthermore, from (3.4), (3.22), (B5), and (B6),

$$|g_0(k,r,r') - (rr')^{1/2}\log(r'/r)|$$
$$\leqslant C(rr')^{1/2}(1+|\log r|)$$
$$\times(1+|\log kr'|)[kr'/(1+kr')]^2$$
$$\leqslant C(rr')^{1/2}(1+|\log r|)(kr')^{\alpha-1},$$
$$1<\alpha<3, \quad \text{for } r'\geqslant r. \tag{B9}$$

We now insert (B7) in the integral equation (3.5) and obtain

$$f_0(k,r) = k^{1/2}\log k(2/\pi)^{1/2}e^{3i\pi/4}g(r)$$
$$+ k^{1/2}(\pi/2)^{1/2}e^{i\pi/4}h(r) + R(k,r), \tag{B10}$$

where $g(r)$ and $h(r)$ satisfy (3.19)–(3.21) and

$$R(k,r) = R_0(k,r) + \int_r^\infty dr'\left[g_0(k,r,r')\right.$$
$$\left. - (rr')^{1/2}\log\left(\frac{r'}{r}\right)\right]V(r')f_0(k,r')$$
$$+ \int_r^\infty dr'(rr')^{1/2}\log\left(\frac{r'}{r}\right)V(r')R(k,r').$$

The Volterra equations (3.19) and (3.21) have unique solutions and determine g and h.

By (B4) and (B9) we have

$$\left|\int_r^\infty dr'\left[g_0(k,r,r') - (rr')^{1/2}\log\left(\frac{r'}{r}\right)\right]V(r')f_0(k,r')\right|$$
$$\leqslant Ck^{\alpha-1/2}r^{1/2}(1+|\log r|)$$
$$\times\int_r^\infty dr'\, r'^{a}|V(r')|(1+|\log r'|)$$
$$\leqslant Ck^{\alpha-1/2}(1+r^{\alpha-1})r^{1/2}$$

if we choose $\alpha<\sigma$. As a result the integral equation for R yields by iteration

$$|R|\leqslant Ck^{\alpha-1/2}(1+r^{\alpha-1})r^{1/2},$$

for all $\alpha\in(1,\sigma)$. Insertion of (B10) in the second form of (3.8) then gives (3.15) with (3.17) and (3.18), and (3.16).

It follows from (3.19) that as $r\to\infty$, $g=r^{1/2}+o(r^{1/2})$, and as $r\to 0$, $g=-(\pi/2)Ar^{1/2}\log r+O(r^{1/2})$. Therefore $A=0$ is the exceptional case (3.14). □

Note that $\mathscr{F}_0(ke^{i\pi})=\bar{\mathscr{F}}_0(k)$ for real k because $\operatorname{Re} h=g$.

Proof of Lemma 3.3: The "regular solution" $\phi_\lambda(r)$ satisfies the Volterra equation

$$\phi_\lambda(r)=r^{(1/2)+\lambda}-\int_0^r dr'\, g_\lambda(r,r')V(r')\phi_\lambda(r') \quad \text{(B11)}$$

with

$$g_\lambda(r,r')=(1/2\lambda)(rr')^{1/2}[(r'/r)^\lambda-(r/r')^\lambda]. \quad \text{(B12)}$$

Using

$$|g_\lambda(r,r')|\leqslant Cr^{(1/2)+\lambda}r'^{(1/2)-\lambda}$$

for $r'\leqslant r$, we iterate and obtain

$$|\phi_\lambda(r)|\leqslant Cr^{(1/2)+\lambda}\exp\left[C\int_0^r dr'\, r'|V(r')|\right]\leqslant Cr^{(1/2)+\lambda} \quad \text{(B13)}$$

if $V\in L^1_1$. Thus ϕ_λ is continuous and bounded by $Cr^{(1/2)+\lambda}$.

The integral equation (B11) now shows that as $r\to\infty$,

$$\phi_\lambda(r)=r^{(1/2)+\lambda}\Gamma_\lambda+o(r^{(1/2)+\lambda}).$$

Define a solution $g_\lambda(r)$ of (3.23) by the equation

$$g_\lambda(r)=r^{(1/2)-\lambda}-\int_r^\infty dr'\, g_\lambda(r,r')V(r')g_\lambda(r'),$$

which can be solved by iteration. One easily finds that

$$W(\phi_\lambda,g_\lambda)=\phi_\lambda g_\lambda'-\phi_\lambda' g_\lambda=-2\lambda\Gamma_\lambda.$$

Hence, if $\Gamma_\lambda=0$ then ϕ_λ is a multiple of g_λ. Furthermore by its boundary condition ϕ_λ cannot vanish identically. Therefore if $\Gamma_\lambda=0$, as $r\to\infty$,

$$\phi_\lambda(r)=cr^{1/2-\lambda}+o(r^{1/2-\lambda}),$$

where $c\neq 0$. □

Proof of Theorem 5: For $\lambda\geqslant 1$ the function W_λ defined by (3.22) behaves as follows[7] as $x\to 0$:

$$W_\lambda(x)=-(1/\pi)(\tfrac{1}{2}x)^{-\lambda}(\lambda-1)!+O(x^{2-\lambda}).$$

One readily finds that

$$|J_\lambda(x)|\leqslant C\left(\frac{x}{1+x}\right)^\lambda(1+x)^{-1/2}, \quad \text{(B14)}$$

$$|Y_\lambda(x)|\leqslant C\left(\frac{x}{1+x}\right)^{-\lambda}(1+x)^{-1/2}, \quad \text{(B15)}$$

$$|W_\lambda(x)|\leqslant C\left(\frac{x}{1+x}\right)^{-\lambda}\frac{1+|\log x|}{1+|\log x|+x^{1/2}}, \quad \text{(B16)}$$

$$\left|J_\lambda(x)-\frac{(\frac{1}{2}x)^\lambda}{\lambda!}\right|\leqslant C\left(\frac{x}{1+x}\right)^{2+\lambda}(1+x)^{\lambda-1/2}, \quad \text{(B14')}$$

$$\left|W_\lambda(x)+\frac{1}{\pi}\left(\frac{1}{2}x\right)^{-\lambda}(\lambda-1)!\right|$$
$$\leqslant C\left(\frac{x}{1+x}\right)^{2-\lambda}\frac{1+|\log x|}{1+|\log x|+x^{1/2}}. \quad \text{(B16')}$$

As a result, for $r'\leqslant r$,

$$|g_\lambda(k,r,r')\leqslant Ck^{-1}\left(\frac{kr}{1+kr}\right)^{1/2+\lambda}\left(\frac{kr'}{1+kr'}\right)^{1/2-\lambda}, \quad \text{(B17)}$$

and eliminating Y_λ by (3.22) from (3.4),

$$|g_\lambda(k,r,r')-g_\lambda(r,r')|\leqslant C\left(\frac{kr}{1+kr}\right)^2\left(\frac{r}{r'}\right)^\lambda(rr')^{1/2}, \quad \text{(B17')}$$

where $g_\lambda(r,r')$ is given by (B12).

The "regular solution" of (3.1) satisfies the Volterra equation (3.2). Using the above inequalities we find that the equation is solvable by iteration and

$$|\phi_\lambda(k,r)|\leqslant Ck^{-1/2-\lambda}[kr/(1+kr)]^{1/2+\lambda}, \quad \text{(B18)}$$

if $V\in L^1_1$. We use this bound in the first form of (3.8) and then let $k\to 0$, so that $\phi_\lambda(k,r)\to\phi_\lambda(r)$. Comparison with (3.24) shows that $\mathscr{F}_\lambda(0)=\Gamma_\lambda$. Thus the exceptional case is the one for which $\mathscr{F}_\lambda(0)=0$.

Next we use the inequalities (B14) and (B18) in the representation (3.10) if $\Gamma_\lambda\neq 0$:

$$|S_\lambda(k)-1|\leqslant Ck^{-1}\int_0^\infty dr|V(r)|\left(\frac{kr}{1+kr}\right)^{1+2\lambda}.$$

Since

$$\left(\frac{kr}{1+kr}\right)^{1+2\lambda}\leqslant(kr)^\sigma\left(\frac{kr}{1+kr}\right)^{1+2\lambda-\sigma},$$

if $1\leqslant\sigma\leqslant 1+2\lambda$, it follows that

$$|S_\lambda-1|\leqslant Ck^{\sigma-1},$$

if $V\in L^1_\sigma$. Since one easily proves, by splitting the integral

$$\int_0^\infty=\int_0^a+\int_a^\infty,$$

where $a=k^{-1/2}$, that

$$\int_0^\infty dr\, r^\sigma|V|\left(\frac{kr}{1+kr}\right)^\epsilon=o(1),$$

for $\epsilon>0$, we find for $\sigma<1+2\lambda$

$$S_\lambda(k)=1+o(k^{\sigma-1}),$$

for $1\leqslant\sigma<1+2\lambda$, and for $\sigma=1+2\lambda$

$$S_\lambda(k)=1+O(k^{2\lambda}).$$

The theorem follows by the definition (3.9). □

Proof of Lemma 3.4: By the inequalities (B17') and (B13) we have

$$\left|\int_0^r dr'\, V(r')\,[g_\lambda(k,r,r') - g_\lambda(r,r')]\phi_\lambda(r')\right|$$
$$\leqslant C\int_0^r dr'\,|V(r')|r'r^{1/2+\lambda}\left(\frac{kr}{1+kr}\right)^2$$
$$\leqslant Cr^{1/2+\lambda}\left(\frac{kr}{1+kr}\right)^2,$$

if $V\in L^1_1$. Now subtract (B11) from (3.2) and use the above inequality, (B14'), and (B17):

$$|\phi_\lambda(k,r) - \phi_\lambda(r)|$$
$$\leqslant Cr^{1/2+\lambda}\left(\frac{kr}{1+kr}\right)^2 + C\int_0^r dr'\, r^{1/2+\lambda}\, r'^{(1/2)-\lambda}$$
$$\times|V(r')|\,|\phi_\lambda(k,r') - \phi_\lambda(r')|$$
$$\leqslant Cr^{1/2+\lambda}\left(\frac{kr}{1+kr}\right)^2 \qquad \text{(B19)}$$

by iteration if $V\in L^1_1$.

Next we subtract (3.24) from the first form of (3.8):

$$\mathscr{F}_\lambda(k) - \Gamma_\lambda = \frac{i\pi}{2\lambda!}\int_0^\infty dr\, r^{1/2}V(r)\left[\left(\frac{1}{2}k\right)^\lambda H_\lambda^{(1)}(kr)\right.$$
$$\left. + \frac{i}{\pi}(\lambda-1)!r^{-\lambda}\right]\phi_\lambda(k,r)$$
$$- \frac{1}{2\lambda}\int_0^\infty dr\, r^{(1/2)-\lambda}V(r)$$
$$\times[\phi_\lambda(k,r) - \phi_\lambda(r)].$$

By (3.22), the fact that $H_\lambda^{(1)} = J_\lambda + iY_l$, (3.14) and (3.16'),

$$\left|\left(\frac{1}{2}k\right)^\lambda H_\lambda^{(1)}(kr) + \frac{i}{\pi}(\lambda-1)!r^{-\lambda}\right|$$
$$\leqslant Ck^\lambda\left[\left(\frac{kr}{1+kr}\right)^{2-\lambda} + \left(\frac{kr}{1+kr}\right)^\lambda(1+|\log kr|)\right],$$

for $1\leqslant\sigma\leqslant 3$ and $\sigma < 2\lambda + 1$. Therefore by (B18) and (B19)

$$|\mathscr{F}_\lambda(k) - \Gamma_\lambda|$$
$$\leqslant Ck^{\sigma-1}\int_0^\infty dr\, r^\sigma|V|\left(\frac{kr}{1+kr}\right)^{3-\sigma}(1+|\log kr|).$$

By the usual argument, the integral tends to zero as $k\to 0$ if $V\in L^1_\sigma$. □

APPENDIX C: PROOFS FROM SEC. IV

Let us differentiate (4.1) with respect to k, indicating the derivative by an overdot, and allow $k\to 0$. Since ϕ_0 is an analytic function of k^2 we have $\dot\phi_0(0,r) = 0$ and hence

$$\dot{\mathscr{F}}_0(0) = W(\dot f_0(0,r), \phi_0(0,r)).$$

Suppose now that $\mathscr{F}_0(0) = 0$. Then $f_0(0,r) = c\phi_0(0,r)$, $c\neq 0$, and hence

$$\dot{\mathscr{F}}_0(0) = (1/c)\,W(\dot f_0(0,r), f_0(0,r)).$$

This formula holds, provided that $V\in L^1_2$. Otherwise there is no assurance that $\dot f_0(0,r)$ exists, as can be seen from the integral equation

$$\dot f_0(0,r) = ir - \int_r^\infty dr'(r-r')V(r')\dot f_0(0,r').$$

If $V\in L^1_2$, iterating this equation shows that

$$|\dot f_0(0,r) - ir| \leqslant C\int_r^\infty dr'\, r'^2|V(r')| \underset{r\to\infty}{\to} 0,$$
$$|\dot f_0'(0,r) - i| \leqslant C\int_r^\infty dr'\, r'|V(r')| \underset{r\to\infty}{\to} 0.$$

Since furthermore

$$|f_0(0,r) - 1| \leqslant C\int_r^\infty dr'|V(r')|r' \underset{r\to\infty}{\to} 0,$$
$$|f_0'(0,r)| \leqslant C\int_r^\infty dr'|V(r')| \underset{r\to\infty}{\to} 0,$$

we may evaluate the Wronskian of $\dot f_0(0,r)$ and $f_0(0,r)$ at $r\to\infty$, replacing $\dot f_0$ by ir and f_0 by 1:

$$\dot{\mathscr{F}}_0(0) = -i/c \neq 0.$$

This demonstration of the well-known fact that if $\mathscr{F}_0(0) = 0$, then near $k = 0$, $\dot{\mathscr{F}}_0(k) = -iak + o(k)$, where $a\neq 0$, clearly shows its dependence on the assumption that $V\in L^1_2$.

Proof of Lemma 4.1: Suppose $V\in L^1_1$,

$$f_0(r) = 1 + \int_r^\infty dr'\,(r'-r)V(r')f_0(r').$$

This can be iterated and converges. One easily gets $|f(r)|\leqslant C$ for all r.

Next consider

$$f_0(kr) - f_0(r)$$
$$= e^{ikr} - 1 + \int_r^\infty dr'(r'-r)$$
$$\times\left[\frac{\sin k(r'-r)}{k(r'-r)} - 1\right]V(r')f_0(k,r')$$
$$+ \int_r^\infty dr'(r'-r)V(r')[f_0(k,r') - f_0(r')].$$

Here we use the inequalities

$$|e^{ikr} - 1| \leqslant C\frac{kr}{1+kr},$$
$$\left|\frac{\sin k(r'-r)}{k(r'-r)} - 1\right| \leqslant C\frac{kr'}{1+kr'}, \quad r'\geqslant r,$$

and iterate to conclude that

$$|f_0(k,r) - f_0(r)|$$
$$\leqslant C\left[\frac{kr}{1+kr} + \int_r^\infty dr'\,\frac{kr'^2}{1+kr'}|V(r')|\right].$$

Therefore

$$|f_0(0,r) - f_0(0)| \leqslant C\int_0^\infty dr\,\frac{kr^2}{1+kr}|V(r)|$$
$$\leqslant Ck^{\sigma-1}\int_0^\infty dr|V|r^\sigma\left(\frac{kr}{1+kr}\right)^{2-\sigma},$$

if $V\in L^1_\sigma$, $1\leqslant\sigma<2$. By the usual argument, splitting

$$\int_0^\infty = \int_0^a + \int_a^\infty, \quad a = k^{-1/2},$$

one sees that the integral is $o(1)$ as $k\to 0$. For $\sigma = 2$, one directly obtains $O(k)$. □

[1]L. D. Faddeev, Trudy Mat. Inst. Steklov **73**, 314 (1964) [Am. Math. Soc. Transl. **2**, 65, 139 (1967)]; P. Deift and E. Trubowitz, Commun. Pure Appl. Math. **32**, 121 (1979).
[2]D. Bollé and F. Gesztesy, Phys. Rev. A **30**, 1279 (1984); Phys. Rev. Lett. **52**, 1459 (1984).
[3]R. G. Newton, *Scattering Theory of Waves and Particles* (Springer, Berlin, 1982), 2nd ed.
[4]In three dimensions the assumption that $V(x) = v(|x|) \in L^1_2(\mathbb{R}_+)$ is natural because it is equivalent to $V(a+x) \in L^1(\mathbb{R}^3)$ for all a.
[5]$\overline{T}$ denotes the complex conjugate of T.
[6]See I. R. Lapidus, Am. J. Phys. **50**, 44 (1982) and S. K. Adhikari, Am. J. Phys. **54**, 362 (1986).
[7]A. Erdelyi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953), Vol. 2, p. 8, Eq. (32).
[8]This is the "surprise" discovered by D. Bollé, F. Gesztesy, C. Danneels, and S. F. J. Wilk, Phys. Rev. Lett. **53**, 900 (1986).

New classical properties of quantum coherent states

Gérard G. Emch[a)]
Institut für Theoretische Physik, Universität Göttingen, Göttingen, West Germany

Gerhard C. Hegerfeldt[b)]
A. P. Sloan Laboratory of Mathematics and Physics, California Institute of Technology, Pasadena, California 91125

(Received 10 July 1985; accepted for publication 16 July 1986)

A noncommutative version of the Cramer theorem is used to show that if two quantum systems are prepared independently, and if their center of mass is found to be in a coherent state, then each of the component systems is also in a coherent state, centered around the position in phase space predicted by the classical theory. Thermal coherent states are also shown to possess properties similar to classical ones.

I. INTRODUCTION

The coherent states ϕ to be studied in this paper have expectation values of the form

$$\langle \phi; e^{-i(uP+vQ)} \rangle = \exp\{-\Theta(\lambda u^2 + \lambda^{-1} v^2)/4\} e^{-i(u\langle P\rangle + v\langle Q\rangle)}, \quad (1.1)$$

where P and Q are the momentum and position operator for a quantum particle in one dimension; the generalization to $\mathbb{R}^n$ is straightforward. The physical interpretation of the parameters $\langle P\rangle$, $\langle Q\rangle$, Θ, and λ, characterizing the state ϕ, is obtained from (1.1) by differentiation; namely,

$$\begin{aligned} \langle P\rangle &= \langle \phi; P\rangle, \\ \langle Q\rangle &= \langle \phi; Q\rangle, \\ \Theta\lambda/2 &= \langle \phi; (P-\langle P\rangle)^2\rangle, \\ \Theta\lambda^{-1}/2 &= \langle \phi; (Q-\langle Q\rangle)^2\rangle, \end{aligned} \quad (1.2)$$

so that

$$\langle (P-\langle P\rangle)^2\rangle \langle (Q-\langle Q\rangle)^2\rangle = \Theta^2/4. \quad (1.3)$$

We must therefore have

$$\lambda > 0 \quad \text{and} \quad \Theta \geqslant \hbar. \quad (1.4)$$

The case $\Theta = \hbar$ corresponds to the class of coherent states introduced by Schrödinger[1]: they have minimal dispersion, compatible with the Heisenberg uncertainty relation, around the point $(\langle P\rangle, \langle Q\rangle)$ of the classical phase space $T^*\mathbb{R} \simeq \mathbb{R}^2$. They are pure states and are characterized by the existence of a vector $\hat{\Phi} \in \mathscr{H} \equiv \mathscr{L}^2(\mathbb{R}, dx)$ such that,

$$\langle \phi; e^{-i(uP+vQ)} \rangle = (\hat{\Phi}, e^{-i(uP+vQ)}\hat{\Phi}) \quad (1.5)$$

and

$$\hat{a}\hat{\Phi} = 0, \quad (1.6)$$

where

$$\hat{a} = \hat{Q} + i\lambda^{-1}\hat{P} \quad (1.7)$$

with

$$\hat{P} = P - \langle P\rangle, \quad \hat{Q} = Q - \langle Q\rangle. \quad (1.8)$$

Note that (1.6)–(1.8) is equivalent to saying that $\hat{\Phi}$ is the wave function for the ground state of the harmonic oscillator with Hamiltonian

$$\hat{H} = (1/2m)\hat{P}^2 + (1/2)k\hat{Q}^2, \quad (1.9)$$

where m and k satisfy the relations

$$\lambda = m\omega \quad \text{with} \quad \omega^2 = k/m. \quad (1.10)$$

When we further have

$$\langle P\rangle = 0 = \langle Q\rangle, \quad (1.11)$$

let us denote by Φ_0 the vector $\hat{\Phi}$ characterized by (1.5)–(1.8). Since the Schrödinger representation of the canonical commutation relations is irreducible, every vector $\Phi \in \mathscr{H}$ is cyclic. In particular we thus have that the (algebraic) vector space

$$\mathrm{Span}\{e^{-i(uP+vQ)}\Phi_0 | u, v \in \mathbb{R}\} \quad (1.12)$$

is dense in $\mathscr{H}$, and for general values of $\langle P\rangle$ and $\langle Q\rangle$ the corresponding vector $\hat{\Phi}$ is linked to Φ_0 by

$$\hat{\Phi} = e^{-i(\langle Q\rangle P - \langle P\rangle Q)/\hbar}\Phi_0. \quad (1.13)$$

In this sense, the vectors $\hat{\Phi}$, obtained by letting $(\langle P\rangle, \langle Q\rangle)$ run over the classical phase space $T^*\mathbb{R} \simeq \mathbb{R}^2$, form an overcomplete basis in $\mathscr{H}$, a mathematical property that has been given much attention[2] in connection with the theory of reproducing kernel Hilbert spaces.

When $\Theta > \hbar$, the change of variables

$$\Theta \equiv \hbar\omega \coth(\beta\hbar\omega/2), \quad \lambda \equiv m\omega, \quad (1.14)$$

allows one, as explained in Sec. III, to interpret the corresponding coherent state as the canonical equilibrium state, at inverse temperature β, for a quantum harmonic oscillator (1.9) with frequency defined as in (1.10). These states are therefore not pure.

All coherent states ($\Theta \geqslant \hbar$) have in common the property that they allow one,[3] upon controlling the limit $\hbar \to 0$, to derive from Mackey's formulation of quantum mechanics the formalism of classical mechanics, complete with its Jordan and Lie products, i.e., with the algebraic structures corresponding to (a) the pointwise multiplication of functions on the classical phase space T^*M, and (b) the Poisson bracket associated with the canonical symplectic form on T^*M.

In this paper we focus our attention on other classical

[a)] Permanent address: Department of Mathematics, The University of Florida, Gainesville, Florida 32601.
[b)] Permanent address: Institut für Theoretische Physik, Universität Göttingen, Göttingen, West Germany.

 0022-2488/86/112731-07$02.50

properties of quantum coherent states, exploring what can be said about the individual states of two quantum systems when these are prepared independently and when their center of mass is found to be in a coherent state. In Sec. II we limit our attention to the usual case of pure coherent states ($\Theta = \hbar$); the general case ($\Theta \geqslant \hbar$) is presented in Sec. II.

The mathematical motivation for this paper is a quantum version of the classical Cramer theorem.[5] The latter asserts that *if* the sum of two independent random variables is normally distributed, *then* each of the two random variables entering in this sum must also be normally distributed. Lemma 4.1 allows a simple derivation of a quantum version of this theorem adapted to the case $\Theta = \hbar$; in Sec. III, however, we need the general quantum version established by one of us in Refs. 6 and 7. The mathematical proofs, pertinent to the results stated in Secs. II and III are collected in Sec. IV.

II. PURE COHERENT STATES

If two *classical* particles, Σ_1 and Σ_2 say, are prepared independently and if their center of mass is found to be at the point $\{p_{\rm CM}, q_{\rm CM}\}$ of phase space one concludes immediately that the state of each of the component systems is described by a point $\{p_\kappa, q_\kappa\}$ ($\kappa = 1, 2$) and that

$$p_1 + p_2 = p_{\rm CM}, \quad \mu_1 q_1 + \mu_2 q_2 = q_{\rm CM}, \tag{2.1}$$

with

$$\mu_\kappa = m_\kappa / m_{\rm CM} \quad \text{and} \quad m_{\rm CM} = m_1 + m_2, \tag{2.2}$$

where m_κ is the mass of the κth particle.

If, however, the two particles are *quantum* systems, and one knows the wave function $\Psi_{\rm CM}$ describing the state of their center of mass, one cannot in general conclude anything about the shape of the wave function Ψ_κ ($\kappa = 1, 2$) of the two component systems, beyond consistency relations between expectation values, e.g.,

$$\langle P_1 \rangle + \langle P_2 \rangle = \langle P_{\rm CM} \rangle, \quad \mu_1 \langle Q_1 \rangle + \mu_2 \langle Q_2 \rangle = \langle Q_{\rm CM} \rangle, \tag{2.3}$$

$$\begin{aligned} &\langle (P_1 - \langle P_1 \rangle)^2 \rangle + \langle (P_2 - \langle P_2 \rangle)^2 \rangle \\ &\quad = \langle (P_{\rm CM} - \langle P_{\rm CM} \rangle)^2 \rangle, \\ &\mu_1^2 \langle (Q_1 - \langle Q_1 \rangle)^2 \rangle + \mu_2^2 \langle (Q_2 - \langle Q_2 \rangle)^2 \rangle \\ &\quad = \langle (Q_{\rm CM} - \langle Q_{\rm CM} \rangle)^2 \rangle, \end{aligned} \tag{2.4}$$

where μ_κ ($\kappa = 1,2$) are as in (2.2). To establish (2.3) one uses the linearity of the state, while to establish (2.4) one also uses the fact that when the two systems Σ_1 and Σ_2 are prepared independently, there are (by definition) no correlations between the observables A_1 relative to Σ_1 and the observable A_2 relative to Σ_2.

The purpose of this section is to show that if in addition $\Psi_{\rm CM}$ describes a coherent state, centered around the point $\{\langle P_{\rm CM} \rangle, \langle Q_{\rm CM} \rangle\}$ in the classical center-of-mass phase space, then each of the component system must be in a pure coherent state, centered precisely around the points $\{\langle P_\kappa \rangle, \langle Q_\kappa \rangle\}$ ($\kappa = 1,2$) satisfying (2.3), and with dispersion parameter λ_κ given by the now unique solution of (2.4), namely,

$$\lambda_\kappa = \mu_\kappa \lambda \quad (\text{with } \kappa = 1, 2), \tag{2.5}$$

where λ is the dispersion parameter of $\Psi_{\rm CM}$, determined uniquely from

$$\begin{aligned} &\langle (P_{\rm CM} - \langle P_{\rm CM} \rangle)^2 \rangle = \lambda \hbar / 2, \\ &\langle (Q_{\rm CM} - \langle Q_{\rm CM} \rangle)^2 \rangle = \lambda^{-1} \hbar / 2. \end{aligned} \tag{2.6}$$

As a consequence, the wave functions Ψ_1 and Ψ_2 will inherit both the Gaussian character of $\Psi_{\rm CM}$ *and* minimal dispersion, i.e., equality sign in the Heisenberg uncertainty relation,

$$\langle (P_\kappa - \langle P_\kappa \rangle)^2 \rangle \langle (Q_\kappa - \langle Q_\kappa \rangle)^2 \rangle = \hbar^2 / 4, \tag{2.7}$$

for $\kappa = 1,2$. We have, in fact,

$$\begin{aligned} &\langle (P_\kappa - \langle P_\kappa \rangle)^2 \rangle = \lambda_\kappa (\hbar / 2), \\ &\langle (Q_\kappa - \langle Q_\kappa \rangle)^2 \rangle = \lambda_\kappa^{-1} (\hbar / 2), \end{aligned} \tag{2.8}$$

with λ_κ as in (2.5).

As discussed in Sec. I, this is the closest one can possibly come to the classical result: when the scale of the phenomena one observes is such that $\hbar$ can be neglected, our quantum states are well approximated by the corresponding, dispersion-free classical states.

We now turn to the mathematical formulation of these results. For the general mathematical concepts underlying the following brief presentation, see, e.g., §8.3 and §9.1 in Ref. 4. In order to streamline our nomenclature, we systematically use the following abbreviations. By an "algebra" $\mathscr{A}$ we mean a W^*-algebra, with unit denoted by I, i.e., a C^*-algebra (with unit) that is the dual of a Banach space $\mathscr{A}_*$; by a "state" ϕ on $\mathscr{A}$, we mean a completely additive state, i.e., a positive linear functional

$$\phi: A \in \mathscr{A} \mapsto \langle \phi; A \rangle \in \mathbb{C} \tag{2.9}$$

that is normalized to 1 and belongs to $\mathscr{A}_*$ (these states are called "normal" in the literature on von Neumann algebras; we will, however, avoid this adjective here, as it may create confusion with the concept of "normal" distribution, familiar in the literature on classical statistics to which we also refer). The following particular case will be of central interest in the sequel: if $\mathscr{A}$ is (isomorphic, as a W^*-algebra, to) the algebra $\mathscr{B}(\mathscr{H})$ of all bounded linear operators on a separable Hilbert space $\mathscr{H}$, then completely additive states ϕ on $\mathscr{A}$ characterized by the fact that they are of the following form, familiar to physicists:

$$\phi: A \in \mathscr{A} \mapsto \mathrm{tr}(\rho A) \in \mathbb{C}, \tag{2.10}$$

where ρ is a density matrix, i.e., ρ is a positive trace-class operator on $\mathscr{H}$, of trace 1, uniquely determined by ϕ. In the present section, we are primarily concerned with (completely additive!) states ϕ on $\mathscr{A} \simeq \mathscr{B}(\mathscr{H})$ which are *pure*, i.e., states for which ρ is a one-dimensional projector, and we denote by Φ any unit vector in the range of ρ. For this section and the next, it is nevertheless useful to recall that for every state (whether pure or not) ϕ on $\mathscr{A}$, there exists a representation, unique up to unitary equivalence,

$$\pi_\phi: A \in \mathscr{A} \mapsto \pi_\phi(A) \in \mathscr{B}(\mathscr{H}_\phi) \tag{2.11}$$

called the GNS representation canonically associated to ϕ and characterized by the existence of a vector $\Phi \in \mathscr{H}_\phi$, such that

$$(\Phi, \pi_\phi(A)\Phi) = \langle \phi; A \rangle, \quad \forall A \in \mathscr{A}, \tag{2.12}$$

$$\overline{\text{Span}}\ \{\pi_\phi(A)\Phi | A\in\mathscr{A}\} = \mathscr{H}_\phi. \tag{2.13}$$

While the existence of the GNS representation does not require that ϕ be completely additive, the latter property ensures that π_ϕ is ultraweakly continuous so that $\pi_\phi(\mathscr{A})$ is a W^*-algebra. If in addition the state ϕ on $\mathscr{A}\simeq\mathscr{B}(\mathscr{H})$ is faithful, i.e., if

$$\langle\phi;A^*A\rangle = 0 \quad \text{implies} \quad A = 0, \tag{2.14}$$

then $\pi_\phi(\mathscr{A})$ is a factor, i.e.,

$$\pi_\phi(\mathscr{A})\cap\pi_\phi(\mathscr{A})' = \mathbb{C}I, \tag{2.15}$$

isomorphic to its commutant

$$\pi_\phi(\mathscr{A})' \equiv \{B\in\mathscr{B}(\mathscr{H}_\phi) | [B,\pi_\phi(A)] = 0,\ \forall A\in\mathscr{A}\}. \tag{2.16}$$

Moreover, every (completely additive!) state ψ on $\mathscr{A}\simeq\mathscr{B}(\mathscr{H})$ is then a vector state for this representation, i.e., there exists a vector $\Psi\in\mathscr{H}_\phi$ such that

$$(\Psi,\pi_\phi(A)\Psi) = \langle\psi;A\rangle, \quad \forall A\in\mathscr{A}. \tag{2.17}$$

Finally, by the Weyl CCR algebra for a particle with one degree of freedom, we mean the abstract W^*-algebra, defined by its realization on $L^2(\mathbb{R},dx)$, namely,

$$\mathscr{A} = \{e^{i(uP+vQ)} | u,v\in\mathbb{R}\}'', \tag{2.18}$$

where P and Q are the self-adjoint operators defined by their restriction to the Schwartz space $\mathscr{S}(\mathbb{R})$, i.e.,

$$(P\Psi)(x) = -i\hbar(\partial_x\Psi)(x),$$
$$(Q\Psi)(x) = x\Psi(x).$$

For two particles of mass m_1 and m_2 with Weyl CCR algebras $\mathscr{A}_1$ and $\mathscr{A}_2$, the Weyl CCR algebra $\mathscr{A}_{\text{CM}}$ for the center-of-mass motion is the subalgebra of $\mathscr{A}_1\otimes\mathscr{A}_2$ generated, in the L^2-realization, by

$$\{e^{i(uP_{\text{CM}}+vQ_{\text{CM}})} | u,v\in\mathbb{R}\}, \tag{2.19}$$

where, in analogy to (2.1) and (2.2),

$$P_{\text{CM}} \equiv P_1\otimes I + I\otimes P_2,$$
$$Q_{\text{CM}} \equiv \mu_1 Q_1\otimes I + \mu_2 I\otimes Q_2. \tag{2.20}$$

The results of this section can now be expressed mathematically as follows.

Theorem 2.1: Let $\mathscr{A}_\kappa$ ($\kappa = 1,2$) be the Weyl CCR algebras for two particles with one degree of freedom; let $\mathscr{A}_{\text{CM}}$ be the Weyl CCR algebra for the center-of-mass motion; let ϕ_κ be a state on $\mathscr{A}_\kappa$ ($\kappa = 1,2$); let

$$\phi_0 \equiv \phi_1\otimes\phi_2 \quad \text{on} \quad \mathscr{A}_0 \equiv \mathscr{A}_1\otimes\mathscr{A}_2; \tag{2.21}$$

and let ϕ_{CM} be the restriction of ϕ_0 to $\mathscr{A}_{\text{CM}}\subset\mathscr{A}_0$, i.e.,

$$\phi_{\text{CM}} \equiv \phi_0 \upharpoonright \mathscr{A}_{\text{CM}} \equiv \phi_1{}^*\phi_2. \tag{2.22}$$

If ϕ_{CM} is a pure coherent state, then ϕ_1 and ϕ_2 are also pure coherent states.

The relations between the characteristic parameters of ϕ_1 and ϕ_2 and those of ϕ_{CM} are now specified.

Corollary 2.2 With the notation of the theorem, the pure coherent state ϕ_{CM} is completely described by

$$\langle\phi_{\text{CM}};e^{-i(uP_{\text{CM}}+vQ_{\text{CM}})}\rangle = \exp\{-\hbar(\lambda_{\text{CM}}u^2 + \lambda_{\text{CM}}^{-1}v^2)/4\}e^{-i(u\langle P_{\text{CM}}\rangle + v\langle Q_{\text{CM}}\rangle)} \tag{2.23}$$

valid for all $u, v\in\mathbb{R}$; the characteristic parameters $\langle P_{\text{CM}}\rangle$, $\langle Q_{\text{CM}}\rangle$, and λ_{CM} of ϕ_{CM} are determined by the relations

$$\langle P_{\text{CM}}\rangle = \langle\phi_{\text{CM}};P_{\text{CM}}\rangle, \quad \langle Q_{\text{CM}}\rangle = \langle\phi_{\text{CM}};Q_{\text{CM}}\rangle, \tag{2.24}$$

$$\langle\phi_{\text{CM}};(P_{\text{CM}} - \langle P_{\text{CM}}\rangle)^2\rangle = \lambda_{\text{CM}}\hbar/2,$$
$$\langle\phi_{\text{CM}};(Q_{\text{CM}} - \langle Q_{\text{CM}}\rangle)^2\rangle = \lambda_{\text{CM}}^{-1}\hbar/2. \tag{2.25}$$

The ϕ_κ ($\kappa = 1,2$) are then of the form

$$\langle\phi_\kappa;e^{-i(uP_{\text{CM}}+vQ_{\text{CM}})}\rangle = \exp\{-\hbar(\lambda_\kappa u^2 + \lambda_\kappa^{-1}v^2)/4\}e^{-i(u\langle P_\kappa\rangle + v\langle Q_\kappa\rangle)}, \tag{2.26}$$

where

$$\lambda_\kappa = \mu_\kappa\lambda_{\text{CM}}, \quad \mu_\kappa = m_\kappa/m_{\text{CM}}, \quad m_{\text{CM}} = m_1 + m_2, \tag{2.27}$$

$$\langle P_\kappa\rangle = \langle\phi_\kappa;P_\kappa\rangle, \quad \langle Q_\kappa\rangle = \langle\phi_\kappa;Q_\kappa\rangle, \tag{2.28}$$

and

$$\langle P_1\rangle + \langle P_2\rangle = \langle P_{\text{CM}}\rangle,$$
$$\mu_1\langle Q_1\rangle + \mu_2\langle Q_2\rangle = \langle Q_{\text{CM}}\rangle. \tag{2.29}$$

The following information on the state of relative motion is also available.

Corollary 2.3: With the notation and assumptions of the theorem and with ϕ_{rel} the restriction of ϕ_0 to $\mathscr{A}_{\text{rel}}\subset\mathscr{A}_0$, we have that ϕ_0 is also a pure coherent state, and

$$\phi_0 = \phi_{\text{CM}}\otimes\phi_{\text{rel}}. \tag{2.30}$$

Note that (2.30) means physically that there are no correlations between the observables for the center of mass and those for the relative motion.

The above three results follow directly from the noncommutative extension of the classical Cramer theorem obtained in Ref. 6, the essence of which, for the case of interest here, is captured in Lemma 4.1 below.

III. THERMAL COHERENT STATES

Let Σ be a *classical* ideal gas in canonical equilbrium at inverse temperature β; its partition function Z and density function f are thus, by definition

$$Z = \int\cdots\int dp_1\cdots dp_N\, dq_1\cdots dq_N \times\exp[-\beta H(p_1,\ldots,p_N,q_1,\ldots,q_N)], \tag{3.1}$$

$$f(p_1,\ldots,p_N,q_1,\ldots,q_N) = Z^{-1}\exp[-\beta H(p_1,\ldots,p_N,q_1,\ldots,q_N)], \tag{3.2}$$

with

$$H(p_1,\ldots,p_N,q_1,\ldots,q_N) = \sum_{\kappa=1}^{N} H_\kappa(p_\kappa,q_\kappa) \tag{3.3}$$

and, for $\kappa = 1,2,\ldots,N$,

$$H_\kappa(p_\kappa,q_\kappa) = (1/2m_\kappa)p_\kappa^2 + V_\kappa(x_\kappa). \tag{3.4}$$

Suppose now that the center of mass of this ideal gas is observed to be distributed according to the canonical equilibrium density of a harmonic oscillator, i.e,

$$f_{\text{CM}}(p_{\text{CM}},q_{\text{CM}}) = Z_{\text{CM}}^{-1}e^{-\beta H_{\text{CM}}(p_{\text{CM}},q_{\text{CM}})}, \tag{3.5}$$

with

$$H_{\rm CM}(p_{\rm CM},q_{\rm CM}) = (1/2m_{\rm CM})p_{\rm CM}^2 + \tfrac{1}{2}k_{\rm CM}q_{\rm CM}^2, \tag{3.6}$$

$$Z_{\rm CM} = \int\int dp_{\rm CM}\, dq_{\rm CM}\, \exp\{-\beta H_{\rm CM}(p_{\rm CM},q_{\rm CM})\}. \tag{3.7}$$

It then follows, by repeated application of the classical Cramer theorem, that the situation described by (3.5)–(3.7) occurs if and only if the individual particles of the ideal gas are displaced harmonic oscillators, in equilibrium at the inverse temperature β. Specifically, one finds, for $\kappa = 1,2,...,N$,

$$H_\kappa = (1/2m_\kappa)(p_\kappa - \langle p_\kappa\rangle)^2 + \tfrac{1}{2}k_\kappa(q_\kappa - \langle q_\kappa\rangle)^2, \tag{3.8}$$

with

$$\sum_{\kappa=1}^{N} m_\kappa = m_{\rm CM}, \tag{3.9}$$

$$\sum_{\kappa=1}^{N} \mu_\kappa \omega_\kappa^{-2} = \omega_{\rm CM}^{-2}, \tag{3.10}$$

$$\sum_{\kappa=1}^{N} \mu_\kappa \langle q_\kappa\rangle = 0 \quad \text{and} \quad \sum_{\kappa=1}^{N} \langle p_\kappa\rangle = 0, \tag{3.11}$$

where

$$\mu_\kappa = m_\kappa/m_{\rm CM} \tag{3.12}$$

$$\omega_\kappa^2 = k_\kappa/m_\kappa \quad \text{and} \quad \omega_{\rm CM}^2 = k_{\rm CM}/m_{\rm CM}. \tag{3.13}$$

Note that

$$\omega = \omega_{\rm CM}, \quad \forall\kappa = 1,2,...,N, \tag{3.14}$$

is always a solution of (3.10) with μ_κ defined by (3.12) and (3.9). Mathematically, this particular solution is characterized by the condition that the independent, $\mathbb{R}^2$-valued random variables

$$(\tilde p_\kappa,\tilde q_\kappa), \quad \kappa = 1,2,...,N, \tag{3.15}$$

defined by

$$\tilde p_\kappa = \mu_\kappa^{-1/2} p_\kappa, \tag{3.16}$$

$$\tilde q_\kappa = \mu_\kappa^{1/2}\hat q_\kappa \quad \text{with} \quad \hat q_\kappa = q_\kappa - \langle q_\kappa\rangle \tag{3.17}$$

be identically distributed, with density

$$f(p,q) = (2\pi)^{-1}\beta\omega_{\rm CM}^{1/2} \times \exp[-\beta((1/2m_{\rm CM})p^2 + \tfrac{1}{2}k_{\rm CM}q^2)]. \tag{3.18}$$

Alternatively, this condition can be expressed by saying that for any pair $\kappa_1 \neq \kappa_2$ of indices 1, 2,...,N, the two $\mathbb{R}^2$-valued random variables

$$(\bar p_{\rm CM},\bar q_{\rm CM}) \quad \text{and} \quad (\bar p_{\rm rel},\bar q_{\rm rel}) \tag{3.19}$$

are statistically independent, where

$$\bar p_{\rm CM} = p_{\kappa_1} + p_{\kappa_2}, \tag{3.20}$$

$$\bar q_{\rm CM} = \bar\mu_1\hat q_{\kappa_1} + \bar\mu_2\hat q_{\kappa_2}, \tag{3.21}$$

$$\bar p_{\rm rel} = \bar m_{\rm rel}\left(\frac{1}{m_{\kappa_1}}p_{\kappa_1} - \frac{1}{m_{\kappa_2}}p_{\kappa_2}\right), \tag{3.22}$$

$$\bar q_{\rm rel} = \hat q_{\kappa_1} - \hat q_{\kappa_2} \tag{3.23}$$

with

$$\bar\mu_\kappa = m_\kappa/\bar m_{\rm CM} \quad (\kappa = \kappa_1,\kappa_2), \tag{3.24}$$

$$\bar m_{\rm CM} = m_{\kappa_1} + m_{\kappa_2}, \tag{3.25}$$

$$m_{\rm rel} = m_{\kappa_1}m_{\kappa_2}/\bar m_{\rm CM}. \tag{3.26}$$

Physically, the condition (3.14) means thus that for every pair $\Sigma_{\kappa_1} \neq \Sigma_{\kappa_2}$ of oscillators in the gas one has

$$H_{\kappa_1}(p_{\kappa_1},q_{\kappa_1}) + H_{\kappa_2}(p_{\kappa_2},q_{\kappa_2}) = \bar H_{\rm CM}(\bar p_{\rm CM},\bar q_{\rm CM}) + \bar H_{\rm rel}(\bar p_{\rm rel},\bar q_{\rm rel}), \tag{3.27}$$

where $\bar H_{\rm CM}$ and $\bar H_{\rm rel}$ are harmonic oscillator Hamiltonians. Specifically

$$\bar H_{\rm CM} = (1/2\bar m_{\rm CM})\bar p_{\rm CM}^{\,2} + \tfrac{1}{2}\bar k_{\rm CM}\bar q_{\rm CM}^{\,2}, \tag{3.28}$$

$$\bar H_{\rm rel} = (1/2\bar m_{\rm rel})\bar p_{\rm rel}^{\,2} + \tfrac{1}{2}\bar k_{\rm rel}\bar q_{\rm rel}^{\,2}, \tag{3.29}$$

where the masses $\bar m_{\rm CM}$ and $\bar m_{\rm rel}$ are defined in (3.25) and (3.26) and the oscillator strengths $\bar k_{\rm CM}$ and $\bar k_{\rm rel}$ are given by

$$\bar k_{\rm CM}/\bar m_{\rm CM} = \omega_{\rm CM}^2 = \omega_{\rm rel}^2 = k_{\rm rel}/m_{\rm rel}. \tag{3.30}$$

The purpose of this section is to analyze the corresponding *quantum* situation. Let

$$V:\ x\in\mathbb{R}\mapsto V(x)\in\mathbb{R} \tag{3.31}$$

be such that

$$H = -\hbar^2\frac{1}{2m}\frac{d^2}{dx^2} + V(x) \tag{3.32}$$

defines a self-adjoint operator in $\mathscr{H} = \mathscr{L}^2(\mathbb{R},dx)$ with $\exp(-\beta H)$ of trace class for all $\beta > 0$.

The density matrix

$$\rho = Z^{-1}\exp(-\beta H) \tag{3.33}$$

with

$$Z = \mathrm{Tr}\exp(-\beta H) \tag{3.34}$$

is then interpreted as the canonical equilibrium state, at inverse temperature β, of a quantum particle in the potential V. In particular, for a harmonic oscillator

$$H = -\hbar^2\frac{1}{2m}\frac{d^2}{dx^2} + \frac{1}{2}kx^2, \tag{3.35}$$

the state

$$\phi:\ B\in\mathscr{B}(\mathscr{H})\mapsto \mathrm{Tr}\,\rho B\in\mathbb{C} \tag{3.36}$$

is faithful and is uniquely determined by its restriction on the Weyl algebra; specifically, with P and Q defined as in (2.19 and 2.20) one has[4]

$$\langle\phi;\exp[-i(uP+vQ)]\rangle = \exp\{-\Theta(\lambda u^2 + \lambda^{-1}v^2)/4\}, \tag{3.37}$$

where

$$\Theta = \hbar\coth(\beta\hbar\omega/2) \tag{3.38}$$

$$\lambda = m\omega \quad \text{with} \quad \omega^2 = k/m. \tag{3.39}$$

It is worthwhile for the sequel to note that (i) ϕ is Gaussian; (ii) one recovers the classical result

$$\lim_{\hbar\to 0}\langle P^2\rangle = \lim_{\hbar\to 0}\Theta\lambda/2 = m/\beta, \qquad \lim_{\hbar\to 0}\langle Q^2\rangle = \lim_{\hbar\to 0}\Theta\lambda^{-1}/2 = 1/k\beta; \tag{3.40}$$

and (iii) one recovers the low-temperature limit of Sec. II, namely,

$$\lim_{\beta\to\infty}\langle P^2\rangle\langle Q^2\rangle=\hbar^2/4. \tag{3.41}$$

We therefore extend the definition of coherent states on the Weyl CCR algebra for one degree of freedom to include states that satisfy

$$\langle\phi;\exp[-i(uP+vQ)]\rangle = \exp\{-\Theta(\lambda u^2+\lambda^{-1}v^2)/4\}e^{-i(u\langle P\rangle+v\langle Q\rangle)} \tag{3.42}$$

with $\Theta/\hbar\geqslant 1$; such states are pure coherent states (in the sense of Sec. II) if and only if $\Theta/\hbar=1$, i.e., $\beta=\infty$ in Eq. (3.38). The following result is stated for two-particle systems although it extends trivially, as does its classical counterpart, to an n-particle system.

Theorem 3.1: Let $\mathscr{A}_\kappa$ (with $\kappa=1,2,\mathrm{CM}$) be as in Theorem 2.1. For $\kappa=1,2$ let ϕ_κ be a state on $\mathscr{A}_\kappa$, and let

$$\phi_0\equiv\phi_1\otimes\phi_2 \quad\text{on}\quad \mathscr{A}_0\equiv\mathscr{A}_1\otimes\mathscr{A}_2. \tag{3.43}$$

Then the restriction ϕ_{CM} of ϕ_0 to $\mathscr{A}_{\mathrm{CM}}$ is a coherent state of the form

$$\langle\phi_{\mathrm{CM}};\exp[-i(uP_{\mathrm{CM}}+vQ_{\mathrm{CM}})]\rangle = \exp\{-\Theta_{\mathrm{CM}}(\lambda_{\mathrm{CM}}u^2+\lambda_{\mathrm{CM}}^{-1}v^2)/4\}e^{-i(u\langle P_{\mathrm{CM}}\rangle+v\langle Q_{\mathrm{CM}}\rangle)},$$
$$\text{with}\quad \Theta_{\mathrm{CM}}/\hbar\geqslant 1 \quad\text{and}\quad \lambda_{\mathrm{CM}}>0, \tag{3.44}$$

if and only if ϕ_κ ($\kappa=1,2$) are coherent states of the form

$$\langle\phi_\kappa;\exp[-i(uP_\kappa+vQ_\kappa)]\rangle = \exp\{-\Theta_\kappa(\lambda_\kappa u^2+\lambda_\kappa^{-1}v^2)/4\}\times\exp[-i(u\langle P_\kappa\rangle+v\langle Q_\kappa\rangle)],$$
$$\text{with}\quad \Theta_\kappa/\hbar\geqslant 1 \quad\text{and}\quad \lambda_\kappa>0 \tag{3.45}$$

with the compatibility relations

$$\langle P_1\rangle+\langle P_2\rangle=\langle P_{\mathrm{CM}}\rangle,\qquad \mu_1\langle Q_1\rangle+\mu_2\langle Q_2\rangle=\langle Q_{\mathrm{CM}}\rangle, \tag{3.46}$$

$$\Theta_1\lambda_1+\Theta_2\lambda_2=\Theta_{\mathrm{CM}}\lambda_{\mathrm{CM}},\qquad \mu_1^2\Theta_1\lambda_1^{-1}+\mu_2^2\Theta_2\lambda_2^{-1}=\Theta_{\mathrm{CM}}\lambda_{\mathrm{CM}}^{-1}, \tag{3.47}$$

where

$$\mu_\kappa=m_\kappa/m_{\mathrm{CM}} \quad\text{and}\quad m_{\mathrm{CM}}=m_1+m_2. \tag{3.48}$$

The physical meaning of the compatibility condition (3.47) is given by the following result.

Scholium 3.2: With (3.43)–(3.46) taken into account, (3.47) is equivalent to

$$\langle(P_1-\langle P_1\rangle)^2\rangle+\langle(P_2-\langle P_2\rangle)^2\rangle = \langle(P_{\mathrm{CM}}-\langle P_{\mathrm{CM}}\rangle)^2\rangle,$$
$$\mu_1^2\langle(Q_1-\langle Q_1\rangle)^2+\mu_2^2\langle Q_2-\langle Q_2\rangle)^2\rangle = \langle(Q_{\mathrm{CM}}-\langle Q_{\mathrm{CM}}\rangle)^2\rangle, \tag{3.49}$$

where μ_1 and μ_2 are given by (3.48)

Note that these results are in conformity with the classical results; see in particle (3.9) and (3.10).

The results of Sec. II ("low-temperature limit") are recovered from (3.45) and (3.46) and the following consequence of (3.47).

Scholium 3.3: With the notation of Theorem 3.1, the following two conditions are equivalent:

$$\Theta_{\mathrm{CM}}=\hbar, \tag{3.50}$$

$$\text{for}\ \kappa=1,2,\quad \Theta_\kappa=\hbar \quad\text{and}\quad \lambda_\kappa=\mu_\kappa\lambda. \tag{3.51}$$

The following change of variables allows us to interpret our results in terms of canonical equilibrium states of harmonic oscillators, in particular in the nontrivial Corollary 3.5.

Scholium 3.4: With the notation of Theorem 3.1, there exist (for $\kappa=1,2,\mathrm{CM}$) $\beta_\kappa\in(0,\infty]$ and $\omega_\kappa\in(0,\infty)$ such that

$$\Theta_\kappa=\hbar\coth(\beta_\kappa\hbar\omega_\kappa/2),\quad \lambda_\kappa=m_\kappa\omega_\kappa. \tag{3.52}$$

Corollary 3.5: With the notations of Theorem 3.1 and Scholium 3.4 assume that

$$\beta_1=\beta_2\equiv\beta\in(0,\infty), \tag{3.53}$$

and

$$\text{either}\quad \beta_{\mathrm{CM}}=\beta \quad\text{or}\quad \omega_1=\omega_2\equiv\omega. \tag{3.54}$$

Then

$$\omega_1=\omega_2=\omega_{\mathrm{CM}} \quad\text{and}\quad \beta_1=\beta_2=\beta_{\mathrm{CM}}, \tag{3.55}$$

and

$$\phi_0=\phi_{\mathrm{CM}}\otimes\phi_{\mathrm{rel}}, \tag{3.56}$$

where ϕ_{rel} is the coherent state

$$\langle\phi_{\mathrm{rel}};\exp[-i(uP_{\mathrm{rel}}+vQ_{\mathrm{rel}})]\rangle = \exp\{-\Theta_{\mathrm{rel}}(\lambda_{\mathrm{rel}}u^2+\lambda_{\mathrm{rel}}^{-1}v^2)/4\}\times\exp[-i(u\langle P_{\mathrm{rel}}\rangle+v\langle Q_{\mathrm{rel}}\rangle)], \tag{3.57}$$

with

$$\Theta_{\mathrm{rel}}=\hbar\coth(\beta_{\mathrm{rel}}\hbar\omega_{\mathrm{rel}}/2),\qquad \omega_{\mathrm{rel}}=\omega_{\mathrm{CM}},\quad \beta_{\mathrm{rel}}=\beta_{\mathrm{CM}},\qquad \lambda_{\mathrm{rel}}=[m_1m_2/(m_1+m_2)]\lambda_{\mathrm{CM}}. \tag{3.58}$$

IV. PROOFS

The proofs of Theorem 2.1 and of its Corollaries 2.2 and 2.3 follow directly from the introductory remarks presented in Sec. I–see in particular (1.5) and (1.6) and (1.9) and (1.10)–and the next simple lemma, an analog of Lemma 2.2 in Ref. 6. The reader interested in domain questions may consult Lemma 2.1 in Ref. 6.

Lemma 4.1: With the notation and assumptions of Theorem 2.1, let (for $\kappa=1,2$) π_κ be the GNS representation of $\mathscr{A}_\kappa$ associated to ϕ_κ, let Φ_κ be the corresponding cyclic vector, and let

$$\hat{P}_\kappa\equiv\pi_\kappa(P_\kappa)-\langle P_\kappa\rangle,\qquad \hat{Q}_\kappa\equiv\pi_\kappa(Q_\kappa)-\langle Q_\kappa\rangle,\qquad \hat{a}_\kappa\equiv\hat{Q}_\kappa+i\lambda_\kappa^{-1}\hat{P}_\kappa.$$

Then, for λ_κ as in (2.27), one has

$$a_\kappa\Phi_\kappa=0 \quad(\kappa=1,2).$$

Proof: Let π_κ ($\kappa=1,2,0$) be the GNS representation of $\mathscr{A}_\kappa$ associated with ϕ_κ, and let Φ_κ be the corresponding cyclic vector. Note that

$$\Phi_0=\Phi_1\otimes\Phi_2. \tag{4.1}$$

Define

$$\hat{P}_{\rm CM} \equiv \pi_0(P_{\rm CM}) - \langle P_{\rm CM}\rangle,$$
$$\hat{Q}_{\rm CM} \equiv \pi_0(Q_{\rm CM}) - \langle Q_{\rm CM}\rangle, \tag{4.2}$$
$$\hat{a}_{\rm CM} \equiv \hat{Q}_{\rm CM} + i\lambda_{\rm CM}^{-1}\hat{P}_{\rm CM},$$

and, for $\kappa = 1,2$,

$$\hat{P}_\kappa \equiv \pi_\kappa(P_\kappa) - \langle P_\kappa\rangle,$$
$$\hat{Q}_\kappa \equiv \pi_\kappa(Q_\kappa) - \langle Q_\kappa\rangle, \tag{4.3}$$
$$\hat{a}_\kappa \equiv \hat{Q}_\kappa + i\lambda_\kappa^{-1}P_\kappa,$$

where

$$\langle P_\kappa\rangle \equiv \langle \phi_\kappa; P_\kappa\rangle, \quad \langle Q_\kappa\rangle \equiv \langle \phi_\kappa; Q_\kappa\rangle,$$
$$\lambda_\kappa \equiv \mu_\kappa \lambda_{\rm CM}, \quad \mu_\kappa \equiv m_\kappa/(m_1+m_2). \tag{4.4}$$

We then have

$$\hat{a}_{\rm CM} = \mu_1 \hat{a}_1 \otimes I + \mu_2 I \otimes \hat{a}_2, \tag{4.5}$$
$$(\Phi_\kappa, \hat{a}_\kappa \Phi_\kappa) = 0 \quad (\kappa = 1,2) \tag{4.6}$$

and, from the fact that $\phi_{\rm CM}$ is a pure coherent state,

$$\hat{a}_{\rm CM}\Phi_0 = 0. \tag{4.7}$$

Upon inserting (4.5) and (4.7), taking the norm of the resulting expression, and taking (4.6) into account, we obtain

$$\mu_1^2 ||\hat{a}_1\Phi_1||^2 + \mu_2^2 ||\hat{a}_2\Phi_2||^2 = 0 \tag{4.8}$$

and thus, since $\mu_\kappa > 0$,

$$\hat{a}_\kappa \Phi_\kappa = 0 \quad (\kappa = 1,2). \tag{4.9}$$

This proves Lemma 4.1.

The proof of (3.45) in Theorem (3.1) is a straightforward application of the general quantum version of Cramer's Theorem established by one of us.[6,7] The consistency relations (3.46)–(3.48) follow then by inspection, we replace $P_{\rm CM}$ and $Q_{\rm CM}$ in (3.44) by their definition (2.20) and match then (3.44) and (3.45), taking into account (3.43).

Scholium 3.2 follows immediately from (3.44), (3.45), and (1.2).

Proof of Scholium 3.3: We multiply the two equations in (3.47) by one another to obtain

$$(\mu_1\Theta_1 + \mu_2\Theta_2)^2 + \mu_1\mu_2\Theta_1\Theta_2(\lambda_1\lambda_2)^{-1}(\mu_2\lambda_1 - \mu_1\lambda_2)^2 = \Theta_{\rm CM}^2. \tag{4.10}$$

From the facts that $\Theta_{\rm CM} = \hbar$, $\mu_1 + \mu_2 = 1$, and $\Theta_\kappa \geqslant \hbar$ ($\kappa = 1, 2$), we conclude from (4.10) that

$$\Theta_\kappa = \hbar \quad (\kappa = 1,2) \tag{4.11}$$

and

$$\mu_1^{-1}\lambda_1 = \mu_2^{-1}\lambda_2. \tag{4.12}$$

Upon inserting (4.12) in the first (or the second) of the consistency relations (3.47), upon taking into account that $\mu_1 + \mu_2 = 1$ and that $\Theta_1 = \Theta_2 = \Theta_{\rm CM}$, we obtain

$$\lambda_\kappa = \mu_\kappa \lambda \quad (\kappa = 1,2). \tag{4.13}$$

This completes the proof.

Scholium 3.4 is only an adaptation of the change of variables (1.14) to the situation now under consideration. Note that $\beta_\kappa = \infty$ corresponds to the pure case $\Theta_\kappa = 1$.

Proof of Corollary 3.5: With the change of variables (3.52) the consistency relation (3.47) reads

$$\mu_1\Theta_1\omega_1 + \mu_2\theta_2\omega_2 = \Theta_{\rm CM}\omega_{\rm CM},$$
$$\mu_1\Theta_1\omega_1^{-1} + \mu_2\Theta_2\omega_2^{-1} = \Theta_{\rm CM}\omega_{\rm CM}^{-1}, \tag{4.14}$$

with

$$\Theta_\kappa = \hbar \coth(\beta_\kappa\hbar\omega_\kappa/2) \quad (\kappa = 1,2,{\rm CM}). \tag{4.15}$$

In case we assume

$$\beta_1 = \beta_2 = \beta_{\rm CM} \equiv \beta \in (0,\infty), \tag{4.16}$$

it is useful to write the consistency equations (4.14) in the vector form

$$\mu_1 X(\omega_1) + \mu_2 X(\omega_2) = X(\omega_{\rm CM}), \tag{4.17}$$

with

$$X(\omega) \equiv \begin{pmatrix} \Xi(\omega) \\ \xi(\omega) \end{pmatrix} \tag{4.18}$$

and

$$\Xi\colon\ \omega \in (0,\infty) \mapsto \omega\coth(\beta\hbar\omega/2) \in (0,\infty),$$
$$\xi\colon\ \omega \in (0,\infty) \mapsto \omega^{-1}\coth(\beta\hbar\omega/2) \in (0,\infty). \tag{4.19}$$

Upon noticing that ξ is bijective, we can use ξ as a variable, and define

$$\hat{\Xi}(\xi) \equiv \Xi \circ \omega(\xi) \tag{4.20}$$

and

$$\hat{X}(\xi) \equiv \begin{pmatrix} \hat{\Xi}(\xi) \\ \xi \end{pmatrix}. \tag{4.21}$$

We then verify that

$$\frac{d^2}{d\xi^2}\hat{\Xi}(\xi) > 0, \tag{4.22}$$

i.e., that $\hat{\Xi}$ is strictly convex. As a consequence, the equation

$$\mu_1\hat{X}(\xi_1) + \mu_2\hat{X}(\xi_2) = \hat{X}(\xi_{\rm CM}), \tag{4.23}$$

where

$$\mu_\kappa > 0 \quad \text{and} \quad \mu_1 + \mu_2 = 1, \tag{4.24}$$

admits a unique solution, namely,

$$\xi_1 = \xi_2 = \xi_{\rm CM}, \tag{4.25}$$

i.e.,

$$\omega_1 = \omega_2 = \omega_{\rm CM}. \tag{4.26}$$

We have thus proven the first part of (3.54) and (3.55).

If we now assume

$$\beta_1 = \beta_2 \equiv \beta \in (0,\infty] \quad \text{and} \quad \omega_1 = \omega_2 \equiv \omega \in (0,\infty), \tag{4.27}$$

we have

$$\Theta_1 = \Theta_2 \equiv \Theta \equiv \hbar\coth(\beta\hbar\omega/2), \tag{4.28}$$

so that (4.26) reduces to

$$\Theta\omega = \Theta_{\rm CM}\omega_{\rm CM}, \quad \Theta\omega^{-1} = \Theta_{\rm CM}\omega_{\rm CM}^{-1}, \tag{4.29}$$

from which we obtain, upon using (4.15) and (4.28),

$$\Theta_{\rm CM} = \Theta, \quad \omega_{\rm CM} = \omega, \quad \beta_{\rm CM} = \beta. \tag{4.30}$$

We have thus proved (3.55). The remainder of the corollary follows then by straightforward inspection. Q.E.D.

ACKNOWLEDGMENTS

This work was supported in part by the Akademie für Wissenschaften zu Göttingen (GGE) and by the Stiftung Volkswagenwerk (GCH).

[1]E. Schrödinger, Naturwissenschaften **14**, 664 (1926).
[2]J. R. Klauder, Ann. Phys. (NY) **11**, 123 (1960); R. J. Glauber, Phys. Rev. **131**, 2766 (1963); J. R. Klauder and B.-S. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985) and references therein.
[3]G. G. Emch, Int. J. Theor. Phys. **22**, 397 (1983); see, also, Sec. 8.3.g in Ref. 4; for an extension to not necessarily flat manifolds, see G. G. Emch, J. Math. Phys. **23**, 1785 (1982).
[4]G. G. Emch, *Mathematical and Conceptual Foundations of 20th Century Physics*, Mathematical Studies Vol. 100 (North-Holland, Amsterdam, 1984).
[5]H. Cramer, Math. Z. **41**, 405 (1936); see, also, E. Lukacs, *Characteristic Functions* (Griffin, London, 1970), 2nd ed., p. 243.
[6]G. C. Hegerfeldt, J. Math. Phys. **26**, 62 (1985).
[7]G. C. Hegerfeldt, "Noncommutative analogs of probabilistic notions and results," J. Funct. Anal. (in press).

Phase-integral formulas for Bessel functions and their relation to already existing asymptotic formulas

Per Olof Fröman, Finn Karlsson, and Staffan Yngve
Institute of Theoretical Physics, University of Uppsala, Thunbergsvägen 3, S-752 38 Uppsala, Sweden

(Received 22 January 1986; accepted for publication 11 July 1986)

The phase-integral method devised by Fröman and Fröman [N. Fröman and P. O. Fröman, *JWKB Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965); Ann. Phys. (NY) **83**, 103 (1974); Nuovo Cimento B **20**, 121 (1974); N. Fröman, Ark. Fys. **32**, 541 (1966); Ann. Phys. (NY) **61**, 451 (1970)], involving a general phase-integral approximation of arbitrary order, which is generated from an unspecified base function, is used for deriving first- and higher-order phase-integral formulas for Bessel functions. For different choices of the base function one thus obtains in a systematic way different kinds of asymptotic formulas. By series expansion of these formulas one obtains already existing asymptotic formulas presented is standard handbooks. The phase-integral formulas are seen to have certain advantages that those latter formulas do not possess.

I. BACKGROUND

Consider the differential equation

$$\frac{d^2\psi}{dz^2} + R(z)\psi = 0, \tag{1.1}$$

where $R(z)$ is an analytic function of z. For the approximate (but in general very accurate) solution of this differential equation one can use the phase-integral method developed by Fröman and Fröman.[1–5] For the advantages of this method and its relation to the so-called WKB method we refer to papers by Dammert and P. O. Fröman[6] and by Fröman and Fröman.[7] Briefly speaking one can say that the method consists first in the solution of the local problem, where one determines two linearly independent phase-integral functions, generated from an unspecified base function, that are approximate solutions of the differential equation, and second in the solution of the global problem, where one determines the appropriate linear combinations of these phase-integral functions, which approximately represent the exact solution in various regions of the complex z plane.

A. Phase-integral approximation of arbitrary order generated from an unspecified base function

The phase-integral approximation to be described now was introduced in Ref. 4 and on pp. 126–131 in Ref. 5, and it was summarized in a somewhat more lucid way in Ref. 7.

