Uniqueness of the Virasoro algebra

Jack L. Uretsky
High Energy Physics Division, Argonne National Laboratory, Argonne, Illinois 60439 and Department of Natural Sciences, College of DuPage, Glen Ellyn, Illinois 60137

(Received 6 October 1989; accepted for publication 18 April 1990)

It is shown that the structure constants of the Virasoro algebra are uniquely specified, up to rescaling of the generators, by the requirement that they be nonvanishing. Permitting some of the structure constants to vanish leads to other Lie algebras, including the Witt algebras. Given the grading of the Virasoro algebra, nonvanishing of the structure constants is both necessary and sufficient for the algebra to be the Virasoro algebra.

I. INTRODUCTION

Fairlie, Nuyts, and Zachos[1] and I[2] have shown that the Virasoro algebra[3] may be constructed by starting with two generators and imposing no more than six conditions upon the repeated commutators of the generators. It is apparently an open question as to whether our number of conditions is minimal.

I investigate here the related question of the uniqueness of the algebra. That is, given a Z-graded Lie algebra[4] of the form

$$[L_m,L_n]=C(m,n)L_{m+n}, \tag{1}$$

what are the nontrivially different possible choices for the structure constants $C(m,n)$? The Virasoro algebra corresponds to

$$C(m,n)=m-n. \tag{2}$$

A trivially different structure constant would be one that corresponds to a rescaling of the elements L_m, that is, multiplication of each L_m by some constant C_m.

It turns out that if the $C(m,n)$ are required to be nonvanishing for m and n different, then Virasoro is the only possibility consistent with the Jacobi identities and the antisymmetry of $C(m,n)$. If $C(n,0)$ is permitted to vanish for some n greater than 1, then a resulting algebra is a Witt algebra[5] on a field of characteristic n. If $C(1,0)$ is allowed to vanish, then at least one of the resulting algebras, which I have not investigated in detail, is nonsimple, unlike the Virasoro algebra. The condition that the structure constants do not vanish is accordingly both necessary and sufficient for an algebra satisfying Eq. (1) to have the structure constants of the Virasoro algebra.

The Jacobi conditions may be written in terms of the $C(m,n)$ as

$$C(l,m)C(l+m,n)+C(n,l)C(n+l,m) + C(m,n)C(m+n,l)=0, \tag{3}$$

so that setting n equal to zero and using the antisymmetry of the C's gives

$$C(l,m)[C(l+m,0)-C(m,0)-C(l,0)]=0. \tag{4}$$

II. THE FIRST CASE: $C(M,N)\neq 0$ FOR $M\neq N$

Let (l,m) in Eq. (4) take the values $(1,-1)$ and $(2,-1)$, respectively, to find that

$$C(-1,0)=-C(1,0) \quad \text{and} \quad C(2,0)=2C(1,0). \tag{5a}$$

Induction on l and replacement of L_0 by $L_0/C(1,0)$ then leads to the conclusion that

$$C(l,0)=l, \quad \text{all} \quad l\in Z. \tag{5b}$$

Next define for all l,

$$C(l,-1)=(l+1)F(l), \tag{6a}$$

where $F(l)$ is an arbitrary function defined on the integers. Note that $F(l)$ must of course be nonzero for l different from -1. Equation (5b) implies that $F(0)$ is unity. The remaining $F(l)$'s may, however, be absorbed into the scaling of the L_1's according to the prescription

$$L_1\to L_1\prod_{k=0}^{l}F(k), \quad l>0 \tag{6b}$$

and

$$L_1\to L_1\prod_{k=2}^{l+1}F(-k)^{-1}, \quad l<-2, \tag{6b'}$$

with the result that the redefined operators satisfy

$$C(l,-1)=(l+1). \tag{6c}$$

It is important to notice, however, that the scale of the operators L_{-1} and L_{-2} is not yet established.

Choosing the value -1 for n in Eq. (3) leads to the equation, making use of Eq. (6c),

$$(l+m+1)C(l,m)=(l+1)C(l-1,m) + (m+1)C(l,m-1), \tag{7}$$

with the values of $C(l,0)$, $C(0,l)$, and $C(l,l)$ already specified for all l. It is then a straightforward matter to do induction on l and m to find that Eq. (2) holds for all positive (l,m) values. Also, induction on l for m equal to 1 in Eq. (7) leads to the same result for all values of l (positive and negative) with two notable exceptions. The coefficients $C(1,-2)$ and $C(2,-2)$ are not determined by the

 0022-2488/90/092059-02$03.00

Jacobi relations. The remark following Eq. (6c) shows, however, that any choice of these two arbitrary functions may be incorporated into the scaling of L_{-1} and L_{-2} in a way that preserves Eq. (2). Induction on l and m in Eq. (7) and its counterpart for n equal to -1 then demonstrates that the structure constants all satisfy Eq. (2).

I have therefore showed that the Virasoro algebra is the unique Lie algebra (except for trivial rescaling of the operators) satisfying Eq. (1) with nonvanishing structure constants.

III. THE SECOND CASE: WITT ALGEBRAS

I now consider the possibility that $C(l,0)$ vanishes for some value of l greater than unity. Let p be the value of l for which $C(l,0)$ vanishes, $C(l,m)$ being nonzero for $p>l>m>0$. Equation (4) shows that when the $C(l,m)$ are nonvanishing then the set $\{C(l,0)\}$ is mapped homomorphically by the additive group of integers. In the first case the mapping was an isomorphism. In the present case, with $C(p,0)$ vanishing, the $C(l,0)$ are isomorphic to the cyclic group of order p.

I choose the scale of L_0, just as before, to make $C(1,0)$ equal to unity. Equation (5b) then holds for integers l modulo p. Equation (6a) becomes

$$C(l,p-1)=(l+1)F(l), \quad l\leqslant p-1, \tag{8}$$

where all arguments are now to be understood modulo p. Equations (6b) and (6c) are then valid for l less than $p-1 \pmod p$.

The induction implied by Eq. (7) may be performed in the field of characteristic p to obtain Eq. (2) modulo p.

I conclude that the assumption that $C(p,0)$ vanishes for some p greater than unity leads to the structure constants of a Witt algebra. (See Ref. 6 for the interesting case of $p=3$ with the range of the structure constants shifted from $\{1,2',0\}$ to $\{\pm\sqrt{3/2},0\}$.)

IV. THE THIRD CASE: ALL THE $C(l,0)$ VANISH

Another way to satisfy Eq. (4) is to have the structure constant $C(1,0)$ vanish. Equation (4) is then satisfied if all the $C(l,0)$ equal zero (this does not appear to be the only possibility, however).

Suppose further that the $C(l,1)$ coefficients are nonvanishing. According to the discussion relating to Eqs. (6a) and (6b) and the definitions of the $C(l,-1)$ structure constants, I can scale the L-operators so that

$$C(l,1)=l(l-1), \quad l\geqslant 0. \tag{9}$$

I next set n equal to $-m$ in Eq. (3) to get

$$C(l,m)C(l+m,-m)=C(l,-m)C(l-m,m), \tag{10}$$

so that choosing m equal to 1 shows that the $C(l,-1)$ are all zero for l greater than 2. Further, $C(2,-1)$ is seen to be zero by setting n equal to $-m+1$ with $l\geqslant 3$ in Eq. (3).

Finally, by setting m equal to successively larger negative integers in Eq. (7) and using induction on l, I conclude that

$$C(l,-m)=0, \quad l>m>0. \tag{11}$$

Equation (11) shows that there are no "step-down" operators in the subalgebra generated by the $\{L_n | n>0\}$. The subalgebra is accordingly an algebraic ideal. The Lie algebra characterized by vanishing structure constants $C(l,0)$ is therefore nonsimple.

V. CONCLUSION

Nonvanishing of the structure constants of a Z-graded Lie algebra satisfying Eq. (1) is both necessary and sufficient for the algebra to have the structure constants of the Virasoro algebra.

ACKNOWLEDGMENTS

I am indebted to C. Zachos for many stimulating conversations and for a critical reading of an earlier version of this paper, and to the High Energy Physics Division of Argonne National Laboratory for its hospitality.

This work was supported in part by the U.S. Department of Energy, Division of High Energy Physics, Contract No. W-31-109-ENG-38.

[1] D. B. Fairlie, J. Nuyts, and C. K. Zachos, Commun. Math. Phys. **117**, 595 (1988).

[2] J. L. Uretsky, Commun. Math. Phys. **122**, 171 (1989).

[3] See I. Kaplansky, Commun. Math. Phys. **86**, 49 (1982); O. Mathieu, Invent. Math. **86**, 371–426 (1986) (especially pp. 374 and 392); V. G. Kac, *Infinite Dimensional Lie Algebras* (Cambridge U.P., New York, 1989), p. 96, and references cited there.

[4] Reference 2, p. 82.

[5] See, e.g., N. Jacobson, *Lie Algebras* (Dover, New York, 1979), p. 196, exercise 21.

[6] D. Fairlie, P. Fletcher, and C. Zachos, Phys. Lett. B **218**, 203 (1989).

A unitary representation of SL(2,R)

Henri Bacry
CNRS, Centre de Physique Theorique-Sec. 2, Luminy-Case 907, F-13288 Marseille Cedex 9, France

(Received 17 January 1990; accepted for publication 28 March 1990)

A given unitary representation of the group SL(2,R), belonging to the discrete series, is shown to involve necessarily some special functions (in particular, Laguerre and Hardy–Pollaczek polynomials). Various realizations of this representation are investigated, including the coherent states one. More generally, it is shown that the representations of the discrete series of the universal covering of SL(2,R) involves generalized Laguerre and Pollaczek polynomials. The Riemann zeta function is shown to be concerned with these representations.

I. INTRODUCTION

It is rather unusual to write an article devoted to a given representation of a Lie group. On one hand, mathematicians are more often interested in classifying all irreducible representations (defined up to an equivalence). This search leads them, secondarily, to the description of special functions in relation with the theory of characters or with the Peter–Weyl theorem. On the other hand, physicists, who are essentially concerned with the notions of observable and state, are often playing with different realizations of a given representation. For them, functions are not essential. As an example, a plane-wave state can be described either by an exponential $\exp(\mathbf{ik}\cdot\mathbf{x})$ or by a Dirac delta function $\delta^{(3)}(\mathbf{k})$. Both functions are identified in this context and denoted by a single ket$|\mathbf{k}\rangle$. Our point of view is close to the physicist one, although slightly different in that our aim is to know which functions are involved in a given representation. This can be described as follows. Let $X_1, X_2, \ldots, X_n$ be a basis of the Lie algebra of a group G. These elements are Hermitian operators corresponding to an irreducible unitary representation of G. They have either a discrete or a continuous spectrum. With physicists' notation, one can write, for instance,

$$X_1|x_1\rangle = x_1|x_1\rangle, \quad \text{where } x_1 \in \text{Spectrum } (X_1),$$

$$X_2|x_2\rangle = x_2|x_2\rangle, \quad \text{where } x_2 \in \text{Spectrum } (X_2).$$

The "matrix element" $\langle x_1|x_2\rangle$ is a function mapping Spectrum $(X_1)\times$Spectrum (X_2) in $\mathbb{C}$. This shows how functions are naturally involved in a given class of representations and how they can be effectively used to build a concrete example of this class of representations.

In the above example, the functions are functions of a real variable. The unitary representation of SL(2,R) we are interested in can be continued to a complex nonunitary representation of SL(2,C). This is a way to extend our study to functions of a complex variable. Such a procedure is analogous to the one that permits the physicist to define coherent states from the real Heisenberg algebra.

Another kind of generalization is possible and will be explored. It consists in replacing operators of the Lie algebra by operators of the enveloping algebra. This procedure permits us to enlarge the set of functions involved in the representation.

Many of the formulas presented in this article were obtained in collaboration with Michael Boon and presented elsewhere without proof.[1] The representation of SL(2,R) we are interested in was studied some years ago by Itzykson[2] in a few realizations, in particular the one involving the Hardy–Pollaczek polynomials. It is an irreducible and unitary representation belonging to the so-called discrete series and, due to the eigenvalue of the Casimir operator, it can be said to be of spin $-\frac{1}{2}$ and is denoted $D_+(-\frac{1}{2})$ in the present article.

The Lorentz group in three dimensions acts canonically on the (real) Lie algebra of SL(2,R). It follows that this Lie algebra has five kinds of nonzero elements (timelike future, timelike past, lightlike future, lightlike past, and spacelike). In the representation we are investigating, a timelike element has a discrete spectrum. Given one of them, any other element is associated with a set of orthogonal polynomials as follows:

timelike	(discrete spectrum)	Meixner polynomials of the first kind
lightlike	(continuous spectrum)	Laguerre polynomials
spacelike	(continuous spectrum)	Meixner polynomials of the second kind

This is a foretaste of our approach.

In the next section, we investigate some general properties of the representation *class* $D_+(-\frac{1}{2})$. It is shown how the Meixner and Laguerre polynomials appear in a natural way. Section III is devoted to some properties of the Hardy–Pollaczek polynomials,[3–5] a special case of the Meixner polynomials.[5] They have a slightly different writing (Pidduck polynomials[6]) that is of interest regarding special functions. In Secs. IV and V, we examine realizations where Hardy–Pollaczek and Laguerre polynomials are involved. Section VI relates our point of view with the usual realizations of the representation, especially the z realization. Section VII is devoted to the so-called coherent states realization and its dual. In Sec. VIII, we show how the Riemann zeta function is involved in the representation. Then, in the last section, we generalize our investigation to the so-called discrete series $D_+(-\frac{1}{2}+\epsilon)$ of the universal covering group of SL(2,R). It is necessary to emphasize that our notation is the one familiar to physicists (Hermitian operators as elements of the Lie algebra). Many of the results of these sections are collected in some tables.

Three appendices are added. Appendix A shows the re-

 0022-2488/90/092061-17$03.00

lationship between the three-dimensional Lorentz group and SL(2,R). Appendix B gives a physicist derivation of the unitary dual of the universal covering of SL(2,R). It permits us to have an idea of the topology of this space. The interest is explained in our conclusion. Appendix C shows the interest of what we call the x realization. Many special functions are described in a compact way with the aid of this realization.

II. DESCRIPTION OF THE REPRESENTATION

For convenience, the word "representation" is used here for the *abstract* representation (equivalence class of representations). The word "realization" will be employed exclusively for a *concrete* representation. In the present section, we are only interested in abstract aspects of the representation.

With the physicists definition, the Lie algebra of SL(2,R) is made of traceless pure-imaginary 2×2 matrices. We choose the three following elements as a basis:

$$J=-\tfrac{1}{2}\sigma_2,\quad K=-(i/2)\sigma_3,\quad L=(i/2)\sigma_1,$$

where the matrices σ_i are the standard Pauli matrices. We have the following commutators:

$$[J,K]=iL,\quad [J,L]=-iK,\quad [K,L]=-iJ. \tag{1}$$

Note that J generates a group isomorphic to U(1); K and L generate groups isomorphic to $\mathbb{R}$. In this two-dimensional representation, J has $\{-\frac{1}{2},\frac{1}{2}\}$ as a spectrum but K and L have a pure imaginary spectrum: $\{-i/2,i/2\}$. We will be interested later on in another kind of basis, namely,

$$R=J+K,\quad S=J-K,\quad L, \tag{2}$$

where R and S only have 1 as an eigenvalue. The commutators are

$$[R,S]=-2iL,\quad [R,L]=-iR,\quad [S,L]=iS. \tag{3}$$

The representation $D_+(-\frac{1}{2})$ we are interested in is entirely defined with the aid of three operators J,K,L (denoted for convenience by the same letters) verifying the properties

$$J>0,\quad J^2-K^2-L^2=-\tfrac{1}{4}, \tag{4}$$

with the conditions that it is irreducible and unitary. In the present section, we do not give an explicit construction of the representation Hilbert space. Our aim is to obtain properties common to all realizations of this representation. We say that the spin has the value $-\frac{1}{2}$ because the relation $j(j+1)=-\frac{1}{4}$ implies $j=-\frac{1}{2}$. We note that this number is the only root of that equation. This is probably why the representation presents interesting fact. [We note that the mapping $j\to -j-1$ does not change the value $j(j+1)$ and change the sign of the "dimension" of the representation $2j+1$.]

The elements of the Lie algebra are of the form

$$X=aJ+bK+cL=aJ+\tfrac{1}{2}(\beta K_-+\beta^* K_+), \tag{5}$$

with

$$K_\pm=K\pm iL,\quad \beta=b+ic. \tag{6}$$

We define the Killing form as

$$(X,X)=a^2-b^2-c^2=a^2-|\beta|^2. \tag{7}$$

The SL(2,R) group acts on the Lie algebra in preserving this quadratic form. One recognizes the action of the three-dimensional Lorentz group (see Appendix A). If it is positive (resp. negative, zero), the (nonzero) element will be said to be timelike (resp. spacelike, lightlike). It is clear that J is timelike and K and L are spacelike. Here, R and S are examples of lightlike elements (The interchange of the operators R and S is performed by the so-called Cartan linear mapping.)

Given two elements $X=aJ+bK+cL$ and $Y=a'J+b'K+c'L$, the Killing form provides us with a hyperbolic scalar product

$$(X,Y)=aa'-bb'-cc'. \tag{8}$$

The plane spanned by X and Y will be said to be

(i) of hyperbolic type if $(X,X)(Y,Y)-(X,Y)^2<0$,

(ii) spacelike if $(X,X)(Y,Y)-(X,Y)^2>0$ (Schwarz inequality),

(iii) tangent (to the light cone) if $(X,X)(Y,Y)-(X,Y)^2=0$.

We state, without proof, the following lemma.

Lemma 1: Set $[X,Y]=iZ$. The type of Z and the type of the plane Π spanned by X and Y are related as follows:

Π is hyperbolic, Z is spacelike,

Π is spacelike, Z is timelike,

Π is tangent, Z is lightlike,

and Z is orthogonal to Π.

This lemma is of general value since it describes a property of the abstract Lie algebra of SL(2,R). It is interesting to emphasize the following interesting fact of the representation we are interested in: The categories of timelike and lightlike elements split into positive and negative operators, corresponding to the classical distinction between future and past Lorentz vectors (see Fig. 1). This last property is shared by all unitary irreducible representations of the class $D_+(-\frac{1}{2}+\epsilon)$ of the universal covering group of SL(2,R). (In all these representations (belonging to the so-called discrete series), investigated in Sec. IX, J is positive.)

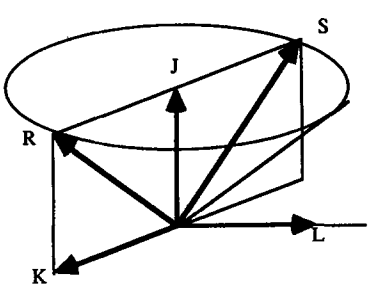


FIG. 1. J is timelike positive, R and S are lightlike positive, and K and L are spacelike.

Theorem 1: J has $N + \frac{1}{2}$ as a spectrum. Moreover, with standard Dirac notation,

$$J|n\rangle = (n+\tfrac{1}{2})|n\rangle, \quad n = 0,1,2,..., \tag{9a}$$

$$K_+|n\rangle = (n+1)|n+1\rangle, \tag{9b}$$

$$K_-|n\rangle = n|n-1\rangle, \tag{9c}$$

$$|n\rangle = (1/n!)K_+^n|0\rangle, \tag{10}$$

Remark (*oscillator realization*): Before giving the proof of Theorem 1, we emphasize that the spectrum of J is the one of the Hamiltonian of the harmonic oscillator. We readily see that there exists a realization in terms of annihilation and creation operators a and a^*. One has

$$J = a^*a + \tfrac{1}{2}, \quad K_- = \sqrt{a^*a+1}\,a, \quad K_+ = a^*\sqrt{a^*a+1}. \tag{11a}$$

These equations can be solved as follows:

$$a = (J+\tfrac{1}{2})^{-1/2}K_-, \quad a^* = K_+(J+\tfrac{1}{2})^{-1/2}. \tag{11b}$$

We will come back to these operators in Sec. VII.

Proof of Theorem 1: From Eqs. (4), we get

$$K_+K_- = J^2 - J + \tfrac{1}{4}, \quad K_-K_+ = J^2 + J + \tfrac{1}{4}. \tag{12}$$

Let $|n\rangle$ be a normed eigenvector of J with $n + \frac{1}{2}$ as an eigenvalue (we do not assume that n is an integer). From the commutators

$$[J, K_\pm] = \pm K_\pm,$$

it is easy to see that $K_+|n\rangle$ (resp. $K_-|n\rangle$) is an eigenvector of J with eigenvalue $n + \frac{3}{2}$ (resp. $n - \frac{1}{2}$). From Eqs. (12), we obtain the conditions

$$\|K_+|n\rangle\|^2 = \langle n|K_-K_+|n\rangle = (n+\tfrac{1}{2})^2 + (n+\tfrac{1}{2}) + \tfrac{1}{4} = (n+1)^2, \tag{13a}$$

$$\|K_-|n\rangle\|^2 = \langle n|K_+K_-|n\rangle = (n+\tfrac{1}{2})^2 - (n+\tfrac{1}{2}) + \tfrac{1}{4} = n^2. \tag{13b}$$

According to assumption (4), we cannot obtain, by an iterated action of K_-, a negative eigenvalue of J. This implies that for some value of n, $K_-|n\rangle$ is the null vector. This means that n does take the value 0 and, consequently, our statement about the spectrum of J is proved. The irreducibility condition implies that the vectors $|n\rangle$ span the whole Hilbert space. They form an orthogonal basis. Equations (10)–(13) are readily obtained. ■

Corollary: Any timelike operator $X = aJ + bK + cL$ of the Lie algebra has a discrete spectrum and has $\mathrm{sgn}(a)\sqrt{(X,X)}\,(n+\frac{1}{2})$ as eigenvalues (with $n = 0,1,2,...$).

Proof: It is well known that every Lorentz transformation maps a timelike vector on a timelike vector with the same values of Δ and $\mathrm{sgn}(a)$ and that such elements of the Lie algebra are conjugate. Therefore they have the same spectrum. ■

The element J provided us with a natural basis in the representation space. It is natural to look for the "eigenvectors" of an arbitrary element $X = aJ + bK + cL$ of the Lie algebra and to express them with the aid of such a basis. For that purpose, we write

$$X|x\rangle = x|x\rangle, \tag{14}$$

with

$$|x\rangle = \sum_{n=0}^{\infty} p_n(x)|n\rangle, \tag{15}$$

where $|x\rangle$ is a unit vector of the Hilbert space iff X is timelike. Otherwise, it is an "improper element" defined up to a factor.

For obvious reasons, we assume $\beta = b + ic \neq 0$. We then have the following theorem.

Theorem 2: If $p_0(x)$ is a constant, the p_n's are orthogonal polynomials in x. They are essentially:

(i) the Laguerre polynomials if $|a| = |\beta|$ (lightlike elements). As special cases, the operators R and S of Eq. (2) are associated, respectively, with the polynomials

$$(-)^n L_n(2x), \quad \text{and} \quad L_n(2x);$$

(ii) the Meixner polynomials of the first kind if $|a| > |\beta|$ (timelike elements). In particular, if $(X,X) = \mathrm{sgn}(a) = 1$, we have $x = n + \frac{1}{2}$, with $n = 0,1,2,...$;

(iii) the Meixner polynomials of the second kind if $|a| < |\beta|$. As special cases, the operators K and L are associated, respectively, with the Hardy–Pollaczek polynomials $P_n(x)$ and $(-i)^n P_n(x)$, where

$$(1-iz)^{-1/2+ix}(1+iz)^{-1/2-ix} = \sum_{n=0}^{\infty} P_n(x)z^n. \tag{16}$$

Proof: Equations (14) and (15) provide us with the infinite set of equations

$$\begin{aligned}
&ap_0(x) + \beta p_1(x) = 2xp_0(x),\\
&\beta^* p_0(x) + 3ap_1(x) + 2\beta p_2(x) = 2xp_1(x),\\
&2\beta^* p_1(x) + 5ap_2(x) + 3\beta p_3(x) = 2xp_2(x),\\
&\vdots\\
&n\beta^* p_{n-1}(x) + (2n+1)ap_n(x) + (n+1)\beta p_{n+1}(x) = 2xp_n(x).\\
&\vdots
\end{aligned} \tag{17}$$

It is a remarkable fact that if $p_0(x)$ is zero, all the p_n's are zero. Therefore, we are sure that, whatever are X and its "eigenvector" $|x\rangle$, one has $\langle 0|x\rangle \neq 0$. It follows that, if $p_0(x)$ is a constant, p_n is a polynomial of degree n in x. Equation (17) shows us that they are orthogonal polynomials. It is more usual, in this context, to write the recurrence relation (17) in the following way:

$$(n+1)\beta p_{n+1}(x) = (2x - (2n+1)a)p_n(x) - n\beta^* p_{n-1}(x). \tag{17'}$$

The case $\beta = 1$, $a = 0$ corresponds to the Hardy–Pollaczek polynomials. Let us set

$$Q_n(x) = \left(\frac{\beta}{|\beta|}\right)^n p_n\left(-\frac{|\beta|}{2}x\right), \quad \sigma = \frac{a}{|\beta|}.$$

The recurrence relation becomes

$$(n+1)Q_{n+1}(x) = [-x-(2n+1)\sigma]Q_n(x) - nQ_{n-1}(x). \quad (18)$$

For $\sigma = -1$ (resp. $\sigma = 1$), one recognizes the recurrence formula of Laguerre's polynomials $L_n(x)$ (resp. $(-)^n L_n(-x)$). For $|\sigma| > 1$ (resp. $|\sigma| < 1$), we obtain special cases of the Meixner polynomials of the first (resp. second) kind. Thus, the representation $D_+(-\frac{1}{2})$ of SL(2,R) gives a nice interpretation of the well-known relationship of the Laguerre and Meixner polynomials: The set of Laguerre polynomials is "between" the two kinds of Meixner's sets (lightlike vectors separate spacelike vectors from timelike vectors).

It is a simple matter to derive a differential equation obeyed by the generating function

$$F(x,z) = \sum_{n=0}^{\infty} Q_n(x) z^n,$$

$$(1 + 2\sigma z + z^2)\frac{\partial F}{dz} + (x + \sigma + z)F = 0.$$

We see that the three kinds of polynomials correspond to the three following situations, regarding the polynomial in z: $1 + 2\sigma z + z^2$:

(i) Two real roots: The $Q_n(x)$ are Meixner's polynomials of the first kind generated by

$$F(x,z) = [1 + (\sigma - \sqrt{\sigma^2 - 1})z]^{-1/2-\gamma} \times [1 + (\sigma + \sqrt{\sigma^2 - 1})z]^{-1/2+\gamma},$$

with

$$\gamma = x/2\sqrt{\sigma^2 - 1}.$$

It is important to underline that the only eigenvalues of an operator satisfying $a^2 - |\beta|^2 = 1$ and $a > 0$ are of the form $x = n + \frac{1}{2}$. The corresponding eigenvectors are vectors of the Hilbert space.

(ii) Two equal roots ($\sigma = \pm 1$): Laguerre polynomials generated by

$$F(x,z) = [1/(-\sigma - z)]\exp[-\sigma x z/(\sigma + z)].$$

(iii) Two imaginary conjugate roots: Meixner's polynomials of the second kind, generated by

$$F(x,z) = [1 + (\sigma - i\sqrt{1 - \sigma^2})z]^{-1/2 - i\gamma'} \times [1 + (\sigma + i\sqrt{1 - \sigma^2})z]^{-1/2 + i\gamma'},$$

with

$$\gamma' = x/2\sqrt{1 - \sigma^2}.$$

Then we have proved Theorem 2. ∎

It is a simple matter to write down the generating functions of the polynomials associated with the operators R, S, K, and L. We give in a Résumé the main properties concerning the eigenvectors of these operators.