In the original differential equation (1.1) there appears no small parameter. One of the essential ideas behind the phase-integral approximation to be discussed now is the realization of how a "small" bookkeeping parameter can conveniently be introduced in a flexible way. To this purpose we introduce an unspecified function $Q(z)$, called the base function, and write the function $R(z)$ in the differential equation (1.1) as the sum of the two functions $Q^2(z)$ and $[R(z) - Q^2(z)]$, of which $Q^2(z)$ is considered to be dominant in some sense. To account in an explicit way for this dominance we introduce a "small" parameter λ, which will at the end be put equal to unity. Thus, instead of the original differential equation (1.1), we now consider the auxiliary differential equation

$$\frac{d^2\psi}{dz^2} + \left(\frac{Q^2(z)}{\lambda^2} + [R(z) - Q^2(z)]\right)\psi = 0, \tag{1.2}$$

where λ is a "small" bookkeeping parameter, and $Q(z)$ is the so far unspecified base function. The functions $R(z)$ and $Q^2(z)$ are assumed to be independent of λ and to have such properties that the phase-integral solution of (1.2) remains valid when one puts λ equal to unity. The choice of the base function $Q(z)$ obviously determines how the "small" parameter λ appears in the auxiliary differential equation. By putting $\lambda = 1$ in a solution of (1.2), one obviously obtains the corresponding solution of (1.1).

The differential equation (1.2) has the two linearly independent, approximate solutions

$$\psi = q^{-1/2}(z)\exp[\pm iw(z)], \tag{1.3}$$

where

$$w(z) = \int^z q(z)dz. \tag{1.4}$$

When the order of the approximation is $2N + 1$, the expression for $q(z)$ is

$$q(z) = \sum_{n=0}^{N} Y^{(2n+1)}Q(z)\lambda^{2n-1}, \tag{1.5}$$

with (cf. Ref. 8)

$$Y^{(1)} = 1, \tag{1.6a}$$

$$Y^{(3)} = \tfrac{1}{2}\epsilon_0, \tag{1.6b}$$

$$Y^{(5)} = -\tfrac{1}{8}(\epsilon_0^2 + \epsilon_2), \tag{1.6c}$$

$$Y^{(7)} = \tfrac{1}{32}(2\epsilon_0^3 + 6\epsilon_0\epsilon_2 + 5\epsilon_1^2 + \epsilon_4), \tag{1.6d}$$

$$Y^{(9)} = -\tfrac{1}{128}(5\epsilon_0^4 + 30\epsilon_0^2\epsilon_2 + 50\epsilon_0\epsilon_1^2 + 10\epsilon_0\epsilon_4 + 28\epsilon_1\epsilon_3 + 19\epsilon_2^2 + \epsilon_6), \tag{1.6e}$$

where

$$\epsilon_0 = \frac{R(z) - Q^2(z)}{Q^2(z)} + Q^{-3/2}(z)\frac{d^2}{dz^2}Q^{-1/2}(z) \tag{1.7}$$

and

$$\epsilon_\mu = \frac{1}{Q(z)}\frac{d\epsilon_{\mu-1}}{dz}, \quad \mu \geqslant 1. \tag{1.8}$$

 0022-2488/86/112738-10$02.50

For the usefulness and flexibility of the approximation now described it is very important that the base function $Q(z)$ is unspecified and can be chosen in a way appropriate for the particular problem under consideration. For the particular choice $Q(z) = R^{1/2}(z)$ one obtains the special phase-integral approximation of arbitrary order introduced by N. Fröman.[2,3] In the applications in Sec. II we shall see how one can obtain approximations with different regions of validity by choosing the base function in different, appropriate ways.

As regards the notations we remark that the function $R(z)$ in (1.1), (1.2), and (1.7) was denoted by $Q^2(z)$ in Refs. 1–6 and 8, the base function $Q(z)$ in (1.2), (1.5), (1.7), and (1.8) was denoted by $Q_{\text{mod}}(z)$ in Refs. 4–6, and the quantities $Y^{(2n+1)}$ in (1.5) and (1.6a)–(1.6e) were denoted by Y_{2n} in Refs. 2–8. Of these changes of notation the first two are made in order to simplify the writing, while the third is made since in the present paper the notation Y_{2n} is used for the Bessel function of the second kind and of the order $2n$.

To prepare for the applications to be made in Secs. II and III of the present paper we insert (1.5) into (1.4), getting

$$w(z) = \sum_{n=0}^{N} w^{(2n+1)}(z)\lambda^{2n-1}, \tag{1.9}$$

where

$$w^{(2n+1)}(z) = \int^z Y^{(2n+1)}Q(z)dz. \tag{1.10}$$

Inserting (1.5) and (1.9) into (1.3) and recalling (1.6a), we get

$$\begin{aligned}\psi &= \left[\frac{Q(z)}{\lambda}\right]^{-1/2}\left(\sum_{n=0}^{N} Y^{(2n+1)}\lambda^{2n}\right)^{-1/2}\exp\left(\pm\frac{i}{\lambda}\sum_{n=0}^{N} w^{(2n+1)}\lambda^{2n}\right)\\ &= \left[\frac{Q(z)}{\lambda}\right]^{-1/2}\exp\left[\frac{\pm iw^{(1)}(z)}{\lambda}\right]\left(1+\sum_{n=1}^{N} Y^{(2n+1)}\lambda^{2n}\right)^{-1/2}\exp\left(\pm\frac{i}{\lambda}\sum_{n=1}^{N} w^{(2n+1)}\lambda^{2n}\right),\end{aligned} \tag{1.11}$$

where (in the last member) N is assumed to be $\geqslant 1$; the corresponding formula for $N = 0$ is obtained by leaving out the sums over n in (1.11). Using the power series expansion of the exponential function, one obtains for the product of the last two factors on the right-hand side of (1.11) the formula

$$\begin{aligned}&\left(1+\sum_{n=1}^{N} Y^{(2n+1)}\lambda^{2n}\right)^{-1/2}\exp\left(\pm\frac{i}{\lambda}\sum_{n=1}^{N} w^{(2n+1)}\lambda^{2n}\right)\\ &\quad= 1 \pm iw^{(3)}\lambda - \tfrac{1}{2}(Y^{(3)} + [w^{(3)}]^2)\lambda^2 \pm i(w^{(5)} - \tfrac{1}{2}w^{(3)}Y^{(3)} - \tfrac{1}{6}[w^{(3)}]^3)\lambda^3\\ &\qquad - \tfrac{1}{2}(Y^{(5)} - \tfrac{3}{4}[Y^{(3)}]^2 + 2w^{(3)}w^{(5)} - \tfrac{1}{2}[w^{(3)}]^2Y^{(3)} - \tfrac{1}{12}[w^{(3)}]^4)\lambda^4\\ &\qquad \pm i(w^{(7)} - \tfrac{1}{2}w^{(3)}Y^{(5)} - \tfrac{1}{2}w^{(5)}Y^{(3)} + \tfrac{3}{8}w^{(3)}[Y^{(3)}]^2 - \tfrac{1}{2}[w^{(3)}]^2w^{(5)} + \tfrac{1}{12}[w^{(3)}]^3Y^{(3)} + \tfrac{1}{120}[w^{(3)}]^5)\lambda^5\\ &\qquad - \tfrac{1}{2}(Y^{(7)} - \tfrac{3}{2}Y^{(3)}Y^{(5)} + \tfrac{5}{8}[Y^{(3)}]^3 + 2w^{(3)}w^{(7)} + [w^{(5)}]^2 - \tfrac{1}{2}[w^{(3)}]^2Y^{(5)}\\ &\qquad - w^{(3)}w^{(5)}Y^{(3)} + \tfrac{3}{8}[w^{(3)}]^2[Y^{(3)}]^2 - \tfrac{1}{3}[w^{(3)}]^3w^{(5)} + \tfrac{1}{24}[w^{(3)}]^4Y^{(3)} + \tfrac{1}{360}[w^{(3)}]^6)\lambda^6 + \cdots.\end{aligned} \tag{1.12}$$

In this formula, where we have assumed that $N \geqslant 3$, i.e., $2N+1 \geqslant 7$, the simple structure of the left-hand member should be compared to the complicated structure of the right-hand member (which appears in asymptotic expansions of conventional form). Furthermore, the Wronskian of the two linearly independent approximate solutions corresponding to the upper sign and the lower sign in (1.11) is constant, whereas the Wronskian of the corresponding approximate solutions, obtained by using (1.12) with the series in the right-hand member truncated, is in general not constant. If one rewrites the expression in the left-hand member of (1.12) as the exponential function of a power series in λ, that power series (which gives rise to the higher-order terms in the WKB expansion) would also have a complicated structure, and the Wronskian of the corresponding two linearly independent approximate solutions (with the series truncated) would in general not be constant. The simple structure of the phase-integral solutions (1.11) thus stands out in contrast to the complicated structure of the usual asymptotic solutions and of the WKB solutions. This fact illuminates one of the advantages of the phase-integral approximation. One of the other advantages is the presence of the unspecified base function in the phase-integral approximation and the freedom to choose this function conveniently; in the WKB approximation (of higher order) one has no such flexibility.

For higher orders of approximation, i.e., when $2N+1 > 1$, the function $q(z)$ has singularities at the zeros of $Q^2(z)$, and therefore we cannot choose the constant lower limit of integration in the definition (1.4) of $w(z)$ to be a zero of $Q^2(z)$. When $Q^2(z)$ has a simple zero, or more generally a zero of odd order, it is instead convenient to express $w(z)$ by means of a certain contour integral on a two-sheet Riemann surface on which $q(z)$ is single valued. The two sheets of this Riemann surface are cut and joined appropriately along a line emerging from the zero of $Q^2(z)$. Calling this zero t, we thus define

$$w(z) = \frac{1}{2}\int_{\Gamma_t(z)} q(z)dz, \tag{1.13}$$

where $\Gamma_t(z)$ is a contour of integration starting at the point corresponding to z but lying on the Riemann sheet adjacent to the complex z plane under consideration, encircling the point t in the negative or in the positive sense, and ending at

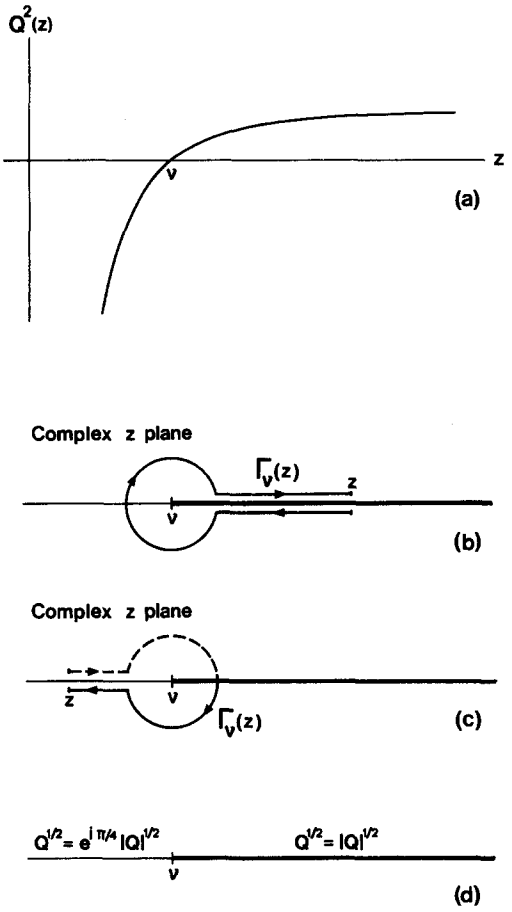


FIG. 1. This figure refers to the case treated in Sec. II B, where the order ν and the argument z are of the same order of magnitude and sufficiently large. (a) When ν is positive, the behavior of the function (2.14), i.e., $Q^2(z) = 1 - \nu^2/z^2$, is shown for real, positive values of z. (b) and (c) Contour of integration $\Gamma_\nu(z)$ when $z > \nu$ [(b)], and when $0 < z < \nu$ [(c)]. The part of the latter contour that lies on the Riemann sheet adjacent to the complex z plane under consideration is indicated by a broken line. The heavy line along the part of the real axis, where $z > \nu$, indicates the cut necessary to make the base function $Q(z)$ single valued. When the contour $\Gamma_\nu(z)$ in (b) is turned through the angle π in the positive sense around the point ν, it goes over into the contour $\Gamma_\nu(z)$ in (c). (d) The phase chosen for $Q^{1/2}(z)$ on the real axis to the left of ν and on the upper lip of the cut along the anti-Stokes line emerging from $z = \nu$ towards the right is indicated.

the point z. Examples of such contours $\Gamma_t(z)$, with the point t ($= \nu$) encircled in the negative sense, are shown in Fig. 1(b) and Fig. 1(c). For the first-order approximation the contour $\Gamma_t(z)$ can be deformed to coincide with a line joining t and z, and hence $w(z)$ is given by the integral (1.4) with t as the lower limit of integration. When $R(z)$ and $Q^2(z)$ are real on the real z axis, the functions $Y^{(2n+1)}$ are real there, and hence, on the real z axis, the function $w(z)$, given by (1.13) with t real, is (for any order of approximation) real in the interval where $Q^2(z)$ is positive (classically allowed region in the generalized sense) but purely imaginary in the interval where $Q^2(z)$ is negative (classically forbidden region in the generalized sense). Inserting (1.5) into (1.13), we obtain (1.9) with

$$w^{(2n+1)}(z) = \frac{1}{2}\int_{\Gamma_t(z)} Y^{(2n+1)}Q(z)dz. \tag{1.14}$$

B. Connection formulas

We shall now consider the particular global problem or connection problem associated with a well-isolated transition point t [i.e., a point t where $Q^2(t) = 0$] on the real axis, when $R(z)$ and $Q^2(z)$ are real on this axis. Such a point t is called a generalized classical turning point. We assume $w(z)$ to be given by (1.13), which is equivalent to (1.9) with (1.14).

The connection formula for tracing (on the real z axis) an approximate solution across the generalized classical turning point t from the classically allowed region to the classically forbidden region is [cf. Eq. (20) in Ref. 3]

$$|q^{-1/2}(z)|\cos[|w(z)| + \gamma - \pi/4] \rightarrow \sin\gamma\,|q^{-1/2}(z)|\exp[|w(z)|], \tag{1.15}$$

where γ is a real constant subjected only to the restriction that $\sin\gamma$ must not be too close to zero.

The connection formula for tracing (on the real z axis) an approximate solution across the turning point t in the opposite direction, i.e., from the classically forbidden region to the classically allowed region is, in a somewhat simplified form [cf. Eqs. (21) and (22) in Ref. 3],

$$|q^{-1/2}(z)|\exp[-|w(z)|] \rightarrow 2|q^{-1/2}(z)|\cos[|w(z)| - \pi/4]. \tag{1.16}$$

We emphasize the one-directional character of the connection formulas (1.15) and (1.16), which means that the tracing of a solution must always be made in the direction of the arrows in (1.15) and (1.16).

The above connection formulas for the phase-integral approximation of arbitrary order, generated from an unspecified base function $Q(z)$, are of the same form as the corresponding, well-known connection formulas for the first-order WKB approximation. Before the former connection formulas had actually been derived, it was, however, far from trivial that this important, simple fact should be true. In this connection we also remark that connection formulas for the WKB approximation of higher order have seldom been treated rigorously (cf., however, Ref. 9), although the corresponding first-order connection formulas have sometimes uncritically been generalized with the hope that such a generalization is immediately possible. It is an important but far from trivial fact that this assumption can be shown to be true.

II. PHASE-INTEGRAL FORMULAS FOR BESSEL FUNCTIONS

In his preface to the Russian translation of Ref. 1, where a rigorous method for handling the connection problems associated with the first-order WKB approximation was developed, the editor of that translation, Professor A. A. Sokolov, remarked that the authors did not include among the applications the highly interesting question of obtaining approximations for special functions of mathematical physics with the aid of the WKB approximation, although the results obtained in Ref. 1 apparently allow this to be done. Later the phase-integral approximation of arbitrary order generated from an unspecified base function was introduced in Refs. 4 and 5. This approximation, which has been briefly

described in Sec. I A of the present paper, is related to the WKB approximation of corresponding order, but it has nicer properties and is in higher order of simpler form; it is also very flexible, since the base function, from which it is generated, is *a priori* unspecified. Furthermore, the connection problems can be handled by means of the method developed in Ref. 1; thus one obtains the connection formulas presented in Sec. I B of the present paper. The applications suggested by Sokolov are still more promising when one uses, instead of the WKB approximation, the arbitrary-order approximation just mentioned, but very few such applications have so far been made. In the present section we shall demonstrate the power of the approach in question for the application to Bessel functions.

The general solution of the differential equation

$$\frac{d^2\psi}{dz^2}+\left(1+\frac{\frac{1}{4}-\nu^2}{z^2}\right)\psi=0 \tag{2.1}$$

is an arbitrary linear combination of the spherical Bessel functions (also called the Riccati–Bessel functions)

$$\psi_1=(\tfrac{1}{2}\pi z)^{1/2}J_\nu(z) \tag{2.2a}$$

and

$$\psi_2=(\tfrac{1}{2}\pi z)^{1/2}Y_\nu(z), \tag{2.2b}$$

$J_\nu(z)$ and $Y_\nu(z)$ being the usual Bessel functions of the first and second kind, respectively.

Using the phase-integral approximation of arbitrary order generated from an unspecified base function, which has been described in Sec. I A, and the connection formulas, which have been presented in Sec. I B, we shall now obtain accurate phase-integral formulas for the Bessel functions. For the sake of simplicity we shall assume ν and z to be positive. The phase-integral formulas to be derived in Sec. II A, which are valid when ν is fixed and z is sufficiently large, are related to the usual asymptotic formulas for the Bessel functions; see Sec. III A. The phase-integral formulas to be derived in Sec. II B, which are valid when ν and z are of the same order of magnitude and sufficiently large, and z does not lie too close to ν, are related to Debye's asymptotic formulas for the Bessel functions; see Sec. III B. In Sec. II C phase-integral formulas are derived for the case of fixed argument z and sufficiently large order ν. These formulas are related to existing but not so well-known asymptotic formulas; see Sec. III C.

A. The case of fixed order and sufficiently large argument

In the present subsection we shall assume ν to be fixed (positive), while z ($>\nu$) is sufficiently large. To account for this assumption, we replace in the differential equation (2.1) z by z/λ, where λ is a "small," positive "bookkeeping" parameter, which is introduced only to indicate in a formal way orders of magnitude, and which will at the end be put equal to unity. Thus we obtain from (2.1) the differential equation

$$\frac{d^2\psi}{dz^2}+\left(\frac{1}{\lambda^2}+\frac{\frac{1}{4}-\nu^2}{z^2}\right)\psi=0, \tag{2.3}$$

which reduces to (2.1) when λ is put equal to unity. Choosing

$$Q^2(z)=1 \tag{2.4}$$

and

$$R(z)-Q^2(z)=(\tfrac{1}{4}-\nu^2)/z^2 \tag{2.5}$$

in (1.2), we obtain the differential equation (2.3). Inserting (2.4) and (2.5) into (1.7), we get $\epsilon_0=(\frac{1}{4}-\nu^2)/z^2$. Using this expression for ϵ_0, choosing in agreement with (2.4) the base function to be

$$Q(z)=1, \tag{2.6}$$

and recalling the definition (1.8), we obtain

$$\epsilon_\mu=\frac{(-1)^\mu(\mu+1)!(\frac{1}{4}-\nu^2)}{z^{\mu+2}},\quad \mu\geqslant 0. \tag{2.7}$$

Inserting (2.7) into (1.6a)–(1.6e), we get

$$Y^{(2n+1)}=a_{2n}z^{-2n},\quad n\geqslant 0, \tag{2.8}$$

where

$$a_0=1, \tag{2.9a}$$

$$a_2=\tfrac{1}{2}(\tfrac{1}{4}-\nu^2), \tag{2.9b}$$

$$a_4=-\tfrac{1}{8}((\tfrac{1}{4}-\nu^2)^2+6(\tfrac{1}{4}-\nu^2)), \tag{2.9c}$$

$$a_6=\tfrac{1}{16}((\tfrac{1}{4}-\nu^2)^3+28(\tfrac{1}{4}-\nu^2)^2+60(\tfrac{1}{4}-\nu^2)), \tag{2.9d}$$

$$a_8=-\tfrac{1}{128}(5(\tfrac{1}{4}-\nu^2)^4+380(\tfrac{1}{4}-\nu^2)^3+3228(\tfrac{1}{4}-\nu^2)^2+5040(\tfrac{1}{4}-\nu^2)). \tag{2.9e}$$

Inserting (2.6) and (2.8) into (1.10) with the constant lower limit of integration chosen conveniently (different for $n=0$ and $n>0$), we obtain [cf. (2.9a)]

$$w^{(1)}(z)=a_0z=z, \tag{2.10a}$$

$$w^{(2n+1)}(z)=[a_{2n}/(1-2n)]z^{1-2n},\quad n\geqslant 1. \tag{2.10b}$$

Recalling (2.2a) and the fact (cf. Ref. 10, p. 199) that $(\pi z/2)^{1/2}J_\nu(z)$ tends to $\cos(z-\nu\pi/2-\pi/4)$ as $z\to+\infty$ (for fixed ν), we obtain with due regard to (1.11) and (2.6), when λ is finally put equal to unity, the phase-integral formula of the $(2N+1)$th-order approximation

$$\begin{aligned}J_\nu(z)&=\left(\frac{2}{\pi z}\right)^{1/2}\left(\sum_{n=0}^{N}Y^{(2n+1)}\right)^{-1/2}\\&\quad\times\cos\left(\sum_{n=0}^{N}w^{(2n+1)}(z)-\left(\nu+\frac{1}{2}\right)\frac{\pi}{2}\right)\\&=\left(\frac{2}{\pi z}\right)^{1/2}\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\\&\quad\times\cos\left(z-\left(\nu+\frac{1}{2}\right)\frac{\pi}{2}+\sum_{n=1}^{N}w^{(2n+1)}(z)\right),\\&\quad z\gg\nu,\end{aligned} \tag{2.11}$$

where $Y^{(2n+1)}$ and $w^{(2n+1)}$ are given by (2.8) and (2.10a) and (2.10b), respectively. Recalling (2.2b) and (cf. Ref. 10, p. 199) the behavior of $(\pi z/2)^{1/2}Y_\nu(z)$ as $z\to+\infty$ (for fixed ν), we similarly obtain a phase-integral formula for $Y_\nu(z)$ for fixed ν and sufficiently large z:

$$\begin{aligned}Y_\nu(z)&=\left(\frac{2}{\pi z}\right)^{1/2}\left(\sum_{n=0}^{N}Y^{(2n+1)}\right)^{-1/2}\\&\quad\times\sin\left(\sum_{n=0}^{N}w^{(2n+1)}(z)-\left(\nu+\frac{1}{2}\right)\frac{\pi}{2}\right)\end{aligned}$$

$$= \left(\frac{2}{\pi z}\right)^{1/2} \left(1 + \sum_{n=1}^{N} Y^{(2n+1)}\right)^{-1/2} \times \sin\left(z - \left(\nu + \frac{1}{2}\right)\frac{\pi}{2} + \sum_{n=1}^{N} w^{(2n+1)}(z)\right), \quad z \gg \nu. \tag{2.12}$$

One obtains this formula from (2.11) by replacing there $J_\nu(z)$ by $Y_\nu(z)$ and cos by sin. In writing down the last members of (2.11) and (2.12) we have assumed that $N \geqslant 1$. One obtains the corresponding formulas for $N = 0$ by deleting the sums over n in the last members of (2.11) and (2.12).

B. The case when the argument and the order are both sufficiently large

In the present subsection we shall assume ν and z to be sufficiently large (positive) and of comparable order of magnitude. To account for this assumption, we replace in the differential equation (2.1) ν by ν/λ and z by z/λ, where λ is a "small," positive parameter that will at the end be put equal to unity. Thus we obtain from (2.1) the differential equation

$$\frac{d^2\psi}{dz^2} + \left(\frac{1 - \nu^2/z^2}{\lambda^2} + \frac{1}{4z^2}\right)\psi = 0, \tag{2.13}$$

which reduces to (2.1) when λ is put equal to unity. Choosing

$$Q^2(z) = 1 - \nu^2/z^2 \tag{2.14}$$

and

$$R(z) - Q^2(z) = 1/(4z^2) \tag{2.15}$$

in (1.2), we obtain the differential equation (2.13). When the functions in (1.2) are given by (2.14) and (2.15) with $\nu \neq 0$, the phase-integral approximation remains valid in the neighborhood of the origin $z = 0$ but breaks down in the neighborhood of the generalized turning point $z = \nu$. Inserting (2.14) and (2.15) into (1.7), we get

$$\epsilon_0 = \frac{1}{(1 - \nu^2/z^2)^3}\left(\frac{1}{4z^2} + \frac{\nu^2}{z^4}\right). \tag{2.16a}$$

From (1.8), (2.14), and (2.16a) we then obtain

$$\epsilon_1 = -\frac{1}{(1 - \nu^2/z^2)^{9/2}}\left(\frac{1}{2z^3} + \frac{5\nu^2}{z^5} + \frac{2\nu^4}{z^7}\right), \tag{2.16b}$$

$$\epsilon_2 = \frac{1}{(1 - \nu^2/z^2)^6}\left(\frac{3}{2z^4} + \frac{28\nu^2}{z^6} + \frac{34\nu^4}{z^8} + \frac{4\nu^6}{z^{10}}\right), \tag{2.16c}$$

$$\epsilon_3 = -\frac{1}{(1 - \nu^2/z^2)^{15/2}}\left(\frac{6}{z^5} + \frac{180\nu^2}{z^7} + \frac{440\nu^4}{z^9} + \frac{176\nu^6}{z^{11}} + \frac{8\nu^8}{z^{13}}\right), \tag{2.16d}$$

$$\epsilon_4 = \frac{1}{(1 - \nu^2/z^2)^9}\left(\frac{30}{z^6} + \frac{1320\nu^2}{z^8} + \frac{5400\nu^4}{z^{10}} + \frac{4576\nu^6}{z^{12}} + \frac{808\nu^8}{z^{14}} + \frac{16\nu^{10}}{z^{16}}\right), \tag{2.16e}$$

where the phase of the square root of $1 - \nu^2/z^2$ is to be chosen in agreement with that of $Q(z)$; see Fig. 1(d). Inserting (2.16a)–(2.16e) into (1.6a)–(1.6d), we get

$$Y^{(1)} = 1, \tag{2.17a}$$

$$Y^{(3)} = \pm\frac{1}{(\pm 1 \mp \nu^2/z^2)^3}\left(\frac{1}{8z^2} + \frac{\nu^2}{2z^4}\right), \tag{2.17b}$$

$$Y^{(5)} = -\frac{1}{(\pm 1 \mp \nu^2/z^2)^6}\left(\frac{25}{128z^4} + \frac{57\nu^2}{16z^6} + \frac{35\nu^4}{8z^8} + \frac{\nu^6}{2z^{10}}\right), \tag{2.17c}$$

$$Y^{(7)} = \pm\frac{1}{(\pm 1 \mp \nu^2/z^2)^9}\left(\frac{1073}{1024z^6} + \frac{11171\nu^2}{256z^8} + \frac{11511\nu^4}{64z^{10}} + \frac{611\nu^6}{4z^{12}} + \frac{213\nu^8}{8z^{14}} + \frac{\nu^{10}}{2z^{16}}\right), \tag{2.17d}$$

where it is convenient to use the upper signs when z is real, positive, and larger than ν and to use the lower signs when z is real, positive, and smaller than ν.

Consulting Fig. 1(d) as regards the phase of the base function $Q(z)$, we obtain from (2.14)

$$Q(z) = \begin{cases} (1 - \nu^2/z^2)^{1/2}, & z > \nu, \\ i(\nu^2/z^2 - 1)^{1/2}, & 0 < z < \nu, \end{cases} \tag{2.18}$$

the square roots in the right-hand member of (2.18) being positive. For $w(z)$ it is, in the present subsection, convenient to use (1.9) with (1.14) for $t = \nu$. Inserting (2.17a)–(2.17d) and (2.18) into (1.14), we obtain

$$w^{(1)}(z) = (z^2 - \nu^2)^{1/2} - \nu \arccos(\nu/z) \ (>0), \quad z > \nu,$$

$$iw^{(1)}(z) = \nu \ln\left[\frac{\nu}{z} + \left(\frac{\nu^2}{z^2} - 1\right)^{1/2}\right] - (\nu^2 - z^2)^{1/2} \ (>0), \quad 0 < z < \nu, \tag{2.19a}$$

$$\left.\begin{matrix} w^{(3)}(z) \\ iw^{(3)}(z) \end{matrix}\right\} = \mp\frac{1}{(\pm 1 \mp \nu^2/z^2)^{3/2}}\left(\frac{1}{8z} + \frac{\nu^2}{12z^3}\right), \tag{2.19b}$$

$$\left.\begin{matrix} w^{(5)}(z) \\ iw^{(5)}(z) \end{matrix}\right\} = \frac{1}{(\pm 1 \mp \nu^2/z^2)^{9/2}}\left(\frac{25}{384z^3} + \frac{203\nu^2}{320z^5} + \frac{21\nu^4}{80z^7} - \frac{\nu^6}{360z^9}\right), \tag{2.19c}$$

$$\left.\begin{matrix} w^{(7)}(z) \\ iw^{(7)}(z) \end{matrix}\right\} = \mp\frac{1}{(\pm 1 \mp \nu^2/z^2)^{15/2}}\left(\frac{1073}{5120z^5} + \frac{21\,269\nu^2}{3584z^7} + \frac{14827\nu^4}{1008z^9} + \frac{985\nu^6}{168z^{11}} + \frac{41\nu^8}{168z^{13}} + \frac{\nu^{10}}{1260z^{15}}\right), \tag{2.19d}$$

where the upper and lower expressions in the left-hand members and the upper and lower signs in the right-hand members in (2.19b)–(2.19d) apply for $z > \nu$ and $0 < z < \nu$, respectively.

Introducing the positive quantity c by the definition

$$c = \begin{cases} (z^2/\nu^2 - 1)^{-1/2}, & \text{for } z > \nu, \\ (1 - z^2/\nu^2)^{-1/2}, & \text{for } 0 < z < \nu, \end{cases} \tag{2.20}$$

we obtain from (2.18)

$$Q(z) = \begin{cases} (c^2 + 1)^{-1/2}, & \text{for } z > \nu, \\ i(c^2 - 1)^{-1/2}, & \text{for } 0 < z < \nu, \end{cases} \tag{2.21}$$

and we can rewrite (2.17a)–(2.17d) and (2.19a)–(2.19d) into the following alternative forms.

For $z > \nu$,

$$Y^{(1)} = 1, \tag{2.22a}$$

$$Y^{(3)} = (c^2/8\nu^2)(1 + 6c^2 + 5c^4), \tag{2.22b}$$

$$Y^{(5)} = -(c^4/128\nu^4)(25 + 556c^2 + 2078c^4 + 2652c^6 + 1105c^8), \quad (2.22c)$$

$$Y^{(7)} = (c^6/1024\nu^6)(1073 + 51\,122c^2 + 423\,691c^4 + 1361\,420c^6 + 2064\,503c^8 + 1490\,850c^{10} + 414\,125c^{12}); \quad (2.22d)$$

$$w^{(1)} = \nu(1/c - \operatorname{arccot} c) \ (>0), \quad (2.23a)$$

$$w^{(3)} = -(c/\nu)(\tfrac{1}{8} + \tfrac{5}{24}c^2), \quad (2.23b)$$

$$w^{(5)} = (c^3/\nu^3)(\tfrac{25}{384} + \tfrac{531}{640}c^2 + \tfrac{221}{128}c^4 + \tfrac{1105}{1152}c^6), \quad (2.23c)$$

$$w^{(7)} = -(c^5/\nu^5)(\tfrac{1073}{5120} + \tfrac{50049}{7168}c^2 + \tfrac{186821}{4608}c^4 + \tfrac{44899}{512}c^6 + \tfrac{82825}{1024}c^8 + \tfrac{82825}{3072}c^{10}). \quad (2.23d)$$

For $0<z<\nu$,

$$Y^{(1)} = 1, \quad (2.24a)$$

$$Y^{(3)} = -(c^2/8\nu^2)(1 - 6c^2 + 5c^4), \quad (2.24b)$$

$$Y^{(5)} = -(c^4/128\nu^4)(25 - 556c^2 + 2078c^4 - 2652c^6 + 1105c^8), \quad (2.24c)$$

$$Y^{(7)} = -(c^6/1024\nu^6)(1073 - 51\,122c^2 + 423\,691c^4 - 1361\,420c^6 + 2064\,503c^8 - 1490\,850c^{10} + 414\,125c^{12}); \quad (2.24d)$$

$$iw^{(1)} = -\nu\left(\frac{1}{c} - \frac{1}{2}\ln\frac{c+1}{c-1}\right) \ (>0), \quad (2.25a)$$

$$iw^{(3)} = -(c/\nu)(\tfrac{1}{8} - \tfrac{5}{24}c^2), \quad (2.25b)$$

$$iw^{(5)} = -(c^3/\nu^3)(\tfrac{25}{384} - \tfrac{531}{640}c^2 + \tfrac{221}{128}c^4 - \tfrac{1105}{1152}c^6), \quad (2.25c)$$

$$iw^{(7)} = -(c^5/\nu^5)(\tfrac{1073}{5120} - \tfrac{50049}{7168}c^2 + \tfrac{186821}{4608}c^4 - \tfrac{44899}{512}c^6 + \tfrac{82825}{1024}c^8 - \tfrac{82825}{3072}c^{10}). \quad (2.25d)$$

It is easily seen that (2.24a)–(2.24d) and (2.25a)–(2.25d) are obtained from (2.22a)–(2.22d) and (2.23a)–(2.23d), respectively, by the replacement of c by $-ic$. We also note that, when c is kept fixed, $Y^{(2n+1)}$ is proportional to ν^{-2n}, and $w^{(2n+1)}$ is proportional to ν^{1-2n}.

Recalling how $J_\nu(z)$ behaves when z is close to zero, we realize that in the classically forbidden region between the origin and the turning point ν the function (2.2a) is given by the approximate formula

$$\psi_1 = \tfrac{1}{2}C|q^{-1/2}(z)|\exp[-|w(z)|], \quad 0<z<\nu, \quad (2.26)$$

where $w(z)$ is given by (1.13) with $t=\nu$, λ is to be put equal to unity, and C is a not yet determined quantity, which is independent of z. Since the function ψ_1 defined by (2.2a) is positive immediately to the right of $z=0$ when ν is positive (which follows from its power series expansion), C is positive. With the aid of the connection formula (1.16) we get the following approximate expression for ψ_1 on the real axis to the right of the turning point ν:

$$\psi_1 = C|q^{-1/2}(z)|\cos[|w(z)| - \pi/4], \quad z>\nu. \quad (2.27)$$

From the facts that the function ψ_1, defined by (2.2a), has the amplitude $(1-\nu^2/z^2)^{-1/4}$ for large positive values of z [see Eq. (7) on p. 229 in Ref. 10], that ν is assumed to be sufficiently large, and that the constant C is positive, we obtain the aid of (2.27) with (1.5), (2.18), (2.20), and (2.22a)–(2.22d) for $\lambda=1$

$$C = 1. \quad (2.28)$$

The function ψ_2, defined by (2.2b), has also the amplitude $(1-\nu^2/z^2)^{-1/4}$ for large positive values of z but has the asymptotic phase shifted by $-\pi/2$ from that of ψ_1; see Eq. (7) on p. 229 in Ref. 10. Hence we obtain from (2.27) and (2.28) the approximate formula

$$\psi_2 = |q^{-1/2}(z)|\cos[|w(z)| - 3\pi/4], \quad z>\nu. \quad (2.29)$$

Using the connection formula (1.15), we obtain from (2.29) the approximate formula

$$\psi_2 = -|q^{-1/2}(z)|\exp[|w(z)|], \quad 0<z<\nu. \quad (2.30)$$

Recalling (1.5) and (1.9) with $\lambda=1$, and taking the phases of $Q(z)$ and $w^{(1)}(z)$ into account, we obtain from (2.26), (2.27), (2.28), (2.29), and (2.30) approximate formulas for ψ_1 and ψ_2, which, with the aid of (2.2a) and (2.2b), yield

$$J_\nu(z) = \left(\frac{2}{\pi z}\right)^{1/2}\left(Q(z)\sum_{n=0}^{N}Y^{(2n+1)}\right)^{-1/2} \times\cos\left(\sum_{n=0}^{N}w^{(2n+1)}(z) - \frac{\pi}{4}\right), \quad z>\nu, \quad (2.31a)$$

$$J_\nu(z) = \frac{1}{2}\left(\frac{2}{\pi z}\right)^{1/2}\left(-iQ(z)\sum_{n=0}^{N}Y^{(2n+1)}\right)^{-1/2} \times\exp\left(-\sum_{n=0}^{N}iw^{(2n+1)}(z)\right), \quad 0<z<\nu, \quad (2.31b)$$

$$Y_\nu(z) = \left(\frac{2}{\pi z}\right)^{1/2}\left(Q(z)\sum_{n=0}^{N}Y^{(2n+1)}\right)^{-1/2} \times\sin\left(\sum_{n=0}^{N}w^{(2n+1)}(z) - \frac{\pi}{4}\right), \quad z>\nu, \quad (2.32a)$$

$$Y_\nu(z) = -\left(\frac{2}{\pi z}\right)^{1/2}\left(-iQ(z)\sum_{n=0}^{N}Y^{(2n+1)}\right)^{-1/2} \times\exp\left(\sum_{n=0}^{N}iw^{(2n+1)}(z)\right), \quad 0<z<\nu, \quad (2.32b)$$

The quantities $Q(z)$, $Y^{(2n+1)}$, and $w^{(2n+1)}$ in (2.31a), (2.31b), (2.32a), and (2.32b) are given by (2.18), (2.17a)–(2.17d), and (2.19a)–(2.19d), respectively. Alternative expressions, in terms of the positive quantity c defined by (2.20), for $Y^{(2n+1)}$ and $w^{(2n+1)}$ are (2.22a)–(2.22d) and (2.23a)–(2.23d) when $z>\nu$ and (2.24a)–(2.24d) and (2.25a)–(2.25d) when $0<z<\nu$; see also (2.21).

We have thus, by means of the phase-integral approximation generated from a conveniently chosen base function, derived approximate formulas for Bessel functions when their order ν and their argument z are both sufficiently large and of the same order of magnitude. The formulas are, however, not valid when z lies too close to ν.

Recalling that in the present subsection z and ν are both assumed to be sufficiently large, we now introduce the further assumption that $z \gg \nu$. Assuming that $N>0$, we then expand $Q(z)\,\Sigma_{n=0}^{N}Y^{(2n+1)}$ and $\Sigma_{n=0}^{N}w^{(2n+1)}(z)$ in powers of ν/z, keeping in the former function powers up to $2N$ and in the latter function powers up to $2N-1$. The formulas one obtains by inserting these truncated power series expansions

into (2.31a) and (2.32a) are identical to (2.11) and (2.12), respectively, with $Y^{(2n+1)}$ and $w^{(2n+1)}(z)$ given by (2.8) and by (2.10a) and (2.10b). When z and ν are both sufficiently large, and furthermore, $z \gg \nu$, (2.31a) and (2.32a) should be more accurate than (2.11) and (2.12), and this expectation is confirmed by numerical tests.

C. The case of fixed argument and sufficiently large order

In the present subsection we shall assume z (positive) to be fixed, while ν ($>z$) is sufficiently large. To account for this assumption, we replace in the differential equation (2.1) ν by ν/λ, where λ is a "small," positive parameter, which will at the end be put equal to unity. Thus we obtain from (2.1) the differential equation

$$\frac{d^2\psi}{dz^2} + \left(-\frac{\nu^2/z^2}{\lambda^2} + 1 + \frac{1}{4z^2} \right)\psi = 0, \tag{2.33}$$

which reduces to (2.1) when λ is put equal to unity. Choosing

$$Q^2(z) = -\nu^2/z^2 \tag{2.34}$$

and

$$R(z) - Q^2(z) = 1 + 1/(4z^2) \tag{2.35}$$

in (1.2), we obtain the differential equation (2.33). When the functions in (1.2) are given by (2.34) and (2.35) with $\nu \neq 0$, the phase-integral approximation remains valid in the neighborhood of the origin $z = 0$. Inserting (2.34) and (2.35) into (1.7), we get $\epsilon_0 = -z^2/\nu^2$. Using this expression for ϵ_0, choosing in agreement with (2.34) the base function to be

$$Q(z) = i\nu/z, \tag{2.36}$$

and recalling the definition (1.8), we obtain

$$\epsilon_\mu = 2^\mu z^2/(i\nu)^{\mu+2}, \quad \mu \geqslant 0. \tag{2.37}$$

Inserting (2.37) into (1.6a)–(1.6e), we get

$$Y^{(1)} = 1, \tag{2.38a}$$

$$Y^{(3)} = -\frac{z^2}{2\nu^2}, \tag{2.38b}$$

$$Y^{(5)} = -\frac{4z^2 + z^4}{8\nu^4}, \tag{2.38c}$$

$$Y^{(7)} = -\frac{8z^2 + 22z^4 + z^6}{16\nu^6}, \tag{2.38d}$$

$$Y^{(9)} = -\frac{64z^2 + 912z^4 + 320z^6 + 5z^8}{128\nu^8}. \tag{2.38e}$$

Inserting (2.36) and (2.38a)–(2.38e) into (1.10), where the integration constant is chosen conveniently (different for $n = 0$ and $n > 0$), we obtain

$$iw^{(1)} = \ln(1/z^\nu), \tag{2.39a}$$

$$iw^{(3)} = \frac{z^2}{4\nu}, \tag{2.39b}$$

$$iw^{(5)} = \frac{8z^2 + z^4}{32\nu^3}, \tag{2.39c}$$

$$iw^{(7)} = \frac{24z^2 + 33z^4 + z^6}{96\nu^5}, \tag{2.39d}$$

$$iw^{(9)} = \frac{768z^2 + 5472z^4 + 1280z^6 + 15z^8}{3072\nu^7}. \tag{2.39e}$$

According to (2.38b)–(2.38e) and (2.39b)–(2.39e) $Y^{(2n+1)}$ and $w^{(2n+1)}$ tend to zero as z tends to zero when $n > 0$. Recalling (1.11), (2.2a) and (2.2b), and the behaviors of $J_\nu(z)$ and $Y_\nu(z)$ for sufficiently small values of z, and using (2.36) and (2.39a), we therefore obtain (for $N \geqslant 1$) the phase-integral formulas

$$J_\nu(z) = \frac{z^\nu}{2^\nu \Gamma(\nu+1)} \left(1 + \sum_{n=1}^{N} Y^{(2n+1)} \right)^{-1/2} \times \exp\left(\sum_{n=1}^{N} -iw^{(2n+1)}(z) \right), \tag{2.40a}$$

$$Y_\nu(z) = -\frac{2^\nu \Gamma(\nu)}{\pi z^\nu} \left(1 + \sum_{n=1}^{N} Y^{(2n+1)} \right)^{-1/2} \times \exp\left(\sum_{n=1}^{N} +iw^{(2n+1)}(z) \right), \tag{2.40b}$$

where $Y^{(2n+1)}$ and $iw^{(2n+1)}(z)$ are to be obtained from (2.38b)–(2.38e) and (2.39b)–(2.39e). In writing (2.40a) and (2.40b) we have assumed that $N \geqslant 1$. The corresponding formulas for $N = 0$ are obtained from (2.40a) and (2.40b) by leaving out the sums over n ($\geqslant 1$). The additive part of $Y_\nu(z)$, which, for integer values of ν, has a branch point at $z = 0$, does not appear in (2.40b), since the corresponding contribution to $Y_\nu(z)$ is not significant in that approximate formula.

III. ASYMPTOTIC FORMULAS OF CONVENTIONAL FORM FOR BESSEL FUNCTIONS

In the present section we shall use (1.12) to rewrite the phase-integral formulas obtained in Sec.II into asymptotic formulas of conventional form.

A. The case of fixed order and sufficiently large argument

From (2.11) and (2.12) we easily obtain the formulas

$$J_\nu(z) = \left(\frac{2}{\pi z}\right)^{1/2} \left[\left(1 + \sum_{n=1}^{N} Y^{(2n+1)} \right)^{-1/2} \cos\left(\sum_{n=1}^{N} w^{(2n+1)}(z) \right) \cos\left(z - \frac{(\nu + \frac{1}{2})\pi}{2} \right) - \left(1 + \sum_{n=1}^{N} Y^{(2n+1)} \right)^{-1/2} \sin\left(\sum_{n=1}^{N} w^{(2n+1)}(z) \right) \sin\left(z - \frac{(\nu + \frac{1}{2})\pi}{2} \right) \right], \tag{3.1}$$

$$Y_\nu(z)=\left(\frac{2}{\pi z}\right)^{1/2}\left[\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\sin\left(\sum_{n=1}^{N}w^{(2n+1)}(z)\right)\cos\left(z-\frac{(\nu+\frac{1}{2})\pi}{2}\right)\right.$$
$$\left.+\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\cos\left(\sum_{n=1}^{N}w^{(2n+1)}(z)\right)\sin\left(z-\frac{(\nu+\frac{1}{2})\pi}{2}\right)\right]. \tag{3.2}$$

Inserting (2.8) and (2.10b) with (2.9a)–(2.9c) into (1.12) with $\lambda=1$, we obtain for $N\geqslant 2$

$$\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\exp\left(\pm i\sum_{n=1}^{N}w^{(2n+1)}\right)$$
$$=1\pm\frac{i}{2z}\frac{\Gamma(\nu+1+\frac{1}{2})}{1!\Gamma(\nu-1+\frac{1}{2})}-\frac{1}{(2z)^2}\frac{\Gamma(\nu+2+\frac{1}{2})}{2!\Gamma(\nu-2+\frac{1}{2})}\mp\frac{i}{(2z)^3}\frac{\Gamma(\nu+3+\frac{1}{2})}{3!\Gamma(\nu-3+\frac{1}{2})}+\frac{1}{(2z)^4}\frac{\Gamma(\nu+4+\frac{1}{2})}{4!\Gamma(\nu-4+\frac{1}{2})}\pm\cdots. \tag{3.3}$$

Using the equations one obtains by taking the real and imaginary parts of (3.3), one can rewrite (3.1) and (3.2) to obtain the usual asymptotic expansions for $J_\nu(z)$ and $Y_\nu(z)$ for fixed order ν and large argument z (see, e.g., p. 199 in Ref. 10):

$$J_\nu(z)=\left(\frac{2}{\pi z}\right)^{1/2}\left[\cos\left(z-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\left(1-\frac{\Gamma(\nu+2+\frac{1}{2})}{2!\Gamma(\nu-2+\frac{1}{2})}\frac{1}{(2z)^2}+\frac{\Gamma(\nu+4+\frac{1}{2})}{4!\Gamma(\nu-4+\frac{1}{2})}\frac{1}{(2z)^4}-\cdots\right)\right.$$
$$\left.-\sin\left(z-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\left(\frac{\Gamma(\nu+1+\frac{1}{2})}{1!\Gamma(\nu-1+\frac{1}{2})}\frac{1}{2z}-\frac{\Gamma(\nu+3+\frac{1}{2})}{3!\Gamma(\nu-3+\frac{1}{2})}\frac{1}{(2z)^3}+\cdots\right)\right], \tag{3.4}$$

$$Y_\nu(z)=\left(\frac{2}{\pi z}\right)^{1/2}\left[\sin\left(z-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\left(1-\frac{\Gamma(\nu+2+\frac{1}{2})}{2!\Gamma(\nu-2+\frac{1}{2})}\frac{1}{(2z)^2}+\frac{\Gamma(\nu+4+\frac{1}{2})}{4!\Gamma(\nu-4+\frac{1}{2})}\frac{1}{(2z)^4}-\cdots\right)\right.$$
$$\left.+\cos\left(z-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\left(\frac{\Gamma(\nu+1+\frac{1}{2})}{1!\Gamma(\nu-1+\frac{1}{2})}\frac{1}{2z}-\frac{\Gamma(\nu+3+\frac{1}{2})}{3!\Gamma(\nu-3+\frac{1}{2})}\frac{1}{(2z)^3}+\cdots\right)\right]. \tag{3.5}$$

B. The case when the argument and the order are both sufficiently large (Debye's asymptotic formulas)

Debye's asymptotic formulas for the Bessel functions were originally given in Ref. 11 and can be found, for instance, also on pp. 241–245 in Ref. 10, on p. XXXV in Ref. 12, and on pp. 130–134 and 382 in Ref. 13. It is also worth mentioning that in 1817, i.e., almost a century before Debye,[11] Carlini[14] (see also Ref. 15) derived an approximate formula that, when expressed in terms of the Bessel function $J_\nu(\xi)$, when ξ ($<\nu$) is proportional to ν, is essentially equivalent to the next lowest order of the corresponding asymptotic formula derived by Debye[11] in a quite different way; see Ref. 16.

A new, alternative derivation of Debye's asymptotic formulas is achieved by appropriate series expansion of the phase-integral formulas (2.31a) and (2.31b) and (2.32a) and (2.32b). To demonstrate this alternative way of deriving Debye's formula for $J_\nu(z)$ when $z>\nu$, we start from (2.31a) and obtain, when $N\geqslant 1$, the approximate formula

$$J_\nu(z)=\left(\frac{2}{\pi z}\right)^{1/2}Q^{-1/2}(z)\left[\cos\left(w^{(1)}(z)-\frac{\pi}{4}\right)\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\cos\left(\sum_{n=1}^{N}w^{(2n+1)}(z)\right)\right.$$
$$\left.-\sin\left(w^{(1)}(z)-\frac{\pi}{4}\right)\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\sin\left(\sum_{n=1}^{N}w^{(2n+1)}(z)\right)\right],\quad z>\nu. \tag{3.6}$$

Inserting (2.22b)–(2.22d) and (2.23b)–(2.23d) into (1.12) with $\lambda=1$, we obtain (when $2N+1\geqslant 7$)

$$\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\exp\left(\pm i\sum_{n=1}^{N}w^{(2n+1)}\right)$$
$$=1\mp i\frac{c}{24\nu}(3+5c^2)-\frac{1}{2}\left(\frac{c}{24\nu}\right)^2(81+462c^2+385c^4)\pm\frac{i}{30}\left(\frac{c}{24\nu}\right)^3(30\,375+369\,603c^2+765\,765c^4+425\,425c^6)$$
$$+\frac{1}{120}\left(\frac{c}{24\nu}\right)^4(4465\,125+94\,121\,676c^2+349\,922\,430c^4+446\,185\,740c^6+185\,910\,725c^8)$$
$$\mp\frac{i}{840}\left(\frac{c}{24\nu}\right)^5(1519\,035\,525+49\,286\,948\,607c^2+284\,499\,769\,554c^4+614\,135\,872\,350c^6$$
$$+566\,098\,157\,625c^8+188\,699\,385\,875c^{10})$$
$$-\frac{1}{25200}\left(\frac{c}{24\nu}\right)^6(2757\,049\,477\,875+127\,577\,298\,354\,750c^2+1050\,760\,774\,457\,901c^4$$
$$+3369\,032\,068\,261\,860c^6+5104\,696\,716\,244\,125c^8$$
$$+3685\,299\,006\,138\,750c^{10}+1023\,694\,168\,371\,875c^{12})+\cdots. \tag{3.7}$$

Taking the real and imaginary parts of (3.7), and inserting the resulting two formulas into (3.6), where now $2N+1$ is assumed to be $\geqslant 7$, and using also (2.20), (2.21), and (2.23a), we obtain

$$J_\nu(z)=\left(\frac{2c}{\pi\nu}\right)^{1/2}\left\{\left[1-\frac{1}{2}\left(\frac{c}{24\nu}\right)^2(81+462c^2+385c^4)\right.\right.$$
$$+\frac{1}{120}\left(\frac{c}{24\nu}\right)^4(4465\,125+94\,121\,676c^2+349\,922\,430c^4+446\,185\,740c^6+185\,910\,725c^8)$$
$$-\frac{1}{25200}\left(\frac{c}{24\nu}\right)^6(2757\,049\,477\,875+127\,577\,298\,354\,750c^2+1050\,760\,774\,457\,901c^4+3369\,032\,068\,261\,860c^6$$
$$\left.+5104\,696\,716\,244\,125c^8+3685\,299\,006\,138\,750c^{10}+1023\,694\,168\,371\,875c^{12})+\cdots\right]\cos\left[\nu\left(\frac{1}{c}-\operatorname{arccot}c\right)-\frac{\pi}{4}\right]$$
$$+\left[\frac{c}{24\nu}(3+5c^2)-\frac{1}{30}\left(\frac{c}{24\nu}\right)^3(30\,375+369\,603c^2+765\,765c^4+425\,425c^6)\right.$$
$$+\frac{1}{840}\left(\frac{c}{24\nu}\right)^5(1519\,035\,525+49\,286\,948\,607c^2+284\,499\,769\,554c^4+614\,135\,872\,350c^6$$
$$\left.\left.+566\,098\,157\,625c^8+188\,699\,385\,875c^{10})+\cdots\right]\sin\left[\nu\left(\frac{1}{c}-\operatorname{arccot}c\right)-\frac{\pi}{4}\right]\right\},\quad z>\nu. \tag{3.8}$$

This formula is equivalent to the formula (39a) with the definitions (38) and (39c) in Ref. 12. Since the asymptotic expansion of $Y_\nu(z)$ for large positive values of ν and z ($>\nu$) differs from that of $J_\nu(z)$ only in the shift of phase by $-\pi/2$ [see Eq. (7) on p. 229 in Ref. 10], one easily obtains from (3.8) an approximate formula for $Y_\nu(z)$ when $z>\nu$; this formula is equivalent to the formula (39b) with the definitions (38) and (39c) in Ref. 12. Analogously as we have for $z>\nu$ obtained (3.8) from (2.31a), we can for $0<z<\nu$ obtain corresponding approximate formulas for $J_\nu(z)$ from (2.31b) and for $Y_\nu(z)$ from (2.32b); these formulas are equivalent to the formulas (41a) and (41b), respectively, with the definitions (39c) and (40) in Ref. 12.

The above derivation of Debye's asymptotic formulas is much simpler than the derivation by means of the method of steepest descents. In fact, it is a herculean task to obtain further terms in the expansion (3.8) with the method of steepest descents, whereas the amount of work is reasonable with the above derivation, if a desk calculator is used to obtain the numerical coefficients.

C. The case of fixed argument of sufficiently large order

Inserting (2.38b) and (2.38c) and (2.39b) and (2.39c) into (1.12) with $\lambda=1$, we obtain for $N\geqslant 2$

$$\left(1+\sum_{n=1}^{N}Y^{(2n+1)}\right)^{-1/2}\exp\left(\pm i\sum_{n=1}^{N}w^{(2n+1)}\right)$$
$$=1\pm\frac{z^2}{4\nu}+\frac{8z^2+z^4}{32\nu^2}\pm\frac{96z^2+36z^4+z^6}{384\nu^3}+\frac{1536z^2+1344z^4+96z^6+z^8}{6144\nu^4}\pm\cdots. \tag{3.9}$$

Inserting then (3.9) into (2.40a) and (2.40b), we obtain

$$J_\nu(z)=\frac{z^\nu}{2^\nu\Gamma(\nu+1)}\left(1-\frac{z^2}{4\nu}+\frac{8z^2+z^4}{32\nu^2}-\frac{96z^2+36z^4+z^6}{384\nu^3}+\frac{1536z^2+1344z^4+96z^6+z^8}{6144\nu^4}-\cdots\right), \tag{3.10a}$$

$$Y_\nu(z)=-\frac{2^\nu\Gamma(\nu)}{\pi z^\nu}\left(1+\frac{z^2}{4\nu}+\frac{8z^2+z^4}{32\nu^2}+\frac{96z^2+36z^4+z^6}{384\nu^3}+\frac{1536z^2+1344z^4+96z^6+z^8}{6144\nu^4}-\cdots\right). \tag{3.10b}$$

These formulas can also be obtained from the power series expansion of $J_\nu(z)$ and the corresponding series expansion of $Y_\nu(z)$, where the additive part of $Y_\nu(z)$, which for integer values of ν, has a branch point at $z=0$, is to be deleted, since it is not significant in the approximate formula (3.10b). Using Stirling's formula for evaluating the gamma functions in (3.10a) and (3.10b), and keeping only the leading term in each one of these two formulas, we obtain

$$J_\nu(z)\approx[1/(2\pi\nu)^{1/2}]\,(ez/2\nu)^\nu, \tag{3.11a}$$

$$Y_\nu(z)\approx-(2/\pi\nu)^{1/2}(2\nu/ez)^\nu. \tag{3.11b}$$

IV. CONCLUSIONS

A great advantage of using the phase-integral method instead of the method of steepest descents for deriving asymptotic formulas is that the amount of work is much reduced, so that higher-order terms are obtained comparatively easily. Furthermore, one can obtain different kinds of asymptotic formulas by choosing the base function in different ways. For every appropriate choice of the base function, it is possible to calculate two linearly independent approximate solutions in a straightforward way and to handle the connection problems for these approximate solutions efficiently. We have illustrated this in Sec. II for Bessel functions under the assumption that ν and z are both positive, but the treatment can be generalized to complex values of ν and z.

The phase-integral formulas of Sec. II have certain very important properties, which sometimes make these formulas more convenient than the conventional asymptotic formulas

of Sec. III. For corresponding order of approximation, the former formulas are much simpler than the latter ones. This is obviously seen when one compares the formulas in Sec. II with the correponding formulas in Sec. III. Furthermore, the phase-integral formulas of Sec. II have such a form that the two linearly independent functions ψ_1 and ψ_2 defined in (2.2a) and (2.2b) have a Wronskian that is exactly constant. This is an important property, which the asymptotic formulas of Sec. III in general do not possess. In fact, apart from the particular case when ν is a half-integer number ($\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$,...), the Wronskian of the functions ψ_1 and ψ_2 in (2.2a) and (2.2b) with $J_\nu(z)$ and $Y_\nu(z)$ given by (3.4) and (3.5) is not constant, if the series are truncated in corresponding ways and higher-order contributions are included. Furthermore, neither Debye's asymptotic formulas nor the formulas of Sec. III C have the property of exact constancy of the Wronskian of the functions ψ_1 and ψ_2 in higher orders of approximation.

It should be pointed out that there are, however, situations when the phase-integral formulas of Sec. II are less convenient than the conventional asymptotic formulas of Sec. III. Assume, for instance, that we have an integral with a Bessel function in the integrand. Sometimes it may then be possible to evaluate the integral analytically, if the conventional asymptotic formula for the Bessel function is used, while this may not be possible if the corresponding phase-integral formula is used.

One can consider the phase-integral formula as a useful asymptotic representation of the function in question. When this representation is advantageous, one uses it directly, but when the corresponding asymptotic representation of conventional form is preferable, one goes over to it by using the expansion (1.12) with $\lambda = 1$, as we have shown in Sec. III.

[1]N. Fröman and P. O. Fröman, *JWKB Approximation, Contributions to the Theory* (North-Holland, Amsterdam, 1965) [Russian translation (MIR, Moscow, 1967)].
[2]N. Fröman, Ark. Fys. **32**, 541 (1966).
[3]N. Fröman, Ann. Phys. (NY) **61**, 451 (1970).
[4]N. Fröman and P. O. Fröman, Ann. Phys. (NY) **83**, 103 (1974).
[5]N. Fröman and P. O. Fröman, Nuovo Cimento B **20**, 121 (1974).
[6]Ö. Dammert and P. O. Fröman, J. Math. Phys. **21**, 1683 (1980).
[7]N. Fröman and P. O. Fröman, *Methodes Semi-Classiques en Mecanique Quantique*, Colloque du 10 au 15 septembre 1984, CIRM (Luminy) (l'Université de Nantes, Institut de Mathématiques et d'Informatiques, 2, rue de la Houssinière, 44072, Nantes, France), pp. 47–53.
[8]J. A. Campbell, J. Comp. Phys. **10**, 308 (1972).
[9]H. Seifert, Math. Z. **48**, 173 (1942/1943).
[10]G. N. Watson, *Theory of Bessel Functions* (Cambridge U.P., London, 1944), 2nd ed.
[11]P. Debye, Math. Ann. **67**, 535 (1909).
[12]British Association for the Advancement of Science, *Mathematical Tables*, Vol. X, *Bessel Functions*, Part II, *Functions of Positive Integer Order*, prepared by W. G. Bickley, L. J. Comrie, J. C. P. Miller, D. H. Sadler, and A. J. Thompson (Cambridge U.P., London, 1952).
[13]F. W. J. Olver, *Asymptotics and Special Functions* (Academic, New York, 1974).
[14]F. Carlini, "Ricerche sulla convergenza della serie che serve alla soluzione del problema di Keplero," in *Appendice all' Effemeridi Astronomiche di Milano per l'anno 1818* (Dall' Imp. Regia Stamperia, Milan, 1817), pp. 3–48.
[15]F. Carlini, Astron. Nachr. **30**, 197 (1850). Translation into German and revision by Jacobi of the paper by Carlini[14] of 1817; also published in C. G. J. Jacobi, *Gesammelte Werke*, Band 7, herausgegeben von K. Weierstrass (Druck und Verlag von Georg Reimer, Berlin, 1891), pp. 189–245.
[16]N. Fröman and P. O. Fröman, historical paper in *Proceedings of the Niels Bohr Centennial Conference on "Semiclassical Descriptions of Atomic and Nucelar Collisions,"* Copenhagen, 25–28 March 1985, edited by J. Bang and J. de Boer (North-Holland, Amsterdam, 1985), pp. 1–7.

Rotation in cosmology: Comments on "Imparting rotation to a Bianchi type II space-time," by M. J. Rebouças and J. B. S. d'Olival [J. Math. Phys. 27, 417 (1986)] and similar papers

Robert T. Jantzen
Department of Mathematical Sciences, Villanova University, Villanova, Pennsylvania 19085

(Received 19 March 1986; accepted for publication 4 June 1986)

The aim of this paper is to clarify confusing notions of the word "rotation" as applied to cosmological solutions of metric theories of gravity, both in general and in the specific case addressed by the article in which these confusing notions have recently reappeared.

Flat Minkowski space-time or open submanifolds of it may be sliced by a family of three-dimensional spacelike orbits of three-dimensional subgroups of the Poincaré group of Bianchi types I, III, V, VII_0, and $VII_{h\neq 0}$ and thus be made to appear as a spatially homogeneous Bianchi-type cosmological model.[1] Timelike congruences that are spatially homogeneous with respect to any of the non-Abelian groups of this list (all but type I) are in general rotating congruences: they have nonzero vorticity. No one would correctly call Minkowski space-time a rotating cosmology because of this fact, yet articles in the literature continue to do exactly this in similar circumstances.

Rotation in cosmology can refer to one of two distinct notions that are often related. Either (1) the space-time possesses an intrinsically defined timelike congruence with nonzero vorticity,[2] or (2) a natural slicing exists in terms of which an orthonormal basis of eigenvectors of the extrinsic curvature necessarily rotates as one moves along the congruence normal to the slicing.

The first idea is relevant to stationary space-times where rotation is usually first met in studying relativity; unfortunately intuition about this case is often extended to other situations where it is no longer appropriate. Nonstatic stationary space-times possess a Killing vector field that is timelike on an open submanifold of the space-time and has nonzero vorticity, i.e., the corresponding one-form is not hypersurface forming.[2] On the other hand, perfect fluid filled space-times whose fluid velocity vector has nonzero vorticity are often justifiably referred to as rotating cosmologies. In both cases the rotation refers to a component of the motion along the congruence of the perpendicular projections of Lie dragged "connecting vectors" associated with the congruence relative to a Fermi-propagated triad of orthonormal vectors spanning the local rest space relative to that congruence (a "nonrotating spatial frame").[3,4] In the famous Gödel solution,[5] which originally challenged people's ideas about rotation in relativity, the fluid velocity vector is a timelike Killing vector field, combining both of these possibilities into a single example.

The second idea is relevant to space-times where a natural slicing exists, since it refers to quantities defined not by the space-time but by a slicing of the space-time. A "Kasner frame[6]" could be defined as an orthogonal spatial frame consisting of eigenvectors of the extrinsic curvature relative to a particular slicing. The orthonormal frame obtained by normalizing such a frame (a unit Kasner frame) can then be compared to an orthonormal spatial frame that is Fermi-propagated along the congruence of unit normals to the slicing. If the unit Kasner frame rotates relative to the nonrotating spatial frame and is unique (nondegenerate eigenvalues), the slicing might be called a rotating slicing of the space-time. When the eigenvalues of the extrinsic curvature are degenerate, one may freeze out the rotational freedom in the eigenvectors due to this degeneracy by minimizing the square of the angular velocity vector, which describes the rotation. If the rotation is still nonzero, the term rotating slicing may again be used.

Like rotating congruences, all space-times have such rotating slicings; for this to be significant the rotating slicing must be intrinsically defined by the space-time. Probably the best candidate for such a slicing is one for which the trace of the extrinsic curvature ($\mathrm{Tr}\,K$), also called the mean extrinsic curvature, is constant on each slice.[7–14] Such a slicing is referred to as a constant mean curvature slicing or a "$\mathrm{Tr}\,K = \text{const}$" slicing, and in the case of vanishing mean curvature, a maximal slicing, and is a choice preferred by the simplifications that occur both in the initial value problem[15] and in geometric coordinate conditions.[16,17] A space-time with a synchronous spacelike singularity also has a unique slicing associated with the maximum lifetime function.[18,19]

For a nonstatic, stationary, axially symmetric space-time, an example of which is the Kerr rotating black hole,[20] a unique maximal slicing[21] exists consisting of the hypersurfaces orthogonal to the congruence of locally nonrotating observers.[20–23] Some thought shows that this slicing is a rotating slicing, suggesting that the idea of describing rotation of a space-time by an intrinsically defined slicing rather than a congruence is not unreasonable.

Spatially homogeneous space-times have a natural constant mean curvature slicing by the family of spacelike orbits of the homogeneity group. When these space-times have an initial big bang or final big crunch singularity, this slicing coincides with the maximum lifetime slicing. It therefore makes sense to classify spatially homogeneous space-times as rotating or nonrotating according to the second sense using the natural slicing. Such a space-time is rotating in this sense if one cannot diagonalize (for all time) the matrix of components of the spatial metric with respect to an invariant spatial frame that is comoving with the normal vector field to the natural slicing. This means that in orthogonal spatial

 0022-2488/86/112748-03$02.50

gauge (zero shift vector field), the invariant spatial frame cannot remain an orthogonal frame if it is chosen so initially. If it does, the space-time is nonrotating, a term which is therefore synonymous with "diagonalizable" as used in the literature.[2]

All Bianchi types including Bianchi type I can be rotating in this sense provided the source is general enough. For example, an electromagnetic Bianchi type I space-time will be rotating as long as the electric and magnetic field densities are not eigenvectors of the extrinsic curvature. As one might expect, all of the spatially homogeneous slicings of Minkowski space-time are nonrotating.

For a spatially homogeneous perfect fluid space-time both notions of rotation are relevant but not synonymous. If the fluid has nonzero vorticity, the natural slicing is necessarily rotating, but the converse is not true. For certain symmetry types the slicing may be rotating without the fluid having nonzero vorticity. This is true of the class B "symmetric case" models, which rotate in the Kasner frame sense even in vacuum but which do not admit a rotating fluid source.

In the case of Bianchi type II space-times, a general homogeneous perfect fluid has nonzero vorticity before one imposes the Einstein equations. If no other source is present with nonzero supermomentum, the degeneracy of the gravitational supermomentum components that occurs for Bianchi type II, requires the single component of the fluid velocity vector which is responsible for the vorticity to vanish. If one includes a general spatially homogeneous electromagnetic field as a source, one can have general values of the individual supermomenta of the fluid and the electromagnetic field while still satisfying the supermomentum constraints and thus have a rotating fluid. Bianchi type I is the only symmetry type that cannot support a rotating fluid under any conditions.

The form of the metric presented by Rebouças and d'Olival is a locally rotationally symmetric (LRS) Bianchi type II metric with an LRS electromagnetic field. This is a member of a continuous family of exact nonrotating LRS Einstein–Maxwell solutions of Bianchi types I, II, VIII, and IX known as the Brill solution.[24] The type VIII solution follows from the type IX solution by the Weyl unitary trick, which relates these two semisimple groups, while the type I and II solutions are obtained by Lie algebra contraction of the semisimple case. All of these may be obtained from Taub's original vacuum solutions[25,26] by a "variation of parameters" trick discussed for the semisimple case by Jantzen.[27] One may easily write the structure constant tensor parameters back into the equations of that discussion and thus extend them to the type I and II cases.

By introducing the new spatial coordinate $\bar{x} = x - \int C^{-2} A\, dt$ and defining $\bar{\omega}^1 = d\bar{x} + y\, dz$, the Rebouças–d'Olival metric takes the usual orthogonal gauge form

$$-ds^2 = -N^2\, dt^2 + C^2(\bar{\omega}^1)^2 + B^2[(\omega^2)^2 + (\omega^3)^2],$$
$$N = B^m C^n, \quad (m,n) = (1,-1),$$

which is explicitly diagonal, where $\{\bar{\omega}^1, \omega^2, \omega^3\}$ are time independent one-forms in the new spatial coordinates having the same exterior derivative relations as the original one-forms. As originally noticed by Bonanos,[28] different choices of (m,n) lead to different decouplings of the various equations determining B and C. The choices $(m,n) = (2,1)$, $(2,-1)$, and $(0,-1)$ were made, respectively, by Taub,[25] Brill,[24] and Misner[26] for the type IX case. For the choice $n = -1$, the vacuum zero cosmological constant equation

$$0 = \bar{G}^0{}_0 + \bar{R}^3{}_3 = \tfrac{1}{2}(\bar{R}^0{}_0 - \bar{R}^3{}_3),$$

which is also valid for an LRS electromagnetic source and nonzero cosmological constant, decouples and provides a second-order equation for B alone which simplifies for $m = 1$, where the solution is $B = (2\beta)^{-1}\cosh \beta t$, neglecting an integration constant associated with the origin of the time variable. (This equation is identical in the Bianchi type VIII and IX cases.) The vacuum zero cosmological constant equation

$$0 = \bar{R}^1{}_1 + \bar{R}^3{}_3,$$

also valid for an LRS electromagnetic source, then provides an easily integrated equation for the natural variable C (not the unnatural variable A) which has the solution $C = \gamma\, \mathrm{sech}\, \beta t$. (This changes for the semisimple case.) The super-Hamiltonian constraint then relates the two parameters β and γ to the single conserved quantity determining the LRS electromagnetic energy-momentum tensor exactly as in the case $(m,n) = (0,-1)$ used by Misner. The same "variation of parameters" enables one to insert a stiff perfect fluid as well.

The initial Rebouças–d'Olival ansatz

$$\begin{aligned} -ds^2 &= -(dt + A\omega^1)^2 + B^2[(\omega^1)^2 + (\omega^2)^2 + (\omega^3)^2] \\ &= -B^2(B^2 - A^2)^{-1}\, dt^2 + (B^2 - A^2)^2 \\ &\quad \times (\omega^1 - C^{-2} A\, dt)^2 + B^2[(\omega^2)^2 + (\omega^3)^2] \\ &\equiv -N^2\, dt^2 + g_{ab}(\omega^a + N^a\, dt)(\omega^b + N^b\, dt) \end{aligned}$$

is motivated by the stationary case for which $-(dt + A\omega^1)$ is the covariant form of a Killing vector with nonzero vorticity and is not particularly relevant to rotation in the nonstationary case. It manifestly expresses the metric in an obvious orthonormal frame, which is tilted with respect to the slicing and expressed in coordinates that are comoving with respect to the timelike member of the frame. The above coordinate transformation to orthogonal gauge coordinates represents a translation of the group manifold that eliminates the nonzero shift vector field while leaving the restrictions of the spatial one-forms to the slicing unchanged.

The above ansatz is a special case of a slightly more general ansatz introduced for the same LRS family of type I, II, VIII, and IX space-times by Bradley and Sviestens[29] for the purpose of studying rotating imperfect fluids where one no longer has deterministic field equations. This form of the metric

$$\begin{aligned} -ds^2 &= -(dt + A\omega^1)^2 + B^2(\omega^1)^2 \\ &\quad + D^2[(\omega^2)^2 + (\omega^3)^2], \\ d\omega^a &= n^{(a)}\omega^b \wedge \omega^c, \quad n^{(2)} = n^{(3)}, \end{aligned}$$

where (a,b,c) is a cyclic permutation of $(1,2,3)$ and $n^{(a)}$ are constants, may easily be reexpressed in lapse/shift form with

$C^2 = B^2 - A^2$:

$$-ds^2 = -B^2C^{-2}\,dt^2 + C^2(\omega^1 - C^{-2}A\,dt)^2 + D^2((\omega^2)^2 + (\omega^3)^2),$$

$$= -B^2C^{-2}\,dt^2 + C^2(\overline{\omega}^1)^2 + D^2((\overline{\omega}^2)^2 + (\overline{\omega}^3)^2).$$

The new one-forms $\overline{\omega}^a$, defined by

$$\mathbf{S} = e^{\theta \mathbf{k}_1}, \quad \dot{\theta} = N^1 \equiv -C^{-2}A, \quad N^2 = N^3 = 0,$$

$$\mathbf{k}_1 = -n^{(2)}\mathbf{e}^3{}_2 + n^{(3)}\mathbf{e}^2{}_3, \quad \overline{\omega}^a = S^{-1a}{}_b(\omega^b + N^b\,dt),$$

where $\mathbf{e}^b{}_a$ is the matrix with whose only nonzero entry is a one in the ath row and b th column, satisfy the same exterior differential relations as the original one-forms and correspond to a new spatial frame which is comoving with the normal vector field to the homogeneous slicing. This shows the ansatz to be entirely equivalent to the usual orthogonal gauge form of the metric for this family of space-times.

The velocity vector u for a perfect fluid must coincide with the normal vector field for this class of space-times. By choosing u to be $\partial/\partial t$ in the original coordinate system (a vector field that is tilted with respect to the normal as long as A is nonzero), one cannot satisfy the perfect fluid Einstein equations. However, one can impose a single condition on the Einstein tensor of this metric to be able to define an isotropic pressure and let the other independent components of the Einstein tensor determine a heat flow vector field. The original frame and coordinates then comove with the fluid velocity vector of this imperfect fluid, which has nonzero vorticity as long as $An^{(3)} \neq 0$. This procedure is entirely *ad hoc* and done only to investigate rotating fluids. One should not forget that these are more appropriately described as "nonsolutions" than as "solutions" of the Einstein equations in the usual sense in which the word "solution" is used.

This ansatz was misunderstood by Grøn[30] who integrated the Bradley–Sviestens equations in the vacuum type IX case with a positive cosmological constant to obtain an exact solution already found by Brill and Flaherty[8] for the Misner choice of time corresponding to $(m,n) = (0, -1)$. (The same solution with a "variation of parameters" allows an electromagnetic source and extends to the other Bianchi types of the Brill family. The general vacuum type IX case with cosmological constant was studied qualitatively by Sirousse-Zia.[31]) Then using the imperfect fluid vorticity formula, Grøn claims the solution is rotating when in fact it is a vacuum solution which is nonrotating in every sense.

ACKNOWLEDGMENTS

Frank Tipler and Jim York are thanked for their comments prior to publication.

[1] R. T. Jantzen, Nuovo Cimento B **59**, 287 (1980).
[2] D. Kramer, H. Stephani, M. MacCallum, and E. Hertl, *Exact Solutions of Einstein's Field Equations* (Cambridge U. P., Cambridge, 1980).
[3] G. F. R. Ellis, "Relativistic Cosmology," in *General Relativity and Cosmology*, edited by R. K. Sachs (Academic, New York, 1971).
[4] G. F. R. Ellis and S. W. Hawking, *The Large Scale Structure of Space-time* (Cambridge U. P., Cambridge, 1973).
[5] K. Gödel, Rev. Mod. Phys. **21**, 447 (1949).
[6] V. A. Belinsky, I. M. Khalatnikov, and E. M. Lifshitz, Ann. Phys. (NY) **19**, 225 (1970); **31**, 639 (1983).
[7] D. Brill and F. Flaherty, Commun. Math. Phys. **50**, 157 (1976).
[8] D. Brill and F. Flaherty, Ann. Inst. H. Poincaré A **28**, 335 (1978).
[9] A. J. Goddard, Gen. Relativ. Gravit. **8**, 525 (1977).
[10] J. E. Marsden and F. J. Tipler, Phys. Rep. **66**, 109 (1980).
[11] S. M. Stumbles, Ann. Phys. (NY) **133**, 28 (1981).
[12] C. Gerhardt, Commun. Math. Phys. **89**, 523 (1983).
[13] R. Bartnik, Commun. Math. Phys. **89**, 523 (1984).
[14] J. D. Barrow and F. J. Tipler, Mon. Not. R. Astron. Soc. **216**, 395 (1985).
[15] J. W. York, Jr., Phys. Rev. Lett. **28**, 1082 (1972).
[16] L. Smarr and J. W. York, Jr., Phys. Rev. D **17**, 1945, 2529 (1978).
[17] J. W. York, Jr., "Kinematics and dynamics of general relativity," in *Sources of Gravitational Radiation*, edited by L. Smarr (Cambridge U. P., Cambridge, 1979).
[18] J. D. Barrow and F. J. Tipler, Phys. Rep. **56**, 371 (1979).
[19] R. M. Wald and P. Yip, J. Math. Phys. **22**, 2658 (1981).
[20] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[21] R. D. Greene, E. L. Schucking, and C. V. Vishveshwara, J. Math. Phys. **16**, 153 (1975).
[22] J. M. Bardeen, Astrophys. J. **162**, 71 (1970).
[23] J. M. Bardeen, W. H. Press, and S. A. Teukolsky, Astrophys. J. **178**, 347 (1972).
[24] D. Brill, Phys. Rev. B **133**, 845 (1964).
[25] A. H. Taub, Ann. Math. **53**, 472 (1951).
[26] C. W. Misner and A. H. Taub, Sov. Phys. JETP **28**, 122 (1969).
[27] R. T. Jantzen, Ann. Phys. (NY) **127**, 302 (1980).
[28] S. Bonanos, Commun. Math. Phys. **22**, 190 (1971).
[29] J. M. Bradley and E. Sviestens, Gen. Relativ. Gravit. **16**, 1119 (1984); E. Sviestens, *ibid.* **17**, 521 (1985).
[30] Ø. Grøn, Phys. Rev. D **33**, 1204 (1986).
[31] H. Sirousse-Zia, Gen. Relativ. Gravit. **14**, 751 (1982).

Spherically symmetric dyonic solutions for vacuum Einstein equations in $4+K$ dimensions

S.-C. Lee
Institute of Physics, Academia Sinica, Taipei, Taiwan 11529, Republic of China and Department of Physics, National Taiwan University, Taipei, Taiwan 10764, Republic of China

S.-L. Lou
Department of Physics, National Tsing Hua University, Hsinchu, Taiwan 30015, Republic of China

(Received 11 April 1986; accepted for publication 18 June 1986)

The general spherically symmetric dyonic solutions of vacuum Einstein equations compatible with a K-dimensional toroidal fiber structure in $4+K$ dimensions are found. Solutions with no electric and/or magnetic charges can be obtained by suitable limiting procedures.

I. INTRODUCTION

In Kaluza–Klein theories,[1] we try to unify the gravitational interaction and the gauge interactions by considering only the gravitational interaction but in a higher-dimensional manifold. This manifold is assumed to have the structure of a fiber bundle with a four-dimensional base manifold and a compact homogeneous fiber. Furthermore, the metric is assumed to have a special form compatible with the bundle structure and, when restricted to the fiber, to be invariant under the "internal" group G that acts on the fiber. The question then arises as to why nature chose such a (pseudo) Riemannian manifold rather than, say, a flat higher-dimensional Minkowski space. In particular, if one assumes that the Lagrangian governing the higher-dimensional gravity is the Einstein–Hilbert Lagrangian generalized to higher dimensions, as the geometrical arguments would suggest, then it is natural to ask if there are any stationary points of the Lagrangian that display the characteristics of our ansatz. Moreover, those stationary points that cannot be continuously deformed to the trivial one, the higher-dimensional Minkowski space, are of particular interest. The monopole solution of the vacuum Einstein equation discovered by Sorkin[2] and by Gross–Perry[3] is such an example in five dimensions. This solution has the special property of being regular everywhere. In five dimensions, the internal group G is necessarily Abelian.

In this work, we shall investigate the general spherically symmetric dyonic solutions of vacuum Einstein equations under the Kaluza–Klein (KK) ansatz with an Abelian Lie group G. This may not be of direct physical relevance, however, it is one of the simplest cases where one may analyze the predictions of classical KK theories in an analytical way. Moreover, in other works,[4–6] we have pointed out the relations of such systems to the nonlinear sigma models in two dimensions, which have received a lot of interest for other reasons.

In a previous work,[7] we found a Lax form for the field equations we are considering. A method of integrating these equations was outlined later.[8] Using the results of these works, we can express a general solution of our system in terms of parameters that satisfy *nonlinear* constraint equations. The purpose of this work is to fill in the details and to solve the nonlinear constraint equations. The investigation of the various properties of the explicit solutions will be carried out in future works.

In the next section, we review very briefly the previous work[7] following largely the same notation. In Sec. III, we show how to integrate the field equations and write out the explicit solutions in terms of constrained parameters. The constraint equations are solved in Sec. IV. The general solutions can then be expressed explicitly in terms of independent parameters. A brief discussion is given in the final section.

II. FIELD EQUATIONS

The metric of the $(4+K)$-dimensional manifold is assumed to have the following form:

$$\bar{g} = g_{\mu\nu}(x)dx^\mu \otimes dx^\nu + \Phi_{ab}(x)\theta^a \otimes \theta^b, \tag{1}$$

where

$$\theta^a = dy^a + A_\mu{}^a(x)dx^\mu, \quad a = 1,\ldots,K, \tag{2}$$

and

$$g_{\mu\nu}(x)dx^\mu dx^\nu = -\exp(2\Psi(r))dt^2 + \exp(2\Lambda(r))dr^2 + r^2 d\Omega^2, \tag{3}$$

$$F^a{}_{\theta\phi} = g^a \sin(\theta), \quad F^{atr} = p^a(r), \quad F^a_{\mu\nu} = 0, \quad \text{otherwise}, \tag{4}$$

$$\Phi_{ab}(x) = (\exp(2\chi(r)))_{ab}, \quad \hat{\chi}(r) = \mathrm{Tr}\,\chi(r). \tag{5}$$

The field equations are given by the vanishing of the Ricci tensor. The equation $\bar{R}_{\mu a} = 0$ can be integrated once to give

$$\Phi_{ab}\, p^b r^2 \exp(\Psi + \hat{\chi} + \Lambda) = c_a, \tag{6}$$

where the c_a are the integration constants. The vanishing of $\bar{R}_{ab}$, $\bar{R}_{tt}$, $\bar{R}_{rr}$, and $\bar{R}_{\theta\theta}$ gives

$$\frac{d}{dr}\left[r^2 \exp(\Psi + \hat{\chi} - \Lambda)(\Phi^{-1}\partial_r \Phi)^a{}_b\right] = \frac{1}{r^2}\left(g^a g_b \exp(\Psi + \hat{\chi} + \Lambda) - c^a c_b \exp(\Psi + \Lambda - \hat{\chi})\right), \tag{7}$$

$$\frac{d}{dr}\left(r^2 \exp(\Psi + \hat{\chi} - \Lambda)\Psi'\right) = \frac{1}{2r^2} c^a c_a \exp(\Psi + \hat{\chi} - \Lambda), \tag{8}$$

 0022-2488/86/112751-05$02.50

$$\frac{d}{dr}(r\exp(\Psi+\hat{\chi}-\Lambda))$$
$$=\exp(\Psi+\hat{\chi}+\Lambda)-\frac{1}{2r^2}g^a g_a\exp(\Psi+\hat{\chi}+\Lambda), \tag{9}$$

$$\hat{\chi}''-\hat{\chi}'(\Psi'+\Lambda')+\frac{1}{4}\mathrm{Tr}(\Phi^{-1}\partial_r\Phi)^2=\frac{2}{r}(\Psi'+\Lambda'), \tag{10}$$

where $f'=df/dr$,

$$g_a=\Phi_{ab}\,g^b,\quad c^a=\Phi^{ab}g_b,\quad \Phi^{ac}\Phi_{cb}=\delta^a{}_b, \tag{11}$$

and summation for repeated indices is implied.

Let us change variables from r to z defined by the equation

$$\frac{dz}{dr}=\frac{1}{r^2}\exp(\Lambda-\Psi-\hat{\chi}). \tag{12}$$

Combining Eqs. (8) and (9) and the trace of Eq. (7), we find

$$\frac{d^2}{dz^2}(\Psi+\hat{\chi}+\ln r)=r^2\exp(2\Psi+2\hat{\chi}), \tag{13}$$

which can be integrated to give

$$r^2\exp(2\Psi+2\hat{\chi})=(k/\sinh kz)^2, \tag{14}$$

where k is the integration constant. This equation gives the relation between r and z once Ψ and $\hat{\chi}$ are known. Our solution will express Ψ, $\hat{\chi}$, and Φ as functions of z. Λ and p^a then follow from Eqs. (12) and (6), respectively. We find

$$\exp(-2\Lambda)=(\sinh kz/k)^2[\dot{\Psi}+\dot{\hat{\chi}}+k\coth kz]^2. \tag{15}$$

Introducing the vielbein $e_a{}^m$ for Φ so that

$$\Phi_{ab}=e_a{}^m e_b{}^m,$$

we find that Eqs. (7)–(9) can be written as a single matrix equation[8]

$$\dot{B}+[L,B]=0, \tag{16}$$

where B and L are $(K+2)\times(K+2)$ matrices given as

$$B=\tfrac{1}{2}[G^{-1}\dot{G}+(G^{-1}\dot{G})^T], \tag{17}$$

$$L=\tfrac{1}{2}[G^{-1}\dot{G}-(G^{-1}\dot{G})^T]. \tag{18}$$

The matrix G is an extension of the vielbein e and is defined by

$$G_{0,0}=\exp(\Psi),\quad G_{0,m}=i(Y^T e)_m,$$
$$G_{0,k+1}=w\exp(-(\Psi+\hat{\chi})),$$
$$G_{a,0}=0,\quad G_{am}=e_a{}^m,\quad G_{a,k+1}=iX_a\exp(-(\Psi+\hat{\chi})),$$
$$G_{k+1,0}=0,\quad G_{k+1,m}=0,$$
$$G_{K+1,K+1}=\exp(-(\Psi+\hat{\chi})),\quad 1\leqslant m,\quad a\leqslant K, \tag{19}$$

where X,Y are column vectors and w is a scalar defined by

$$\dot{X}=\Phi g\exp(2(\Psi+\hat{\chi})),\quad \dot{Y}=\Phi^{-1}c\exp(2\Psi),$$
$$\dot{w}=-X^T Y=-g^T\Phi Y\exp(2(\Psi+\hat{\chi})). \tag{20}$$

Note the use of matrix notations for Φ_{ab}, g^a, and c_a. It follows from Eq. (5) that $\det\Phi=\exp(2\hat{\chi})$ so that

$$\det G=1. \tag{21}$$

From this and Eq. (17), we find that

$$\mathrm{Tr}\,B=0. \tag{22}$$

We still have to check that Eq. (10) is satisfied. One can show that it is equivalent to the equation

$$\mathrm{Tr}\,B^2=2k^2. \tag{23}$$

Since $\mathrm{Tr}\,B^n$ are constants of motion, as follows from Eq. (16), Eq. (10) merely relates k^2 to other integration constants. Let the initial value of B be B_0. Since B_0 is symmetric we can write

$$B_0=K_0HK_0^T,\quad K_0K_0^T=1,\quad \overline{K}_0=IK_0I, \tag{24}$$

where H and I are diagonal,

$$H_{0a}=-\tfrac{1}{2}w_a,\quad I_{0,0}=I_{k+1,k+1}=-1,$$
$$I_{aa}=1,\quad \text{otherwise}, \tag{25}$$

where $\overline{K}_0$ is the complex conjugate of K_0. The condition on $\overline{K}_0$ is necessary to guarantee that the real and imaginary elements of B_0 appear in the proper places. Equations (22) and (23) become, respectively,

$$\sum_a w_a=0,\quad \sum_a w_a{}^2=8k^2. \tag{26}$$

In the following section, we shall indicate how to integrate Eqs. (16)–(18).

III. INTEGRATING FIELD EQUATIONS

The integration of the field equations is based on a theorem known in the mathematical literature in a more general context.[9] For simplicity, we shall choose a gauge so that the vielbein e is upper triangular. Then G is also upper triangular. The theorem states that if

$$K_0\exp(zH)=S(z)K(z), \tag{27}$$

where S is upper triangular and $K(z)$ satisfies the same constraints as K_0, i.e.,

$$KK^T=1,\quad \overline{K}=IKI, \tag{28}$$

then

$$B(z)=K(z)HK(z)^T, \tag{29}$$

$$G(z)=G_0S(z), \tag{30}$$

where G_0 is the initial value of G if the initial value of S, S_0 is chosen to be identity.

The proof of this theorem is quite simple and can be found in Ref. 8.

This theorem reduces the integration of the field equations to the problem of matrix decomposition. The required decomposition was carried out in the previous work.[8] It was also pointed out there that to find "gauge invariant" quantities, i.e., those independent of the choice of the extended vielbein G, such as the metric components Φ_{ab}, there is no need to do the matrix decomposition. Indeed, from Eqs. (27) and (30), we find

$$GG^T=U\exp(2zH)U^T, \tag{31}$$

where

$$U=G_0K_0. \tag{32}$$

It is also convenient to introduce the matrix

$$V=U\exp(zH). \tag{33}$$

We shall denote the rows of U,V by U_a,V_a, respectively, and consider them as $(K+2)$-dimensional vectors.

From the definitions of G in Eqs. (19) and (20) and using Eq. (31), we can express the metric components in terms of V as follows:

$$\exp(-2(\Psi+\hat{\chi}))=|V_{K+1}|^2, \tag{34}$$

$$\Phi_{ab}=\tilde{V}_a\cdot\tilde{V}_b, \tag{35}$$

$$\exp(2\Psi)=|v|^2, \tag{36}$$

$$iX_a=(V_a\cdot V_{K+1})/|V_{K+1}|^2, \tag{37}$$

$$i(\Phi Y)_a=\tilde{V}_a\cdot\tilde{V}_0, \tag{38}$$

$$w=(V_0\cdot V_{K+1})/|V_{K+1}|^2, \tag{39}$$

where $a,b=1,...,K$ and

$$\tilde{V}_a=V_a-iX_aV_{K+1}, \tag{40}$$

$$\tilde{V}_0=V_0-wV_{K+1}, \tag{41}$$

$$v=\tilde{V}_0-i\sum_{a=1}^{K}Y^a\tilde{V}_a. \tag{42}$$

Similarly, if we take the inverse on both sides of Eq. (31), and define

$$V^*=(V^T)^{-1},\quad U^*=(U^T)^{-1}, \tag{43}$$

then we can obtain alternate formulas to those presented in Eqs. (34)–(36). In particular, we have

$$\exp(-2\Psi)=|V^*{}_0|^2, \tag{44}$$

$$(\Phi^{-1})_{ab}=\tilde{V}^*_a\cdot\tilde{V}^*_b, \tag{45}$$

$$-iY_a=(V^*_0\cdot V^*_a)/|V^*_0|^2, \tag{46}$$

where

$$\tilde{V}^*_a=V^*_a+iY_aV^*_0. \tag{47}$$

Let us observe that X, Y, and w are auxiliary variables that depend on the metric components through their definitions in Eq. (20). This has two consequences. First, Eq. (20) will give constraints among the initial data U. Next, since Eq. (20) involves only the first derivatives of X, Y, and w, the metric components will not depend on their initial values. This is also obvious from Eqs. (35) and (36). We shall refer to this as gauge degrees of freedom. More explicitly, the metric components are invariant under the following gauge transformations:

$$U_a\to U_a+ik_aU_{K+1},\quad a=1,...,K, \tag{48}$$

$$U_0\to U_0+i\sum_{a=1}^{K}k'^aU_a-(k')^{K+1}U_{K+1}, \tag{49}$$

where k^a, k'^a, are arbitrary constant vectors of dimensions K and $K+1$, respectively.

The constraints following from Eq. (20) are the following:

$$i\sum_{a=1}^{K}g^aV_{an}+(w_n-\alpha)V_{K+1,n}=0,\quad a=1,...,K, \tag{50}$$

$$ic_a|v|^2+\sum_{n=0}^{K+1}w_n(\tilde{V}_a)_nv_n=0,\quad a=1,...,K, \tag{51}$$

where α is a constant parameter.

These constraints need only be satisfied at $z=0$. Equations of motion will guarantee that they be satisfied at any z. This is also obvious from the explicit solutions (34)–(39). Finally, we have, following from Eqs. (21) and (32),

$$(\det U)^2=1. \tag{52}$$

In the following section, we shall show how to solve the constraint Eqs. (50)–(52) on the initial data U.

IV. SOLVING CONSTRAINTS ON INITIAL DATA

Let us begin with Eqs. (51). This equation can be written as

$$\begin{aligned}-ic_a&=V'_a\cdot v/|v|^2\\&=\mathrm{Det}(V'_a,V_1,...,V_{K+1})/\mathrm{Det}(v,V_1,...,V_{K+1})\\&=\mathrm{Det}(U'_a,U_1,...,U_{K+1})/\mathrm{Det}(U_0,U_1,...,U_{K+1}),\end{aligned} \tag{53}$$

where V'_a is the $(K+2)$-vector with the components $w_n(\tilde{V}_a)_n$ and U'_a is the $(K+2)$-vector with the components w_nU_{an}. In the arguments of "Det," we have written out the rows of the determinant. To obtain the second equality, we used the fact that v is orthogonal to $V_1,...,V_{K+1}$. To obtain the last equality, we used Eq. (50), and factored out the z dependence. It follows from Eq. (53) that we can write

$$\sum_{b=1}^{K}\gamma_a{}^bU_{bn}+i\beta_aU_{K+1,n}-ic_aU_{0n}=w_nU_{an},\quad a=1,...,K, \tag{54}$$

where $\gamma_a{}^b$ and β_a are constants.

Equations (50) and (54) can now be combined into a single matrix equation

$$(w_n-\bar{\gamma})U^t{}_n=-iU_{0n}\bar{c},\quad n=0,...,K+1, \tag{55}$$

where $\bar{\gamma}$ is the $(K+1)\times(K+1)$ matrix

$$\bar{\gamma}=\begin{pmatrix}\gamma_a{}^b & \beta_a\\ g^b & \alpha\end{pmatrix} \tag{56}$$

and $U^t{}_n$, $\bar{c}$ are $(K+1)$-vectors with the components

$$(U^t{}_n)_a=U_{an},\quad (U^t{}_n)_{K+1}=iU_{K+1,n},\quad a=1,...,K, \tag{57}$$

$$(\bar{c})_a=c_a,\quad (\bar{c})_{K+1}=0, \tag{58}$$

respectively.

It is convenient to write

$$U_{0n}=if_n\det(w_n-\bar{\gamma})=if_n\sum_{m=0}^{K+1}w_n{}^{K+1-m}a_m(\bar{\gamma}), \tag{59}$$

where the last equality defines $a_m(\bar{\gamma})$.

Using the Hamilton–Caley equation,[10] i.e., if $P(\lambda)=\det(\lambda-\bar{\gamma})$, then $P(\bar{\gamma})=0$, we can compute the inverse of $w_n-\bar{\gamma}$ to get

$$(\det(w_n-\bar{\gamma}))(w_n-\bar{\gamma})^{-1}=\sum_{m=0}^{K}w_n{}^{K-m}R_m, \tag{60}$$

where R_m are the $(K+1)\times(K+1)$ matrices

$$R_m=\sum_{n=0}^{m}a_n(\bar{\gamma})\bar{\gamma}^{m-n}. \tag{61}$$

Defining the $(K+1)$-vectors $\bar{u}_m$ by

$$\bar{u}_m(\bar{\gamma},\bar{c})=R_m\bar{c} \tag{62}$$

and using Eqs. (55), (59) we obtain

$$U_{an} = f_n \sum_{m=0}^{K} w_n{}^{K-m} (\bar{u}_m)_a, \quad a = 1,...,K, \tag{63}$$

$$U_{K+1,n} = -if_n \sum_{m=0}^{K} w_n{}^{K-m} (\bar{u}_m)_{K+1}. \tag{64}$$

Note that the matrix U can be decomposed as

$$U = I^{1/2} \cdot \bar{\Gamma} \cdot A \cdot W \cdot F \cdot I^{-1/2}, \tag{65}$$

where $I^{1/2}$ and F are diagonal matrices

$$(I^{1/2})_{00} = i, \quad (I^{1/2})_{K+1,K+1} = -i, \quad (I^{1/2})_{aa} = 1,$$
$$F_{00} = if_0, \quad F_{K+1,K+1} = -if_{K+1},$$
$$F_{aa} = f_a, \quad a = 1,...,K; \tag{66}$$

$\bar{\Gamma}$ is the matrix

$$\bar{\Gamma}_{00} = 1, \quad \bar{\Gamma}_{0a} = \bar{\Gamma}_{a0} = 0,$$
$$\bar{\Gamma}_{ab} = (\bar{\gamma}^{b-1}\bar{c})_a, \quad a,b = 1,...,K+1; \tag{67}$$

and A is upper triangular, and is given by

$$A_{i,i+j} = a_j(\bar{\gamma}), \quad i,j = 0,...,K+1. \tag{68}$$

Finally, W is the matrix

$$W_{ij} = (w_j)^{K-i+1}, \quad \det W = \Delta(w_0,...,w_{K+1}). \tag{69}$$

It is now trivial to find the determinant of U, we get

$$\det U = \left(\prod_{n=0}^{K+1} f_n\right) \Delta(w_0,...,w_{K+1}) \det \bar{\Gamma}. \tag{70}$$

The constraint equation (51) can be solved by

$$(F_{nn})^2 = bh_n(-1)^n \prod_{m=0}^{K+1}{}' (w_n - w_m)^{-1},$$
$$n = 0,1,...,K+1, \tag{71}$$

$$\prod_{n=0}^{K+1} h_n = 1, \quad h_n > 0 \quad \text{if } w_0 > w_1 > \cdots > w_{K+1}, \tag{72}$$

where the superscript "$'$" in Eq. (71) means the term $m = n$ is to be omitted and the constant b is given by

$$b^{K+2} = (\det \bar{\Gamma}\,\bar{\Gamma}^T)^{-1} \tag{73}$$

and is a function of $\bar{\gamma}$,c only.

Computing VV^T, we find

$$(VV^T)_{ij} = \epsilon_{ij}\,\Delta_{ij}/\Delta, \quad \epsilon_{ij} = (I^{1/2})_{ii}(I^{1/2})_{jj}, \tag{74}$$

where Δ_{ij} is the determinant obtained from $\Delta(w_0,...,w_{K+1})$ by replacing the first row of

$$w_n{}^{K+1} \to bh_n\, I_{nn}\, P_{ij}(w_n)\exp(-w_n z); \tag{75}$$

P_{ij} is the polynomial

$$P_{ij}(w_m) = \sum_{n=0}^{2K+2} w_m{}^{2K+2-n} a_{ij,n} \tag{76}$$

with

$$a_{ij,n} = \sum_{\gamma=r_1}^{r_2} (\bar{\Gamma}A)_{ir}(\bar{\Gamma}A)_{j,n-r},$$
$$r_1 = \max(0,n-K-1), \quad r_2 = \min(n,K+1). \tag{77}$$

Similarly, we can compute V^*V^{*T} to get

$$(V^*V^{*T})_{ij} = \epsilon^*_{ij}\Delta^*_{ij}/\Delta, \quad \epsilon^*_{ij} = (I^{-1/2})_{ii}(I^{-1/2})_{jj}, \tag{78}$$

where again Δ^*_{ij} is obtained from Δ by the replacement of its first row

$$w_m{}^{K+1} \to b^{-1}h_m{}^{-1} I_{mm} P^*_{ij} \exp(w_m z); \tag{79}$$

P^*_{ij} is now given by

$$P^*_{ij} = \sum_{r,s=0}^{K+1} (\bar{\Gamma}^* A^*)_{ir}(\bar{\Gamma}^* A^*)_{js} a_r(D_m) a_s(D_m), \tag{80}$$

where D_m is the $(K+1)\times(K+1)$ diagonal matrix obtained from the diagonal matrix D, $D_{nn} = w_n$, by deleting the mth row and column. The matrices $\bar{\Gamma}^*, A^*$ are defined similarly to V^*, U^*. Note that we have, from Eq. (26),

$$a_1(D_m) = w_m, \quad a_2(D_m) = -4k^2 + w_m{}^2. \tag{81}$$

In general, $a_r(D_m)$ can be written as a polynomial in w_m with coefficients depending only on $\mathrm{Tr}\, D^n$.

The components of the metric tensor then follow from Eqs. (34), (35), (44), and (45). We have

$$\exp(-2(\Psi + \hat{\chi})) = -\Delta_{K+1,K+1}/\Delta, \tag{82}$$

$$\exp(-2\Psi) = -\Delta^*_{0,0}/\Delta, \tag{83}$$

$$\Phi_{ab} = (\Delta_{a,K+1}\Delta_{b,K+1} - \Delta_{ab}\Delta_{K+1,K+1})/(\Delta\Delta_{K+1,K+1}), \tag{84}$$

$$(\Phi^{-1})_{ab} = (\Delta^*_{0,a}\Delta^*_{0,b} - \Delta^*_{ab}\Delta^*_{0,0}). \tag{85}$$

The remaining components $\exp(-2\Lambda)$ follow from Eq. (15) and the electric fields $p^a(r)$ follows from Eq. (6).

We have expressed the general solution of the field equations in terms of $\bar{\gamma}$, c_a, w_n, and h_n, which are related to the initial conditions. Here w_n and h_n satisfy constraints that are trivial to solve. There are, however, still redundant parameters due to the gauge degrees of freedom as expressed in Eqs. (48) and (49). Substituting these equations in Eq. (55), we find that the gauge transformations are equivalent to the following transformations on $\bar{\gamma}$:

$$\bar{\gamma} \to \begin{pmatrix} 1, & -k \\ 0, & 1 \end{pmatrix} \bar{\gamma} \begin{pmatrix} 1, & k \\ 0, & 1 \end{pmatrix} + \bar{c}(k')^T, \tag{86}$$

where the vectors k^a, $(k')^i$ are gauge parameters that appear in Eqs. (48) and (49).

Let c^* be an arbitrary K-vector that may depend only on c_a, g^a and is such that $(c^*)^T c = 0$. Then one can choose a gauge such that

$$\beta_a = 0, \quad \alpha = 0, \quad \sum_{a=1}^{K} c^{*a}\gamma_a{}^b = -g^b. \tag{87}$$

In this gauge, it is easy to see that

$$\det \bar{\Gamma} = a_K(\gamma)(c^{*T} \cdot c)\mathrm{Det}(c,\gamma c,...,\gamma^{K-1}c), \tag{88}$$

$$a_{K+1}(\bar{\gamma}) = 0, \quad a_n(\bar{\gamma}) = a_n(\gamma), \quad n = 0,1,...,K. \tag{89}$$

Since the gauge constraints [Eqs. (87)] are linear, it is easy to solve explicitly. However, as in the case of w_n,h_n, we may leave it here to preserve the symmetric appearance in the indices.

Equations (82)–(85) represent our principal results. An application of these formulas to the six-dimensional case to obtain an explicit expression for all dyonic solutions can be found in Ref. 11.

V. CONCLUDING REMARKS

By adding $2K+1$ auxiliary variables, namely X, Y, and w, we have shown that the system of field equations for

$(4+K)$-dimensional KK dyons is equivalent to a Toda type system[9] based on the symmetric space $SL(K+2,R)/SO(K,2)$. (Although pure imaginary quantities seem to appear in our formulas, they can all be removed by conjugation with $I^{1/2}$.) In five dimensions ($K=1$), it is well known that the group $SO(1,2)$ appears[12] and $SO(K,2)$ appears to be the proper generalization to higher dimensions.

The fact that we have to add auxiliary variables means that the true physical system is a reduction of the corresponding Toda system. This reduction is carried out by algebraic manipulations in this work. A more geometrical approach will bring the symmetric aspects forward and may help in global analysis of the solutions.

Our initial motivation is to look for regular dyonic solutions in higher dimensions. We have succeeded in obtaining an expression for the general solutions. The analysis of regularity, even around the origin, by examining the curvature scalars is quite tedious. We hope to reformulate the problem of regularity in the language of Toda flows and see if a more powerful method can be applied. This is still under investigation.

ACKNOWLEDGMENT

This work is supported in part by a grant from the National Science Council, Taiwan, Republic of China, under Contract No. NSC74-0208-M001-15.

[1]Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. Berlin, Math. Phys. K **1**, 966 (1921); O. Klein, Z. Phys. **37**, 895 (1926).
[2]R. Sorkin, Phys. Rev. Lett. **51**, 87 (1983).
[3]D. J. Gross and M. J. Perry, Nucl. Phys. B **226**, 29 (1983).
[4]S.-C. Lee, Lett. Nuovo Cimento **44**, 133 (1985).
[5]S.-C. Lee, Phys. Lett. B **158**, 413 (1985).
[6]S.-C. Lee, Phys. Lett. B **164**, 75 (1985).
[7]S.-C. Lee and S.-L. Lou, Lett. Nuovo Cimento **44**, 69 (1985).
[8]S.-C. Lee, Class. Quantum Grav. **3**, 373 (1986).
[9]W. W. Symes, Invent. Math. **59**, 13 (1980); Physica D **1**, 339 (1980); R. Goodman and N. R. Wallach, Commun. Math. Phys. **83**, 355 (1982); **94**, 177 (1984).
[10]P. R. Halmos, *Finite-Dimensional Vector Spaces* (Van Nostrand, New York, 1958), or other textbooks in linear algebra.
[11]S.-L. Lou, Thesis, National Tsing Hua University, Hsinchu, Taiwan.
[12]A. Salam and J. Strathdee, Ann. Phys. (NY) **141**, 316 (1982).

Wavelike solutions to the Einstein equations coupled to neutrino and gauge fields

G. F. Torres del Castillo

Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, Puebla, Mexico and Departamento de Física, Centro de Investigación y de Estudios Avanzados del IPN, Apartado Postal 14-740, 07000 México, D. F. Mexico

(Received 14 April 1986; accepted for publication 18 June 1986)

Starting from the plane-wave metric, solutions to the Einstein field equations coupled to a Weyl neutrino field and to a Yang–Mills field are found. These solutions can be superposed to yield a solution with both sources if the direct interaction between them is neglected. A solution to the coupled Einstein–Yang–Mills–Weyl equations that represents a multiplet of neutrino fields interacting with the gauge field and the gravitational field is also found. All these solutions contain arbitrary functions.

I. INTRODUCTION

In a previous paper,[1] some type D solutions to the Einstein field equations coupled to different sources were given which, in the gauge employed there, where the metric has a Kerr–Schild form, can be superposed in the sense that, neglecting any interaction between the matter fields, the form of the solution found for each of the matter fields in interaction with the space-time metric is unaltered by the presence of the other sources, while the structural function contained in the metric is the sum of the expressions corresponding to each source separately. Thus, in those cases, the indirect interaction between the matter fields via the space-time metric does not change the form of the solutions as expressed in the basis used.

In the present paper a similar result is given starting from the plane-wave metric, which is a type N metric of the Kerr–Schild form, with a Weyl neutrino field and a Yang–Mills field as sources. We also give a solution of the coupled Einstein–Yang–Mills–Weyl equations, which represents a multiplet of Weyl neutrino fields interacting with the gauge field. The limiting case where the gauge field is absent and there is only one neutrino field corresponds to a solution previously found by Audretsch and Graf,[2] while when the Weyl neutrino fields are absent, the solution is a non-Abelian generalization (for an arbitrary gauge group) of the plane-wave solution of the Einstein–Maxwell equations found by Robinson.[3] In the flat-space limit, this second limiting case corresponds to the non-Abelian plane wave found by Coleman.[4] We find that the gravitational field produced by a plane wave (Abelian or non-Abelian) can also be produced by a neutrino plane wave.

The formalism used in this paper is mostly the null tetrad formalism as presented in Ref. 5. We also make use of the spinor formalism; the necessary information concerning its connection with the null tetrad formalism is summarized here (see also Ref. 1).

II. INTEGRATION OF THE FIELD EQUATIONS

We shall assume that the metric of the space-time has the Kerr–Schild form

$$g = 2\,d\zeta\,d\bar{\zeta} + 2\,du\,dv + 2h(k_\mu\,dx^\mu)^2\,, \tag{1a}$$

where u and v are real coordinates, ζ is a complex coordinate, and $\bar{\zeta}$ denotes its complex conjugate, with the choice

$$k_\mu\,dx^\mu = du \tag{1b}$$

and

$$h = h(\zeta,\bar{\zeta},u)\,. \tag{1c}$$

Writing $g = g_{ab}e^a e^b = 2e^1e^2 + 2e^3e^4$, with

$$e^1 = d\zeta\,,\quad e^2 = d\bar{\zeta}\,,\quad e^3 = du\,,\quad e^4 = dv + h\,du\,, \tag{2}$$

one finds that the tangent null tetrad ∂_a, defined by $e^a(\partial_b) = \delta^a_b$, is

$$\partial_1 = \partial_\zeta\,,\quad \partial_2 = \partial_{\bar{\zeta}}\,,\quad \partial_3 = \partial_u - h\,\partial_v\,,$$

$$\partial_4 = \partial_v = k^\mu\frac{\partial}{\partial x^\mu} \tag{3}$$

and that the independent connection one-forms, which satisfy $de^a = e^b \wedge \Gamma^a{}_b$, are given by

$$\Gamma_{42} = 0 = \Gamma_{12} + \Gamma_{34}\,,\quad \Gamma_{31} = (\partial_1 h)e^3\,. \tag{4}$$

Therefore, the curves defined by $(\zeta,\bar{\zeta},u) = \text{const}$, which have ∂_4 as tangents, form a nonexpanding and nontwisting shear-free congruence of null geodesics.

With respect to this tetrad, the nonvanishing independent components of the curvature are $R_{3131} = -\partial_1\partial_1 h$ and $R_{3132} = -\partial_1\partial_2 h$. Hence if $C^{(1)} \equiv 2R_{3131}$ is different from zero the metric is of type N and ∂_4 defines a quadruple principal null direction of the conformal curvature. The only nonvanishing component of the Ricci tensor $R_{ab} = R^c{}_{abc}$ is then given by

$$R_{33} = 2\,\partial_\zeta\,\partial_{\bar{\zeta}} h\,, \tag{5}$$

therefore,

$$R_{\mu\nu} = \sigma k_\mu k_\nu\,, \tag{6}$$

where σ is a real-valued function. The metric (1) admits the Killing vector field ∂_v, which is covariantly constant.

We now solve the field equations for the gravitational field represented by (1) coupled to a Weyl neutrino field and to a gauge field separately.

A. Weyl neutrino field

The Weyl neutrino equation—$\nabla^{A\dot{B}}\Psi_A = 0$ in spinor notation—can be written explicitly in terms of a null tetrad ∂_a as

 0022-2488/86/112756-03$02.50

$$
\begin{aligned}
&(\partial_4 - \Gamma_{421} + \tfrac{1}{2}(\Gamma_{124} + \Gamma_{344}))\Psi_2 \\
&\quad - (\partial_1 - \tfrac{1}{2}(\Gamma_{121} + \Gamma_{341}) + \Gamma_{314})\Psi_1 = 0\,, \\
&(\partial_2 + \Gamma_{423} + \tfrac{1}{2}(\Gamma_{122} + \Gamma_{342}))\Psi_2 \\
&\quad + (\partial_3 - \tfrac{1}{2}(\Gamma_{123} + \Gamma_{343}) - \Gamma_{312})\Psi_1 = 0\,,
\end{aligned} \tag{7}
$$

where $\Gamma_{abc} = \Gamma_{ab}(\partial_c)$ and Ψ_A denotes the components of the neutrino field. The energy-momentum tensor is

$$
\begin{aligned}
T_{ab} = (i\hbar/8)\,[&g_a{}^{A\dot{B}}(\Psi_{\dot{B}}\nabla_b\Psi_A - \Psi_A\nabla_b\Psi_{\dot{B}}) \\
&+ g_b{}^{A\dot{B}}(\Psi_{\dot{B}}\nabla_a\Psi_A - \Psi_A\nabla_a\Psi_{\dot{B}})\,]\,,
\end{aligned} \tag{8}
$$

where $\Psi_{\dot{A}} = \overline{\Psi_A}$, $g_1{}^{2\dot{1}} = g_2{}^{1\dot{2}} = -g_3{}^{2\dot{2}} = g_4{}^{1\dot{1}} = \sqrt{2}$, and all other $g_a{}^{A\dot{B}}$ are equal to zero. The covariant derivatives of Ψ_A are obtained from

$$
\nabla_a\Psi_A = \partial_a\Psi_A - \omega^B{}_A(\partial_a)\Psi_B\,,
$$

with

$$
\omega^1{}_1 = -\omega^2{}_2 = \tfrac{1}{2}(\Gamma_{12} + \Gamma_{34})\,,\quad \omega^1{}_2 = \Gamma_{31}\,,\quad \omega^2{}_1 = -\Gamma_{42}\,,
$$

and

$$
\nabla_a\Psi_{\dot{A}} = \partial_a\Psi_{\dot{A}} - \omega^{\dot{B}}{}_{\dot{A}}(\partial_a)\Psi_{\dot{B}},
$$

where

$$
\omega^{\dot{B}}{}_{\dot{A}} = \overline{\omega^B{}_A}.
$$

In order to satisfy the Einstein field equations

$$
R_{ab} - \tfrac{1}{2}Rg_{ab} = -8\pi T_{ab}\,, \tag{9}
$$

with R_{ab} given in (6), we impose the condition $\Psi_1 = 0$, which implies, using (4) and (8), that $T_{44} = T_{42} = T_{22} = 0$. Then from Eqs. (4) and (7) it follows that $\partial_4\Psi_2 = \partial_2\Psi_2 = 0$, which implies that $\Psi_2 = \Psi_2(\zeta,u)$, and from $T_{31} = 0$ we get $\Psi_2 = \Psi_2(u)$. Now, expressing Ψ_2 as $\Psi_2 = Re^{i\theta}$, with R and θ being real-valued functions of u, from Eqs. (5), (8), and (9) we find that

$$
h = -2\sqrt{2}\pi\hbar R^2\zeta\bar{\zeta}\frac{d\theta}{du} + f(\zeta,u) + \bar{f}(\bar{\zeta},u)\,, \tag{10}
$$

where f is an arbitrary function. When $\theta = \text{const}$, with $R \neq 0$, Ψ_A is a ghost field and the space-time corresponds to the general plane fronted gravitational wave.[6]

The neutrino field given by $\Psi_1 = 0$, $\Psi_2 = \Psi_2(u)$ has an energy-momentum tensor of the form $T_{\mu\nu} = \rho k_\mu k_\nu$, with k_μ being proportional to its flux vector, which is analogous to that of an electromagnetic plane wave, but, in contrast with the electromagnetic case where $\rho > 0$, for the neutrino field ρ can be positive, negative, or zero, depending on the value of $d\theta/du$. This solution, together with the metric (1) and (10), was obtained previously in Ref. 2; in the flat space-time, corresponding to $h = 0$, it represents a plane wave.

B. Yang–Mills field

A gauge (Yang–Mills) field is described locally by a matrix-valued one-form, $A = A_\mu\,dx^\mu$, which can be regarded as defining a connection on a principal fiber bundle with a certain structure group G. The field strength $F = \frac{1}{2}F_{\mu\nu}\,dx^\mu\wedge dx^\nu$ that corresponds to the curvature of the connection defined by A, is given by

$$
F = dA + \tfrac{1}{2}[A,A]\,, \tag{11}
$$

with the definition $[A,A] = [A_\mu,A_\nu]\,dx^\mu\wedge dx^\nu$. The Yang–Mills equations are

$$
d{}^*F + [A,{}^*F] = 4\pi{}^*j\,, \tag{12}
$$

where ${}^*F$ denotes the (Hodge) dual of the two-form F, $[A_\mu\,dx^\mu, \frac{1}{2}{}^*F_{\nu\rho}\,dx^\nu\wedge dx^\rho] = [A_\mu, \frac{1}{2}{}^*F_{\nu\rho}]\,dx^\mu\wedge dx^\nu\wedge dx^\rho$, j is a matrix-valued one-form and ${}^*j$ denotes its dual. The current one-form j is constructed out of the matter field interacting with the gauge field. If the gauge group G consists of unitary matrices, then A, F, and j are skew-Hermitian.

The energy-momentum tensor of the Yang–Mills field is

$$
4\pi T_{ab} = -\operatorname{tr}(F_{ac}F_b{}^c - \tfrac{1}{4}F_{cd}F^{cd}g_{ab})\,, \tag{13}
$$

where tr denotes the trace. In order to have all $T_{ab} = 0$ except for T_{33}, as required by (5) and (9), F_{31} and F_{32} must be the only nonvanishing independent components of the field strength; then

$$
4\pi T_{33} = -2\operatorname{tr}(F_{31}F_{32})\,. \tag{14}
$$

We shall assume that A_3 is the only nonvanishing component of A (in some specific gauge), then $A = A_3\,du$, and $F = dA_3\wedge du$. From the condition $F_{43} = 0$ it follows that A_3 must be a function of u, ζ, and $\bar{\zeta}$ only.

Defining the components of ${}^*F$ by ${}^*F_{ab} = (i/2)\times\epsilon_{abcd}F^{cd}$, with $\epsilon_{1234} = 1$, we get

$$
{}^*F = -i\,\partial_1A_3e^1\wedge e^3 + i\,\partial_2A_3e^2\wedge e^3;
$$

therefore

$$
d{}^*F + [A,{}^*F] = 2i\,\partial_\zeta\,\partial_{\bar{\zeta}}A_3e^1\wedge e^2\wedge e^3 \tag{15}
$$

[cf. Eq. (5)]. In the sourceless case ($j = 0$) from Eqs. (12) and (15) it follows that $A_3 = \gamma(\zeta,u) + \delta(\bar{\zeta},u)$, where γ and δ are arbitrary matrix-valued functions. If the elements of G are unitary matrices, then A must be skew-Hermitian and, therefore, $\delta = -\gamma^\dagger$, where $\gamma^\dagger$ is the Hermitian adjoint of γ. The Einstein field equations (9) with (5) and (14) give

$$
h = -2\operatorname{tr}\gamma\gamma^\dagger + f(\zeta,u) + \bar{f}(\bar{\zeta},u)\,, \tag{16}
$$

where f is an arbitrary function.

The particular solution $\gamma(\zeta,u) = c(u)\zeta$ represents, in the flat space-time corresponding to $h = 0$, a non-Abelian plane wave.[4] The gravitational field corresponding to this wave, according to (16) is determined by

$$
h = -2(\operatorname{tr}cc^\dagger)\zeta\bar{\zeta} + f(\zeta,u) + \bar{f}(\bar{\zeta},u)\,, \tag{17}
$$

which is similar to the expression given by Eq. (10). Thus, when $d\theta/du > 0$, the gravitational field given by Eq. (10) can be considered as produced by a neutrino field or by a plane wave, Abelian or non-Abelian. An analogous dual interpretation for the source of a gravitational field has been found in Ref. 1.

III. SOLUTION OF THE EINSTEIN–YANG–MILLS–WEYL EQUATIONS

It is easy to see that, expressed with respect to the tetrad (2) and (3), the fields

$$
A = [\gamma(\zeta,u) + \delta(\bar{\zeta},u)]\,du\,,
$$

$$
\Psi_1 = 0\,,\quad \Psi_2 = R(u)\exp i\theta(u)\,,
$$

with

$$h = 2\,\mathrm{tr}(\gamma\delta) - 2\sqrt{2}\pi\hbar R^2\zeta\bar{\zeta}\frac{d\theta}{du} + f(\zeta,u) + \bar{f}(\bar{\zeta},u)\,,$$

satisfy the Einstein–Yang–Mills–Weyl equations provided that any interaction between the Yang–Mills field and the Weyl field be neglected. This means that the perturbation of the space-time geometry produced by each of these matter fields does not change the expression of the other. This result is similar to that given in Ref. 1.

In the present case we can also take into account the interaction between the Yang–Mills field and a multiplet of Weyl neutrino fields. Labeling with Latin indices, $i,j,\ldots$, the components with respect to a basis of the "internal space," a multiplet of Weyl fields has components $\Psi_A{}^i$ that, interacting with a gauge field $A_a = (A^j_{ai})$, satisfy the equation obtained from Eq. (7) by replacing ∂_a by $\partial_a + A_a$ (regarding Ψ_A as a column with entries $\Psi_A{}^i$). Expression (8) must be modified by replacing $\nabla_a\Psi_A$ by $\nabla_a\Psi_A + A_a\Psi_A$ and $\nabla_a\Psi_{\dot{B}}$ by $\nabla_a\Psi_{\dot{B}} - \Psi_{\dot{B}}A_a$ (regarding $\Psi_{\dot{B}}$ as a row with entries $\Psi_{\dot{B}i}$) and placing the dotted components to the left of the undotted ones.

Assuming, as before, that $\Psi_1{}^i = 0$ and that $A_3 \neq 0$ only, one gets essentially the same equations as in Sec. II, which imply that $\Psi_2{}^i$ are (complex-valued) functions of ζ and u only. Then condition $T_{31} = 0$ requires that $\Psi_{\dot{2}i}\,\partial_\zeta\Psi_2{}^i = 0$ (summed over i); thus, $\Psi_{\dot{2}i}\Psi_2{}^i$ has to be a function of u only. The Weyl field multiplet acts as a source of both the gauge field and the gravitational field. The (matrix-valued) current one-form corresponding to the multiplet $\Psi_A{}^i$ is given by $j^k_i = i\epsilon g_a{}^{A\dot{B}}\Psi_{\dot{B}i}\Psi_A{}^k e^a$, where ϵ is a (real) coupling constant. [The value of ϵ is already fixed by the normalization used in Eqs. (8), (12), and (13).] In the present case we have

$$j^k_i = -\sqrt{2}i\epsilon\Psi_{\dot{2}i}\Psi_2{}^k e^3,$$

therefore

$$*j^k_i = -6\sqrt{2}\epsilon\Psi_{\dot{2}i}\Psi_2{}^k e^1\wedge e^2\wedge e^3,$$

and from Eqs. (12) and (15) it follows that the gauge field is determined by

$$\partial_\zeta\,\partial_{\bar{\zeta}}A_{3i}{}^k = 12\sqrt{2}i\epsilon\Psi_{\dot{2}i}\Psi_2{}^k\,.$$

Then, from Eqs. (5), (8), (9), and (14), one finds that the solution of the Einstein equations is given by

$$\partial_\zeta\partial_{\bar{\zeta}}h = 2\,\mathrm{tr}(\partial_\zeta A_3)(\partial_{\bar{\zeta}}A_3) + \sqrt{2}i\pi\hbar(\Psi_{\dot{2}j}\,\partial_u\Psi_2{}^j - \Psi_2{}^j\,\partial_u\Psi_{\dot{2}j} + 2\Psi_{\dot{2}j}A_{3k}{}^j\Psi_2{}^k)\,.$$

IV. CONCLUDING REMARKS

We have shown that the metric considered in this paper is compatible with a Weyl neutrino field, with an electromagnetic field, and with a Yang–Mills field. Furthermore, the effect produced by each of these fields on the metric of the space-time does not change the form of the other solutions. This is due to the fact that, with the proposed alignment of the fields, the function h included in the metric does not appear in the expressions for these matter fields. The metric studied in Ref. 1 also has these properties; moreover, according to Ref. 6, these two metrics are the only ones that admit a ghost neutrino field and, in the auxiliary Minkowski metric $2\,d\zeta\,d\bar{\zeta} + 2\,du\,dv$, the corresponding geodesic shear-free null congruences defined by k_μ [see Eq. (1a)] constitute geometric representations of twistors (see, e.g., Refs. 7 and 8).

ACKNOWLEDGMENT

The author acknowledges support from the Sistema Nacional de Investigadores (México).

[1] G. F. Torres del Castillo and J. F. Plebański, J. Math. Phys. **26**, 477 (1985).
[2] J. Audretsch and W. Graf, Commun. Math. Phys. **19**, 315 (1970).
[3] I. Robinson, Lecture at King's College, London, 1956.
[4] S. Coleman, Phys. Lett. B **70**, 59 (1977).
[5] G. C. Debney, R. P. Kerr, and A. Schild, J. Math. Phys. **10**, 1842 (1969).
[6] C. D. Collinson and P. B. Morris, J. Phys. A: Math. Nucl. Gen. **6**, 915 (1973).
[7] R. Penrose, J. Math. Phys. **8**, 345 (1967).
[8] G. F. Torres del Castillo, "Ghost neutrino fields in flat space-time," preprint.

Calculation of radiated gravitational energy using the second-order Einstein tensor

Chris X. Habisohn
Enrico Fermi Institute and Department of Physics, University of Chicago, 5640 S. Ellis Avenue, Chicago, Illinois 60637

(Received 12 March 1986; accepted for publication 2 July 1986)

In general relativity there is a well-defined prescription for defining a quantity that represents the radiated energy of an exact, asymptotically flat solution of Einstein's equation. This quantity is called the Bondi energy flux. However, in linearized gravity off a stationary and asymptotically flat background, the second-order Einstein tensor has been used as a stress-energy tensor for the perturbed gravitational field, enabling one to calculate the energy radiated away in gravitational radiation. It is natural to ask how this method compares to the exact method for calculating the Bondi energy flux. In this paper, it is shown that if the metric perturbation satisfies certain falloff and gauge conditions, then the radiated energy calculated using the second-order Einstein tensor equals the second-order contribution to the Bondi energy flux associated with the perturbation. As an application, the second-order Einstein tensor is used to demonstrate gravitational superradiance from a Kerr black hole. Also, the Appendix contains a theorem that makes precise the notion that if $\nabla_{(a}\xi_{b)}$ and its derivative is "small," then ξ_a is close to a Killing field.

I. INTRODUCTION

For asymptotically flat space-times there is available a natural, well-defined prescription for calculating the total energy carried off in gravitational radiation. We call this prescription the Bondi prescription and the energy carried off in gravitational radiation is called the total Bondi energy flux.[1] The Bondi prescription is natural in the sense that no additional geometric structures, other than those present for all asymptotically flat spaces, must be introduced on the space-time to define the procedure. The Bondi prescription is also satisfying in the sense that the Bondi energy flux is equal to the change in the Bondi mass of the system. In practice, however, the Bondi energy flux is difficult to calculate. This is because the Bondi energy flux is defined as an integral over $\mathscr{I}^+$, the future null boundary of the space-time. If one wants to calculate the Bondi energy flux given a certain asymptotically flat solution of Einstein's equation, he must first perform the tedious process of constructing the manifold $\mathscr{I}^+$ and its relevant fields before the integral may be evaluated.

A different method has been proposed to calculate the energy contained in gravitational waves that employs the Landau–Lifshitz complex. In a suitable coordinate system this method has been shown to agree with the Bondi prescription.[2]

In this paper we propose another method for calculating the total energy flux in gravitational radiation, which we call the second-order Einstein method. This method is defined in connection with linearized perturbations off a stationary asymptotically flat background. The idea is the following. To any solution γ_{ab} of the linearized Einstein equation we associate a divergence-free, symmetric tensor field $G^{(2)}[\gamma_{cd}]_{ab}$ called the second-order Einstein tensor, defined by

$$G^{(2)}[\gamma_{cd}]_{ab} = \frac{1}{2}\frac{d^2}{d\lambda^2} G[g_{cd} + \lambda\gamma_{cd}]_{ab}\,|_{\lambda=0}\,, \quad (1.1)$$

where $G[g_{cd} + \lambda\gamma_{cd}]_{ab}$ is the Einstein tensor for the metric $g_{ab} + \lambda\gamma_{ab}$. We formally treat $-8\pi G^{(2)}[\gamma_{cd}]_{ab}$ as if it were a stress-energy tensor for the field γ_{ab}. The total radiated energy flux, denoted $\zeta[\gamma_{ab}]$, is now calculated by contracting the timelike Killing field into $-8\pi G^{(2)}[\gamma_{cd}]_{ab}$ and integrating the resulting mass-energy current over a timelike three-surface surrounding the source of the gravitational radiation. This method has been used[3,4] to calculate the energy flux associated with linearized gravitational waves on a Minkowski background.

The main goals of this paper are to define rigorously the second-order Einstein method and to demonstrate that, under suitable conditions, the energy flux calculated with this method equals the second-order contribution to the total Bondi flux. With these goals in mind, we begin in Sec. II with a review of the mathematical machinery needed to define the total Bondi energy flux.

In Sec. III we discuss some properties of the second-order Einstein tensor and define the second-order Einstein method for perturbations off a stationary asymptotically flat background. One problem that arises in using the second-order Einstein tensor as a stress-energy tensor for the perturbation γ_{ab} is that it is gauge dependent. That is, two different metric perturbations, γ_{ab} and γ'_{ab}, which differ by a symmetrized derivative of a covector field, generally produce different second-order Einstein tensors. Hence, it is possible that $\zeta[\gamma_{ab}]$ will not equal $\zeta[\gamma'_{ab}]$ even though γ_{ab} and γ'_{ab} represent the same physical situation. This threatens to destroy the uniqueness of the second-order Einstein method. However, in Sec. III we show that if one restricts attention to perturbations that satisfy a certain set of falloff conditions, called the weak falloff conditions, then the second-order Einstein method is unique. Roughly, the weak falloff conditions require γ_{ab}, in a neighborhood of null infinity, to vanish identically in the past of some spacelike slice and it and its first and second derivatives are required to approach zero in the future of the slice.

 0022-2488/86/112759-11$02.50

While imposing the weak falloff conditions on γ_{ab} is enough to insure that the second-order Einstein method is gauge invariant, an additional restriction is imposed on γ_{ab} in Sec. IV in order to prove that the radiated energy calculated using the second-order Einstein method equals the second-order contribution to the total Bondi energy flux, denoted $E^{(2)}[\gamma_{ab}]$. This restriction requires γ_{ab} to be gauge related to some perturbation γ'_{ab} that is in the Geroch–Xanthopoulos gauge and satisfies a strengthened set of falloff conditions. These strengthened falloff conditions, which imply the weak falloff conditions, apply to the conformally related perturbation $\tilde{\gamma}'_{ab} = \tilde{\Omega}^2 \gamma'_{ab}$, which is defined on the "unphysical" conformally related space-time used to defined the asymptotic flatness of the background space-time, where $\tilde{\Omega}$ is the conformal factor. The Geroch–Xanthopoulos gauge choice requires certain components of $\tilde{\gamma}'_{ab}$ to fall off to zero on $\mathscr{I}^+$ like $\tilde{\Omega}$, $\tilde{\Omega}^2$, and $\tilde{\Omega}^3$. Geroch and Xanthopoulos[5] have shown that any perturbation satisfying the weak falloff conditions can be brought into the Geroch–Xanthopoulos gauge through a gauge transformation. However, it is not clear when a perturbation satisfying the weak falloff conditions is gauge related to a γ'_{ab} that satisfies both the Geroch–Xanthopoulos gauge conditions and the strengthened falloff conditions; therefore we impose this restriction on γ_{ab} in the hypothesis of our theorem.

The key steps in the proof of our main theorem and their motivation are best illustrated by first describing a seemingly reasonable approach to the proof of the theorem, pointing out its major flaw, and then describing how this flaw is dealt with in the acutal proof in Sec. IV. The seemingly reasonable approach is the following. First, write the integral for defining $\zeta[\gamma_{ab}]$ in terms of "unphysical" quantities, such as $\tilde{\gamma}_{ab}$, on the unphysical conformally related space-time used to define the asymptotic flatness of the background space-time. Second, because this integral is independent of surface (since $G^{(2)}[\gamma_{cd}]_{ab}$ is divergence-free), we may push the surface of integration out to $\mathscr{I}^+$. We then compare this integrand on $\mathscr{I}^+$ to that used to define $E^{(2)}[\gamma_{ab}]$. If the integrands are the same, the theorem is proved. The major flaw that appears when one tries to implement the above approach is that the integrand may not smoothly extend to $\mathscr{I}^+$. In our proof in Sec. IV, we get around this difficulty by first gauge transforming the perturbation so that it satisfies the Geroch–Xanthopoulos gauge and the strengthened falloff conditions. With our perturbation in the Geroch–Xanthopoulos gauge, it turns out that the integrand used to define the second-order Einstein method now extends smoothly to $\mathscr{I}^+$. Comparing this integrand on $\mathscr{I}^+$ to the integrand that defines $E^{(2)}[\gamma_{ab}]$, we find that they differ by a term that integrates to zero; thus proving the theorem. We have yet to mention the role that strengthened falloff conditions have in the proof. Essentially, these conditions insure that we can push the integration surface out to $\mathscr{I}^+$ without changing the value of the integral. Without these conditions, we knew only that the integral was independent of surface within the physical space-time.