Résumé: If P_n and L_n denote, respectively, the Hardy–Pollaczek and Laguerre polynomials, we have, for the operators J, K, L, R, S:

$$J|n\rangle = (n + \tfrac{1}{2})|n\rangle,$$

$$K|k\rangle = k|k\rangle, \quad |k\rangle = \sum P_n(k)|n\rangle, \quad (19a)$$

$$L|\lambda\rangle = \lambda|\lambda\rangle, \quad |\lambda\rangle = \sum (-i)^n P_n(\lambda)|n\rangle, \quad (19b)$$

$$R|r\rangle = r|r\rangle, \quad |r\rangle = \sum (-)^n \sqrt{2}e^{-r} L_n(2r)|n\rangle, \quad (19c)$$

$$S|s\rangle = s|s\rangle, \quad |s\rangle = \sum \sqrt{2}e^{-s} L_n(2s)|n\rangle. \quad (19d)$$

Here, distinct arbitrary functions have been chosen for p_0. We will show later on why it is "natural" to choose $p_0(r) = \sqrt{2}e^{-r}$ and $p_0(s) = \sqrt{2}e^{-s}$ but $p_0(k)$ and $p_0(\lambda) = 1$ as normalization factors. In the formulas, the only kets that are elements of the Hilbert space are the $|n\rangle$'s. They are vectors of norm 1.

Theorem 3: The operators R and S are positive.

Proof: Let $|\psi\rangle$ be an arbitrary unit vector. We can always write it in the form

$$|\psi\rangle = \sum_{n=0}^{\infty} \psi_n |n\rangle.$$

We get

$$\langle\psi|(2J \pm 2K)|\psi\rangle = \sum_{n=0}^{\infty} \psi_n^* \sum_{m=0}^{\infty} \psi_m \langle n|(2J \pm 2K)|m\rangle$$

$$= \sum [(2n+1)|\psi_n|^2 \pm n(\psi_n^* \psi_{n-1} + \psi_{n-1}^* \psi_n)]$$

$$= \sum n|\psi_n \pm \psi_{n-1}|^2,$$

which is always positive. ∎

Corollary: Lightlike elements of the Lie algebra are either positive or negative.

The proof follows from the transitive action of SL(2,R) on future (resp. past) lightlike vectors. ∎

It is interesting to underline that the Laguerre polynomials are also involved in other relations in the next theorem where a kind of duality is shown:

(i) between $-K_+$ and $2R$,

(ii) between K_+ and $2S$.

Theorem 4: We have the following formulas:

$$(1/n!)(2R)^n|0\rangle = L_n(-K_+)|0\rangle,$$
$$(-)^n|n\rangle = (1/n!)(-K_+)^n|0\rangle = L_n(2R)|0\rangle, \quad (20a)$$

$$(1/n!)(2S)^n|0\rangle = L_n(K_+)|0\rangle,$$
$$|n\rangle = (1/n!)K_+^n|0\rangle = L_n(2S)|0\rangle. \quad (20b)$$

Proof: First, we need to prove the relations

$$\frac{1}{N!}(2R)^N|0\rangle = \sum_{n=0}^{N} \binom{N}{n}|n\rangle,$$

$$\frac{1}{N!}(2S)^N|0\rangle = \sum_{n=0}^{N} \binom{N}{n}(-)^n|n\rangle. \quad (21)$$

This can be shown recurrently, with the aid of

$$2R|n\rangle = (K_+ + K_- + 2J)|n\rangle$$
$$= (n+1)|n+1\rangle + n|n-1\rangle + (2n+1)|n\rangle,$$
$$2S|n\rangle = (-K_+ - K_- + 2J)|n\rangle$$
$$= -(n+1)|n+1\rangle - n|n-1\rangle + (2n+1)|n\rangle.$$

Then, we use (10) and the Laguerre formula

$$L_N(x) = \sum_{n=0}^{N} \frac{N!}{n!n!(N-n)!}(-x)^n, \tag{22}$$

to obtain the first part of Eqs. (20). The second part is obtained by use of the generating function

$$\sum L_n(x)u^n = \frac{1}{1-u}\exp\left(-x\frac{u}{1-u}\right). \tag{23}$$

One gets

$$\exp(2uR)|0\rangle = \sum \frac{u^N(2R)^N}{N!}|0\rangle$$
$$= \sum L_N(-K_+)u^N|0\rangle$$
$$= \frac{1}{1-u}\exp\left(\frac{u}{1-u}K_+\right)|0\rangle,$$
$$\exp(2uS)|0\rangle = \sum \frac{u^N(2S)^N}{N!}|0\rangle$$
$$= \sum L_N(K_+)u^N|0\rangle$$
$$= \frac{1}{1-u}\exp\left(-\frac{u}{1-u}K_+\right)|0\rangle.$$

If we set $t = -u/(1-u)$, these relations become

$$[1/(1-t)]\exp(-[t/(1-t)]2R)|0\rangle$$
$$= \exp(-tK_+)|0\rangle, \tag{24a}$$
$$[1/(1-t)]\exp(-[t/(1-t)]2S)|0\rangle = \exp(tK_+)|0\rangle, \tag{24b}$$

and, using again (23), we obtain, in identifying the terms t^n of both sides, the second part of Eqs. (20). ■

Corollary:

$$R^N|0\rangle = \sum_{n=0}^{N} \frac{N!N!}{2^N n!(N-n)!}|n\rangle, \tag{25a}$$
$$S^N|0\rangle = \sum_{n=0}^{N} (-)^n \frac{N!N!}{2^N n!(N-n)!}|n\rangle. \tag{25b}$$

Proof: These relations are a direct consequence of Eqs. (10), (20), and (22). ■

III. THE HARDY–POLLACZEK POLYNOMIALS

The oldest reference we know about these polynomials is the one of Pidduck,[6] who used them in order to solve a physical problem, but did not study their properties. A slightly different form of them was investigated by Hardy.[3] They are the ones that are given here the name of Hardy–Pollaczek, and are defined by Eq. (16); in associating the name of Pollaczek, we are referring to the class of orthogonal polynomials studied by this last author.[4] The whole class is shown to play a role in the representations of the discrete series (Sec. IX).

Theorem 5: The Hardy–Pollaczek polynomials [our polynomials are related with the F_n's of Itzykson by the formula: $P_n(\lambda) = i^n F_n(\lambda)$] satisfy the orthogonality relation

$$\int_{-\infty}^{+\infty} \frac{P_n(\lambda)P_m(\lambda)}{\cosh(\pi\lambda)}d\lambda = \delta_{nm}. \tag{26}$$

Proof: Let us set $z = i\tanh(\pi u)$ in Eq. (16). We get easily

$$\exp(2i\pi u\lambda) = \sum_{n=0}^{\infty} P_n(\lambda)\frac{(i\tanh(\pi u))^n}{\cosh(\pi u)}. \tag{27}$$

It is well known that the function sech is its own Fourier transform. More precisely, we have

$$\frac{1}{\cosh(\pi u)} = \int_{-\infty}^{+\infty} \frac{e^{2i\pi u\lambda}}{\cosh(\pi\lambda)}d\lambda. \tag{28}$$

From these two formulas we obtain directly the orthogonality relation of the Hardy–Pollaczek polynomials. For this purpose, it suffices to expand the expression $e^{2i\pi(u-v)\lambda}$ in the integral

$$\int_{-\infty}^{+\infty} \frac{e^{2i\pi(u-v)\lambda}}{\cosh(\pi\lambda)}d\lambda,$$

according to (27), to use the relation

$$\frac{1}{\cosh(\pi(u-v))}$$
$$= \frac{1}{\cosh(\pi u)\cosh(\pi v)(1-\tanh(\pi u)\tanh(\pi v))}$$
$$= \frac{1}{\cosh(\pi u)\cosh(\pi v)}\sum_{n=0}^{\infty}(\tanh(\pi u)\tanh(\pi v))^n,$$

and to identify the terms in $(\tanh(\pi u))^n(\tanh(\pi v))^m$. ■

A. Consequences

(i) Symmetry properties of the H–P polynomials. By taking the complex conjugate of (27) and changing u into $-u$, we see that the polynomials are real. If we change the signs of u and λ, we obtain the property

$$P_n(-\lambda) = (-)^n P_n(\lambda). \tag{29}$$

(ii) Normalization of the $|k\rangle$'s. Equation (26) can be written as

$$\int_{-\infty}^{+\infty} f_n^*(\lambda)f_m(\lambda)d\lambda = \delta_{nm}, \tag{26'}$$

where

$$f_n(\lambda) = \Gamma(\tfrac{1}{2} + i\lambda)P_n(\lambda).$$

Here we used the properties of the Γ function

$$\Gamma(\tfrac{1}{2}+i\lambda)\Gamma(\tfrac{1}{2}-i\lambda) = 1/\cosh(\pi\lambda),$$
$$\Gamma(\tfrac{1}{2}+i\lambda)^* = \Gamma(\tfrac{1}{2}-i\lambda).$$

Obviously, the f_n's satisfy the recurrence relation satisfied by the H–P polynomials. If we define, instead of (19a),

$$|k\rangle = \sum f_n(k)|n\rangle = \sum \Gamma(\tfrac{1}{2}+i\lambda)P_n(\lambda)|n\rangle,$$

the Dirac formalism leads to

$$\sum_n f_n(k)^* f_n(k') = \delta(k-k'),$$

and

$$|n\rangle = \int_{-\infty}^{+\infty} \Gamma(\tfrac{1}{2} - ik) P_n(k) |k\rangle dk. \tag{30}$$

Theorem 6: The moments of the Hardy–Pollaczek polynomials are $|E_n|/2^n$, where E_n is the nth Euler number defined by

$$\frac{1}{\cosh(\pi u)} = \sum_{n=0}^{\infty} E_n \frac{(\pi u)^n}{n!}.$$

Proof: From (28) and (30), we readily get

$$\int_{-\infty}^{+\infty} \sum_{n=0}^{\infty} \frac{(2i\pi u\lambda)^n}{n! \cosh(\pi\lambda)} d\lambda = \sum_{n=0}^{\infty} E_n \frac{(\pi u)^n}{n!},$$

and, by identification,

$$\int_{-\infty}^{+\infty} \frac{\lambda^n}{\cosh(\pi\lambda)} d\lambda = \frac{|E_n|}{2^n}.$$

It is clear that the odd Euler numbers vanish. The even ones have alternate signs ($E_0 = 1, E_2 = -1, E_4 = 5, E_6 = -61, E_8 = 1385, E_{10} = -50521, \ldots$). ■

We note the following property:

$$E_{2n} = (-)^n \frac{2^{2n+2}(2n)!}{\pi^{2n+1}} \left(1 - \frac{1}{3^{2n+1}} + \frac{1}{5^{2n+1}} - \cdots\right),$$

which is obtained with the aid of the expansion $1/2\cosh x = e^{-x} - e^{-3x} + e^{-5x} - \cdots$.

B. The H-P polynomials as characteristic polynomials

It is a simple matter to prove, with the aid of the recurrence relation (17a),

$$(n+1)P_n(\lambda) = 2\lambda P_n(\lambda) - nP_{n-1}(\lambda),$$

that

$$n! P_n(\lambda) = \det \begin{pmatrix} 2\lambda & 1 & 0 & 0 & 0\cdots \\ 1 & 2\lambda & 2 & 0 & 0\cdots \\ 0 & 2 & 2\lambda & 3 & 0\cdots \\ 0 & 0 & 3 & 2\lambda & 4\cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

where the matrix is $n \times n$.

C. Pidduck polynomials

They are the polynomials $\mu_n(x)$ related with the Hardy–Pollaczek ones by the following formula:

$$P_n(\lambda) = (-i)^n \mu_n(-\tfrac{1}{2} + i\lambda). \tag{31}$$

They have interesting properties that are less transparent under the form of the Hardy–Pollaczek polynomials. Their generating functions are

$$\frac{(1+u)^x}{(1-u)^{x+1}} = \sum_{n=0}^{\infty} \mu_n(x) u^n, \tag{32a}$$

$$e^t M(-x, 1, -2t) = \sum_0^{\infty} \mu_n(x) \frac{t^n}{n!}, \tag{32b}$$

where M denotes the confluent hypergeometric function.

We note that the left-hand side of Eq. (32a) is invariant under the change

$$u \to -u, \quad x \to -x-1.$$

It follows that (29) now reads:

$$\mu_n(-x-1) = (-)^n \mu_n(x).$$

The Pidduck polynomials can be written with the aid of the hypergeometric function as

$$\mu_n(x) = F(-n, -x; 1; 2) = \sum_{m=0}^{m=n} \frac{n! 2^m}{m! m! (n-m)!} x(x-1)(x-2) \cdots (x-m+1) \tag{33}$$

or, shortly,

$$\mu_n(x) = \sum_{m=0}^{m=n} 2^m \binom{n}{m}\binom{x}{m}.$$

The proof is easy. We only have to write

$$\frac{(1+u)^x}{(1-u)^{x+1}} = \frac{1}{1-u}\left(1 + \frac{2u}{1-u}\right)^x = \sum_{m=0}^{m=\infty} \frac{x(x-1)(x-2)\cdots(x-m+1)}{m!} \times \frac{(2u)^m}{(1-u)^{m+1}},$$

and to expand $1/(1-u)^{m+1}$.

Equation (17a) provides us with the recurrence formula obeyed by the Pidduck polynomials:

$$(n+1)\mu_{n+1}(x) = (2x+1)\mu_n(x) + n\mu_{n-1}(x). \tag{34}$$

The symmetry between n and x in $F(-n, -x; 1; 2)$ permits us to write immediately

$$(x+1)\mu_n(x+1) = (2n+1)\mu_n(x) + x\mu_n(x-1), \tag{35a}$$

or, equivalently,

$$\frac{1}{2}\left[\exp\left(\frac{d}{dx}\right) x - x \exp\left(-\frac{d}{dx}\right)\right] \mu_n(x) = (n + \tfrac{1}{2})\mu_n(x). \tag{35b}$$

It follows that these polynomials are the eigenfunctions of an operator with $N + \frac{1}{2}$ as a spectrum. We are going to show that this operator is involved in the representation of $SL(2,R)$ we are interested in.

IV. A REALIZATION WITH THE H–P POLYNOMIALS AS EIGENFUNCTIONS OF J (L IS DIAGONAL)

The realization we are going to describe has been obtained by Itzykson[2] through a Mellin transform from the z realization described later on. The reason is that the z realization is one of the most "natural" when we are interested in group transformations of $SL(2,R)$. However, the one involving the Hardy–Pollaczek polynomials seems more interesting because it has nice properties related with classical special functions.

We define the realization of the Lie algebra by

$$R=\exp\left(\frac{d}{dx}\right)x,\quad S=-x\exp\left(-\frac{d}{dx}\right),$$

$$L=-i(x+\tfrac{1}{2}). \tag{36a}$$

[These operators give simple ways of writing some relations involving the hypergeometric functions, the factorial polynomials, and the Pidduck polynomials (see Appendix C).]

According to (35b), the eigenfunctions of J are the Pidduck polynomials. If we set

$$x=-\tfrac{1}{2}+i\lambda,$$

we have

$$R=\exp\left(-i\frac{d}{d\lambda}\right)\left(-\frac{1}{2}+i\lambda\right),$$

$$S=\left(\frac{1}{2}-i\lambda\right)\exp\left(i\frac{d}{d\lambda}\right),\quad L=\lambda. \tag{36b}$$

We note that R becomes S when we replace x by $-x-1$ or λ by $-\lambda$. This transformation is the Cartan mapping since it maps J on itself and change the operators K and L into $-K$ and $-L$, respectively.

As shown by Itzykson, the H–P polynomials span the Hilbert space of the square-integrable functions on the real line with measure $d\lambda/\cosh(\pi\lambda)$. It is interesting to mention that the physicist would have defined abstractly, from Eq. (19b), the Hardy–Pollaczek polynomials as the $P_n(\lambda)=(-i)^n\langle\lambda|n\rangle=i^n\langle n|\lambda\rangle$. For that reason we will refer to this realization as the λ realization, the one which "diagonalizes" the operator L.

A. Eigenfunctions of *R* and *S*

If we denote by $[r^x/\Gamma(x+1)]\rho_r(x)$ an eigenfunction of R associated with the eigenvalue r, we readily see that $\rho_r(x)$ must be a periodic function of x of period 1. It follows from Eq. (19c) that

$$\frac{r^x}{\Gamma(x+1)}\rho_r(x)=\sum(-)^nL_n(2r)\mu_n(x),$$

where $\rho_r(x)$ is a periodic function of x, of period 1. In fact, we have the following theorem.

Theorem 7: The eigenfunction of R (resp. S) with eigenvalue r (resp. s) is

$$\sum(-)^ne^{-r}L_n(2r)\mu_n(x)=\frac{r^x}{2\Gamma(x+1)}, \tag{37a}$$

$$\sum e^{-s}L_n(2s)\mu_n(x)=\frac{1}{2\Gamma(-x)s^{x+1}}=-\frac{\sin\pi x\Gamma(x+1)}{2\pi s^{x+1}}. \tag{37b}$$

Proof: The fact that they are eigenfunctions of R (resp. S) is easy to check. Since (37b) follows from (37a), in replacing x by $-x-1$ and r by s and using the identity

$$\Gamma(-x)\Gamma(x+1)=-/\sin\pi x,$$

we only have to prove (37a). First, we note that, if $\operatorname{Re}p>-1$,

$$\int_0^\infty e^{-(p+1)t}L_n(2t)dt=\frac{(p-1)^n}{(p+1)^{n+1}}.$$

This is a direct consequence of (22):

$$\sum_{k=0}^n\frac{n!(-2)^k}{(n-k)!k!k!}\int_0^\infty e^{-(p+1)t}t^k\,dt=\sum_{k=0}^n\binom{n}{k}\frac{(-2)^k}{(p+1)^{k+1}}=\frac{1}{p+1}\left(1-\frac{2}{p+1}\right)^n.$$

From (32a) we get

$$p^{-z-1}=2\sum(-)^n\mu_n(z)\frac{(p-1)^n}{(p+1)^{n+1}}.$$

Since p^{-z-1} is the Laplace transform of $t^z/\Gamma(z+1)$, we readily obtain the proof of (37a). ■

Remark: It is easy to show that $\mu_{2n}(-\frac{1}{2})=(2n-1)!!/(2n)!!$ and $\mu_{2n+1}(-\frac{1}{2})=0$. For $x=-\frac{1}{2}$, the formulas (37a) and (37b) give both the following property of the Laguerre polynomials:

$$\sum L_{2n}(2s)\frac{(2n-1)!!}{(2n)!!}=\frac{e^s}{2\sqrt{\pi s}}.$$

V. A REALIZATION INVOLVING THE LAGUERRE POLYNOMIALS (*R* IS DIAGONAL)

We note the relation

$$\int_0^\infty e^{-tu}u^xL_n(u)du=\Gamma(x+1)t^{-x-1}F(-n,x+1,1;1/t),\quad \operatorname{Re}x>0,$$

and, since $\mu_n(x)=(-)^nF(-n,x+1,1;2)$, we easily get

$$\mu_n(x)=\frac{(-)^n}{\Gamma(x+1)}\int_0^\infty e^{-r}r^xL_n(2r)dr. \tag{38a}$$

The inverse formula of this Mellin transformation is

$$(-)^ne^{-r}L_n(2r)=\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty}\Gamma(x+1)\mu_n(x)r^{-x-1}\,dx,\quad \sigma\ \text{real}>-1. \tag{38b}$$

This formula permits us to deduce the action of the Lie algebra on the Laguerre functions $\sqrt{2}e^{-r}L_n(2r)$. For instance, according to (36a), R transforms $\mu_n(x)$ into $(x+1)\mu_n(x+1)$; the integrand of (38b) becomes

$$\Gamma(x+1)(x+1)\mu_n(x+1)r^{-x-1}=r\Gamma(x+2)\mu_n(x+1)r^{-x-2},$$

and the integral is multiplied by r.

This indicates the way we are able to obtain the realization that diagonalizes the operator R and that will be referred to as the r realization. It follows that we can interpret the function $\sqrt{2}(-)^ne^{-r}L_n(2r)$ as the scalar product $\langle r|n\rangle$ and $\sqrt{2}e^{-s}L_n(2s)$ as the scalar product $\langle s|n\rangle$. From the differential equations obeyed by the Laguerre polynomials, it is an easy task to show that these functions are eigenfunctions of the operator

$$J=-\frac{r}{2}\frac{d^2}{dr^2}-\frac{1}{2}\frac{d}{dr}+\frac{r}{2}.$$

Finally, we get, for the r realization (This realization is used in Ref. 7.)

$$R=r,\quad S=-r\frac{d^2}{dr^2}-\frac{d}{dr},\quad L=i\left(r\frac{d}{dr}+\frac{1}{2}\right). \tag{39}$$

The Laguerre functions are orthogonal according to the following formula:

$$2\int_0^\infty e^{-2r}(-)^n L_n(2r)(-)^m L_m(2r)dr=\delta_{nm}.$$

The choice of our eigenfunctions of J corresponds to $p_0(r)=\sqrt{2}e^{-r}$ in Eq. (19c). Similarly, we take $p_0(s)=\sqrt{2}e^{-s}$ in Eq. (19d). It follows that the eigenfunctions of the operator S, in the r realization are given by[8]

$$\sum_{n=0}^{\infty}2e^{-(r+s)}(-)^n L_n(2s)L_n(2r)=J_0(2\sqrt{sr}),$$

where J_0 is the Bessel function. With the physicist notation, the function $J_0(2\sqrt{sr})$ represents the scalar products $\langle r|s\rangle$ and $\langle s|r\rangle$.

Other relations can be found. First, the eigenfunctions (the scalar products $\langle r|\lambda\rangle$) of the operator L are

$$(1/\sqrt{2})/\left[r^{-1/2-i\lambda}/\Gamma(\tfrac{1}{2}-i\lambda)\right].$$

The normalization [this corresponds to $p_0(r)=1$ in Eq. (19c)] is such that the eigenfunction (associated with the eigenvalue r_0) of the operator R is the distribution

$$(1/\sqrt{r_0 r})\delta(\mathrm{Log}\, r-\mathrm{Log}\, r_0),$$

where δ denotes the Dirac distribution.

VI. THE z AND θ REALIZATIONS

In the book by Gelfand *et al.*,[9] all representations of $SL(2,R)$ are built with the aid of the space of homogeneous infinitely differentiable functions of two real variables x and y [except at point (0,0)], satisfying the property

$$f(ax,ay)=|a|^{s-1}(\mathrm{sgn}\, a)^{\epsilon}f(x,y),$$

where s is an arbitrary complex number and $\epsilon=0$ or 1.

In our case, $s=0$ and $\epsilon=1$; that is

$$f(ax,ay)=(1/a)f(x,y),$$

the representation is irreducible. The homogeneity of the functions can be expressed in requiring them to be eigenfunctions of the operator $x\,\partial/\partial x+y\,\partial/\partial y$ with eigenvalue -1. The space is spanned by the functions $(x+iy)^n/(x-iy)^{n+1}$ with $n=0,1,2,\ldots$. A physicist will note immediately that they are the eigenfunctions of the operator $J=-(i/2)(x(\partial/\partial y)-y(\partial/\partial x))$ with eigenvalues $n+\frac{1}{2}$. One is tempted to replace $x+iy$ by ζ [the homogeneity condition reads: $\zeta(\partial/d\zeta)+\zeta^*(\partial/\partial\zeta^*)=-1$] or to replace x by $\rho\cos\theta$ and y by $\rho\sin\theta$ [that permits to eliminate one variable, since $\rho(\partial/\partial\rho)$ equals -1]. One could also suggest to use the variables x and $z=y/x$ (we are left with the only variable z). One gets for all these cases

Operator J	*Eigenfunctions*
$-\frac{i}{2}\left(x\frac{\partial}{dy}-y\frac{\partial}{dx}\right)$,	$\frac{(x+iy)^n}{(x-iy)^{n+1}}$,
$\frac{1}{2}\left(\zeta\frac{\partial}{\partial\zeta}-\zeta^*\frac{\partial}{\partial\zeta^*}\right)$,	$\frac{\zeta^n}{\zeta^{*n+1}}$,
$-\frac{i}{2}\frac{\partial}{\partial\theta}$,	$\frac{1}{\rho}e^{i(2n+1)\theta}$,
$\frac{i}{2}(z^2+1)\frac{d}{dz}+\frac{i}{2}z$,	$\frac{(z+i)^n}{(z-i)^{n+1}}$.

Here we referred to the two last realizations as the θ and the z realizations. It is a Mellin transformation on the z realization which leads Itzykson to the λ realization involving the Hardy–Pollaczek polynomials. The operators R, S, and L read as follows in the z realization:

$$R=i\frac{d}{dz},\quad S=i\left(z^2\frac{d}{dz}+z\right),\quad L=-i\left(z\frac{d}{dz}+\frac{1}{2}\right). \tag{40}$$

We know that L and R are strictly positive operators. It follows that the number z corresponds to the eigenvalue of an operator that can be written as the quotient of L and R. More precisely, this operator is

$$Z=-(LR^{-1}+R^{-1}L)/2. \tag{41}$$

In the θ realization, we have

$$J=-\frac{i}{2}\frac{\partial}{\partial\theta},\quad K=-\frac{i}{2}\left(-\sin 2\theta+\cos 2\theta\frac{\partial}{\partial\theta}\right),$$
$$L=-\frac{i}{2}\left(\cos 2\theta+\sin 2\theta\frac{\partial}{\partial\theta}\right). \tag{42}$$

As it could be expected, the variable θ being an angle, cannot be associated with an operator of the enveloping Lie algebra. Let us examine the operators $e^{\pm 2i\theta}$. The operator $e^{2i\theta}$ can be expressed in terms of J and K_+. One obtains

$$e^{2i\theta}=K_+(J+\tfrac{1}{2})^{-1},\quad e^{2i\theta}|n\rangle=|n+1\rangle.$$

The operator $e^{-2i\theta}$ is such that $K_-=e^{-2i\theta}(J-\frac{1}{2})$. We cannot solve this equation since the operator $J-\frac{1}{2}$ has zero as an eigenvalue. We can only write $e^{-2i\theta}|n\rangle=|n-1\rangle$ for $n\neq 0$.

The various realizations we have studied permit us to classify relations between special functions. For instance, the relations expressing an eigenstate of R (or S) in the basis $|n\rangle$, namely,

$$|r\rangle=\sqrt{2}\sum(-)^n e^{-r}L_n(2r)|n\rangle,$$

$$|s\rangle=\sqrt{2}\sum e^{-s}L_n(2s)|n\rangle,$$

that reads in the x realization:

$$\sum\sqrt{2}(-)^n e^{-r}L_n(2r)\mu_n(x)=\frac{r^x}{\sqrt{2}\Gamma(x+1)}, \tag{37a}$$

$$\sum\sqrt{2}e^{-s}L_n(2s)\mu_n(x)=-\frac{\sin\pi x\Gamma(x+1)}{\sqrt{2}\pi s^{x+1}}, \tag{37b}$$

and reads in the θ realization:

$$\sum\sqrt{2}(-)^n e^{-r}L_n(2r)\frac{e^{i(2n+1)\theta}}{\sqrt{2}\pi}=\frac{e^{ir\mathrm{tg}\theta}}{2\cos\theta\sqrt{\pi}}, \tag{43a}$$

TABLE I. Some realizations of the Lie algebra.

	$SL(2,R)$ matrices	n (J diagonal)	λ (L diagonal)	x	r (R diagonal)	z $\left(\frac{LR^{-1}+R^{-1}L}{2}\text{diag}\right)$
R	$\begin{pmatrix} 0 & 0 \\ -i & 0 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & 0 & . \\ 1 & 3 & 2 & 0 & . \\ 0 & 2 & 5 & 3 & . \\ 0 & 0 & 3 & 7 & . \\ . & . & . & . & . \end{pmatrix}$	$\left(\frac{1}{2}+i\lambda\right)\exp\left(-i\frac{d}{d\lambda}\right)$	$\exp\left(\frac{d}{dx}\right)x$	r	$-i\frac{d}{dz}$
S	$\begin{pmatrix} 0 & i \\ 0 & 0 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} 1 & -1 & 0 & 0 & . \\ -1 & 3 & -2 & 0 & . \\ 0 & -2 & 5 & -3 & . \\ 0 & 0 & -3 & 7 & . \\ . & . & . & . & . \end{pmatrix}$	$\left(\frac{1}{2}-i\lambda\right)\exp\left(i\frac{d}{d\lambda}\right)$	$-x\exp\left(\frac{d}{dx}\right)$	$-r\frac{d^2}{dr^2}-\frac{d}{dr}$	$-i\left(z^2\frac{d}{dz}+z\right)$
L	$\frac{i}{2}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$		λ	$-i(x+\frac{1}{2})$	$i\left(r\frac{d}{dr}+\frac{1}{2}\right)$	$-i\left(z\frac{d}{dz}+\frac{1}{2}\right)$
J	$\frac{1}{2}\begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & 0 & . \\ 0 & 3 & 0 & 0 & . \\ 0 & 0 & 5 & 0 & . \\ 0 & 0 & 0 & 7 & . \\ . & . & . & . & . \end{pmatrix}$				

TABLE II. Eigenfunctions of R, S, L, J in some realizations.