To complete our discussion of the second-order Einstein method, we use it to demonstrate that the Kerr space-time possesses superradiant gravitational modes. Press and Teukolsky[6] have previously shown that Kerr possesses superradiant modes by a different method involving a separated version of the spin-2 field equation in Kerr.

Finally, the Appendix contains a theorem that is used in Sec. III. Essentially the theorem makes precise the notion that if $\nabla_{(a}\xi_{b)}$ is "small" then ξ_a is close to a Killing field.

II. ASYMPTOTICS AND THE GEROCH–XANTHOPOULOS GAUGE

In this section we review the asymptotic machinery needed to define the total Bondi energy flux. A more extensive account of the material contained in this review can be found in Ref. 1. The Geroch–Xanthopoulos gauge choice for the linearized metric perturbation γ_{ab} is also defined, and the expression for the second-order contribution to the Bondi energy in this gauge is presented.

Recall that to an asymptotically flat space-time (M, g_{ab}) we associate an asymptote $(\tilde{M}, \tilde{g}_{ab}, \tilde{\Omega}, \mathscr{I})$ that consists of a space-time $(\tilde{M}, \tilde{g}_{ab})$ with null boundary $\mathscr{I}$ and a conformal factor $\tilde{\Omega}$ that conformally relates the interior of $(\tilde{M}, \tilde{g}_{ab})$ to (M, g_{ab}), i.e., $\tilde{\Omega}^2 g_{ab} = \tilde{g}_{ab}$. The conformal factor $\tilde{\Omega}$ is required to satisfy the following four properties: $\tilde{\Omega} = 0$ on $\mathscr{I}$, $\tilde{\nabla}_a \tilde{\Omega} = \tilde{n}_a \neq 0$ on $\mathscr{I}$, $\tilde{n}_a \tilde{n}_b \tilde{g}^{ab} = 0$ on $\mathscr{I}$, and $\tilde{n}^a = \tilde{n}_b \tilde{g}^{ab}$ is to be a complete vector field when restricted to $\mathscr{I}$. The boundary $\mathscr{I}$ is required to be diffeomorphic to two disconnected copies of $S^2 \times R$. The components of $\mathscr{I}$ are denoted $\mathscr{I}^+$ and $\mathscr{I}^-$, where $\mathscr{I}^+$ is called future null infinity and $\mathscr{I}^-$ is called past null infinity. Here (M, g_{ab}) is called the physical space-time while $(\tilde{M}, \tilde{g}_{ab})$ is called the unphysical space-time. Heuristically we think of the asymptote as a means of attaching the points at infinity (the boundary $\mathscr{I}$) to the physical space-time.

The physical Ricci tensor R_{ab} is related to the unphysical Ricci tensor $\tilde{R}_{ab}$ through

$$\tilde{\Omega} S_{ab} = \tilde{\Omega} \tilde{S}_{ab} + 2\tilde{\nabla}_a \tilde{n}_b - \tilde{f} \tilde{g}_{ab} , \tag{2.1}$$

where $\tilde{\nabla}_a$ is the derivative operator associated with the unphysical metric, and S_{ab}, $\tilde{S}_{ab}$, and $\tilde{f}$ are defined by the following expressions:

$$S_{ab} = R_{ab} - \tfrac{1}{6} R g_{ab} , \tag{2.2}$$

$$\tilde{S}_{ab} = \tilde{R}_{ab} - \tfrac{1}{6} \tilde{R} \tilde{g}_{ab} , \tag{2.3}$$

$$\tilde{f} = \tilde{\Omega}^{-1} \tilde{n}^a \tilde{n}^b \tilde{g}_{ab} , \tag{2.4}$$

with R and $\tilde{R}$ being the physical and unphysical scalar curvature, respectively.

As a manifold, $\mathscr{I}^+$ inherits many geometrical structures from $\tilde{M}$, including a vector field, a degenerate metric, and a volume element. We denote tensor fields defined on $\mathscr{I}^+$ by underlining their symbol. The vector field, denoted $\underline{n}^a$, is simply the restriction of $\tilde{n}^a$ to $\mathscr{I}^+$. The degenerate metric $\underline{g}_{ab}$ is obtained by restricting the action of $\tilde{g}_{ab}$ to $\mathscr{I}^+$. The volume element $\underline{\epsilon}_{abc}$ on $\mathscr{I}^+$ is defined by the requirement that $\tilde{n}_a \tilde{\epsilon}^{abcd} \underline{\epsilon}_{bcd} = 3!$, where $\tilde{\epsilon}^{abcd}$ is the contravariant volume element on $\tilde{M}$. It follows from (2.1) that if $\tilde{\Omega}$ is chosen so that $\tilde{f}$ vanishes on $\mathscr{I}^+$ then both $\underline{g}_{ab}$ and $\underline{\epsilon}_{abc}$ are Lie derived by $\underline{n}^a$. Such a conformal factor is said to be in the Bondi gauge. We shall impose this gauge choice on our $\tilde{\Omega}$.

Even though $\underline{g}_{ab}$ is not invertible, it is still possible to

define an object $\underline{g}^{ab}$, which is similar to an inverse. Let $\underline{g}^{ab}$ be any tensor on $\mathscr{I}^+$ that satisfies $g_{ac}\underline{g}^{ab}g_{bd} = g_{cd}$. We note that $\underline{g}^{ab}$ is unique up to addition of any tensor of the form $\underline{n}^{(a}\underline{v}^{b)}$. Of course, we must only use g_{ab} in places where the freedom in choosing $\underline{g}^{ab}$ does not affect the result.

If our physical space-time satisfies the vacuum Einstein equation in a neighborhood of $\mathscr{I}^+$, then $\mathscr{I}^+$ also possesses a symmetric tensor field $\underline{N}_{ab}$ called the Bondi news tensor. Here $\underline{N}_{ab}$ contains information about the asymptotic gravitational radiation field. In particular the total outgoing Bondi energy flux is obtained by integrating the square of the Bondi news over $\mathscr{I}^+$ with the volume element $\underline{\epsilon}_{abc}$. A few important properties of the news tensor are the following:

$$\text{(i)}\ \underline{N}_{ab}\underline{n}^a = 0\,, \tag{2.5a}$$

$$\text{(ii)}\ \underline{N}_{ab}\underline{g}^{ab} = 0\,, \tag{2.5b}$$

$$\text{(iii)}\ \underline{N}_{ab} = \underline{S}_{ab} - \underline{\rho}_{ab}\,, \tag{2.5c}$$

where $\underline{S}_{ab}$ is the pullback of $\tilde{S}_{ab}$ to $\mathscr{I}^+$ and $\underline{\rho}_{ab}$ is a tensor field on $\mathscr{I}^+$ uniquely defined by $\underline{n}^a$ and $\underline{g}_{ab}$. The expression for $\underline{\rho}_{ab}$ in terms of $\underline{n}^a$ and $\underline{g}_{ab}$ is complicated and will not be given here.

(iv) if $\tilde{t}^a$ is an asymptotic time translation vector field on $\tilde{M}$ and $\tilde{t}^a$ restricted to $\mathscr{I}^+$ equals $\underline{n}^a$, then the total Bondi energy flux at $\mathscr{I}^+$, denoted E, associated with $\tilde{t}^a$ is given by

$$E = (32\pi)^{-1}\int_{\mathscr{I}^+} \underline{N}_{ab}\underline{N}_{cd}\underline{g}^{ac}\underline{g}^{bd}\underline{\epsilon}_{efh}\,. \tag{2.5d}$$

We note that the use of the inverse metric $\underline{g}^{ab}$ is justified in (2.5d) since $\underline{n}^a\underline{N}_{ab} = 0$. Also since the signature of g_{ab} is $(0,+,+)$, E vanishes if and only if $\underline{N}_{ab}$ vanishes.

Fix (M,g_{ab}) as a stationary asymptotically flat space-time with timelike Killing field t^a and asymptote $(\tilde{M},\tilde{g}_{ab},\tilde{\Omega},\mathscr{I})$. We would now like to calculate the lowest-order change in the total Bondi energy flux due to a first-order metric perturbation in the Geroch–Xanthopoulos gauge. For this calculation we keep the Bondi gauge choice in force and we further choose our conformal factor so that the extension of the timelike Killing field to $\mathscr{I}^+$ equals $\underline{n}^a$. Let γ_{ab} be a solution of the linearized Einstein equation in a neighborhood of $\mathscr{I}^+$. Because the Bondi news tensor for a stationary space-time vanishes, the lowest-order contribution to the total Bondi energy flux, denoted $E^{(2)}[\gamma_{ab}]$, of the metric perturbation γ_{ab} is second order and is given by

$$E^{(2)}[\gamma_{ab}] = (32\pi)^{-1}\int_{\mathscr{I}^+} \underline{N}^1{}_{ab}\underline{N}^1{}_{cd}\underline{g}^{ac}\underline{g}^{bd}\underline{\epsilon}_{efh}\,, \tag{2.6}$$

where $\underline{N}^1{}_{ab}$ is the first-order contribution to the Bondi news due to γ_{ab}. We need not include any changes in $\underline{g}^{ab}$, $\underline{\epsilon}_{efh}$, or changes in the conformal factor $\tilde{\Omega}$ in (2.6) because these would only contribute to the total Bondi energy to third order or higher. Hence the integral in (2.6) is performed using the inverse metric g^{ab}, volume element $\underline{\epsilon}_{efh}$, and conformal factor $\tilde{\Omega}$ associated with the background space-time. To compute $\underline{N}^1{}_{ab}$, we will employ the properties of the Geroch–Xanthopoulos gauge.

The perturbation γ_{ab} is said to be in the Geroch–Xanthopoulos gauge if $\tilde{\gamma}_{ab} = \tilde{\Omega}^2\phi_{ab}$ is smoothly extendable to and satisfies the following three properties on $\mathscr{I}^+$:

$$\text{(i)}\ \tilde{\gamma}_{ab}\big|_{\mathscr{I}^+} = 0\,, \tag{2.7a}$$

$$\text{(ii)}\ \tilde{\Omega}^{-1}\tilde{\gamma}_{ab}\tilde{n}^a\big|_{\mathscr{I}^+} = 0\,, \tag{2.7b}$$

$$\text{(iii)}\ \tilde{\Omega}^{-2}\tilde{\gamma}_{ab}\tilde{n}^a\tilde{n}^b\big|_{\mathscr{I}^+} = 0\,. \tag{2.7c}$$

As a consequence of the Geroch–Xanthopoulos gauge, we have

$$\underline{N}^1{}_{ab} = \pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}_{ab}) \quad \text{pulled back to } \mathscr{I}^+\,. \tag{2.8}$$

Indeed, (2.8) follows directly from (2.7a)–(2.7c), (2.1), and (2.5c). Equation (2.7a) implies that to first order g_{ab} and $\underline{n}^a$ are unchanged; hence $\underline{\rho}_{ab}$ (which is dependent on $\underline{g}_{ab}$ and $\underline{n}^a$) is unchanged to first order. Therefore (2.5c) implies $\underline{N}^1{}_{ab} = \underline{S}^1{}_{ab}$, where $\underline{S}^1{}_{ab}$ is the first-order contribution to $\underline{S}_{ab}$. That $\tilde{S}^1{}_{ab} = \pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}_{ab})$ on $\mathscr{I}^+$, where $\tilde{S}^1{}_{ab}$ is the first-order contribution to $\tilde{S}_{ab}$, follows from (2.7a)–(2.7c), (2.1), and the fact that γ_{ab} satisfies the linearized Einstein equation. Therefore $\underline{N}^1{}_{ab} = \pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}_{ab})$ pulled back to $\mathscr{I}^+$.

Using Eqs. (2.6) and (2.8) we can now express $E^{(2)}[\gamma_{ab}]$ as

$$E^{(2)}[\gamma_{ab}] = (32\pi)^{-1}\int_{\mathscr{I}^+} \pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}^a{}_b)\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}^b{}_a)\underline{\epsilon}_{efh}\,. \tag{2.9}$$

We were allowed to convert the inverse metric $\underline{g}^{ab}$ on $\mathscr{I}^+$ in (2.6) to the full Lorentz metric $\tilde{g}^{ab}$ on $\tilde{M}$ in (2.9) because of (2.7b). One nice feature of Eq. (2.9), besides the fact that it is the formula for $E^{(2)}[\gamma_{ab}]$ that we have sought, is that the only geometrical object it contains that is intrinsic to $\mathscr{I}^+$ is the volume element. Even though the integral is over $\mathscr{I}^+$, the integrand $\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}^a{}_b)\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}^b{}_a)$ is a function on $\tilde{M}$. This fact helps us in comparing $E^{(2)}[\gamma_{ab}]$ to $\zeta[\gamma_{ab}]$ in the proof of our main theorem.

At this time, we also introduce some auxiliary structures on $\tilde{M}$ that we will need in order to state and prove our theorems in Secs. III and IV. The auxiliary structures are a neighborhood of $\mathscr{I}^+$ denoted U, a three surface σ, a time coordinate t on U, and a norm on tensors on U. We require U to be a neighborhood of $\mathscr{I}^+$ in which g_{ab} satisfies the vacuum Einstein equation and which contains complete integral curves of t^a, the timelike Killing field. We require σ to be a

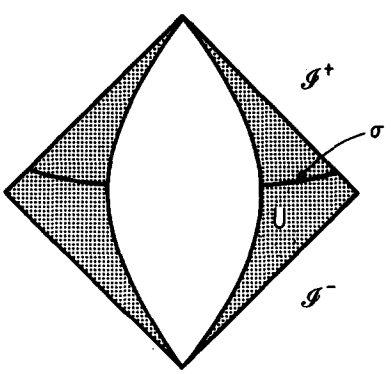


FIG. 1. The unphysical space-time $(\tilde{M},\tilde{g}_{ab})$ with future null boundary $\mathscr{I}^+$, past null boundary $\mathscr{I}^-$, open neighborhood U of $\mathscr{I}^+$, and spacelike slice of U denoted σ.

spacelike three-surface in U that intersects $\mathscr{I}^+$ in a cross section and intersects each integral curve of t^a in U; see Fig. 1. Also let us define t on U using the following two equations:

$$t|_\sigma = 0\,, \tag{2.10}$$

$$\pounds_t t = 1\,. \tag{2.11}$$

Finally, let h_{ab} be a positive definite metric on U that satisfies the property

$$\pounds_t h_{ab} = 0\,. \tag{2.12}$$

We now define the norm of an arbitrary tensor $T^{a\cdots b}{}_{c\cdots d}$ in the tensor space of U by the formula

$$|T^{a\cdots b}{}_{c\cdots d}| = \{T^{a_1\cdots b_1}{}_{c_1\cdots d_1} T^{a_2\cdots b_2}{}_{c_2\cdots d_2} h_{a_1a_2}\cdots h_{b_1b_2} h^{c_1c_2}\cdots h^{d_1d_2}\}^{1/2}\,, \tag{2.13}$$

where h^{ab} is the inverse of h_{ab}.

III. THE SECOND-ORDER EINSTEIN TENSOR

The second-order Einstein tensor associated with the metric perturbation γ_{ab} is defined by (1.1). Expressed in terms of the background derivative operator, the second-order Einstein tensor is given by

$$\begin{aligned} G^{(2)}[\gamma_{cd}]_{ab} &= \tfrac{1}{2}\gamma^{cd}\nabla_a\nabla_b\gamma_{cd} + \tfrac{1}{4}(\nabla_a\gamma^{cd})\nabla_b\gamma_{cd} \\ &\quad + (\nabla^{[c}\gamma^{d]a})(\nabla_c\gamma_{db}) - \tfrac{1}{4}C^d(2\nabla_{(b}\gamma_{a)d} - \nabla_d\gamma_{ba}) \\ &\quad - \tfrac{1}{2}\gamma^{cd}(2\nabla_c\nabla_{(b}\gamma_{a)d} - \nabla_c\nabla_d\gamma_{ba}) \\ &\quad + \{-\tfrac{1}{4}\gamma^{cd}\nabla^e\nabla_e\gamma_{cd} - \tfrac{3}{8}(\nabla^e\gamma^{cd})\nabla_e\gamma_{cd} + \tfrac{1}{8}C^dC_d \\ &\quad + \tfrac{1}{4}\gamma^{cd}\nabla_cC_d + \tfrac{1}{4}(\nabla^d\gamma^{ce})(\nabla_c\gamma_{de})\}g_{ab}\,, \end{aligned} \tag{3.1}$$

where C_d represents the combination $2\nabla^c\gamma_{cd} - \nabla_d\gamma^c{}_c$. The expression (3.1) is rather lengthy. It is one of the goals of this paper to present a simple method, in some cases, of calculating quantities involving the second-order Einstein operator without actually writing out this lengthy expression. Essentially this is achieved by replacing $G^{(2)}[\gamma_{cd}]_{ab}$ wherever it appears by the Einstein tensor associated with the metric $g_{ab} + \lambda\gamma_{ab}$. One then calculates the appropriate quantities as functions of λ and takes their second derivative to arrive at the desired result. This is done in Sec. V to demonstrate superradiance off a Kerr black hole.

The second-order Einstein operator arises naturally in perturbation theory. Recall that in perturbation theory one considers a one-parameter family of solutions to Einstein's equation, $g(\lambda)_{ab}$. The perturbation equations are generated by expanding out Einstein's equation in powers of λ and equating like powers. In the absence of matter the first few lowest-order equations become

$$0 = G[g_{cd}]_{ab}\,, \tag{3.2a}$$

$$0 = G^{(1)}[\gamma_{cd}]_{ab}\,, \tag{3.2b}$$

$$0 = G^{(1)}[\gamma^{(2)}{}_{cd}]_{ab} + G^{(2)}[\gamma_{cd}]_{ab}\,, \tag{3.2c}$$

where γ_{cd} and $\gamma^{(2)}{}_{cd}$ are the first- and second-order contributions to $g(\lambda)_{cd}$, respectively, and G_{ab} and $G^{(1)}{}_{ab}$ are the Einstein operator and the linearized Einstein operator, respectively. The second-order Einstein operator first appears in Eq. (3.2c). Let us define the tensor field τ_{ab} as

$$\tau_{ab} = -(8\pi)^{-1}G^{(2)}[\gamma_{cd}]_{ab} \tag{3.3}$$

and rewrite Eq. (3.2c) as

$$G^{(1)}[\gamma^{(2)}{}_{cd}]_{ab} = 8\pi\tau_{ab}\,. \tag{3.4}$$

We are now tempted to regard τ_{ab} as a stress-energy tensor for γ_{cd} for the following three reasons. First, τ_{ab} is quadratic in its dependence on γ_{cd}. Second, τ_{ab} is symmetric and divergence-free with respect to the background derivative operator. And, third, τ_{ab} is sourcing the higher-order correction $\gamma^{(2)}{}_{ab}$ in the same way that a first-order matter stress-energy term $T^{(1)}{}_{ab}$ would have sourced γ_{ab} if a one-parameter family of matter stress-energy tensors, $T(\lambda)_{ab}$, would have been included in the perturbation expansion (i.e., $G^{(1)}[\gamma_{cd}] = 8\pi T^{(1)}{}_{ab}$).

The divergence-free property of τ_{ab} is a consequence of the definition of $G^{(2)}[\gamma_{cd}]_{ab}$ and the first two perturbation equations (3.2a) and (3.2b) in the absence of matter. To show this let ∇_a be the background derivative operator and let $\Gamma'(\lambda)^a{}_{bc}$ be the connection between ∇_a and the derivative operator associated with the metric $g'(\lambda)_{ab}$ defined by $g'(\lambda)_{ab} = g_{ab} + \lambda\gamma_{ab}$. Taking the second derivative with respect to λ of each side of the identity

$$\begin{aligned} 0 = g'(\lambda)^{ae}\{&\nabla_a G[g'(\lambda)_{cd}]_{eb} + \Gamma'(\lambda)^h{}_{ae} \\ &\times G[g'(\lambda)_{cd}]_{hb} + \Gamma'(\lambda)^h{}_{ab}G[g'(\lambda)_{cd}]_{eh}\}\,, \end{aligned} \tag{3.5}$$

setting $\lambda = 0$, and using (3.2a) and (3.2b) we get

$$\begin{aligned} 0 = \nabla^a G^{(2)}[\gamma_{cd}]_{ab} &+ \Gamma'(0)^{ae}{}_e G^{(2)}[\gamma_{cd}]_{ab} \\ &+ \Gamma'(0)^e{}_b{}^a G^{(2)}[\gamma_{cd}]_{ae}\,. \end{aligned} \tag{3.6}$$

Substituting $\Gamma'(0)^a{}_{bc} = 0$ into (3.6) produces the desired result

$$0 = -(8\pi)^{-1}\nabla^a G^{(2)}[\gamma_{cd}]_{ab} = \nabla^a\tau_{ab}\,. \tag{3.7}$$

We are cautioned against taking this analogy too far, though. Probably the main drawback in using τ_{ab} as a stress tensor for the perturbed gravitational field is that it is gauge dependent. That is, two different γ_{ab}'s, which differ by a symmetrized derivative of a covector field, and hence represent the same physical perturbation, will not in general produce equivalent second-order Einstein tensors.

We are now in a position to describe the second-order Einstein method for calculating radiated gravitational energy. Basically, the second-order Einstein method is a procedure for assigning a total radiated energy flux $\zeta[\gamma_{ab}]$ to a linearized metric perturbation γ_{ab} off (M, g_{ab}), which satisfies the linearized Einstein equation in U. In this method, Eq. (3.3) is used to construct τ_{ab}, which is then contracted into the timelike Killing field t^a to produce a conserved mass-energy covector $j_a = -\tau_{ab}t^b$. This covector is integrated over a timelike three-surface Σ in U surrounding the source of the radiation to obtain $\zeta[\gamma_{ab}]$.

To insure that Σ catches all the outgoing radiation, we require Σ to approach future and past timelike infinity and to stay away from $\mathscr{I}$, see Fig. 2 (there are timelike surfaces that intersect $\mathscr{I}^+$). A reasonable restriction that we shall impose on Σ which will insure that this condition is met is

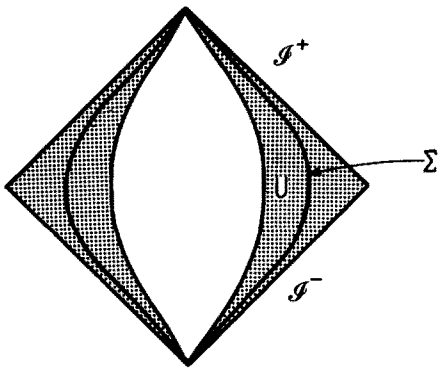


FIG. 2. The unphysical space-time $(\tilde{M},\tilde{g}_{ab})$ with a typical timelike surface of integration Σ.

that the surface should lie tangent to and contain complete orbits of the timelike Killing field.

In order for the second-order Einstein method to be uniquely defined, we impose a few relatively minor constraints on the metric perturbation. The uniqueness of the second-order Einstein method is assured if $\zeta[\gamma_{ab}]$ exists and is independent of choice of integration surface Σ and if a class of gauges for γ_{ab} can be specified such that if γ_{ab} and γ'_{ab} are gauge related and both belong to the specified gauge class then $\zeta[\gamma_{ab}]=\zeta[\gamma'_{ab}]$. We will find that both these requirements will be satisfied by requiring γ_{ab} to satisfy a set of falloff conditions called the weak falloff conditions.

The perturbation γ_{ab} is said to satisfy the weak fall-off conditions if γ_{ab} vanishes in the past of σ and $t^{(1+\epsilon)/2}\,|\gamma_{ab}|$, $t^{(1+\epsilon)/2}\,|\nabla_a\gamma_{bc}|$, and $t^{(1+\epsilon)/2}\,|\nabla_a\nabla_b\gamma_{cd}|$ each approach zero uniformly to the future of U, where ϵ is some positive constant and t is the time coordinate described at the end of Sec. II. By a function ρ approaching zero uniformly to the future of U we mean that for each $\delta>0$ there exists a number $t(\delta)$ such that $|\rho|<\delta$ in that part of U for which $t>t(\delta)$.

Let γ_{ab} satisfy the weak falloff conditions in U and let $j_a=G^{(2)}[\gamma_{cd}]_{ab}t^b$. Because j_a depends algebraically on γ_{ab}, $\nabla_a\gamma_{bc}$, and $\nabla_a\nabla_b\gamma_{cd}$, we conclude that $|j_a|$ will fall off uniformly faster than $t^{-(1+\epsilon)}$, where ϵ is some positive constant. This is enough to guarantee the existence of the integral defining $\zeta(\gamma_{ab})$. Now consider two surfaces Σ_1 and Σ_2 satisfying the above criterion imposed on Σ. Since $\nabla^a j_a=0$, the integral of j_a over Σ_1 will equal the integral of j_a over Σ_2 provided the past and future boundary integrals of j_a vanish. However, the boundary integrals are assured to vanish due to the uniform falloff of $|j_a|$. Therefore $\zeta[\gamma_{ab}]$ is independent of the surface Σ.

The fact that requiring γ_{ab} to satisfy the weak falloff conditions is sufficient to make the second-order Einstein method independent of gauge on γ_{ab} is proven in the following theorem.

Theorem: Let γ_{ab} satisfy the linearized Einstein equation and the weak falloff conditions in U. Also let $\gamma'_{ab}=\gamma_{ab}+2\nabla_{(a}\xi_{b)}$ satisfy these same falloff conditions. Then $\zeta[\gamma_{ab}]=\zeta[\gamma'_{ab}]$.

Proof: Let $\varphi(\lambda)$ be the one-parameter family of diffeomorphisms on M associated with ξ^a. Because g_{ab} satisfies the vacuum Einstein equation in U and γ_{ab} satisfies the linearized Einstein equation in U, the second-order part of the expression $G[g_{cd}+\lambda\gamma_{cd}]_{ab}$ will equal the second-order part of the expression $\varphi_*(\lambda)[G[g_{cd}+\lambda\gamma_{cd}]_{ab}]$. Therefore we can express the second-order Einstein tensor in U as

$$G^{(2)}[\gamma_{cd}]_{ab}=\frac{1}{2}\frac{d^2}{d\lambda^2}\varphi_*(\lambda)[G[g_{cd}+\lambda\gamma_{cd}]_{ab}]\Big|_{\lambda=0}. \tag{3.8}$$

Now $\varphi_*(\lambda)$ commutes with the Einstein operator, hence

$$G^{(2)}[\gamma_{cd}]_{ab}=\frac{1}{2}\frac{d^2}{d\lambda^2}G[\varphi_*(\lambda)[g_{cd}]+\lambda\varphi_*(\lambda)[\gamma_{cd}]]_{ab}\Big|_{\lambda=0}. \tag{3.9}$$

Computing the right side of (3.9) we get

$$G^{(2)}[\gamma_{cd}]_{ab}=G^{(2)}[\gamma_{cd}+2\nabla_{(c}\xi_{d)}]_{ab}+G^{(1)}[\pounds_\xi(\gamma_{cd}+\nabla_{(c}\xi_{b)})]_{ab}. \tag{3.10}$$

From (3.10) we conclude that

$$\zeta[\gamma_{ab}]-\zeta[\gamma'_{ab}]=(8\pi)^{-1}\int_\Sigma G^{(1)}[\pounds_\xi(\gamma_{cd}+\nabla_{(c}\xi_{b)})]_{ab}t^b\epsilon^a{}_{efh}. \tag{3.11}$$

Now the integrand in (3.11) is divergence-free, hence it is locally the divergence of some two-form F_{ab}. In fact, we can find a global F_{ab} as follows. Let

$$H_{cd}=\pounds_\xi(\gamma_{cd}+\nabla_{(c}\xi_{b)}) \tag{3.12}$$

and define F_{ab} as

$$F_{ab}=-t_{[a}\nabla_{b]}H^c{}_c+t_{[a}\nabla^cH_{b]c}-\tfrac{1}{2}H^c{}_c\nabla_{[a}t_{b]}-t^c\nabla_{[a}H_{b]c}+H_{[a}{}^c\nabla_{|c|}t_{b]}. \tag{3.13}$$

Then

$$\nabla^aF_{ab}=G^{(1)}[\pounds_\xi(\gamma_{cd}+\nabla_{(c}\xi_{d)})]_{ab}t^a. \tag{3.14}$$

Applying Stokes' theorem, we get

$$\xi[\gamma_{ab}]-\xi[\gamma'_{ab}]=(8\pi)^{-1}\int_{\partial\Sigma}F_{cd}\epsilon^{cd}{}_{ab}. \tag{3.15}$$

To prove that the integral in (3.15) is zero, we first note that F_{ab} is algebraically dependent on γ_{ab} and its first and second derivatives, $\nabla_{(a}\xi_{b)}$ and its first and second derivatives, ξ_a, $\nabla_{[a}\xi_{b]}$, and $\nabla_a\nabla_{[b}\xi_{c]}$. All but the last three terms listed vanish uniformly to the future of U due to the hypothesis. Let $\partial\Sigma_1,\partial\Sigma_2,\ldots$ be a sequence of compact two-surfaces in Σ, which are time translates of each other and approach the future of Σ. The theorem in the Appendix implies that there exists a $c>0$ such that if the norm of $\nabla_{(a}\xi_{b)}$ and its first derivative is less than some δ on $\partial\Sigma_i$, then there exists a $\xi^{(i)}{}_a$ that differs from ξ_a by a Killing field and for which both $|\xi^{(i)}{}_a|$ and $|\nabla_{[a}\xi^{(i)}{}_{b]}|$ are less than $c\delta$ on $\partial\Sigma_i$. Also since $\nabla_a\nabla_{[b}\xi^{(i)}{}_{c]}$ is algebraically dependent on $R_{bca}{}^d\xi^{(i)}{}_d$ and $\nabla_a\nabla_{(b}\xi_{c)}$ through the formula

$$\nabla_a\nabla_{[b}\xi^{(i)}{}_{c]}=-R_{bca}{}^d\xi^{(i)}{}_d+2L_{a[bc]}, \tag{3.16}$$

where $L_{abc}=\nabla_b\nabla_{(c}\xi_{a)}$, we could choose c such that $|\nabla_a\nabla_{[b}\xi^{(i)}{}_{c]}|$ is also less than $c\delta$ on $\partial\Sigma_i$. Now the integral of $F_{cd}\epsilon^{cd}{}_{ab}$ over $\partial\Sigma_i$ [Eq. (3.15)] is unchanged by adding a Killing vector to ξ_a in (3.12). This is because the integral in (3.11) depends only on ξ_a through its symmetrized deriva-

tive. Therefore the integral of $F_{cd}\epsilon^{cd}{}_{ab}$ over $\partial\Sigma_i$ is equal to the integral of $F^{(i)}{}_{cd}\epsilon^{cd}{}_{ab}$ over $\partial\Sigma_i$, where $F^{(i)}{}_{cd}$ is defined by (3.12) and (3.13) except that $\xi^{(i)}{}_a$ is substituted for ξ_a. But it is clear that the norm of $F^{(i)}{}_{cd}\epsilon^{cd}{}_{ab}$ on $\partial\Sigma_i$ approaches zero as $i\to\infty$. Hence the integral of $F_{cd}\epsilon^{cd}{}_{ab}$ over $\partial\Sigma_i$ approaches zero as $i\to\infty$ (i.e., as $\partial\Sigma_i$ approaches the future of Σ). Therefore $\zeta[\gamma_{ab}]=\zeta[\gamma'_{ab}]$.

Q.E.D.

IV. MAIN THEOREM

In this section we will prove a result that states sufficient conditions on γ_{ab} so that $\zeta[\gamma_{ab}]=E^{(2)}[\gamma_{ab}]$. However, to state the result we need to introduce a new set of falloff conditions on γ_{ab}, called the strengthened falloff conditions. Let γ_{ab} be in the Geroch–Xanthopoulos gauge. Then γ_{ab} is said to satisfy the strengthened falloff conditions if γ_{ab} vanishes in the past of σ and the quantities $t^{(1+\epsilon)/2}|\tilde\Omega^{-1}\tilde\gamma_{ab}|$, $t^{(1+\epsilon)/2}|\tilde\nabla_a\tilde\Omega^{-1}\tilde\gamma_{bc}|$, and $t^{(1+\epsilon)/2}|\tilde\nabla_a\tilde\nabla_b\tilde\Omega^{-1}\tilde\gamma_{cd}|$ approach zero uniformly to the future of U, where ϵ is any positive constant. The strengthened falloff conditions are slightly stronger than the weak falloff conditions mainly because they require that the perturbation also obey falloff conditions on $\mathscr{I}^+$. The strengthened falloff conditions imply the weak falloff conditions.

Theorem: Let γ_{ab} be a solution of the lineaized Einstein equation in U that satisfies the weak falloff conditions and is gauge equivalent to a perturbation that satisfies the Geroch–Xanthopoulos gauge conditions and the strengthened falloff conditions. Then $\zeta[\gamma_{ab}]$ equals the second-order contribution to the total Bondi energy flux $E^{(2)}[\gamma_{ab}]$.

Proof: We begin by choosing our conformal factor $\tilde\Omega$ judiciously so that subsequent calculations are simplified. We choose $\tilde\Omega$ so that it is Lie derived by t^a and so that the extension of the timelike Killing field t^a to $\mathscr{I}^+$ (which we still denote as t^a), when restricted to $\mathscr{I}^+$, equals $\underline{n}^a$. With this choice of conformal factor, it follows that

$$t^a\tilde n_a=0\,, \tag{4.1a}$$

$$\tilde\nabla_a t^b|_{\mathscr{I}^+}=0\,, \tag{4.1b}$$

$$\tilde\Omega^{-1}t^a\tilde\nabla_a\tilde n^b=\tilde\Omega^{-1}\tilde n^a\tilde\nabla_a\tilde n^b=\tfrac12\tilde\nabla^b\tilde f+\tfrac12\tilde\Omega^{-1}\tilde f\tilde n^b \text{ on } \mathscr{I}^+\,. \tag{4.1c}$$

Furthermore, t^a is a Killing vector field for $\tilde g_{ab}$ and the Bondi gauge condition is satisfied. We now proceed to prove these claims. Equation (4.1a) is obvious. We next prove that the Bondi gauge condition, $\tilde f$ vanishing on $\mathscr{I}^+$, is satisfied. Evaluating (2.1) on $\mathscr{I}^+$ we get

$$\tilde\nabla_a\tilde n_b=\tfrac12\tilde f\tilde g_{ab}\,. \tag{4.2}$$

Furthermore, since $\tilde n^a=t^a$ on $\mathscr{I}^+$, there exists a $\tilde v^b$ such that on $\mathscr{I}^+$ we have

$$\tilde\nabla_a\tilde n^b-\tilde\nabla_a t^b=\tilde n_a\tilde v^b\,. \tag{4.3}$$

Lowering the index on this expression with $\tilde g_{ab}$, symmetrizing, and using the fact that t^b is a Killing field, we get

$$\tilde\nabla_a\tilde n_b=\tilde n_{(a}\tilde v_{b)} \tag{4.4}$$

on $\mathscr{I}^+$. Comparing (4.2) and (4.4) we conclude that $\tilde n_{(a}\tilde v_{b)}=\frac12\tilde f\tilde g_{ab}$ on $\mathscr{I}^+$, which can only be satisfied when $\tilde f$ and $\tilde v^b$ both vanish on $\mathscr{I}^+$; thus satisfying the Bondi gauge condition. Furthermore, (4.3), (4.4), and the vanishing of $\tilde v^b$ on $\mathscr{I}^+$ imply (4.1b). Equation (4.1c) follows the definition of $\tilde f$. This completes the proof of the claims. For further reference, let us also denote the timelike three-surfaces of constant $\tilde\Omega$ in U by $\Sigma(\tilde\Omega)$. We note $\Sigma(0)=\mathscr{I}^+$.

Because the second-order Einstein method is gauge independent, we may assume that γ_{ab} satisfies the Geroch–Xanthopoulos gauge conditions and the strengthened falloff conditions. Furthermore, because the second-order Einstein method is independent of surface $\zeta[\gamma_{ab}]$ is given by

$$\zeta[\gamma_{ab}]=(8\pi)^{-1}\int_{\Sigma(\tilde\Omega)}G^{(2)}[\gamma_{cd}]_{ab}t^b\epsilon^a{}_{efh}\,, \tag{4.5}$$

where $\tilde\Omega$ can take any value, except possibly $\tilde\Omega=0$, where $G^{(2)}[\gamma_{cd}]_{ab}t^b\epsilon^a{}_{efh}$ will not, in general, be smooth.

We now demonstrate that, as a consequence of the Geroch–Xanthopoulos gauge choice, if we view $G^{(2)}[\gamma_{cd}]_{ab}\times t^b\epsilon^a{}_{efh}$ not as a three-form on $\tilde M$, but rather as a three-form defined on $\Sigma(\tilde\Omega)$ then it is smoothly extendable to $\mathscr{I}^+$. Using the fact that $\tilde n^a$ is normal to $\Sigma(\tilde\Omega)$, we may write the integral in (4.2) as

$$\zeta[\gamma_{ab}]=-(8\pi)^{-1}\int_{\Sigma(\tilde\Omega)}\tilde\Omega^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde n^a\tilde\epsilon_{efh}\,, \tag{4.6}$$

where $\tilde\epsilon_{efh}$ is the intrinsic volume element to $\Sigma(\tilde\Omega)$ defined by $\tilde n_a\tilde\epsilon^{abcd}\tilde\epsilon_{efh}=3!$. The minus sign was introduced in (4.6) because $\tilde n^a$ is the inward normal. Expressing the integrand in (4.6) in terms of $\tilde\gamma_{ab}$ and its unphysical derivatives we have

$$\begin{aligned}\tilde\Omega^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde n^a=\tilde\Omega^{-2}t^b\tilde n^a\{&\tilde\nabla_{[a}(\Lambda^d{}_{c]b}\tilde\gamma^c{}_d)+\tfrac12\Lambda^e{}_{b[a}\Lambda^c{}_{c]e}+\tilde\Omega^{-1}\tilde\gamma^{cd}\Lambda_{cab}\tilde n_d-\tfrac12\tilde\Omega^{-1}\Lambda^{de}{}_e\tilde n_d\tilde\gamma_{ab}\\&-\tilde\Omega^{-1}\tilde\gamma^{me}(\tilde\nabla_m\tilde n_e)\tilde\gamma_{ab}+3\tilde\Omega^{-2}\tilde\gamma^{cd}\tilde n_c\tilde n_d\tilde\gamma_{ab}+\text{terms proportional to }\tilde g_{ab}\}\,,\end{aligned} \tag{4.7}$$

where Λ_{abc} represents the combination $2\tilde\nabla_{(b}\tilde\gamma_{c)a}-\tilde\nabla_a\tilde\gamma_{bc}$. Now the variables $\tilde K_{ab}$, $\tilde K_a$, $\tilde K$, $\tilde M_a$, and $\tilde M$ defined by $\tilde\Omega^{-1}\tilde\gamma_{ab}$, $\tilde\Omega^{-2}\tilde\gamma_{ab}\tilde n^b$, $\tilde\Omega^{-3}\tilde\gamma_{ab}\tilde n^a t^b$, $\tilde\Omega^{-2}\tilde\gamma_{ab}t^b$, and $\tilde\Omega^{-3}\tilde\gamma_{ab}\tilde n^a t^b$, respectively, are smooth on $\mathscr{I}^+$ due to the Geroch–Xanthopoulos gauge choice. Expressing (4.7) in terms of these variables, and using the fact that t^b is Killing and Lie derives $\tilde\Omega$, we get

$$\begin{aligned}&\tilde\Omega^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde n^a\\&=\tfrac14\pounds_t\pounds_{\tilde n}(\tilde K^c{}_d\tilde K^d{}_c)-\tfrac14(\pounds_t\tilde K^a{}_b)\pounds_{\tilde n}\tilde K_a{}^b+\tilde\nabla_c(\tilde K^{cd}\tilde K_{bd}t^a\tilde\nabla_a\tilde n^b)+\tfrac38\tilde f\pounds_t(\tilde K^c{}_d\tilde K^d{}_c)+(t^b\tilde\nabla_c\tilde n^a)\tilde K^{cd}\nabla_{(a}\tilde K_{b)d}\\&\quad+\{2\tilde K^d{}_{bd}\tilde n^c+\tilde n^a\tilde K^{cd}\nabla_{(a}\tilde K_{b)d}+\tfrac12\tilde f\tilde K_{bd}\tilde K^{cd}+\tfrac12\tilde K_d\tilde n_b\tilde K^{cd}-\tfrac14\tilde n^c\tilde\nabla_b\tilde K^a{}_d\tilde K^d{}_a-\tfrac12\tilde K_{be}\tilde n^c\tilde\nabla^e\tilde K^a{}_a\\&\quad-\tfrac12\tilde K^a{}_a\tilde K_b\tilde n^c-2\tilde K^d{}_{bd}\tilde n^c+\tilde n^c\tilde K^d{}_{bd}+\tilde K^{[d}{}_a(\tilde\nabla^{c]}\tilde n^a)\tilde K_{bd}-\tfrac14\tilde K^e{}_b\pounds_{\tilde n}\tilde K^c{}_e+\tfrac14\tilde K^c{}_d\pounds_{\tilde n}\tilde K^d{}_b\}\tilde\nabla_c t^b\\&\quad+\tilde\Omega(\text{terms smooth at }\mathscr{I}^+\text{ involving }\tilde K_{ab},\tilde K_a,\tilde K,\tilde M_a,\tilde M,\tilde f,\tilde n^a,t^a,\tilde\Omega,\text{ and their derivatives}).\end{aligned} \tag{4.8}$$

Clearly, (4.8) is smooth at $\mathscr{I}^+$.

Next, we evaluate $\tilde{\Omega}^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde{n}^a$ on $\mathscr{I}^+$. Using the fact that $\tilde{f}$, $\tilde{\nabla}_a\tilde{n}_b$, and $\tilde{\nabla}_a t^b$ vanish on $\mathscr{I}^+$, we get, from (4.8), that

$$\tilde{\Omega}^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde{n}^a\big|_{\mathscr{I}^+} = \tfrac{1}{4}\pounds_t\pounds_{\tilde{n}}(\tilde{K}^c{}_d\tilde{K}^d{}_c) - \tfrac{1}{4}(\pounds_t\tilde{K}^a{}_b)\pounds_{\tilde{n}}\tilde{K}_a{}^b + \tilde{\nabla}_c(\tilde{K}^{cd}\tilde{K}_{bd}t^a\tilde{\nabla}_a\tilde{n}^b)\big|_{\mathscr{I}^+}. \tag{4.9}$$

The last term in (4.9), it turns out, also vanishes. Indeed, (4.1c) implies that $\tilde{\Omega}^{-1}t^a\tilde{\nabla}_a\tilde{n}^b$ is proportional to $\tilde{n}^b$ on $\mathscr{I}^+$. This, together with the fact that $K_{bd}\tilde{n}^b$ vanishes on $\mathscr{I}^+$, implies that $\tilde{K}^{cd}\tilde{K}_{bd}t^a\tilde{\nabla}_a\tilde{n}^b$ falls off on $\mathscr{I}^+$ like $\tilde{\Omega}^2$. Hence $\tilde{\nabla}_c(\tilde{K}^{cd}\tilde{K}_{bd}t^a\tilde{\nabla}_a\tilde{n}^b)$ vanishes on $\mathscr{I}^+$. Therefore

$$\tilde{\Omega}^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde{n}^a\big|_{\mathscr{I}^+} = \tfrac{1}{4}\pounds_t\pounds_{\tilde{n}}(\tilde{K}^c{}_d\tilde{K}^d{}_c) - \tfrac{1}{4}(\pounds_t\tilde{K}^a{}_b)\pounds_{\tilde{n}}\tilde{K}_a{}^b\big|_{\mathscr{I}^+}. \tag{4.10}$$

Now, if the value of the integral in (4.6), as a function of $\tilde{\Omega}$, was continuous at $\tilde{\Omega}=0$, then we could evaluate $\zeta[\gamma_{ab}]$ by substituting (4.10) into (4.6) and integrating over $\mathscr{I}^+$. To show that the integral is continuous at $\tilde{\Omega}=0$, we invoke the strengthened falloff conditions. Let t_δ be some positive real parameter, and consider the following two integrals, I_1 and I_2:

$$I_1 = (8\pi)^{-1}\int_{\Sigma(\tilde{\Omega},t_\delta<t)}\tilde{\Omega}^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde{n}^a\tilde{\epsilon}_{efh}, \tag{4.11}$$

$$I_2 = (8\pi)^{-1}\int_{\Sigma(\tilde{\Omega},0<t<t_\delta)}\tilde{\Omega}^{-2}G^{(2)}[\gamma_{cd}]_{ab}t^b\tilde{n}^a\tilde{\epsilon}_{efh}, \tag{4.12}$$

where $\Sigma(\tilde{\Omega},t_\delta<t)$ is that region of $\Sigma(\tilde{\Omega})$ for which $t>t_\delta$, and similarly, $\Sigma(\tilde{\Omega},0<t<t_\delta)$ is that region of $\Sigma(\tilde{\Omega})$ for which $0<t<t_\delta$. Clearly, $I_1+I_2=\zeta[\gamma_{ab}]$ for $\tilde{\Omega}\neq 0$. The strengthened falloff conditions imply that the integrands in (4.11) and (4.12) fall off uniformly in U like $t^{-(1+\epsilon)}$ from which one can show that for every $\delta>0$ there exists a value for t_δ such that $I_1<\delta$ for any value of $\tilde{\Omega}$, including $\tilde{\Omega}=0$. Furthermore, since the region in U that satisfies $0<t<t_\delta$ is compact we have that I_2 is continuous in $\tilde{\Omega}$ for any value of $\tilde{\Omega}$, including $\tilde{\Omega}=0$. Since δ is arbitrary and I_2 is continuous, it follows that (4.6) is a continuous function of $\tilde{\Omega}$, even at $\tilde{\Omega}=0$.

Using (4.10) to evaluating (4.6) on $\mathscr{I}^+$, we get

$$\zeta[\gamma_{ab}] = (32\pi)^{-1}\int_{\mathscr{I}^+}\{\pounds_t(\tilde{\Omega}^{-1}\tilde{\gamma}^c{}_d)\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}_c{}^d) - \pounds_t\pounds_{\tilde{n}}(\tilde{\Omega}^{-2}\tilde{\gamma}^{cd}\tilde{\gamma}_{cd})\}\tilde{\epsilon}_{efh}. \tag{4.13}$$

The second term in (4.13) integrates by parts to zero. Making use of the fact that $\underset{\sim}{\epsilon}_{abc}=\tilde{\epsilon}_{abc}$ and $\pounds_t(\tilde{\Omega}^{-1}\tilde{\gamma}^c{}_d)=\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}^c{}_d)$ on $\mathscr{I}^+$, we can write $\zeta[\gamma_{ab}]$ as

$$\zeta[\gamma_{ab}] = (32\pi)^{-1}\int_{\mathscr{I}^+}\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}^c{}_d)\pounds_{\tilde{n}}(\tilde{\Omega}^{-1}\tilde{\gamma}_c{}^d)\,\underset{\sim}{\epsilon}_{efh}. \tag{4.14}$$

The expression for $\zeta[\gamma_{ab}]$ in (4.14) is identical to the expression for $E^{(2)}[\gamma_{ab}]$ given in (2.9). Therefore $\zeta[\gamma_{ab}]=E^{(2)}[\gamma_{ab}]$. Q.E.D.

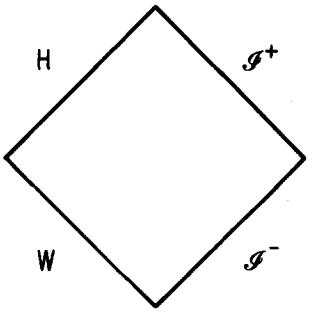


FIG. 3. A space-time diagram of that portion of the Kerr space-time exterior to the black hole. $\mathscr{I}^+$ and $\mathscr{I}^-$ are the future and past null boundaries and H and W are the black hole and white hole horizons, respectively.

V. SUPERRADIANCE FROM A KERR BLACK HOLE

In this section we use the second-order Einstein method to demonstrate that, in the framework of linearized gravity, there are superradiant gravitational modes in the Kerr space-time.

Consider that portion of the Kerr space-time that is exterior to the black hole as shown in Fig. 3. Let γ_{ab} be a solution of the linearized Einstein equation in Kerr that vanishes in a neighborhood of the white hole horizon and let $E^{(2)}{}_{\text{fut}}[\gamma_{ab}]$ and $E^{(2)}{}_{\text{past}}[\gamma_{ab}]$ denote the second-order contribution to the total Bondi energy flux evaluated on $\mathscr{I}^+$ and $\mathscr{I}^-$, respectively. We say that γ_{ab} is superradiant if $E^{(2)}{}_{\text{fut}}[\gamma_{ab}]>E^{(2)}{}_{\text{past}}[\gamma_{ab}]$.

Now, the second-order Einstein method has been presented as a method for calculating $E^{(2)}{}_{\text{fut}}[\gamma_{ab}]$ when γ_{ab} vanishes in the past of some spacelike three-surface intersecting $\mathscr{I}^+$. We could also use this method to calculate $E^{(2)}{}_{\text{past}}[\gamma_{ab}]$ when γ_{ab} vanishes in the future of some spacelike three-surface intersecting $\mathscr{I}^-$. This particular falloff condition (i.e., the vanishing of γ_{ab} in the past or future of some spacelike surface intersecting $\mathscr{I}$) essentially required γ_{ab} to vanish in a neighborhood of spacelike infinity. However, this condition was imposed to make the proofs of our theorems manageable. We expect that if γ_{ab} does not vanish near spacelike infinity, but instead falls off at an appropriate rate at spacelike infinity, then the second-order Einstein method would actually calculate $E^{(2)}{}_{\text{fut}}[\gamma_{ab}]-E^{(2)}{}_{\text{past}}[\gamma_{ab}]$. We will assume this is the case. Hence γ_{ab} is superradiant if its associated total energy flux calculated using the second-order Einstein method is positive. Also because the second-order Einstein tensor is divergence-free, the integral defining $\zeta[\gamma_{ab}]$ may be evaluated on the black hole event horizon, which we denote as H. We conclude that γ_{ab} is superradiant if

$$(8\pi)^{-1}\int_H G^{(2)}{}_{ab}[\gamma_{cd}]t^a\epsilon^b{}_{efh}>0, \tag{5.1}$$

where t^a is the timelike Killing field in Kerr (even though it is not timelike on the horizon).

Because Kerr also has a rotational Killing field φ^a, we may decompose γ_{ab} into modes. A mode is a solution to the linearized Einstein equation of the form

$$\gamma^{(\omega,m)}{}_{ab} = \text{Re}\{f_{ab}\exp(-i\omega t+im\varphi)\}, \tag{5.2}$$

where t and φ are the time and rotational Boyer–Linquist coordinates in Kerr, and f_{ab} is a complex tensor field that is Lie derived by t^a and φ^a. Since the contribution to the inte-

gral in (5.1) due to each mode that comprises γ_{ab} is independent of the other modes comprising γ_{ab}, it is sufficient to study the value of the integral in (5.1) for each mode. In order to avoid infinities when evaluating the integral in (5.1) for a mode we will only integrate over a compact portion of the horizon that corresponds to one period of the mode in the timelike direction. We denote this compact region as H'. Our project, then, is to discover which modes, if any, superradiate.

We start by listing the relevant geometric structures in the Kerr space-time that we will use. Let γ_{ab} be a mode of frequency ω and azimuthal quantum number m and define $\zeta'[\gamma_{ab}]$ by the formula

$$\zeta'[\gamma_{ab}] = (8\pi)^{-1}\int_{H'} G^{(2)}_{ab}[\gamma_{cd}]\, t^a \epsilon^b{}_{efh}. \tag{5.3}$$

Also let

$$\psi^a = t^a + \Omega_H \varphi^a \tag{5.4}$$

be the Killing field that is normal to, and null on, the horizon. Ω_H is a positive quantity sometimes referred to as the angular velocity of the horizon. Let

$$K^a = t^a + \omega\psi^a/(m\Omega_H - \omega) \tag{5.5}$$

denote the Killing field that Lie derives γ_{ab}. The surface gravity, which is a constant, is denoted as κ. It is defined by

$$\tfrac{1}{2}\nabla_a(\psi^b\psi_b) = -\kappa\psi_a \quad \text{on the horizon.} \tag{5.6}$$

One other fact we will find useful is that the derivative of ψ_a on the horizon has the form

$$\nabla_a\psi_b = \psi_{[a}v_{b]} \tag{5.7}$$

for some covector field v_a. It follows that on the horizon we get

$$\psi^a v_a = -2\kappa. \tag{5.8}$$

We will demonstrate that, for a mode,

$$\operatorname{sgn}\{\zeta'[\gamma_{ab}]\} = \operatorname{sgn}\{m\Omega_H - \omega\}. \tag{5.9}$$

Therefore, the mode is superradiant if $m\Omega_H > \omega$. Below we present three lemmas that we will employ to demonstrate the validity of (5.9).

Lemma 1: The integral in (5.3) is gauge invariant in the following sense:

$$\zeta'[\gamma_{ab}] = \zeta'[\gamma_{ab} + \nabla_{(a}\xi_{b)}],$$

as long as ξ_b is of the form

$$\xi_a = \operatorname{Re}\{f_a \exp(-i\omega t + im\varphi)\}, \tag{5.10}$$

where f_a is a complex convector Lie derived by t^a and φ^a.

The proof that (5.3) is gauge invariant is similar to the proof of gauge invariance in Sec. III except for the way the boundary integrals are treated. In Sec. III the boundary integrals were shown to go to zero. Here the boundary integrals cancel each other because of the periodicity of γ_{ab} and ξ_a.

Lemma 2: There exists a gauge transformation satisfying the hypothesis of Lemma 1 which brings γ_{ab} into a gauge satisfying (5.11) and (5.12) on the horizon:

$$\gamma_{ab}\psi^a\big|_H = 0, \tag{5.11}$$

$$\gamma^a{}_a\big|_H = 0. \tag{5.12}$$

For the proof of this lemma assume that γ_{ab} still contains its imaginary part. That is, γ_{ab} is given by (5.2) without taking the real part. We will now proceed to find a complex gauge transformation that will bring γ_{ab} into a gauge satisfying (5.11) and (5.12). By taking the real part of this expression we produce the desired real gauge transformation.

Consider the expression

$$\gamma'_{ab} = \gamma_{ab} + \nabla_{(a}(p\gamma_{b)c}\psi^c), \tag{5.13}$$

where p is a complex constant. Contracting (5.13) with $\psi^a\psi_b$ and using (5.8) and evaluating the expression on the horizon we get

$$\psi^a\psi^b\gamma'_{ab} = \psi^a\psi^b\gamma_{ab} + p\pounds_\psi(\psi^a\psi^b\gamma_{ab}) - p\kappa\psi^a\psi^b\gamma_{ab}. \tag{5.14}$$

Since $\pounds_\psi(\psi^a\psi^b\gamma_{ab}) - \kappa\psi^a\psi_b\gamma_{ab}$ is nonvanishing (unless $\psi^a\psi_b\gamma_{ab}$ vanishes) and is proportional to $\psi^a\psi^b\gamma_{ab}$, we can choose p such that $\psi^a\psi^b\gamma'_{ab}$ vanishes on the horizon. Now consider a further gauge transformation as follows:

$$\gamma''_{ab} = \gamma'_{ab} + \nabla_{(a}(q\gamma'_{b)c}\psi^c), \tag{5.15}$$

where q is a complex constant. Contracting with ψ^b we get

$$\gamma''_{ab}\psi^b = \gamma'_{ab}\psi^b + \tfrac{1}{2}q\pounds_\psi(\psi^b\gamma'_{ab}) + \tfrac{1}{2}q\nabla_a(\psi^b\psi^c\gamma'_{bc}) - q\psi^b\gamma'_{cb}\nabla_a\psi^c. \tag{5.16}$$

By choosing q so that the first two terms in (5.16) cancel we get

$$\gamma''_{ab}\psi^b = \tfrac{1}{2}q\nabla_a(\psi^b\psi^c\gamma'_{bc}) - q\psi^b\gamma'_{cb}\nabla_a\psi^c. \tag{5.17}$$

Using (5.8) and the fact that $\psi^b\psi^c\gamma'_{bc} = 0$ on the horizon, it follows that $\gamma''_{ab}\psi^b$ is proportional to ψ_a on the horizon. Using (5.8) again we find that

$$\gamma''_{ab}\psi^b = \psi_a\gamma''_{bc}(\nabla^d\psi^b)(\nabla^c\psi_d)/(\kappa^2) \tag{5.18}$$

on the horizon. Now consider our third and final gauge transformation

$$\gamma'''_{ab} = \gamma''_{ab} + \nabla_{(a}f\psi_{b)}, \tag{5.19}$$

where f is given by

$$f = r\gamma''_{bc}(\nabla^d{}_\psi{}^b)(\nabla^c\psi_d)/(\kappa^2) \tag{5.20}$$

and r is a complex constant. Contracting (5.19) with ψ^b and evaluating the expression on the horizon we get

$$\gamma'''_{ab}\psi^b = \gamma''_{ab}\psi^b + \tfrac{1}{2}\pounds_\psi(\psi_a f) = \gamma''_{ab}\psi^b + \tfrac{1}{2}r\pounds_\psi(\gamma''_{ab}\psi^b). \tag{5.21}$$

Clearly r can be chosen so that the expression in (5.21) vanishes on the horizon. We have transformed γ_{ab} into γ'''_{ab}, a gauge in which $\gamma'''_{ab}\psi^b$ vanishes on the horizon. $\gamma'''^a{}_a = 0$ on the horizon is a consequence of this particular gauge choice, the fact that γ'''_{ab} is a mode and the fact that γ'''_{ab} satisfies the linearized Einstein equation in Kerr.

Lemma 3: The value of the integral in (5.3) is unchanged if we substitute $-\omega\psi^a/(m\Omega_H - \omega)$ for t^a.

To prove Lemma 3, consider the integral

$$J = (8\pi)^{-1}\int_H G^{(2)}{}_{ab}[\gamma_{cd}]K^a\epsilon^b{}_{efh}, \tag{5.22}$$

which is similar to (5.3) except t^a is replaced with K^a. We will show this integral vanishes. Because γ_{ab} satisfies the

linearized Einstein equation we may rewrite J as

$$J=(8\pi)^{-1}\int_{H'}\left(R^{(2)}{}_{ab}[\gamma_{cd}]-\frac{1}{2}R^{(2)}[\gamma_{cd}]g_{ab}\right)K^a\epsilon^b{}_{efh}, \tag{5.23}$$

where $R^{(2)}{}_{ab}[\gamma_{cd}]$ and $R^{(2)}[\gamma_{cd}]$ are the second-order Ricci tensor and second-order scalar curvature defined in a similar manner to the second-order Einstein tensor. Since K^a is tangent to the horizon, the second term in (5.23) vanishes. We now write the first term in (5.23) as

$$J=(16\pi)^{-1}\frac{d^2}{d\lambda^2}\int_{H'}(\nabla(\lambda)_a\nabla(\lambda)_b K^a - \nabla(\lambda)_b\nabla(\lambda)_a K^a)g(\lambda)^{bc}\epsilon(\lambda)_{cefh}\big|_{\lambda=0}, \tag{5.24}$$

where $\epsilon(\lambda)_{abcd}$ and $\nabla(\lambda)_a$ are the volume element and derivative operator associated with the metric $g(\lambda)_{ab}$ defined by

$$g(\lambda)_{ab}=g_{ab}+\lambda\gamma_{ab} \tag{5.25}$$

with inverse $g(\lambda)^{ab}$. We note that parametrizing the volume element does not affect the value of J because the fact that the background Ricci tensor and the linearized Ricci tensor vanish implies that perturbations of the volume element contribute in (5.24) to terms third order or higher in λ. For the same reason, we are able to raise the index on the volume element in (5.24) with $g(\lambda)^{bc}$. For the rest of this section we will, by convention, raise and lower indices on tensors parametrized by λ with $g(\lambda)_{ab}$.

Because K^a Lie derives γ_{ab} we have

$$\nabla(\lambda)_a K^a=\tfrac{1}{2}g(\lambda)^{ab}\lambda\pounds_K\gamma_{ab}=0. \tag{5.26}$$

Hence, the second term in (5.24) vanishes. The first term in (5.24) can be split into a symmetric and antisymmetric part as follows:

$$J=(16\pi)^{-1}\frac{d^2}{d\lambda^2}\int_{H'}\{\nabla(\lambda)^a\nabla(\lambda)_{(b}K(\lambda)_{a)} + \nabla(\lambda)^a\nabla(\lambda)_{[b}K(\lambda)_{a]}\}\epsilon(\lambda)^b{}_{efh}\big|_{\lambda=0}, \tag{5.27}$$

where $K(\lambda)_a=K^b g(\lambda)_{ab}$ and the index on the λ-dependent derivative operator was raised with the λ-dependent metric. The symmetric term in (5.27) is zero because K^a Lie derives γ_{ab} and the antisymmetric term integrates by parts to zero because of the periodicity of γ_{ab}. Therefore J vanishes. Using (5.5) we can rewrite (5.3) as

$$\zeta'[\gamma_{ab}] = -\omega(8\pi)^{-1}(m\Omega_H-\omega)^{-1}\int_{H'}G^{(2)}{}_{ab}[\gamma_{cd}]\psi^a\epsilon^b{}_{efh}. \tag{5.28}$$

This proves the lemma.

We now evaluate $\zeta'[\gamma_{ab}]$ to demonstrate the validity of (5.9). In light of Lemmas 1 and 2 we may assume that γ_{ab} is in the gauge discussed in Lemma 2. Because ψ^a is tangent to the horizon, the scalar curvature part of the second-order Einstein tensor vanishes in (5.28). We can write the second-order Ricci part of (5.28), using the parametrized derivative operators introduced in the proof of Lemma 3, as follows:

$$\zeta'[\gamma_{ab}] = -\omega(16\pi)^{-1}(m\Omega_H-\omega)^{-1}\frac{d^2}{d\lambda^2}\times\int_{H'}\{\nabla(\lambda)^a\nabla(\lambda)_{(b}\psi(\lambda)_{a)} + \nabla(\lambda)^a\nabla(\lambda)_{[b}\psi(\lambda)_{a]} - \nabla(\lambda)_b\nabla(\lambda)_a\psi^a\}\epsilon(\lambda)^b{}_{efh}\big|_{\lambda=0}, \tag{5.29}$$

where $\psi(\lambda)_a=\psi^b g(\lambda)_{ab}$. The second and third terms in (5.29) both integrate by parts to zero due to the periodicity of γ_{ab}. The first term in (5.29) can be written as

$$\zeta'[\gamma_{ab}] = -\omega(32\pi)^{-1}(m\Omega_H-\omega)^{-1}\frac{d^2}{d\lambda^2}\times\int_{H'}(\lambda\nabla(\lambda)^a\pounds_\psi\gamma_{ab})\epsilon(\lambda)^b{}_{efh}\Big|_{\lambda=0}. \tag{5.30}$$

Now, the vanishing of the trace of our perturbation implies that the first-order change in $\epsilon(\lambda)_{befh}$ vanishes. Likewise, the λ dependence introduced in (5.30) by raising the index of the volume element with $g(\lambda)^{ab}$ does not contribute to the λ dependence of the integral because our perturbation is orthogonal to the normal of the horizon [i.e., raising and lowering the index of the horizon's normal vector field is independent of λ due to (5.11)]. Therefore, we replace $\epsilon(\lambda)^b{}_{efh}$ with $\psi^b\epsilon_{efh}$, where ϵ_{efh} is the appropriate, unparametrized, volume element on H. Equation (5.30) now becomes

$$\zeta'[\gamma_{ab}] = -\omega(32\pi)^{-1}(m\Omega_H-\omega)^{-1}\frac{d^2}{d\lambda^2}\times\int_{H'}\{\lambda\nabla(\lambda)^a\pounds_\psi(\gamma_{ab}\psi^b) - \lambda\pounds_\psi(\gamma_{ab})\nabla(\lambda)^{(a}\psi^{b)}\}\epsilon_{efh}\big|_{\lambda=0}. \tag{5.31}$$

Because $\pounds_\psi(\gamma_{ab}\psi^b)$ vanishes on the horizon we may use the background derivative operator to evaluate the first term in (5.31) as

$$\frac{d^2}{d\lambda^2}(\lambda\nabla(\lambda)^a\pounds_\psi(\gamma_{ab}\psi^b))\Big|_{H,\lambda=0} = -2\gamma^{ac}\nabla_c\pounds_\psi(\gamma_{ab}\psi^b). \tag{5.32}$$

Here $\pounds_\psi(\gamma_{ab}\psi^b)$ vanishing on the horizon implies that $\nabla_c\pounds_\psi(\gamma_{ab}\psi^b)$ on the horizon is proportional to $\psi_c r_a$ for some r_a. Therefore (5.11) implies that (5.32) vanishes. Using

$$\left(\frac{d}{d\lambda}\right)\nabla(\lambda)^{(a}\psi^{b)}\Big|_{\lambda=0}=\left(\frac{1}{2}\right)\pounds_\psi\gamma_{ab}, \tag{5.33}$$

we evaluate (5.31) as

$$\zeta'[\gamma_{ab}] = \omega(32\pi)^{-1}(m\Omega_H-\omega)^{-1}\int_{H'}(\pounds_\psi\gamma^{ab})(\pounds_\psi\gamma_{ab})\epsilon_{efh}. \tag{5.34}$$

Now $(\pounds_\psi\gamma^{ab})(\pounds_\psi\gamma_{ab})$ is positive due to our gauge choice (5.11). Therefore $\zeta'[\gamma_b]$ has the same sign as $m\Omega_H-\omega$. We conclude that a mode will be superradiant if $m\Omega_H>\omega$.

ACKNOWLEDGMENTS

I would like to thank R. Wald for many helpful discussions. This work was submitted in partial fulfillment of the

requirements for a Ph.D. degree at the University of Chicago.

This research was supported in part by National Science Foundation Grant No. PHY84-16691 to the University of Chicago.

APPENDIX: A KILLING FIELD THEOREM

We prove a theorem that makes precise the notion that if $\nabla_{(a}\xi_{b)}$ is "small" then ξ_a is close to a Killing field.

Theorem: Let M be a connected compact manifold, possibly with boundary, with metric g_{ab} and associated derivative operator ∇_a. Also equip M with a positive definite metric h^{ab} for taking norms of tensors, i.e.,

$$|T_{ab}| = \{T_{ab}T_{cd}h^{ac}h^{bd}\}^{1/2}.$$

Then there exists a $c>0$ depending only on M, g_{ab}, and h^{ab} such that given a covector field ξ_a with $V_{ab} = \nabla_{(a}\xi_{b)}$ satisfying

$$\sup\{|V_{ab}| + |\nabla_{[a}V_{b]c}|\} < \epsilon, \tag{A1}$$

for some ϵ, then there exists a ξ'_a satisfying

$$\sup\{|\xi'_a| + |\nabla_{[a}\xi'_{b]}|\} < c\epsilon, \tag{A2}$$

which differs from ξ_a by a Killing field, that is, $V_{ab} = \nabla_{(a}\xi'_{b)}$.

Proof: For the purpose of obtaining a contradiction suppose there is no such c. Then there will exist a sequence $\{\xi^{(i)}{}_a\}$ of covector fields satisfying

$$\sup\{|\xi^{(i)}{}_a + k_a| + |\nabla_{[a}\xi^{(i)}{}_{b]} + \nabla_{[a}k_{b]}|\} \geqslant 1, \tag{A3}$$

for all Killing fields k_a and its sequence of symmetrized derivatives $\{V^{(i)}{}_{ab}\}$, where $V^{i}{}_{ab} = \nabla_{(a}\xi^{(i)}{}_{b)}$ will satisfy

$$\lim_{i\to\infty}[\sup\{|V^{(i)}{}_{ab}| + |\nabla_{[a}V^{(i)}{}_{b]c}|\}] = 0. \tag{A4}$$

We will show that (A3) implies there exists a subsequence of $\{\sup\{|V^{(i)}{}_{ab}| + |\nabla_{[a}V^{(i)}{}_{b]c}|\}\}$ bounded away from zero, hence (A4) cannot hold.

Without loss of generality we may assume that the equality is attained in (A3) for all i when k_a vanishes. Also, since M is compact the supremum in (A3) is always attained. Therefore for each $\xi^{(i)}$ there is a point p^i for which

$$(|\xi^{(i)}{}_a| + |\nabla_{[a}\xi^{(i)}{}_{b]}|)|_p i = 1.$$

Let

$$\eta^i = (\xi^{(i)}{}_a, \nabla_{[a}\xi^{(i)}{}_{b]})|_p i.$$

For convenience let us denote the vector bundle that is the direct sum of the cotangent bundle and the bundle of two-forms as B. Then $\eta^i \in B$. Now, the subbundle of B which consists of all possible pairs of covectors and two-forms having norms that sum to 1 form an s^9 bundle over M that is compact. Hence, the sequence $\{\eta^i\}$ will have a point of accumulation, say η. Let the base point of η in M be denoted p.

We will arrive at our contradiction by arguing that there exists a neighborhood U_η of η in B and a $\delta>0$ such that if the cross section $(\xi^{(i)}{}_a, \nabla_{[a}\xi^{(i)}{}_{b]})$ of B intersects U_η, then

$$\sup\{|V^{(i)}{}_{ab}| + |\nabla_{[a}V^{(i)}{}_{b]c}|\} \geqslant \delta. \tag{A5}$$

To prove the existence of U_η and δ we consider the following two cases.

First suppose that there is no Killing field k_a for which $(k_a, \nabla_{[a}k_{b]})|_p = \eta$. We will argue that there must exist a closed curve α in M beginning and ending at p such that transporting η around α using the transport equations for (ξ_a, F_{ab})

$$\alpha^a\nabla_a\xi_b = \alpha^a F_{ab}, \tag{A6a}$$

$$\alpha^a\nabla_a F_{bc} = \alpha^a R_{cba}{}^d\xi_d \tag{A6b}$$

(where α^a is the tangent vector to α) results in a point in B, say η', different from η. Suppose α did not exist. Then this type of transport, which is called Killing transport, would be independent of path. Hence we could uniquely construct the entire covector field ξ_a by transporting η over the entire manifold using Killing transport. Therefore ξ_a must be Killing by virtue of (A6a) and (A6b). But we have assumed that η is not data for any Killing field. Therefore α must exist.

Solutions to transport equations like (A6a) and (A6b) are continuous in both initial data and coefficients. Therefore sufficiently small modifications of the transport equations, of the curve α, and of the initial data η, will not affect the result that transporting the initial data around the curve produces a pair of tensors different from the initial data.

We cast this result in a useful form in the following statement. There is a neighborhood U_η of η and a $\delta>0$ such that if W_{ab} is a tensor field satisfying

$$\sup\{|W_{ab}| + |\nabla_{[a}W_{b]c}|\} < \delta \tag{A7}$$

then for every point $v \in U_\eta$ there is a closed curve $\beta(v)$ around which we can transport v using

$$\beta(v)^a\nabla_a\xi_b = \beta(v)^a F_{ab} + \beta^a W_{ab}, \tag{A8a}$$

$$\beta(v)^a\nabla_a F_{bc} = \beta(v)^a R_{cba}{}^d\xi_d + 2\beta(v)^a\nabla_{[b}W_{c]a}, \tag{A8b}$$

and the result of this transport is an element of B different from v.

But $\xi_a = \xi^{(i)}{}_a$, $F_{ab} = \nabla_{[a}\xi^{(i)}{}_{b]}$, and $W_{ab} = \nabla_{(a}\xi^{(i)}{}_{b)}$ identically satisfy the transport equations (A8a) and (A8b). Set $W_{ab} = \nabla_{(a}\xi^{(i)}{}_{b)}$ in (A8a) and (A8b) and use these equations to transport initial data that is attained by $(\xi^{(i)}{}_a, \nabla_{[a}\xi^{(i)}{}_{b]})$ at p around any closed loop. The result must be equal to the initial data. We conclude that if $(\xi^{(i)}{}_a, \nabla_{[a}\xi^{(i)}{}_{b]})$ intersects U_η then $V^{(i)}{}_{ab}$ must satisfy (A5).

Now suppose $(k_a, \nabla_{[a}k_{b]})|_p = \eta$ for some Killing field k_a, and denote $\eta_0 = (0_a, 0_{[ab]})|_p$ as the pair consisting of the zero covector and the zero two-form at p. Let us connect every point in M to p by curves. Because M is compact, we can assume the lengths of the curves measured with h^{ab} are bounded. Using Killing transport to transport η_0 along any of these curves will result in the zero pair. By continuity of the transport equations there will exist a $\delta>0$ with the property that if W_{ab} satisfies (A7) and $\chi = (f_a, f_{ab})|_q$ is the result of transporting η_0 along one of these curves using the modified transport equations (A8a) and (A8b), then

$$(|f_a| + |f_{ab}|)|_q < 1. \tag{A9}$$

Again because the transport equations are continuous in their parameters, we can expand the above result to a neighborhood of η_0 in the following sense. There exists a $\delta>0$ and a neighborhood U_0 of η_0 such that if $(\xi_a, \nabla_{[a}\xi_{b]})$ intersects U_0 and satisfies

$$\sup\{|\xi_a| + |\nabla_{[a}\xi_{b]}|\} \geq 1, \tag{A10}$$

then (A5) holds where $\nabla_{(a}\xi_{b)}$ replaces $V^{(i)}{}_{ab}$. We are now ready to finish the proof. Translate U_0 by k_a to get U_η, a neighborhood of η. If $(\xi^{(i)}{}_a, \nabla_{[a}\xi^{(i)}{}_{b]})$ intersects U_η then $(\xi^{(i)}{}_a - k_a, \nabla_{[a}\xi^{(i)}{}_{b]} - \nabla_{[a}k_{b]})$ intersects U_0. Hence $V^{(i)}{}_{ab}$ satisfies (A5). Q.E.D.

[1] R. P. Geroch, "Asymptotic structure of space-time," in *Asymptotic Structure of Space-Time*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977).

[2] S. Persides and D. Papadopoulos, Gen. Relativ. Gravit. **11**, 233 (1979).

[3] R. M. Wald, *General Relativity* (Univ. Chicago P., Chicago, 1984).

[4] S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).

[5] R. P. Geroch and B. C. Xanthopoulos, J. Math. Phys. **19**, 714 (1978).

[6] W. H. Press and S. A. Teukolsky, Astrophys. J. **193**, 443 (1974).

Short-time nonstationary character of correlation functions in classical equilibrium ensembles

C. A. Chatzidimitriou-Dreismann
Churchill College, University of Cambridge, Cambridge CB3 0DS, England and I. N. Stranski-Institute for Physical and Theoretical Chemistry, Technical University of Berlin, D-1000 Berlin 12, Federal Republic of Germany

(Received 25 March 1986; accepted for publication 2 July 1986)

Dynamical quantities $b(t)$ are considered that depend on the canonical variables of a small number n of particles of a classical N-particle condensed system being in thermal equilibrium. It is proved that equalities like $d^2\langle b(0)b(t)\rangle/dt^2 = -\langle \dot{b}(0)\dot{b}(t)\rangle$ are in general violated, if the interparticle interactions have a finite range and sufficiently short times t are considered. This violation reflects the continuous-in-time creations and destructions of s-particle correlations $n < s \ll N$, which are due to the thermal motion.

I. INTRODUCTION

In previous work,[1] we have studied some quantum mechanical aspects of the molecular motion in liquids, in the picosecond time range. In particular, the connection between (i) the molecular reorientation of small polar molecules and (ii) the corresponding far-infrared (FIR) band shapes has been treated within the formal framework of linear response theory (LRT). In these investigations we have found that the equation

$$\frac{d^2}{dt^2}\langle \hat{e}(0)\hat{e}(t)\rangle = -\langle \dot{\hat{e}}(0)\dot{\hat{e}}(t)\rangle , \tag{1}$$

which seems to hold true very generally in the case of thermal equilibrium, may be violated in the physical context under consideration. [The unit vector $\hat{e}(t)$ represents the orientation of some characteristic molecular axis, e.g., the axis defined by the two nuclei of a diatomic molecule; the brackets represent equilibrium ensemble averages.] The physical reasons for the violation of Eq. (1) have been discussed in detail. In particular, van Vliet's[2] and van Kampen's[3] critical remarks concerning the standard version[4] of LRT have been explicitly taken into account; their importance in the interpretation of the anomalous temperature dependence[5,6] of the spectral FIR absorption bands has been shown.

Now let us consider a liquid as a *classical* dynamical system that contains N interacting particles, $N \sim 10^{23}$; the system is assumed to be in thermal equilibrium. In this paper it will be shown that Eq. (1) may be (and, in general, is) violated in the picosecond time range, if standard experimental conditions are considered. This statement appears to be surprising since

$$\frac{d^2}{dt^2}\langle a(0)b(t)\rangle = \frac{d}{dt}\langle a(0)\dot{b}(t)\rangle = \frac{d}{dt}\langle a(-t)\dot{b}(0)\rangle = -\langle \dot{a}(-t)\dot{b}(0)\rangle = -\langle \dot{a}(0)\dot{b}(t)\rangle . \tag{2}$$

(a and b may represent two arbitrary one- or many-body dynamical functions,[7] i.e., they are functions of specific numbers of generalized coordinates and/or momenta.) These equations seem to be valid, because (i) the complete equilibrium distribution function does not depend on time explicitly and (ii) the above correlation functions are invariant under time translation. Similar calculations can be found in many textbooks of statistical mechanics.

II. N-PARTICLE DYNAMICAL FUNCTIONS

For reasons of clarity and simplification, let us consider in the following a monoatomic liquid (or dense fluid) with two-body interaction potentials $H_2(x_i,x_j)$ between the particles i and j. Using the abbreviation[8] $x_j \equiv (\mathbf{q}_j,\mathbf{p}_j)$ for the set of canonical variables pertaining to particle j, the total Hamiltonian reads

$$H(x_1,\ldots,x_N) = \sum_{j=1}^{N} H_1(x_j) + \sum_{j<n=1}^{N}\sum H_2(x_i,x_j) . \tag{3}$$

Additionally it is assumed that the interactions H_2 have a *finite* spatial range R_0; i.e.,

$$H_2(x_i,x_j) = 0 , \quad \text{if } |\mathbf{q}_i - \mathbf{q}_j| > R_0 . \tag{4}$$

The corresponding Liouvillian reads[7]

$$L_N \equiv L_N(x_1,\ldots,x_N) = \sum^{N} L_j^0 + \sum\sum^{N} L'_{jn} , \tag{5}$$

with obvious notation. The complete distribution function $F \equiv F(x_1,\ldots,x_N)$ represents an equilibrium ensemble, and thus

$$\frac{\partial}{\partial t}F = L_N F = 0 . \tag{6}$$

The observables of the system are described by dynamical functions b of the canonical variables x_j. As the particles in the system are identical, we may restrict our interest to those functions in which all particles play the same role; these are the only ones that represent physically relevant quantities. [This remark implies the validity of Eq. (10), see below.] In order to make the presentation self-contained, let us just mention the following points.[9]

(i) The general form of a dynamical function reads

$$b(x_1,\ldots,x_N) = b_0 + \sum_{j=1}^{N} b_1(x_j) + \sum\sum^{N}_{j<n=1} b_2(x_j,x_n) + \cdots + b_N(x_1,\ldots,x_N) . \tag{7}$$

 0022-2488/86/112770-04$02.50

Here b_s is a function of s variables only, which cannot be decomposed into a sum of functions depending on less variables.

(ii) The definition of the reduced s-particle distribution functions $s \leqslant N$ is

$$f_s = \frac{N!}{(N-s)!} \int dx_{s+1} \cdots dx_N \, F(x_1,...,x_N) \, . \tag{8}$$

Hence, we may express the ensemble average of the dynamical function b completely in terms of reduced distribution functions[10]:

$$\langle b \rangle \equiv \int dx_1 \cdots dx_N \, F(x_1,...,x_N) b(x_1,...,x_N)$$
$$= \sum_{s=0}^{N} (s!)^{-1} \int dx_1 \cdots dx_s \, b_s(x_1,...,x_s) f_s(x_1,...,x_s) \, . \tag{9}$$

(iii) All functions f_s and b_s, $s \leqslant N$, are *symmetric* under permutation of any two variables.[10] For example, one has

$$b_s(\cdots x_i \cdots x_j \cdots) = b_s(\cdots x_j \cdots x_i \cdots) \, . \tag{10}$$

The particles of the dynamical system under consideration undergo thermal motions. Thus the canonical variables and the dynamical functions must fulfill the appropriate equations of motion. As is well known,[8]

$$\dot{b}_s = -L_N b_s \, , \tag{11}$$

$$x_j[t] = \exp(-L_N t) x_j \, , \tag{12}$$

and also

$$b_s(x_1[t],...,x_s[t]) = \exp(-L_N t) b_s(x_1,...,x_s) \, , \tag{13}$$

due to the fact that the time-evolution operator preserves the algebraic structure of the set of all the dynamical functions.[11] Here the abbreviation $x_j \equiv x_j[0]$ has been used.

Now let us consider the special case of the dynamical function $B(x_1,...) \equiv b_N(x_1,...,x_N)$, cf. Eq. (7). By definition one has[12]

$$\langle B(0)B(t) \rangle \equiv \int dx_1 \cdots dx_N \, B(x_1,...) \cdot B(x_1[t],...) \cdot F(x_1,...,x_N) \, , \tag{14}$$

$$\langle \dot{B}(0)\dot{B}(t) \rangle \equiv \int dx_1 \cdots dx_N [-L_N\{B(x_1,...)\}] \cdot [-L_N\{B(x_1[t],...)\}] \times F(x_1,...,x_N) \, . \tag{15}$$

From Eq. (14) it follows immediately

$$\frac{d^2}{dt^2} \langle B(0)B(t) \rangle = \int dx_1 \cdots dx_N \, B(x_1,...) \cdot [(-1)^2 L_N\{L_N\{B(x_1[t],...)\}\}] \times F(x_1,...,x_N) \, . \tag{16}$$

Since (i) L_N is a linear differential operator and (ii) F together with a sufficient number of its derivatives can be assumed to vanish at the boundaries of the system in configuration space and also for $\mathbf{p}_j = \pm \infty$,[7] one can carry out the appropriate partial integrations straightforwardly and obtain from Eq. (15)

$$\langle \dot{B}(0)\dot{B}(t) \rangle = -\int dx_1 \cdots dx_N \, B(x_1,...) \cdot [-L_N\{-L_N\{B(x_1[t],...)\} \times F(x_1,...,x_N)\}] \, , \tag{17}$$

which, with the aid of Eq. (6), proves that

$$\frac{d^2}{dt^2} \langle B(0)B(t) \rangle = -\langle \dot{B}(0)\dot{B}(t) \rangle \, . \tag{18}$$

It is easily seen that this equation can be proved also by starting with Eq. (16), and then by carrying out the appropriate partial integrations that bring L_N "on the left" of $B(x_1,...)$.

III. s-PARTICLE DYNAMICAL FUNCTIONS ($s \ll N$)

Now we are in the position to study in detail the validity of equations like (1), (2), (18), etc. For simplicity, let us for the moment consider a *one*-particle dynamical function b. The second time derivative of the correlation function $\langle b(0)b(t) \rangle$ is[7]

$$\frac{d^2}{dt^2} \langle b(0)b(t) \rangle = \int dx_1 \cdots dx_N \, F(x_1,...,x_N) \times \sum_{j=1}^{N} b_1(x_j) \cdot [L_N\{L_N\{b_1(x_j[t])\}\}] \, . \tag{19}$$

We also have

$$\langle \dot{b}(0)\dot{b}(t) \rangle = \int dx_1 \cdots dx_N \, F(x_1,...,x_N) \times \sum_{j=1}^{N} [-L_N\{b_1(x_j)\}] \cdot [-L_N\{b_1(x_j[t])\}] \, . \tag{20}$$

Before the integrations on the right-hand side can be carried out, however, the variables $x_j[t]$ must be substituted with expressions depending on the integration variables x_j, $j = 1,...,N$. But it is clear that $x_j[t]$ is, in general, a complicated function of several variables, which are the "initial values" in the dynamical problem involving $x_j[t]$. This is, of course, due to the interactions L'_{jn}, Eq. (4). Thus, in the limit $t \to \infty$, $x_j[t]$ becomes a function of all the variables x_j ($j = 1,...,N$). Therefore the following identities hold true:

$$x_j[t] \equiv g(x_j,...;j;t) \, . \tag{21}$$

The functions g depend parametrically on the particular particle number j as well as the time t.

Now let t be sufficient "small," in the sense that the function $g(x_j,...;j;t)$ depends on a small number $u(j)$ of variables, $u(j) \ll N$. The terms

$$V_{jt} \equiv L_N\{b_1(x_j[t])\} \quad \text{and} \quad W_{jt} \equiv L_N\{V_{jt}\} \tag{22}$$

depend then on specific numbers $v(j)$ and $w(j)$ of variables, respectively. Due to the dynamical couplings which are caused by the interaction Hamiltonian, one has very generally

$$u(j) \leqslant v(j) \leqslant w(j) \ll N \, , \tag{23}$$

for all j (and for sufficient small t and R_0). For the jth summation term appearing in the right-hand side (rhs) of Eq. (19), one obtains

$$S_j \equiv \int dx_1 \cdots dx_N \, F(x_1,\ldots,x_N) \cdot b_1(x_j) \cdot L_N\{L_N\{b_1(x_j[t])\}\}$$
$$= \int dx_1 \cdots dx_N \, F(x_1,\ldots,x_N) \cdot b_1(x_1) \cdot L_N\{L_N\{b_1(g(x_1,\ldots,x_{u(j)};j;t))\}\} . \quad (24)$$

The renumbering of the integration variables on the rhs of this equation is permitted, because of the aforementioned symmetry properties of the functions b_s and F under permutation of the variables. Correspondingly, the jth summation term T_j in the rhs of Eq. (20) reads

$$T_j \equiv \int dx_1 \cdots dx_N \, F(x_1,\ldots x_N) \cdot L_N\{b_1(x_1)\} \cdot L_N\{b_1(g(x_1,\ldots x_{u(j)};j;t))\} . \quad (25)$$

From the above considerations it immediately follows that the factor $L_N\{b_1(x_1)\}$ depends on those variables that appear in the factor $L_N\{b_1(g(\ldots))\}$, too. Thus, those two factors depend on the $v(j)$ variables $x,\ldots,x_{v(j)}$. Let m be the maximum of $v(j)$ and $w(j)$. One is now in the position to carry out immediately the *same* number $(N-m)$ of integrations over the variables $x_{m+1},\ldots,x_N$, in the rhs of equations (24) and (25). Thus,

$$S_j = \frac{N!}{(N-m)!} \int dx_1 \cdots dx_m \, f_m(x_1,\ldots,x_m) \cdot b_1(x_1) \cdot L_N\{V_{jt}\} , \quad (26)$$

$$T_j = \frac{N!}{(N-m)!} \int dx_1 \cdots dx_m \, f_m(x_1,\ldots,x_m) \cdot L_N\{b_1(x_1)\} \cdot V_{jt} . \quad (27)$$

In these formulas, of course, only a part of the complete Liouvillian L_N becomes "active," since, by definition of the number m, it is true that

$$L'_{kl}\{V_{jt}\} \equiv 0 , \quad \text{for } k,l > m . \quad (28)$$

For this reason, the replacement

$$L_N \triangleq L_m \quad (29)$$

in Eqs. (26) and (27) is permitted; cf. definition (5).

In order to prove the desired equation $T_j = -S_j$, one would try to make use of the partial integrations that proved Eq. (18). Because of the linearity of L_m and the vanishing of f_m at the "boundaries" of the system [cf. the remarks after Eq. (16)], one obtains

$$S_j = -\frac{N!}{(N-m)!} \int dx_1 \cdots dx_m \, V_{jt} \cdot L_m\{f_m(x_1,\ldots,x_m) b_1(x_1)\}$$
$$= -\frac{N!}{(N-m)!} \int dx_1 \cdots dx_m \, V_{jt} \cdot b_1(x_1) \cdot L_m\{f_m(x_1,\ldots,x_m)\} - T_j . \quad (30)$$

But, in general, the first term on the rhs does *not vanish identically*. This follows with the aid of the *BBGKY hierarchy*[7]: in the special case of equilibrium ensembles $(\partial f_s/\partial t) = 0$, $s \leqslant N$, and one obtains explicitly

$$L_m\{f_m\} = \sum_{i=1}^{m} \int dx_{m+1} \, L'_{im+1}\{f_{m+1}(x_1,\ldots,x_{m+1})\} . \quad (31)$$

Thus it follows that, in general $S_j + T_j \neq 0$, for each $j = 1,\ldots,N$. Therefore we conclude that, for $t \neq 0$ (cf. Ref. 13),

$$\frac{d^2}{dt^2} \langle b(0)b(t)\rangle \neq -\langle \dot{b}(0)\dot{b}(t)\rangle , \quad (32)$$

in opposition to the specific result (18). Q.E.D.

IV. DISCUSSION

From the above considerations one obtains easily the following *generalizations*.

(1) The result (32) holds true also in the case of polyatomic particles; in this case, however, the interaction Liouvillians must be characterized with sufficiently more indices.

(2) The explicit introduction of the "reduced" Liouvillians L_m is illustrative but not crucial. This follows easily from the above Eq. (28) and Eqs. (3.4.4) of Ref. 7.

(3) The quantities b can also depend at $t = 0$ on the degrees of freedom of several particles.

(4) The result (32) holds true also for the "cross-correlation function" $\langle a(0)b(t)\rangle$, cf. Eqs. (2).

(5) Similar relations concerning many-time correlation functions $\langle a(0)b(t)c(t+t_1)\cdots\rangle$ may also be affected by the above considerations, if strict stationarity [as in Eqs. (2)] is used by their derivation.

(6) The violation (32) also holds true for *stationary* nonequilibrium ensembles that represent systems in steady states. This is due to the fact that in the presented derivation only the conditions $\partial f_s/\partial t = 0$, $s \leqslant N$ have been used. Equation (31) still remains valid in this case; cf. Ref. 7.

It can be shown that the assumption (4) concerning the finite range R_0 of the interactions is also not necessary.[14] Here, however, let us merely state on physical grounds that (i) the repulsive potentials do have a very short R_0 and (ii) many theoretical long-range potentials have also a relatively small R_0 in condensed matter, due to shielding effects.

The result (32) reflects the continuous-in-time creations and destructions of s-particle correlations between each particle j and $(s-1)$ particles of its environment $s \geqslant m+1$; this process is due to the thermal motion. It is easily seen that, with increasing time, the numbers $w(j)$ and m will also increase. Thus, for sufficient large t, both sides in (32) will become equal, because (i) it will be $m = N$ or (ii) the BBGKY hierarchy can be truncated.

These considerations also illustrate the physical reason for the violation (32) together with the validity of Eq. (18): The Liouvillian L, acting on f_s, $s \leqslant N$, reveals information concerning correlations in an N-particle system (that are time independent for equilibrium ensembles, by definition). On the other hand, the action of L on $b(t)$ determines the dynamical function $\dot{b}(t)$ (which, in general, is not a constant of the motion). But the partial integrations leading to Eqs.

(17) and (30) also interchange the quantity on which L acts, drastically revealing this subtle difference.[15]

ACKNOWLEDGMENTS

I wish to thank the Fonds der Chemischen Industrie (Frankfurt am Main) and the Churchill College (Cambridge) for the award of the "German Fellowship 1985/86." Financial support by the Fonds der Chemischen Industrie und the Deutsche Forschungsgemeinschaft (Bonn) is gratefully acknowledged.

[1]For a review, see Chap. 3 in E. Lippert, C. A. Chatzidimitriou-Dreismann, and K.-H. Naumann, Adv. Chem. Phys. **57**, 311 (1984).
[2]K. M. van Vliet, J. Math. Phys. **19**, 1345 (1978).
[3]N. G. van Kampen, Phys. Norv. **5**, 279 (1971).
[4]R. Kubo, J. Phys. Soc. Jpn. **12**, 570 (1957).
[5]S. G. Kroon and J. van der Elsken, Chem. Phys. Lett. **1**, 285 (1967).
[6]See also references cited in Ref. 1.
[7]R. Balescu, *Equilibrium and Nonequilibrium Statistical Mechanics* (Wiley, New York, 1975).
[8]The notations of Ref. 7 are widely used in this paper.
[9]For an excellent presentation, see Ref. 7.
[10]Cf. Ref. 7, Chap. 3.
[11]Cf. Ref. 7, Chap. 1. This set is called the *dynamical algebra*.
[12]The braces{ } indicate the range of action of each L_N.
[13]For smooth interactions, it is trivial to show that both sides in (32) become equal for $t = 0$. This time point, however, should be excluded from the context under consideration, because in the case of hard-core particles a singularity appears in the velocity correlation function; e.g., see I. M. de Schepper, Phys. Rev. A **24**, 2789 (1981).
[14]This proof is more involved and is in preparation.
[15]Molecular dynamics simulations would be able to show the effect under consideration.

The free energy of the three-dimensional Zamolodchikov model

M. L. Glasser
Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13676

(Received 8 April 1986; accepted for publication 18 June 1986)

Details are given for the reduction of Baxter's expression for the exact partition function per site of the Zamolodchikov model to explicit form.

I. INTRODUCTION

In his remarkable exact solution for the free energy per site of the Zamolodchikov model on a cubic lattice, Baxter[1] obtains the expression

$$\kappa = \xi(1-v_1^2)^{-1/4}(1-v_3^2)^{-1/4}\exp\{(i/4\pi)G_-(v_1^2) - G_-(v_3^{-2}) + G_+(v_2^{-2}) + G_+(v_4^2)\}, \tag{1}$$

where

$$G_\pm(x) = \int_0^x \left[\frac{\ln(1\pm y)}{y} - \frac{\ln y}{y\pm 1}\right] dy + \text{const.}$$

The parameters, defined in Ref. 1, are expressible in terms of the angles θ_1, θ_2, and θ_3 of a spherical triangle on the unit sphere and the corresponding sides a_1, a_2, and a_3. After a short, but indirect transformation (1) can be put in the form

$$\ln\kappa = -\ln\gamma + (1/\pi)L + \frac{1}{2}\sum_{i=1}^{3}\left[\ln\cos\left(\frac{\theta_i}{2}\right) - \pi^{-1}a_i\ln\tan\left(\frac{\theta_i}{2}\right)\right], \tag{2}$$

where

$$L = \mu\int_0^{a_1}\frac{x\sin x}{1-(v-\mu\cos x)^2}dx,$$
$$\mu = \sin\theta_2\sin\theta_3, \quad v = \cos\theta_2\cos\theta_3. \tag{3}$$

The purpose of this paper is to present the details of a reduction of (2) to the symmetric form presented in Eq. (23) of Ref. 1. This result is of independent mathematical interest in that it is certainly not evident that the integral L is symmetric with respect to the angles θ_1, θ_2, and θ_3.

II. CALCULATION AND RESULTS

First, by a partial fraction decomposition and integration by parts

$$L = \frac{1}{2}a_1\ln\left(\frac{1-\mu\cos a_1/(1+v)}{1+\mu\cos a_1/(1-v)}\right) + \frac{1}{2}I,$$
$$I = \int_0^{a_1}\ln\left(\frac{1+\mu\cos x/(1-v)}{1-\mu\cos x/(1+v)}\right)dx. \tag{4}$$

Next, let

$$\lambda_1 = (1/\mu)\left[(1-v) - \sqrt{(1-v-\mu)(1-v+\mu)}\right],$$
$$\lambda_2 = (1/\mu)\left[\sqrt{(1+v-\mu)(1+v+\mu)} - (1+v)\right]$$

or

$$\lambda_1 = [1 - \cos\theta_2\cos\theta_2 - 2\sin\tfrac{1}{2}|\theta_2-\theta_3| \times\sin\tfrac{1}{2}(\theta_2+\theta_3)]/\sin\theta_2\sin\theta_3,$$
$$\lambda_2 = [2\cos\tfrac{1}{2}(\theta_2+\theta_3)\cos\tfrac{1}{2}(\theta_2-\theta_3) - 1 - \cos\theta_2\cos\theta_3]/\sin\theta_2\sin\theta_3.$$

We have the spherical triangle identities[2]

$$\tan\frac{1}{2}B\cot\frac{1}{2}A = \frac{\cos(b-s)\cos\frac{1}{2}c}{\sin\frac{1}{2}A\cos\frac{1}{2}B} - \cos c, \tag{5a}$$

$$\tan\frac{1}{2}B\tan\frac{1}{2}C = \cos c - \frac{\cos s\sin\frac{1}{2}C}{\cos\frac{1}{2}B\cos\frac{1}{2}A}, \tag{5b}$$

$$\frac{\sin(s-b)}{\sin a} = \sin\frac{1}{2}C\cos\frac{1}{2}B\sec\frac{1}{2}A, \tag{5c}$$

$$\frac{\sin s}{\sin a} = \cos\frac{1}{2}B\cos\frac{1}{2}C\csc\frac{1}{2}A, \tag{5d}$$

in the notation of Fig. 1.

With $A=\theta_2$, $B=\theta_3$, and $C=\theta_1$ in (5a); $A=\theta_3$, $B=\theta_2$, and $C=\theta_1$, in (5b); $A=\theta_1$, $B=\theta_3$, and $C=\theta_3$ in (5c); and $A=\theta_1$, $B=\theta_2$, and $C=\theta_3$ in (5d); we get the simpler forms

$$\lambda_1 = \frac{\sin(s-a_2)}{\sin(s-a_3)}, \quad \lambda_2 = -\frac{\sin(s-a_1)}{\sin s}. \tag{6}$$

By means of the identity

$$1 \pm \frac{\mu}{1\mp v}\cos x = \frac{\sin^2 x}{1+\lambda_\pm^2}\left[1 + \left(\frac{\lambda_\pm + \cos x}{\sin x}\right)^2\right]$$

with $\lambda_+ = \lambda_1$, $\lambda_- = \lambda_2$, the integral in (4) is reduced to

$$I = a_1\ln\left(\frac{1+\lambda_2^2}{1+\lambda_1^2}\right) + \int_0^{a_1}\ln\left(\frac{1+\cot^2 u_1}{1+\cot^2 u_2}\right)dx, \tag{7}$$

where

$$\tan u_j = \sin x/(\lambda_j + \cos x),$$
$$\sin u_j = \sin(x-u_j)/\lambda_j.$$

Now,

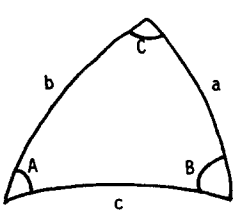


FIG. 1. Spherical triangle for Eq. (5). The perimeter is $s = a + b + c$.

 0022-2488/86/112774-02$02.50

$$\int_0^{a_1} \ln(1+\cot^2 u_j)\,dx$$
$$= 2[a_1 - \xi_j]\ln|\lambda_j| - 2\int_0^{a_1-\xi_j} \ln|\sin t\,|dt - 2\int_0^{\xi_j} \ln|\sin t\,|dt, \qquad (8)$$

where

$$\xi_j = \tan^{-1}\left(\frac{\sin a_1}{\lambda_j + \cos a_1}\right).$$

Equation (8) was not easy to derive, but is easily verified by differentiating both sides with respect to a_1. With the aid of (6) we have

$$\xi_1 = s - a_3, \quad \xi_2 = s.$$

In terms of Clausen's function

$$\phi(x) = -2\int \ln 2|\sin t\,|dt = \sum_{k=1}^{\infty} \frac{\sin 2kx}{k^2},$$

Eq. (3) becomes

$$L = \frac{1}{2}a_1 \ln\left|\frac{1-\mu\cos a_1/(1+v)}{1+\mu\cos a_1/(1-v)} \cdot \frac{\sin^2 s + \sin^2(s-a_1)}{\sin^2(s-a_2)+\sin^2(s-a_3)} \frac{\sin^2(s-a_3)}{\sin^2(s-a_1)}\right|$$
$$+ s\ln\left|\frac{\sin(s-a_1)\sin(s-a_2)}{\sin s \sin(s-a_3)}\right| + a_2 \ln\left|\frac{\sin(s-a_3)}{\sin(s-a_2)}\right| + \tfrac{1}{2}[\phi(s-a_1)+\phi(s-a_2)+\phi(s-a_3)-\phi(s)]. \qquad (9)$$

Finally, by means of the identities in (5), the first two factors in the argument of the first logarithm in (9) can be shown to cancel and the remaining terms can be manipulated into the form

$$L = \frac{1}{2}\left\{a_1 \ln\left|\frac{\sin(s-a_2)\sin(s-a_3)}{\sin s \sin(s-a_1)}\right| + a_2 \ln\left|\frac{\sin(s-a_1)\sin(s-a_3)}{\sin s \sin(s-a_2)}\right|\right.$$
$$\left. + a_3 \ln\left|\frac{\sin(s-a_1)\sin(s-a_2)}{\sin s \sin(s-a_3)}\right| + \phi(s-a_1)+\phi(s-a_2)+\phi(s-a_3)-\phi(s)\right\}. \qquad (10)$$

By combining (2) and (10) with the further identity[2]

$$\frac{\sin s \sin(s-a_1)}{\sin(s-a_2)\sin(s-a_3)} = \cot^2\frac{1}{2}\theta_1,$$

we arrive at Eq. (23) of Ref. 1. An alternative proof of this result starting from the form (1) has been found by Baxter[3] by using the properties of the Euler dilogarithm function.

ACKNOWLEDGMENTS

The author thanks Professor R. Baxter for bringing this problem to his attention and for extending the hospitality of the Australian National University.

This work was supported in part by the National Science Foundation under Grant No. DMS 8503296.

[1] R. J. Baxter, Phys. Rev. Lett. **53**, 1795 (1984).
[2] I. Todhunter and J. G. Leathem, *Spherical Trigonometry* (MacMillan, London, 1949).
[3] R. J. Baxter, Physica D **18**, 321 (1985).

Automorphisms of algebraic varieties and Yang–Baxter equations

Jean-Marie Maillard
Laboratoire de Physique Théorique et Hautes Energies,[a] *Université Pierre et Marie Curie, Tour 16-ler étage, 4, place Jussieu, 75230 Paris Cedex 05, France*

(Received 12 February 1986; accepted for publication 4 June 1986)

It is shown that commuting transfer matrix models in statistical mechanics are parametrized by algebraic varieties having a set of automorphisms deduced from the so-called "inversion relation." In general this set of automorphisms is infinite: this shows that for algebraic varieties of dimension 1, the models are parametrized by algebraic curves of genus 0 or 1.

I. INTRODUCTION

Commuting transfer matrices provide the best known criterion for exactly solvable lattice models in statistical mechanics (or models of quantum field theory). A key role is played by a special system of algebraic equations, the so-called Yang–Baxter equations (or star triangle relation or factorization equations): the underlying reason is that the (local) star-triangle relation is a sufficient (and, to some extent, necessary[1]) condition for the commutation of (global) transfer matrices.

These Yang–Baxter equations can be seen as certain homological conditions that describe the structure of the exactly solvable models. A large number of solutions of the Yang–Baxter equations have been found and recorded.[2,3] One should, however, note that all these solutions are parametrized in terms of elliptic, trigonometric, or rational functions. The few examples that gave some hope to elaborating more sophisticated structures seem to confirm a somewhat disappointing situation: the two-dimensional vertex models for which a uniformization by theta functions of genus $g>1$ was introduced do not satisfy the Yang–Baxter equations despite the fact that a Zamolodchikov algebra does exist for these models (because of the Frobenius relation on theta functions)[4,5]; on the other hand, the remarkable solution to the three-dimensional generalization of the star triangle equation, namely the "tetrahedron equation," obtained by Zamolodchikov and Baxter, turned out to be closely related with the two-dimensional free fermion Ising model (for which an elliptic parametrization occurs).[6,7] The star-triangle relation appears to be a very stringent structure (overdetermined set of equations) and this fully legitimatizes the attempts to classify exhaustively these remarkable nontrivial solutions. Along this line one should recall the beautiful papers of Belavin–Drinfeld (in which an exhaustive classification of some "classical" limit of the Yang–Baxter equations related to simple Lie algebras is displayed[8]) as well as Jimbo's success at "quantizing" this classical limit by introducing a q-analog of the universal enveloping algebra and an associated Hecke algebra.[9] But an exhaustive list of solutions is still unavailable.

We will not deal in this paper with the (infinite-dimensional) Lie algebra aspects of the problem. The aim here is rather to suggest an approach to this classification problem that concentrates on the parametrization of the Yang–Baxter equations in the framework of algebraic geometry. We shall show that the parameter space of the exactly solvable models of statistical mechanics is naturally foliated by algebraic varieties that are stable under the action of a generically infinite number of birational transformations. Our problem then reduces to classical problems of algebraic geometry (algebraic varieties possessing an infinite set of automorphisms, diophantine equations, etc.) for which numerous results are available.

In that generic case, the existence of an infinite set of automorphisms does not allow these algebraic varieties to be of the so-called "general type." In particular when these are of dimension 1 it means that the model can be parametrized by curves of genus 0 or 1 only (elliptic or rational parametrization). The study of these varieties, which are not of the general type, will lead us to make a distinction between the varieties obtained by a complete and an incomplete intersection.

The requirement that the group of automorphisms be finite very sharply constrains the model: for instance, in the case of the anisotropic q-state Potts model this imposes a restriction to the values

$$q = 2 + 2\cos 2\pi m/n \quad (m,n\in\mathbb{Z}).$$

These particular values have already been singled out by many authors (Tutte–Beraha numbers, two-dimensional models with conformal covariance, rational critical exponents, etc.[10,11]).

The results of that paper are not restricted to two-dimensional exactly solvable models. No assumption is made on the existence of a particular classical limit for the Yang–Baxter (or tetrahedron) equations.

II. THE BAXTER MODEL

Let us recall briefly some basic results concerning one of the most important exactly solvable model: the symmetric eight-vertex Baxter model.[12] It is parametrized by four homogeneous variables $(a,b,c,d)\in\mathbb{P}_3$, and the Yang–Baxter equations take the form of six trilinear homogeneous equations for three sets of points in $\mathbb{P}_3$: (a,b,c,d), (a',b',c',d'), and (a'',b'',c'',d''). This system of homogeneous equations has nontrivial solutions if

$$F_1(a,b,c,d) = F_1(a',b',c',d') = F_1(a'',b'',c'',d'') \quad (1)$$

and

$$F_2(a,b,c,d) = F_2(a',b',c',d') = F_2(a'',b'',c'',d''), \quad (2)$$

where

[a] Laboratory associated with CNRS VA 280.

 0022-2488/86/112776-06$02.50

$$F_1 = (a^2 + b^2 - c^2 - d^2)/ab$$

and

$$F_2 = cd/ab.$$

The Yang–Baxter equations imply the commutation of the $2^N \times 2^N$ row-to-row transfer matrices for arbitrary N (N is the number of vertices in a row), that is,

$$[T_N(a,b,c,d), T_N(a',b',c',d)] = 0,$$

when Eqs. (1) and (2) are satisfied.

The integrability of the model leads thus to the following foliation of the parameter space:

$$\begin{aligned} F_1(a,b,c,d) &= K_1 = \text{const}, \\ F_2(a,b,c,d) &= K_2 = \text{const}. \end{aligned} \tag{3}$$

One recognizes the well-known projective representation of an elliptic curve as an intersection of quadrics in $\mathbf{P}_3$ (Clebsch's biquadratic). One can introduce the following elliptic parametrization:

$$\begin{aligned} a &= \rho\cdot\mathrm{sn}(v+\eta,k), \\ b &= \rho\cdot\mathrm{sn}(\eta - v,k), \\ c &= \rho\cdot\mathrm{sn}(2\eta,k), \\ d &= \rho k\cdot\mathrm{sn}(2\eta,k)\mathrm{sn}(\eta - v,k)\mathrm{sn}(v+\eta,k), \end{aligned} \tag{4}$$

with

$$\begin{aligned} K_1 &= 2\,\mathrm{cn}(2\eta,k)\cdot\mathrm{dn}(2\eta,k), \\ K_2 &= k\,\mathrm{sn}^2(2\eta,k), \end{aligned}$$

where sn, cn, and dn are the Jacobian elliptic functions of modulus k. With that elliptic parametrization the Yang–Baxter equations simply read

$$v + v' + v'' = \eta. \tag{5}$$

In this particular case we have an obvious connection between the Yang–Baxter structure and the Abelian character of the algebraic curve. There also exist exact symmetries on the model, the so-called inversion relations,[13] which correspond to rational transformations on the parameters of the model. These transformations are involutions and will be denoted by I and J:

$$I:\ a \to a/(a^2 - d^2),\quad b \to b/(b^2 - c^2),$$
$$c \to -c/(b^2 - c^2),\quad d \to -d/(a^2 - d^2), \tag{6a}$$

$$J:\ a \to a/(a^2 - c^2),\quad b \to b/(b^2 - d^2),$$
$$c \to -c/(a^2 - c^2),\quad d \to -d(b^2 - d^2). \tag{6b}$$

Here F_1 and F_2 are invariant under I and J. With the elliptic parametrization I and J reduce to

$$I:\ v \to +2\eta - v,\quad J:\ v \to -2\eta - v.$$

They are conjugate via the "crossing" symmetry on the model

$$a \leftrightarrow b,\quad v \to -v.$$

These involutions generate an infinite discrete group G of symmetries of the model isomorphic to the semidirect product

$$\mathbf{Z}_2 \oplus \mathbf{Z}\quad (v \to \pm v \pm 2n\eta,\quad n \in \mathbf{Z}).$$

This infinite set of birational transformations preserve the elliptic curve (3) and the modulus of the elliptic functions. One should not confuse these transformations with the isogenies of the elliptic curve (Landen, Jacobi, Legendre transformations). One of these isogenies, the Landen transformation $k \to k_L = 2\sqrt{k}/(1+k)$ can be identified with a generator of the renormalization group for that model (a fixed point of that transformation is $k = 1$, the critical point of the model): the group G and the renormalization group act in an "orthogonal" way.

Finally the Baxter model trivializes on the so-called disorder varieties of the parameter space, on which the partition function reduces to that of an isolated vertex. For this model these varieties have a very simple expression; one of these varieties, for instance, reads

$$a + d = b + c. \tag{7}$$

The partition function per size Z is then very simple:

$$Z = a + d. \tag{8}$$

Of course these disorder varieties correspond to a trivialization of the parametrization: equation (7) corresponds to a relation between F_1 and F_2 and a value of the modulus of the elliptic functions for which this parametrization trivializes

$$F_1 = 2 - 2F_2 \Rightarrow k = -1 \quad (\text{or } k_L = \infty).$$

III. INTRODUCTION TO THE GENERAL SITUATION

For the sake of simplicity we restrict ourselves to the q-state IRF model[3] but the ideas we develop here also apply straightforwardly to two-dimensional vertex models, three- (or higher-) dimensional models. In order to fix the notations let us first recall the definition of the q-state IRF model. The spin variable associated to each site i of a square lattice are assumed to take q values: $W(\sigma_i, \sigma_j, \sigma_k, \sigma_l)$ is the Boltzmann weight associated to each of the q^4 spin configurations around a face with sites i,j,k,l (see Fig. 1). The model depends therefore on q^4 homogeneous parameters $(x_1,\ldots,x_i,\ldots,x_{q^4})$. The partition function per site Z is defined by

$$Z^N = \sum_{\{\sigma_i\}} \prod_{\square} W(\sigma_i,\sigma_j,\sigma_k,\sigma_l) \quad (\sigma_i \in \mathbf{Z}_q), \tag{9}$$

where the product is taken over all the elementary square of the lattice and N is the number of these squares.

More accurately the partition function (or even the transfer matrices) are invariant under some "gauge" transformations

$$W(\sigma_i,\sigma_j,\sigma_k,\sigma_l) \to W(\sigma_i,\sigma_j,\sigma_k,\sigma_l)\,\frac{D(\sigma_i,\sigma_l)}{D(\sigma_j,\sigma_k)}\,\frac{\Delta(\sigma_i,\sigma_j)}{\Delta(\sigma_l,\sigma_k)}. \tag{10}$$

The analysis made in this paper forgets these trivial transformations. There exist two inversion relations I and J. They act on the Boltzmann weight to give W_I and W_J defined by (see Fig. 2)

$$\sum_{\sigma_k} W(\sigma_i,\sigma_j,\sigma_k,\sigma_l)\cdot W_I(\sigma_l,\sigma_k,\sigma_j,\sigma_m) = \lambda\cdot\delta\sigma_i,\sigma_m, \tag{11a}$$

$$\sum_{\sigma_k} W(\sigma_i,\sigma_j,\sigma_k,\sigma_l)\cdot W_J(\sigma_j,\sigma_m,\sigma_l,\sigma_k) = \lambda\cdot\delta\sigma_i,\sigma_m. \tag{11b}$$

These transformations amount (*up to a rotation of the elementary square*) to looking at W, in two different ways, as q^2

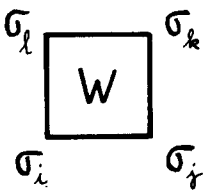


FIG. 1. The Boltzmann weight $W(\sigma_i, \sigma_j, \sigma_k, \sigma_l)$ associated to each of the q^4 spin configurations $(\sigma_i, \sigma_j, \sigma_k, \sigma_l)$ around a face with sites i,j,k,l.

$q \times q$ matrices, and taking the inverse of these q^2 matrices as

$$(W(\sigma_i,\sigma_j,\sigma_k,\sigma_l) \to W_{\sigma_i,\sigma_k}(\sigma_j,\sigma_l) \quad \text{or} \quad W_{\sigma_j,\sigma_l}(\sigma_i,\sigma_k)).$$

Because of this composition by a rotation these transformations are not involutions as the one previously introduced for the Baxter model; they are generally of infinite order.

These transformations I and J are both birational transformations

$$x_i \to \frac{P_i(x_1,\ldots,x_{q^4})}{Q_i(x_1,\ldots,x_{q^4})} = I(x_i) \text{ or } J(x_i), \tag{12}$$

where P_i and Q_i are two homogeneous polynomials of degree $q-1$ and q in the x_i's, respectively, with integer coefficients ($+1$ or -1).

This model may seem to be too general, depending on a too large number of parameters. The usual practice corresponds to imposing different symmetries or constraints on the model in order to restrict the number of homogeneous parameters of the model (equalities between different x_i's, exclusion of some configurations $x_j = 0$, etc) from q^4 to n. In the following we will restrict the parameter space to such a homogeneous space $\mathbb{P}_{n-1}$ with the condition that the (rational) transformations I and J leave that subspace invariant. Heuristic arguments based on the transfer matrix formalism enable us to show the partition function per site presents some automorphy properties with respect to these two transformations (11a) and (11b) and of course the group G generated by these two transformations[14]:

$$Z(x_1,\ldots,x_n)\cdot Z(I(x_1),\ldots,I(x_n)) = \lambda, \tag{13a}$$

$$Z(x_1,\ldots,x_n)\cdot Z(J(x_1),\ldots,J(x_n)) = \lambda'. \tag{13b}$$

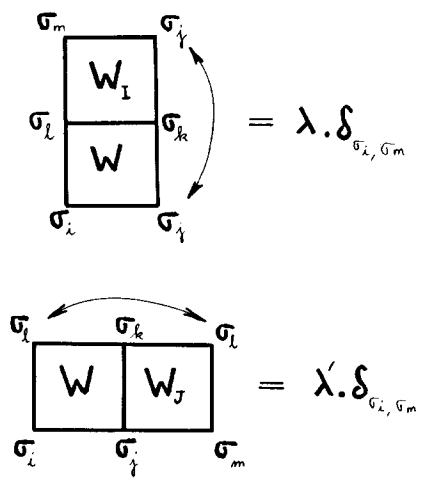


FIG. 2. Pictorial representation of the definition of the two inverse Boltzmann weights W_I and W_J.

The group G is, in general, an infinite discrete group. We now suppose that the model is exactly solvable in the sense that the Yang–Baxter equations are satisfied for the model. This leads to the commutation of the row-to-row (*and also column-to-column*) transfer matrices for arbitrary size N ($[T_N(W), T_N(W')] = 0$). The commutation of transfer matrices of specific sizes N leads to a set of algebraic equations[1,14,15] (see Appendix A for a simpler demonstration than in Ref. 1):

$$F_{\alpha,N}(x_1,\ldots,x_n) = F_{\alpha,N}(x'_1,\ldots,x'_n), \tag{14}$$

where

$$F_{\alpha,N}(x_1,\ldots,x_n) = \frac{U_{\alpha,N}(x_1,\ldots,x_n)}{V_{\alpha,N}(x_1,\ldots,x_n)},$$

where $U_{\alpha,N}$ and $V_{\alpha,N}$ are homogeneous polynomials (of degree $d_{\alpha,N}$) with integer coefficients. It can be shown that the algebraic varieties defined by the intersection of the expressions $F_{\alpha,N}$ corresponding to the row-to-row and column-to-column transfer matrices are invariant under the transformations I and J (see Refs. 14 and 16):

$$F_{\alpha,N}(x_1,\ldots,x_n) = F_{\alpha,N}(I(x_1),\ldots,I(x_n)) = \cdots. \tag{15}$$

This is a consequence of the fact that if a Yang–Baxter equation exists for the Boltzmann weight (W,W',W'') there necessarily exists another one involving W'_I and W and in fact an infinite set of other triplets of Boltzmann weights corresponding to some transformations of the initial triplet (W,W',W'') under the action of the group G (see Ref. 14). In the previous example of the Baxter model this corresponds to saying that Eq. (5) is also satisfied if one replaces (v,v',v'') by $(2n_1\eta + v, 2n_2\eta + v', 2n_3\eta + v'')$ with $n_1 + n_2 + n_3 = 0$.

An integrable model must therefore present the two following remarkable features.

(i) The infinite set of equations (14) corresponding to the various values of N must be redundant and equivalent to a finite set of m equations ($m \leqslant n-2$) we will denote from now on by F_α ($\alpha = 1,\ldots,m$) (if this is not the case we are reduced to the trivial commutation of a matrix with itself).

(ii) The algebraic variety $\mathscr{V}$ defined by the intersection of these m equations (of dimension $n-1-m$) has to be invariant under the infinite discrete group G of birational transformations in $\mathbb{P}_{n-1}$.

Therefore one sees that the exactly solvable models are naturally parametrized in terms of algebraic varieties that have (in the general case) an infinite group of automorphisms.

IV. RESULTS

A. Algebraic curves

In almost all the examples of exactly solvable models known in statistical mechanics the algebraic varieties $\mathscr{V}$ turn out to be of dimension 1 (i.e., an algebraic curve). The following result is well-known: the only algebraic curves with an infinite group of automorphisms are of genus 0 or 1 (see Ref. 17).

In other words, if the group G does not degenerate into a finite group G, one has to deal with a rational or elliptic pa-

rametrization. This result can be understood in the following heuristic way: the main distinction between the curves of genus 0 or 1 and curves of the general type of genus $g \geqslant 2$ (for which one would have to envisage a uniformization in terms of automorphic functions) lies in that there exists a finite number of particular points, called the Weierstrass points, for the curves of general type. (A point is called an ordinary point if the gap values are $1,2,\ldots,g$; otherwise it is called a Weierstrass point.) The group G that leaves invariant the algebraic curve must leave invariant these points. One understands that it is difficult for an infinite discrete group to leave invariant such a finite set of points. An old demonstration of Hettner (and also Noether) is based on these ideas. It is amusing to notice that if we consider a rational point in $\mathbb{P}_{n-1}$ ($x_i \in \mathbb{Q} \Rightarrow F_\alpha \in \mathbb{Q}$), the images of that point by the infinite group G are also rational points. We are thus led to an algebraic curve with a (generically) infinite set of rational points: Falting's theorem confirms that the curve has to be of genus 0 or 1 (see Ref. 18).

Now that we have a precise characterization of the curves that can possibly arise in the context of exactly solvable models it is useful to study the projective representation of an elliptic curve (in $\mathbb{P}_n$); the results are the following: the only case when a curve of genus 1 is given by a complete intersection are the plane cubic in $\mathbb{P}_2$ and the previous Clebsch's biquadratic in $\mathbb{P}_3$; the other representations are in $\mathbb{P}_n$ ($n \geqslant 4$) and correspond to incomplete intersections. The case of incomplete intersection may, at first sight, seem rather academic as far as statistical mechanics is concerned. However, there does exist at least one interesting example of model corresponding to that situation: for the hard hexagon model[19] the elliptic curve that parametrizes the model is given by an incomplete intersection of a quadric $F_1 = \text{const} = C_1$, a cubic $F_2 = \text{const} = C_2$, and a quartic $F_3 = \text{const} = C_3$ in $\mathbb{P}_4$,

$$F_1 = \frac{x_1^2 - x_4 x_5}{x_2 x_3}, \quad F_2 = \frac{x_4 x_3^2 + x_5 x_2^2 - x_1 x_4 x_5}{x_1 x_2 x_3}, \qquad F_3 = \frac{x_1 x_2^2 x_5 + x_1 x_3^2 x_4 - x_4^2 x_5^2 - x_2^2 x_3^2}{x_2 x_3 x_4 x_5}. \tag{16}$$

On these expressions one verifies immediately that the intersection is incomplete (as it should) because it contains the spurious varieties $x_1 = x_2 = x_4 = 0$ and $x_1 = x_3 = x_5 = 0$. The genus of the algebraic curve defined by this intersection can be calculated from the formula of addition of the characteristic of Euler–Poincaré:

$$\begin{aligned} 1 - g = {} & \chi(O_{\mathbb{P}_4}) - \chi(O(-2)) \\ & - \chi(O(-3)) - \chi(O(-4)) + \chi(O(-5)) \\ & + \chi(O(-6)) + \chi(O(-7)) - \chi(O(-9)), \end{aligned} \tag{17}$$

with

$$\chi(O(n)) = [(n+1)(n+2)(n+3)(n+4)]/4!$$

leading to a rather high genus if there were no singularities. The $g = 1$ case of the hard hexagon model corresponds to two relations between the previous constants C_i that raise the number of singularities to a maximum and thus reduce the genus to a minimum ($g = 1$):

$$C_1 \cdot C_2 = 1 \text{ and } C_1 + C_2 = C_3.$$

B. Algebraic surfaces

The problem of the classification of algebraic surfaces is much more complicated.[20] There exist invariants playing a role similar to the genus for curves (Kodeira's dimension, etc.) One can sketch the classification that way: first come surfaces of "general type," which have only a finite number of automorphisms. This case is excluded when G is infinite.

The surfaces that are not of the general type fall into five different classes (up to a birational correspondence): (a) the rational surfaces birationally isomorphic to $\mathbb{P}_2$; (b) the ruled surfaces ($\Gamma \times \mathbb{P}_1$) (these are surfaces that can be mapped onto a curve in such a way that all fibers of this mapping are isomorphic to $\mathbb{P}_1$); (c) the elliptic surfaces (fibrations by elliptic curves); (d) Abelian surfaces; and (e) K 3 surfaces. The K 3 surfaces have the property in common with Abelian surfaces that their canonical class is 0. However, in contrast with Abelian surfaces there are no regular one-dimensional forms on them.

These five sets of surfaces can all admit an infinite set of automorphisms.

Let us now assume that the algebraic variety $\mathscr{V}$ is given by a complete intersection (this corresponds *a priori* to the simplest situation in statistical mechanics).

A classical theorem (see Ref. 21, pp. 401 and 402) shows that complete intersection of dimension 2 has a trivial homotopy group ($\pi_1 = 0$). Thus the assumption of complete intersection excludes the Abelian surfaces and imposes that the variety $\mathscr{V}$ has singularities. To be more specific, this situation of complete intersection occurs for a cubic or a quartic in $\mathbb{P}_3$, for the intersection of two quadrics in $\mathbb{P}_4$ corresponding to a rational surface, and for the intersection of a quadric and a cubic in $\mathbb{P}_4$ or the intersection of three quadrics in $\mathbb{P}_5$ that correspond to a surface of type K 3.

In the case of a surface of type K 3 any explicit parametrization of the surface is, of course, hopeless.

C. Algebraic varieties of dimension > 2

Little information is available concerning the classification of these varieties. However, remarkable progress has been made during the past few years.[21] It is possible to define some invariants that unfortunately play only partially the role of the genus for algebraic curves (Betti numbers, etc.). Despite this complexity it is possible to single out varieties of a "general type" for which the number of automorphisms is finite.

The varieties that are not of a general type constitute a jungle, which is, however, fairly well understood in the simplest case of complete intersection.

Thus the situation seems rather unsatisfactory: one would like to be able to find other algebraic varieties invariant under the action of the group G that would make it possible (by taking the intersection with the algebraic varieties $\mathscr{V}$) to restrict the problem to an algebraic variety of lower dimension (eventually of dimension 1, leading to a foliation of the algebraic variety by curves of genus 0 or 1).

Fortunately such varieties can be obtained taking into account the fact that the inversion relations correspond (up to rotations of the elementary square) to taking the inverse of a set of matrices (see Appendix B). Of course, this approach applies only for algebraic varieties of dimension $\geqslant 3$. In the case of the Baxter model one can, for instance, exhibit in this way algebraic varieties defined by an intersection actually invariant under the group G:

$$abcd/[(a^2-c^2)(b^2-d^2)] = \text{const}, \tag{18a}$$

$$abcd/[(a^2-d^2)(b^2-c^2)] = \text{const}. \tag{18b}$$

However, the curve given by the intersection of these two quartics has, in general, no intersection with the elliptic curve (3).

V. G IS A FINITE GROUP

The previous analysis is based on the infinite character of the group G. When the group G is finite this leads to algebraic constraints on the parameter space that characterize the model very precisely. For every element g of G there exists an integer p such that g^P is equal to the unit element of G. This equality translated on the homogeneous parameters x_i means that the model is restricted to some very particular algebraic varieties.

Let us now recall the hexagon model, which can be seen as a subcase of the S.O.S. eight-vertex Baxter model[22]: despite the fact that this model has a finite group G, it presents (as we have mentioned already) an elliptic uniformization, which can be seen as a restriction of the elliptic uniformization of the Baxter model.[23] Nevertheless, it is true that it is difficult to specify the algebraic varieties corresponding to a model, with a finite group G, that is not obviously embedded into a larger model with an already known uniformization. It is, however, possible, in the case of algebraic curves of genus g, to give an upper bound of the order of the finite group G (see Ref. 17): $[G] \leqslant 84g - 3$.

VI. DISORDER VARIETIES

We have already remarked that the Baxter model trivializes on a simple disorder variety (7). In fact such disorder varieties are quite easy to find[24] and their corresponding codimension is rather low. For instance, in the case of the 16-vertex model, there exist disorder varieties of codimension 1 in the parameter space. This should be compared with the codimension of the parameter space of the exactly solvable subcase of that model, the Fan and Wu free-fermion model[7] and the Baxter model of codimensions 4 and 5, respectively.

For instance let us consider a subcase of the 16-vertex model that has the two previous integrable models as subcases (but is not integrable in general): the asymmetric eight-vertex model. The homogeneous parameters of that model are usually denoted a,a',b,b',c,c',d, and d' (the symmetric eight-vertex model corresponds to $a=a'$, $b=b'$, $c=c'$, and $d=d'$). That model has a disorder solution on the (disorder) variety given by the quartic equation (this result has also been obtained recently by Giacomini[25])

$$(a+a') + ((a+a')^2 - 4(aa'-dd'))^{1/2} = (b+b') + ((b+b'')^2 - 4(bb'-cc'))^{1/2}. \tag{19}$$

If the model were integrable, there should occur a trivialization of the parametrization on this disorder variety and also on the images of this variety by the infinite group G generated by the two inversion relations. It is a simple and instructive exercise to verify that there does, in fact, exist an infinite number of such images except in the two already-mentioned cases of the Baxter model and the free-fermion model ($aa'+bb'=cc'+dd'$), where the number of images of (19) under the action of the infinite group G is finite. The checkerboard Potts is another example of an infinite number of images of a disorder variety under the action of G (see Ref. 26); moreover one has a remarkable and instructive agreement between the exact expressions of the (analytical continuation of the) partition function on this infinite set of algebraic varieties and the exact expression of the partition function on the critical variety where the model is exactly solvable.[27]

The existence of such an infinite set of varieties at first seems hardly compatible with the exact solvability of the model. An obvious situation where this set is finite is when the group G is itself finite. Let us consider the checkerboard Ising model: this model has an elliptic uniformization and the modulus of the elliptic functions that occur is given (in terms of the high-temperature variables $t_i = th\, K_i$ and the dual variable

$$t_i^* = \frac{1-t_i}{1+t_i}, \quad \text{by Eq. (20):}$$

$$k = \prod_{i=1}^{4} \frac{t_i(t_i^* + t_j^* t_k^* t_l^*)(1-t_i^2)}{t_i^*(t_i + t_j t_k t_l)(1-t_i^{*2})}, \tag{20}$$

$$(i,j,k,l) = (1,2,3,4).$$

This algebraic expression trivializes on the disorder varieties of the model, on the dual of these disorder varieties (and of course when the coupling constant of the model trivializes $t_i = 0$, $t_i = \pm 1$, $t_i^* = 1$). From this example it is rather tempting (if one is willing to bet on the exact solvability of the noncritical three-state Potts model) to guess an algebraic expression k associated to that model from the known equations of the disorder varieties and their images under the group G (see Refs. 28 and 29).

VII. CONCLUSION, PROSPECTS

The exactly solvable models are parametrized by means of algebraic varieties having a group of automorphisms deduced from the so-called "inversion relations." It is very constraining for a model of statistical mechanics to ask for this group to be finite. It is, in general, infinite and this shows that these algebraic varieties are not of the "general type" (but this does not prove that they should be Abelian varieties). For algebraic varieties of dimension 1, this sheds a new light on the occurence of curves of genus 0 or 1 only for all the exact models known at the present moment. Of course this is just a preliminary work and these ideas will be pursued in forthcoming publications. The ideas we have developed here also apply, *mutatis mutandis*, to statistical models in d dimensions with the difference that the number of inversion relations that generate the group G grows with the dimension d. *A priori* there is no relation between these generators. Therefore the group G is in general a very "large" one (infi-

nite discrete of course): is it possible for algebraic varieties to have such a large group of automorphisms?

ACKNOWLEDGMENTS

The author would like to thank A. Beauville, D. Bertrand, and J. L. Verdier for fruitful discussions and for careful reading of the manuscript. O. Babelon, P. Lochak, A. Sznitman, and L. Szpiro are also acknowledged for providing very helpful comments.

APPENDIX A: ALGEBRAIC CONDITIONS FOR COMMUTATION

The commutation of two $n \times n$ matrices T and T' leads to the existence of algebraic expressions in the coefficients of these two matrices $F_\alpha(T) = F_\alpha(T')$. They can be seen as some symmetric functions of the eigenvectors shared by T and T'. We sketch here a simple way to get these F_α's: Let us denote by C and C' two matrices that are linear combinations of powers of T and T',

$$C = \sum_{p=0}^{n-1} \alpha_p \cdot T^p, \quad C' = \sum_{p=0}^{n-1} \alpha'_p \cdot T'^p.$$

We have

$$[T,T'] = 0 \Rightarrow [C,C'] = 0. \tag{A1}$$

Let us denote by C_{ij}, C'_{ij}, T_{ij}, and T'_{ij} the coefficients of these matrices. We can choose α_p and α'_p some algebraic expressions of the T_{ij} and T'_{ij} such that

$$C_{1j} = 0, \quad j = 1,\ldots,n-1, \quad C_{1n} \neq 0,$$

$$C'_{1j} = 0, \quad j = 1,\ldots,n-1, \quad C'_{1n} \neq 0.$$

Equation (20) then leads to

$$\forall i: \quad C'_{1n}/C'_{ni} = C_{1n}/C_{ni}. \tag{A2}$$

Similar algebraic expressions can be obtained imposing C such that

$$C_{ij} = 0, \quad j = 1,\ldots,n-1, \quad C_{in} \neq 0.$$

APPENDIX B: G-INVARIANT VARIETIES

The characteristic polynomial $P_M(\lambda)$ of an $n \times n$ matrix M and of its inverse matrix M^{-1} are related:

$$\lambda^n P_{M^{-1}}(1/\lambda) = P_M(\lambda). \tag{B1}$$

We denote by c_i the coefficients of $P_M(\lambda)$ and obtain

$$P_M(\lambda) = \lambda^n + c_1 \cdot \lambda^{n-1} + \cdots + c_i \lambda^{n-i} + \cdots + c_n.$$

An immediate consequence of (B1) is that the expressions $\phi_i = c_i c_{n-i}/c_n$ are invariant under the transformation $M \to M^{-1}$. These expressions are the ratio of two homogeneous polynomials of degree n in the coefficients of the matrix M. The inversion relations I and J correspond (up to a permutation of the homogeneous parameters of the model x_i) to taking the inverse of a set of q^2 matrices M_α. One can associate to each of these matrices the corresponding expressions ϕ_i^α.