	Spectrum	λ	x	r	z
R	$\mathbb{R}_+^*$	$\frac{r^{-1/2+i\lambda}e^r}{2\Gamma(1/2+i\lambda)}$	$\frac{r^x e^r}{2\Gamma(x+1)}$	$\frac{\delta(\mathrm{Log}(r/r_0))}{\sqrt{r_0 r}}$	$\exp(-irz)$
S	$\mathbb{R}_+^*$	$\frac{s^{-1/2-i\lambda}e^s}{2\Gamma(1/2-i\lambda)}$	$\frac{e^s \sin \pi x \Gamma(x+1)}{2\pi s^{x+1}}$	$J_0(2\sqrt{sr})$	$\frac{1}{z}\exp\left(\frac{is}{z}\right)$
L	$\mathbf{R}$	$\delta(\lambda-\lambda_0)$		$\frac{r^{-1/2-i\lambda_0}}{\sqrt{2}\Gamma(1/2-i\lambda_0)}$	$z^{1/2-i\lambda_0}$
J	$\mathbf{N}+\frac{1}{2}$	$P_n(\lambda)$	$\mu_n(x)$	$\sqrt{2}(-)^n e^{-r}L_n(2r)$	$(-)n\frac{1}{\sqrt{\pi}}\frac{(z+i)^n}{(z-i)^{n+1}}$

$$\sum \sqrt{2}e^{-s}L_n(2s)\frac{e^{i(2n+1)\theta}}{2\pi} = \frac{ie^{-is\mathrm{tg}\theta}}{2\sin\theta\sqrt{\pi}}. \tag{43b}$$

This similitude between different formulas is an example of the ones that are collected in Tables I–III.

VII. THE COHERENT STATE REALIZATION

Coherent states were defined for this group by Barut and Girardello.[10] Here we make use of the generalized definition of sets of coherent states given by Perelemov.[11] Then, we define a coherent state as a vector belonging to the orbit of the "ground state" $|0\rangle$ under the action of the SL(2,R) group. We have

$$J|0\rangle = \tfrac{1}{2}|0\rangle,$$

and, if $g\in$SL(2,R), g transforms $|0\rangle$ into an eigenvector of gJg^{-1} with the same eigenvalue:

$$(gJg^{-1})g|0\rangle = (aJ+\beta K_- + \beta^* K_+)g|0\rangle = \tfrac{1}{2}|0\rangle.$$

From the classification of elements of the Lie algebra [Eq. (7)], it follows that

$$a^2 - |\beta|^2 = 1.$$

We choose the following parametrization: $a=\cosh\phi$, $\beta = \sinh\phi\, e^{i\psi}$, with $\phi>0$. Bringing these expressions in Eq. (17a), with $x=\frac{1}{2}$, we get

$$p_n = (-)^n e^{-in\psi}\tanh^n(\phi/2)p_0,$$

and the normalization of the state $\Sigma_0^\infty |p_n|^2 = 1$ is obtained in choosing $p_0 = \cosh(\phi/2)$. Finally, the coherent states of the "ground-state" type are given by the formula

$$|\phi,\psi\rangle = \sum_0^\infty (-)^n e^{-in\psi}\frac{\tanh^n(\phi/2)}{\cosh(\phi/2)}|n\rangle. \tag{44}$$

It is easy to verify that these vectors are eigenvectors of the annihilation operator A defined by

$$A = (J+\tfrac{1}{2})^{-1}K_- \tag{45}$$

To check that statement, it suffices to use Eqs. (9)

$$A|n\rangle = (J+\tfrac{1}{2})^{-1}K_-|n\rangle$$
$$= (J+\tfrac{1}{2})^{-1}n|n-1\rangle = |n-1\rangle, \quad \text{for } n\neq 0,$$
$$A|0\rangle = 0.$$

One obtains

$$A|\phi,\psi\rangle = -e^{-i\psi}\tanh(\phi/2)|\phi,\psi\rangle. \tag{46}$$

Remarks:

(i) The operator A we have defined is different of the one, denoted by a, we have introduced in the harmonic oscillator realization [Eqs. (11)].

(ii) In contradistinction with the harmonic oscillator coherent states that fill up the whole complex plane, the spectrum of A fills the unit open disk $|z|<1$.

These two remarks are an invitation to produce new kinds of "coherent states." Here, we proposed two sets, denoted ξ states and ζ states, respectively.

A. The ξ states

They are defined by the relation

$$|\xi\rangle = \exp\left(-\frac{1}{2}|\xi|^2\right)\sum_0^\infty \frac{\xi^n}{\sqrt{n!}}|n\rangle. \tag{47}$$

TABLE III. Transformation formulas involving Laguerre and Hardy–Pollaczek polynomials.

$$\frac{(z-i)^n}{(z+i)^{n+1}} = \frac{(-i)^{n+1}}{2}\int_{-\infty}^{+\infty}\frac{P_n(\lambda)(-iz)^{-1/2+i\lambda}}{\cosh(\pi\lambda)d\lambda} \quad (-\pi/2<\arg(-iz)<\pi/2)$$

$$e^{-r}L_n(2r) = \frac{1}{2\pi i}\int_{(-i)+}e^{-irz}\frac{(z-i)^n}{(z+i)^{n+1}}dz$$

$$P_n(\lambda) = \frac{i^n}{\Gamma(\frac{1}{2}+i\lambda)}\int_0^\infty e^{-r}r^{i\lambda-1/2}L_n(2r)dr$$

$$P_n(\lambda) = i^n\frac{\cosh(\pi\lambda)}{\pi}\int_0^\infty \frac{(p-1)^n}{(p+1)^{n+1}}p^{-i\lambda-1/2}$$

$$\frac{(p-1)^n}{(p+1)^{n+1}} = \int_0^\infty e^{-(p+1)r}L_n(2r)dr$$

$$(-)^n e^{-r}L_n(2r) = \frac{i^n}{2\pi}\int_{-i\sigma-\infty}^{-i\sigma+\infty}\Gamma\left(\frac{1}{2}+i\lambda\right)P_n(\lambda)r^{i\lambda-1/2}d\lambda \quad \left(\sigma\,\text{real} > -\frac{1}{2}\right)$$

They are vectors of the Hilbert space since they have a unit norm. Let

$$\langle \xi | \xi \rangle = \sum_0^\infty \exp(-|\xi|^2) \frac{|\xi|^{2n}}{n!} = 1.$$

It is a simple exercise to verify that they are eigenstates of the anihilation operator a of the harmonic oscillator realization and that two ξ states cannot be orthogonal.

B. The ζ states

$$|\zeta\rangle = N(\zeta) \sum_0^\infty \frac{\zeta^n}{n!} |n\rangle, \tag{48}$$

where ζ is an arbitrary complex number and $N(\zeta)$ a normalization factor. Since

$$\langle \zeta | \zeta \rangle = N(\zeta)^2 \sum_0^\infty \frac{|\zeta|^{2n}}{n!n!} = N(\zeta)^2 I_0(2|\zeta|),$$

we see that $N(\zeta) = [I_0(2|\zeta|)]^{-1/2}$, where I_0 denotes the Bessel function. It is a simple matter to verify that the vectors $|\zeta\rangle$ are eigenvectors of the operator K_-.

The scalar product of two coherent states is

$$\langle \zeta' | \zeta \rangle = N(\zeta')N(\zeta) \sum_0^\infty \frac{(\zeta'^* \zeta)^n}{n!n!}$$

$$= \frac{I_0(2\sqrt{\zeta'^* \zeta})}{\sqrt{I_0(2|\zeta'|) I_0(2|\zeta|)}}. \tag{49}$$

It is known that the zeros of the Bessel function I_0 are all pure imaginary. Let us denote by $\pm 2i\alpha_k$ these zeros. [The roots α_k $(k = 0,1,2,...)$ are close to $\pm i(3\pi/4 + k\pi)$.] For two coherent states to be orthogonal, it is necessary to have $\zeta'^* \zeta = -\alpha_k^2$. If we set $\zeta = \xi + i\eta$ and $\zeta' = \xi' + i\eta'$, this condition reads

$$\xi'\eta - \xi\eta' = 0, \quad \xi\xi' + \eta\eta' = -\alpha_k^2,$$

which means (a) that the points ζ and ζ' lie on a line containing the origin (the origin lying in between) and (b) that $|\zeta||\zeta'| = \alpha_k^2$.

Let us see how a coherent state is expressed in the λ realization. We have, according to (19b) and (32b),

$$\langle \lambda | \zeta \rangle = \sum_0^\infty \frac{\zeta^n}{n!} \langle \lambda | n \rangle$$

$$= \sum_0^\infty \frac{\zeta^n}{n!} i^n P_n(\lambda) = e^\zeta M(\tfrac{1}{2} - i\lambda, 1; -2\zeta). \tag{50}$$

The scalar product formula gives

$$\int_{-\infty}^{\infty} \frac{e^{\zeta'^* + \zeta} M(\tfrac{1}{2} + i\lambda, 1; -2\zeta'^*) M(\tfrac{1}{2} i\lambda, 1; -2\zeta)}{\cosh(\pi\lambda)} d\lambda$$

$$= I_0(2\sqrt{\zeta'^* \zeta}). \tag{51}$$

VIII. RIEMANN'S ZETA FUNCTION

The Riemann zeta function is implied in a natural way in the representation we are concerned with. More precisely, as we will see, we have

$$\langle x | 1/(1 + e^{-R}) | 0 \rangle = (1 - 2^{-x}) \zeta(x+1), \tag{52}$$

or, if we take into account the relation $x = -\frac{1}{2} + i\lambda$,

$$\langle \lambda | 1/(1 + e^{-R}) | 0 \rangle = (1 - 2^{-1/2 - i\lambda}) \zeta(\tfrac{1}{2} + i\lambda).$$

Before investigating the relationship between the Riemann zeta function and our representation, we state a few lemmas leading to properties of the operators $e^{\rho R}$.

Lemma 2: Any function of the type $(at+b)^x/(ct+d)^{x+1}$ can be expanded in the μ_n basis. We get

$$\frac{(at+b)^x}{(ct+d)^{x+1}} = 2 \sum_{n=0}^\infty \frac{[(a-c)t + b - d]^n}{[(a+c)t + b + d]^{n+1}} \mu_n(x).$$

Proof: The proof is easy. We only have to set

$$u = [(a-c)t + b - d]/[(a+c)t + b + d],$$

in Eq. (32a). ■

Lemma 3: The numbers $\nu_{m,n}$ defined by the expansion

$$\frac{(\alpha t + \beta)^n}{(\gamma t + \delta)^{n+1}} = \sum_{m=0}^\infty \nu_{m,n} t^m,$$

have as a double generating function

$$\sum_{m,n} \nu_{m,n} t^m u^n = \frac{1}{\delta + \gamma t - \beta u - \alpha t u}, \tag{53}$$

and verify the following recurrence formula

$$\delta \nu_{m+1,n+1} = \beta \nu_{m+1,n} - \gamma \nu_{m,n+1} + \alpha \nu_{m,n}. \tag{54}$$

This formula permits to build the whole table of numbers from the "initial conditions"

$$\nu_{0,n} = \beta^n/\delta^{n+1}, \quad \nu_{m,0} = (-\gamma)^m/\delta^{m+1}.$$

Proof: The double generating function comes from the relation

$$\sum_{n=0}^\infty \frac{(\alpha t + \beta)^n}{(\gamma t + \delta)^{n+1}} u^n = \frac{1}{\delta + \gamma t - \beta u - \alpha t u},$$

and the recurrence relation comes from the identity

$$(\delta + \gamma t - \beta u - \alpha t u) \sum \nu_{m,n} t^m u^n = 1,$$

in matching the term $t^{m+1} u^{n+1}$ on both sides. Finally, the numbers $\nu_{m,0}$ and $\nu_{0,n}$ are easily obtained by setting $u = 0$ (resp. $t = 0$) in the double generating function. ■

Remark: We note that the numbers are symmetric $(\mu_{m,n} = \mu_{n,m})$ provided $\gamma = -\beta$. That is the case for the so-called Delannoy numbers[12-14] for which we have $\alpha = \beta = -\gamma = \delta = 1$. These numbers are the values taken by the Pidduck polynomials for positive integral arguments: $\mu_m(n) = \mu_n(m) = \mu_{nm}$. If the Delannoy numbers are simple, it is essentially due to their simple initial conditions $\mu_{m,0} = \mu_{0,n} = 1$.

Lemma 4: The matrix elements $\nu_{m,n}(\rho) = \langle m | e^{2\rho R} | n \rangle$ of the operator $e^{2\rho R}$ in the basis $|n\rangle$ have as a double generating function

$$\sum \nu_{m,n}(\rho) t^m u^n = \frac{1}{(1-\rho) - \rho t - \rho u - (1+\rho) t u}.$$

Proof: The proof becomes simple if we make use of the x realization of the Lie algebra. We remind the reader that $R = \exp(d/dx)\, x$. We have, according to Eq. (32a):

$$\sum_{n=0}^{\infty} e^{2\rho R}\mu_n(x)t^n$$
$$= e^{2\rho R}\frac{(1+t)^x}{(1-t)^{x+1}}$$
$$= \sum_{m=0}^{\infty}\frac{(2\rho)^m R^m}{m!}\frac{(1+t)^x}{(1-t)^{x+1}}$$
$$= \sum_{m=0}^{\infty}\frac{(2\rho)^m}{m!}(x+1)(x+2)\cdots(x+m)$$
$$\times\frac{(1+t)^{x+m}}{(1-t)^{x+m+1}}$$
$$= \frac{1}{(1-2\rho(1+t)/(1-t))^{x+1}}\frac{(1+t)^x}{(1-t)^{x+1}}$$
$$= \frac{(1+t)^x}{[1-2\rho-t(1+2\rho)]^{x+1}}.$$

The end of the proof consists in a direct application of Lemmas 2 and 3. ■

Lemma 5: The numbers $\nu_{m,n}$ defined by Eq. (53) are related to the Jacobi polynomials by the following relation:

$$\nu_{m,n} = \frac{\alpha^m\beta^{n-m}}{\delta^{n+1}}P_m^{(n-m,0)}\frac{1-2\beta\gamma}{\alpha\delta}.$$

Proof: The proof is based on the following recurrence relation obeyed by the Jacobi polynomials:

$$P_{m+1}^{(n-m,q)}(s) = P_{m+1}^{(n-m-1,q)}(s) + P_m^{(n-m,q)}(s) + [(s-1)/2]P_m^{(n+1-m,q)}(s).$$

This nonstandard recurrence relation can be deduced easily from the standard list of the recurrence relations involving three polynomials. The following generating functions can be deduced:

$$\sum_{m=0}^{\infty} P_m^{(n-m,q)}(s)t^m = \frac{(1+t)^n}{(1+t[(1-s)/2])^{n+q+1}},$$

$$\sum_{m=0}^{\infty} P_m^{(n-m,q)}(s)t^m u^n = \frac{1}{(1+(1-s)/2)^q}\frac{1}{1+[(1-s)/2]t-u-tu}.$$

This last function is of type (53) when $q=0$. It is not difficult to show from this formula that the $\nu_{m,n}$ of Lemma 3 have the generating function (53). In particular, the Delannoy numbers[12-14] $\mu_{m,n}=\mu_n(m)=\mu_m(n)$ are the values taken by the Jacobi polynomials for the argument 3. ■

The infinite matrix $e^{2\rho R}$ reads:

$$\begin{pmatrix}
\frac{1}{1-\rho} & \frac{\rho}{(1-\rho)^2} & \frac{\rho^2}{(1-\rho)^3} & \frac{\rho^3}{(1-\rho)^4} & \cdots \\
\frac{\rho}{(1-\rho)^2} & \frac{\rho^2+1}{(1-\rho)^3} & \frac{\rho(\rho^2+2)}{(1-\rho)^4} & \frac{\rho^2(\rho^2+3)}{(1-\rho)^5} & \cdots \\
\frac{\rho^2}{(1-\rho)^3} & \frac{\rho(\rho^2+2)}{(1-\rho)^4} & \frac{\rho^4+4\rho^2+1}{(1-\rho)^5} & \frac{\rho(\rho^4+6\rho^2+3)}{(1-\rho)^6} & \cdots \\
\frac{\rho^3}{(1-\rho)^4} & \frac{\rho^2(\rho^2+3)}{(1-\rho)^5} & \frac{\rho(\rho^4+6\rho^2+3)}{(1-\rho)^6} & \frac{\rho^6+8\rho^4+6\rho^2+1}{(1-\rho)^7} & \cdots \\
\frac{\rho^4}{(1-\rho)^5} & \cdots & \cdots & \cdots & \cdots
\end{pmatrix}.$$

The matrix entries are, according to Lemma 5, the numbers

$$\frac{\rho^{n-m}(1+\rho)^m}{(1-\rho)^{n+1}}P_m^{(n-m,0)}\left(\frac{1+\rho^2}{1-\rho^2}\right).$$

Let us note that the operator $e^{2\rho R}$ is an element of the group $SL(2,R)$ provided ρ is pure imaginary. We note that this matrix reads, for $\rho=-\frac{1}{2}$,

$$\exp(-R) = 2\begin{pmatrix}
\frac{1}{3} & -\frac{1}{3^2} & \frac{1}{3^3} & -\frac{1}{3^4} & \cdots \\
-\frac{1}{3^2} & \frac{5}{3^3} & -\frac{9}{3^4} & \frac{13}{3^5} & \cdots \\
\frac{1}{3^3} & -\frac{9}{3^4} & \frac{33}{3^5} & -\frac{73}{3^6} & \cdots \\
-\frac{1}{3^4} & \frac{13}{3^5} & -\frac{73}{3^6} & \frac{193}{3^7} & \cdots \\
\frac{1}{3^5} & \cdots & \cdots & \cdots & \cdots
\end{pmatrix}.$$

These matrix entries are the numbers $[(-)^{m+n}/(3^{n+1})]2P_m^{(n-m,0)}(\frac{5}{3})$. For the matrix $\exp(-S)$, they would be the numbers $(1/3^{n+1})2P_m^{(n-m,0)}(\frac{5}{3})$.

Remark: The same lemmas could be used for computing the matrix elements of the operator $\exp(2i\phi L)$. The natural realization to be used for this calculation is the Laguerre one. One obtains

$$e^{2i\phi L}=\frac{1}{\mathrm{ch}}\begin{pmatrix}1 & -\mathrm{th} & \mathrm{th}^2 & -\mathrm{th}^3 & \cdots\\ \mathrm{th} & -2\,\mathrm{th}^2+1 & 3\,\mathrm{th}^3-2\,\mathrm{th} & -4\,\mathrm{th}^4+3\,\mathrm{th}^2 & \cdots\\ \mathrm{th}^2 & -3\,\mathrm{th}^3+2\,\mathrm{th} & 6\,\mathrm{th}^4-6\,\mathrm{th}^2+1 & -10\,\mathrm{th}^5+12\,\mathrm{th}^3-3\,\mathrm{th} & \cdots\\ \mathrm{th}^3 & -4\,\mathrm{th}^4+3\,\mathrm{th}^2 & 10\,\mathrm{th}^5-12\,\mathrm{th}^3+3\,\mathrm{th} & -20\,\mathrm{th}^6+30\,\mathrm{th}^4-12\,\mathrm{th}^2+1 & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots\end{pmatrix},$$

where ch is put for $\cosh(\phi)$ and th for $\tanh(\phi)$. The matrix elements are generated by the function $[(1-tu)\cosh(\phi)+(t-u)\sinh(\phi)]^{-1}$, which is of the type (53). They obey the recurrence relation (54).

A. The coefficients ζ_n

They are defined by the series

$$\zeta_n=(-)^n\sum_{k=2}^{\infty}(-)^{k-1}\frac{(k-1)^n}{(k+1)^{n+1}}. \tag{55}$$

It is not difficult to prove that these alternate series converge. For $n=0$, we have $\zeta_0=1-\mathrm{Log}\,2$. Unfortunately, these series are not convenient for practical calculations because the convergence is very slow. Fortunately, we have the following theorem.

Theorem 8: The numbers ζ_n obey the following relations:

$$\zeta_n=(-)^n\int_0^{\infty}\frac{e^{-r}L_n(2r)}{1+e^r}\,dr, \tag{56}$$

$$(-)^n\zeta_n=-\log 2+\sum_{k=1}^{n}\binom{n}{k}(-)^k(2^k-1)[\zeta(k+1)-1],$$

$$n\neq 0, \tag{57}$$

$$\zeta_0=1-\log 2.$$

They are involved in the following expansions:

$$(1-2^{-x})\zeta(x+1)=2\sum_{k=0}^{\infty}\zeta_n\mu_n(x), \tag{58}$$

$$\frac{1}{1+e^{-r}}=2\sum_{n=0}^{\infty}\zeta_n(-)^nL_n(2r), \tag{59}$$

$$\frac{1}{1+e^{-R}}|0\rangle=2\sum_{n=0}^{\infty}\zeta_n|n\rangle. \tag{60}$$

Proof: We start with the proof of (60). The operator $(1+e^{2\rho R})^{-1}=\Sigma_{k=0}^{\infty}(-)^ke^{2k\rho R}$ has as matrix elements, the numbers given by the double generating function:

$$\sum_{k=0}^{\infty}\frac{1}{1-k\rho-k\rho(t+u)-(1+k\rho)tu}=-\frac{1}{\rho(1+t)(1+u)}\times\sum_{k=0}^{\infty}\frac{(-)^k}{k-(1-tu)/\rho(1+t)(1+u)}.$$

According to Ref. 8, p. 20, Eq. (6), this function is related with the hypergeometric function. It is

$$\frac{1}{1-tu}\,{}_2F_1\left(1,\frac{tu-1}{\rho(1+t)(1+u)};1+\frac{tu-1}{\rho(1+t)(1+u)};-1\right).$$

The formula (60) concerns the elements $\langle 0|[1+\exp(-R)]^{-1}|n\rangle$. They are obtained with the aid of the function we just obtained in making $u=0$ and $\rho=-\frac{1}{2}$, that is, the function

$${}_2F_1\left(1,\frac{2}{1+t};1+\frac{2}{1+t};-1\right)=1-\frac{2}{t+3}+\frac{2}{2t+4}-\frac{2}{3t+5}+\frac{2}{4t+6}-\cdots=1+2\sum_{k=2}^{\infty}\frac{(-)^{k-1}}{(k-1)t+k+1}.$$

We want to write this function as a power series in t. By taking the nth derivative on both sides and making $t=0$, one gets the series

$${}_2F_1\left(1,\frac{2}{1+t};1+\frac{2}{1+t};-1\right)=2\sum_{n=0}^{\infty}\zeta_nt^n, \tag{61}$$

with the ζ_n's given by Eq. (55). This proves Eq. (60). Before giving the proof of the other equations, let us give the values of the first coefficients ζ_n.

$$\begin{aligned}\zeta_0&=+0.30685\ 28194\ 40054=1-\mathrm{Log}\,2,\\ \zeta_1&=+0.04821\ 31137\ 11718,\\ \zeta_2&=-0.00944\ 97563\ 42275,\\ \zeta_3&=+0.00059\ 44724\ 73647,\\ \zeta_4&=+0.00069\ 90473\ 19902,\\ \zeta_5&=-0.00049\ 57044\ 38210,\\ \zeta_6&=+0.00020\ 94452\ 93504,\\ \zeta_7&=-0.00005\ 34608\ 70199,\\ \zeta_8&=-0.00000\ 57666\ 01662,\\ \zeta_9&=+0.00001\ 88215\ 01375,\\ \zeta_{10}&=-0.00001\ 56240\ 80434.\end{aligned}$$

Obviously, these numbers are not computed with the aid of series (55) that converges very slowly, but with Eq. (57)

and tables for the Riemann zeta function. Equation (59) is a direct consequence of Eq. (60) since it is Eq. (60) with $|n\rangle$ replaced by its representative in the r realization (Laguerre polynomials). Equation (56) follows. Equation (58) can be proved in the following way. It is well known that

$$(1-2^{-x})\zeta(x+1)=\sum_{k=1}^{\infty}\frac{(-)^{k-1}}{k^{x+1}}. \tag{62}$$

This is an entire function (the limit exists when x goes to zero: it is Log 2). If we replace $(1+u)/(1-u)$ by $(1/k)$ in Eq. (32a), we get

$$\frac{1}{k^{x+1}}=2\sum_{n=1}^{\infty}\mu_n(x)\frac{(1-k)^n}{(1+k)^{x+1}}.$$

We note that for $k=1$, the right-hand side is 1. We bring this expression in Eq. (62). We obtain, after having shown that the order of the two summations can be exchanged, our formula (58):

$$(1-2^{-x})\zeta(x+1)=\sum_{n=0}^{\infty}\zeta_n\mu_n(x).$$

We are left with Eq. (57). It follows from our definition of the coefficients ζ_n. We have

$$\begin{aligned}\frac{(k-1)^n}{(k+1)^{n+1}}&=\frac{1}{k+1}\left(1-\frac{2}{k+1}\right)^n\\&=\sum_{m=0}^{n}\binom{n}{m}\frac{(-2)^m}{(k+1)^{m+1}}\\&=\frac{1}{k+1}+\sum_{m=1}^{n}\binom{n}{m}\frac{(-2)^m}{(k+1)^{m+1}}\end{aligned}$$

Then, with the aid of (62), we get

$$\begin{aligned}(-)^n\zeta_n&=\sum_{k=2}^{\infty}(-)^{k-1}\frac{(k-1)^n}{(k+1)^{n+1}}\\&=\frac{1}{2}-\text{Log}\,2-\sum_{m=1}^{n}\binom{n}{m}(-2)^m[(1-2^{-m})\\&\quad\times\zeta(m+1)-1+2^{-m-1}],\end{aligned}$$

which leads to Eq. (57). ■

Other properties of the coefficients ζ_n can be proved. Here we give some of them without proof.

$$\zeta_n=-\frac{1}{n}\int_0^{\infty}e^{-2t}t^n\frac{d^n}{dt^n}\left(\frac{1}{1+e^t}\right)dt,$$

$$(-)^n\zeta_n=(-1)^n-\log 2-\int_0^{\infty}\frac{L_n(2t)-L_n(t)}{e^t-1}\,dt$$

$$(2^n-1)\zeta(n+1)=2^n-\sum_{m=0}^{n}\binom{n}{m}\zeta_m,$$

$$\sum_{n=0}^{\infty}\zeta_n=\frac{1}{2}\log 2,$$

$$\sum_{n=0}^{\infty}\zeta_n(-)^n=\frac{1}{4},$$

$$\sum_{n=0}^{\infty}\zeta_n^2=2\sum_{n=0}^{\infty}\zeta_{2n+1}=\sum_{n=0}^{\infty}\frac{(-)^n\zeta_n}{3^{n+1}}=\frac{1}{2}\text{Log}\,2-\frac{1}{4},$$

$$\sum_{n=0}^{\infty}\frac{\zeta_n}{3^{n+1}}=\frac{1}{2}-\frac{\pi}{8}.$$

To give an idea of the convergence of some of these series, we note that

$$\sum_{n=1}^{10}\zeta_n^2=0.096\,573\,590\,162\cdots,$$

the infinite sum gives, instead, 0.096 573 590 279⋯.

IX. THE REPRESENTATIONS $D_+(-\frac{1}{2}+\epsilon)$ FOR $\epsilon>-\frac{1}{2}$

It is natural to give a generalization of the above results to other representations of SL(2,R). In order to make a kind of synthesis, we want to include some of the representations that lie in the neighborhood of $D_+(-\frac{1}{2})$. That is why we give some results on the representations of the universal covering of SL(2,R) that belongs to the series $D_+(-\frac{1}{2}+\epsilon)$ for $\epsilon>-\frac{1}{2}$. They are representations of SL(2,R) provided 2ϵ is an integer. We only give some indications about this investigation. The method is simply a generalization of the one used in the above sections in the case $\epsilon=0$.