Let us consider ϕ_i the product of the ϕ_i^α and algebraic expressions A_j invariant under the previous permutation of the x_i; the algebraic variety defined by the intersection of equations

$$\phi_i(x_1,\ldots,x_n) = \text{const}, \quad A_j(x_1,\ldots,x_n) = \text{const}$$

is invariant under the inversion relations I and J.

[1] P. Lochak and J. M. Maillard, J. Math. Phys. **27**, 593 (1986).
[2] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic, New York, 1982); *Proceedings of the 1980 Enschede Summer School, Fundamental Problems in Statistical Mechanics*, Vol. V (North-Holland, Amsterdam, 1980).
[3] M. Jimbo and T. Miwa, "Classification of solutions to the star triangle relation for a class of 3- and 4-state IRF models," NSF-ITP-84-154, 1984.
[4] I. V. Cherednik, Theor. Math. Phys. **43**, 117, 356 (1980).
[5] H. C. Öttinger and J. Honerkamp, Phys. Lett. A **88**, 339 (1982); R. E. Shrock, *ibid.* **91**, 157 (1982).
[6] A. B. Zamolodchikov, Commun. Math. Phys. **79**, 489 (1981); R. J. Baxter, *ibid.* **88**, 185 (1983); Physica D **18**, 321 (1986).
[7] C. Fan and F. Y. Wu, Phys. Rev. B **2**, 723 (1970).
[8] A. A. Belavin and V. G. Drinfeld D., Funct. Anal. Appl. **17**, 220 (1983).
[9] M. Jimbo, "A q-analogue of U (gl($N+1$)), Hecke algebra and the Yang–Baxter equation," RIMS preprint $n°$517, 1985.
[10] S. Beraha, J. Kahane, and N. J. Weiss, J. Comb. Theory B **28**, 52 (1980).
[11] A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov, J. Stat. Phys. **36**, 775 (1984).
[12] R. J. Baxter, Ann. Phys. (NY) **70**, 193 (1972); **71**, 1 (1973).
[13] Y. G. Stroganov, Phys. Lett. A **74**, 116 (1979).
[14] J. M. Maillard and T. Garel, J. Phys. A. **17**, 1251 (1983).
[15] M. T. Jaekel and J. M. Maillard, J. Phys. A **16**, 3105 (1983).
[16] J. M. Maillard, "The star-triangle relation and the inversion relation in statistical mechanics," in *Proceedings of the Brasov International Summer School 1983* (Birkhauser, Bäsel, 1985).
[17] R. Hartshorne, *Algebraic Geometry* (Springer, New York, 1977); J. Griffiths and J. Harri, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[18] L. Szpiro, Séminaire Bourbaki $n°$619, 1983/84.
[19] R. J. Baxter, Phys. A **13**, L61 (1980).
[20] A. Beauville, "Surface algébriques complexes," Astérisque 54, Société Mathématique de France, 1984; Séminaire Bourbaki $n°$609, 1982/83.
[21] I. R. Shafarevich, *Basic Algebraic Geometry* (Springer, New York, 1977).
[22] H. Esnault, Séminaire Bourbaki, Lecture Notes 901, 1981.
[23] G. E. Andrews, R. J. Baxter, and P. J. Forrester, J. Stat. Phys. **35**, 193 (1984).
[24] M. T. Jaekel and J. M. Maillard, J. Phys. A **18**, 1229 (1985).
[25] H. J. Giacomini, submitted to J. Phys. A.
[26] M. T. Jaekel and J. M. Maillard, J. Phys. A **17**, 2079 (1984).
[27] C. Domb and M. S. Green, *Phase Transitions and Critical Phenomena* (Academic, New York, 1972), Vol. 1, pp. 285 and 323.
[28] D. Hansel, A. Georges, and P. Le Doussal (private communication).
[29] R. J. Baxter, "Free fermion checkerboard and Z-invariant lattice models in statistical mechanics," to appear in Proc. R. Soc./Philos. Trans. R. Soc.

The fractional diffusion equation

Walter Wyss
Department of Physics, University of Colorado, Boulder, Colorado 80309

(Received 25 November 1985; accepted for publication 16 July 1986)

In one space—and in one time—dimension a diffusion equation is solved, where the first time derivative is replaced by the λ-fractional time derivative, $0<\lambda\leqslant 1$. The solution is given in closed form in terms of Fox functions.

I. INTRODUCTION

The concept of fractional derivatives, expressed as a convolution with ϕ_λ (see Ref. 1) and the theory of Fox functions[2–5] are used to solve the following two fractional diffusion equations.

Problem I:

$$a^2\frac{\partial^2}{\partial x^2}T(x,t)=\phi_{-\lambda}(t)*T(x,t),\quad x\in R_+,$$
$$a>0,\quad 0<\lambda\leqslant 1, \tag{1.1}$$
$$T(x,0)=\theta(x),\quad T(0,t)=0,\quad t>0.$$

Problem II:

$$a^2\frac{\partial^2}{\partial x^2}T(x,t)=\phi_{-\lambda}(t)*T(x,t),\quad x\in R_+,$$
$$a>0,\quad 0<\lambda\leqslant 1, \tag{1.2}$$
$$T(x,0)=\theta(x),\quad T(0,t)=-1,\quad t>0.$$

Observe that problem I and problem II only differ by a translation in temperature. However, by treating the two problems separately, we will see the different information contained in the corresponding Fox function solutions. This reflects relations amongst Fox functions.

For $\lambda=1$, the problems reduce to classical diffusion problems. In this case we have the solutions.

Problem I ($\lambda=1$):

$$T(x,t)=\mathrm{Erf}[(1/2a)t^{-1/2}x),\quad x\in R_+. \tag{1.3}$$

Problem II ($\lambda=1$):

$$T(x,t)=\mathrm{Erf}[(1/2a)t^{-1/2}x)-1,\quad x\in R_+. \tag{1.4}$$

The long time behavior in problem I for $\lambda=1$ is given by

$$T(x,t)\approx(1/a)\pi^{-1/2}(xt^{-1/2}). \tag{1.5}$$

For $0<\lambda<1$, we have a nonstandard diffusion, possibly being due to impurities. We solve problem I and problem II exactly and discuss the asymptotic behavior.

II. THE FRACTIONAL DIFFUSION EQUATION

The fractional diffusion equation for both of our problems is the following integrodifferential equation

$$a^2\frac{\partial^2}{\partial x^2}T(x,t)=\int_0^t T(x,\tau)\frac{(t-\tau)^{-\lambda-1}}{\Gamma(-\lambda)}d\tau,\quad x\in R_+, \tag{2.1}$$

where $0<\lambda\leqslant 1$. Here we take into account[1] that

$$\phi_\lambda(x)\equiv x_+^{\lambda-1}/\Gamma(\lambda) \tag{2.2}$$

are distributions in $(D')_+$ and entire in λ.

Both of our problems are invariant under the following scaling:

$$\sigma>0,\quad x\to\sigma x\quad\text{and}\quad t\to\sigma^{2/\lambda}t, \tag{2.3}$$

and thus

$$T(x,t)=F(t^{-\lambda/2}x). \tag{2.4}$$

This is shown by using the substitutions

$$y=t^{-\lambda/2}x,\quad z=\tau^{-\lambda/2}x \tag{2.5}$$

in the integrodifferential equation (2.1). We get now an equation for $F(y)$:

$$a^2\frac{d^2F(y)}{dy^2}=\frac{2}{\lambda\Gamma(-\lambda)}\int_y^\infty dz\,y^{2/\lambda}z^{-1-2/\lambda}F(z)\times\left[1-\left(\frac{y}{z}\right)^{2/\lambda}\right]^{-1-\lambda}. \tag{2.6}$$

Introduce the distribution

$$g(\omega)=\begin{cases}\dfrac{2}{\lambda\Gamma(-\lambda)}(1-\omega^{2/\lambda})^{-\lambda-1}, & 0<\omega<1,\\ 0, & \omega>1,\end{cases} \tag{2.7}$$

then Eq. (2.6) reads

$$a^2\frac{d^2}{dy^2}F(y)=y^{2/\lambda}\int_0^\infty dz\,z^{-1-2/\lambda}F(z)g\left(\frac{y}{z}\right). \tag{2.8}$$

The boundary conditions are as follows in the problems below.

Problem I:

$$F(0)=0,\quad F(\infty)=1. \tag{2.9}$$

Problem II:

$$F(0)=-1,\quad F(\infty)=0. \tag{2.10}$$

III. SOLUTIONS OF THE PROBLEMS

We first compute the Mellin transform[6] of Eq. (2.8):

$$\hat F(s)\equiv\int_0^\infty F(y)\,y^{s-1}\,dy, \tag{3.1}$$

$$a^2(s-1)(s-2)\hat F(s-2)=\hat F(s)\hat g[s+(2/\lambda)], \tag{3.2}$$

with

$$\hat g(s)=\frac{\Gamma[(\lambda/2)s]}{\Gamma[-\lambda+(\lambda/2)s]}. \tag{3.3}$$

We then get the difference equation

$$a^2(s-1)(s-2)\hat F(s-2)=\frac{\Gamma[1+(\lambda/2)s]}{\Gamma[1-\lambda+(\lambda/2)s]}\hat F(s). \tag{3.4}$$

 0022-2488/86/112782-04$02.50

Taking into account the path and pole structure, explained in the Appendix, the solutions, compatible with the boundary conditions, are as follows.

Problem I:

$$\hat{F}(s) = \pi^{-1/2}(2a)^s \frac{\Gamma(-s)\Gamma(\frac{1}{2}+s/2)\Gamma(1+s/2)}{\Gamma(1-s)\Gamma(1+(\lambda/2)s)}. \tag{3.5}$$

Problem II:

$$\hat{F}(s) = -\pi^{-1/2}(2a)^s \frac{\Gamma(s)\Gamma(\frac{1}{2}+2/s)\Gamma(1+s/2)}{\Gamma(1+s)\Gamma(1+(\lambda/2)s)}. \tag{3.6}$$

We now compute the inverse Mellin transforms: First let

$$\hat{F}(s) = \pi^{-1/2}(2a)^s \hat{h}(s). \tag{3.7}$$

Then

$$F(y) = \pi^{-1/2} h((1/2a)y), \tag{3.8}$$

where $h(z)$ is given by the following problems.

Problem I:

$$h(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{\Gamma(-s)\Gamma(\frac{1}{2}+\frac{1}{2}s)\Gamma(1+\frac{1}{2}s)}{\Gamma(1-s)\Gamma(1+(\lambda/2)s)} z^{-s}\,ds. \tag{3.9}$$

Problem II:

$$h(z) = -\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{\Gamma(s)\Gamma(\frac{1}{2}+\frac{1}{2}s)\Gamma(1+\frac{1}{2}s)}{\Gamma(1+s)\Gamma(1+(\lambda/2)s)} z^{-s}\,ds. \tag{3.10}$$

Replacing s by $-s$ and using the theory of Fox functions (see the Appendix) we get for our two problems the following solutions.

Problem I:

$$h(z) = \frac{1}{2\pi i}\int_L \frac{\Gamma(s)\Gamma(\frac{1}{2}-\frac{1}{2}s)\Gamma(1-\frac{1}{2}s)}{\Gamma(1+s)\Gamma(1-(\lambda/2)s)} z^s\,ds. \tag{3.11}$$

The Fox parameters are

$$\begin{aligned} &m=2, \quad b_1=\tfrac{1}{2}, \quad \beta_1=\tfrac{1}{2},\\ &\qquad b_2=1, \quad \beta_2=\tfrac{1}{2},\\ &n=1, \quad a_1=1, \quad \alpha_1=1,\\ &q=3, \quad b_3=0, \quad \beta_3=1,\\ &p=2, \quad a_2=1, \quad \alpha_2=\lambda/2, \end{aligned}$$

and thus

$$h(z) = H_{23}^{21}\left(z\left|\begin{matrix}(1,1);(1,\lambda/2)\\(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2});(0,1)\end{matrix}\right.\right). \tag{3.12}$$

Problem II:

$$h(z) = -\frac{1}{2\pi i}\int_L \frac{\Gamma(-s)\Gamma(\frac{1}{2}-\frac{1}{2}s)\Gamma(1-\frac{1}{2}s)}{\Gamma(1-s)\Gamma(1-(\lambda/2)s)} z^s\,ds. \tag{3.13}$$

The Fox parameters are

$$\begin{aligned} &m=3, \quad b_1=0, \quad \beta_1=1,\\ &\qquad b_2=\tfrac{1}{2}, \quad \beta_2=\tfrac{1}{2},\\ &\qquad b_3=1, \quad \beta_3=\tfrac{1}{2},\\ &n=0,\\ &q=3,\\ &p=2, \quad a_1=1, \quad \alpha_1=1,\\ &\qquad a_2=1, \quad a_2=\lambda/2, \end{aligned}$$

and thus

$$h(z) = -H_{23}^{30}\left(z\left|\begin{matrix}-;(1,1),(1,\lambda/2)\\(0,1),(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2});-\end{matrix}\right.\right). \tag{3.14}$$

IV. POWER SERIES EXPANSIONS, ASYMPTOTIC EXPANSIONS

According to the theory of Fox functions we identify the following expressions.

Problem I:

$$\begin{aligned} A(s) &= \Gamma(\tfrac{1}{2}-\tfrac{1}{2}a)\Gamma(1-\tfrac{1}{2}s),\\ B(s) &= \Gamma(s),\\ C(s) &= \Gamma(1+s),\\ D(s) &= \Gamma(1-(\lambda/2)s). \end{aligned} \tag{4.1}$$

The poles of A are given by

$$P(A) = \{s=k,\ k=1,2,...\} \tag{4.2}$$

and the poles of B by

$$P(B) = \{s=-k,\ k=0,1,2,...\} \tag{4.3}$$

and thus $P(A)\wedge P(B)=\phi$.

Problem II:

$$\begin{aligned} A(s) &= \Gamma(-s)\Gamma(\tfrac{1}{2}-\tfrac{1}{2}s)\Gamma(1-\tfrac{1}{2}s),\\ B(s) &= 1,\\ C(s) &= 1,\\ D(s) &= \Gamma(1-s)\Gamma(1-(\lambda/2)s). \end{aligned} \tag{4.4}$$

The poles of A are given by

$$P(A) = \{s=k,\ k=0,1,2,...\} \tag{4.5}$$

and the poles of B by

$$P(B) = \phi, \tag{4.6}$$

and thus indeed $P(A)\wedge P(B)=\phi$.

For our two problems we now get the following power series expansions.

Problem I: First notice that $\mu = 1-\lambda/2 > 0$. Then

$$H_{23}^{21}(z) = -\sum_{s\in P(A)} \operatorname{Res}\left(\frac{A(s)B(s)}{C(s)D(s)} z^s\right) \tag{4.7}$$

or explicitly

$$\begin{aligned} h(z) = 2\sum_{k=0}^{\infty}(-1)^k \frac{1}{k!} &\times\left[\frac{1}{2k+1}\frac{\Gamma(\frac{1}{2}-k)}{\Gamma(1-(\lambda/2)-\lambda k)} z^{2k+1}\right.\\ &\left.+\frac{1}{2k+2}\frac{\Gamma(-\frac{1}{2}-k)}{\Gamma(1-\lambda-\lambda k)} z^{2k+2}\right]. \end{aligned} \tag{4.8}$$

Problem II: First notice that $\mu = 1 - \lambda/2 > 0$. Then

$$H_{23}^{30}(z) = -\sum_{s\in P(A)} \operatorname{Res}\left(\frac{A(s)B(s)}{C(s)D(s)} z^s\right) \tag{4.9}$$

or explicitly

$$h(z) = -\sqrt{\pi} + 2\sum_{k=0}^{\infty} (-1)^k \frac{1}{k!} \times\left[\frac{1}{2k+1}\frac{\Gamma(\frac{1}{2}-k)}{\Gamma(1-(\lambda/2)-\lambda k)} z^{2k+1} + \frac{1}{2k+2}\frac{\Gamma(-\frac{1}{2}-k)}{\Gamma(1-\lambda-\lambda k)} z^{2k+2}\right]. \tag{4.10}$$

We now determine the asymptotic behavior of our solutions.

Problem I: First observe that $\delta = (1-\lambda/2)\pi = \mu\pi$ and thus $\delta > (\pi/2)\mu$. Then

$$H_{23}^{21}(z) \approx \sum_{s\in P(B)} \operatorname{Res}\left(\frac{A(s)B(s)}{C(s)D(s)} z^s\right) \tag{4.11}$$

or explicitly

$$h(z) \approx \sqrt{\pi}, \tag{4.12}$$

as $|z|\to\infty$, uniformly on every closed subsector of

$$|\arg z| < (\pi/2)(1-\lambda/2).$$

Problem II: First observe that $n = 0$ and $q = m$ and thus $\delta = \mu\pi$. Therefore we have an exponentially small behavior. Then

$$H_{23}^{30}(z) \approx -2\pi E(ze^{i\pi\mu}) \tag{4.13}$$

with

$$\alpha = \tfrac{3}{2}, \quad \beta = 2(\lambda/2)^{\lambda/2}, \quad \mu = 1-\lambda/2.$$

We find

$$H_{23}^{30}(z) \approx \mu^{-3/2}\beta^{-1/2\mu}\sum_{k=0}^{\infty}(-1)^k A_k \mu^{-k} \times \beta^{-k/\mu} z^{(-1-2k)/2\mu} e^{-\mu\beta^{1/\mu}z^{1/\mu}}, \tag{4.14}$$

as $|z|\to\infty$ uniform on every closed subsector of $|\arg z| < \mu(\pi/2)$. The coefficient A_0 is given by

$$A_0 = 2^{1/2}\mu(2/\lambda)^{1/2}. \tag{4.15}$$

V. EXPLICIT SOLUTIONS AND DISCUSSION

From (2.4), (3.8), (3.12), and (3.14) we find the solution to our problems.

Problem I:

$$T(x,t) = \pi^{-1/2} H_{23}^{21}\left(\frac{1}{2a} t^{-\lambda/2} x \,\middle|\, \begin{matrix}(1,1);(1,\lambda/2)\\(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2});(0,1)\end{matrix}\right). \tag{5.1}$$

Problem II:

$$T(x,t) = -\pi^{-1/2} H_{23}^{30}\left(\frac{1}{2a} t^{-\lambda/2} x \,\middle|\, \begin{matrix}-;(1,1),(1,\lambda/2)\\(0,1),(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2});-\end{matrix}\right). \tag{5.2}$$

From (4.8) and (4.10) we see that the boundary conditions are satisfied. From (4.12) and (4.14) we see that also the initial conditions are satisfied.

For $x > 0$ we have the following long-time behavior.

Problem I:

$$T(x,t) \approx \frac{1}{\Gamma(1-\lambda/2)}\frac{1}{a} t^{-\lambda/2} x. \tag{5.3}$$

Problem II:

$$T(x,t) \approx -1 + \frac{1}{\Gamma(1-\lambda/2)}\frac{1}{a} t^{-\lambda/2} x. \tag{5.4}$$

For the special case $\lambda = 1$, we have the usual diffusion equation. From (4.8), (4.10), and (4.14) we get the following results:

Problem I ($\lambda = 1$):

$$T(x,t) = \frac{2}{\sqrt{\pi}}\sum_{k=0}^{\infty}(-1)^k \frac{1}{k!}\frac{1}{2k+1}\left(\frac{1}{2a} t^{-1/2} x\right)^{2k+1} = \operatorname{Erf}\left(\frac{1}{2a} t^{-1/2} x\right). \tag{5.5}$$

Problem II ($\lambda = 1$):

$$T(x,t) = -1 + \operatorname{Erf}[(1/2a)t^{-1/2}x] \tag{5.6}$$

and the exponentially small asymptotic behavior

$$T(x,t) \approx -\frac{1}{\sqrt{\pi}}\frac{1}{2a} t^{-1/2} x \exp\left(-\frac{x^2}{4a^2 t}\right), \tag{5.7}$$

as

$$x/\sqrt{t} \to \infty.$$

We have treated the two problems in a parallel way to exhibit the richness of Fox functions especially as related to the asymptotic behavior.

Remarks: (1) The above analysis is also valid for $1 \leqslant \lambda < 2$.

(2) The fractional diffusion equation has been derived by Nigmatullin[7] for a medium with fractal geometry (porous medium).

APPENDIX: FOX FUNCTIONS[3-5]

The Fox function

$$H_{pq}^{mn}(z) = H_{pq}^{mn}\left(z \,\middle|\, \begin{matrix}(a_1,\alpha_1),\ldots,(a_n,\alpha_n);(a_{n+1},\alpha_{n+1}),\ldots,(a_p,\alpha_p)\\(b_1,\beta_1),\ldots,(b_m,\beta_m);(b_{m+1},\beta_{m+1}),\ldots,(b_q,\beta_q)\end{matrix}\right)$$

is defined by the contour integral

$$H_{pq}^{mn}(z) = \frac{1}{2\pi i}\int_L \frac{A(s)B(s)}{C(s)D(s)} z^s\, ds,$$

with

$$A(s) = \prod_{k=1}^{m}\Gamma(b_k - \beta_k s),$$

$$B(s) = \prod_{k=1}^{n}\Gamma(1-a_k+\alpha_k s),$$

$$C(s) = \prod_{k=m+1}^{q}\Gamma(1-b_k+\beta_k s),$$

$$D(s) = \prod_{k=n+1}^{p}\Gamma(a_k - \alpha_k s).$$

Here m,n,p,q are integers satisfying

$$0 \leqslant n \leqslant p, \quad 1 \leqslant m \leqslant q.$$

For $n = 0$ we put $B(s) = 1$, for $q = m$ we put $C(s) = 1$, and for $p = n$ we put $D(s) = 1$. The Fox parameters $(a_1,...,a_p)$ and $(b_1,...,b_q)$ are complex, whereas $(\alpha_1,...,\alpha_p)$ and $(\beta_1,...,\beta_q)$ are positive numbers.

These parameters are restricted by the condition

$$P(A) \wedge P(B) = \phi,$$

where

$$P(A) = \{s = (b_j + k)/\beta_j,\ j = 1,...,m;\ k = 0,1,2,...\},$$

$$P(B) = \{s = (a_j - 1 - k)/\alpha_j,\ j = 1,...,n;\ k = 0,1,2,...\}$$

are the set of poles of $A(s)$ and $B(s)$, respectively. The integration contour L runs between $s = \infty + ic$ and $s = \infty - ic$, where

$$c > \max_{1 \leqslant j \leqslant m} \{|\mathrm{Im}\, b_j|/\beta_j\},$$

and such that $P(A)$ lies to the left of L, and $P(B)$ to the right of L. From now on we assume that

$$\mu \equiv \sum_{k=1}^{q} \beta_k - \sum_{k=1}^{p} \alpha_k > 0.$$

In Ref. 3, also the case $\mu = 0$ is treated. Under these conditions $H_{pq}^{mn}(z)$ is an analytic function for $z \neq 0$, in general multivalued (one-valued on the Riemann surface of $\ln z$). It is given by

$$H_{pq}^{mn}(z) = -\sum_{s \in P(A)} \mathrm{Res}\left(\frac{A(s)B(s)}{C(s)D(s)} z^s\right).$$

The asymptotic behavior of $H_{pq}^{mn}(z)$ for $|z| \to \infty$ is determined by analytic continuation. Here we give the results for the two cases.

First let

$$\delta \equiv \left(\sum_{j=1}^{m} \beta_j - \sum_{j=n+1}^{p} \alpha_j\right) \pi.$$

Case 1 ($n > 0$, $\delta > \mu\pi/2$): Then

$$H_{pq}^{mn}(z) \approx \sum_{s \in P(B)} \mathrm{Res}\left(\frac{A(s)B(s)}{C(s)D(s)} z^s\right)$$

as $|z| \to \infty$, uniformly on every closed subsector of

$$|\arg z| < \delta - \mu \cdot (\pi/2).$$

Case 2 ($n = 0$, $q = m$):

$$H_{pm}^{mo}(z) \approx (2\pi)^{m-p} e^{i\pi(\alpha - 1/2)} E(z e^{i\pi\mu})$$

as $|z| \to \infty$, uniformly on every closed sector (vertex at 0) contained in

$$|\arg z| < \mu\pi/2.$$

Here

$$E(z) = \frac{1}{2\pi i \mu} \sum_{k=0}^{\infty} A_k (\beta\mu^{\mu} z)^{(1-\alpha-k)/\mu} e^{(\beta\mu^{\mu} z)^{(1/\mu)}},$$

where the constants α and β are given by

$$\alpha = \sum_{k=1}^{p} a_k - \sum_{k=1}^{q} b_k + \frac{1}{2}(q - p + 1)$$

and

$$\beta_k = \prod_{k=1}^{p} \alpha_k^{\alpha_k} \prod_{k=1}^{q} \beta_k^{-\beta_k}$$

and $q = m$.

The coefficients $A_k, k = 0,1,2,...$, are determined by

$$\frac{A(s)B(s)}{C(s)D(s)} (\beta\mu^{\mu})^{-s} \approx \sum_{k=0}^{\infty} \frac{A_k}{\Gamma(\mu s + \alpha + k)}.$$

In particular

$$A_0 = (2\pi)^{(1/2)(p-q+1)} \mu^{\alpha - 1/2} \prod_{k=1}^{p} \alpha_k^{1/2 - a_k} \prod_{k=1}^{q} \beta_k^{b_k - 1/2}$$

at $q = m$.

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

[2] C. Fox, "The G and H functions as symmetrical Fourier kernels," Trans. Am. Math. Soc. **98**, 395 (1961).

[3] B. L. S. Braaksma, "Asymptotic expansions and analytic continuations for a class of Barnes-integrals," Compos. Math. **15**, 239 (1964).

[4] K. G. Gupta and U. C. Jain, "The H-function-II.," Proc. Nat. Acad. Sci. India A **36**, 594 (1966).

[5] W. R. Schneider, "Stable distributions: Fox function representation and generalization," to be published in Lect. Notes Phys.

[6] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Tables of Integral Transforms* (McGraw–Hill, New York, 1954).

[7] R. R. Nigmatullin, "The realization of the generalized transfer equation in a medium with fractal geometry," Phys. Status. Solidi B **133**, 425 (1986).

Spinor propagators in anti-de Sitter space-time

H. Janssen and C. Dullemond
Institute for Theoretical Physics, University of Nijmegen, Nijmegen, The Netherlands

(Received 30 April 1985; accepted for publication 11 July 1986)

Expressions are found for homogeneous and inhomogeneous propagators for spinor fields of arbitrary mass in anti-de Sitter space-time.

I. INTRODUCTION

One of the main problems in high energy physics today is to understand why quarks and gluons are confined. It is generally assumed that quantum chromodynamics (QCD) alone is responsible for confinement, but so far no proof has ever been given. It is therefore expedient to make models in which confinement has been built in.

In analogy to models put forward by several authors[1] a model was studied in which this was done by geometrical means.[2] Here, quarks and gluons move inside a spherical bag with anti-de Sitter (AdS) metric, and carry out harmonic oscillations with a universal frequency equal to c/R, where R is the radius of the bag. In quantized form this frequency can be related to the apparent universal level spacing of quarkonium spectra.

In another article the connection between QCD was suggested via spontaneous symmetry breaking of the conformal symmetry of the QCD Lagrangian to SO(3,2) or AdS symmetry.[3]

In order to take quantum effects into account, like gluon exchange between quarks, etc., we need the use of propagators. Early work has been done by Fronsdal[4] for homogeneous AdS scalar propagators and Fronsdal and Haugen[5] for spinor fields. The massless case for arbitrary spin has been studied by Fronsdal[6] and Fang and Fronsdal.[7] In an earlier paper, expressions for SO(3,2) symmetric massive scalar propagators, homogeneous as well as inhomogeneous, were found, using configuration space methods[8] (see also Ref. 9).

As is well known, in order to write down a meaningful Dirac equation in a curved space, one has to define an orthonormal vierbein field that cannot be specified by a unique covariant prescription and is therefore arbitrary to a large degree. Such a vierbein field must therefore also play a role in the definition of spinor propagators if these are to be genuinely independent of embedding spaces.

In our view, this point has obtained insufficient attention in the present literature on propagators in curved spaces.[9] Meanwhile we want to stress again the importance of inhomogeneous propagators like the Feynman propagators, which are vacuum expectation values of time-ordered field products. Time is here a many-valued reference function, and the introduction of a covering of AdS space is necessary. Discussion on this subject can be found in Refs. 4 and 8. It is important to note that these propagators are not just trivial extensions of the homogeneous propagators.[8] In the present paper we use the results obtained in Ref. 8 to find homogeneous and inhomogeneous propagators for spinor fields in AdS configuration space.

Furthermore, we have to specify implicit boundary conditions in order to obtain unique propagators.[8,10] In general this can be done by taking the solution, which goes fastest to zero, when a certain invariant quantity approaches minus infinity.[11,12]

In Sec. II we give a review of the scalar propagators obtained in Ref. 8 and discuss some properties.

In Sec. III we obtain spinor propagators as a function of the coordinates ξ^M ($M = 1,...,5$) of five-space, on the hyperboloid $\xi^M\xi_M = R^2 = \text{const} > 0$, which describes (a covering space of) AdS space. The generalized Dirac equation we use can be seen to arise from a fünfbein formalism, with trivial fünfbein $V_A^M = \delta_A^M$ (see also Ref. 4).

In Sec. IV we obtain the spinor propagators for a natural vierbien by performing a local rotation of the fünfbein, such that one of its "legs" points in the radial direction. This rotation can be performed in such a way that the AdS vierbein becomes independent of time. Strictly speaking, this is not a necessity, but it is certainly convenient. For practical purposes, it is better to stick to one choice of the vierbein field for all propagators independent of the reference point. Although we can use any coordinate system, in order to obtain a bag structure we choose special coordinates, which are obtained by performing a central projection.[13] This is done in Sec. V.

II. SCALAR PROPAGATORS

Consider a five-dimensional space with coordinates ξ^M ($M = 1,...,5$) and metric

$$\eta_{MN} = \text{diag}(-1,-1,-1,+1,+1)\,. \tag{2.1}$$

Anti-de Sitter space can be visualized as (the covering space of) the hyperboloid

$$\xi_M\xi^M = -\xi^{1^2} + \xi^{4^2} + \xi^{5^2} = R^2 = \text{const} > 0\,. \tag{2.2}$$

Transformations leaving this hyperboloid invariant form the covering group of SO(3,2). Therefore a winding number must be introduced, or alternatively we can use a many-valued reference function, which plays the role of time.

Consider the equation

$$\Box_5\psi = \left(\Box_4 + \frac{\partial^2}{\partial\xi^{5^2}}\right)\psi = 0\,. \tag{2.3}$$

We shall take as reference point $\xi_0^M = (0,0,0,0,R)$ and define the invariant quantity

$$\lambda = \xi_\rho\xi^\rho/R^2 = 1 - (\xi^{5^2}/R^2)\,. \tag{2.4}$$

We are interested in those functions that are invariant under transformations leaving the reference point invariant. They have the form

 0022-2488/86/112786-05$02.50

$$\psi(R,\lambda,n) = R^m \phi_{(n)}(\lambda,m)\,, \tag{2.5}$$

where n is a winding number.

Introduce the angular momentum operator

$$M_{MN} = i\left(\xi_M \frac{\partial}{\partial \xi^N} - \xi_N \frac{\partial}{\partial \xi^M}\right) \tag{2.6}$$

and

$$M^2 = \tfrac{1}{2} M_{MN} M^{MN}\,. \tag{2.7}$$

Then $\phi_{(n)}(\lambda,m)$ satisfies

$$(M^2 - m(m+3))\phi_{(n)}(\lambda,m) = 0\,. \tag{2.8}$$

The time variable t is introduced by

$$\begin{aligned} \xi^4 &= \sqrt{R^2 + \xi^2}\,\sin(t/R), \\ \xi^5 &= \sqrt{R^2 + \xi^2}\,\cos(t/R), \end{aligned} \tag{2.9}$$

and is many-valued in ξ^M-space. Because m and $-(m+3)$ are interchangeable, we limit ourselves to $m \geqslant -\frac{3}{2}$. The exceptional point $m = -\frac{3}{2}$ will not be considered. For the homogeneous propagators, satisfying (2.8) we have,[8] for $|\lambda| < 1$,

$$\begin{aligned} G^m_{(n)}(\xi^M) &= \cos[\pi(m+1)n] G^m_{(0)}(\xi^M) \\ &\quad - \sin[\pi(m+1)n] \overline{G}^m_{(0)}(\xi^M)\,, \\ \overline{G}^m_{(n)}(\xi^M) &= \sin[\pi(m+1)n] G^m_{(0)}(\xi^M) \\ &\quad + \cos[\pi(m+1)n] \overline{G}^m_{(0)}(\xi^M)\,, \end{aligned} \tag{2.10}$$

where

$$\begin{aligned} G^m_{(0)}(\xi^M) &= \frac{1}{2\pi R^2}\epsilon(\xi^4\xi^5)\left[\delta(\lambda) - \frac{(m+1)(m+2)}{4}\right. \\ &\quad \left.\times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right)\theta(\lambda)\right], \\ \overline{G}^m_{(0)}(\xi^M) &= \frac{1}{2\pi^2 R^2}\left[P\left(\frac{1}{\lambda}\right) - \frac{(m+1)(m+2)}{4}\right. \\ &\quad \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right) \\ &\quad \left.\times \ln|\lambda| - \chi(m,\lambda)\right], \end{aligned} \tag{2.11}$$

and $\chi(m,\lambda)$ is a real analytic function, regular in the domain $|\lambda| < 1$, with appropriate analytic continuations for $|\lambda| \geqslant 1$. The discrete function $\theta(x)$ is defined by $\int_{-\infty}^{x} \delta(x')\,dx'$ and $\epsilon(x)$ is defined by $2\theta(x) - 1$.

We can also define positive and negative frequency parts by

$$G^m_{\pm} = \tfrac{1}{2}(\overline{G}^m \pm iG^m)\,. \tag{2.12}$$

Then

$$G^m_{\pm(n)}(\xi^M) = e^{\mp \pi i(m+1)n} G^m_{\pm(0)}(\xi^M)\,, \tag{2.13}$$

where

$$\begin{aligned} G^m_{\pm(0)}(\xi^M) &= \lim_{\epsilon \to 0} \frac{1}{4\pi^2 R^2}\left[\frac{1}{\lambda_{\mp\epsilon}} - \frac{(m+1)(m+2)}{4}\right. \\ &\quad \times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right) \\ &\quad \left.\times \ln(-\lambda_{\mp\epsilon}) - \chi(m,\lambda)\right], \end{aligned} \tag{2.14}$$

for $|\lambda| < 1$ and

$$\lambda_{\pm\epsilon} = \lambda \pm i\epsilon\xi^4\xi^5/R^2\,. \tag{2.15}$$

The inhomogeneous propagators satisfying

$$(M^2 - m(m+3))G^m_{(n)}(\xi^M) = -R^2\delta^{(4)}(\xi^\mu)\delta_{n0}\,, \tag{2.16}$$

are the retarded and advanced Green's functions

$$G^m_{\substack{\text{ret}\\\text{adv}}} = \pm\theta(\pm t) G^m\,, \tag{2.17}$$

and the Feynman propagator

$$G^m_F = (1/2i)[\overline{G}^m + i\epsilon(t)G^m]\,. \tag{2.18}$$

Note that the singularity on the right-hand side of (2.16) occurs only on the principal sheet of the covering.

In the conformal invariant "massless" case ($m = -1$) we obtain

$$G_F^{-1}(\xi^M) = (1/4\pi^2 iR^2)[1/(\lambda - i\epsilon t\xi^4\xi^5)]\,, \quad \text{for } \epsilon \to 0\,. \tag{2.19}$$

To find the propagators for an arbitrary reference point ξ_0^M, we use the invariant forms

$$\lambda = 1 - [\gamma/R]^2 \tag{2.20}$$

and $\epsilon(\gamma S)$, where

$$S = \xi^4\xi_0^5 - \xi^5\xi_0^4 \tag{2.21}$$

and $\gamma = \xi^M \xi_{0M}$.

The definition of winding number n can be given as follows.

(i) $n = 0$ when ξ^M can be obtained from ξ_0^M by continuous displacement within the allowed domain without changing the sign of γ.

(ii) $\Delta n = \pm 1$, whenever γ changes sign and $\Delta t \geqslant 0$ with t given by (2.9).

For $|\lambda| < 1$, we find

$$\begin{aligned} G^m_{(0)}(\xi^M;\xi_0^M) &= \frac{1}{2\pi R^2}\epsilon(\gamma S)\left[\delta(\lambda) - \frac{(m+1)(m+2)}{4}\right. \\ &\quad \left.\times F\left(-\frac{m}{2}, \frac{m+3}{2}; 2; \lambda\right)\theta(\lambda)\right] \end{aligned} \tag{2.22}$$

and $\overline{G}^m_{(0)}$ is given by (2.11).

$G^m_{(n)}$ and $\overline{G}^m_{(n)}$ are obtained from these using (2.10). We can find the inhomogeneous propagators as follows.

Define t by (2.9) and t_0 by

$$\begin{aligned} \xi_0^4 &= \sqrt{R^2 + \xi_0^2}\,\sin(t_0/R), \\ \xi_0^5 &= \sqrt{R^2 + \xi_0^2}\,\cos(t_0/R). \end{aligned} \tag{2.23}$$

Then $t - t_0$ can only change sign when $\lambda \leqslant 0$, $n = 0$. Moreover, either $G^m(\xi^M;\xi_0^M) = 0$ or $\xi^M = \xi_0^M$, which makes $\epsilon(t-t_0)$, $\theta(t-t_0)$ effectively invariant functions when multiplied by $G^m(\xi^M;\xi_0^M)$. We find

$$G^m_{\substack{\text{ret}\\\text{adv}}} = \pm\theta[\pm(t-t_0)]G^m\,, \tag{2.24}$$

$$G^m_F = (1/2i)[\overline{G}^m + i\epsilon(t-t_0)G^m]\,. \tag{2.25}$$

If we define for all propagators

$$G^m(\xi,t;\xi_0,t_0) = G^m_{(n)}(\xi^M;\xi_0^M)\,, \tag{2.26}$$

then

$$[M^2 - m(m+3)]G(\xi,t;\xi_0,t_0) = -R^2\delta^{(3)}(\xi-\xi_0)\delta(t-t_0)\,. \tag{2.27}$$

III. SPINOR PROPAGATORS

Introduce the four-dimensional Dirac matrices

$$\bar{\gamma}^M = (i\gamma^5\gamma^\mu, \gamma^5)\,, \tag{3.1}$$

where γ^μ are the usual Minkowski-space Dirac matrices and $\gamma^5 = -i\gamma^1\gamma^2\gamma^3\gamma^4$. They satisfy

$$\{\bar{\gamma}^M, \bar{\gamma}^N\} = 2\eta^{MN}\,. \tag{3.2}$$

Consider the equation

$$\Box_5\psi = \bar{\gamma}^M\partial_M\bar{\gamma}^N\partial_N\psi = 0\,, \tag{3.3}$$

and write

$$\psi(R,\lambda,n) = R^m\phi_{(n)}(\lambda,m)\,. \tag{3.4}$$

The $\phi_{(n)}(\lambda,m)$ satisfies

$$\{R\bar{\gamma}^M\partial_M + [(m-1)/R\,]\bar{\gamma}^M\xi_M\} \times\{R\bar{\gamma}^N\partial_N + (m/R)\bar{\gamma}^N\xi_N\}\phi_{(n)}(\lambda,m) = 0\,, \tag{3.5}$$

which implies that

$$S_{(n)}(\lambda,m) \equiv \{R\bar{\gamma}^N\partial_N + (m/R)\bar{\gamma}^N\xi_N\}\phi_{(n)}(\lambda,m) \tag{3.6}$$

is a homogeneous spinor propagator if $\phi_{(n)}(\lambda,m)$ is a homogeneous scalar propagator. It satisfies the equation

$$\{R\bar{\gamma}^M\partial_M + [(m-1)/R\,]\bar{\gamma}^M\xi_M\}S_{(n)}(\lambda,m) = 0\,. \tag{3.7}$$

This result is the same as that found in Ref. 5.

Here a remark is in order. From (3.6) we see that a differential operation must be applied to a singular function [see Eq. (2.11)]. However, as has been shown in Ref. 8, the singular function can always be written as the limit of a differentiable function that is an exact solution of the original homogeneous equation. The correct procedure is to apply the operator $R\bar{\gamma}^N\partial_N$ to this function and then to take the appropriate limit.

We find the inhomogeneous spinor propagators by using inhomogeneous scalar propagators. They are given by the same formula (3.6) and satisfy the equation

$$(R\bar{\gamma}^M\partial_M + [(m-1)/R\,]\bar{\gamma}^M\xi_M)S^m_{(n)}(\xi^M) = R^2\delta^{(4)}(\xi^\mu)\delta_{n0}\,. \tag{3.8}$$

Also here a differential operation must be applied to a singular function that is the limit of a regular function. The latter satisfies the original inhomogeneous equation, but with the δ-like source function smeared out. The proper procedure for carrying out the differentiation is the same as for the homogeneous case.

As an example we give the Feynman propagator for $m = -1$,

$$S_F^{-1}(\xi^M) = \frac{1}{4\pi i R^3}\left\{\frac{\bar{\gamma}^\mu\xi_\mu(\lambda-2) + \bar{\gamma}^5\xi_5\lambda}{(\lambda - i\epsilon t\xi^4\xi^5)^2}\right\}, \tag{3.9}$$

in the limit $\epsilon\to 0$.

For an arbitrary reference point ξ_0^M we find

$$S_F^{-1}(\xi^M;\xi_0^M) = \frac{1}{4\pi i R^3}\left\{\frac{\bar{\gamma}^M\xi_M(\lambda-2) + 2(\gamma/R^2)\bar{\gamma}^M\xi_{0M}}{(\lambda - i\epsilon(t-t_0)\gamma S)^2}\right\} \tag{3.10}$$

in the limit $\epsilon\to 0$.

In order to obtain unique propagators, we take the solution that goes fastest to zero when $\lambda\to-\infty$, and limit ourselves to $m > -\frac{3}{2}$.[8,11]

IV. SPINOR PROPAGATORS FOR A NATURAL VIERBEIN FIELD

The generalized Dirac equation obtained in Sec. III can be seen to arise from a fünfbein formalism, with a special choice of the fünfbein:

$$V_A^M = \delta_A^M\,, \tag{4.1}$$

where

$$\bar{\gamma}^M = V_A^M\gamma^A\,, \tag{4.2}$$

and

$$\gamma^A = (i\gamma^5\gamma^\mu, \gamma^5)\,. \tag{4.3}$$

Here M is the "world index" and A the local index. We make a rotation of this fünfbein field, such that one of its "legs" points in a direction perpendicular to the hyperboloid [radial (R) direction]. Since the square of the Dirac operator in curved spaces is not simply related to the scalar field operator the introduction of this rotation is necessary. Therefore we introduce the transformation matrix

$$U(\xi,t) = e^{-(1/2)i\chi\hat{\xi}\cdot\gamma}e^{i\gamma^4 t/2R}\,, \tag{4.4}$$

where

$$\chi = \sinh^{-1}(|\xi|/R). \tag{4.5}$$

Then

$$\tilde{\gamma}^M \equiv V_A^M\gamma^A = U\bar{\gamma}^M U^{-1}\,, \tag{4.6}$$

where V_5^M points in the radial direction. Equation (3.7) can be written as

$$R\tilde{\gamma}^M[\partial_M + U(\partial_M U^{-1}) + [(m-1)/R^2]\xi_M] \times U(\xi,t)S^m(\xi,t) = 0\,. \tag{4.7}$$

This implies

$$\{-i\tilde{\gamma}^\mu D_\mu + [(m+1)/R\,]\}\gamma^5 U(\xi,t)S^m(\xi,t) = 0\,, \tag{4.8}$$

where

$$\tilde{\gamma}^\mu = V_\alpha^\mu\gamma^\alpha \quad (\mu \text{ world index}, \alpha \text{ local index})\,, \tag{4.9}$$

and where the covariant derivative D_μ is given by

$$D_\mu = \partial_\mu + \Gamma_\mu\,. \tag{4.10}$$

Here Γ_μ is defined as follows:

$$\Gamma_\mu = \tfrac{1}{2}\Delta_{\mu,\alpha\beta}\sigma^{\alpha\beta}\,, \tag{4.11}$$

with

$$\Delta_{\mu,\alpha\beta} = \Gamma_{\nu\sigma\mu}V_\alpha^\nu V_\beta^\sigma - V_{\sigma\alpha,\mu}V_\beta^\sigma\,, \tag{4.12}$$

and

$$\sigma^{\alpha\beta} = \tfrac{1}{4}[\gamma^\alpha,\gamma^\beta]\,,$$

and $\Gamma_{\nu\sigma\mu}$ is the affine connection. The matrices γ^{α} are the usual Minkowski-space Dirac matrices satisfying

$$\{\gamma^{\alpha},\gamma^{\beta}\} = 2\eta^{\alpha\beta} . \tag{4.13}$$

Equation (4.8) is the Dirac equation for anti-de Sitter space whose metric is given by

$$g^{\mu\nu} = V^{\mu}_{\alpha} V^{\nu}_{\beta} \eta^{\alpha\beta} . \tag{4.14}$$

The vierbein V^{μ}_{α} is given by

$$\begin{aligned} V^0_0 &= (1+\xi^2/R^2)^{-1/2} , \\ V^0_i &= V^i_0 = 0 , \\ V^i_j &= \{\delta^i_j + (\xi^i\xi^j/|\xi|^2)[(1+\xi^2/R^2)^{1/2} - 1]\} . \end{aligned} \tag{4.15}$$

We obtain for $g^{\mu\nu}$:

$$\begin{aligned} g^{00} &= (1+\xi^2/R^2)^{-1} , \quad g^{0i} = g^{i0} = 0 , \\ g^{ij} &= (\eta^{ij} - (\xi^i\xi^j/R^2)) . \end{aligned} \tag{4.16}$$

From Eq. (4.8) we see that the "massless" Dirac equation corresponds to $m = -1$, as should be the case. The solution to Eq. (4.8) can in this special frame be written as

$$US^m(\xi,t) = \{ -i\tilde{\gamma}^{\mu} D_{\mu} + [(m+2)/R]\}\gamma^5 U G^m(\xi,t) , \tag{4.17}$$

with $G^m(\xi,t)$ a scalar propagator. The solution for the inhomogeneous equation

$$\begin{aligned} &\{ -i\tilde{\gamma}^{\mu} D_{\mu} + [(m+1)/R]\} S^m(\xi,t) \\ &\quad = U(\xi,t)\delta^3(\xi)\delta(t) = \delta^3(\xi)\delta(t) \end{aligned} \tag{4.18}$$

is given by

$$S^m(\xi,t) = \{i\tilde{\gamma}^{\nu} D_{\nu} + [(m+2)/R]\} U G^m(\xi,t) , \tag{4.19}$$

where $G^m(\xi,t)$ is an inhomogeneous scalar propagator. Consider the Lagrangian

$$\mathcal{L} = \bar{\psi}^m(\bar{\gamma}^N \partial_N + [(m-1)/R]\bar{\gamma}^N \xi_N)\psi^m , \tag{4.20}$$

which can be written as

$$\begin{aligned} \mathcal{L}' &= \sqrt{-g}\,\mathcal{L} = \mathcal{L} \\ &= \bar{\psi}'^m(-i\tilde{\gamma}^{\mu} D_{\mu} + [(m+1)/R])\psi'^m , \end{aligned} \tag{4.21}$$

where

$$\bar{\psi}' = \bar{\psi} U^{-1} , \quad \psi' = \gamma^5 U \psi . \tag{4.22}$$

When quantized they satisfy the following commutation relation:

$$\{\psi'(\xi,t), \bar{\psi}'(\xi',t)\}\tilde{\gamma}^4 = \delta^3(\xi - \xi') . \tag{4.23}$$

We can write (3.8) as follows for the Feynman propagator with arbitrary reference point

$$\begin{aligned} &R(\bar{\gamma}^M \partial_M + [(m-1)/R]\bar{\gamma}^M \xi_M) \\ &\quad \times R i\langle 0|T\psi^m(\xi^M)\bar{\psi}^m(\xi^M_0)|0\rangle \\ &\quad = R^2 \delta^3(\xi - \xi_0)\delta(t - t_0) , \end{aligned} \tag{4.24}$$

which implies

$$\begin{aligned} &\{ -i\tilde{\gamma}^{\mu} D_{\mu} + [(m+1)/R]\}\gamma^5 U(\xi,t) \\ &\quad \times i\langle 0|T\psi^m(\xi^M)\bar{\psi}^m(\xi^M_0)|0\rangle \\ &\quad = \delta^3(\xi - \xi_0)\delta(t - t_0) U(\xi_0,t_0) , \end{aligned} \tag{4.25}$$

or

$$\begin{aligned} &\{ -i\tilde{\gamma}^{\mu} D_{\mu} + [(m+1)/R]\} i\langle 0|T\psi'^m(\xi,t)\bar{\psi}'^m(\xi_0,t_0)|0\rangle \\ &\quad = \delta^3(\xi - \xi_0)\delta(t - t_0) , \end{aligned} \tag{4.26}$$

which is consistent with (4.23).

Thus, for an arbitrary reference point ξ^M_0 the solution for the inhomogeneous propagators is given by

$$\begin{aligned} S^m(\xi,t;\xi_0,t_0) = &\{i\tilde{\gamma}^{\nu} D_{\nu} + [(m+2)/R]\} \\ &\times U(\xi,t) G^m(\xi,t;\xi_0,t_0) U^{-1}(\xi_0,t_0) . \end{aligned} \tag{4.27}$$

Note that since for $m = -1$ the scalar propagators are not unique,[8,10] the massless spinor propagators are not unique either. Uniqueness can be restored by putting $m \neq -1$ and taking the limit $m \to -1$.

V. SPINOR PROPAGATORS FOR CENTRAL PROJECTION COORDINATES

We can perform a central projection[13] defined by

$$t \to t , \quad \xi^i = x^i \cosh\chi . \tag{5.1}$$

For the metric $g^{\mu\nu} = V^{\mu}_{\alpha} V^{\nu}_{\beta} \eta^{\alpha\beta}$ we thus obtain

$$\begin{aligned} g^{00} &= (1 - \alpha r^2) , \quad \alpha = 1/R^2 , \\ g^{0j} &= g^{j0} = 0 , \\ g^{ij} &= (1 - \alpha r^2)(\eta^{ij} + \alpha x^i x^j) , \end{aligned} \tag{5.2}$$

with vierbein V^{μ}_{α}:

$$\begin{aligned} V^0_0 &= (1 - \alpha r^2)^{1/2} , \quad V^0_i = V^i_0 = 0 , \\ V^i_j &= (1 - \alpha r^2)^{1/2}\{\delta^i_j - \alpha x^i x^j [1 + (1 - \alpha r^2)^{1/2}]^{-1}\} . \end{aligned} \tag{5.3}$$

The anti-de Sitter metric is thus confined to a spherical bag of radius R in x^{μ}-space, which can be used as a geometrical description of hadrons.[2,3,13] The spinor propagators can be obtained from the propagators of Sec. IV by performing the coordinate transformation (5.1). In the flat-space limit ($R \to \infty$) we obtain for the massless Feynman propagator

$$S_F(\mathbf{x},t;\mathbf{x}_0,t_0) = -\frac{1}{2\pi^2} \frac{\gamma^{\mu}(x_{\mu} - x^0_{\mu})}{[(x - x_0)^2 - i\epsilon]^2} , \tag{5.4}$$

which is a solution of

$$-i\gamma^{\nu} \partial_{\nu} S_F(\mathbf{x},t;\mathbf{x}_0,t_0) = \delta^3(\mathbf{x} - \mathbf{x}_0)\delta(t - t_0) . \tag{5.5}$$

This is the correct result.

ACKNOWLEDGMENTS

We wish to thank Dr. E. van Beveren and Ir. K. Metzger for useful discussions.

Part of this work was included in the research program of the Stichting voor Fundamenteel Onderzoek der Materie (F.O.M.) with financial support from the Nederlandse Organisatie voor Zuiver-Wetenschappelijk Onderzoek (Z.W.O.).

[1] C. Isham, A. Salam, and J. Strathdee, Phys. Rev. D **3**, 867 (1971); S. Fubini, Nuovo Cimento A **34**, 521 (1976); B. Zumino, Nucl. Phys. B **127**, 189 (1977); A. Salam and J. Strathdee, Phys. Rev. D **18**, 4596 (1978); Phys. Lett. B **67**, 429 (1977); Z. Haba, Phys. Rev. D **18**, 4610 (1978); Phys. Lett. B **78**, 421 (1978); C. Sivaram, and K. P. Sinha, Phys. Rep. **51**, 111 (1979).

[2] C. Dullemond and E. van Beveren, Phys. Rev. D **28**, 1028 (1983); C.

Dullemond, J. Math. Phys. **25**, 2638 (1984); "New particle production," in *Proceedings of the XIX Rencontre de Moriond*, edited by J. Tran Tranh Van (Editions Frontières, Gif sur Yvette, 1984), p. 847; E. van Beveren, T. A. Rijken, C. Dullemond, and G. Rupp, in *Proceedings of the Bielefeld Workshop on Resonances*, edited by S. Albeverio, L. S. Ferreira, and L. Streit (Springer, Berlin, 1984), p. 331.

[3]C. Dullemond, T. A. Rijken, and E. van Beveren, Nuovo Cimento A **80**, 401 (1984).

[4]C. Fronsdal, Phys. Rev. D **10**, 589 (1974).

[5]C. Fronsdal and R. B. Haugen, Phys. Rev. D **12**, 3810 (1975).

[6]C. Fronsdal, Phys. Rev. D **20**, 848 (1979).

[7]J. Fang and C. Fronsdal, Phys. Rev. D **22**, 1361 (1980).

[8]C. Dullemond and E. van Beveren, J. Math. Phys. **26**, 2050 (1985).

[9]*Relativity, Groups and Topology*, edited by B. S. DeWitt and C. DeWitt (Gordon and Breach, New York, 1963); P. Candelas and D. J. Raine, Phys. Rev. D **12**, 965 (1975); E. Mottola, *ibid.* **31**, 754 (1985); N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge U. P., Cambridge, 1982).

[10]C. Fronsdal, Phys. Rev. D **12**, 3819 (1975); S. J. Avis, C. J. Isham, and D. Storey, *ibid.* **18**, 3565 (1978).

[11]For certain physically interesting systems this may be a too stringent condition. See P. Breitenlohner and D. Z. Freedman, Ann. Phys. (NY) **144**, 249 (1982).

[12]For a general discussion see, e.g., S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space Time* (Cambridge U. P., Cambridge, 1973).

[13]E. van Beveren, T. A. Rijken, and C. Dullemond, Phys. Rev. D **30**, 1103 (1984).

An infinite number of infinite hierarchies of conserved quantities of the Federbush model

P. H. M. Kersten
Department of Applied Mathematics, Twente University of Technology, P. O. Box 217, 7500 AE Enschede, The Netherlands

H. M. M. Ten Eikelder
Department of Mathematics and Computing Science, Eindhoven University of Technology, P. O. Box 513, 5600 MB Eindhoven, The Netherlands

(Received 14 March 1986; accepted for publication 11 June 1986)

The construction of two Lie–Bäcklund transformations is given, which are Hamiltonian vector fields leading to an infinite number of hierarchies of conserved functionals and associated Lie–Bäcklund transformations.

I. INTRODUCTION AND GENERAL

In two recent papers[1,2] we constructed eight [in effect four, $Y_i^+, Y_i^-, Z_i^+, Z_i^-$ $(i \in \mathbb{Z})$] infinite hierarchies of Lie–Bäcklund transformations of the Federbush model.[3] We conjectured that the hierarchies Y_i^+, Y_i^- $(i \in \mathbb{Z})$ are (x,t) independent, while the hierarchies Z_i^+, Z_i^- $(i \in \mathbb{Z})$ are linear in x and t. These Lie–Bäcklund transformations turned out to be Hamiltonian vector fields[4,5] and the corresponding Hamiltonian densities were given. In this way we obtained t-independent and t-dependent conserved functionals for the Federbush model.

Now we shall construct two (x,t)-dependent Lie–Bäcklund transformations of degree 0, with respect to the grading, which are polynomial in x,t of degree 2 and from which we can obtain the creating and annihilating Lie–Bäcklund transformations $Z_{\pm 1}^{\pm}$, by taking the Lie bracket with the (x,t)-independent vector fields $Y_{\pm 1}^{\pm}$ (cf. the Appendix). Moreover these two vector fields turn out to be Hamiltonian vector fields and the associated Hamiltonian densities are given. This will be done in Sec. II. In Sec. III we prove a theorem from which we obtain an infinite number of infinite hierarchies of Hamiltonian vector fields, where the $Y_i^+, Y_i^-, Z_i^+ Z_i^-$ $(i \in \mathbb{Z})$ are just the first four of this infinite number of hierarchies. The Hamiltonian densities of the vector fields $Z_i^{\pm}$ $(i = -1,0,1)$, $Y_j^{\pm}$ $(j = -2,-1,0,1,2)$ are surveyed in an Appendix at the end of this paper for reasons of completeness. In this section we shall introduce the notions needed in Secs. II and III. All computations have been carried through on a DEC-system 20 computer, using the symbolic language REDUCE[6] and software packages[7,8] to do the huge computations at hand.

Lie–Bäcklund transformations are vector fields V defined on the infinite jet bundle of M,N, $J^\infty(M,N)$, where M is the space of independent variables and N the space of the dependent variables. A Lie–Bäcklund transformation of a differential equation is a vector field V defined on $J^\infty(M,N)$ satisfying the condition

$$\mathcal{L}_V(D^\infty I) \subset D^\infty I, \tag{1.1}$$

where I denotes a differential ideal associated to the differential equation at hand, while $D^\infty I$ denotes its infinite prolongation to $J^\infty(M,N)$; $\mathcal{L}_V$ is the Lie derivative with respect to the vector field V. Since the vector field V is supposed to depend only on a finite number of variables, condition (1.1) reduces to

$$\mathcal{L}_V I \subset D^r I \quad \text{for some } r. \tag{1.2}$$

Using this method we computed Lie–Bäcklund transformations of the Federbush model.[1] It can be shown that the Lie–Bäcklund transformations in this setting are just symmetries in the works of Magri[4] and Ten Eikelder[5] where (generators of) symmetries of partial differential equations of evolutionary type are described as transformations on special types of infinite-dimensional spaces. Suppose that

$$\frac{du}{dt} = \Omega^{-1}\, dH \tag{1.3}$$

is an infinite-dimensional Hamiltonian system, where Ω is the symplectic operator, H is the Hamiltonian, and dH is the Fréchet derivative of H. Then to each Hamiltonian symmetry (also called canonical symmetry) Y there corresponds by definition a Hamiltonian $F(Y)$ such that

$$Y = \Omega^{-1}\, dF(Y) \tag{1.4}$$

and the Poisson bracket of F and H vanishes.[4,5] Suppose that Y_1, Y_2 are two Hamiltonian symmetries, then $[Y_1, Y_2]$ is a Hamiltonian symmetry and

$$F([Y_2, Y_1]) = \{F(Y_1), F(Y_2)\}, \tag{1.5}$$

where $\{\cdot,\cdot\}$ is the Poisson bracket defined by

$$\{F(Y_1), F(Y_2)\} = \langle dF(Y_1), Y_2 \rangle, \tag{1.6}$$

where $\langle \cdot,\cdot \rangle$ denotes the contraction of a one-form and a vector field.

II. CONSTRUCTION OF TWO NEW LIE–BÄCKLUND TRANSFORMATIONS OF THE FEDERBUSH MODEL

We construct two Lie–Bäcklund transformations of the Federbush model. This model is described by

$$\begin{pmatrix} i(\partial_t + \partial_x) & -m(s) \\ -m(s) & i(\partial_t - \partial_x) \end{pmatrix} \begin{pmatrix} \psi_{s,1} \\ \psi_{s,2} \end{pmatrix} = 4s\pi\lambda \begin{pmatrix} |\psi_{-s,2}|^2 & \psi_{s,1} \\ -|\psi_{-s,1}|^2 & \psi_{s,2} \end{pmatrix} \quad (s = \pm 1), \tag{2.1}$$

where the $\psi_s(x,t)$ are two-component complex valued func-

 0022-2488/86/112791-06$02.50

tions. Suppressing the factor $4\pi(\lambda' = 4\pi\lambda)$ and introducing eight real variables $u_1,v_1,u_2,v_2,u_3,v_3,u_4,v_4$ by

$$\psi_{1,1} = u_1 + iv_1, \quad \psi_{-1,1} = u_3 + iv_3, \quad m(+1) = m_1,$$
$$\psi_{1,2} = u_2 + iv_2, \quad \psi_{-1,2} = u_4 + iv_4, \quad m(-1) = m_2, \tag{2.2}$$

Eq. (2.1) is rewritten as a system of eight nonlinear partial differential equations for the functions $u_1,...,v_4$, i.e.,

$$\begin{aligned}
&u_{1t} + u_{1x} - m_1 v_2 = \lambda R_4 v_1,\\
&-v_{1t} - v_{1x} - m_1 u_2 = \lambda R_4 u_1,\\
&u_{2t} - u_{2x} - m_1 v_1 = -\lambda R_3 v_2,\\
&-v_{2t} + v_{2x} - m_1 u_1 = -\lambda R_3 u_2,\\
&u_{3t} + u_{3x} - m_2 v_4 = -\lambda R_2 v_3,\\
&-v_{3t} - v_{3x} - m_2 u_4 = -\lambda R_2 u_3,\\
&u_{4t} - u_{4x} - m_2 v_3 = \lambda R_1 v_4,\\
&-v_{4t} + v_{4t} - m_2 u_3 = \lambda R_1 u_4,
\end{aligned} \tag{2.3}$$

where in (2.3)

$$\begin{aligned}
&R_1 = u_1^2 + v_1^2, \quad R_2 = u_2^2 + v_2^2,\\
&R_3 = u_3^2 + v_3^2, \quad R_4 = u_4^2 + v_4^2,
\end{aligned} \tag{2.4a}$$

and, for further use,

$$R_{12} = R_1 + R_2, \quad R_{34} = R_3 + R_4. \tag{2.4b}$$

In two recent papers we obtained Lie–Bäcklund transformations for this model; results that are surveyed in the Appendix for reasons of completeness. Motivated by the results obtained previously,[2] i.e., the existence of four infinite hierarchies of Hamiltonian vector fields, two hierarchies probably being independent of x and t [Y_i^+, Y_i^- ($i \in \mathbb{Z}$)] and two hierarchies probably being linear in x and t [Z_i^+, Z_i^- ($i \in \mathbb{Z}$)]; we now want to search for a Lie–Bäcklund transformation that is polynomial in x,t of degree 2.