Instead of Eqs. (9), we get

$$J|n\rangle=(n+\tfrac{1}{2}+\epsilon)|n\rangle, \tag{63a}$$

$$K_+|n\rangle=\sqrt{(n+1)(n+1+2\epsilon)}\,|n+1\rangle, \tag{63b}$$

$$K_-|n\rangle=\sqrt{n(n+2\epsilon)}\,|n-1\rangle. \tag{63c}$$

We readily see a characteristic of the representation $D_+(-\frac{1}{2})$: it is the only representation of this series for which these expressions involve rational functions of n. Unfortunately, the orthogonal polynomials associated with Eqs. (63) are more complicated than Laguerre and Meixner ones. Equation (18) would read

$$\begin{aligned}&\sqrt{(n+1)(n+1+\epsilon)}\,Q_{n+1}(x)\\&\quad=[-x-(2n+1+2\epsilon)\sigma]Q_n(x)\\&\qquad-\sqrt{n(n+\epsilon)}\,Q_{n-1}(x).\end{aligned}$$

However, it is possible to simplify the calculation in choosing, instead of the *orthonormal* basis $|n\rangle$, the *orthogonal* basis $|\bar{n}\rangle$ defined by the relation

$$|n\rangle=\sqrt{\Gamma(2\epsilon+n+1)/n!}\,|\bar{n}\rangle. \tag{64}$$

In that case the polynomials involved become rational. (Let us emphasize that the operators K_+ and K_- are still Hermitian conjugate but the corresponding matrices, in this new basis, are not.) We get

$$K_+|\bar{n}\rangle=(n+1+2\epsilon)|\,\overline{n+1}\,\rangle, \tag{65a}$$

$$K_-|\bar{n}\rangle=n|\,\overline{n-1}\,\rangle, \tag{65b}$$

and, instead of the recurrence formula (17a), we get

$$\beta_{n+1}(n+1)p_{n+1}(x) = [2x - a(2n+1+2\epsilon)]p_n(x) - \beta^*(n+2\epsilon)p_{n-1}(x).$$

Let us consider the cases of K $(a=0,\beta=1)$ and R $(a=\beta=1)$. With $p_0(x)=1$, one obtains, for the generating functions;

$$K: \quad \sum_{n=0}^{\infty} p_n(x)z^n = (1-iz)^{ix-\epsilon-1/2}(1+iz)^{ix-\epsilon-1/2},$$

$$R: \quad \sum_{n=0}^{\infty} p_n(x)z^n = (1+z)^{-1-2\epsilon}\exp\left(-\frac{2xz}{1+z}\right).$$

This last function is the generating function of the Laguerre polynomials $(-)^n L_n^{(2\epsilon)}(2x)$.

Let us give the description of the x realization. We have

$$R = \exp\left(\frac{d}{dx}\right)(x+2\epsilon),$$

$$S = -x\exp\left(-\frac{d}{dx}\right), \quad L = -i(x+\epsilon+\tfrac{1}{2}). \tag{66}$$

The polynomials

$$\mu_n^{(\epsilon)}(x) = \frac{2^n\Gamma(x+1)}{n!\Gamma(x+1-n)} \times F(-n,-n-2\epsilon;x+1-n;\tfrac{1}{2}),$$

obey the recurrence relation

$$(n+1)\mu_{n+1}^{(\epsilon)}(x) = (2x+2\epsilon+1)\mu_n^{(\epsilon)}(x) + (n+2\epsilon)\mu_{n-1}^{(\epsilon)}(x). \tag{67}$$

Their generating function is

$$\sum \mu_n^{(\epsilon)}(x)t^n = \frac{(1+t)^x}{(1-t)^{x+2\epsilon+1}}. \tag{68}$$

They are eigenfunctions of J:

$$J\mu_n^{(\epsilon)}(x) = (n+\epsilon+\tfrac{1}{2})\mu_n^{(\epsilon)}(x). \tag{69}$$

Their symmetry property reads

$$\mu_n^{(\epsilon)}(x) = (-)^n\mu_n^{(\epsilon)}(-x-2\epsilon-1).$$

These polynomials are related with the Pollaczek polynomials[4] as follows (Our labeling is different from the standard one.):

$$P_n^{(\epsilon)}(\lambda) = i^n\mu_n^{(\epsilon)}(-\tfrac{1}{2}-\epsilon+i\lambda). \tag{70}$$

They are orthogonal;

$$\int_{-\infty}^{\infty} 2^{2\epsilon}\frac{\Gamma(\frac{1}{2}+\epsilon+i\lambda)\Gamma(\frac{1}{2}+\epsilon-i\lambda)}{\pi\Gamma(2\epsilon+1)}P_n^{(\epsilon)}(\lambda)P_m^{(\epsilon)}(\lambda)d\lambda = \delta_{nm}.$$

Finally, we note the relation of these polynomials with the Laguerre ones, through a Mellin transformation;

$$\mu_n^{(\epsilon)}(x) = \frac{(-)^n}{\Gamma(x+2\epsilon+1)}\int_0^{\infty} e^{-r}r^{x+2\epsilon}L_n^{(2\epsilon)}(2r)dr. \tag{71}$$

From that transformation, one can obtain the following realization:

$$R = r, \quad S = -r\frac{d^2}{dr^2} - (2\epsilon+1)\frac{d}{dr},$$

$$L = i\left(r\frac{d}{dr} + \epsilon + \frac{1}{2}\right).$$

The eigenfunctions of J are now given by

$$\frac{(z-1)^n}{(z+1)^{n+1}} = -\frac{i^n}{2}\int_{-\infty}^{+\infty}\frac{P_n(\lambda)z^{-1/2+i\lambda}}{\cosh(\pi\lambda)}d\lambda.$$

X. CONCLUSION

In the previous sections, we have given ten realizations of the unitary representation $D_+(-\frac{1}{2})$ of the group $SL(2,R)$. Let us recall them, in mentioning the operator that is "diagonalized" and the corresponding name of the realization (Note that the angle operator Θ cannot be diagonalized in a strict sense):

Operator diagonalized	*Name of the realization*
J	n(matrix)
L	λ(Hardy–Pollaczek)
$X = -\frac{1}{2} + iL$	x(Pidduck)
R	r(Laguerre)
S	s(Laguerre)
$Z = -(LR^{-1} + R^{-1}L)/2$	z
$\Theta = -(i/2)\times\log[K_+(J+\frac{1}{2})^{-1}]$	θ(Fourier)
$A = (J+\frac{1}{2})^{-1}K_-$	(ϕ,ψ)(coherent state)
$a = (J+\frac{1}{2})^{-1/2}K_-$	ξ(harmonic oscillator)
K_-	ζ

Apart the interest due to Fourier, Laplace, and Mellin transformations relating various special functions in a group theoretical context, we have to underline that there exist physical implications. We have already mentioned the one of Bracken *et al.*[7] We must also underline that the J spectrum is the one of the harmonic oscillator. Let us say a few words about that. The usual group associated with this physical system is the metaplectic group $Mp(2,R)$, the double covering of $SL(2,R)$. The harmonic oscillator states span the Hilbert space of a *reducible* representation of $Mp(2,R)$, namely the representations denoted $D_+(-\frac{1}{4})$ and $D_+(-\frac{3}{4})$ in our notation. Both correspond to the eigenvalue $-\frac{3}{16}$ of the Casimir operator.[15] The reduction of this representation separate even states from odd states. The representation $D_+(-\frac{1}{2})$ has the advantage to unify them. It is probably involved in an article of Leyvraz and Seligman [see Eq. (3.5) of their paper].[16]

Section IX was devoted to the Riemann zeta function. The link between this famous function and the operators R and S in the x realization has been considered as a relevant fact by de Branges in his attempt to furnish a proof of the Riemann conjecture. The representation $D_+(-\frac{1}{2})$ seems to be related with the "critical line" of the zeta function. The representations $D_+(-\frac{1}{2}+\epsilon)$ permit an exploration of the whole strip $-1<\mathrm{Re}(\epsilon)<0$, where the critical zeros (The "critical" zeros are the nonreal zeros.) of the zeta function are known to lie.

APPENDIX A: HOMOMORPHISM OF SL(2,R) IN THE LORENTZ GROUP L(1,2)

Let v be a real traceless 2×2 matrix. [Here, v is chosen as $2i(aJ+bK+cL)$, with $2J=-\sigma_2$, $2K=-i\sigma_3$, $2L=i\sigma_1$, in agreement with the choice made in Sec. II.] We make an element g of $SL(2,R)$ acting on it as follows:

$$\mathrm{v}\to g\mathrm{v}g^{-1}, \tag{A1}$$

(one verifies that $g\mathrm{v}g^{-1}$ is real and traceless). The set of these matrices is a real three-dimensional vector space. Each v can be written in the form

$$\mathrm{v}=\begin{pmatrix} b & -a-c \\ a-c & -b \end{pmatrix},$$

where a, b, c are real numbers. We see that $\det(\mathrm{v})=a^2-b^2-c^2$ is conserved by the transformation. Therefore there exists a homomorphism

$$SL(2,R)\to L(1,2).$$

The question is to know if parity and time reversal are implemented in $L(1,2)$. The answer is *no*. To prove it we have to check that there is no g such that

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

and no g such that

$$g\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}=\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}g.$$

The proof is easy.

Any g can be written in the form

$$g=\tfrac{1}{2}\begin{pmatrix} t+\beta & -\alpha-\gamma \\ \alpha-\gamma & t-\beta \end{pmatrix}, \quad \text{with } t^2+\alpha^2-\beta^2-\gamma^2=4$$

(where t is the trace). It is not difficult to obtain, from Eq. (1), the formula

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}\to\begin{pmatrix} \frac{t^2+\alpha^2+\beta^2+\gamma^2}{4} & -\frac{\alpha\beta+t\gamma}{2} & \frac{t\beta-\alpha\gamma}{2} \\ \frac{\alpha\beta-t\gamma}{2} & \frac{t^2-\alpha^2-\beta^2+\gamma^2}{4} & \frac{t\alpha-\beta\gamma}{2} \\ \frac{\alpha\gamma+t\beta}{2} & -\frac{\beta\gamma+t\alpha}{2} & \frac{t^2-\alpha^2+\beta^2-\gamma^2}{4} \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

This formula permits to see that the only g's that act as the identity on R^3 are

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

which proves that the kernel of the homomorphism is the group with two elements:

$$SL(2,R)\sim L(1,2)/Z_2.$$

APPENDIX B: UNITARY DUAL OF THE COVERING GROUP OF SL(2,R)

In the same physicist spirit, we give here a description of the set of unitary representations of the universal covering group of $SL(2,R)$.

Let J, K, L the basis elements of the Lie algebra in a unitary irreducible representation for which the Casimir operator has $\tau(\tau+1)$ as an eigenvalue. The fact that it is a real number obliged us to consider the two following cases:

(i) τ is real (we can impose $\tau\geqslant -\frac{1}{2}$)

(ii) τ is imaginary; then $\mathrm{Re}(\tau)=-\frac{1}{2}$ and $\mathrm{Im}(\tau)\neq 0$.

Let $|m\rangle$ be a normed eigenvector of J, with eigenvalue m. We have the relations

$$J|m\rangle=m|m\rangle,$$

$$JK_\pm|m\rangle=(m\pm 1)|m\pm 1\rangle.$$

This shows that, provided that $m\pm 1\neq 0$, $K_\pm|m\rangle$ is an eigenvector of J. *A priori*, we have four possibilities for the spectrum of J. It can be

(i) unbounded: $m=m_0, m_0\pm 1, m_0\pm 2,\ldots$, where m_0 can be chosen in an arbitrary interval of length one.

(ii) bounded below: $m=m_0, m_0+1, m_0+2,\ldots$. This implies that $K_-|m_0\rangle=0$.

(iii) bounded above: $m=m_0, m_0+1, m_0+2,\ldots$. This implies that $K_+|m_0\rangle=0$.

(iv) bounded both sides: $m=m_0, m_0+1, m_0+2,\ldots, |m_0+N\rangle$, which implies $K_+|m_0+N\rangle=0$ and $K_-|m_0\rangle=0$.

From the relations

$$K_+K_-=J^2-J-\tau(\tau+1)Id,$$

$$K_-K_+=J^2+J-\tau(\tau+1)Id,$$

we get

$$\|K_+|m\rangle\|^2=(m-\tau)(m+\tau+1), \tag{B1}$$

$$\|K_-|m\rangle\|^2=(m+\tau)(m-\tau-1). \tag{B2}$$

Now, let us examine separately the two cases:

1. τ real

(i) Let us suppose that the spectrum is not bounded. We know that the expressions (B1) and (B2) must be strictly positive, whatever m is. This situation corresponds to the *open* shaded regions on Fig. 2. We readily see that a sequence of eigenvalues m_0, $m_0\pm 1$, $m_0\pm 2,\ldots$ is only possible if $-1<\tau<0$. Since we can impose the relation $\tau\geqslant -\frac{1}{2}$, we see that m_0 can be chosen in the triangle OMN (OM and ON

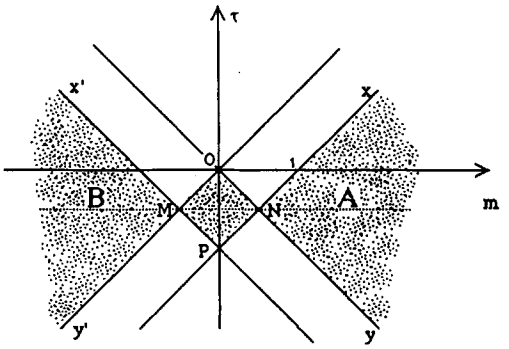


FIG. 2. Supplementary series.

excluded). Such a point defines a unique representation of the so-called supplementary series.

(ii) Let us now turn to the case where the spectrum is bounded below. We know that $K_-|m_0\rangle = 0$. We are in the region defined by the union of A and the two lines Px and Oy. Because of the symmetry $\tau \to -\tau - 1$, we can restrict ourselves to the case where the spectrum is bounded below by the line Px (point P excluded). Then the spectrum of J is $\tau + 1, \tau + 2, \tau + 3, \ldots$. It is the series of representations $D_+(\tau)$.

(iii) Symmetrically, we can define the series $D_-(\tau)$ in replacing the region A by the region B. The spectrum of J is $\tau - 1, \tau - 2, \tau - 3, \ldots$.

(iv) If the spectrum is bounded both sides, we must meet both lines Px and Px'. The only possibility is the point P itself. This corresponds to the trivial representation.

2. $\mathrm{Re}(\tau) = -\frac{1}{2}$ and $\mathrm{Im}(\tau) > 0$

We get the so-called principal series. In that case, $\|K_+|m\rangle\| = \|K_-|m\rangle\| = (m + \frac{1}{2})^2 + (\mathrm{Im}\,\tau)^2 > 0$. The spectrum of J is $m_0, m_0 \pm 1, m_0 \pm 2, \ldots$. One can always impose m_0 to lie in some given interval of width one.

APPENDIX C: SOME RELATIONS EXPRESSED WITH THE AID OF THE OPERATORS R AND S

The operators we are speaking about are the ones defined by Eqs (36a), namely

$$R(x) = \exp\left(\frac{d}{dx}\right)x,$$

$$S(x) = -x\exp\left(-\frac{d}{dx}\right) = R(-x-1).$$

It is not difficult to prove the following relations:

$$R(x)^n 1 = (x+1)(x+2)\cdots(x+n)$$

$$= n!2^{-n}\sum_{k=0}^{k=n}\binom{n}{k}\mu_k(x),$$

$$S(x)^n 1 = (-)^n x(x-1)(x-2)\cdots(x-n+1)$$

$$= n!(-2)^n\sum_{k=0}^{k=n}\binom{n}{k}(-)^{n-k}\mu_k(x),$$

$$\mu_n(x) = (-)^n L_n(2R(x))1 = L_n(2S(x))1.$$

The hypergeometric functions have a shorthand writing, namely

$$F(a+1,b+1;c;z) = \exp\{z[R(a)R(b)/S(-c)]\}1.$$

The Bessel function is easily related with the confluent hypergeometric function as follows:

$$I_\nu(2\sqrt{zS})1 = \sum_0^\infty \frac{(zS)^n}{n!(n+\nu)!}1$$

$$= \sum_0^n \frac{x(x-1)\cdots(x-n+1)}{n!(n+\nu)!}(-z)^n$$

$$= M(-x,\nu+1;z).$$

We also note that the expression

$$f(x) = \frac{-S(x)}{\exp[-S(x)]^{-1}}1 = \sum_0^\infty \frac{B_n}{n!}[-S(x)]^n 1$$

takes, for the natural integers, the values $f(n) = B_n$, where the B_n's are the Bernoulli numbers. This property follows from the relation

$$B_N = \sum_0^N \binom{N}{n}(-)^n B_n.$$

Finally, we write again the compact formula that expressed the entire function $(1-2^{-x})\zeta(x+1)$, with the aid of $R(x)$;

$$(1-2^{-x})\zeta(x+1) = [1/(1+e^{-R(x)})]1.$$

[1] H. Bacry and M. Boon, C. R. Acad. Sci. Paris **301**, 273 (1985); and *Proceedings of the Second International Symposium on Orthogonal Polynomials and their Applications* (Monografias de la Academia de Ciencias de Zaragoza, Zaragoza, 1988).
[2] C. Itzykson, J. Math. Phys. **10**, 1109 (1969).
[3] G. H. Hardy, Proc. Cambridge Philos. Soc. **36**, 1 (1940); or *Complete Works* (Clarendon, Oxford, 1974), p. 549.
[4] F. Pollaczek, C. R. Acad. Sci. Paris **230**, 1563 (1950).
[5] G. Szegö, *Orthogonal Polynomials* (Am. Math. Soc., Providence, 1978).
[6] F. B. Pidduck, Proc. R. Soc. London Ser. A **83**, 347 (1910) and **86**, 396 (1912).
[7] A. J. Bracken, H. S. Green, and L. Bass, J. Aust. Math. Soc. B **30**, 101 (1988).
[8] Erdélyi, *Higher Transcendental Functions* (Krieger, Malabar, FL, 1981), Vol. 1.
[9] I. M. Gelfand, M. I. Graev and N. Ja. Vilenkin, *Generalized Functions* (Academic, New York, 1966), Vol. V.
[10] A. O. Barut and L. Girardello, Commun. Math. Phys. **21**, 41 (1971).
[11] A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[12] L. Comtet, *Advanced Combinatorics* (Reidel, Dordrecht, 1974).
[13] R. G. Stanton and D. D. Cowan, Siam Rev. **12**, 277 (1970).
[14] H. Bacry and M. Boon, J. Math. Phys. **28**, 2639 (1987).
[15] H. Bacry and J.-L. Richard, J. Math. Phys. **8**, 2230 (1967).
[16] F. Leyvraz and T. H. Seligman, J. Math. Phys. **30**, 2512 (1989).

Branching rules for the Weyl group $W(D_n)$

G. Iommi Amunátegui
Instituto de Física, Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile

(Received 13 October 1989; accepted for publication 18 April 1990)

Reduction theorems for the decomposition of induced and irreducible characters of $W(D_n)$ in terms of induced and irreducible characters of $W(D_{n-1})$, respectively, are given.

I. INTRODUCTION

In recent years there has been a number of situations in which the Weyl groups W generated by the reflections of the root system of the classical Lie groups G have played an important role. This importance grew out of the various possibilities of application to physical problems, especially on lattices (for instance, discrete σ models, lattice gauge theories, chiral models; see Refs. 1 and 2).

Moreover, the set of maps from the circle to the Lie group G forms an infinite-dimensional group, called the loop group of G. The algebra of this group of maps is the untwisted affine Kac–Moody algebra $\hat{g}$, whose root system is infinite but which spans a finite-dimensional space. With each Kac–Moody algebra is associated a Virasoro algebra (for details, see Ref. 3). Infinite-dimensional algebras of this sort occur in certain areas of physics such as the string theories of particle interactions, two-dimensional statistical models (systems of spins on lattices), and two-dimensional σ models.[4]

The Weyl group $\hat{W}$ of $\hat{g}$ is the semidirect product of W and the root lattice obtained by interchanging the root lengths of the Lie algebra of G. Here, W is the subgroup of $\hat{W}$ that fixes any given point of the lattice of the Lie algebra dual to g.

In fact, the groups $\hat{W}$ were classified many years ago by Coxeter.[5]

The reduction of a representation of a group into representations of one of its subgroups is the subject of an extended literature. For Lie groups there exist general branching rules derived using tensor and spinor methods (see Ref. 6) as well as results given by explicit algebraically closed expressions (for instance, see Refs. 7 and 8). The reduction of the general linear group into the symmetric group S_n has also been considered.[9] A rule for the restriction $S_n \to S_{n-1}$ was obtained by Weyl.[10] The symmetric group is the Weyl group of the Lie group $SU(n)$.

In a previous article, hereafter referred to as I (see Ref. 11), the structure of the hyperoctahedral group $W(B_n)$ has been considered, and branching rules for its simple (irreducible) and induced characters have been established. Here $W(B_n)$ is the Weyl group of the classical Lie groups $B_n = SO(2n+1)$ and $C_n = Sp(2n)$ and $W(B_n) = Z_2^n \otimes S_n$ is the semidirect product of the Abelian group Z_2^n generated by the n sign changes $(+i, -i)$, $1 \leqslant i \leqslant n$, and the symmetric group S_n. The present paper is concerned with $W(D_n)$, the Weyl group of the Lie groups $SO(2n)$. Then, $W(D_n)$ is a subgroup of $W(B_n)$ of index 2 and consists of those elements of $W(B_n)$ that contain in their cycle decomposition an even number of changes of signs. Its order is $2^{n-1}n!$

The group $W(D_n)$ has already been studied (in particular, see Refs. 12 and 13; also Ref. 14, Chaps. 4 and 5, Refs. 15 and 16). Clearly, many of the results that will be stated below on the structure of $W(D_n)$ are well known. Our aim is to derive reduction rules for the simple and induced characters of $W(D_n)$. Some results obtained in I will form the substratum, so to say, of this work. We shall follow, as far as possible, the notions and notations displayed therein.

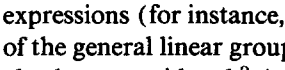
In Sec. II, the characters and classes of $W(D_n)$ are considered. It must be remarked that $W(D_{2n+1})$ is isomorphic to the factor group $W(B_{2n+1})/\{1, -1\}$ while the analogous property is not true for $W(D_{2n})$. Hence, this difference between $W(D_{2n+1})$ and $W(D_{2n})$ will appear subsequently.

In Sec. III the tables of the induced characters of $W(D_n)$ are constructed and in Sec. IV the branching rule $W(D_n) \to W(D_{n-1})$ for the induced characters is established. Section V is dedicated to the simple characters of $W(D_n)$ and Sec. VI to their reduction properly.

II. THE CHARACTERS AND CLASSES OF $W(D_n)$

A. The characters of $W(D_n)$

The first step will be to establish a correspondence between the notation employed by Mayer[13] and our notation.[11] Mayer denotes an irreducible character of $W(B_n)$ by a pair of subpartitions λ and μ of n such that $\lambda + \mu = n$. An irreducible character may then be written $X^{(\lambda;\mu)}$. For instance, in the case of $W(B_2)$ the irreducible characters are $X^{(2;0)}$, $X^{(0;2)}$, $X^{(11;0)}$, $X^{(1;1)}$, and $X^{(0;11)}$. In I, we denote these characters, respectively, by



and

i.e., to the subpartition λ is ascribed the sign $-$, and to the subpartition μ, the sign $+$. Bearing this in mind, we transcribe the main result of Mayer concerning the characters of $W(D_n)$.

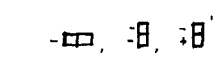
Theorem 2.1: Let $(\lambda;\mu)$ be a pair of subpartitions of n. Then, (i) $X^{(\lambda;\mu)}$ is an irreducible character of $W(D_n)$ if $\lambda \neq \mu$; (ii) $X^{(\lambda;\mu)} = X^{(\mu;\lambda)}$; (iii) $X^{(\lambda;\lambda)}$ is the sum of two distinct irreducible characters of $W(D_n)$ of the same degree; (iv) every irreducible character of $W(D_n)$ has the form $X^{(\lambda;\mu)}$ $(\lambda \neq \mu)$ or is the component of $X^{(\lambda;\lambda)}$ for some λ,μ;

 0022-2488/90/092078-07$03.00

(v) all the irreducible characters of $W(D_n)$ mentioned in (iv) are distinct, subject to (ii). (For a proof see Ref. 13.)

Note that (iii) occurs when n = even.

As an example, for $W(D_2)$ the irreducible characters are in the present notation:



and

Remark: $X^{(\lambda;\mu)}$ has been chosen instead of $X^{(\mu;\lambda)}$.

Moreover, $\lambda \geqslant \mu$ (lexicographical order). The equality case is related to point (iii) of Mayer's result.

Hereafter, we shall say that $X^{(\mu;\lambda)}$ do not belong to $W(D_n)$. To clarify this fact, consider the irreducible characters $X^{(3;21)}$ and $X^{(21;3)}$ of $W(B_6)$: In $W(D_6)$, only $X^{(3;21)}$ will be taken into account.

B. The classes of $W(D_n)$

If α_i^+ and α_i^- are, respectively, the number of positive and negative cycles of length i of a permutation, $\alpha = (\alpha_1^+, \alpha_1^-, \ldots, \alpha_l^+, \alpha_l^-)$ is called the α system of cycles. A class of $W(B_n)$ with such an α system is denoted $C(\alpha)$. A class $C(\alpha)$ is even (+) if all the α_i are positive or if the number of negative α_i is even. If this is not the case, $C(\alpha)$ is odd (−). A table for the number of classes of $W(B_n)$ for $1 \leqslant n \leqslant 150$ may be found in Ref. 17.

The classes of the subgroup $W(D_n)$ correspond to positive classes of $W(B_n)$. A distinction must be made between n = odd and n = even.

The number of classes of $W(D_n)$ for n odd is equal to

[number of classes + of $W(B_n)$]

$= \frac{1}{2}$ [number of classes of $W(B_n)$],

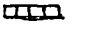
and for n even, is equal to

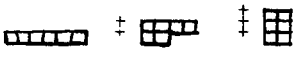
[number of classes + of $W(B_n)$] + $P(n)$ inferior type even parts,

where $P(n)$ inferior type even parts is the number of partitions of n in even parts.

Besides, the classes corresponding to the partitions of n in even parts are divided in two classes of equal order of $W(D_n)$. For these classes all the α_i are positive. That is, the classes that suffer such a subdivision are

$W(D_2)$ 

$W(D_4)$ 

$W(D_6)$ 

$\cdots$.

(See Appendix A.)

III. THE INDUCED CHARACTERS OF $W(D_n)$

Let $\lambda = (\lambda_1, \ldots, \lambda_k)$ be a partition of n ($\lambda_1 \geqslant \cdots \geqslant \lambda_k$) and $b = (b_1, \ldots, b_k)$ be such that $b_i = 1$ or 0 (if $\lambda_i = \lambda_{i+1}$, then $b_i \leqslant b_{i+1}$).

TABLE I. Induced characters of $W(D_2)$.

Order	1	1	1	1
Classes				
	1	1	1	1
	2	2	0	0
	2	0	0	0
	2	0	0	0

$= \phi\{W(D_2)\}$

A canonical subgroup of $W(B_n) = Z_2^n \otimes S_n$ is defined as $(Z_2^{(\lambda_1 - b_1)} \otimes S_1) \times \cdots \times (Z_2^{(\lambda_k - b_k)} \otimes S_k)$. In a similar manner, the canonical subgroups of $W(D_n)$ may be constructed at once from the irreducible characters defined in Sec. II A. For example, for $W(D_2)$, the canonical subgroups are

 $\lambda_1 = 2,\ b_1 = 0 \quad Z_2^{(2-0)} \otimes S_2,$

 $\lambda_1 = \lambda_2 = 1,\ b_1 = b_2 = 0$

$(Z_2^{(1-0)} \otimes S_1) \times (Z_2^{(1-0)} \otimes S_1),$

 $\lambda_1 = \lambda_2 = 1,\ b_1 = 0\ b_2 = 1$

$(Z_2^{(1-0)} \otimes S_1) \times (Z_2^{(1-1)} \otimes S_1).$

In I, an algorithm for the character of the representation of $W(B_n)$ induced by the identity representation of a canonical subgroup is given. For the subgroup of $W(B_n)$ we are dealing with, i.e., $W(D_n)$, it suffices to consider the induced characters of the canonical subgroups, defined via the corresponding partitions (see Sec. II A).