We require the vector field to be of degree 0 with respect to the grading of (2.3),

$$\begin{aligned}
&\deg(u_1) = \cdots = \deg(v_4) = 1,\\
&\deg(x) = \deg(t) = -2,\\
&\deg(\partial_x) = \deg(\partial_t) = 2,\\
&\deg(m_1) = \deg(m_2) = 2.
\end{aligned} \tag{2.5}$$

The vector field has the following required structure:

$$\begin{aligned}
Y^+(2,0) = {} & x^2(\alpha_1 Y_2^+ + \alpha_2 m_1 Y_1^+ + \alpha_3 m_1^2 Y_0^+ + \alpha_4 m_1 Y_{-1}^+ + \alpha_5 Y_{-2}^+) + 2xt(\beta_1 Y_2^+ + \beta_2 m_1 Y_1^+ + \beta_3 m_1^2 Y_0^+\\
& + \beta_4 m_1 Y_{-1}^+ + \beta_5 Y_{-2}^+) + t^2(\gamma_1 Y_2^+ + \gamma_2 m_1 Y_1^+ + \gamma_3 m_1^2 Y_0^+ + \gamma_4 m_1 Y_{-1}^+ + \gamma_5 Y_{-2}^+)\\
& + xC_1^+ + tC_2^+ + C_0^+,
\end{aligned} \tag{2.6}$$

where the Y_i^+ ($i = -2, -1, 0, 1, 2$) are the vector fields associated to the conserved functionals $F(Y_i^+)$ surveyed in the Appendix; $\alpha_i, \beta_i, \gamma_i$ ($i = 1,...,5$) being constant, while C_1^+, C_2^+, C_0^+ are vector fields of degree 2,2, and 0, respectively. Substituting (2.6) into the Lie–Bäcklund condition (1.2),

$$\mathcal{L}_V I \subset D^2 I, \tag{2.7}$$

and solving the resulting overdetermined system of partial differential equations for the coefficients $\alpha_i, \beta_i, \gamma_i$ ($i = 1,...,5$) and the vector fields C_0^+, C_1^+, C_2^+ using (2.4), we obtained the following result:

$$Y^+(2,0) = x^2(Y_2^+ - \tfrac{1}{2}m_1^2 Y_0^+ + Y_{-2}^+) + 2xt(Y_2^+ - Y_{-2}^+) + t^2(Y_2^+ + \tfrac{1}{2}m_1^2 Y_0^+ + Y_{-2}^+) + xC_1^+ + tC_2^+, \tag{2.8}$$

where in (2.8)

$$\begin{aligned}
C_1^+ = {} & (-2v_{1x} - m_1 u_2 - \lambda R_{34} u_1)\partial_{u_1} + (+2u_{1x} - m_1 v_2 - \lambda R_{34} v_1)\partial_{v_1} + (-2v_{2x} + m_1 u_1 - \lambda R_{34} u_2)\partial_{u_2}\\
& + (+2u_{2x} + m_1 v_1 - \lambda R_{34} v_2)\partial_{v_2},\\
C_2^+ = {} & (+2v_{1x} + m_1 u_2 + \lambda R_{34} u_1)\partial_{u_1} + (-2u_{1x} + m_1 v_2 + \lambda R_{34} v_1)\partial_{v_1} + (-2v_{2x} + m_1 u_1 - \lambda R_{34} u_2)\partial_{u_2}\\
& + (+2u_{2x} + m_1 v_1 - \lambda R_{34} v_2)\partial_{v_2},
\end{aligned} \tag{2.9}$$

while in (2.6)

$$C_0^+ = 0. \tag{2.10}$$

In a similar way,[1,2] motivated by the structure of the Lie algebra, we obtain another Lie–Bäcklund transformation, i.e.,

$$Y^-(2,0) = x^2(Y_2^- - \tfrac{1}{2}m_2^2 Y_0^- + Y_{-2}^-) + 2xt(Y_2^- - Y_{-2}^-) + t^2(Y_2^- + \tfrac{1}{2}m_2^2 Y_0^- + Y_{-2}^-) + xC_1^- + tC_2^-, \tag{2.11}$$

where in (2.11)

$$\begin{aligned}
C_1^- = {} & (-2v_{3x} - m_2 u_4 + \lambda R_{12} u_3)\partial_{u_3} + (+2u_{3x} - m_2 v_4 + \lambda R_{12} v_3)\partial_{v_3} + (-2v_{4x} + m_2 u_3 + \lambda R_{12} u_4)\partial_{u_4}\\
& + (+2u_{4x} + m_2 v_3 + \lambda R_{12} v_4)\partial_{v_4},\\
C_2^- = {} & (+2v_{3x} + m_2 u_4 - \lambda R_{12} u_3)\partial_{u_3} + (-2u_{3x} + m_2 v_4 - \lambda R_{12} v_3)\partial_{v_3} + (-2v_{4x} + m_2 u_3 + \lambda R_{12} u_4)\partial_{u_4}\\
& + (+2u_{4x} + m_2 v_2 + \lambda R_{12} v_4)\partial_{v_4}.
\end{aligned} \tag{2.12}$$

To give an idea of the action of the vector fields $Y^+(2,0)$ and $Y^-(2,0)$, we compute the Lie bracket with the vector fields $Y_1^+, Y_0^+, Y_{-1}^+, Y_1^-, Y_0^-, Y_{-1}^-$ yielding the following results:

$$[Y^+(2,0),Y_1^+] = +2Z_1^+, \quad [Y^-(2,0),Y_1^-] = +2Z_1^-, \quad [Y^+(2,0),Y_0^+] = 0, \quad [Y^-(2,0),Y_0^-] = 0,$$
$$[Y^+(2,0),Y_{-1}^+] = -2Z_{-1}^+, \quad [Y^-(2,0),Y_{-1}^-] = -2Z_{-1}^-, \quad [Y^+(2,0),Y_i^-] = 0, \tag{2.13}$$
$$[Y^-(2,0),Y_i^+] = 0 \quad (i = -1,0,1).$$

These results suggest setting

$$Y^{\pm}(1,i) = Z_i^{\pm} \quad \text{and} \quad Y^{\pm}(0,i) = Y_i^{\pm} \quad (i \in \mathbf{Z}).$$

Now we arrive at the following remarkable fact: the vector fields $Y^+(2,0)$ and $Y^-(2,0)$ are again Hamiltonian vector fields, the corresponding Hamiltonian densities being given by

$$\begin{aligned}&\tilde{F}(Y^-(2,0))\\ &= x^2(\tilde{F}(Y_2^-) - \tfrac{1}{2}m_2^2\tilde{F}(Y_0^-) + \tilde{F}(Y_{-2}^-)) + 2xt(\tilde{F}(Y_2^-) - \tilde{F}(Y_{-2}^-)) + t^2(\tilde{F}(Y_2^-) + \tfrac{1}{2}m_2^2\tilde{F}(Y_0^-) + \tilde{F}(Y_{-2}^-))\\ &= (x+t)^2\tilde{F}(Y_2^-) - \tfrac{1}{2}m_2^2(x+t)(x-t)\tilde{F}(Y_0^-) + (x-t)^2\tilde{F}(Y_{-2}^-)\end{aligned} \tag{2.14a}$$

and

$$\tilde{F}(Y^+(2,0)) = (x+t)^2\tilde{F}(Y_2^+) - \tfrac{1}{2}m_1^2(x+t)(x-t)\tilde{F}(Y_0^+) + (x-t)^2\tilde{F}(Y_{-2}^+), \tag{2.14b}$$

where the densities $\tilde{F}(Y_i^{\pm})$ $(i = -2,0,2)$ are given in the Appendix.

This result shows a remarkable resemblance to the results for the Benjamin–Ono equation.[9]

III. PROOF OF THE EXISTENCE OF AN INFINITE NUMBER OF HIERARCHIES

In this section we shall first prove a generalization of a lemma proved in Ref. 2. The main theorem of this section is a direct application of Lemma 3.1 to the special cases at hand and leads to the existence of an infinite number of infinite hierarchies of algebraically independent conserved functionals for the Federbush model. The associated Lie–Bäcklund transformations are obtained from these results by application of formula (1.4).

We state the following lemma.

Lemma 3.1: Let $H_n^r(u,v)$, $K_n^r(u,v)$ be defined by

$$\begin{aligned}H_n^r(u,v) &= \int_{-\infty}^{\infty} x^r(u_n^2 + v_n^2) \quad (r,n = 0,1,\ldots),\\ K_n^r(u,v) &= \int_{-\infty}^{\infty} x^r(u_{n+1}v_n - v_{n+1}u_n) \quad (r,n = 0,1,\ldots),\end{aligned} \tag{3.1}$$

where in (3.1)

$$u_n = \left(\frac{d}{dx}\right)^n u, \quad v_n = \left(\frac{d}{dx}\right)^n v,$$

and r,n such that the degree of H_n^r, K_n^r is positive. Define the Poisson bracket of functionals F,L by

$$\{F,L\} = \int_{-\infty}^{\infty} \left(+ \frac{\delta F}{\delta v}\frac{\delta L}{\delta u} - \frac{\delta F}{\delta u}\frac{\delta L}{\delta v}\right), \tag{3.2}$$

then

$$\{H_1^1, H_n^r\} = 4(n-r)K_n^r, \tag{3.3a}$$
$$\{H_1^1, K_n^r\} = (4(n-r)+2)H_{n+1}^r + r(r-1)(r-n-1)H_n^{r-2}, \tag{3.3b}$$
$$\{H_1^2, H_n^r\} = 4(2n-r)K_n^{r+1}, \tag{3.3c}$$
$$\{H_1^2, K_n^r\} = (2n+1-r)(4H_{n+1}^{r+1} - r^2H_n^{r-1}) \quad (r,n = 0,1,\ldots). \tag{3.3d}$$

Proof: Relations (3.3a) and (3.3b) are generalizations of formulas given in Ref. 2 and can be proved in a similar way. We now prove (3.3c) and (3.3b). Calculation of the Fréchet derivatives of H_n^r, K_n^r yields

$$\frac{\delta H_n^r}{\delta u} = \left(-\frac{d}{dx}\right)^n (2x^r u_n), \quad \frac{\delta H_n^r}{\delta v} = \left(-\frac{d}{dx}\right)^n (2x^r v_n), \tag{3.4a}$$

$$\frac{\delta K_n^r}{\delta u} = \left(-\frac{d}{dx}\right)^{n+1}(x^r v_n) - \left(-\frac{d}{dx}\right)^n (x^r v_{n+1}), \quad \frac{\delta K_n^r}{\delta v} = -\left(-\frac{d}{dx}\right)^{n+1}(x^r u_n) + \left(-\frac{d}{dx}\right)^n (x^r u_{n+1}). \tag{3.4b}$$

Substitution of (3.4a) into (3.2) results in

$$\{H_1^2,H_n^r\}=\int_{-\infty}^{\infty}-\frac{d}{dx}(2x^2v_1)\cdot(-1)^n\left(\frac{d}{dx}\right)^n(2x^ru_n)+\frac{d}{dx}(2x^2u_1)\cdot(-1)^n\left(\frac{d}{dx}\right)^n(2x^rv_n)$$

$$=(-1)^n(-1)^{n-1}\int_{-\infty}^{\infty}\left(\frac{d}{dx}\right)^n(2x^2u_1)\frac{d}{dx}(2x^rv_n)-\left(\frac{d}{dx}\right)^n(2x^2v_1)\frac{d}{dx}(2x^ru_n)$$

$$=-4\int_{-\infty}^{\infty}(x^2u_{n+1}+2nxu_n+n(n-1)u_{n-1})(x^rv_{n+1}+rx^{r-1}v_n)$$

$$-(x^2v_{n+1}+2nxv_n+n(n-1)v_{n-1})(x^ru_{n+1}+rx^{r+1}u_n)$$

$$=-4\int_{-\infty}^{\infty}rx^{r+1}(u_{n+1}v_n-v_{n+1}u_n)-2nx^{r+1}(u_{n+1}v_n-v_{n+1}u_n)$$

$$+n(n-1)x^r(v_{n+1}u_{n-1}-u_{n+1}v_{n-1})+n(n-1)rx^{r-1}(v_nu_{n-1}-u_nv_{n-1})=4(2n-r)K_n^{r+1},\quad(3.5)$$

which proves relation (3.3c). The last equality in (3.5) results from the fact that the last two terms are just a total derivative of $n(n-1)x^r(v_nu_{n-1}-u_nv_{n-1})$.

In order to prove (3.3d) we substitute (3.4a) and (3.4b) into (3.2), which results in

$$\{H_1^2,K_n^r\}=\int_{-\infty}^{\infty}-\frac{d}{dx}(2x^2v_1)\cdot\left[(-1)^{n+1}\left(\frac{d}{dx}\right)^{n+1}(x^rv_n)-(-1)^n\left(\frac{d}{dx}\right)^n(x^rv_{n+1})\right]$$

$$+\frac{d}{dx}(2x^2u_1)\cdot\left[(-1)^n\left(\frac{d}{dx}\right)^{n+1}(x^ru_n)+(-1)^n\left(\frac{d}{dx}\right)^n(x^ru_{n+1})\right].\quad(3.6)$$

Integration, n times, of the terms in brackets leads to

$$\{H_1^2,K_n^r\}=2\int_{-\infty}^{\infty}\left(\frac{d}{dx}\right)^{n+1}(x^2v_1)\cdot\left(\frac{d}{dx}(x^rv_n)+x^rv_{n+1}\right)+\left(\frac{d}{dx}\right)^{n+1}(x^2u_1)\cdot\left(\frac{d}{dx}(x^ru_n)+x^ru_{n+1}\right)$$

$$=2\int_{-\infty}^{\infty}(x^2v_{n+2}+2(n+1)xv_{n+1}+n(n+1)v_n)(2x^rv_{n+1}+rx^{r-1}v_n)$$

$$+(x^2u_{n+2}+2(n+1)xu_{n+1}+n(n+1)u_n)(2x^ru_{n+1}+rx^{r-1}u_n).\quad(3.7)$$

Expanding the expressions in (3.7), we arrive after a short computation at

$$\{H_1^2,K_n^r\}=(2n+1-r)(4H_{n+1}^{r+1}-r^2H_n^{r-1}),\quad(3.8)$$

which proves (3.3d).

We are now in a position to prove the main theorem of this section.

Theorem 3.1: The conserved functionals $F(Y^{\pm}(2,0))$ associated to the Lie–Bäcklund transformations $Y^{\pm}(2,0)$ generate an infinite number of hierarchies, starting at the $F(Y_j^+)_{j\in\mathbb{Z}}, F(Y_j^-)_{j\in\mathbb{Z}}$ hierarchies by repeated action of the Poisson bracket.

The $F(Z_j^+)_{j\in\mathbb{Z}}, F(Z_j^-)_{j\in\mathbb{Z}}$ hierarchies are obtained by the first step of this procedure [cf. (2.13)]. Moreover the $F(Y_j^+)_{j\in\mathbb{Z}}$, $F(Y_j^-)_{j\in\mathbb{Z}}$ hierarchies are obtained from $F(Y_{+1}^-)$ by repeated action of the conserved functional

$$F(Z_{\pm1}^{\pm})=\pm\tfrac{1}{2}F([Y^{\pm}(2,0),Y_{\pm1}^{\pm}])\quad(3.9)$$

(cf. Table I).

Proof: The proof of theorem 3.1 is a straightforward application of Lemma 3.1 and the observation that the (λ,m_1,m_2)-independent parts of the conserved densities associated to $Y_{\pm1}^{\pm}$, $Y(+2,0)$, $Y(-2,0)$, (A3), (A4), (2.14a), and (2.14b) are given by

$$Y_1^+\to-\tfrac{1}{2}(u_{2x}v_2-v_{2x}u_2),\quad Y_{-1}^+\to-\tfrac{1}{2}(u_{1x}v_1-v_{1x}u_1),\quad(3.10)$$

$$Y_1^-\to-\tfrac{1}{2}(u_{4x}v_4-v_{4x}u_4),\quad Y_{-1}^-\to-\tfrac{1}{2}(u_{3x}v_3-v_{3x}u_3),$$

$$Y^+(2,0)\to-\tfrac{1}{2}(x+t)^2(u_{2x}^2+v_{2x}^2)-\tfrac{1}{2}(x-t)^2(u_{1x}^2+v_{1x}^2),\quad(3.11)$$

$$Y^-(2,0)\to-\tfrac{1}{2}(x+t)^2(u_{4x}^2+v_{4x}^2)-\tfrac{1}{2}(x-t)^2(u_{3x}^2+v_{3x}^2).$$

Note that in applying Lemma 3.1 we have to choose $(u,v)\equiv(u_2,v_2)$, $(u,v)=(u_1,v_1)$,..., where now (u_i,v_i) $(i=1,...,4)$ refer to (2.2)!

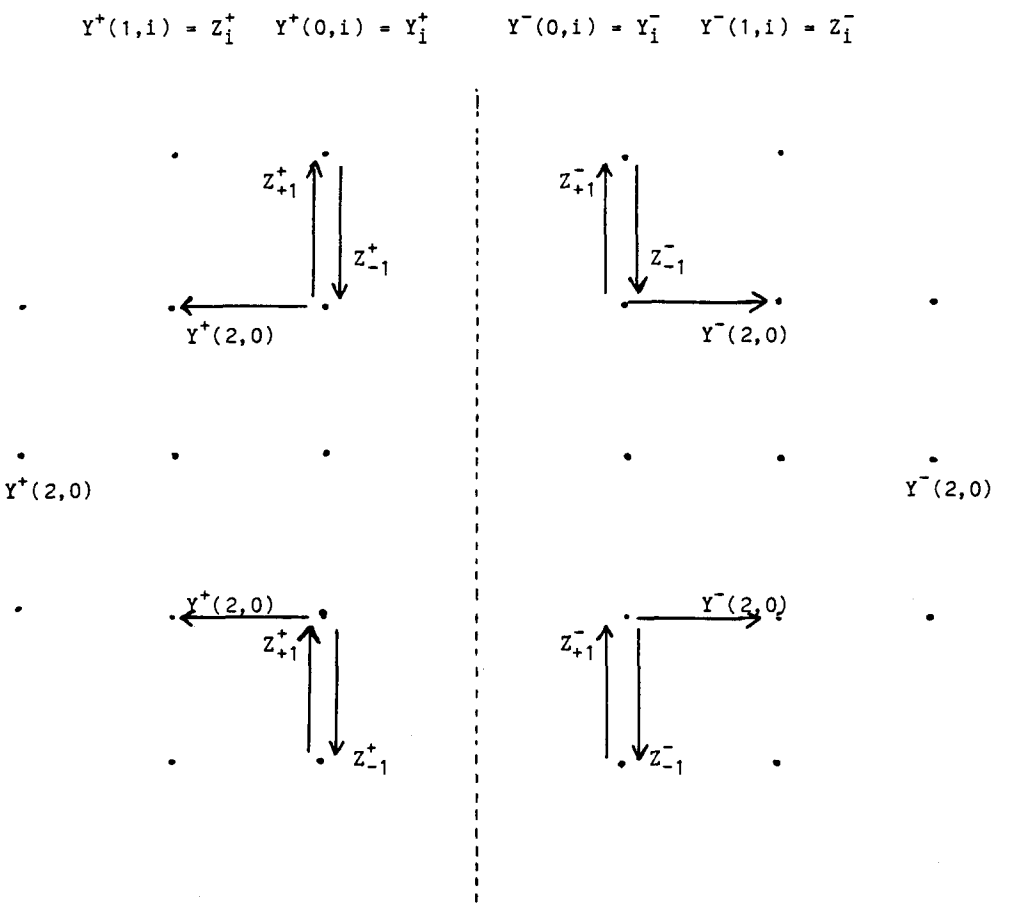
Remark: The Lie–Bäcklund transformations of degree 0, $Y_0^+=Y^+(0,0)$, $Z_0^+=Y^+(1,0)$, $Y^+(2,0)$ and $Y_0^-=Y^-(0,0)$, $Z_0^-=Y^-(1,0)$, $Y^-(2,0)$ being just the first few of them, can probably be obtained by the action of $Z_{\pm1}^{\pm}$ on the vector fields of degree 1 (cf. Ref. 1), i.e.,

$$Y^{\pm}(k,0)=\alpha_k[Z_{\pm1}^{\pm},Y^{\pm}(k,\pm1)].$$

IV. CONCLUSION

By the construction of two Hamiltonian vector fields $Y^+(2,0)$ and $Y^-(2,0)$ we construct an infinite number of infinite hierarchies, the elements of which are all Hamiltonian vector fields. The associated conserved functionals are obtained by the action of the Poisson bracket.

TABLE I. The Lie algebraic picture of the Lie–Bäcklund transformations.



APPENDIX: CONSERVED FUNCTIONALS FOR THE FEDERBUSH MODEL

We summarize here some of the results obtained in Ref. 2 that are of interest in Sec. II. We derived the following conserved functionals:

$$F(*)=\int_{-\infty}^{\infty}\tilde{F}(*)dx, \tag{A1}$$

where the densities $\tilde{F}(*)$ are given by

$$\tilde{F}(Y_0^+)=\tfrac{1}{2}(R_1+R_2),\quad \tilde{F}(Y_0^-)=\tfrac{1}{2}(R_3+R_4), \tag{A2}$$

$$\begin{aligned}
\tilde{F}(Y_1^+) &= -\tfrac{1}{2}(u_{2x}v_2-u_2v_{2x})+(\lambda/4)R_{34}R_2\\
&\quad -\tfrac{1}{2}m_1(u_1u_2+v_1v_2),\\
\tilde{F}(Y_{-1}^+) &= -\tfrac{1}{2}(u_{1x}v_1-u_1v_{1x})+(\lambda/4)R_{34}R_1\\
&\quad +\tfrac{1}{2}m_1(u_1u_2+v_1v_2),\\
\tilde{F}(Y_1^-) &= -\tfrac{1}{2}(u_{4x}v_4-u_4v_{4x})-(\lambda/4)R_{12}R_4\\
&\quad -\tfrac{1}{2}m_2(u_3u_4+v_3v_4),\\
\tilde{F}(Y_{-1}^-) &= -\tfrac{1}{2}(u_{3x}v_3-u_3v_{3x})-(\lambda/4)R_{12}R_3\\
&\quad +\tfrac{1}{2}m_2(u_3u_4+v_3v_4),
\end{aligned} \tag{A3}$$

and

$$\begin{aligned}
\tilde{F}(Y_2^+) &= -\tfrac{1}{2}(u_{2x}^2+v_{2x}^2)+(\lambda/2)R_{34}(u_{2x}v_2-u_2v_{2x})\\
&\quad -\tfrac{1}{2}m_1(u_{2x}v_1-u_1v_{2x})-\tfrac{1}{8}\lambda^2R_{34}^2R_2\\
&\quad +\tfrac{1}{4}m_1\lambda R_{34}(u_1u_2+v_1v_2)-\tfrac{1}{8}m_1^2R_{12},\\
\tilde{F}(Y_{-2}^+) &= -\tfrac{1}{2}(u_{1x}^2+v_{1x}^2)+(\lambda/2)R_{34}(u_{1x}v_1-u_1v_{1x})\\
&\quad +\tfrac{1}{2}m_1(u_{1x}v_2-u_2v_{1x})-\tfrac{1}{8}\lambda^2R_{34}^2R_1\\
&\quad -\tfrac{1}{4}m_1\lambda R_{34}(u_1u_2+v_1v_2)-\tfrac{1}{8}m_1^2R_{12},\\
\tilde{F}(Y_2^-) &= -\tfrac{1}{2}(u_{4x}^2+v_{4x}^2)-(\lambda/2)R_{12}(u_{4x}v_4-u_4v_{4x})\\
&\quad -\tfrac{1}{2}m_2(u_{4x}v_3-u_3v_{4x})-\tfrac{1}{8}\lambda^2R_{12}^2R_4\\
&\quad -\tfrac{1}{4}m_2\lambda R_{12}(u_3u_4+v_3v_4)-\tfrac{1}{8}m_2^2R_{34},\\
\tilde{F}(Y_{-2}^-) &= -\tfrac{1}{2}(u_{3x}^2+v_{3x}^2)-(\lambda/2)R_{12}(u_{3x}v_3-u_3v_{3x})\\
&\quad +\tfrac{1}{2}m_2(u_{3x}v_4-u_4v_{3x})-\tfrac{1}{8}\lambda^2R_{12}^2R_3\\
&\quad +\tfrac{1}{4}m_2\lambda R_{12}(u_3u_4+v_3v_4)-\tfrac{1}{8}m_2^2R_{34}.
\end{aligned} \tag{A4}$$

The t-dependent conserved functionals are

$$\tilde{F}(Z_0^+) = (x+t)\tilde{F}(Y_1^+) - (x-t)\tilde{F}(Y_{-1}^+),$$
$$\tilde{F}(Z_0^-) = (x+t)\tilde{F}(Y_1^-) - (x-t)\tilde{F}(Y_{-1}^-), \tag{A5}$$

and

$$\tilde{F}(Z_1^+) = (x+t)\tilde{F}(Y_2^+) - \tfrac{1}{4}m_1^2(x-t)\tilde{F}(Y_0^+),$$
$$\tilde{F}(Z_{-1}^+) = -(x-t)\tilde{F}(Y_{-2}^+) + \tfrac{1}{4}m_1^2(x+t)\tilde{F}(Y_0^+), \tag{A6}$$
$$\tilde{F}(Z_1^-) = (x+t)\tilde{F}(Y_2^-) - \tfrac{1}{4}m_2^2(x-t)\tilde{F}(Y_0^-),$$
$$\tilde{F}(Z_{-1}^-) = -(x-t)\tilde{F}(Y_{-2}^-) + \tfrac{1}{4}m_2^2(x+t)\tilde{F}(Y_0^-).$$

The vector fields $Y_i^{\pm}$ $(i=-2,-1,0,1,2)$ and $Z_j^{\pm}$ $(j=-1,0,1)$ obtained from (A2)–(A6) by

$$Y = \Omega^{-1}\, dF(Y). \tag{A7}$$

[1] P. H. M. Kersten, "Creating and annihilating Lie–Bäcklund transformations of the Federbush model," J. Math. Phys. **27**, 1139 (1986).
[2] P. H. M. Kersten and H. M. M. Ten Eikelder, "Infinite hierarchies of t-independent and t-dependent conserved functionals of the Federbush model," J. Math. Phys. **27**, 214 (1986).
[3] S. N. M. Ruijsenaars, "The Wightman Axioms for the Fermionic Federbush model," Commun. Math. Phys. **87**, 181 (1982).
[4] F. Magri, "A simple model of the integrable Hamiltonian equation," J. Math. Phys. **19**, 1156 (1978).
[5] H. M. M. Ten Eikelder, "Symmetries for dynamical and Hamiltonian systems," CWI Tract 17, Centre of Mathematics and Computer Science, Amsterdam, 1985.
[6] A. C. Hearn, *REDUCE User's Manual* (Version 3.0) (The Rand Corporation, Santa Monica, CA, 1983).
[7] P. K. H. Gragert, Ph.D. thesis, Twente University of Technology, Enschede, The Netherlands, 1981.
[8] P. H. M. Kersten, Ph.D. thesis, Twente University of Technology, Enschede, The Netherlands, 1985.
[9] L. J. F. Broer and H. M. M. Ten Eikelder, "Constants of the motion for the Benjamin–Ono and related equations," Phys. Lett. A **92**, 56 (1982).

Gauge theories, the holonomy operator, and the Riemann–Hilbert problem

E. T. Newman
Department of Physics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260 and Institute for Theoretical Physics, University of California, Santa Barbara, California 93106

(Received 13 May 1986; accepted for publication 11 July 1986)

The so-called Riemann–Hilbert problem has arisen and plays a major role in the study of many nonlinear integrable systems, such as the sine–Gordon equation, stationary axial symmetric Einstein equations, etc. Here it is shown how the Riemann–Hilbert problem arises naturally in the study of self-dual Yang–Mills fields in Minkowski space via a simple geometric construction of the holonomy operator on anti-self-dual planes. This Riemann–Hilbert problem is then converted to a linear homogeneous differential equation that is considerably simpler to study than the original problem. Finally it is shown that the nonlinear equation of Yang for self-dual fields is easily understood from the holonomy point of view.

I. INTRODUCTION

Over the past seven or eight years there has been, for a variety of reasons, considerable interest in those solutions of the Yang–Mills equations[1,2] and Einstein equations[3,4] that are self-dual (or anti-self-dual). The reasons have varied from the importance of these fields in quantum field theory[5] to their use in a quantized version of general relativity.[6] Though interest in these uses remains, fresh interest in these self-dual fields has developed in the past few years because of the discovery[7] that many of the nonlinear equations of mathematical physics (e.g., the Bogomolny equation, the sine–Gordon equation, the stationary axial symmetric Einstein equations, etc.) turn out to be symmetry reductions of the self-dual Yang–Mills equations for different gauge groups.

It is mainly for this latter reason that we have reexamined one of the important solution generating techniques,[1] which is based on a version of the Riemann–Hilbert (RH) problem.[8] (We know of two RH techniques for solution generation which appear superficially unrelated. The first is based on the use of "seed" solutions,[9] the latter is based on the use of arbitrary characteristic data.[10] It is this latter one we will be concerned with though there is evidence that these two methods are intimately related.)

In particular we wish to describe in detail the relationship of this RH problem with the holonomy operator associated with the connection one-form of the self-dual field. This brings out clearly the geometry of the situation and the role played in the RH problem by the characteristic initial data. We will furthermore show how, from the RH problem, one can easily derive a simple linear differential equation (the Sparling equation) the solution of which solves the RH problem and provides a generating function for the self-dual Yang–Mills field.

In Sec. II we introduce our basic notation for gauge fields, define the closed loop parallel propagator (or holonomy operator), introduce the theory of complex null cones and the associated α and β planes, and finally give a brief discussion of the Riemann–Hilbert problem. In Sec. III we continue the discussion of the holonomy operator now associated with certain particular loops in the anti-self-dual β-planes, and relate this to a RH problem. This RH problem is converted, in Sec. IV, to a linear partial differential equation, which appears to be considerably simpler to solve than the original RH problem. In Sec. V we show how one can formulate self-dual Yang–Mills equations in terms of the holonomy operator associated with certain loops in the (self-dual) α planes as well as how to construct the YM field itself, with no use of potentials, from the holonomy operator. In the conclusion we discuss the prospects for solving these equations for different gauge groups.

II. NOTATION

We consider on Minkowski space M a trivial n-complex-dimensional vector bundle $B = M \times C^n$ associated with some gauge group. A choice of the global vector fields e_A $(A = 1,...,n)$ as a basis set is made. Covariant differentiation is defined by

$$\nabla_a e_A = \gamma_{aA}{}^B e_B \tag{2.1}$$

and parallel transport of vectors $V = V^A e_A$ from point 1 to 2 along a path P is given by

$$V^A(x_1^a) G_A^B(x_1^a, x_2^a, P) = V^B(x_2^a) \tag{2.2}$$

with

$$G = P \exp \int_1^2 \gamma_a \, dx^a , \tag{2.3}$$

P indicating the path-ordered integral and with the matrix indices suppressed. The holonomy operator $h_A^B(x_1, P)$ is a special case of (2.3), involving a closed path P beginning and ending at x_1^a, i.e.,

$$h = P \exp \oint \gamma_a \, dx^a . \tag{2.4}$$

The gauge fields are defined by

$$F_{ab} = \nabla_a \gamma_b - \nabla_b \gamma_a - [\gamma_a, \gamma_b] \tag{2.5}$$

with the dual given by

$$F^*_{ab} = \tfrac{1}{2}\sqrt{-g}\,\epsilon_{abcd} F^{cd}, \tag{2.6}$$

where ϵ_{abcd} is the alternating symbol with $\epsilon_{0123} = -1$.

If

$$F^*_{ab} = iF_{ab} \tag{2.7}$$

or

 0022-2488/86/112797-06$02.50

$$F^*_{ab} = -iF_{ab} \tag{2.8}$$

the fields are, respectively, self-dual or anti-self-dual.

We wish now to study the Minkowski space light cones or more accurately some of the properties of their complexification.

Consider a point x^a and its light cone, the points of which can be described by

$$y^a = x^a + rl^a(\zeta,\bar{\zeta}), \tag{2.9}$$

where $l^a(\zeta,\bar{\zeta})$ is the null tangent vector to the null geodesics on the cone. The complex stereographic coordinates $(\zeta,\bar{\zeta})$ parametrize the sphere of null geodesics while r is an affine parameter along the geodesics. A convenient form for l^a is

$$l^a(\zeta,\bar{\zeta}) = [1/\sqrt{2}(1+\zeta\bar{\zeta})] \times [(1+\zeta\bar{\zeta}),(\zeta+\bar{\zeta}),i(\bar{\zeta}-\zeta),(-1+\zeta\bar{\zeta})]. \tag{2.10}$$

Note the important point that l^a is a real null vector when $\bar{\zeta}$ is the complex conjugate of ζ, but even when $\bar{\zeta}$ is independent of ζ, we still have that $l_a l^a = 0$, but now for complex l^a. Notationally in this case we will replace $\bar{\zeta}$ by $\tilde{\zeta}$.

If we now allow r to be complex and $\bar{\zeta}\to\tilde{\zeta}$ then Eq. (2.9) defines the complex null cone C_x of the point x^a. The points y^a in general live in complex Minkowski space M_c. We would like now to concern ourselves with certain features of these complex cones.

There are two families of two-complex-dimensional submanifolds that span a specific C_x. They are defined from (2.9) as follows.

(i) α planes are the set of points determined by varying both r and $\tilde{\zeta}$ keeping x^a and ζ fixed.

(ii) β planes are determined by varying r and ζ with x^a and $\tilde{\zeta}$ fixed.

It is easy to see that both α and β planes have a linear structure and hence are really planes.

We will refer to specific α or β planes associated with, for example, ζ_1 or $\tilde{\zeta}_2$ by α_{ζ_1} or $\beta_{\tilde{\zeta}_2}$. The intersection of an α_ζ and $\beta_{\tilde{\zeta}}$ plane for the same C_x is simply the generator $(\zeta,\tilde{\zeta})$ of C_x. When we are dealing with a different cone, e.g., C_y, we will use $(\eta,\tilde{\eta})$ to label its generators and α_η and $\beta_{\tilde{\eta}}$ to label its α and β planes.

Both the α and β planes are totally null in the sense that any pair of vectors in either an α or β plane have a zero Minkowski space scalar product. The α and β planes are, respectively, self-dual and anti-self-dual in the sense that any pair of vectors in the α or β plane, when skewed together yield, respectively, self-dual or anti-self-dual tensors. [See (2.7) and (2.8).]

These facts are easily checked by varying $(r,\tilde{\zeta})$ and (r,ζ), respectively, in (2.9) obtaining for the α plane

$$dy^a = dr\, l^a + r\bar{m}^a[d\tilde{\zeta}/(1+\zeta\tilde{\zeta})] \tag{2.11}$$

and for the β plane

$$dy^a = dr\, l^a + rm^a[d\zeta/(1+\zeta\tilde{\zeta})], \tag{2.12}$$

with

$$m^a = \eth l^a \equiv (1+\zeta\tilde{\zeta})\frac{\partial}{\partial\zeta} l^a,$$

$$\bar{m}^a = \bar{\eth} l^a \equiv (1+\zeta\tilde{\zeta})\frac{\partial}{\partial\tilde{\zeta}} l^a, \tag{2.13}$$

$$n^a = l^a + \eth\bar{\eth} l^a \equiv l^a + (1+\zeta\tilde{\zeta})^2\frac{\partial^2}{\partial\zeta\partial\tilde{\zeta}} l^a,$$

where $(l^a,\bar{m}^a)$ and (l^a,m^a) are, respectively, two independent tangent vectors in the α and β planes. One easily checks that $l_a l^a = l^a\bar{m}^a = \bar{m}^a\bar{m}_a = l^a m_a = m^a m_a = 0$, showing that the planes are each totally null. Furthermore, the pair of two-forms $l_{[a}\bar{m}_{b]}$ and $l_{[a}m_{b]}$ are self-dual and anti-self-dual, respectively. (The n^a which has a zero scalar product with all the others except $l^a n_a = ,$ is given here for later use.)

[We point out that the description of the α and β planes in terms of fixed (x^a,ζ) and $(x^a,\tilde{\zeta})$ is highly redundant. Any point $\hat{x}^a$ in the same plane could have been used in the description of the same plane. There, however, does exist an important nonredundant description leading to what is known as twistor theory.[11,12] The set of all α planes in M_c is known as twistor space, while the β planes form dual twistor space. The use of any prior knowledge of twistor theory will be avoided.]

We will conclude this section with a brief discussion of a special case of the RH problem.

First consider a holomorphic function $a(z)$ given on an annular region in the neighborhood of the equator of the Riemann sphere (or extended complex plane, $CU\{\infty\}$) with $a(z)$ having singularities in both the northern and southern hemispheres and $a(z)\neq 0$ in the annular region. The problem then is to "split" $a(z)$ in the annular region such that

$$a(z) = G_N^{-1}(z)G_S(z), \tag{2.14}$$

with G_N and G_S holomorphic in, respectively, the northern and southern hemispheres. The solution is quite simple and in fact is given by

$$G_N(z_N) = \exp\oint_C \frac{\log(a)}{(z-z_N)}\,dz,$$
$$G_S(z_S) = \exp\oint_C \frac{\log(a)}{(z-z_S)}\,dz, \tag{2.15}$$

where the path integral is taken along a curve C in the holomorphic (or annular) region surrounding the singular regions with z_N,z_S being, respectively, points to the north or south of C.

The proof of (2.15) is a simple application of the Cauchy integral theorem applied after taking the logarithm of (2.14).

We are now interested in a generalization of (2.14) to the case where $a(z)$ is a holomorphic matrix valued function and the two "splitting" functions $G_N(z)$ and $G_S(z)$ are also matrix valued. Though theorems for the existence of G_N and G_S are known, there does not appear to be any known method of, *in general* reducing this problem to the form of a simple quadrature as in (2.15). An important special case where this can be done is when the matrix $a(z)$ is in either upper or lower triangular form.

In the next section we will show how the holonomy operator for self-dual fields applied to curves on the β planes (anti-self-dual) leads immediately to the matrix version of (2.14) where the $a(z)$ is the characteristic data for the fields.

III. HOLONOMY AND THE RIEMANN–HILBERT PROBLEM

We continue our discussion of the complexified null cone and in particular study the intersection properties of two separate cones and of their α and β planes.

We first consider a fixed reference or data cone C_Y based on an arbitrary but fixed point Y^a, given by

$$y^a = Y^a + Rl^a(\zeta,\tilde{\zeta}). \tag{3.1}$$

It will be on C_Y that characteristic data will be given. In addition we consider the cone C_x based on an arbitrary real field point x^a in the interior of C_Y given by

$$y'^a = x^a + rl^a(\eta,\tilde{\eta}). \tag{3.2}$$

For the present the only two cones we will be concerned with are C_Y and C_x. Their associated α (and β planes) will be denoted by α_ζ and α_η, respectively (and $\beta_{\tilde{\zeta}}$ and $\beta_{\tilde{\eta}}$). The intersection of C_x and C_Y is determined by

$$y^a = y'^a. \tag{3.3}$$

Since points of C_Y and C_x are determined by $(R,\zeta,\tilde{\zeta})$ and $(r,\eta,\tilde{\eta})$, respectively, (3.3) determines a relationship between the triplets of the form

$$\begin{aligned} &\zeta = \zeta(x^a,Y^a,r,\eta,\tilde{\eta}),\\ &\tilde{\zeta} = \tilde{\zeta}(x^a,Y^a,r,\eta,\tilde{\eta}),\\ &R = R(x^a,Y^a,r,\eta,\tilde{\eta}). \end{aligned} \tag{3.4}$$

Equation (3.4) can in fact be explicitly determined. Defining

$$w^a \equiv x^a - Y^a, \tag{3.5}$$

$$U(w^a,\eta,\tilde{\eta}) \equiv l_a(\eta,\tilde{\eta})w^a, \tag{3.6}$$

$$M(w^a,\eta,\tilde{\eta}) \equiv m_a(\eta,\tilde{\eta})w^a, \tag{3.7}$$

$$\tilde{M}(w^a,\eta,\tilde{\eta}) \equiv \bar{m}_a(\eta,\tilde{\eta})w^a, \tag{3.8}$$

one can by direct calculation find from (3.1), (3.2), and (3.3)

$$\zeta = (\eta M - U)/(M + U\tilde{\eta}), \tag{3.9}$$

$$\tilde{\zeta} = (\tilde{\eta}\tilde{M} - U)/(\tilde{M} + U\eta), \tag{3.10}$$

$$R = U(1 + (M/U)(\tilde{M}/U)). \tag{3.11}$$

We could easily have found $r = r(\eta,\tilde{\eta},w^a)$ but it is not needed. One can go further and show, using (2.10) and (2.13) that (3.9) and (3.10) really have the simple form

$$\zeta = \zeta(w^a,\tilde{\eta}) = -(u + \omega\tilde{\eta})/(\tilde{\omega} + v\tilde{\eta}), \tag{3.12}$$

$$\tilde{\zeta} = \tilde{\zeta}(w^a,\eta) = -(u + \tilde{\omega}\eta)/(\omega + v\eta), \tag{3.13}$$

with $u = w^0 - w^3$, $v = w^0 + w^3$, $\omega = w^1 + iw^2$, and $\tilde{\omega} = w^1 - iw^2$ and inverses

$$\eta = \eta(w^a,\tilde{\zeta}) = -(u + \omega\tilde{\zeta})/(\tilde{\omega} + v\tilde{\zeta}), \tag{3.14}$$

$$\tilde{\eta} = \tilde{\eta}(w^a,\zeta) = -(u + \tilde{\omega}\zeta)/(\omega + v\zeta). \tag{3.15}$$

These relationships are very important. For example, we see immediately that if we are given the two α planes (or two β planes) α_ζ and α_η (or $\beta_{\tilde{\zeta}}$ and $\beta_{\tilde{\eta}}$), they have a unique intersection point given by (3.13), (3.15), and (3.11) [or by (3.12), (3.14), and (3.11)]. Or from (3.12) given the β plane $\beta_{\tilde{\eta}}$, there is a unique α plane α_ζ which it intersects. The intersection is not at a point but along a curve.

Consider the following construction which is basic for what follows.

On C_Y choose two β planes $\beta_{\tilde{\zeta}_1}$ and $\beta_{\tilde{\zeta}_2}$, which are arbitrary but fixed for now. On the cone C_x let $\beta_{\tilde{\eta}}$ determine a variable β plane that intersects $\beta_{\tilde{\zeta}_1}$ and $\beta_{\tilde{\zeta}_2}$ on the generators $(\eta_1,\tilde{\eta})$ and $(\eta_2,\tilde{\eta})$ of C_x, with, from (3.14),

$$\eta_1 = \eta(w^a,\tilde{\zeta}_1), \quad \eta_2 = \eta(w^a,\tilde{\zeta}_2). \tag{3.16}$$

We now form the following triangle, starting at x^a, two legs are the generators $(\eta_1,\tilde{\eta})$ and $(\eta_2,\tilde{\eta})$ ending at p_1 and p_2 on C_Y, with the base B being the unique curve on C_Y determined by (3.12) and (3.11), connecting p_1 and p_2 (see Fig. 1).

We will refer to this triangle as $\Delta(x,\tilde{\eta})$.

If we consider self-dual Yang–Mills fields then parallel propagation around closed curves, as for example $\Delta(x,\tilde{\eta})$, lying in a β plane, is trivial, i.e., the holonomy operator is the identity. The reason for this is that the self-dual F_{ab} when projected onto the anti-self-dual β plane vanishes so that the connection, in the plane, is integrable.

Defining the parallel propagator from the point x^a, along a generator $(\eta,\tilde{\eta})$, to a point p on C_Y by

$$G = G(x^a,\eta,\tilde{\eta}) = P\exp\int_{x^a}^{p}\gamma_a\,dy^a \tag{3.17}$$

and the parallel propagator from p_1 to p_2 by

$$a = a(\tilde{\eta};\eta_1,\eta_2) = P\exp\int_{p_1}^{p_2}\gamma_a\,dy^a, \tag{3.18}$$

we have that the holonomy operator

$$h(x^a,\Delta(x,\tilde{\eta})) = G(x^a,\eta_1,\tilde{\eta})a(\tilde{\eta};\eta_1\eta_2)G^{-1}(x^a,\eta_2,\tilde{\eta}). \tag{3.19}$$

Since $h = I$ for self-dual fields we have

$$G_N^{-1}(\tilde{\eta})G_S(\tilde{\eta}) = a(\tilde{\eta};\eta_1,\eta_2), \tag{3.20}$$

$$G_N = G(x^a,\eta_1;\tilde{\eta}), \quad G_S = G(x^a,\eta_2;\tilde{\eta}). \tag{3.21}$$

Comparison of (3.20) with (2.14) shows that if we identify $\tilde{\eta}$ with z then, at least formally, (3.20) is a RH prob-

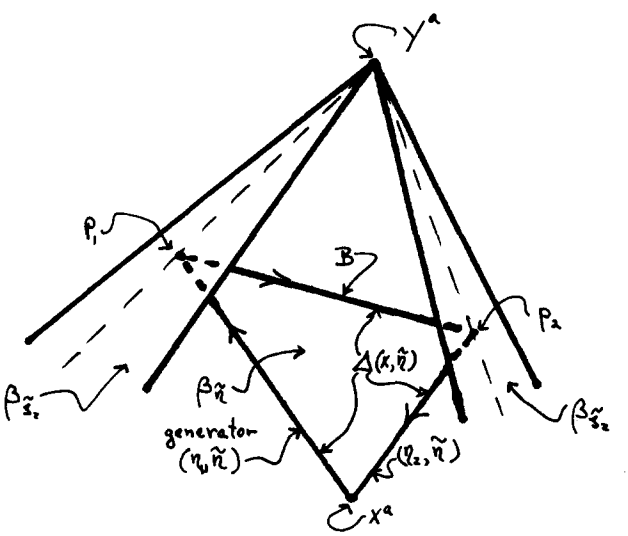


FIG. 1. The β plane $\beta_{\tilde{\eta}}$ through x^a intersecting the two β planes $\beta_{\tilde{\zeta}_1}$ and $\beta_{\tilde{\zeta}_2}$ on C_Y at the points p_1 and p_2 yielding the triangle $\Delta(x,\tilde{\eta})$.

lem for the determination of $G(x^a,\eta,\tilde{\eta})$ for a given $a(\tilde{\eta};\eta_1,\eta_2)$.

The function $a(\tilde{\eta};\eta_1,\eta_2)$ is determined from the (holomorphic) characteristic data, namely the components of the connection on C_Y via (3.18). The values of η_1 and η_2 (or equivalently $\tilde{\zeta}_1$ and $\tilde{\zeta}_2$), along with the choice of characteristic data, determine the singularities of a as a function of $\tilde{\eta}$. We will assume that these singularities occur when $\tilde{\eta}$ is in a neighborhood of both

$$\tilde{\eta} = \tilde{\eta}_S \equiv -1/\eta_1, \tag{3.22}$$

$$\tilde{\eta} = \tilde{\eta}_N \equiv -1/\eta_2. \tag{3.23}$$

This occurs when the YM field is regular on finite regions of C_Y and when Y^a is a sufficiently large timelike displacement from x^a. Geometrically this means that $a(\tilde{\eta};\eta_1,\eta_2)$ is singular when the $\beta_{\tilde{\eta}}$ plane is parallel to either the $\beta_{\tilde{\zeta}_1}$ or $\beta_{\tilde{\zeta}_2}$ plane; the intersection points p_1 and p_2 both being then at ∞. When Y^a tends to timelike infinity Eqs. (3.14) and (3.15) simplify to $\eta = -\tilde{\zeta}^{-1}$ and $\tilde{\eta} = -\zeta^{-1}$. [The assumption (3.23) does not appear to be very restrictive.]

If we consider $\tilde{\eta}_S$ and $\tilde{\eta}_N$ as two points of the Riemann sphere, then in the annular region between them, $a(\tilde{\eta},\eta_1,\eta_2)$ is holomorphic in $\tilde{\eta}$ and we do have in fact a RH problem with

$$G_N = G(x^a,\eta_1;\tilde{\eta}), \quad G_S = G(x^a,\eta_2;\tilde{\eta})$$

as the splitting matrices. Note that since G_N is to be holomorphic in the region far from $\tilde{\eta}_S = -\eta_1^{-1}$, it must be holomorphic in the region antipodal to $\tilde{\eta}_S$, namely in the region $\tilde{\eta} \approx \bar{\eta}_1$. (Two points, on the Riemann sphere, η and η' are antipodal if $\eta' = -\bar{\eta}^{-1}$.) We then have the important result that $G(x^a,\eta,\tilde{\eta})$ is to be holomorphic in *both* variables η and $\tilde{\eta}$ in the region $\tilde{\eta} \approx \bar{\eta}$. This becomes our boundary condition for the differential equation derived from the RH problem in the next section.

We see that we have a variation on the usual RH problem. Instead of having the data being given with *fixed* singularities and having the two "splitting" functions G_N and G_S, we have the situation where the positions of the singularities are variable and determined by the value of η. This leads to a single "splitting" function $G(x^a,\eta,\tilde{\eta})$ holomorphic in the two variables $(\eta,\tilde{\eta})$ in the antiholomorphic strip $\tilde{\eta} \approx \bar{\eta}$. The two usual "splitting" functions G_N and G_S arise by choosing two fixed values for η, i.e., η_1 and η_2.

This new point of view towards the RH problem allows us in the next section to derive a simple linear differential equation for the $G(x^a,\eta,\tilde{\eta})$, which is considerably simpler to study than the original RH problem.

To conclude this section we mention for use in Sec. V that from knowledge of $G(x^a,\eta,\tilde{\eta})$ it is a simple matter to calculate the Yang–Mills connection. This is done by differentiating (3.17) with respect to x^a yielding

$$l^a \nabla_a G G^{-1} = \gamma_a l^a, \tag{3.24}$$

which becomes, after some manipulation[10]

$$\gamma_a = G_{,a} G^{-1} + \eth h l_a - h m_a \tag{3.25}$$

with

$$h = l^b \eth (G_{,b} G^{-1})$$

and

$$\nabla_a G = G_{,a}.$$

One can rewrite Eq. (3.25) in a somewhat simpler form, which will be useful to us later. We note that the left side of (3.25) is only a function of x^a, and therefore we can use different values of η and $\tilde{\eta}$ on the right side. Thus (3.25) can also be written as

$$\gamma_a(x^a) = G'_{,a} G'^{-1} + \eth' h' l'_a - h' m'_a, \tag{3.26}$$

where the prime means, replace $(\eta,\tilde{\eta})$ by $(\eta',\tilde{\eta}')$. By multiplying (3.25) by l^a and m^a, respectively, and (3.26) by l'^a and m'^a we have

$$\begin{aligned} &\gamma_a l^a = l^a G_{,a} G^{-1}, \quad \gamma_a m^a = m^a G_{,a} G^{-1}, \\ &\gamma_a l'^a = l'^a G'_{,a} G'^{-1}, \quad \gamma_a m'^a = m'^a G'_{,a} G'^{-1}. \end{aligned} \tag{3.27}$$

By choosing $(\eta',\tilde{\eta}') = (\eta, -\bar{\tilde{\eta}}^{-1})$, (3.27) can be rewritten as

$$\begin{aligned} &\gamma_a l^a(\eta,\tilde{\eta}) = l^a(\eta,\tilde{\eta}) G_{,a} G^{-1}, \\ &\gamma_a m^a(\eta,\tilde{\eta}) = m^a(\eta,\tilde{\eta}) G_{,a} G^{-1}, \\ &\gamma_a n^a(\eta,\tilde{\eta}) = n^a(\eta,\tilde{\eta}) G'_{,a} G'^{-1}, \\ &\gamma_a \bar{m}^a(\eta,\tilde{\eta}) = \bar{m}^a(\eta,\tilde{\eta}) G'_{,a} G'^{-1}, \end{aligned} \tag{3.28}$$

where $l^a, n^a, m^a, \bar{m}^a$ form a null tetrad system $l^a n_a = -m^a \bar{m}_a = 1$, with all other products vanishing. [This result follows from (3.27) becuase $l'^a(\eta, -\bar{\tilde{\eta}}^{-1})$ and $m'^a(\eta,\bar{\tilde{\eta}}^{-1})$ are linear combinations of $n^a(\eta,\tilde{\eta})$ and $\bar{m}^a(\eta,\tilde{\eta})$.] [See (2.13).]

Equation (3.28) can be solved for γ_a, i.e.,

$$\begin{aligned} \gamma_a = &(l^b G_{,b} G^{-1}) n_a - (m^b G_{,b} G^{-1}) \bar{m}_a \\ &+ (\eta^b \hat{G}_{,b} \hat{G}^{-1}) l_a - (\bar{m}^b \hat{G}_{,b} G^{-1}) m_a, \end{aligned} \tag{3.29}$$

where we have defined $\hat{G} = G(x^a,\eta, -\bar{\tilde{\eta}}^{-1})$.

One can by a gauge transformation make $\hat{G} = I$ in which case we have (in a particular gauge)

$$\gamma_a = (l^b G_{,b} G^{-1}) n_a - (m^a G_{,b} G^{-1}) \bar{m}_a. \tag{3.30}$$

IV. THE RIEMANN–HILBERT PROBLEM AS A DIFFERENTIAL EQUATION

Returning to the RH problem (3.20) written as

$$G^{-1}(x^a,\eta_1,\tilde{\eta}) G(x^a,\eta_2,\tilde{\eta}) = a(\tilde{\eta};\eta_1,\eta_2),$$

we wish to convert it to a differential equation for $G(x^a,\eta,\tilde{\eta})$. Though this can easily be done, it turns out that the resulting equation is simpler if, via (3.12) and (3.13), we use ζ and $\tilde{\zeta}$ as the basic variables and ask for an equation for $G(x^a,\zeta,\tilde{\zeta})$. Equation (3.20) then has the form

$$G^{-1}(x^a,\zeta,\tilde{\zeta}_1) G(x^a,\zeta,\tilde{\zeta}_2) = a(\zeta,\tilde{\zeta}_1,\tilde{\zeta}_2). \tag{4.1}$$

We obtain the differential equation by taking the $\tilde{\zeta}_1$ derivative of (4.1), evaluated at $\tilde{\zeta}_1 = \tilde{\zeta}_2$, i.e.,

$$-G^{-1} G_{,\tilde{\zeta}} G^{-1} G = a_{,\tilde{\zeta}_1 | \tilde{\zeta}_1 = \tilde{\zeta}_2 = \tilde{\zeta}}. \tag{4.2}$$

From (3.18) and (2.11) we have

$$a_{,\tilde{\zeta}_1 | \tilde{\zeta}_1 = \tilde{\zeta}_2 = \tilde{\zeta}} = \gamma_a \frac{dy^a}{d\tilde{\zeta}} = \gamma_a \bar{m}^a(\zeta,\tilde{\zeta}) R (1 + \zeta\tilde{\zeta})^{-1}$$

and thus (4.2) becomes

$$\eth_{\tilde\zeta} G = -GA(R,\zeta,\tilde\zeta) \tag{4.3}$$

with

$$\eth_{\tilde\zeta} G \equiv (1+\zeta\tilde\zeta)G_{,\tilde\zeta}$$

and

$$A = \gamma_a \bar m^a R. \tag{4.4}$$

Equation (4.3), a first-order homogeneous linear partial differential equation well known in the literature[10,13] and referred to as the Sparling equation, is our sought for equation.

One seeks solutions of the Sparling equation $G(x^a,\zeta,\tilde\zeta)$ holomorphic in the strip $\tilde\zeta \approx \bar\zeta$. These solutions solve the original RH problem.

We, however, must say a bit more about $a(\zeta;\tilde\zeta_1,\tilde\zeta_2)$ and $A(R,\zeta,\tilde\zeta)$. Our notation has suppressed the fact that they are both dependent on x^a, the field point. It is, in fact, this dependence which forces G to be also a function of x^a. We first point out that we can always, for self-dual fields, make a choice of gauge such that on the data surface C_Y we have $\gamma_a l^a(\zeta,\tilde\zeta) = \gamma_a m^a(\zeta,\tilde\zeta) = 0$, the only nonvanishing component on C_Y being $\gamma_a \bar m^a$. Though this choice of gauge is not necessary, it does make the exposition simpler and henceforth will be used.

It is easily seen from (3.18) and (4.4) that $A(R,\zeta,\tilde\zeta)$ is the component of the connection on C_Y which is tangent to the curve B connecting p_1 and p_2. Here A, which is our free characteristic data given on C_Y, is to be thought of as an arbitrary (spin-weight -1) function of $R,\zeta,\tilde\zeta$. If we, however, use (3.11) with (3.14) and (3.15), we obtain a relationship

$$R = R(w^a,\zeta,\tilde\zeta), \tag{4.5}$$

i.e., that value of R on the intersection of C_x with C_Y which is to be used in (4.3). We thus have

$$A(R,\zeta,\tilde\zeta) = A\{R(w^a,\zeta,\tilde\zeta),\zeta,\tilde\zeta\} = A(x^a - Y^a,\zeta,\tilde\zeta), \tag{4.6}$$

which is the appropriate form to be used in the Sparling equation.

For completeness we give the form for (4.5), namely

$$R = w^a l_a(-\tilde\zeta^{-1},\tilde\eta) \tag{4.7}$$

with

$$\tilde\eta = -(u+\omega\tilde\zeta)/(\tilde\omega + v\tilde\zeta).$$

In the very important case of C_Y becoming null infinity, i.e., when $Y^a \to$ timelike infinity and $\tilde\eta = -\zeta^{-1}$, $\eta = -\tilde\zeta^{-1}$, Eq. (4.3) is shown easily to be the usual Sparling equation for $G(x^a,\eta,\tilde\eta)$, i.e.,

$$\eth_\eta G = -GA\{x^a l_a(\eta,\tilde\eta),\eta,\tilde\eta\}. \tag{4.8}$$

V. FURTHER COMMENTS ON THE HOLONOMY OPERATOR

We saw that the triviality of the holonomy operator $h = I$ on the β planes led to the RH problem (3.20). A natural question is what use is the holonomy operator (which we will refer to as $\tilde h$) on the α planes.

The closed paths we use on the α_η planes are the direct analogs of the paths we used on the $\beta_{\tilde\eta}$ planes.

From a field point x^a on α_η we construct a triangle $\tilde\Delta(x,\eta)$ with two legs given by the generators $(\eta,\tilde\eta_1)$ and $(\eta,\tilde\eta_2)$ from x^a to p_1' and p_2' on C_Y and the base $\tilde B$ being the unique curve on C_Y connecting p_1' and p_2' on α_η.

The holonomy operator around $\tilde\Delta(x,\eta)$ (starting from a point on C_Y) is simply given by

$$\tilde h(x,\eta;\tilde\eta_1\tilde\eta_2) = G^{-1}(x^a,\eta,\tilde\eta_1)G(x^a,\eta_1\tilde\eta_2), \tag{5.1}$$

which, in general, is nontrivial. Note that there is no contribution to $\tilde h$ from the curve $\tilde B$ since by our choice of gauge $\gamma_a m^a(\zeta,\tilde\zeta) = 0$. Note further that $\tilde h$ is gauge invariant, i.e., from $G \to g(x)G$, we have $\tilde h \to \tilde h$.

$\tilde h(x,\eta;\tilde\eta_1,\tilde\eta_2)$ is an interesting and important variable. Directly from $\tilde h$ alone one can calculate the Yang–Mills field. Furthermore $\tilde h$ satisfies equations equivalent to the self-dual YM equations. In other words, the nonlocal holonomy operator $\tilde h$ contains all the information about the gauge field.

One can look at $\tilde h$ in two ways; we can first consider it as a function of the seven variables $(x^a,\eta,\tilde\eta_1,\tilde\eta_2)$ and second, for some *fixed* values of $(\eta,\tilde\eta_1,\tilde\eta_2)$, consider it simply as a function of x^a alone. From the first point of view one can derive a generalized RH problem and its "localized" differential equation analogous to the Sparling equation. The second point of view leads to a simple second-order nonlinear differential equation for $\tilde h(x^a)$ obtained in a different context by Yang,[14] which replaces the usual self-dual equations.

To elucidate these claims we begin by writing the RH problem for two β planes, $\beta_{\tilde\eta_1}$ and $\beta_{\tilde\eta_2}$:

$$\begin{aligned} a(\tilde\eta_1;\eta_1\eta_2) &= G^{-1}(x,\eta_1\tilde\eta_1)G(x,\eta_2,\tilde\eta_1), \\ a(\tilde\eta_2;\eta_1\eta_2) &= G^{-1}(x,\eta_1\tilde\eta_2)G(x,\eta_2,\tilde\eta_2), \end{aligned} \tag{5.2}$$

and the holonomy operator $\tilde h$ on two α planes α_{η_1} and α_{η_2}, i.e.,

$$\begin{aligned} \tilde h(x,\eta_1;\tilde\eta_1,\tilde\eta_2) &= G^{-1}(x,\eta_1,\tilde\eta_1)G(x,\eta_1,\tilde\eta_2), \\ h(x,\eta_2;\tilde\eta_1,\tilde\eta_2) &= G^{-1}(x,\eta_2,\tilde\eta_1)G(x,\eta_2,\tilde\eta_2). \end{aligned} \tag{5.3}$$

By inspection of (5.2) and (5.3) we see that G can be eliminated yielding an expression involving on $\tilde h$ and a, i.e.,

$$a(\tilde\eta_1;\eta_1,\eta_2)\tilde h(\eta_1;\tilde\eta_1,\tilde\eta_2) = \tilde h(\eta_2;\tilde\eta_1,\tilde\eta_2)a(\tilde\eta_2,\eta_1\eta_2), \tag{5.4}$$

which can be thought of as a generalized RH problem in two variables $\tilde\eta_1$ and $\tilde\eta_2$, given $a(\tilde\eta,\eta_1,\eta_2)$. By differentiating (5.4) with respect to η_2 and $\tilde\eta_2$ at $\eta_2 = \eta_1$ and $\tilde\eta_2 = \tilde\eta_1$ one obtains the differential version of (5.4), namely

$$\eth\tilde H + [\tilde H, A] + \eth A = 0 \tag{5.5}$$

with the infinitesimal holonomy operator $\tilde H$ given by

$$\tilde H = (1+\eta_1\tilde\eta_1)\left.\frac{\partial\tilde h}{\partial\tilde\eta_2}\right|_{\tilde\eta_2=\tilde\eta_1}.$$

[For simplicity here we have tacitly assumed that C_Y is null infinity as in (4.8).] As in the case of the Sparling equation versus the RH problem, (3.20), it is easier to study (5.5), a linear inhomogeneous differential equation, than (5.4).

For our final result we derive Yang's equation on $\tilde h$ thought of as a function only of x^a. We fix, in (5.1), η, $\tilde\eta_1$, and $\tilde\eta_2 = -\tilde\eta_1^{-1}$ and write (5.1) as

$$\tilde h(x^a) = G^{-1}\hat G. \tag{5.1'}$$

In a straightforward fashion we compute

$$f_{ab} \equiv (\tilde{h}^{-1}\tilde{h}_{,a})_{,b} \tag{5.6}$$

using

$$(G_{,a}G^{-1})_{,b} = (G_{,b}G^{-1})_{,a} - [G_aG^{-1}, G_bG^{-1}] \tag{5.7}$$

obtaining

$$f_{ab} = \hat{G}^{-1}\{(\hat{G}_{,b}\hat{G}^{-1})_{,a} - (G_{,a}G^{-1})_{,b} + [\hat{G}_{,b}\hat{G}^{-1}, G_{,a}G^{-1}]\}\hat{G}. \tag{5.8}$$

Though f_{ab} is *not* skew on a and b and is *not* the Yang–Mills field, it *does* contain all the Yang–Mills field information as we now demonstrate.

Using either (3.29) or (3.30) one can show that (in a particular gauge) the three nonvanishing self-dual components of the YM field are

$$\begin{aligned} F_{ab}l^a\bar{m}^b &= -f_{ab}l^a\bar{m}^b, \\ F_{ab}(l^an^b + m^a\bar{m}^b) &= -f_{ab}(l^an^b + m^a\bar{m}^b), \\ F_{ab}\eta^am^b &= -f_{ab}m^an^b. \end{aligned} \tag{5.9}$$

(Note that on the right side the placement of the indices are important.)

Furthermore one can calculate

$$F_{ab}(l^an^b - m^a\bar{m}^b) = -f_{ab}(l^an^b - m^a\bar{m}^b), \tag{5.10}$$

which must vanish by virtue of the fact that $l^{[a}n^{b]} - m^{[a}\bar{m}^{b]}$ is anti-self-dual and F_{ab} is self-dual. This leads to the simple differential equation for $\tilde{h}$

$$(\tilde{h}^{-1}\tilde{h}_{,u})_{,v} - (\tilde{h}^{-1}\tilde{h}_{,w})_{\bar{w}} = 0, \tag{5.11}$$

where we have introduced coordinates $u,v,w,\bar{w}$ by

$$\begin{aligned} l^a\frac{\partial}{\partial x^a} &= \frac{\partial}{\partial u}, \quad n^a\frac{\partial}{\partial x^a} = \frac{\partial}{\partial v}, \\ m^a\frac{\partial}{\partial x^a} &= \frac{\partial}{\partial w}, \quad \bar{m}^a\frac{\partial}{\partial x^a} = \frac{\partial}{\partial \bar{w}}. \end{aligned} \tag{5.12}$$

Equation (5.11) is the Lorentzian version of an equation derived by Yang[14] for Euclidean self-dual fields.

VI. CONCLUSION

In a future paper we will show how, for the gauge groups SL(2,C) or SU(2), it is possible to solve in closed form, i.e., by quadratures, the RH problem (3.20) or the related Sparling equation (4.3). This means, at least in principle, one can calculate by quadratures any of the symmetry reductions of the SL(2,C) self-dual equations and in particular the stationary axial symmetric solutions of the Einstein equations. It appears likely, but at the moment unproved, that our method of integrating the SL(2,C) RH problem or Sparling equation can be extended to other groups.

ACKNOWLEDGMENTS

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] R. S. Ward, Phys. Lett. A **61**, 81 (1977).
[2] M. F. Atiyah, *Classical Geometry of Yang–Mills Fields*, Fermi Lectures (Scuola Normale, Pisa, 1980).
[3] R. O. Hansen, E. T. Newman, R. Penrose, and K. P. Tod, Proc. R. Soc. London Ser. A **363**, 445 (1978).
[4] M. Ko, M. Ludvigsen, E. T. Newman, and K. P. Tod, Phys. Rep. **71**, 51 (1981).
[5] A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Yu. S. Tyupkin, Phys. Lett. B **59**, 85 (1975).
[6] R. Penrose, Gen. Relativ. Gravit. **7**, 31 (1976).
[7] R. Ward, "Integrable and solvable systems and relations among them," University of Durham, England, preprint.
[8] I. Hauser, *On the Homogeneous Hilbert Problem, Lecture Notes in Physics*, Vol. 205, edited by C. Hoenselaers and W. Dietz (Springer, Berlin, 1984).
[9] See the previous paper and the papers by F. J. Ernst and D. S. Guo in the same volume.
[10] E. T. Newman, Phys. Rev. D **22**, 3023 (1980).
[11] R. Penrose and W. Rindler, *Spinors and Space-Time* (Cambridge U.P., Cambridge, 1986), Vol. II.
[12] S. Huggett and K. P. Tod, *An Introduction to Twistor Theory*, London Mathematical Society Student Texts 4 (Cambridge U.P., Cambridge, 1985).
[13] C. Kozameh and E. T. Newman, Found. Phys. **15**, 487, 1985.
[14] C. N. Yang, Phys. Rev. Lett. **38**, 1377 (1977).

A gauge symmetric approach to Fierz identities

Frank Reifler and Randall Morris
RCA/Government and Commercial Systems, Moorestown, New Jersey 08057

(Received 6 March 1986; accepted for publication 2 July 1986)

Seventy-five years ago Cartan invented spinors by mapping C^2 onto isotropic (null) vectors in C^3. In recent work this map was extended and it was shown that bispinors are isomorphic to a class of Yang–Mills vector triplets $\mathbf{F}_k = \mathbf{E}_k + i\mathbf{H}_k$ which satisfy the following $SU(2)\times U(1)$ gauge invariant constraint: $\mathbf{F}_j \cdot \mathbf{F}_k = s^2\,\delta_{jk}$, where $s^2 = \frac{1}{3}\mathbf{F}_k \cdot \mathbf{F}_k$ (k summed from 1 to 3). Thus bispinors have inherent $SU(2)\times U(1)$ gauge symmetry. In this paper it is shown, using the extended Cartan map and the gauge symmetry of the constrained Yang–Mills fields, that all the Fierz identities reduce to a single equation. Moreover, this equation includes not only the 75 identities recently derived by Takahashi [Y. Takahashi, J. Math. Phys. **24**, 1783 (1983)] but an additional 75 which come from interchanging gauge and vector components. It is further shown that the Fierz identities for bispinors can be generalized to any multiplet, $\Psi\in C^{2^n}$, consisting of 2^{n-1} spinors ($n = 1$ for spinors, $n = 2$ for bispinors, $n = 3$ for bispinor doublets, etc.). The generalized identities can also be used to show that the 2^{n-1} spinor multiplets are isomorphic to multiplets of constrained Yang–Mills vector fields.

I. INTRODUCTION

A bispinor $\tilde{\Psi} = (\xi,\eta^*)\in C^4$ consists of a spinor ξ and a conjugated spinor η^*. We have shown in recent papers[1–5] that bispinors could be mapped isomorphically onto spinor pairs $\Psi = (\xi,\eta)\in C^4$ or equivalently onto a set of Yang–Mills vector triplets (see Fig. 1),

$$\mathbf{F}_k = \mathbf{E}_k + i\mathbf{H}_k \quad (k = 1,2,3 \text{ and } \mathbf{E}_k,\mathbf{H}_k\in R^3), \tag{1.1}$$

which satisfy the following constraint:

$$\mathbf{F}_j \cdot \mathbf{F}_k = s^2\delta_{jk}, \tag{1.2}$$

where

$$s^2 = \tfrac{1}{3}\mathbf{F}_k \cdot \mathbf{F}_k. \tag{1.3}$$

The isomorphism between the bispinors and the Yang–Mills vector triplets satisfying (1.2), revealed that the bispinors have an *inherent* $SU(2)\times U(1)$ gauge symmetry. The gauge group $SU(2)$ acts on the Yang–Mills triplets $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$ by formal rotations of the $\mathbf{F}_k$. However, the equivalent action on the bispinors is noncomplex linear, which obfuscates their inherent gauge symmetry.[6] In this paper we use the gauge symmetry of the Yang–Mills vector triplets to simplify the Fierz bispinor identities recently derived by Takahashi.[7]

The map from bispinors $\tilde{\Psi}$ to constrained Yang–Mills vector triplets $\mathbf{F}_k$ we called the Cartan map.[8] We defined the Cartan map to be the bilinear map B^α_K from $C^4\times C^4$ into $C^4\otimes C^4$ given as follows (see Sec. II):

$$B^\alpha_K(\Psi,\chi) = \overline{\Psi}^*\tau_K\sigma^\alpha\chi, \tag{1.4}$$

where

$\alpha = 0, 1, 2, 3$ = Lorentz indices,
$K = 0, 1, 2, 3$ = gauge indices,
Ψ, χ = spinor pairs,
Ψ^* = conjugate of Ψ (see Sec. II),
σ^α = Pauli spin matrices,
τ_K = gauge matrices.

For each spinor pair $\Psi\in C^4$, we define

$$F^\alpha_K = iB^\alpha_K(\Psi,\Psi). \tag{1.5}$$

Writing $F^\alpha_K = (F^0_K,\mathbf{F}_K)$, we define $\mathbf{F}_k$ with $k = 1, 2, 3$ as the Yang–Mills vector triplet and $s = F^0_0$ as the scalar corresponding to Ψ. ($\mathbf{F}_0$ and F^0_k identically vanish.)

The complete set of 15 Fierz identities derived by Takahashi[9] can be combined into a *single* gauge symmetric formula as follows:

$$B^\alpha(\Psi,\chi)\bullet B^\beta(\chi',\Psi') = C^{\alpha\beta}_{\gamma\delta}B^\gamma_0(\Psi,\Psi')B^\delta(\chi',\chi), \tag{1.6}$$

where $C^{\alpha\beta}_{\gamma\delta}$ is the *tensor*[10] whose covariant components are given by

$$C_{\alpha\beta\gamma\delta} = g_{\alpha\gamma}g_{\beta\delta} + g_{\alpha\delta}g_{\beta\gamma} - g_{\alpha\beta}g_{\gamma\delta} - i\epsilon_{\alpha\beta\gamma\delta}, \tag{1.7}$$

where

$g_{\alpha\beta}$ = Lorentz metric tensor,
$\epsilon_{\alpha\beta\gamma\delta}$ = permutation tensor,
Ψ, Ψ', χ, χ' = spinor pairs.

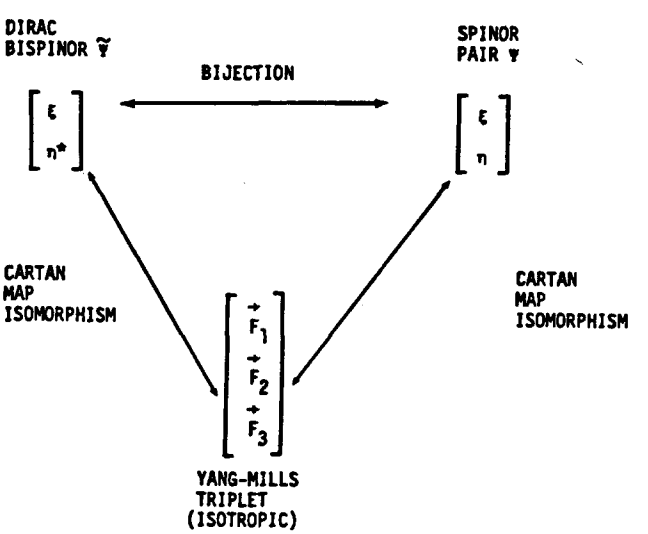


FIG. 1. Isomorphisms.

 0022-2488/86/112803-04$02.50

The operation • denotes the formal Clifford algebra multiplication defined on formal quadruplets ϕ_K, μ_K, and λ_K $(K=0,1,2,3)$ as follows for $\lambda=\phi\bullet\mu$:

$$\begin{aligned}\lambda_0&=\phi_K\,\mu_K,\\ \lambda_k&=\phi_0\mu_k+\phi_k\,\mu_0+i\epsilon_{kpq}\phi_p\,\mu_q\end{aligned}\tag{1.8}$$

(see Sec. II).

Since the Lorentz indices α and the gauge indices K are symmetric in formula (1.4), interchanging them in (1.6) leads to a second gauge symmetric formula given as follows:

$$B_J(\Psi,\chi)\circ B_K(\chi',\Psi')=C^{PQ}_{JK}B^0_P(\Psi,\Psi')B_Q(\chi',\chi),\tag{1.9}$$

where C^{PQ}_{JK} is the same tensor[11] defined in (1.7), i.e.,

$$C^{JKPQ}=g^{JP}g^{KQ}+g^{JQ}g^{KP}-g^{JK}g^{PQ}-i\epsilon^{JKPQ},\tag{1.10}$$

and the operation $\circ$ denotes Clifford algebra multiplication with respect to the Lorentz indices (that is, for four-vectors: p^α, q^α, $r^\alpha\in C^4$ such that $r=p\circ q$, then

$$\begin{aligned}r^0&=p^0q^0+\mathbf{p}\cdot\mathbf{q},\\ \mathbf{r}&=p^0\mathbf{q}+q^0\mathbf{p}+i\mathbf{p}\times\mathbf{q})\end{aligned}\tag{1.11}$$

(see Sec. II).

We first derive formulas (1.6) and (1.9), and then show that the 75 nongauge symmetric identities relating bispinor observables obtained by Takahashi[12] are specific cases of formula (1.6). We further show that these identities, which pertain to a single bispinor, can also be reduced to a *single* gauge symmetric identity.

The Fierz identities (1.6) and (1.9) can be easily generalized to *any* multiplet,

$$\Psi\in C^{2^n}\tag{1.12}$$

$(n=1,2,3,\ldots)$ consisting of 2^{n-1} spinors. The case $n=1$ (spinors) is discussed in Sec. II. Fierz identities (1.6) and (1.9) are for the case $n=2$ (spinor pairs). Special cases of the Fierz identities for $n=3$ (spinor pair doublets) were used in the $SU(2)\times SU(2)\times U(1)$ model of color-electroweak interactions.[13]

To facilitate extending the identities to any n, we consider first the case $n=3$. The Cartan map for $n=3$ is defined to be the bilinear map B^α_{JK} from $C^8\times C^8$ into $C^4\otimes C^4\otimes C^4$ given as follows[14]:

$$B^\alpha_{JK}(\Psi,\chi)=-\bar\Psi^*\tau_J t_K\sigma^\alpha\chi,\tag{1.13}$$

where $\Psi,\chi\in C^8$ and

$\alpha=0,1,2,3=$ Lorentz indices,
$J,K=0,1,2,3=$ gauge indices,
$\Psi^*=$ conjugate of Ψ,
$\sigma^\alpha=$ Pauli spin matrices,
$\tau_K,t_K=$ gauge matrices.

Then for $n=3$ formula (1.9) becomes the following Fierz identity using the formal tensor C^{JKPQ} defined in formula (1.10):

$$\begin{aligned}&B_{JM}(\Psi,\chi)\circ B_{KN}(\chi',\Psi')\\ &\quad=\tfrac{1}{2}C^{PQ}_{JK}C^{RS}_{MN}B^0_{PR}(\Psi,\Psi')B_{QS}(\chi',\chi).\end{aligned}\tag{1.14}$$

Since the Lorentz indices α and the gauge indices J and K are symmetric in formula (1.13), interchanging them leads to further identities.

As an application of formula (1.14) for $n=3$, let $\Psi\in C^8$. Define

$$F^\alpha_{JM}=B^\alpha_{JM}(\Psi,\Psi).\tag{1.15}$$

Writing $F^\alpha_{JM}=(F^0_{JM},\mathbf{F}_{JM})$ we obtain ten Yang–Mills vectors $\mathbf{F}_{00}$ and $\mathbf{F}_{jm}$, and six scalars $s_j=F^0_{j0}$ and $r_m=F^0_{0m}$. The remaining components of F^α_{JM} vanish. From (1.14) and (1.15) we get the following formula for the scalar invariants of the nonet $\mathbf{F}_{jm}$:

$$\begin{aligned}\mathbf{F}_{jm}\cdot\mathbf{F}_{kn}=&-\rho^2\delta_{jk}\,\delta_{mn}+s_js_k\,\delta_{mn}+r_mr_n\,\delta_{jk}\\ &+(\epsilon_{jkp}s_p)(\epsilon_{mnq}r_q),\end{aligned}\tag{1.16}$$

where $\rho^2=s_ks_k=r_kr_k$.

Using the method discussed in Sec. II, the generalization of (1.14) to the cases $n>3$ is a straightforward induction on the products of the formal tensor C^{PQ}_{JK}, which gives the following formula for $n>1$:

$$\begin{aligned}&B_{J_1\cdots J_{n-1}}(\Psi,\chi)\circ B_{K_1\cdots K_{n-1}}(\chi',\Psi')\\ &\quad=(1/2^{n-2})\left[C^{P_1Q_1}_{J_1K_1}\cdots C^{P_{n-1}Q_{n-1}}_{J_{n-1}K_{n-1}}\right]\\ &\qquad\times B^0_{P_1\cdots P_{n-1}}(\Psi,\Psi')B_{Q_1\cdots Q_{n-1}}(\chi',\chi),\end{aligned}\tag{1.17}$$

where $\Psi,\chi,\Psi',\chi'\in C^{2^n}$ and the Cartan map $B^\alpha_{K_1\cdots K_{n-1}}$ from $C^{2^n}\times C^{2^n}$ into $C^4\otimes\cdots\otimes C^4$ (n factors) is defined by

$$B_{K_1\cdots K_{n-1}}(\Psi,\chi)=(-1)^n\bar\Psi^*\tau_{K_1}{}^{(1)}\cdots\tau_{K_{n-1}}{}^{(n-1)}\sigma^\alpha\chi,\tag{1.18}$$

where the $\tau_{K_1}{}^{(1)},\ldots,\tau_{K_{n-1}}{}^{(n-1)}$ are the gauge matrices.

The generalization (1.17) also leads to the following isomorphisms between 2^{n-1} spinor multiplets Ψ and constrained Yang–Mills vector fields:

$$\mathbf{F}_{k_1\cdots k_{n-1}}=\mathbf{B}_{k_1\cdots k_{n-1}}(\Psi,\Psi).\tag{1.19}$$

The constraints may be derived from (1.17) in the same manner that (1.16) was derived from (1.14).

For every n, the Yang–Mills vector fields $\mathbf{F}_{k_1\cdots k_{n-1}}$ satisfy a vector Dirac equation[15] derived from the corresponding spinor (multiplet) equation using the Fierz identity (1.17). Furthermore, the vector Dirac equations are more general than their spinor versions. Indeed, spinor multiplet fields cannot be defined generally for space-time manifolds[16]; whereas the vector fields $\mathbf{F}_{k_1\cdots k_{n-1}}$ are defined on *all* space-time manifolds. Thus, it is the Fierz identity (1.17) that allows us to pass from spinor multiplets (which do not have curvilinear components) to vectors which are fully covariant with respect to all coordinate transformations.

II. FIERZ IDENTITIES

By definition a spinor is a two-dimensional complex vector; i.e., an element of C^2. Let

$$\xi=\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}\in C^2$$

denote a spinor. The conjugate spinor ξ^* is defined as

$$\xi^*=\begin{bmatrix}\bar\xi_2\\ -\bar\xi_1\end{bmatrix},\tag{2.1}$$

where the bar denotes ordinary complex conjugation. The

map $\xi \to \xi^*$ is called spinor conjugation. Since

$$\xi^{**} = -\xi$$

spinor conjugation is a bijection.

The Pauli matrices, denoted σ^α, where $\alpha = 0, 1, 2, 3$, are the 2×2 Hermitian matrices acting on spinors $\xi \in C^2$ defined by

$$\sigma^0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma^1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \sigma^2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma^3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{2.2}$$

We often use the following notation:

$$\boldsymbol{\sigma} = (\sigma^1, \sigma^2, \sigma^3), \quad \sigma^\alpha = (\sigma^0, \boldsymbol{\sigma}).$$

Let $p^\alpha \in C^4$ have the components

$$p^\alpha = (p^0, \mathbf{p}), \quad \mathbf{p} = (p^1, p^2, p^3).$$

Then by definition (2.2)

$$p^\alpha \sigma^\alpha = \begin{bmatrix} p^0 + p^3 & p^1 - ip^2 \\ p^1 + ip^2 & p^0 - p^3 \end{bmatrix}. \tag{2.3}$$

Then 2×2 matrices (2.3) satisfy the following rule for multiplication. Let $p^\alpha, q^\alpha \in C^4$, then

$$(p^\alpha \sigma^\alpha)\ (q^\beta \sigma^\beta) = r^\gamma \sigma^\gamma, \tag{2.4}$$

where $r^\gamma \in C^4$ has the components

$$r^0 = p^\alpha q^\alpha, \quad \mathbf{r} = p^0 \mathbf{q} + q^0 \mathbf{p} + i\mathbf{p} \times \mathbf{q}. \tag{2.5}$$

Writing $r = p \circ q$, formula (2.5) defines an associative multiplication on C^4, which makes C^4 a Clifford algebra.

In order to obtain Lorentz covariant expressions we must consider products of the form $p \circ \hat{q}$, where

$$\hat{q}^\alpha = (q^0, -\mathbf{q}) = q_\alpha. \tag{2.6}$$

Substituting (2.6) into (2.5) we get

$$[p \circ \hat{q}]^0 = p^\alpha q_\alpha, \quad [p \circ \hat{q}]^{\rightarrow} = q^0 \mathbf{p} - p^0 \mathbf{q} - i\mathbf{p} \times \mathbf{q}. \tag{2.7}$$

That is, if $p^\alpha, q^\alpha \in C^4$ transform as Lorentz four-vectors, and if we denote

$$s = [p \circ \hat{q}]^0, \quad \mathbf{F} = [p \circ \hat{q}]^{\rightarrow},$$

then s is a Lorentz scalar and $\mathbf{F} = \mathbf{E} + i\mathbf{H}$ is a Yang–Mills vector.

The Cartan map is defined to be a bilinear map from $C^2 \times C^2$ into C^4 as follows:

$$b^\alpha(\xi, \eta) = \begin{bmatrix} \xi_1 \eta_2 - \xi_2 \eta_1 \\ \xi_1 \eta_1 - \xi_2 \eta_2 \\ i(\xi_1 \eta_1 + \xi_2 \eta_2) \\ -\xi_1 \eta_2 - \xi_2 \eta_1 \end{bmatrix}, \tag{2.8}$$

where $\xi, \eta \in C^2$ are spinors. From (2.1) and (2.2), formula (2.8) may be written as

$$b^\alpha(\xi, \eta) = -\bar{\xi}^* \sigma^\alpha \eta. \tag{2.9}$$

From (2.8), the Cartan map b: $C^2 \times C^2 \to C^4$ has the symmetry property

$$b(\xi, \eta) = -\hat{b}(\eta, \xi) \tag{2.10}$$

and from (2.9), the following conjugation properties:

$$\overline{b(\xi, \eta)} = \hat{b}(\xi^*, \eta^*), \quad \overline{b(\xi^*, \eta)} = -\hat{b}(\xi, \eta^*). \tag{2.11}$$

Moreover, from (2.4), (2.5), and (2.9), the Cartan map commutes with the matrices (2.3) as follows:

$$b(\xi, p^\alpha \sigma^\alpha \eta) = p \circ b(\xi, \eta). \tag{2.12}$$

In addition, from (2.8) a straightforward derivation shows that the Cartan map satisfies the following identity for all $\xi, \eta, \xi', \eta' \in C^2$:

$$b(\xi, \eta) \circ b(\eta', \xi') = 2b^0(\xi, \xi') b(\eta', \eta). \tag{2.13}$$

Formula (2.13) shows that the image of the Cartan map is closed under the operation of Clifford algebra multiplication. Equivalently, using (2.10), we have

$$b(\xi, \eta) \circ \hat{b}(\xi', \eta') = 2b^0(\xi, \xi') \hat{b}(\eta, \eta'). \tag{2.14}$$

From (2.14) we derive Lorentz covariant equations. Using (2.7) and denoting $b^\alpha = (b^0, \mathbf{b})$ we can expand formula (2.14) in component form as follows

$$\begin{aligned} 2b^0(\xi, \xi') b^0(\eta, \eta') &= b^\alpha(\xi, \eta) b_\alpha(\xi', \eta'), \\ 2b^0(\xi, \xi') \mathbf{b}(\eta, \eta') &= b^0(\xi, \eta) \mathbf{b}(\xi', \eta') \\ &\quad - \mathbf{b}(\xi, \eta) b^0(\xi', \eta') \\ &\quad + i\mathbf{b}(\xi, \eta) \times \mathbf{b}(\xi', \eta'). \end{aligned} \tag{2.15}$$

Formula (2.14) gives us all possible Fierz identities for *spinors*. Fierz identities for *bispinors* will now be derived as a straightforward generalization of (2.14).

A bispinor $\tilde{\Psi} = (\xi, \eta^*) \in C^4$ consists of a spinor $\xi \in C^2$ and a conjugated spinor $\eta^* \in C^2$. Associated with the bispinor $\tilde{\Psi} = (\xi, \eta^*)$ is the spinor pair $\Psi = (\xi, \eta)$, where $\eta = -(\eta^*)^*$, and the conjugate spinor pair $\Psi^* = (\eta^*, -\xi^*)$. Note that $\Psi = \Psi^{**}$. The map $\tilde{\Psi} \to \Psi$ and $\Psi \to \Psi^*$ are bijections. The Cartan map (2.8) can be extended to spinor pairs (bispinors) as follows: For each $K = 0, 1, 2, 3$ the Cartan map is defined to be the bilinear map from $C^4 \times C^4$ into C^4 given by[17]

$$B^\alpha_K(\Psi, \chi) = \begin{bmatrix} b^\alpha(\xi, \nu) - b^\alpha(\eta, \kappa) \\ b^\alpha(\xi, \kappa) - b^\alpha(\eta, \nu) \\ i[b^\alpha(\xi, \kappa) + b^\alpha(\eta, \nu)] \\ -b^\alpha(\xi, \nu) - b^\alpha(\eta, \kappa) \end{bmatrix}, \tag{2.16}$$

where $\Psi = (\xi, \eta)$ and $\chi = (\kappa, \nu)$ denote spinor pairs and b^α is defined by (2.8). Note that the extended Cartan map (2.16) can be rearranged as follows:

$$\begin{aligned} b^\alpha(\xi, \kappa) &= -(i/2)(B^\alpha_2 + iB^\alpha_1), \\ b^\alpha(\eta, \nu) &= -(i/2)(B^\alpha_2 - iB^\alpha_1), \\ b^\alpha(\xi, \nu) &= -\tfrac{1}{2}(B^\alpha_3 - B^\alpha_0), \\ b^\alpha(\eta, \kappa) &= -\tfrac{1}{2}(B^\alpha_3 + B^\alpha_0), \end{aligned} \tag{2.17}$$

where we have denoted

$$B^\alpha_K = B^\alpha_K(\Psi, \chi).$$

The $SU(2) \times U(1)$ gauge generators, denoted τ_K, where $K = 0, 1, 2, 3$, are the 4×4 Hermitian matrices acting on spinor pairs $\Psi = (\xi, \eta) \in C^4$ defined by $\tau_0 =$ identity matrix,

$$\tau_1 = \begin{bmatrix} 0 & I_2 \\ I_2 & 0 \end{bmatrix}, \quad \tau_2 = \begin{bmatrix} 0 & -iI_2 \\ iI_2 & 0 \end{bmatrix}, \qquad \tau_3 = \begin{bmatrix} I_2 & 0 \\ 0 & -I_2 \end{bmatrix}, \tag{2.18}$$

where I_2 is the identity of C^2. We often use the following notation:

$$\tau_k = (\tau_1,\tau_2,\tau_3), \quad \tau_K = (\tau_0,\tau_k),$$

with $k = 1, 2, 3$. Using the gauge generators τ_K, formula (2.16) may be written as

$$B^{\alpha}_{K}(\Psi,\chi) = \overline{\Psi}^{*}\tau_K\sigma^{\alpha}\chi. \tag{2.19}$$

The $B^{\alpha}_{K} = (B^{0}_{K}, \mathbf{B}_K)$ have the following symmetries:

$$\begin{aligned} &B^{0}_{0}(\Psi,\chi) = B^{0}_{0}(\chi,\Psi), \quad \mathbf{B}_0(\Psi,\chi) = -\mathbf{B}_0(\chi,\Psi), \\ &B^{0}_{k}(\Psi,\chi) = -B^{0}_{k}(\chi,\Psi), \quad \mathbf{B}_k(\Psi,\chi) = \mathbf{B}_k(\chi,\Psi), \end{aligned} \tag{2.20}$$

with $k = 1, 2, 3$. It follows from (2.20) that $B^{0}_{k}(\Psi,\Psi)$ and $\mathbf{B}_0(\Psi,\Psi)$ vanish. Similar to (2.10), we define

$$\hat{B}(\Psi,\chi) = B(\chi,\Psi). \tag{2.21}$$

Then the B^{α}_{K} satisfy the following conjugation properties:

$$\overline{B(\Psi,\chi)} = \hat{B}(\Psi^{*},\chi^{*}), \quad \overline{B(\Psi^{*},\chi)} = \hat{B}(\Psi,\chi^{*}). \tag{2.22}$$

From (2.13), (2.16), and (2.17), we derive the following Fierz identity for spinor pairs (bispinors):

$$B_J(\Psi,\chi)\circ B_K(\chi',\Psi') = C^{PQ}_{JK}B^{0}_{P}(\Psi,\Psi')B_Q(\chi',\chi), \tag{2.23}$$

with

$$C^{JKPQ} = g^{JP}g^{KQ} + g^{JQ}g^{KP} - g^{JK}g^{PQ} - i\epsilon^{JKPQ}, \tag{2.24}$$

where[18]

$J, K, P, Q = 0, 1, 2, 3,$

g^{JK} = formal Lorentz metric tensor,

ϵ^{JKPQ} = formal permutation tensor,

Ψ, Ψ', χ, χ' = spinor pairs.

[We may substitute $\hat{B}$ for B in the Fierz identity (2.23) to obtain Lorentz covariant equations similar to (2.15).]

We may also define a Clifford algebra product with respect to the gauge indices K, rather than using the Lorentz indices α as in formula (2.5). Thus, consider quadruplets: ϕ_K, μ_K, and λ_K, with $K = 0, 1, 2, 3$, then $\lambda = \phi\bullet\mu$ is defined by

$$\lambda_0 = \phi_K\,\mu_K, \quad \lambda_k = \phi_0\mu_k + \phi_k\,\mu_0 + i\epsilon_{kpq}\phi_p\,\mu_q. \tag{2.25}$$

This associative multiplication makes the quadruplets a formal Clifford algebra. [In order to obtain formal Lorentz expressions we use products of the form $\phi\bullet\hat{\mu}$, where $\hat{\mu} = (\mu_0, -\mu_k)$ similar to formula (2.6).]

Using the symmetry between the Lorentz and gauge indices in formula (2.19) we immediately obtain from (2.23) the following Fierz identity:

$$B^{\alpha}(\Psi,\chi)\bullet B^{\beta}(\chi',\Psi') = C^{\alpha\beta}_{\gamma\delta}B^{\gamma}_{0}(\Psi,\Psi')B^{\delta}(\chi',\chi) \tag{2.26}$$

with

$$C_{\alpha\beta\gamma\delta} = g_{\alpha\gamma}g_{\beta\delta} + g_{\alpha\delta}g_{\beta\gamma} - g_{\alpha\beta}g_{\gamma\delta} - i\epsilon_{\alpha\beta\gamma\delta}, \tag{2.27}$$

and where $g_{\alpha\beta}$ is the Lorentz metric tensor and $\epsilon_{\alpha\beta\gamma\delta}$ is the permutation tensor.[19] [We may substitute $\hat{B}$ for B in the Fierz identity (2.26) to obtain formal Lorentz equations.]

As an application of formulas (2.23) and (2.26), let $\Psi\in C^4$ be a spinor pair. Define

$$j^{\alpha}_{K} = B^{\alpha}_{K}(\Psi^{*},\Psi), \quad F^{\alpha}_{K} = iB^{\alpha}_{K}(\Psi,\Psi). \tag{2.28}$$

Then $s = F^{0}_{0}$ is a scalar and $\mathbf{F}_k$ is a Yang–Mills vector triplet. From (2.20), $\mathbf{F}_0$ and F^{0}_{k} vanish. The j^{α}_{K} are a quadruplet of real Lorentz four-vector currents. Then (2.23) and (2.26) give, using (2.22) and (2.28),

$$j_J\circ\hat{j}_K = C^{PQ}_{JK}\overline{F}^{0}_{P}F_Q, \quad j^{\alpha}\bullet\hat{j}^{\beta} = C^{\alpha\beta}_{\gamma\delta}\overline{F}^{\gamma}_{0}F^{\delta}. \tag{2.29}$$

Using the fact that $F^{\alpha}_{0} = (s,0)$ and $F^{\alpha}_{k} = (0,\mathbf{F}_k)$, and also $\hat{j}^{\alpha}_{0} = j_{0\alpha}$ and $\hat{j}^{\alpha}_{k} = -j_{k\alpha}$, we get from the first equation of (2.29),

$$j^{\alpha}_{J}j_{K\alpha} = |s|^2g_{JK}, \tag{2.30}$$

which shows that $e^{\alpha}_{K} = j^{\alpha}_{K}/|s|$, for $K = 0, 1, 2, 3$, form an orthonormal Lorentz basis for R^4. Also, from the first equation of (2.29) we get

$$\mathbf{F}_k = s(e^{0}_{0}\mathbf{e}_k - e^{0}_{k}\mathbf{e}_0 + i\mathbf{e}_0\times\mathbf{e}_k). \tag{2.31}$$

From the second equation of (2.29) we get

$$e^{\alpha}_{K}e^{\beta K} = g^{\alpha\beta}, \tag{2.32}$$

which again shows that the e^{α}_{K} form an orthonormal Lorentz basis. Also from the second equation of (2.29) we obtain the following equation which is equivalent to (2.31):

$$F^{\alpha\beta}_{k} = s(e^{\alpha}_{0}e^{\beta}_{k} - e^{\alpha}_{k}e^{\beta}_{0} + i\epsilon_{kpq}e^{\alpha}_{p}e^{\beta}_{q}), \tag{2.33}$$

where $F^{\alpha\beta}_{k} = (\mathbf{F}_k, -i\mathbf{F}_k)$ are the antisymmetric tensors associated with the Yang–Mills vectors $\mathbf{F}_k$. Given that the e^{α}_{K} are orthonormal, formula (2.33) reduces to (2.31) and is equivalent to all 75 equations derived by Takahashi.[20]

[1]F. Reifler, J. Math. Phys. **25**, 1088 (1984).
[2]F. Reifler, J. Math. Phys. **26**, 542 (1985).
[3]F. Reifler and R. Morris, J. Math. Phys. **26**, 2059 (1985).
[4]F. Reifler and R. Morris, "A SU(2)×SU(2)×U(1) model for color-electroweak interactions," submitted to J. Math. Phys.
[5]R. Morris and F. Reifler, "Higgs field dynamics in the SU(2)×SU(2)×U(1) model of color-electroweak interactions," submitted to J. Math. Phys.
[6]See Ref. 2, pp. 547 and 548.
[7]Y. Takahashi, J. Math. Phys. **24**, 1783 (1983).
[8]See Ref. 2, p. 548.
[9]Y. Takahashi, J. Math. Phys. **24**, 1783 (1983), pp. 1786 and 1787. Takahashi lists these general Fierz identities as F-1 to F-15.
[10]By substituting Ψ^{*} and Ψ and χ'^{*} for χ', formula (1.6) becomes a *tensor* identity, and hence the components $C^{\alpha\beta}_{\gamma\delta}$ form a tensor. See Ref. 2, p. 548. Note that we can express $C^{\alpha\beta}_{\gamma\delta}$ directly as follows:

$$C^{\alpha\beta}_{\gamma\delta} = g^{\alpha}_{\gamma}g^{\beta}_{\delta} + g^{\alpha}_{\delta}g^{\beta}_{\gamma} + g^{\alpha\beta}g_{\gamma\delta} - ig^{\alpha\alpha'}g^{\beta\beta'}\epsilon_{\alpha'\beta'\gamma\delta},$$

where $g_{\alpha\beta}$ = Lorentz metric tensor and $\epsilon_{\alpha\beta\gamma\delta}$ = permutation tensor.
[11]Conforming with our previous notation in Refs. 1–5, *covariant* gauge indices K are written in the *raised* position, which is the opposite convention used for Lorentz indices α. This helps us separate the two types of indices in the formulas. For example, g^{JK} is the *formal* metric tensor, which is the gauge analog of $g_{\alpha\beta}$, the Lorentz metric tensor.
[12]Y. Takahashi, J. Math. Phys. **24**, 1783 (1983), pp. 1787–1790. Takahashi lists these Fierz identities as JJ-1 to RR̄-15.
[13]See Ref. 4, Sec. IV.
[14]See Ref. 4, Sec. II.
[15]For the cases $n = 1$, 2, and 3, see Refs. 1–5.
[16]R. Geroch, J. Math. Phys. **9**, 1739 (1968).
[17]See Ref. 2, p. 548.
[18]The components C^{PQ}_{JK} form a formal gauge tensor. See footnote 11.
[19]See footnote 10.
[20]Y. Takahashi, J. Math. Phys. **24**, 1783 (1983).

A Green's function approach to third-order elastic constants of disordered solids

T. R. Middya and A. N. Basu
Physics Department, Jadavpur University, Calcutta 700032, India

S. Sengupta
Solid State Physics Research Centre, Physics Department, Presidency College, Calcutta 700073, India

(Received 9 January 1985; accepted for publication 11 June 1986)

Starting from a Lippmann–Schwinger-type equation, which is very similar to that of quantum mechanical multiple scattering theory, Zeller and Dederichs [Phys. Status Solidi B **55**, 831 (1973)] have developed the effective medium theory. This theory has found wide application in understanding the mechanical behavior of disordered solids. However, unlike the problem in quantum mechanics, this equation of the random elasticity is only approximate since this is a linear response theory. So, it is proposed in this work for the first time to go beyond this approximation to treat nonlinear properties of such solids of which the third-order elastic constant is a generic. Again, so far as the nonlinear elastic behavior of these solids is concerned, no work has been done except the simple Voigt- and Reuss-type averaging. Both are extreme approximations and are, moreover, known to lead to violation of the equilibrium condition. The salient feature of the present calculation is to get an exact formal solution of the problem in terms of an appropriate Green's function in a closed form. The result obtained is quite general and may be adopted to treat nonlinearity in any tensor property of disordered materials. Finally several approximations, including a self-consistent solution, have been discussed for obtaining the effective nonlinear static mechanical susceptibility.

I. INTRODUCTION

It is becoming increasingly evident that the different theoretical methods[1–4] developed for calculating the macroscopic elastic properties of disordered materials, though highly instructive and ingenious, are basically intuitive and hence are rather difficult to apply. Even in the case of the most effective and rigorous technique involving the variational procedure[5] an educated guess is essential for the choice of the trial functions. On the other hand, as emphasized by Gubernatis and Krumhansl,[6] starting from the local stress–strain relation a rigorous theoretical formulation of the problem seems to have been inhibited so far by the difficulties of solving the resulting differential equations with randomly fluctuating values of the coefficients. However, the same equation being cast in the framework of an integral equation formalism is amenable to iterative or perturbative methods of solution.

The above idea has been used by many authors[7,8] for calculating various effective properties of a macroscopically homogeneous medium with fluctuating values of the same property on a microscopic scale. In particular Zeller and Dederichs[9,10] have developed an interesting and theoretically satisfying formulation of the problem for the effective elastic properties of disordered materials. They have quite successfully exploited the idea of an effective medium whose mathematical formulation is algebraically equivalent to that of the quantum mechanical scattering theory. Recently Middaya, Sarkar, and Sengupta[11] have applied the method to discuss the elastic properties of noncubic polycrystals. Operationally a self-consistent effective medium solution implies that the average scattering by the crystallites composing the solid is zero, when they scatter independently. This idea has been employed in different areas of condensed matter physics that contain some disorder.[12] In the case of elastic properties, the idea has been somewhat as follows. The strain, at any point in a specimen, is not determined alone by the local stress at the point under consideration. The strains from other parts of the solid, due to different values of the elastic constants at those parts generated owing to the microscopic inhomogeneity of structure, propagate to the given point. Then solving the relevant equations self-consistently we arrive at the effective elastic constants corresponding to which the fictitious medium is called the effective medium.

It is, however, to be noted that all the applications envisaging the idea of the effective medium have so far been confined to only the linear response regime; namely, the second-order elastic constants, the second-rank dielectric tensor, and similar linear response susceptibilities of thermal and magnetic properties of disordered materials. The main purpose of the present investigation is to extend the idea of the effective medium approach to the nonlinear domain. The particular case that we shall treat is the case of the third-order elastic constants, which is a generic for all static nonlinear susceptibilities. Specifically the theory of Zeller and Dederichs[9] will be employed for the calculation. The solution of the resulting integral equation may be suitably adopted to treat nonlinearity in the other physical properties also.

In the next section we develop the general theory including the nonlinear response. It is shown that under some physically valid assumptions the exact formal solution for the effective tensor of the third-order elastic constants is obtained. Then the different approximate methods of solution and a self-consistent solution of the final equation have been discussed.

In the last section we discuss the results and their possible applications.

 0022-2488/86/112807-06$02.50

II. THEORY

A. Strain distribution

In order to calculate the third-order elastic constants of a polycrystal the usual method[13] has so far been to assume either the constant strain (Voigt average) or the constant stress (Reuss average) throughout the specimen. The isotropy observed is satisfied by one of the above averages. Both are extreme approximations and they are not theoretically consistent either, because in both cases the equilibrium condition is violated at the boundaries of the crystallites. Sometimes a mean of the above two averages is calculated (Hill average), which, of course, does not correspond to any defined stress–stress relation. So, it is imperative, for any theory that tries to go beyond these approximations, to take into consideration the basic equilibrium equation to determine the actual strain distribution among the crystallites. In the following discussion we make such an attempt.

Since the nonlinear part of the response cannot be considered separately from the linear part, we relate the total local stress and strain fields by the following relation:

$$\sigma_{ij}(\mathbf{r}) = C_{ijkl}(\mathbf{r}) + \tfrac{1}{2} C_{ijklmn}(\mathbf{r})\epsilon_{kl}(\mathbf{r})\epsilon_{mn}(\mathbf{r}), \tag{1}$$

where σ_{ij} and ϵ_{kl}'s represent the stress and the strain, respectively. The second-order and third-order elastic moduli tensors, C_{ijkl} and C_{ijklmn}, and the corresponding compliances S_{ijkl} and S_{ijklmn} are in general functions of $\mathbf{r}$. For disordered materials these quantities will vary in a statistical manner on a microscopic scale. We shall specialize our calculation in subsequent work for a polycrystal for which the statistical distribution will refer, in particular, to grain orientation. The term "grain" is used generically to mean even dislocations, cracks, voids, and similar other irregularities. Our purpose is similar to what Zeller and Dederichs have achieved with Eq. (1) neglecting the second term on the right-hand side of the equation; namely, to establish a connection between the ensemble averaged stress and strain fields (defined as $\langle\sigma\rangle$ and $\langle\epsilon\rangle$) for macroscopic homogeneous materials through the following sets of nonlocal second- and third-order elastic constants defined as follows:

$$\langle\sigma_{ij}(\mathbf{r})\rangle = \int d\mathbf{r}'\, C^{\text{eff}}_{ijkl}(\mathbf{r},\mathbf{r}')\langle\epsilon_{kl}(\mathbf{r}')\rangle + \frac{1}{2}\int d\mathbf{r}'\, C^{\text{eff}}_{ijklmn}(\mathbf{r},\mathbf{r}')\langle\epsilon_{kl}(\mathbf{r}')\rangle\,\langle\epsilon_{mn}(\mathbf{r}')\rangle. \tag{2}$$

It should be emphasized that Eq. (2), as it may apparently seem, does not imply the approximation

$$\langle\epsilon\epsilon\rangle = \langle\epsilon\rangle\,\langle\epsilon\rangle,$$

which decouples the combined effects of fluctuation and nonlinearity, which will, of course, be evident from the detailed calculation given in Eqs. (11)–(17).

Equation (2) should be interpreted as follows.

The strain field $\epsilon(\mathbf{r})$ is rapidly varying in the sense that it changes value from one crystallite to another. The strain fields $\langle\epsilon\rangle$ and ϵ^0 [the strain field corresponding to the nonfluctuating part; see Eq. (11)] on the other hand are slowly varying. For a given distribution of elastic constants, however, one can map both the ϵ field and $\langle\epsilon\rangle$ field from ϵ^0. Symbolically writing this nonlocal integral relation as

$$\epsilon = J\epsilon^0$$

where J is an integral operator, we get

$$\langle\epsilon\rangle = \langle J\rangle\epsilon^0.$$

Using the above expression for ϵ we get from Eq. (1)

$$\sigma = C^{(2)}J\epsilon^0 + \tfrac{1}{2} C^{(3)}JJ\epsilon^0\epsilon^0$$

and

$$\begin{aligned}\langle\sigma\rangle &= \langle C^{(2)}J\rangle\epsilon^0 + \tfrac{1}{2}\langle C^{(3)}JJ\rangle\epsilon^0\epsilon^0 \\ &= \langle C^{(2)}J\rangle\,\langle J\rangle^{-1}\langle\epsilon\rangle \\ &\quad + \tfrac{1}{2}\langle C^{(3)}JJ\rangle\,\langle J\rangle^{-1}\langle J\rangle^{-1}\langle\epsilon\rangle\,\langle\epsilon\rangle,\end{aligned}$$

where $C^{(2)} \equiv C_{ijkl}$ and $C^{(3)} \equiv C_{ijklmn}$.

It is precisely the above result that is given in Eq. (2) to indicate just the fact that $\langle\sigma\rangle$ and $\langle\epsilon\rangle$ are related by nonlocal integral relation.

As emphasized in Ref. 9, the nonlocality of C^{eff}_{ijkl} and C^{eff}_{ijklmn} are not very important. Therefore, Eq. (2) can be replaced by a local stress–strain relation with the constant effective tensors C^* and C^{**}

$$\langle\sigma_{ij}(\mathbf{r})\rangle = C^*_{ijkl}\langle\epsilon_{kl}(\mathbf{r})\rangle + \tfrac{1}{2} C^*_{ijklmn}\langle\epsilon_{kl}(\mathbf{r})\rangle\,\langle\epsilon_{mn}(\mathbf{r})\rangle, \tag{2'a}$$

where

$$C^*_{ijkl} = \int d\mathbf{r}'\, C^{\text{eff}}_{ijkl}(\mathbf{r},\mathbf{r}') \tag{2'b}$$

and

$$C^{**}_{ijklmn} = \int d\mathbf{r}'\, C^{\text{eff}}_{ijklmn}(\mathbf{r},\mathbf{r}'). \tag{2'c}$$

If we neglect the second term in Eq. (2') we retrieve the relation of Zeller and Dederichs.[9] The present treatment will yield two sets of effective elastic constants. Of them the second order will be identical with that of Ref. 9. The only thing known about the disordered system is the statistical distribution of $C_{ijkl}(\mathbf{r})$ and $C_{ijklmn}(\mathbf{r})$. The equation describing the equilibrium condition is obtained by taking the divergence of Eq. (1),

$$\text{div}(\sigma_{ij}(\mathbf{r})) = 0, \tag{3}$$

which written explicitly gives

$$\text{div}(C_{ijkl}\epsilon_{kl} + \tfrac{1}{2} C_{ijklmn}\epsilon_{kl}\epsilon_{mn}) = 0. \tag{4}$$

This equation satisfies the equilibrium condition regarding the displacement vectors s_k. The boundary conditions are given in the form of either the surface displacements or the surface forces.

For obtaining the solution of the above equation we split both the fluctuating tensors $C_{ijkl}(\mathbf{r})$ and $C_{ijklmn}(\mathbf{r})$ into a constant part and a fluctuating part as follows:

$$C_{ijkl}(\mathbf{r}) = C^0_{ijkl} + \delta C_{ijkl}(\mathbf{r}) \equiv C^{(2)}$$

and

$$C_{ijklmn}(\mathbf{r}) = C^{00}_{ijklmn} + \delta C'_{ijklmn}(\mathbf{r}) \equiv C^{(3)}. \tag{5}$$

Now introducing the displacement vectors we write Eq. (4) in the following form:

$$(C_{ijkl}(\mathbf{r})s_{k|l}(\mathbf{r}))_{|j} + \tfrac{1}{2}(C_{ijklmn}(\mathbf{r})s_{k|l}s_{m|n})_{|j} = 0, \tag{6}$$

where

$$|_j = \frac{\partial}{\partial r_j}.$$

Now considering both the fluctuating parts of the second-order and the third-order elastic tensors, as well as the constant part of the third-order elastic tensor as the inhomogeneous terms in the above equation, we give below the above equation with the homogeneous and inhomogeneous parts

$$C^0_{ijkl}\frac{\partial^2 s_k}{\partial r_i\,\partial r_j}+\frac{\partial}{\partial r_j}\left[\delta C_{ijkl}(\mathbf{r})\frac{\partial s_k}{\partial r_l}+\frac{1}{2}C_{ijklmn}(\mathbf{r})\frac{\partial s_k}{\partial r_l}\frac{\partial s_m}{\partial r_n}\right]=0. \tag{6'}$$

The solution of the homogeneous equation is solved by introducing the Green's function

$$C^0_{ijkl}\,g_{kp|lj}(\mathbf{r},\mathbf{r}') = -\delta_{ip}\,\delta(\mathbf{r}-\mathbf{r}'). \tag{7}$$

This part of the equation remains the same as that given by Zeller and Dederichs. The boundary conditions also remain unaltered, namely $g_{kp}(\mathbf{r},\mathbf{r}')=0$, for all $\mathbf{r}$ on the surface or there are no surface forces. Then for the given surface displacements s^0_k we obtain from Eq. (6)

$$\begin{aligned}s_k = s^0_k(\mathbf{r}) + \int d\mathbf{r}'\, g_{ki}(\mathbf{r},\mathbf{r}')\big[&(\delta C_{ijkl}(\mathbf{r}')s_{k|l}(\mathbf{r}'))_{|j'}\\ &+\tfrac{1}{2}C^{00}_{ijklmn}(s_{k|l}(\mathbf{r}')s_{m|n}(\mathbf{r}'))_{|j'}\\ &+\tfrac{1}{2}(\delta C'_{ijklmn}(\mathbf{r}')s_{k|l}(\mathbf{r}')s_{m|n}(\mathbf{r}'))_{|j'}\big],\end{aligned} \tag{8}$$

where $s^0_k(\mathbf{r})$ is the displacement field corresponding to the constant part of the elastic constant, C^0. By partially integrating the above equation and then differentiating we get the resulting integral equation for the strain tensor:

$$\begin{aligned}\epsilon_{kq}(\mathbf{r}) = \epsilon^0_{kq}(\mathbf{r}) &+ \int d\mathbf{r}'\, G_{kqij}(\mathbf{r},\mathbf{r}')\delta C_{ijkl}(\mathbf{r}')\epsilon_{kl}(\mathbf{r}')\\ &+\frac{1}{2}\int d\mathbf{r}'\, G_{kqij}(\mathbf{r},\mathbf{r}')C^{00}_{ijklmn}\epsilon_{kl}(\mathbf{r}')\epsilon_{mn}(\mathbf{r}')\\ &+\frac{1}{2}\int d\mathbf{r}'\, G_{kqij}(\mathbf{r},\mathbf{r}')\delta C'_{ijklmn}(\mathbf{r}')\epsilon_{kl}(\mathbf{r}')\epsilon_{mn}(\mathbf{r}')\end{aligned} \tag{9}$$

with

$$G_{kqij} = -\tfrac{1}{2}(g_{ki|qj'}+g_{qi|kj'}) \tag{10}$$

and

$$|_{j'} = \frac{\partial}{\partial r'_j}.$$

The terms G_{kqij} and ϵ^0_{kq} depend neither on the fluctuating quantities nor on the constant nonlinear part of the elastic constants. An excellent discussion of the fundamental Eq. (8) excluding third order is given in Ref. 14. Introducing short notation, Eq. (9) may be written in either of the following forms:

$$\epsilon = \epsilon^0 + G\delta C\epsilon + \tfrac{1}{2}GC^{00}\epsilon\epsilon + \tfrac{1}{2}G\delta C'\epsilon\epsilon$$

or (11)

$$\epsilon = \epsilon^0 + G\delta C\epsilon + \tfrac{1}{2}GC^{(3)}\epsilon\epsilon.$$

It is to be remembered that the integral Eq. (9) is the basic equation for most of what follows. We have, however, suppressed the $\mathbf{r}$ dependence and the tensor indices in Eq. (11) for subsequent manipulative purposes.

Let us now have recourse to an iterative method of the solution of Eq. (11). The iteration immediately leads to the following two sets of infinite series in terms of which the solution is given by

$$\begin{aligned}\epsilon = \epsilon^0 &+ G(\delta C + \delta C G\,\delta C + \delta C G\,\delta C G\,\delta C + \cdots\infty)\epsilon^0\\ &+\tfrac{1}{2}G(C^{(3)} + 3C^{(3)}G\,\delta C + 6C^{(3)}G\,\delta C G\,\delta C\\ &+ 10C^{(3)}G\,\delta C G\,\delta C G\,\delta C + \cdots\infty)\epsilon^0\epsilon^0\\ &+ \text{higher-order terms.}\end{aligned} \tag{12}$$

The different type of product operators will have the following meaning. In general the r-dependent quantities are $G(\mathbf{r},\mathbf{r}')$, $C^{(2)}(\mathbf{r},\mathbf{r}') = C^{(2)}(\mathbf{r})\delta(\mathbf{r},\mathbf{r}')$, $C^{(3)}(\mathbf{r},\mathbf{r}') = C^{(3)}(\mathbf{r})\delta(\mathbf{r},\mathbf{r}')$, and the fundamental tensors are of rank 2, 4, 6, etc. The products are defined as follows:

$$(G\,\delta C\epsilon)_{ij} = \int d\mathbf{y}\, G_{ijkl}(\mathbf{x},\mathbf{y})\delta C_{klmn}(\mathbf{y})\epsilon_{mn}(\mathbf{y}), \tag{13a}$$

$$(G\,\delta C G\,\delta C\epsilon)_{ij} = \int d\mathbf{x}\int d\mathbf{y}\, G_{ijkl}(\mathbf{z},\mathbf{x})\delta C_{klmn}(\mathbf{x}) \times G_{mnpq}(\mathbf{x},\mathbf{y})\delta C_{pqrs}(\mathbf{y})\epsilon_{rs}(\mathbf{y}), \tag{13b}$$

$$(GC^{(3)}\epsilon\epsilon)_{ij} = \int d\mathbf{y}\, G_{ijkl}(\mathbf{x},\mathbf{y})C^{(3)}_{klmnop}(\mathbf{y})\epsilon_{mn}(\mathbf{y})\epsilon_{op}(\mathbf{y}), \tag{13c}$$

and

$$(GC^{(3)}G\,\delta C\epsilon\epsilon)_{ij} = \int d\mathbf{x}\int d\mathbf{y}\, G_{ijkl}(\mathbf{z},\mathbf{x})B_{klmnop}\epsilon_{mn}\epsilon_{op}, \tag{13d}$$

where

$$B_{klmnop} = C^{(3)}_{klmnab}(\mathbf{x})G_{abcd}(\mathbf{x},\mathbf{y})\delta C_{cdop}(\mathbf{y}), \tag{13e}$$

etc. It is further important to note that in the second series in Eq. (12) the same quantities occur in mixed order in different terms. The order may be interchanged provided the Voigt symmetry of each term is ensured. This happens because the terms involve products of tensors of ranks 6 and 4. So, in any application of the above formula each term has to be properly symmetrized. There is no such problem with the first series where the symmetry is automatically satisfied. Now, the above two series may be summed up to yield an exact formal solution in closed form. This is possible since on repeated interation of Eq. (12) no terms other than those indicated in the series representing the coefficients of ϵ^0 and $\epsilon^0\epsilon^0$ will be generated. So for obtaining the second- and the third-order elastic constants the cubic and and higher-order terms in ϵ^0 are neglected and we write

$$\epsilon = \epsilon^0 + GT\epsilon^0 + \tfrac{1}{2}GT'\epsilon^0\epsilon^0, \tag{14a}$$

where

$$T = \delta C(I - G\,\delta C)^{-1} \tag{14b}$$

and

$$T' = C^{(3)}(I - G\,\delta C)^{-3} \tag{14c}$$

and the previously defined multiplication rules for operators apply. The unit operator I is defined as

$$I_{ijkl}(\mathbf{r},\mathbf{r}') = \tfrac{1}{2}(\delta_{ik}\,\delta_{jl} + \delta_{il}\,\delta_{jk})\delta(\mathbf{r}-\mathbf{r}'). \tag{15}$$

It is interesting to note that in Eq. (14) the terms linear in ϵ^0 incorporating the fluctuations in the second-order elastic constants remain unchanged, while the higher-order term involves both the fluctuations of the second order and the whole of the third-order elastic constants. A knowledge of T and T' completely specifies both the linear and the nonlinear susceptibilities. In order to avoid the rather unwieldy expressions in the algebra, a shorthand notation is often used to represent the above formulas by diagrams. The associated diagram for calculation is shown in Fig. 1. Next we shall use Eq. (14) to derive the effective elastic constants of the medium.

B. Effective elastic constants

In order to get the expression for the effective nonlocal elastic constants $C^{\text{eff}}_{ijkl} \equiv C^{(2)\text{eff}}$ and $C^{\text{eff}}_{ijklmn} \equiv C^{(3)\text{eff}}$ we require the macroscopic strain given by Eq. (14). Thus we get

$$\langle\epsilon\rangle = \epsilon^0 + \langle GT\rangle\epsilon^0 + \tfrac{1}{2}\langle GT'\rangle\epsilon^0\epsilon^0. \tag{16}$$

The macroscopic stress is given by [averaging Eq. (1)]

$$\langle\sigma\rangle = \langle C^{(2)}\epsilon\rangle + \tfrac{1}{2}\langle C^{(3)}\epsilon\epsilon\rangle, \tag{17}$$

which in terms of the effective elastic constant is

$$\langle\sigma\rangle = C^{(2)\text{eff}}\langle\epsilon\rangle + \tfrac{1}{2}\,C^{(3)\text{eff}}\langle\epsilon\rangle\,\langle\epsilon\rangle. \tag{18}$$

To find expressions for $C^{(2)\text{eff}}$ and $C^{(3)\text{eff}}$ in terms of the T-matrices we have to solve for ϵ^0 in terms of $\langle\epsilon\rangle$. Iterating Eq. (16) in the reverse order we get retaining terms up to second order in macroscopic strains

$$\epsilon^0 = (I + \langle GT\rangle)^{-1}\langle\epsilon\rangle - (I + \langle GT\rangle)^{-3}\langle GT'\rangle\,\langle\epsilon\rangle\,\langle\epsilon\rangle. \tag{19}$$

Next separating the constant and the fluctuating parts of the elastic constants Eq. (17) may be written as

$$\langle\sigma\rangle = C^0\langle\epsilon\rangle + \langle\delta C\epsilon\rangle + \tfrac{1}{2}\,C^{00}\langle\epsilon\epsilon\rangle + \tfrac{1}{2}\langle\delta C'\epsilon\epsilon\rangle. \tag{20}$$

Then with the help of Eqs. (14) and (19) we get the macroscopic stress $\langle\sigma\rangle$ in terms of $\langle\epsilon\rangle$ and the T matrices. Comparing this final expression with Eq. (18) we finally arrive at the following expressions:

$$C^{(2)\text{eff}} = C^0 + \langle T\rangle(I + \langle GT\rangle)^{-1} \tag{21}$$

and

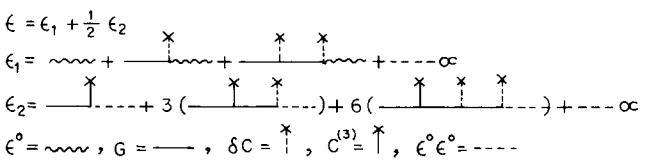


FIG. 1. Diagrammatic representation of integral Eq. (12) for the total strain field. Each vertex (intersection point of different lines) implies operator multiplication. The symbols are not unique.

$$\begin{aligned} C^{(3)\text{eff}} = {} & C^{00}\langle(I+GT)^2\rangle(I+\langle GT\rangle)^{-2} \\ & - \langle T\rangle\,\langle GT'\rangle(I+\langle GT\rangle)^{-3} \\ & + \langle T(I+GT)^{-1}GT'\rangle(I+\langle GT\rangle)^{-2} \\ & + \langle T''(I+GT)^{-1}\rangle(I+\langle GT\rangle)^{-2}, \end{aligned} \tag{22}$$

where Eq. (21) is identical with that obtained by Zeller and Dederichs and Eq. (22) is the new result giving the effective third-order elastic constants. While writing Eq. (22) we have used the following notation:

$$\begin{aligned} T' &= C^{(3)}(I - G\,\delta C)^{-3} = C^{00}(I+GT)^3 + T'' \\ &= C^{00}(I - G\,\delta C)^{-3} + T'' \end{aligned} \tag{23a}$$

and

$$T'' = \delta C'(I - G\,\delta C)^{-3}. \tag{23b}$$

Equations (21) and (22) provide the complete solution for the effective nonlocal second-order and third-order elastic constants if $\langle T\rangle$, $\langle GT\rangle$, $\langle T''\rangle$, and $\langle GT'\rangle$ can be computed. But all these quantities may only be approximated and the nature of the material will supply the necessary guidance.

C. Method of evaluation and different approximations

In this section we discuss some methods for obtaining the effective elastic constants from Eq. (21) and (22) assuming different models of the solid. As shown in Ref. 6 the elastic stiffness constants may be decomposed without any loss of generality into the sum of contributions from different grains (i.e., in general it may include pores, vacancies, dislocations, etc.; the only assumption is that there is a distribution of species)

$$\delta C = \sum_\alpha \delta C^\alpha \theta^\alpha(\mathbf{r}), \tag{24}$$

$$\delta C' = \sum_\alpha \delta C'^\alpha \theta^\alpha(\mathbf{r}), \tag{25}$$

where $\theta^\alpha(\mathbf{r})$ is a step function whose value is unity when $\mathbf{r}$ is in the αth grain and zero otherwise. Here δC^α and $\delta C'^\alpha$ are both constant fourth-and sixth-rank tensors, respectively. Hence forth we drop writing the step function explicitly; the presence of α will imply it. We now indicate some approximations of increasing accuracy.

1. Voigt average

The simplest application of the above theory will lead to results already known.[13,15] In particular if the disorder refers to different orientation of the crystallites in a specimen and if we make the assumption that the strain everywhere is the same, i.e., $\epsilon = \epsilon^0$ we find from Eqs. (21), (22), (23a), and (23b)

$$C^* = (C^{(2)})_{\text{Voigt}} \tag{26}$$

and

$$C^{**} = (C^{(3)})_{\text{Voigt}}, \tag{27}$$

where $(C^{(2)})_{\text{Voigt}}$ and $(C^{(3)})_{\text{Voigt}}$ are simple volume average of the elastic constants. In this approximation both the ensemble averages $\langle\delta C\rangle$ and $\langle\delta C'\rangle$ are zero, i.e., there is no fluctuation in the medium. From a certain point of view the

above results may be seen as the correction to the effective elastic constants of lowest order in δC and $\delta C'$. Similarly with constant stress approximation we get the known result, namely, the Reuss average. As has been indicated earlier these approximations, however, lead to certain difficulties regarding equilibrium equation.

2. Self-consistent solution

In order to keep the discussion general we briefly indicate the method of the self-consistent solution of Eqs. (21) and (22). Although the T-matrices have been extensively used in the quantum mechanical problems connected with the disordered systems there have been only a few attempts to tackle problems related to classical mixtures. A particularly illuminating discussion in the case of the continuum problem is given by Zeller and Dederichs[14] and in another case of network problems by Kirkpatrick.[16] Recently Hori,[17] in a series of papers, has treated the problem of classical mixtures of dielectrics. He has given a critical discussion of the T-matrix expansion. Following these works we define the following two matrices for a single grain α in analogy with the T and T'' matrices for the whole system:

$$t_\alpha = \delta C^\alpha + \delta C^\alpha G t_\alpha = \delta C^\alpha (I - G\,\delta C^\alpha)^{-1} \tag{28}$$

and

$$t''_\alpha = \delta C'^\alpha + \delta C'^\alpha G t''_\alpha = \delta C'^\alpha (I - G\,\delta C^\alpha)^{-3}, \tag{29}$$

where δC's are given by Eqs. (24) and (25). Then the well-known T-matrix expansion as in the multiple scattering theory becomes

$$T = \sum_\alpha t_\alpha + \sum\sum_{\alpha\neq\beta} t_\alpha G t_\beta + \sum\sum\sum_{\alpha\neq\beta\neq\gamma} t_\alpha G t_\beta G t_\gamma \tag{30}$$

and

$$T'' = \sum_\alpha t''_\alpha + \sum\sum_{\alpha\neq\beta} t''_\alpha\, G t''_\beta + \sum\sum\sum_{\alpha\neq\beta\neq\gamma} t''_\alpha\, G t''_\beta\, G t''_\gamma, \tag{31}$$

where each sum is taken where no two successive subscripts are equal. The Green's function depends on the statistical information about the shape, the orientation, and the relative position of the grains. So averaging the terms other than the first one in Eqs. (30) and (31) will include intergranular correlations even for the statistically independent grains. Now, if we assume that the intergranular correlations are small, all the higher-order correlative terms are neglected. This is the usual procedure of truncation, which is found to be a good approximation.

Although the t-matrices in Eqs. (30) and (31) appear formally similar, they are quite distinct both physically and operationally. Both, however, refer to scattering from the same grain and give rise to intragranular scattering. While the former refers to the scattering of stress due to fluctuation of the same tensorial property, namely, of second-order elastic constants, the latter envisages scattering the stress due to fluctuations of both the third-order (sixth-rank tensor) and the second-order elastic constants (fourth-rank tensor). In the latter case also, both the scattering occur from the same grain and hence the grainwise decomposition of terms in the new T''-matrix has been possible. The difference in character of the two types of T-matrices is seen in a more transparent way from the related diagrams shown in Fig. 2. The second diagram of Fig. 2 clearly projects the correlated character of the two types of fluctuations. The absence of fluctuation in the third-order constants immediately leads to a vanishing of the t''-matrix and that of in the second order to a vanishing of both.

After the grainwise decomposition and using the above-mentioned approximations let us write

$$\langle T\rangle \simeq \langle \tau\rangle, \tag{32}$$

$$\langle T''\rangle \simeq \langle \tau''\rangle, \tag{33}$$

where

$$\tau = \sum_\alpha t_\alpha \quad \text{and} \quad \tau'' = \sum_\alpha t''_\alpha,$$

where all the higher-order intergranular correlations are absent. We may now substitute values in the equations for the effective elastic constants, but we have not yet specified the C^0 and C^{00} on which τ, τ'', and G depend. There are several prescriptions possible. For example, if we have this plausible choice, namely,

$$C^0 = (C^{(2)})_{\text{Voigt}}, \quad C^{00} = (C^{(3)})_{\text{Voigt}}, \quad G = (G)_{\text{Voigt}}, \text{ etc.},$$

we have the approximation analogous to the average T-matrix approximation (ATA) in the theory of disordered materials[6,12]; when

$$C^{(2)\text{eff}} = (C^{(2)})_{\text{ATA}} \tag{34}$$

and

$$C^{(3)\text{eff}} = (C^{(3)})_{\text{ATA}}, \tag{35}$$

the ATA averages the strain field by single grain inclusion.

Further refinement of the solution may be possible if we look for a self-consistent solution by taking advantage of the free choice of C^0 and C^{00}. This is equivalent to absorbing the effects of the higher-order terms excluding intergranular

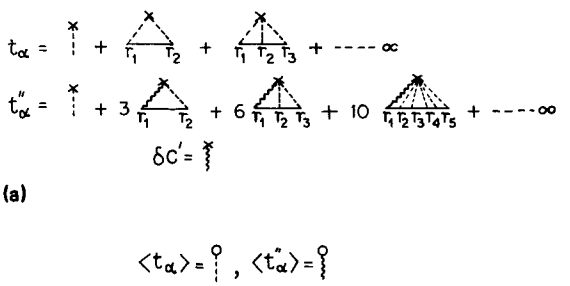


FIG. 2. (a) Diagrammatic representation of Eqs. (28) and (29), which represent two types of the single grain T-matrices. The recipe for calculating the diagrams is as follows. (i) The different points $\mathbf{r}_1,\mathbf{r}_2,\ldots$ are represented by nodes on the horizontal line. (ii) The propagator $G(\mathbf{r}_1,\mathbf{r}_2)$ is assigned to the solid line connecting $\mathbf{r}_1$ and $\mathbf{r}_2$ as in the previous figure. (iii) The crossed vertex with a following dotted line is the elastic constant operator in the second order and the same with a following wavy line is the elastic constant operator in the third order. (iv) Finally one has to take the operator product of all the Green's functions and the elastic operators to evaluate the diagrams. (b) The definition of the ensemble average of a single grain T-matrix in Fig. 2(a). A circle followed by a dotted line and a wavy line, respectively, is used to denote average for the second-order and the third-order cases.

scatterings. Since T and T'' are functions of C^0 and C^0 and C^{00}, a self-consistent solution is achieved by the following implicit coupled equations:

$$\langle \tau \rangle = \langle \tau' \rangle = \langle \tau'' \rangle = 0,$$

where

$$\langle 2G^2\tau^2 + 2G^3\tau^3 \rangle = \langle \tau' \rangle, \tag{36}$$

which finally deliver the nonlocal self-consistent (SC) values of the second-order and third-order elastic constants,

$$C^{(2)\text{eff}} = (C^{(2)})_{\text{SC}} = C^0, \tag{37}$$

$$C^{(3)\text{eff}} = (C^{(3)})_{\text{SC}} = C^{00}. \tag{38}$$

Equations (36) are the central result of the self-consistent approach corresponding to which the medium is an effective medium.

The local effective second-order elastic constants (C^*) and third-order elastic constants (C^{**}) can be obtained by using Eqs. (2'a),(2'b), and (36). The method of solving Eq. (36) on a computer may be developed following the strategy outlined by Gubernatis and Krumhansl[6] for the second-order elastic constants, which is, however, much more complex in the present case.

III. DISCUSSION

In this investigation we have found out the strain distribution, including the nonlinear part for a macroscopic homogeneous material with microscopic inhomogeneities. A formal solution of the problem has been obtained in Eq. (14) by the introduction of a T'-matrix. This solution has been utilized to get an expression of the effective third-order elastic constants of the medium. Equation (22) is quite general and may be applied for the calculation of the nonlinear susceptibilities in a variety of situations including homogeneous mechanical mixtures. A particularly important area of application will be that of single crystal aggregates, namely, polycrystals, which are real materials used in practical situations, which we shall consider in a future communication.

ACKNOWLEDGMENT

One of the authors (TRM) acknowledges the financial support of C.S.I.R., Government of India, in the form of a research fellowship.

[1] B. Kroner, Z. Phys. **151**, 504 (1958); J. Eng. Mech. Div. (Am. Soc. Civil Eng.) **106**, 889 (1980). (A review of the present situation is given in this paper.)
[2] J. D. Eshelby, Proc. R. Soc. London Ser. A **241**, 376 (1957).
[3] B. Budiansky, J. Mech. Phys. Solids **13**, 233 (1965).
[4] R. Hill, J. Mech. Phys. Solids **13**, 213 (1965).
[5] Z. Hashin and S. Shtrikman, J. Mech. Phys. Solids **10**, 343 (1962).
[6] J. H. Gubernatis and J. A. Krumhansl, J. Appl. Phys. **46**, 1875 (1975).
[7] G. Kneer, Phys. Status Solidi **9**, 825 (1965).
[8] P. R. Morris, Int. J. Eng. Sci. **3**, 49 (1970).
[9] R. Zeller and P. H. Dederichs, Phys. Status Solidi B **55**, 831 (1973).
[10] P. H. Dederichs and R. Zeller, Z. Phys. **259**, 103 (1973).
[11] T. R. Middya, A. Sarkar, and S. Sengupta, Acta Phys. Pol. A **66**, 561 (1984).
[12] P. Soven, Phys. Rev. **156**, 809 (1967); Phys. Rev. B **2**, 4715 (1970); S. Kirkpatrick, B. Velicky, and H. Ehrenreich, Phys. Rev. B **1**, 3250 (1970); R. J. Elliot, J. A. Krumhansl, and P. L. Leath, Rev. Mod. Phys. **46**, 465 (1974).
[13] G. R. Barsch, J. Appl. Phys. **39**, 3780 (1968).
[14] P. H. Dederichs and R. Zeller, KFA-Jul-Rep., JUL-877-FF, 1972.
[15] R. A. Hamilton and J. E. Parrott, J. Phys. (Proc. Phys. Soc.) C **1**, 829 (1968); C. S. G. Cousins, J. Phys. (Proc. Phys. Soc.) C **1**, 835 (1968).
[16] S. Kirkpatrick, Rev. Mod. Phys. **43**, 574 (1973).
[17] M. Hori, J. Math. Phys. **18**, 487 (1977). (The references of other works of the author and his collaborators are given in this paper.)