Remarks: For n even,

(1) pairs of identical induced characters appear. Each member of these pairs has a value equal to one-half the value of the corresponding induced character of $W(B_n)$ [this is related to Mayer's theorem, point (iii)].

(2) These pairs of induced characters have the same value for each subdivision of the classes related to the partitions of n in even parts.

We denote the induced character table of $W(D_n)$ by $\phi\{W(D_n)\}$ (see Tables I–III).

TABLE II. Induced characters of $W(D_3)$.

Order	1	3	6	6	8
Classes					
	1	1	1	1	1
	3	3	1	1	0
	6	2	2	0	0
	6	6	0	0	0
	12	4	0	0	0

$= \phi\{W(D_3)\}$

IV. THE REDUCTION $W(D_n) \to W(D_{n-1})$: THE INDUCED CHARACTERS

A brief discussion seems necessary before the statement of the reduction theorem. As already noted, the cases with n odd and n even must be treated separately. In both cases, however, the main pattern of the reduction theorem proved in I for $W(B_n) \to W(B_{n-1})$ remains unaltered. But (1) for $n =$ odd, $n-1 =$ even, and there is a splitting in two of the characters corresponding to equal subpartitions; (2) for $n =$ even, $n-1 =$ odd, consequently there are no partitions formed of equal subpartitions. Besides it may happen that in the reduction process a partition that does not belong to $W(D_{n-1})$ occurs, i.e., in Mayer's notation, (μ,λ) is obtained instead of (λ,μ) (see Sec. II A). From I, we know the reduction of induced and simple (irreducible) characters concerns only classes with $\alpha_1^+ \neq 0$. By inspection of the characters tables, it is seen that for classes with $\alpha_1^+ \neq 0$: (i) for the induced case, the characters of $(\lambda;\mu)$ and $(\mu;\lambda)$ are different; and (ii) for the irreducible case, the characters of $(\lambda;\mu)$ and $(\mu;\lambda)$ are identical.

Hence the partition that does not belong (see above) to $W(D_{n-1})$ must be considered only when we are dealing with the reduction of the induced characters.

To state the reduction theorem we employ the notation of I: $(i^+)^{a_i}$ denote a_i subpartitions of n, of length i and sign $+$, and $(i^-)^{a'_i}$, a'_i subpartitions of n, of length i and sign $-$. Here $\phi\{(1^+)^{a_1}(1^-)^{a'_1}(2^+)^{a_2}(2^-)^{a'_2}\cdots\}$ represents an induced character of $W(D_n)$.

Theorem 4.1: The induced character of $W(D_n)$, $\phi\{(1^+)^{a_1}(1^-)^{a'_1}\cdots\}$ reduces into $W(D_{n-1})$ induced characters as follows:

$$\phi\{(1^+)^{a_1}(1^-)^{a'_1}\ldots\}$$
$$= 2a_1\phi\{(1^+)^{a_1-1}(1^-)^{a'_1}\cdots\}$$
$$+ a'_1\phi\{(1^+)^{a_1}(1^-)^{a'_1-1}\ldots\}$$
$$+ a_2\phi\{(1^+)^{a_1+1}(1^-)^{a'_1}(2^+)^{a_2-1}\cdots\} + \cdots.$$

For n odd, the same decomposition must be repeated for each partition of $n-1 =$ even into equal subpartitions.

For n even, the resulting partitions, which do not belong to $W(D_{n-1})$, must be considered.

We omit the proof, analogous to that of Theorem 1 expressed in I.

This result allows us to construct a reduction matrix $F^{W(D_n)}_{W(D_{n-1})}$; each row of this matrix is given by the corresponding reduction.

In general (see I), the reduction of the induced characters may be written as

$$\phi\{W(D_n)\} = F^{W(D_n)}_{W(D_{n-1})}\phi\{(W(D_{n-1}))\}.$$

To exemplify this equation, let us consider an *even* case reduction: $F^{W(D_4)}_{W(D_3)}\phi\{W(D_3)\} = \phi\{W(D_4)\}$

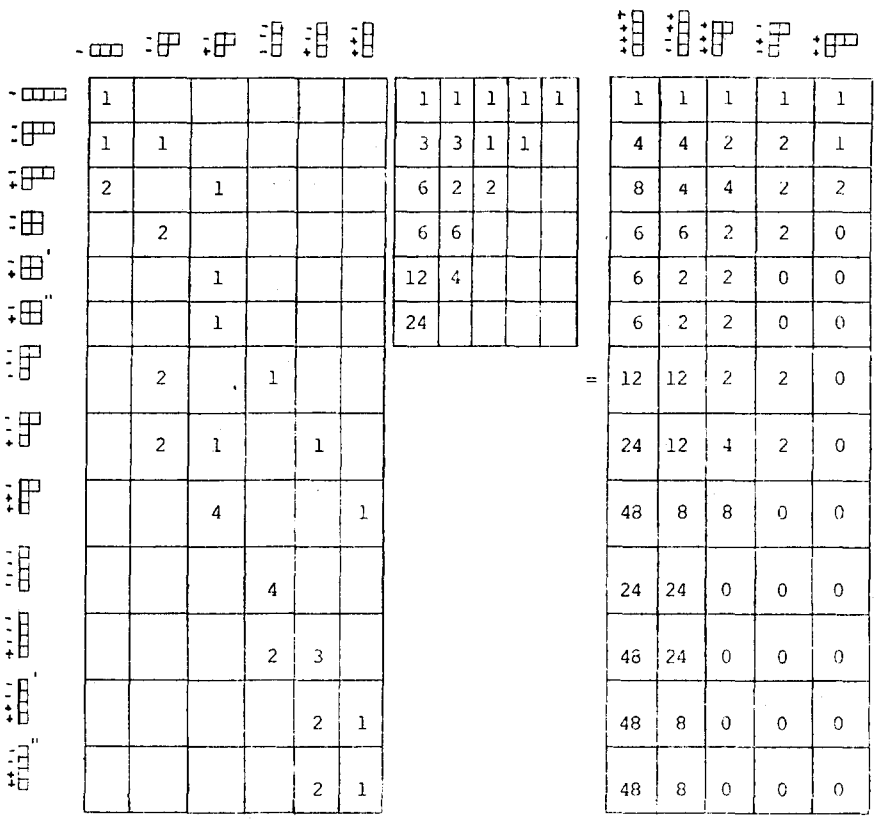


TABLE III. Induced characters of $W(D_4)$.

Order	1	6	1	12	12	24	6	6	12	32	32	24	24	
Classes														
	1	1	1	1	1	1	1	1	1	1	1	1	1	
	4	4	4	2	2	2	0	0	0	1	1	0	0	
	8	4	0	4	0	2	0	0	0	2	0	0	0	
	6	6	6	2	2	2	2	2	2	0	0	0	0	
	6	2	6	2	2	0	2	2	0	0	0	0	0	
	6	2	6	2	2	0	2	2	0	0	0	0	0	
	12	12	12	2	2	2	0	0	0	0	0	0	0	$= \phi\{W(D_4)\}$
	24	12	0	4	0	2	0	0	0	0	0	0	0	
	48	8	0	8	0	0	0	0	0	0	0	0	0	
	24	24	24	0	0	0	0	0	0	0	0	0	0	
	48	24	0	0	0	0	0	0	0	0	0	0	0	
	48	8	0	0	0	0	0	0	0	0	0	0	0	
	48	8	0	0	0	0	0	0	0	0	0	0	0	

Remarks: (1) The last column of F_3^4 corresponds to the partition $\begin{smallmatrix}\cdot\\\cdot\\\cdot\end{smallmatrix}\!\!\square$ which does not belong to $W(D_3)$.

(2) The last line of $\phi\{W(D_3)\}$ corresponds to the character induced by the canonical subgroup corresponding to $\begin{smallmatrix}\cdot\\\cdot\\\cdot\end{smallmatrix}\!\!\square$

(3) In $\phi\{W(D_4)\}$ the characters correspond to the classes with $\alpha_1^+ \neq 0$.

V. THE IRREDUCIBLE (SIMPLE) CHARACTERS OF $W(D_n)$

As in I, the table of irreducible characters of $W(D_n)$, $X\{W(D_n)\}$, can be obtained from $\phi\{W(D_n)\}$: each row ϕ_i of $\phi\{W(D_n)\}$ must be considered as a vector.

It is shown that

$$X_i = \phi_i - \sum_{k=1}^{i-1} (\phi_i K X_k) X_k \quad (\text{for } i = 1, X_1 = \phi_1), \tag{5.1}$$

where X_i and ϕ_i are, respectively, the ith rows of $X\{W(D_n)\}$ and $\phi\{W(D_n)\}$ and K is a diagonal matrix whose elements are

$$(K_{\alpha\beta}) = \delta_{\alpha\beta}(|C(\alpha)|/2^{n-1}n!),$$

$|C(\alpha)|$ is the order of class $C(\alpha)$ of $W(D_n)$. As before, we must distinguish between n = odd and n = even.

A. n=odd

The procedure established in I for $W(B_n)$ may be applied directly, i.e., working out (5.1) the coefficients of the X_i may be written as a matrix $\Delta\{W(D_n)\}$. So finally,

$$\Delta\{W(D_n)\}X\{W(D_n)\} = \phi\{W(D_n)\}.$$

Remark that for n = odd, Δ is nonsingular and det $\Delta = 1$.

B. n=even

This case presents the same special features that must be taken into account to carry out the calculation of $X\{W(D_n)\}$.

The characters of $W(B_n)$ denoted by two equal subpartitions are the sum of two distinct characters of $W(D_n)$, of the same degree. Consequently, a coefficient $\frac{1}{2}$ must precede each corresponding character of $W(D_n)$.

Note that for n even: (i) Δ is singular (see Table IV for the case $n = 4$), and (ii) to obtain the irreducible characters corresponding to the subdivisions of the classes defined by partitions of n in even positive parts use must be made of the fact that their sum is known and of the orthogonality relations.

VI. THE REDUCTION $W(D_n) \to W(D_{n-1})$: THE SIMPLE CHARACTERS

For n odd, the reduction of the irreducible characters may be formulated in terms analogous to those of the corresponding result for $W(B_n)$ (see I, Theorem 2):

Theorem 6.1: The irreducible characters X_n of $W(D_n)$ reduces into irreducible characters X_{n-1} of $W(D_{n-1})$ according to the equation:

$$X_n = W_{n-1}^n X_{n-1},$$

where

$$W_{n-1}^n = \Delta_n^{-1} F_{n-1}^n \Delta_{n-1}.$$

(For a proof, we refer to I.)

For n even, Δ_n is singular and W_{n-1}^n cannot be calculated directly. We know that the matrices F_{n-1}^n and W_{n-1}^n

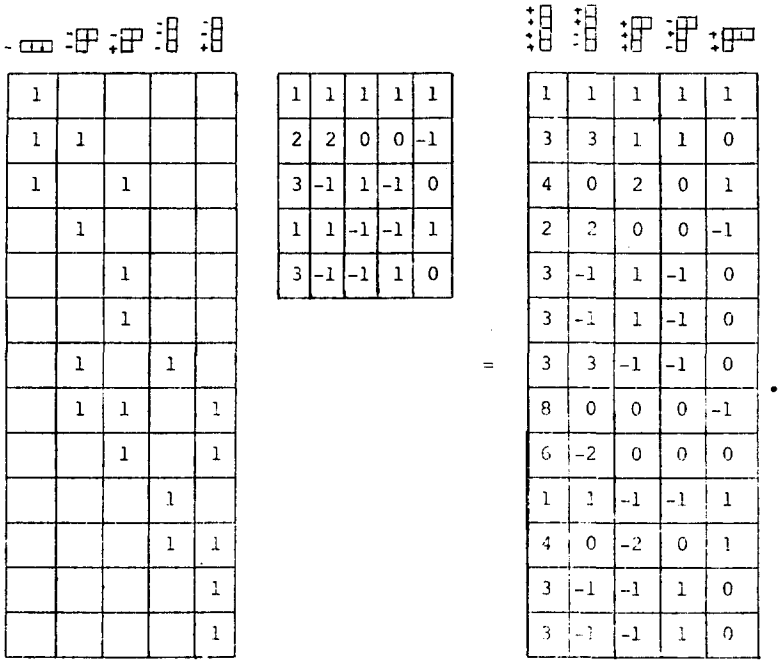
TABLE IV. The matrix Δ for $W(D_4)$.

	1											
	1	1										
	1	1	1									
	1	1	0	1								
$\begin{smallmatrix}\cdot\\\cdot\\\cdot\end{smallmatrix}\!\!\square$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$							
	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$							
	1	2	0	1	0	0	1						
	1	2	1	1	0	0	1	1					
	1	2	2	1	1	1	1	2	1				
	1	3	0	2	0	0	3	0	0	1			
	1	3	1	2	0	0	3	2	0	1	1		
$\begin{smallmatrix}\cdot\\\cdot\\\cdot\end{smallmatrix}\!\!\square$	$\frac{1}{2}$	$\frac{3}{2}$	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	2	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{3}{2}$	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	2	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$

have the same entries $\neq 0$. For W^n_{n-1} these entries are equal to 1. This property is valid $\forall\, n$ (see I). However, for n even, it must be recalled that for classes with $\alpha_1^+ \neq 0$, $(\lambda;\mu) \neq (\mu;\lambda)$ for the induced characters and $(\lambda;\mu) = (\mu;\lambda)$ for the irreducible characters. Hence, the columns of F^n_{n-1} corresponding to $(\mu;\lambda)$ must be omitted in W^n_{n-1}, i.e., the characters $(\mu;\lambda)$ must be placed in the $(\lambda;\mu)$ column.

Example:

$W^4_3 X\{W(D_3)\} = X\{W(D_4)\}$,



In Appendix B, we give a method to verify that W^n_{n-1} has been correctly obtained from F^n_{n-1}.

ACKNOWLEDGMENTS

I am grateful to D. Speiser and J. P. Doeraene for many useful hints. A summary of this article was presented at the XVIII International Colloquium on Group Theoretical Methods in Physics (Moscow, June 1990).

This work was supported in part by FONDECYT (Contract No. 0143-90).

APPENDIX A: THE CLASSES OF $W(D_n)$

Let us analyze the classes decomposition of $W(B_n)$ and $W(D_n)$ for $1 \leqslant n \leqslant 10$:

n	$W(B_n)$ number of classes	classes +	classes −	$W(D_n)$ number of classes
1	2	1	1	1
2	5	3	2	4
3	10	5	5	5
4	20	11	9	13
5	36	18	18	18
6	65	34	31	37
7	110	55	55	55
8	185	95	90	100
9	300	150	150	150
10	481	244	237	251

For n odd the expression given in Sec. II B is obtained at once.

For n even:

n	number of classes of $W(D_n)$	number of classes + of $W(B_n)$	$p(n)_{\text{even parts}}$
2	4	3	1
4	13	11	2
6	37	34	3
8	100	95	5
10	251	244	7

For example, $p(8)_{\text{even parts}} = 5\{8,62,44,422,2222\}$ [obviously, $p(n)_{\text{even parts}} = p(n/2)$].

Hence, the expression given in Sec. II B is obtained [$W(D_4)$ has been considered in Ref. 4].

Remarks: (1) for n even, the relation between the positive and the negative classes of $W(B_n)$ is

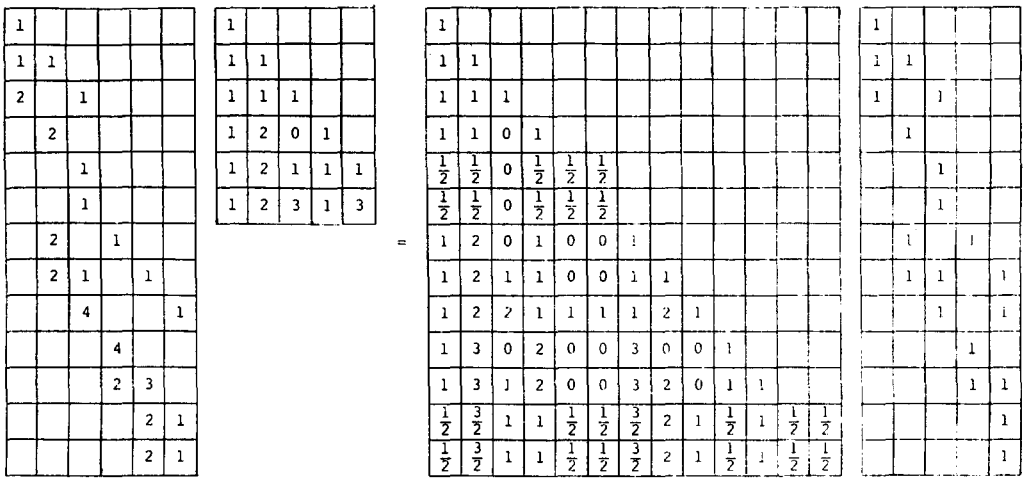
$$\{\text{number of classes} + \} = \{\text{number of classes} - \} + p(n)_{\text{even parts}}$$

if n odd, $p(n)$ even parts $= 0$, $\therefore$ (number of classes +) = (number of classes −)].

(2) Using the expression for the number of classes of $W(D_n)$ and for the number of classes +, we have

$$\text{number of classes of } W(D_n) = \left\{ \begin{matrix} \text{number of classes} - \\ \text{of } W(B_n) \end{matrix} \right\} + 2p(n)_{\text{even parts}}.$$

TABLE V. The equation $F^n_{n-1}\Delta_{n-1} = \Delta_n W^n_{n-1}$ $(n=4)$.



APPENDIX B: THE EQUATION $F^n_{n-1}\,\Delta_{n-1} = \Delta_n W_{n-1}$ (n=even)

In general, we have (see I)

$$X_n = W^n_{n-1} X_{n-1}, \tag{B1}$$

$$\phi_n = F^n_{n-1}\phi_{n-1}, \tag{B2}$$

$$\phi_n = \Delta_n X_n. \tag{B3}$$

Multiplying from the left by Δ_n, Eq. (B1) becomes

$$\Delta_n X_n = \Delta_n W^n_{n-1} X_{n-1}.$$

From Eq. (B3):

$$\phi_n = \Delta_n X_n = \Delta_n W^n_{n-1} X_{n-1}.$$

From (B2):

$$F^n_{n-1}\phi_{n-1} = \Delta_n W^n_{n-1} X_{n-1},$$

but

$$\phi_{n-1} = \Delta_{n-1} X_{n-1}$$

and

$$F^n_{n-1}\Delta_{n-1} X_{n-1} = \Delta_n W^n_{n-1} X_{n-1},$$

hence,

$$F^n_{n-1}\Delta_{n-1} = \Delta_n W^n_{n-1}. \tag{B4}$$

Since F^n_{n-1}, Δ_{n-1}, and Δ_n are known, Eq. (B4) provides a verification for W^n_{n-1} (n even).

Note: F^n_{n-1} has an added column, in consequence Δ_{n-1} must have a supplementary line.

Table V gives expression (B4) for the case $n = 4$.

APPENDIX C: ON THE DIMENSIONS OF THE REPRESENTATIONS OF $W(D_n)$

In I we gave a procedure to evaluate the dimensions of the irreducible and induced representations of the hyperoctahedral group $W(B_n)$. The dimensions were obtained via the products of the reduction matrices W and F, respectively. The same method may be applied to the present case, i.e., the dimensions of the irreducible representations of $W(D_n)$ are given by

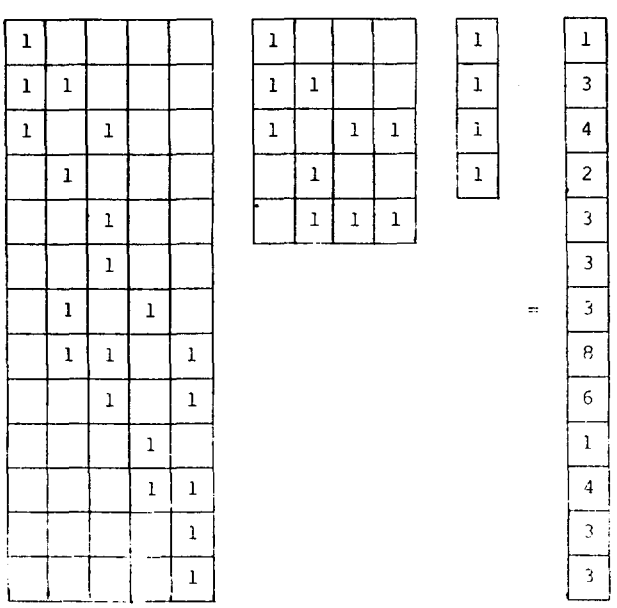
$$W^n_{n-1} W^{n-1}_{n-2} \cdots W^2_1$$

and the dimensions of the induced representations are given by

$$F^n_{n-1} F^{n-1}_{n-2} \cdots F^2_1.$$

Example: (a) Dimensions of the irreducible representations of

$$W(D_4) = W^4_3 W^3_2 W^2_1,$$



(b) Dimensions of the induced representations of $W(D_4) = F_3^4 F_2^3 F_1^2$,

1					
1	1				
2		1			
	2				
		1			
		1			
	2		1		
	2	1		1	
		4			1
			4		
			2	3	
				2	1
				2	1

1			
1	1		
2		1	1
	3		
	2	2	2
	4	4	4

1
2
2
2

=

1
4
8
6
6
6
12
24
48
24
48
48
48

.

[1]J. E. Mandula, G. Zweig, and J. Govaerts, Nucl. Phys. B **228**, 109 (1983).
[2]M. Baake, J. Math. Phys. **25**, 3171 (1984).
[3]P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986).
[4]*Proceedings of the XVIIth International Colloquium on Group Theoretical Methods in Physics*, Sainte-Adele, Canada, 1988, edited by Y. Saint-Aubin and L. Vinet (World Scientific, Singapore, 1989). (Specially the works of Affleck, Creutz, D'Hoker, and Jimbo, therein.)
[5]H. S. M. Coxeter, *Regular Polytopes* (Dover, New York, 1973), 3rd ed., Chap. XI.
[6]R. C. King, J. Phys. A: Math. Gen. **8**(4), 429 (1975).
[7]A. De Baenst-Vandenbroucke, P. De. Baenst, and D. Speiser, Proc. R. Irish Acad. A **73**, 131 (1973).
[8]M. Aguirre, R. Buzzo, and G. Iommi Amunátegui, Proc. R. Irish Acad. A **82**, 33 (1982).
[9]R. C. King, J. Math. Phys. **15**, 258 (1974).
[10]H. Weyl, *The Theory of Groups and Quantum Mechanics*, translated by H. P. Robertson (Dover, New York, 1949), 2nd ed., p. 390.
[11]J. P. Doeraene and G. Iommi Amunátegui, J. Math. Phys. **30**, 2469 (1989).
[12]A. Young, Proc. London Math. Soc. **2**, 31, 273 (1930).
[13]S. J. Mayer, Math. Proc. Cambridge Philos. Soc. **77**, 259 (1975).
[14]G. James and A. Kerber, *The Representation Theory of the Symmetric Group* (Addison-Wesley, Reading, MA, 1981).
[15]A. C. Hurley, Proc. Cambridge Philos. Soc. **47**, 650 (1951).
[16]M. Baake, B. Gemünden, and R. Oedingen, J. Math. Phys. **23**, 944 (1982).
[17]M. Baake, J. Math. Phys. **25**, 3171 (1984).

Point symmetries of conditionally integrable nonlinear evolution equations

J. Rubin and P. Winternitz
Centre de Recherches Mathématiques, Université de Montréal, CP 6128-A, Montréal, Quebec H3C 3J7, Canada

(Received 22 November 1989; accepted for publication 28 March 1990)

The Lie point symmetries of the first two equations in the Kadomtsev–Petviashvili (KP) hierarchy, introduced by Jimbo and Miwa, are investigated. The first is the potential KP equation, the second involves four independent variables and is called the Jimbo–Miwa (JM) equation. The joint symmetry algebra for the two equations is shown to have a Kac–Moody–Virasoro structure, whereas the symmetry algebra of the JM equation alone does not. Subgroups of the joint symmetry group are used to perform symmetry reduction and to obtain invariant solutions.

I. INTRODUCTION

A recent article[1] was devoted to the integrability of equations in the Kadomtsev–Petviashvili (KP) hierarchy of equations, introduced by Jimbo and Miwa.[2] This is an infinite sequence of equations, involving an increasing number of independent variables. Jimbo and Miwa[2] gave a τ function solution to the KP hierarchy that represents an N-soliton solution.

Reference 1 concentrated on the second equation in the KP hierarchy. It was called the "Jimbo–Miwa equation" (JM) and was written in the form

$$w_{xxxy} + 3w_{xy}w_x + 3w_y w_{xx} + 2w_{yt} - 3w_{xz} = 0, \quad (1.1)$$

where w is a real scalar function of the real variables x,y,z,t and the subscripts denote partial derivatives. The first equation in the KP hierarchy is the Kadomtsev–Petviashvili equation itself. In its potential form it can be written as

$$w_{xxxx} + 6w_x w_{xx} + 3w_{yy} - 4w_{xt} = 0. \quad (1.2)$$

It was shown that the JM equation (1.1) alone does not pass any of the conventional integrability tests.[1] Thus the equation, as a PDE, does not have the Painlevé property, as defined by Weiss *et al.*[3] On the other hand, if one considers a subset of solutions of (1.1), that for fixed $z = z_0$ also satisfy the potential KP equation, then these solutions do pass the Painlevé test.[1] Furthermore, Eq. (1.1) was shown to have two types of solitary wave solutions, satisfying two different dispersion relations.[1] One type also satisfies the PKP equation (1.2) for $z = z_0$ fixed, the other does not. Only the first type gives rise to N-soliton solutions with $N \geqslant 3$. The solitary waves of the second type do not interact like solitons, i.e., they do not survive mutual interactions. In view of these properties, the concept of "conditionally integrability" was introduced.[1] The equations in the KP hierarchy of Jimbo and Miwa are integrable, under the condition that they are solved simultaneously with all the preceeding equations in the hierarchy. This implies that the usual linear integration techniques[2,4,5] will only provide a small subclass of solutions of the higher-order equations in the hierarchy.

The aim of this article is to investigate the Lie point symmetries of the JM equation (1.1). More specifically, it was shown that the Lie algebra of the symmetry group of Eq. (1.1) is infinite dimensional,[1] but that it does not have the Kac–Moody–Virasoro loop structure, typical for integrable equations in 2 + 1 or 3 dimensions. The symmetry algebras of the Kadomtsev–Petviashvili equation, the potential KP equation, the Davey–Stewartson equation, the three-wave resonant interaction equations, and several others have all been shown[6–10] to be of the Kac–Moody–Virasoro type.[2,11,12] Here we shall show that the symmetry algebra of the set of *two* equations (1.1) and (1.2) does have a Kac–Moody–Virasoro structure. This further justifies the use of the term "conditional integrability" for such equations.

A concept of "conditional invariance" of differential equations has also been introduced in the literature,[13–16] (not necessarily always under the same name). It refers to a group of local point transformations leaving a subset of solutions of a given equation invariant, rather than the set of all solutions. The symmetry group of the system of equations (1.1) and (1.2) can thus be viewed as the group of "conditional symmetries" of the JM equation, the condition being that the solutions also satisfy the PKP equation (1.2).

II. THE LIE ALGEBRA OF THE CONDITIONAL SYMMETRY GROUP

A. Determination of the Lie algebra

In order to find the symmetry algebra of the JM + PKP system (1.1) and (1.2), we use a standard algorithm,[17] implemented as a MACSYMA package[18] (a REDUCE package is also available[19]). The Lie algebra is realized by vector fields of the form

$$V = \eta_i \frac{\partial}{\partial_{x_i}} + \phi \frac{\partial}{\partial_w}, \quad x_1 = x,\ x_2 = y,\ x_3 = z,\ x_4 = t, \quad (2.1)$$

where η_i and ϕ are functions of x_i and w. We rewrite the system (1.1) and (1.2) in the symbolic form

$$\Delta_i = 0, \quad i = 1,2,$$

and request that the fourth prolongation[17] of the vector field V should annihilate the equations on their joint solution space:

$$\mathrm{pr}^{(4)}\Delta_i|_{\Delta_1 = 0, \Delta_2 = 0} = 0. \quad (2.2)$$

Equation (2.2) implies a system of first-order linear partial differential equations for the coefficients η_i and ϕ in (2.1). The fact that we are restricting ourselves to local point transformations is reflected in that η_i and ϕ depend on x, y, z, t, and w, but not on derivatives of w.

The determining equations are easy to solve and we ob-

 0022-2488/90/092085-06$03.00

tain an infinite-dimensional symmetry algebra, depending on four arbitrary functions of one variable and one function of two variables. More explicitly, we have

$$V = Z(f) + T(g) + Y(h) + X(k) + W(G), \quad (2.3a)$$

where

$$Z(f) = f\partial_z + \tfrac{1}{4}[xf' + \tfrac{3}{2}ytf'' + \tfrac{9}{32}t^3 f''']\partial_x + \tfrac{1}{2}[yf' + \tfrac{9}{16}t^2 f'']\partial_y + \tfrac{3}{4}tf'\partial_t - \tfrac{1}{4}[wf' + xyf'' + \tfrac{3}{4}t(y^2 + \tfrac{3}{4}tx)f''' + \tfrac{9}{32}yt^3 f''']\partial_w, \quad (2.3b)$$

$$T(g) = g\partial_t + \tfrac{1}{32}[16yg' + 9t^2 g'']\partial_x + \tfrac{3}{4}tg'\partial_y - \tfrac{1}{32}[4(3tx + 2y^2)g'' + 9yt^2 g'']\partial_w, \quad (2.3c)$$

$$Y(h) = h\partial_y + \tfrac{3}{4}th'\partial_x - \tfrac{1}{4}[2xh' + 3tyh'']\partial_w, \quad (2.3d)$$

$$X(k) = k\partial_x - yk'\partial_w, \quad (2.3e)$$

$$W(G) = G(z,t)\partial_w. \quad (2.3f)$$

The functions $f(z)$, $g(z)$, $h(z)$, and $k(z)$ are C^∞ on some open set $U \in \mathbb{R}$, $G(z,t)$ is C^∞ on some open subset of $\mathbb{R}\times\mathbb{R}$; the dots denote z derivatives and G_z, G_t below will denote partial derivatives.

The commutation relations for the Lie algebra (2.3) can be summed up as follows:

$$[Z(f_1), Z(f_2)] = Z(f_1 f_2' - f_1' f_2), \quad (2.4a)$$

$$[Z(f), T(g)] = T(fg' - \tfrac{3}{4}f'g),$$

$$[Z(f), Y(h)] = Y(fh' - \tfrac{1}{2}f'h), \quad (2.4b)$$

$$[Z(f), X(k)] = X(fk' - \tfrac{1}{4}kf'),$$

$$[Z(f), W(G)] = W(fG_z + \tfrac{1}{4}f(3tG_t + G)),$$

$$[T(g_1), T(g_2)] = \tfrac{3}{4}Y(g_1 g_2' - g_1' g_2),$$

$$[T(g), Y(h)] = \tfrac{1}{4}X(3gh' - 2g'h),$$

$$[T(g), X(k)] = \tfrac{3}{8}W(t(kg'' - 2k'g')),$$

$$[T(g), W(G)] = W(gG_t), \quad (2.4c)$$

$$[Y(h_1), Y(h_2)] = \tfrac{3}{4}W(t(h_1' h_2 - h_1 h_2')),$$

$$[Y(h), X(k)] = \tfrac{1}{2}W(h'k - 2hk'),$$

$$[Y(h), W(G)] = [X(k_1), X(k_2)] = [X(k), W(G)] = [W(G_1), W(G_2)] = 0. \quad (2.4d)$$

The commutation relations (2.4) show that the algebra L has a Levi decomposition[20] (a nontrivial statement for an infinite-dimensional Lie algebra). Indeed, we have

$$L = S \triangleright R, \quad (2.5)$$

where $S = \{Z(f)\}$ is a simple Lie algebra, namely the centerless Virasoro algebra, isomorphic to the Lie algebra of real smooth vector fields on $\mathbb{R}$ (one of Cartan's infinite-dimensional simple Lie algebras[21]). The radical (maximal solvable ideal) $R = \{T(g), X(k), Y(h), W(G)\}$ is also infinite dimensional.

B. A Kac–Moody–Virasoro subalgebra

Contrary to the case of integrable systems in three dimensions,[6–10] we cannot directly identify R as a subalgebra of a Kac–Moody algebra, in view of the presence of the Abelian ideal $\{W(G)\}$, where $G(x,t)$ depends on two variables.

The radical R does however contain a Kac–Moody type subalgebra $R_{KM} \subset R$. To obtain a basis for R_{KM} we expand the functions $g(z)$, $k(z)$, $h(z)$, and $G(z,t)$ into Laurent series and consider the Lie algebra, spanned by

$$R_{KM} = \{T(z^n), X(z^n), Y(z^n), W(z^n, t^a) \mid n \in \mathbb{Z}, a = 0,1,\ldots,5\}. \quad (2.6)$$

That R_{KM} is a Lie algebra follows from the commutation relations (2.4), as does the fact that R_{KM} is an ideal in the algebra

$$L_{KMV} = S_v \triangleright R_{KM}, \quad S_v = \{Z(z^n) \mid n \in \mathbb{Z}\}. \quad (2.7)$$

The algebra L_{KMV} is a Kac–Moody–Virasoro algebra in which the Kac–Moody part is based on a 19-dimensional solvable Lie algebra $\mathcal{L}_0$ with the following basis:

$$\begin{aligned}
&Z_1 = x\partial_x + 2y\partial_y + 3t\partial_t - w\partial_w,\\
&Z_2 = 12yt\partial_x + 9t^2\partial_y - 8xy\partial_w,\\
&Z_3 = 3t^3\partial_x - 2t(4y^2 + 3tx)\partial_w, \quad Z_4 = yt^3\partial_w,\\
&T_1 = \partial_t, \quad T_2 = 2y\partial_x + 3t\partial_y,\\
&T_3 = 9t^2\partial_x - 4(3tx + 2y^2)\partial_w, \quad T_4 = yt^2\partial_w,\\
&Y_1 = \partial_y, \quad Y_2 = 3t\partial_x - 2x\partial_w, \quad Y_3 = ty\partial_w,\\
&X_1 = \partial_x, \quad X_2 = y\partial_w,\\
&W_k = t^{k-1}\partial_w, \quad k = 1,2,\ldots,6.
\end{aligned} \quad (2.8)$$

The nilradical of $\mathcal{L}_0$ is 18 dimensional, spanned by the basis elements (2.8) without Z_1. The largest Abelian ideal $A_0 \subset \mathcal{L}_0$ is 11 dimensional: $\{W_1, W_2, W_3, W_4, W_5, W_6, X_1, X_2, Y_3, T_4, Z_4\}$. According to Ado's theorem,[22] any finite-dimensional Lie algebra can be imbedded into $sl(n,\mathbb{R})$ for large enough n. The fact that we have $\dim A_0 = 11$ already implies $n \geqslant 7$. If we require that each basis element in (2.6) has a specific degree in a grading, that the degree of Z_1 be zero, and that the degree correspond to the distance of the first entry in the corresponding $sl(n,\mathbb{R})$ matrix from the diagonal, we obtain $n = 14$. The algebra

$$\mathcal{L}_{KMV} = \{Z(z^n), T(z^n), X(z^n), Y(z^n), W(z^n, t^a), n \in \mathbb{R}, a = 0,1,\ldots,5\}, \quad (2.9)$$

is then identified as a subalgebra of the Kac–Moody–Virasoro algebra $\widehat{sl}(14,\mathbb{R})$.

C. Comparison with the symmetry algebras of the PKP and JM equations

The symmetry algebra L_{PKP} of the PKP equation[7] can be summed up as

$$\hat{V} = \hat{T}(a) + \hat{Y}(b) + \hat{X}(c) + \hat{W}(d) + \hat{U}(e), \quad (2.10a)$$

where a, b, c, d, and e are C^∞ functions of t and

$$\hat{T}(a) = a\partial_t + \tfrac{2}{3}ya'\partial_y + \tfrac{1}{3}[a'x + \tfrac{2}{3}a''y^2]\partial_x - [\tfrac{1}{3}wa' + \tfrac{1}{9}x^2 a'' + \tfrac{4}{27}xy^2 a''' + \tfrac{4}{243}y^4 a'''']\partial_w, \quad (2.10b)$$

$$\hat{Y}(b) = b\partial_y + \tfrac{2}{3}b'y\partial_x - \tfrac{4}{9}y[xb'' + \tfrac{2}{9}y^2 b''']\partial_w, \quad (2.10c)$$

$$\hat{X}(c) = c\partial_x - \tfrac{2}{3}[c'x + \tfrac{2}{3}c''y^2]\partial_w, \quad (2.10d)$$

$$\hat{W}(d) = dy\,\partial_w, \quad \hat{U}(e) = e\partial_w. \quad (2.10e,f)$$

The primes denote time derivatives. Notations differ slightly from those of Ref. 7. A basis for the symmetry algebra L_{JM} of the JM equation can, on the other hand, be written as[1]:

$$P_z = \partial_z, \quad P_t = \partial_t,$$
$$D_1 = z\partial_z + y\partial_y, \quad D_2 = 3t\partial_t + x\partial_x - 2y\partial_y - w\partial_w,$$
$$R(g) = g(t)\partial_x + \tfrac{2}{3}\dot{g}(t)x\partial_w,$$
$$Y(h), \quad X(k), \quad W(G), \tag{2.11}$$

where $Y(h)$, $X(k)$, and $W(G)$ are given in (2.3d)–(2.3f), respectively.

Let us now compare the symmetry algebra (2.3) of the joint JM + PKP system with that of the PKP system alone namely (2.10). To do this we must consider the vector fields (2.3) as acting on functions of x, y, t, and w. The derivative ∂/∂_z acts trivially and the functions $f(z),\ldots,\overset{\cdots}{f}(z),g(z),\ldots,\overset{\cdots}{g}(z)$, etc. are to be considered as independent constants. In this restricted sense the symmetry algebra of the JM + PKP system is a subalgebra of the PKP algebra and can be written as

$$\begin{aligned} Z_R(f) &= \tfrac{1}{4}\dot{f}(z)\hat{T}(t) + \tfrac{9}{32}\ddot{f}(z)\hat{Y}(t^2) \\ &\quad + \tfrac{9}{128}\dddot{f}(z)\hat{X}(t^3) - \tfrac{9}{128}\overset{\cdots}{f}(z)\hat{W}(t^3), \\ T_R(g) &= g(z)\hat{T}(1) + \tfrac{3}{4}\dot{g}(z)\hat{Y}(t) \\ &\quad + \tfrac{9}{32}\ddot{g}(z)\hat{X}(t^2) - \tfrac{9}{32}\dddot{g}(z)\hat{W}(t^2), \\ Y_R(h) &= h(z)\hat{Y}(1) + \tfrac{3}{4}\dot{h}(z)\hat{X}(t) - \tfrac{3}{4}\ddot{h}(z)\hat{W}(t), \\ X_R(k) &= k(z)\hat{X}(1) - \dot{k}(z)\hat{W}(1), \\ W(G) &= \sum_n G_n(z)\hat{U}(t^n). \end{aligned} \tag{2.12}$$

On the other hand, the JM + PKP algebra (2.3) is not a subalgebra of the JM algebra, nor is the converse true. The algebras (2.3) and (2.11) have a large intersection, namely $\{Y(h),X(k),W(G),P_z,P_t\}$, but neither one is a subalgebra of the other.

III. THE CONDITIONAL SYMMETRY GROUP

A. The local Lie point transformations

Any one-dimensional subgroup of the symmetry group of the JM + PKP system can be obtained by integrating a vector field (2.3). The vector fields have the form (2.1), so we must solve the differential system:

$$\frac{d\tilde{x}_i}{d\lambda} = \eta_i(\tilde{x},\tilde{w}), \quad \frac{d\tilde{w}}{d\lambda} = \phi(\tilde{x},\tilde{w}), \quad \tilde{x}_i|_{\lambda=0} = x_i,$$
$$\tilde{w}|_{\lambda=0} = w. \tag{3.1}$$

By construction, the result of integrating (3.1) is a local point tranformation of the form

$$\tilde{x} = \Lambda_\lambda(x,w), \quad \tilde{w} = \Omega_\lambda(x,w), \tag{3.2}$$

where the functions Λ and Ω are defined in some neighborhood of the identity element ($\lambda = 0$) and some neighborhood of the origin of the (x,w) space.

We shall construct the one-dimensional subgroups, corresponding to the individual vector fields (2.3b)–(2.3f). More general transformations are obtained by composing a finite, or at least in principle, infinite number of one-dimensional ones.

Let us now run through the individual types of subgroups:

1. The algebra W(G)

The corresponding group transformations is

$$\tilde{x}_i = x_i, \quad \tilde{w}(\tilde{x}_i) = w(x_i) + \lambda G(z,t). \tag{3.3}$$

This is a pure gauge transformation, acting on the solution w, but not on space-time itself.

2. The algebra X(k)

We obtain

$$\tilde{x} = x + \lambda k(z), \quad \tilde{y} = y, \quad \tilde{z} = z, \quad \tilde{t} = t,$$
$$\tilde{w}(\tilde{x},\tilde{y},\tilde{z},\tilde{t}) = w(x,y,z,t) - \lambda y\dot{k}(z). \tag{3.4}$$

3. The algebra Y(h)

The transformation is

$$\begin{aligned} &\tilde{x} = x + \tfrac{3}{4}\lambda t\dot{h}(z), \quad \tilde{y} = y + \lambda h(z), \quad \tilde{z} = z, \quad \tilde{t} = t, \\ &\tilde{w}(\tilde{x},\tilde{y},\tilde{z},\tilde{t}) = w(x,y,z,t) - \tfrac{1}{4}\lambda(2x\dot{h} + 3ty\ddot{h}) \\ &\qquad - \tfrac{3}{16}\lambda^2 t(\dot{h}^2 + 2h\ddot{h}). \end{aligned} \tag{3.5}$$

4. The algebra T(g)

Integrating (2.3c) we obtain

$$\begin{aligned} &\tilde{z} = z, \quad \tilde{t} = t + \lambda g(z), \quad \tilde{y} = y + \tfrac{3}{8}\dot{g}(2\lambda t + \lambda^2 g), \\ &\tilde{x} = x + \tfrac{1}{32}[(16\dot{g}y + 9\ddot{g}t^2)\lambda + 3(2\dot{g}^2 + 3g\ddot{g})t\lambda^2 \\ &\qquad + (2\dot{g}^2 + 3g\ddot{g})\lambda^3 g], \\ &\tilde{w}(\tilde{x},\tilde{y},\tilde{z},\tilde{t}) = w(x,y,z,t) + A\lambda + B\lambda^2 \\ &\qquad + C\lambda^3 + D\lambda^4 + E\lambda^5, \\ &A = -(1/2^5)[4(2y^2 + 3tx)\ddot{g} + 9t^2y\dddot{g}], \\ &B = -(3/2^9)[48yt(\dot{g}\ddot{g} + g\dddot{g}) + 9t^3(\ddot{g}^2 + 2\dot{g}\dddot{g}) \\ &\qquad + 32xg\dddot{g}], \\ &C = -(1/2^8)[18t^2(\dot{g}^2\ddot{g} + g\ddot{g}^2) + 8yg(4\dot{g}\ddot{g} + 3g\dddot{g}) \\ &\qquad + 45t^2g\ddot{g}\dddot{g}], \\ &D = -(3/2^8)tg[5\ddot{g}\dot{g}^2 + 3g\ddot{g}^2 + 9g\dot{g}\dddot{g}], \\ &E = -(3/5.2^8)g^2[5\dot{g}^2\ddot{g} + 3g\ddot{g}^2 + 9g\dot{g}\dddot{g}]. \end{aligned} \tag{3.6}$$

5. The algebra Z(f)

We integrate (2.3b) to obtain

$$\tilde{z} = \phi^{-1}(\lambda + \phi(z)), \quad \phi(z) = \int_{z_0}^{z} \frac{ds}{f(s)},$$
$$\tilde{t} = \left[\frac{f(\tilde{z})}{f(z)}\right]^{3/4} t$$
$$\tilde{y} = \left[\frac{f(\tilde{z})}{f(z)}\right]^{1/2}\left[y + \frac{9}{32}t^2\frac{(\dot{f}(\tilde{z}) - \dot{f}(z))}{f(z)}\right],$$
$$\tilde{x} = \left(\frac{f(\tilde{z})}{f(z)}\right)^{1/4}[x + \alpha(z,\lambda)ty + \beta(z,\lambda)t^3],$$

$$\tilde{w} = \left(\frac{f(\tilde{z})}{f(z)}\right)^{-1/4} [w + \gamma(z,\lambda)xy + \delta(z,\lambda)y^2 t + \mu(z,\lambda)yt^3 + w(z,\lambda)xt^2 + \sigma(z,\lambda)t^5].$$

The expressions $\alpha(z,\lambda),\ldots,\sigma(z,\lambda)$ are easy to calculate, but they are cumbersome to present and not very informative, so we shall not present them here. They all vanish for $\lambda = 0$.

B. A global physical subgroup

Intuitively, or "physically" speaking, the most obvious symmetries of the JM + PKP system are the global ones. The corresponding subalgebra of the symmetry algebra (2.3) is obtained by restricting the functions $f(z)$, $g(z)$, $h(z)$, and $k(z)$ to be first-order polynomials in z. The function $G(z,t)$ will correspond to a global gauge tranformation if it has no singularities for finite values of z and t. The global transformations acting nontrivially on the space-time variables are generated by

$$X(1) = \partial_x, \quad Y(1) = \partial_y, \quad Z(1) = \partial_z, \quad T(1) = \partial_t,$$
$$X(z) = z\partial_x - y\partial_w, \quad Y(z) = z\partial_y + \tfrac{3}{4}t\partial_x - \tfrac{1}{2}x\partial_w,$$
$$Z(z) = z\partial_z + \tfrac{1}{4}x\partial_x + \tfrac{1}{2}y\partial_y + \tfrac{3}{4}t\partial_t - \tfrac{1}{4}w\partial w,$$
$$T(z) = z\partial_t + \tfrac{1}{2}y\partial_x + \tfrac{3}{4}t\partial_y.$$

It is now obvious that $X(1)$, $Y(1)$, $Z(1)$, and $T(1)$ generate translations in the x, y, z, and t directions, respectively. Further, $X(z)$ generates a "shear" transformation, acting on x, parallel to z:

$$x' = x + \lambda z, \quad y' = y, \quad z' = z, \quad t' = t, \quad w' = w - \lambda y.$$

Similarly, $Y(z)$ generates a shear, acting on y, parallel to z, accompanied by a Galilei transformation in the x direction:

$$x' = x + \tfrac{3}{4}\lambda t, \quad y' = y + \lambda z,$$
$$z' = z, \quad t' = t, \quad w' = w - \tfrac{1}{2}x\lambda.$$

The vector field $Z(z)$ generates a dilation with scale factors $1, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and $-\frac{1}{4}$ in the z, x, y, t and w directions, respectively. Finally, $T(z)$ generates the transformation

$$z' = z, \quad t' = t + \lambda z, \quad y' = y + \tfrac{3}{4}\lambda t + \tfrac{3}{8}\lambda^2 z,$$
$$x' = x + \tfrac{1}{2}\lambda y + \tfrac{3}{16}\lambda^2 t + \tfrac{1}{16}\lambda^3 z, \quad w' = w,$$

a linear transformation in the x, y, t plane, accompanied by a shear parallel to z.

IV. SOLUTIONS OBTAINED BY SYMMETRY REDUCTION

In order to obtain all types of group invariant solutions of the PKP + JM system we must perform a classification of the low-dimensional subgroups of the symmetry group. More specifically, we need to classify all subgroups having generic orbits of codimension 2, 3, and 4 in the space $X \times U$ of independent and dependent variables, with nonzero projections onto U. Such a classification is easy to perform, using methods developed for finite-dimensional Lie algebras[23,24] that have also been extended to Kac–Mooody–Virasoro algebras.[9,25]

It is not our aim to study solutions of Eqs. (1.1) and (1.2) systematically, so we shall only give some illustrative examples.

A. Reductions obtained using one-dimensional subgroups

Every one-dimensional subalgebra of the symmetry algebra (2.3) is conjugate, under the adjoint action of the symmetry group of the JM + PKP equations, to one of the following ones:

$$Z(1), \; T(1), \; Y(1) + W(G), \; X(1) + W(G), \; W(G), \tag{4.1}$$

where $G(z,t)$ is an arbitrary function.

Invariance under the group generated by $Z(1)$ simply implies that the solution of the JM + PKP system satisfies $w = w(x,y,t)$, where w simultaneously satisfies the PKP equations and the JM equations (1.1) from which the term w_{xz} is dropped. Invariance under the group generated by $T(1)$ implies $w = w(x,y,z)$, where w satisfies a potential Boussinesq equation, together with Eq. (1.1), in which the term w_{yt} is dropped.

Invariance under the group generated by $Y(1)$ implies

$$w = w(x,z,t) = f(x,t) + g(z,t), \tag{4.2}$$

where $g(z,t)$ is arbitrary and $f(x,t)$ satisfies the once-differentiated potential Korteweg–de Vries equation

$$[f_{xxx} + 3f_x^2 - 4f_t]_x = 0. \tag{4.3}$$

For $Y(1) + W(G)$, $G \neq 0$ invariance implies

$$w = yG(z,t) + f(x,z,t), \tag{4.4}$$

where f satisfies (4.3) (with z considered as a fixed parameter), and

$$3Gf_{xx} - 3f_{xz} + 2G_t = 0. \tag{4.5}$$

The linear equation (4.5) can be solved for $f(x,z,t)$ in the form

$$f = \frac{2}{3}W_t(z,t)x + g(z,t) + \int r(\xi,t)d\xi,$$
$$\xi = x + W(z,t), \quad W(z,t) = \int G(z,t)dz, \tag{4.6}$$

where $g(z,t)$ is arbitrary. From (4.3) we find that $r(\xi)$ must satisfy a Korteweg–de Vries equation with a right-hand side, namely

$$r_{\xi\xi\xi} + 6rr_\xi - 4r_t = \tfrac{8}{3}W_{tt}(z,t). \tag{4.7}$$

Finally, the invariant solution is

$$w(x,y,z,t) = yG(z,t) + \frac{2}{3}x\int G_t(z,t)dz + \int r(\xi,t)d\xi + g(z,t), \tag{4.8}$$

where $G(z,t)$ and $g(z,t)$ are arbitrary functions and $r(\xi,t)$ satisfies (4.7).

The group generated by $X(1)$ leads to the solution

$$w = A(z)y + B(z,t), \tag{4.9}$$

where $A(z)$ and $B(z,t)$ are arbitrary functions.

Finally, the subgroup corresponding to $X(1) + W(G)$ for $G(z,t) \neq 0$ provides the solution

$$w = [\alpha(z)t + \beta(z)]x + \tfrac{2}{3}\alpha(z)y^2 + [\tfrac{3}{4}\dot{\alpha}(z)t^2 + \tfrac{3}{2}\dot{\beta}(z)t + \gamma(z)]y + L(z,t), \tag{4.10}$$

where $\alpha(z)$, $\beta(z)$, $\gamma(z)$, and $L(z,t)$ are arbitrary and $G(z,t)$ is restrained to being

$$G(z,t) = \alpha(z)t + \beta(z), \tag{4.11}$$

and for $G_{tt} \neq 0$ the reduced equations are incompatible.

We shall not consider reductions by two-dimensional subgroups of the symmetry group and instead go over immediately to examples of three-dimensional subgroups, providing reductions to ordinary differential equations.

B. Reductions obtained using three-dimensional subgroups

1. The algebra {Y(1),T(1),Z(1)}

In this case we have $w = w(x)$. The JM equation (1.1) is satisfied trivially, whereas the PKP equation (1.2) reduces to

$$w_{xxxx} + 6w_x w_{xx} = 0. \tag{4.12}$$

Putting $w = u_x$ and integrating twice, we obtain

$$u_x^2 = -2(u-u_1)(u-u_2)(u-u_3), \quad u_1 \leqslant u_2 \leqslant u_3,$$
$$u_1 + u_2 + u_3 = 0, \tag{4.13}$$

where u_1, u_2, and u_3 are constants.

If the roots u_i are all distinct we obtain a finite periodic solution in terms of Jacobi elliptic functions[26] (cnoidal waves):

$$u(x) = u_2 + (u_3 - u_2)cn^2(\sqrt{(u_3-u_1)/2}(x-x_0), k,$$
$$k = [(u_3-u_2)/(u_3-u_1)]^{1/2}. \tag{4.14}$$

If two of the roots coincide, $u_1 = u_2 < u_3$, we obtain a soliton solution

$$u = u_3 - (u_3 - u_1)\left[\tanh\sqrt{(u_3-u_1)/2}(x-x_0)\right]^2, \tag{4.15}$$

satisfying $u \to u_1$ for $x \to \pm\infty$, $u = u_3$ for $x = x_0$. The group transformations of Sec. III can be applied to the solutions (4.14) and (4.15) to introduce a polynomial dependence on y and t into the argument and a virtually arbitrary dependence on z (via the arbitrary functions $k(z)$, $h(z)$, $g(z)$, and $f(z)$).

2. The algebra {Z(1),Z(z),T(1)}

A group invariant solution will in this case have the form

$$w = F(\xi)y^{-1/2}, \quad \xi = xy^{-1/2}. \tag{4.16}$$

The JM and PKP equations reduce to

$$\xi\dddot{F} + 4\ddot{F} + 6\xi\dot{F}\ddot{F} + 3F\ddot{F} + 6\dot{F}^2 = 0, \tag{4.17a}$$
$$4\dddot{F} + 24\dot{F}\ddot{F} + 3\xi^2\ddot{F} + 15\xi\dot{F} + 9F = 0, \tag{4.17b}$$

respectively. Integrating both equations once and eliminating the third-order terms from the two equations, we obtain a second-order equation, that can again be integrated once to yield a Riccati equation:

$$12\dot{F} + 6F^2 - 3\xi^3 F + 6(A\xi^2 + B\xi + C) = 0, \tag{4.18}$$

where A, B, and C are constants. To linearize it we put

$$F = 2\dot{H}/H, \quad 4\ddot{H} - \xi^3 H + (A\xi^2 + B\xi + C)H = 0. \tag{4.19}$$

Using (4.19) we can calculate $F,\ldots,\dddot{F}$ in terms of H and $\dot{H}$ and substitute back into (4.17). This provides a first-order differential equation for $H(\xi)$, which is actually an algebraic equation for $F(\xi)$. More specifically, we obtain a quadratic equation for $F(\xi)$ from which an explicit algebraic solution is obtained, depending on the three integration constants A, B, and C. We shall not reproduce the result here.

V. CONCLUSIONS

"Conditional symmetries" of a differential equation, or system of differential equations, can differ very significantly from ordinary symmetries. The requirement of conditional symmetry is on one hand more restrictive: Instead of requesting that the given system $\Delta_i = 0$ be left invariant, we add further equations, $\Delta_A = 0$, and request that the combined system be left invariant. On the other hand, only a subset of solutions of the original system is transformed into solutions of this system. The domain of application of the conditional symmetry group G_c is smaller than that of the usual symmetry group G. There is no *a priori* reason for one group to contain the other. Moreover, quite often the intersection of the two tranformation groups, $G \cap G_c$, is simply the identity transformation.

The case under consideration is the JM equation (1.2), for which the PKP equation (1.1) is viewed as the supplementary condition. We have found that the conditional symmetries have the Kac–Moody–Virasoro character, typical for multidimensional integrable equations, whereas the ordinary symmetries do not. We do not have $G_{\rm JM} \subset G_c$, nor $G_c \subset G_{\rm JM}$. The conditional symmetry group G_c can hence be used to obtain interesting new reductions of the JM equations to lower-dimensional equations and hence to obtain new solutions.

On the other hand, if we view the PKP equation as the basic equation and the JM equation as a supplementary condition, then we find $G_c \subset G_{\rm PKP}$, at least when G_c is viewed as a transformation group acting on the space $\{x,y,z = z_0,t,w\}$. Hence, symmetry reduction using the group G_c will not provide new solutions of the PKP equation. Instead, we obtain subclasses of solutions obtained from the $G_{\rm PKP}$ group and can then use the G_c group to generate a z dependence compatible with the evolution according to the JM equation.

ACKNOWLEDGMENTS

This research was performed while one of the authors (Jacques Rubin) was visiting the CRM in Montreal, where his stay was supported by the Québec-France scientific exchange program, project no. 20.022089. P. W.'s research is partly supported by research grants from NSERC of Canada and FCAR du Québec.

[1] B. Dorrizzi, B. Grammaticos, A. Ramani, and P. Winternitz, J. Math. Phys. **27**, 2848 (1986).

[2] M. Jimbo and T. Miwa, Publ. Res. Inst. Math. Sci. (Kyoto Univ.) **19**, 943 (1983).

[3] J. Weiss, M. Tabor, and G. Carnavale, J. Math. Phys. **24**, 522 (1983).

[4] M. J. Ablowitz and M. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).

[5] S. P. Novikov, S. V. Manakov, L. P. Pitaevski, and V. E. Zakharov, *Theory of Solitons—The Inverse Method* (Plenum, New York, 1984).

[6] D. David, N. Kamran, D. Levi, and P. Winternitz, Phys. Rev. Lett **55**, 2111 (1985); J. Math Phys. **27**, 1225 (1986).
[7] D. David, D. Levi, and P. Winternitz, Phys. Lett A **118**, 390 (1986).
[8] B. Champagne and P. Winternitz, J. Math. Phys. **29**, 1 (1988).
[9] L. Martina and P. Winternitz, Ann. Phys. **196**, 231 (1989).
[10] P. Winternitz, in *Symmetries and Nonlinear Phenomena* (World Scientific, Singapore, 1988), pp. 358–375.
[11] V. G. Kac, *Infinite-Dimensional Lie Algebras* (Birkhauser, Boston, 1983).
[12] P. Goddard and D. Olive, Int. J. Mod Phys. A **1**, 303 (1986).
[13] G. W. Bluman and J. D. Cole, J. Math. Mech. **18**, 1025 (1969).
[14] P. J. Olver and Ph. Rosenau, Phys. Lett A **114**, 107 (1986); SIAM J. Appl. Math **47**, 263 (1987).
[15] W. I. Fushchich and A. G. Nikitin, *Symmetries of Maxwell's Equations* (Reidel, Dordrecht, 1987).
[16] D. Levi and P. Winternitz, J. Phys. A **22**, 2915 (1989).
[17] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer, New York, 1986).
[18] B. Champagne and P. Winternitz, Preprint CRM-1278, Montréal, 1985.
[19] F. Schwarz, Computing **34**, 91 (1985).
[20] N. Jacobson, *Lie Algebras* (Dover, New York, 1979).
[21] E. Cartan, *Oeuvres Complètes* (CNRS, Paris, 1984), Vol. 2, pp. 893.
[22] I. D. Ado, Usp. Mat. Nauk. **2**, 159 (1947).
[23] J. Patera, P. Winternitz, and H. Zassenhaus, J. Math Phys. **16**, 1597, 1615 (1975).
[24] J. Patera, R. T. Sharp, P. Winternitz, and H. Zassenhaus, J. Math Phys. **18**, 2259 (1977).
[25] P. Winternitz, Preprint CRM-1619, 1989; to appear in *Partially Integrable Nonlinear Evolution Equations and Their Physical Applications* (Kluwer, Dordrecht, 1989).
[26] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists* (Springer, Berlin, 1971).

The two-dimensional magnetic field problem revisited

N. Anghel[a)]
Department of Mathematics, The Ohio State University, Columbus, Ohio 43210

(Received 1 September 1988; accepted for publication 9 May 1990)

The Atiyah–Patodi–Singer index theorem is used to relate the analytic and topological indices of a spin $\frac{1}{2}$-charged particle in a two-dimensional magnetic field.

I. INTRODUCTION

Let D be the first-order elliptic differential operator

$$\frac{\partial}{\partial x}+i\frac{\partial}{\partial y}+\frac{\partial\phi}{\partial x}+i\frac{\partial\phi}{\partial y}, \tag{1.1}$$

defined on $C^\infty(\mathbf{R}^2,\mathbf{C})$, where $\phi\in C^\infty(\mathbf{R}^2)$ is a real valued function such that $\phi = F\ln r$, where F is some real constant, for $r=\sqrt{x^2+y^2}$ large enough.

Despite the fact that the unique closed extension of $D_{|C_0^\infty(\mathbf{R}^2)}$ in $L^2(\mathbf{R}^2)$, still denoted by D, is not a Fredholm operator, D and its formal adjoint D^* do have finite-dimensional L^2 kernels, and then the analytic L^2 index of D is defined, as usual, by

$$L^2\,\mathrm{index}(D) := \dim\ker(D)\cap L^2(\mathbf{R}^2) - \dim\ker(D^*)\cap L^2(\mathbf{R}^2).$$

The kernels and this analytic index were analyzed in detail in Refs. 1 and 2.

In Ref. 3 Bollé *et al.* studied the Witten index associated to D and found that this was precisely the magnetic flux, $(1/2\pi)\int_{\mathbf{R}^2}\Delta\phi = F$, which is also the topological index of Atiyah and Singer.[4]

Since the two indices do not coincide, in general, it was asked[3,5] what the nature of their difference was. The Atiyah–Patodi–Singer index theorem for manifolds with boundary and nonlocal boundary conditions[6] provides an elegant and natural way to answer this question: The difference is essentially the η invariant of the Dirac operator on the unit circle, shifted by the flux F.

II. THE ASSOCIATED BOUNDARY VALUE PROBLEM

Identifying $\mathbf{R}^2$ and $\mathbf{C}$ in the standard way, $(x,y)\leftrightarrow z = x+iy$, we can write

$$D=\frac{\partial}{\partial\bar z}+\frac{\partial\phi}{\partial\bar z}=e^{-\phi}\frac{\partial}{\partial\bar z}e^{\phi},$$

$$D^*=-\frac{\partial}{\partial z}+\frac{\partial\phi}{\partial z}=-e^{\phi}\frac{\partial}{\partial z}e^{-\phi}. \tag{2.1}$$

Now assume that for $r:=\sqrt{x^2+y^2}$ larger than some fixed constant $R>0$, $\phi = F\ln r$, where F is real constant. It is readily seen that in polar coordinates (r,θ) we have, for $r>R$,

$$D=e^{i\theta}\left(\frac{\partial}{\partial r}+\frac{i(\partial/\partial\theta)+F}{r}\right),$$

$$D^*=-e^{-i\theta}\left(\frac{\partial}{\partial r}-\frac{i(\partial/\partial\theta)+F}{r}\right). \tag{2.2}$$

Let $u\in L^2(\mathbf{R}^2)$ be such that $Du=0$. The restriction $u_{|>R}$ of u to the complement $\mathbf{C}B_R$ of $B_R := \{z\in\mathbf{C}\mid |z|\leqslant R\}$ belongs to $L^2((R,\infty)\times\mathbf{S}^1, r\,dr\,d\theta)$ and admits a Fourier series expansion

$$u_{|>R}\cong\sum_{k\in\mathbf{Z}}u_k(r)e^{ik\theta}.$$

Since $(Du)_{|>R}\cong\Sigma_{k\in\mathbf{Z}}\{u_k'+[(-k+F)/r]u_k\}\times e^{i(k+1)\theta}$ is 0, we get for any $k\in\mathbf{Z}$, $u_k'+[(-k+F)/r]u_k=0$, i.e., $u_k=c_kr^{k-F}$, c_k constant. But then $u_{|>R}\in L^2$ only if $c_k=0$ for $k\geqslant F-1$.

This suggests the introduction of the following boundary value problem. Denote by $(D, P_{<F-1})$ the operator whose domain is the set of functions $v\in C^\infty(B_R)$ such that for any $k\in\mathbf{Z}$, $k\geqslant F-1$,

$$\frac{1}{2\pi}\int_0^{2\pi}v(R,\theta)e^{-ik\theta}\,d\theta=0, \tag{2.3}$$

and on which D acts according to the first line in Eq. (2.1). Equation (2.3) amounts to a nonlocal boundary condition. Similarly, one defines $(D^*, P_{>F+1})$.

With these preparations we have the following proposition.

Proposition 2.4:
(a) $\ker(D)\cap L^2(\mathbf{R}^2)\simeq\ker(D, P_{<F-1})$,
(b) $\ker(D^*)\cap L^2(\mathbf{R}^2)\simeq\ker(D^*, P_{>F+1})$.

Proof: (a) From what was said previously the map $u\mapsto u_{|<R}$ from $\ker(D)\cap L^2(\mathbf{R}^2)$ into $\ker(D, P_{<F-1})$ is one to one. The inverse map is seen to be

$$\ker(D, P_{<F-1})\ni v\mapsto u\in\ker(D)\cap L^2(\mathbf{R}^2)$$

$$v\mapsto u=\begin{cases}v(r,\theta), & r\leqslant R,\\ \Sigma_{k<F-1}v_k(r/R)^{k-F}e^{ik\theta}, & r\geqslant R,\end{cases}$$

where v_k is the Fourier coefficient $(1/2\pi)\int_0^{2\pi}v(R,\theta)\times e^{-ik\theta}\,d\theta$. The proof of (b) is similar. □

III. THE RESULT

In order to bring the Atiyah–Patodi–Singer index theorem into the picture we need to take a different look at the operator (1.1) and the boundary condition (2.3).

Let $\mathcal{D}: C^\infty(\mathbf{R}^2,\mathbf{C}^2)\to C^\infty(\mathbf{R}^2,\mathbf{C}^2)$ be the Dirac operator on $\mathbf{R}^2$, i.e.,

$$\mathcal{D}=\begin{pmatrix}0&-1\\1&0\end{pmatrix}\frac{\partial}{\partial x}+\begin{pmatrix}0&i\\i&0\end{pmatrix}\frac{\partial}{\partial y}.$$

Twisting the spinor bundle $\mathbf{R}^2\times\mathbf{C}^2$ with the trivial bundle $\mathbf{R}^2\times\mathbf{C}$ equipped with the connection given by the one-form

$$\omega=i\frac{\partial\phi}{\partial y}\,dx-i\frac{\partial\phi}{\partial x}\,dy,$$

[a)] Present address: Department of Mathematics, University of North Texas, Denton, Texas 76203.

 0022-2488/90/092091-03$03.00

we get the corresponding Dirac operator,

$$\mathscr{D}_\phi = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \frac{\partial}{\partial x} + \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \frac{\partial}{\partial y} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{\partial \phi}{\partial x} + \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \frac{\partial \phi}{\partial y}, \tag{3.1}$$

or equivalently,

$$\mathscr{D}_\phi = \begin{pmatrix} 0 & D^* \\ D & 0 \end{pmatrix}.$$

Now restrict $\mathscr{D}_\phi$ to B_R. It is the index of the associated twisted $\frac{1}{2}$-Dirac operator that relates nicely to $L^2\,\mathrm{index}(D)$. In fact, the boundary conditions described by Eq. (2.3) are equivalent to the spectral boundary conditions introduced by Atiyah, Patodi, and Singer in Ref. 6. To see this notice that from Eq. (2.2) the tangential part of D is the self-adjoint elliptic operator $A = [(1/i)(\partial/\partial\theta) - F]/R$ defined on $C^\infty(\partial B_R)$ and with eigenvalues up to a scaling factor, $\{k - F\}_{k\in\mathbf{Z}}$, corresponding to the eigenfunctions $e^{ik\theta}$.

Let P_+ be the spectral projection on the positive eigenvalues of A and take on $C^\infty(\partial B_R)$ the boundary condition $P_+ v(R,\cdot) = 0$. Then (D, P_+) is exactly our previously defined $(D, P_{\leqslant F})$.

Theorem 3.2:

(a) $\mathrm{index}(D, P_+) = \dim\ker(D, P_{\leqslant F}) - \dim\ker(D^*, P_{>F+1})$

(b) $\mathrm{index}(D, P_+) = F + \frac{1}{2} - \frac{1}{2}(\eta_F(0) - h)$,

where $\eta_F(0)$ is the eta invariant associated to A, i.e., the value at 0 of the analytic continuation of the η function,

$$\eta_F(s) = \sum_{k\in\mathbf{Z}, k\neq F} \frac{\mathrm{sgn}(k-F)}{|k-F|^s}, \quad \Re(s) \geqslant 0$$

and $h = \dim\ker A$.

Proof: (a) Since $(D, P_+) = (D, P_{\leqslant F})$, (a) is a consequence of the fact that the adjoint $(D, P_+)^*$ of (D, P_+) is precisely $(D^*, P_{>F+1})$. This follows right away from the integration by parts formula

$$\int_{B_R} Du\,\bar{v} = \int_{B_R} u\,\overline{D^* v} + \int_{\partial B_R} e^{i\theta} u\bar{v}, \quad u, v \in C^\infty(B_R).$$

(b) We can also interpret (D, P_+) as a twisted version of the $\bar{\partial}$ operator on a compact surface with boundary. Then the Atiyah–Patodi–Singer index theorem for manifolds with boundary[6,7] gives

$$\mathrm{index}(D, P_+) = \frac{1}{2}\int_{B_R} c_1(T_{\mathbf{C}} B_R) + \int_{B_R} c_1(\mathbf{R}^2\times\mathbf{C}, \phi) - \frac{1}{2}(\eta_F(0) - h), \tag{3.3}$$

where $c_1(T_{\mathbf{C}} B_R)$ and $c_1(\mathbf{R}^2\times\mathbf{C}, \phi)$ are the first Chern classes of the complex tangent bundle to B_R, respectively, the twistor bundle $\mathbf{R}^2\times\mathbf{C}$. Now the Chern form of a surface coincides with the Euler form, and so by the Gauss–Bonnet theorem $\int_{B_R} c_1(T_{\mathbf{C}} B_R)$ is the Euler characteristic of B_R, namely, 1. Also, $c_1(\mathbf{R}^2\times\mathbf{C}) = (i/2\pi)d\omega = (1/2\pi)\Delta\phi \times dx\,dy$. Part (b) of the Theorem 3.2 follows. □

Note that the right-hand side of Eq. (3.3) should also contain the $\hat{A}$ genus of B_R and a secondary characteristic class of ∂B_R, which accounts for the fact that geometrically B_R is not a product near the boundary. These disappear since the $\hat{A}$ genus of any surface is 0.

Now we are ready to state our main result.

Theorem 3.4: If D is the operator in (1.1) then

$$L^2\,\mathrm{index}(D) = \frac{1}{2\pi}\int_{\mathbf{R}^2} \Delta\phi - \frac{\mathrm{sgn}(F)}{2} - \tfrac{1}{2}[\eta_F(0) + \mathrm{sgn}(F)h],$$

where $\eta_F(0)$ is the eta invariant associated to $A = (1/i)(\partial/\partial\theta) - F$ on $C^\infty(\mathbf{S}^1)$ and $h = \dim\ker A$.

Proof: Assume first that $F < 0$. From (2.4) and 3.2(a) we obtain

$$L^2\,\mathrm{index}(D) = \mathrm{index}(D, P_+) - [\dim\ker(D, P_{\leqslant F}) - \dim\ker(D, P_{<F-1})].$$

Since $\ker(D, P_{<F-1}) \subseteq \ker(D, P_{\leqslant F})$, the claim is proved if we show that $\ker(D, P_{\leqslant F}) = 0$. Now (2.1) states that an element in $\ker(D, P_{\leqslant F})$ is essentially a holomorphic function, thus its Fourier series at the boundary must contain only $e^{ik\theta}$ with $k \geqslant 0$. However, $P_{\leqslant F}$ prevents this if $F < 0$. We want to stress that $L^2\,\mathrm{index}(D) \neq \mathrm{index}(D, P_+)$, if $F \geqslant 0$.

The case $F > 0$ follows interchanging the roles of D and D^* in all we said so far, and $F = 0$ is trivial. □

A direct evaluation of the eta invariant $\eta_F(0)$ helps recover the Aharonov–Casher result.[1]

Proposition 3.5: If

$$\eta_F(s) = \sum_{k\in\mathbf{Z}, k\neq F} \frac{\mathrm{sgn}(k-F)}{|k-F|^s}, \quad \Re(s) \geqslant 0,$$

then

$$\eta_F(0) = \begin{cases} 2\epsilon - 1, & \text{if } F > 0,\ F = N + \epsilon,\ N \in \mathbf{N},\ 0 < \epsilon < 1, \\ 1 - 2\epsilon, & \text{if } F < 0,\ -F = N + \epsilon,\ N \in \mathbf{N},\ 0 < \epsilon < 1, \\ 0, & \text{if } F \in \mathbf{Z}. \end{cases}$$

Proof: See Ref. 8, or use the following regularization of the eta invariant:

$$\beta_F(t) = \sum_{k\in\mathbf{Z}} \mathrm{sgn}(k-F) e^{-t|k-F|},$$

$$\eta_F(0) = \lim_{t\searrow 0} \beta_F(t).$$ □

Corollary 3.6 (Aharonov–Casher[1]): If D is the operator in (1.1) then we have

$$L^2\,\mathrm{index}(D)=\begin{cases} N, & \text{if } F>0,\ F=N+\epsilon,\ N\in\mathbf{N},\ 0<\epsilon<1,\\ N-1, & \text{if } F>0,\ F=N,\ N\in\mathbf{N},\\ 0, & \text{if } F=0,\\ -N, & \text{if } F<0,\ -F=N+\epsilon,\ N\in\mathbf{N},\ 0<\epsilon<1,\\ -N+1, & \text{if } F<0,\ -F=N,\ N\in\mathbf{N}. \end{cases}$$

Proof: Immediate, using Theorem 3.4 and Proposition 3.5. □

IV. REMARK

A similar result can be proved for surfaces more general than $\mathbf{R}^2$, namely, those with Euclidean ends. A noncompact Riemann surface M is Euclidean at infinity if outside some compact set it is isometric to a disjoint union of finitely many CB_R's. Here the analog of D is $\bar\partial + \bar\partial\phi$. For instance, if M has just one Euclidean end and $F<0$ then

$$L^2\,\mathrm{index}(\bar\partial+\bar\partial\phi)=F+(1-2g)/2-\tfrac{1}{2}(\eta_F(0)-h),$$

where g is the genus of the compact Riemann surface obtained taking the one-point compactification of M.

ACKNOWLEDGMENTS

We gratefully acknowledge stimulating conversations with Professor R. Seeley.

This work was partially supported by the National Science Foundation, Grant No. DMS 8803072.

[1] Y. Aharonov and A. Casher, Phys. Rev. A **19**, 2461 (1979).
[2] H. Cycon, R. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators with Applications in Quantum Mechanics and Global Geometry*, in Texts and Monographs in Physics (Springer, New York, 1987).
[3] D. Bollé, F. Gesztesy, H. Grosse, W. Schweiger, and B. Simon, J. Math. Phys. **28**, 1512 (1987).
[4] M. F. Atiyah, R. Bott, and V. K. Patodi, Invent. Math. **19**, 279 (1973).
[5] F. Gesztesy and B. Simon, J. Funct. Anal. **79**, 91 (1988).
[6] M. F. Atiyah, V. K. Patodi, and I. M. Singer, Math. Proc. Cambridge Philos. Soc. **77**, 43 (1975).
[7] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, No. 6, 213 (1980).
[8] P. B. Gilkey, *Invariance Theory, the Heat Equation, and the Atiyah–Singer Index Theorem* (Publish or Perish, Wilmington, DE, 1984).

Hosotani breaking of E_6 to a subgroup of rank five

Brett McInnes[a)]
Center for Theoretical Physics, Laboratory for Nuclear Science and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 19 September 1989; accepted for publication 28 March 1990)

On multiply connected manifolds, it is possible to construct vacuum gauge configurations with nontrivial holonomy groups. This is the basis of the Hosotani mechanism. This naturally suggests a "Hosotani inverse problem": If we wish to break a gauge group G to a subgroup H, what are the possible finite holonomy groups having this effect, and what can one say about the fundamental groups of the underlying manifolds? Usually, this problem is too difficult to solve, but we show that, for $G = E_6$ and H locally isomorphic to the rank five group $SU(3)\times SU(2)\times U(1)\times U(1)$, a complete solution is possible. It is hoped that the results will aid a search for examples of Calabi–Yau manifolds leading to a low-energy gauge group of rank five.

I. INTRODUCTION

Recent developments in superstring theory,[1] and in the theory of the cosmological constant,[2] have underlined the importance of understanding physical phenomena in the context of multiply connected manifolds. In particular, gauge vacua on such manifolds can differ profoundly from those defined on simply connected spaces: The vacuum holonomy group can be nontrivial, and this can lead to gauge symmetry breaking.

This remarkable phenomenon—gauge fields breaking gauge symmetry, as it were—can be exploited as a symmetry breaking mechanism in circumstances where the Higgs mechanism cannot be employed. True, there is certainly a sense in which the "mechanism" terminology is misleading: It suggests that the phenomenon is entirely at our disposal, or at least can be relied upon not to appear when it is not wanted. In reality, however, one can show that gauge configurations of this type exist on *all* manifolds that are likely to be of physical interest, provided only that the manifold be multiply connected. In this sense, the "Hosotani[3] effect" is a generic, fundamental aspect of gauge theory, and it should always be taken into consideration when dealing with gauge fields on multiply connected manifolds.

Nevertheless, it is not unreasonable to hope that the Hosotani effect will make itself welcome in some way. The best-known application of this kind is in superstring phenomenology, where the method is used to break E_6 to the low-energy gauge group. The latter will be of rank six if the holonomy group is Abelian, and this case has received a great deal of attention. (Contrary to popular belief, it is *not* true in general that Abelian holonomy groups cannot reduce the rank; regrettably, however, this is true of E_6.) On the other hand, a non-Abelian holonomy group necessarily reduces the rank to five. Apart from the fact that this obviously brings us closer to the standard group, these non-Abelian compactifications have many other virtues.[4] Nevertheless, they have received less attention, mainly because very few explicit examples of Calabi–Yau manifolds with non-Abelian fundamental groups are known. It is not difficult to see why that should be the case: Finite non-Abelian groups are less familiar, more complex, and more numerous than their Abelian counterparts. For example, the authors of Ref. 4 begin with the $Y_{7:2,2,2,2}$ space,[5] a nonsingular intersection of four quadrics in CP^7. This space admits a fixed-point-free action by the non-Abelian group $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$, which can give rise to holonomy groups isomorphic either to $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$ itself or to the dihedral group D_8. The latter leads to a more satisfactory theory, although in the end it, too, succumbs to the comparison with phenomenology. The point, however, is this: Among all the various finite non-Abelian groups, just these two, $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$ and D_8, are of interest in this case. In order to unearth further examples of compactifications with non-Abelian fundamental groups, we need some way of focusing on the relevant groups.

We could, for example, ask: What groups can act freely on simply connected six-dimensional Calabi–Yau manifolds? Indeed, this is the question we would need to answer if we were to undertake a complete classification of all Calabi–Yau spaces. Unhappily, such a project is completely unrealistic. We proceed instead as follows. In more detail, the Hosotani effect breaks E_6 to the *centralizer* $C(\Phi)$ of the holonomy group, where $\Phi \subset E_6$ is the holonomy group and

$$C(\Phi) = \{g \in E_6 \text{ such that } g\phi = \phi g \text{ for all } \phi \in \Phi\}.$$

Thus if we know, for example, that the holonomy group is D_8, we find the low-energy gauge group by computing the centralizer of D_8 in E_6. But in reality, of course, our actual situation is just the reverse: We *know* the low-energy group $C(\Phi)$, and we would like to *find* Φ (and the underlying fundamental group). This suggests that we attack the following problem. Given a gauge group G and a subgroup H, together with a simply connected manifold $\tilde{M}$,

(i) find all subgroups $\Phi \subset G$ with $C(\Phi) = H$;

(ii) insist that Φ be the holonomy group of some vacuum gauge configuration constructed on $\tilde{M}/\pi$, and thus deduce a list of candidates for π, the fundamental group.

We call this, for obvious reasons, the "Hosotani inverse problem." Naturally, we must expect a high degree of nonuniqueness in our solution: We can only expect, for example, to obtain a "catalog" of possible holonomy groups and fun-

[a)] Permanent address: Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, 0511 Republic of Singapore.

 0022-2488/90/092094-11$03.00

damental groups leading to the breaking of E_6 to a group (locally) isomorphic to $SU(3)\times SU(2)\times U(1)\times U(1)$. This nonuniqueness must be reduced, in the manner of Ref. 4, by phenomenological considerations.

In fact, even this much less ambitious program cannot be carried out in most cases. Fortunately, however, the case of interest to us—$G = E_6$, $H =$ the above rank five subgroup, $\tilde{M} =$ Calabi–Yau—is one of the few exceptions. The purpose of this paper, then, is to solve the Hosotani inverse problem in this case. The most interesting groups—those corresponding to $\tilde{M}$ with reasonably low Euler characteristic—are presented explicitly, in a form suitable for checking the existence of a free action (or for computing the transformation behavior of various fields, in the case of the holonomy group). We begin by formulating the problem more precisely.

II. GENERAL FORMALISM

In physical language, a vacuum gauge configuration is a gauge field with $F^i{}_{\mu\nu} = 0$ everywhere. Mathematically, it corresponds to a principal fiber bundle with a flat connection, (P,M,G,ω), where M is the base manifold, G is the gauge group, and ω is a connection with a zero curvature form. Gauge transformations correspond to smooth maps $\mu: P\to P$ that map each fiber into itself and that satisfy $\mu\circ R_g = R_g\circ\mu$ for all $g\in G$, where R_g denotes the action of G on P. Now obviously the group of all gauge transformations (which is infinite dimensional) cannot be identified with G; and yet the two are evidently closely related. In the context of vacuum configurations, there is a particularly attractive way of making this relationship explicit. Let us define the *vacuum symmetry group* V to be the set of all gauge transformations that preserve the flat connection ω, in the sense that $\mu^*\omega = \omega$. Thus V is the group of internal symmetries of the whole structure, (P,M,G,ω). Now V can be characterized with the aid of the following theorem.

Theorem 1: Let (P,M,G) be any principal fiber bundle over a connected base manifold, and let ω be an arbitrary connection on P. Then the subgroup of gauge transformations satisfying $\mu^*\omega = \omega$ is isomorphic to $C(\Phi)$, the centralizer (in G) of the holonomy group of ω.

Proof: References 6 and 7.

Thus V, the vacuum symmetry group, is isomorphic to some subgroup of G. Now we obtain the relationship between G and the group of all gauge transformations.

Lemma 2: Let G be any Lie group and M any connected manifold. Then there exists a vacuum gauge configuration over M with symmetry group $V = G$.

Proof: Let P be the trivial bundle $P = M\times G$. The canonical flat connection on P is defined by taking the horizontal subspace at $(m,g)\in M\times G$ to be the tangent space to the submanifold $M\times\{g\}$. The holonomy group consists of a single element, the identity $e\in G$, and so the vacuum symmetry group $V = C\{e\} = G$.

Thus we can interpret G as the largest possible vacuum symmetry group: That is, it is the largest subgroup (of the group of all gauge transformations) that can preserve some vacuum configuration. However, while Lemma 2 guarantees the existence of a maximally symmetric vacuum, there may *also* exist other vacua with symmetry groups V that are proper subgroups of G. In such a case one has the familiar situation in which the vacuum is less symmetric than the Lagrangian, and one says that the maximal symmetry group G has been broken to its subgroup $V = C(\Phi)$. This is the mathematical basis of the Hosotani effect.

III. CALABI–YAU SPACES AND FINITE HOLONOMY GROUPS

Nontrivial vacuum holonomy groups arise as follows. By definition, the holonomy group Φ measures the nontriviality of gauge parallel transport around arbitrary closed loops in M. If we consider only those loops that can be contracted to a point, then the *restricted* holonomy group Φ_0 is obtained. Now Φ_0 is a normal subgroup of Φ, and it can be shown[8] that there is a homomorphism θ that maps the fundamental group $\pi(M)$ onto Φ/Φ_0. The Ambrose–Singer theorem[8] essentially states that the gauge curvature Ω generates Φ_0, so in the vacuum case we have a homomorphism from $\pi(M)$ onto Φ. Unlike $\pi(M)$, Φ is a subgroup of G, and so this situation is often described in the physics literature as "embedding" $\pi(M)$ (henceforth abbreviated to π) in G. This terminology is, however, both inaccurate and misleading. First, there is a profound distinction between π (which reflects purely topological properties of M, and is not related either to geometry or to gauge theory) and Φ (which is a geometric object, dependent on the particular connection under discussion). Second, the homomorphism $\theta:\pi\to\Phi$ need not be one-to-one, so π and Φ *may not even be of the same order*—hardly what one would wish to call an "embedding." Instead of speaking[4] of two "embeddings" of $\mathbf{Z}_4\rtimes\mathbf{Z}_4$ in E_6, for example, one should speak of a manifold, with $\pi = \mathbf{Z}_4\rtimes\mathbf{Z}_4$, on which one may wish to consider two different vacuum gauge fields, one with $\Phi = \mathbf{Z}_4\rtimes\mathbf{Z}_4$, the other with $\Phi = D_8$. In fact, the existence of a homomorphism $\theta:\pi\to\Phi$ means that π admits a normal subgroup N (which may be trivial) such that π/N is isomorphic to Φ. [Indeed, $\mathbf{Z}_4\rtimes\mathbf{Z}_4$ has a normal subgroup $\mathbf{Z}_2$ such that $(\mathbf{Z}_4\rtimes\mathbf{Z}_4)/\mathbf{Z}_2 = D_8$.] Henceforth, we shall always strictly distinguish Φ from π; to do otherwise is to invite confusion.

As π is always countable for a Riemannian manifold, the equation $\pi/N = \Phi$ means that vacuum holonomy groups are always discrete, though not necessarily finite. In general, then, the inverse problem may require us to deal with infinite discrete groups, which often have a subtle structure. Fortunately, that will not occur in our case: The theorem of Cheeger and Gromoll[9] implies that if M is a compact Riemannian manifold with non-negative Ricci tensor and nonzero Euler characteristic, then its universal cover is compact and so its fundamental group is finite. The manifolds in which we are interested are Ricci flat and (since the number of particle generations is proportional to the Euler characteristic) certainly have nonzero characteristic. Thus π and consequently Φ are both finite, not merely discrete. This is an important simplification. [This is, however, the only point at which we use the fact that M is Calabi–Yau; thus all subsequent results apply to any class of manifolds (such as those with positive Ricci tensor bounded away from

zero) with necessarily finite fundamental groups.] Henceforth, we shall consider only finite candidates for Φ and π.

IV. FINITE SUBGROUPS OF E_6: GENERAL RESULTS

Our first problem is to find all finite subgroups of E_6 having as their centralizer the desired low-energy group. There are at least two reasons to expect that this will be difficult. First, a complete classification of all finite subgroups is known for no more than a handful of Lie groups: It is most certainly not known for E_6. Second, it is quite possible for two subgroups of a group to be abstractly isomorphic and yet have entirely different centralizers. For example, take $SU(2)\times SU(2)$. The first factor has centralizer $\mathbf{Z}_2\times SU(2)$, while the diagonal subgroup clearly has centralizer $\mathbf{Z}_2\times\mathbf{Z}_2$—and yet this diagonal subgroup is also isomorphic to $SU(2)$. In the general case, these difficulties rule out an explicit solution of the inverse problem. The purpose of this section is to show that, in our particular case, these problems can be surmounted.

We begin by presenting the embedding of the standard group in E_6. We take this opportunity to emphasize, as strongly as possible, that Hosotani symmetry breaking operates *at the group level*: One must compute the centralizer of the holonomy *group* in the gauge *group*. It is customary in physics, and normally harmless, to deal with Lie groups purely through their algebras—so that, for example, one speaks of the $SO(10)$ "subgroup" of E_6, even though $SO(10)$ as a *group* cannot in fact be embedded in the group E_6. In the context of the Hosotani effect, however, such laxity can easily lead to swift disaster. For example, $SO(4)$ has $SU(2)$ as a subgroup, and it also obviously contains $SO(3)$; but the centralizer of $SU(2)$ in $SO(4)$ [which is $(SU(2)\times SU(2))/\mathbf{Z}_2$ in fact] is just $SU(2)$, while the centralizer of $SO(3)$ is clearly $\mathbf{Z}_2$, the diagonal matrices in $SO(4)$. Thus the centralizer can detect the distinction between $SU(2)$ and $SO(3)$, and so it is crucial to do all computations at the group level.

It is customary to embed the standard group in E_6 through the maximal subgroup $[SU(3)\times SU(3)\times SU(3)]/\mathbf{Z}_3$, but for our purposes it is much clearer to choose an essentially equivalent embedding through $[SU(2)\times SU(6)]/\mathbf{Z}_2$. The factoring by $\mathbf{Z}_2$ means that $-I_2$ in $SU(2)$ is to be identified with $-I_6$ in $SU(6)$. One can see that this is necessary by noting that the 27 of E_6 decomposes as $(2,\bar{6})+(1,15)$, whence it is clear that the pairs (I_2,I_6) and $(-I_2,-I_6)$ are indistinguishable in E_6. (The author is grateful to the referee for this observation.) (Geometrically, it can be explained in terms of the theory of symmetric spaces.[10]) Now consider the subgroup H_4 of $SU(6)$ consisting of matrices of the form

$$\begin{bmatrix} u_2 & & \\ & u_3 & \\ & & 1 \end{bmatrix},$$

where $u_3\in U(3)$, $u_2\in U(2)$, and $(\det u_2)(\det u_3)=1$. Because of this last condition, this group is locally isomorphic to $SU(3)\times SU(2)\times U(1)$, and is in fact the usual standard group. Now obviously it is essential, if we wish to break E_6 to H_4 by means of the Hosotani effect, that it be possible to express H_4 as the centralizer of some subgroup of E_6. But we have the following "well-known" fact (see Ref. 11 for a rigorous proof).

Theorem 3: There exists no subgroup of E_6 with H_4 as its centralizer.

Note that our inability to obtain the standard group follows purely from the group theory of E_6, not from any assumptions about string theory.

More satisfactory results can be obtained if we consider the rank five subgroup H_5, consisting of matrices

$$\begin{bmatrix} u_2 & & \\ & u_3 & \\ & & \gamma \end{bmatrix},$$

where u_2 and u_3 are as before, but $\gamma=(\det u_2)^{-1}(\det u_3)^{-1}$. This group is locally isomorphic to $SU(3)\times SU(2)\times U(1)\times U(1)$. It is possible to prove that H_5 is the centralizer, in E_6, of the group $H_3=[SU(2)\times(U(1)\times U(1))]/\mathbf{Z}_2$, where the $U(1)\times U(1)$ is embedded in $SU(6)$ as the diagonal matrix diag $(\alpha,\alpha,\beta,\beta,\beta,\alpha^{-2}\beta^{-3})$. (We say "it is possible to prove," because we want H_5 *precisely* and not some disconnected group with H_5 as its identity component.)

Our problem now is to find a complete list of finite holonomy groups that can break E_6 to H_5. The problem is simplified by the following very simple result. We shall say that a subgroup H (of a group G) is a *centralizer* in G if it can be expressed as the centralizer of some other subgroup of G. (Thus, for example, H_4 is *not* a centralizer in E_6.)

Lemma 4: Let G be any group, let H be a centralizer in G, and let F be any subgroup of G with $C(F)=H$. Then F must be a subgroup of CH.

Proof: If $CF=H$, then $CCF=CH$. But evidently F is a subgroup of CCF.

Now the centralizer of H_5 in E_6 is just H_3. [The relationship is reciprocal—$C(H_3)=H_5$, and $C(H_5)=H_3$.] Thus we obtain the conclusion that the finite holonomy groups for which we are searching are all subgroups of H_3. Regrettably, the converse is not true; some subgroups of H_3 have larger centralizers. But in this case we can surmount this problem, because H_3 is (barely) small enough for a complete survey of its finite subgroups to be possible. (In general, this is a key point in deciding whether an explicit solution of the Hosotani inverse problem is feasible: To find all finite holonomy groups breaking G to H, one needs at least some information on the finite subgroups of CH.)

We defined H_3 as $[SU(2)\times(U(1)\times U(1))]/\mathbf{Z}_2$, but there is a more concrete way of presenting it.

Lemma 5: H_3 is globally isomorphic to $U(1)\times U(2)$.

Proof: If $s\in SU(2)$, any element of H_3 has the form (s,α,β), where however we must bear in mind that (s,α,β) is identified with $(-s,-\alpha,-\beta)$. Now it is easy to verify that the map $(s,\alpha,\beta)\to(\alpha\beta,\alpha s)$ is a group homomorphism from H_3 onto $U(1)\times U(2)$. In fact it is an isomorphism, because the only triples mapped onto the identity $(1,I_2)$ in $U(1)\times U(2)$ are $(I_2,1,1)$ and $(-I_2,-1,-1)$, and both of these correspond to the identity in H_3. (Note that this proof would break down if the subgroup of E_6 had been $[SU(2)\times SU(6)]$ instead of $[SU(2)\times SU(6)]/\mathbf{Z}_2$.)

Now it is natural to split our study of $U(1)\times U(2)$ into two parts: $SU(2)$ and the remainder.

The finite subgroups of $SO(3)$ can all be found by geometric means.[10] They are isomorphic to (a) the cyclic groups $\mathbf{Z}_n$, (b) the dihedral groups, symmetry groups of the regular polygons, denoted D_{2n} for $n\geqslant 3$, (c) the symmetry groups T_{12}, O_{24}, and I_{60} of the regular tetrahedron, octahedron, and icosahedron, respectively, and finally (d) the noncyclic group $\mathbf{Z}_2\times\mathbf{Z}_2$, which we denote D_4 for formal reasons. All of these are non-Abelian except $\mathbf{Z}_n$ and D_4, and their orders are indicated by the subscripts, a convention to which we adhere henceforth. Now there is of course a well-known two-to-one homomorphism ϕ:$SU(2)\to SO(3)$, and so we define $Q_8=\phi^{-1}D_4$, $Q_{4n}=\phi^{-1}D_{2n}$, $\tilde{T}_{24}=\phi^{-1}T_{12}$, $\tilde{O}_{48}=\phi^{-1}O_{24}$, $\tilde{I}_{120}=\phi^{-1}I_{60}$. The Q_{4n}, $n\geqslant 2$, are non-Abelian groups called the quaternionic groups, while the remainder are called the binary polyhedral groups. It can be shown[10] that every finite subgroup of $SU(2)$ is isomorphic either to a cyclic group or to one of these groups. However, this alone does *not* solve our problem, for we have seen that two different subgroups can have different centralizers even though they be abstractly isomorphic. In short, a classification up to isomorphism is not good enough for our purposes. Again, however, this (serious) problem is tractable in our particular case, as we shall now explain.

Let A and B be two abstractly isomorphic subgroups of a group G. We shall say that A and B are isomorphic by conjugacy if there exists $g\in G$ with $B=g^{-1}Ag$. This is a more restrictive notion than mere isomorphism; for example, it is easy to show that the diagonal $SU(2)$ in $SU(2)\times SU(2)$ is not isomorphic by conjugacy to either of the explicit $SU(2)$ factors. We now have the following result.

Theorem 6: Let G be a group with the property that any two subgroups are isomorphic if and only if they are isomorphic by conjugacy. Let G be embedded as a subgroup of J and let A be a proper subgroup of G with the property $C_JA=C_JG$. Then if B is any subgroup of G isomorphic to A, we have $C_JA=C_JB=C_JG$.

Proof: By hypothesis, there exists $g\in G$ with $B=g^{-1}Ag$. Let $c\in C_JB$. Then $cg^{-1}ag=g^{-1}agc$ for all $a\in A$, and so $gcg^{-1}a=agcg^{-1}$. Hence, $gcg^{-1}\in C_JA$, that is, $c\in g^{-1}(C_JA)g$. That is, $C_JB\subseteq g^{-1}(C_JA)g=g^{-1}(C_JG)g=C_JG$ since $g\in G$. So we have $C_JB\subseteq C_JG$. But the fact that $B\subset G$ clearly implies $C_JG\subseteq C_JB$. Hence, $C_JB=C_JG$. □

Now $SU(2)$ has precisely this property: It can be shown[10] that if two finite subgroups of $SU(2)$ are isomorphic, then they are isomorphic by conjugacy. Hence, these somewhat abstract considerations allow us to deal with our problem in a very concrete way. Take, for example, the group Q_8. Abstractly it may be defined as a group with two generators x, y, satisfying the relations $x^2=y^2$, $y^{-1}xy=x^{-1}$. In $SU(2)$ there is an uncountable infinity of distinct finite subgroups isomorphic to Q_8. In particular, we may take $x=\begin{pmatrix} i & 0\\ 0 & -i\end{pmatrix}$ and $y=\begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}$. Since every other Q_8 in $SU(2)$ is isomorphic by conjugacy to this one, Theorem 6 says that if this representative has the same centralizer in E_6 as $SU(2)$, so do all the others. Hence our problem, as far as the $SU(2)$ part is concerned, is reduced to an easy matrix computation. The quaternionic groups Q_{4n} are given abstractly by $x^n=y^2$, $y^{-1}xy=x^{-1}$, and we can take $x=\begin{pmatrix}\delta & 0\\ 0 & \delta^{-1}\end{pmatrix}$, $y=\begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}$, where $\delta=\exp(i\pi/n)$. A straightforward computation shows that all of these groups have the same centralizer in E_6 as $SU(2)$ [namely, $SU(6)$]. Similarly, the cyclic groups can be represented explicitly by diagonal $SU(2)$ matrices, whereupon it is obvious that they do *not* have the same centralizer as $SU(2)$. Finally, we deal with the binary polyhedral groups as follows. In $SO(3)$ the following inclusions are well known[12] (and have a clear geometric meaning):

$$D_4\subset T_{12}\subset O_{24},$$

$$D_4\subset T_{12}\subset I_{60}.$$

Hence, lifting to $SU(2)$, we have

$$Q_8\subset\tilde{T}_{24}\subset SU(2),$$

$$Q_8\subset\tilde{O}_{48}\subset SU(2),$$

$$Q_8\subset\tilde{I}_{120}\subset SU(2).$$

In general, if $A\subset B\subset G\subset J$, and if we know that $C_JA=C_JG$, then $C_JG\subseteq C_JB\subseteq C_JA$, whence $C_JB=C_JG$. We know that Q_8 has the same centralizer in E_6 as $SU(2)$, and so we conclude that the binary polyhedral groups likewise have the same centralizer in E_6 as $SU(2)$. We have proved the following theorem.

Theorem 7: Let F be any finite subgroup of $SU(2)\subset E_6$. Then F has the same centralizer in E_6 as $SU(2)$ if and only if F is not Abelian.

We now understand how to replace $SU(2)$ by something finite. Now let us turn to the $U(1)\times U(1)$ factor of $U(1)\times U(2)$. It will of course have to be replaced by cyclic groups and products of cyclic groups. $U(1)\times U(1)$ is embedded in $SU(6)$ as diag $(\alpha,\alpha,\beta,\beta,\beta,\alpha^{-2}\beta^{-3})$. Replace α by w, the generator of a $\mathbf{Z}_n$ subgroup of the second $U(1)$, so that $\mathbf{Z}_n$ corresponds to diag $(w,w,w^{-1},w^{-1},w^{-1},w)$. (Recall that the isomorphism of Lemma 5 identified the first $U(1)$ with $\alpha\beta$ and the second with α, so if $\alpha\beta=1$ and $\alpha=w$, then $\beta=w^{-1}$ and $\alpha^{-2}\beta^{-3}=w$.) Similarly, if z generates the $\mathbf{Z}_m$ replacing the first $U(1)$, then $\mathbf{Z}_m$ is diag $(1,1,z,z,z,z^{-3})$. Clearly, not all values of m and n are permissible, in the sense that, for some values of m and n, $\mathbf{Z}_m\times\mathbf{Z}_n$ will not have the same centralizer in $SU(6)$ as $U(1)\times U(1)$. For example, if $z^{-3}=1$, then both diag $(w,w,w^{-1},w^{-1},w^{-1},w)$ and diag $(1,1,z,z,z,1)$ will commute with a subgroup of $SU(6)$ that is locally isomorphic to $SU(3)\times SU(3)\times U(1)$. Thus $z^{-3}=1$ cannot be permitted; that is, $m\neq 1,3$. Similarly, if $n=1$ or 2 (so that $w=\pm 1$), then we cannot allow $z=z^{-3}$, so we must have $m\neq 2$ or 4 in that case. Finally, we shall later find it necessary to ask whether the product diag $(w,w,w^{-1},w^{-1},w^{-1},w)\times$diag$(1,1,z,z,z,z^{-3})$ has the correct centralizer. The answer is yes unless $n=2m$ (because then $w=zw^{-1}$) or $m=2n$ (because then $zw^{-1}=wz^{-3}$). Thus, identifying z and w with their images in $SU(6)$, we have the following:

$$m\neq 1,3. \tag{1}$$

$$\text{If } n=1 \text{ or } 2, \quad \text{then } m\neq 2,4. \tag{2}$$

The product zw has the correct centralizer only if

$$n\neq 2m, \quad m\neq 2n. \tag{3}$$

Finally, let us bring together our findings on SU(2) and U(1)×U(1). In doing so, it is crucial to remember that U(2) is *not* isomorphic to U(1)×SU(2), but rather to [U(1)×SU(2)]/$\mathbf{Z}_2$. This is because $-I_2$ belongs both to U(1) and to SU(2), and must not be "counted" twice. Now $\mathbf{Z}_n$ contains -1 only when n is even, and so the finite subgroups of U(2) are given in two families: $\mathbf{Z}_n \times F_p$ when n is odd [where F_p is any finite non-Abelian subgroup of SU(2)] and $[\mathbf{Z}_n \times F_p]/\mathbf{Z}_2$, which we denote $\mathbf{Z}_n \circ F_p$, when n is even. All of the F_p have $\mathbf{Z}_2$ as their center, and $\mathbf{Z}_n \circ F_p$ is called the central product of $\mathbf{Z}_n$ and F_p. In the case of the quaternionic groups, these central products are given by three generators, w, x, and y, where w commutes with x and y and, for $\mathbf{Z}_n \circ Q_p$,

$$w^{(n/2)} = x^{(p/4)} = y^2, \quad y^{-1}xy = x^{-1}. \tag{4}$$

Note that the order of $\mathbf{Z}_n \circ F_p$ is $\frac{1}{2}np$, not np. Allowing for all this and for (1) and (2) above, we have the following result.

Theorem 8: Every finite subgroup of E_6 having H_5 as its centralizer is a subgroup of a group in the following list:

$$\mathbf{Z}_m \times \mathbf{Z}_n \times F_p, \quad n \text{ odd},$$
$$\mathbf{Z}_m \times (\mathbf{Z}_n \circ F_p), \quad n \text{ even},$$

where F_p is any quaternionic or binary polyhedral group, and where (a) $m \neq 1$ or 3, (b) $m \neq 2$ or 4 if $n = 1$ or 2.

In principle, this solves our problem: We "merely" have to list the non-Abelian subgroups of these groups. In practice, there are two problems here. The first is that some of these groups, and also their subgroups, are presented in an unfamiliar form: The reader who consults a standard catalog[13] of finite groups will not, for example, find $\mathbf{Z}_4 \circ Q_8$ listed under that title. Consequently, we will not know, for example, whether our list has any redundancies. (Two finite groups can appear to be different and yet be isomorphic.) The second problem is that our solution is insufficiently explicit. As is well known, one of the great virtues of Calabi–Yau compactifications is the fact that they give an understanding of the generation problem: The number of generations is $|\chi|/2|\pi|$, where χ is the Euler characteristic of the universal cover, and $|\pi|$ is the order of the fundamental group. Since the holonomy group Φ is given by $\Phi = \pi/N$, it follows that $|\pi|$ is some multiple of $|\Phi|$. Hence, we wish to answer questions such as: "Are there any holonomy groups of order 6 breaking E_6 to H_5?" (answer: no), or "which groups of order 36 occur in the list?" (answer: four of them).

Now in fact it is extremely difficult to solve our problem as explicitly as this. But note that, if we arbitrarily assume the existence of four generations, then $|\chi| = 8|\pi| = 8r|\Phi|$ for some integer r. Hence, the finite holonomy groups of very large order are relevant only to simply connected Calabi–Yau spaces with enormous Euler characteristics; for example, holonomy groups of order 64 are of physical interest only for manifolds with Euler characteristics 512, 1024, and so on. We shall therefore present completely explicit solutions only for the holonomy groups of lowest order. The computations cease to be feasible, at least by hand, at $|\Phi| = 64$; hence, we give explicit solutions for every Φ with $|\Phi| < 64$.

V. CLASSIFICATION OF ALL FINITE HOLONOMY GROUPS OF LOW ORDER BREAKING E_6 TO H_5

As mentioned above, Theorem 8 gives us many groups in unfamiliar forms. We deal with this by settling on a standard list. Reference 14 gives a complete list of all groups of order $\leqslant 32$, in a way that is consistent with Ref. 13. For every group we encounter, we shall establish an isomorphism with one of the groups presented in Ref. 14. For groups of higher order, we give in the Appendix a list of standard forms in a style and notation consistent with those of Ref. 14.

Aside from establishing these isomorphisms (which can be rather difficult, since it cannot be done in a systematic fashion), the procedure is fairly straightforward, if tedious. Let us illustrate with some of the simpler cases. The smallest admissible values of m and n are $m = 2$, $n = 3$, and the smallest F_p is Q_8, so we have $\mathbf{Z}_2 \times \mathbf{Z}_3 \times Q_8$. We must find the subgroups with the correct centralizer. First, of course, is $\mathbf{Z}_2 \times \mathbf{Z}_3 \times Q_8$ itself, a group of order 48 generated by z, w, x, y, where $z^2 = w^3 = 1, x^2 = y^2, y^{-1}xy$. Now it is important to bear in mind that if A and B are finite groups, then subgroups of $A \times B$ *need not* have the form (subgroup of A)×(subgroup of B). That is true,[15] however, if the orders of A and B are relatively prime. So we only need to examine the subgroups of $\mathbf{Z}_2 \times Q_8$. For example, the subgroup generated by x and zy—denote it by $\langle x,zy\rangle$—clearly has the same centralizer in E_6 as $\mathbf{Z}_2 \times Q_8$. Since $x^2 = (zy)^2$ and $(zy)^{-1}xzy = x^{-1}$, this subgroup is clearly isomorphic to Q_8, and so we have obtained a second candidate, the order 24 group $\mathbf{Z}_3 \times Q_8$. We must also consider $\langle zx,y\rangle$; again, however, $(zx)^2 = y^2$, $y^{-1}zxy = zx^{-1} = (zx)^{-1}$ since $z^2 = 1$, and so we obtain nothing new because this subgroup is also isomorphic to Q_8. Again, $\langle zx,zy\rangle$ is isomorphic to Q_8.

Proceeding in this way, one obtains Table I. The groups are arranged in increasing order, from 8 to 60. The first column gives the minimum possible value for $|\chi|$, assuming the existence of either three or four generations. In the second column the conventional name, if any, of the group is given, followed by its index number according either to Ref. 14 or to the Appendix. In most cases it is not easy to see how the group in question arises, so we specify it as a subgroup of one of the groups in Theorem 8, giving the values of m, n, and p. (This is ambiguous for the binary polyhedral groups, but in fact they appear in only one single case, indicated by the asterisk in the table.) In every other case, (m,n,p) refers to $\mathbf{Z}_m \times \mathbf{Z}_n \times Q_p$ for n odd, or to $\mathbf{Z}_m \times \mathbf{Z}_n \circ Q_p$ for n even. The precise embedding is given by the appropriate combination of canonical generators: z generates $\mathbf{Z}_m$, w generates $\mathbf{Z}_n$, and x and y generate Q_p [see Eqs. (4)]. For example, we found above that $\mathbf{Z}_3 \times Q_8$, of order 24, occurs as a subgroup of $\mathbf{Z}_2 \times \mathbf{Z}_3 \times Q_8$; hence, $(m,n,p) = (2,3,8)$. In terms of the canonical generators, it was $\langle w,x,zy\rangle$, and, hence, the entry in the table.

A full proof that Table I is complete, for groups of order < 64, is extremely long and technical. We shall instead concentrate on a few of the more interesting and subtle cases.

A. The dihedral group D_8

Begin with the group $\mathbf{Z}_2 \times (\mathbf{Z}_4 \circ Q_8)$, generated by z,w,x,y with $z^2 = 1$ and $w^2 = x^2 = y^2$, $y^{-1}xy = x^{-1}$. It is

TABLE I. All finite holonomy groups of order less than 64, breaking E_6 to a subgroup of rank five.

$\lvert\chi\rvert_{\min}$		Φ		Subgroup of $\mathbf{Z}_m \times \mathbf{Z}_n$ ($\times$ or $\circ$) Q_p			Canonical generators
$g=3$	$g=4$	Conventional	TW	m	n	p	z,w,x,y
48	64	D_8	8/4	2	4	8	$\langle wx,zy\rangle$
72	96	D_{12}	12/3	2	4	12	$\langle zx^2,wy\rangle$
		Q_{12}	12/5	4	4	12	$\langle x^2,zwy\rangle$
96	128	$(\mathbf{Z}_2\times\mathbf{Z}_4)\rtimes\mathbf{Z}_2$	16/8	2	4	8	$\langle w,x,zy\rangle$
		$\mathbf{Z}_2\times D_8$	16/6	2	4	8	$\langle z,wx,y\rangle$
		$\mathbf{Z}_8\rtimes\mathbf{Z}_2$	16/11	2	8	8	$\langle zx,wy\rangle$
		$\mathbf{Z}_8\rtimes\mathbf{Z}_2$	16/13	2	4	16	$\langle wx,zy\rangle$
		D_{16}	16/12	2	4	16	$\langle zx,wy\rangle$
		$(\mathbf{Z}_2\times\mathbf{Z}_4)\rtimes\mathbf{Z}_2$	16/9	4	4	8	$\langle zx,wy\rangle$
		$\mathbf{Z}_4\rtimes\mathbf{Z}_4$	16/10	4	4	8	$\langle x,zwy\rangle$
108	144	$\mathbf{Z}_3\times D_6$	18/3	2	12	12	$\langle x^2,zwy\rangle$
120	160	D_{20}	20/3	2	4	20	$\langle zx^2,wy\rangle$
		Q_{20}	20/4	4	4	20	$\langle x^2,zwy\rangle$
144	192	$\mathbf{Z}_3\times Q_8$	24/8	2	3	8	$\langle w,x,zy\rangle$
		$\mathbf{Z}_3\times D_8$	24/7	2	12	8	$\langle w^4,w^3x,zy\rangle$
		$\mathbf{Z}_4\times D_6$	24/9	2	4	12	$\langle w,x^2,zwy\rangle$
		$\mathbf{Z}_2\times D_{12}$	24/4	2	4	12	$\langle z,x,wy\rangle$
		$\mathbf{Z}_3\rtimes\mathbf{Z}_8$	24/14	2	8	12	$\langle x^2,zwy\rangle$
		D_{24}	24/10	2	4	24	$\langle zx,wy\rangle$
		$\mathbf{Z}_3\rtimes D_8$	24/15	2	4	24	$\langle wx,zy\rangle$
		$\mathbf{Z}_2\times Q_{12}$	24/6	4	4	12	$\langle x,zwy\rangle$
168	224	D_{28}	28/3	2	4	28	$\langle zx^2,wy\rangle$
		Q_{28}	28/4	4	4	28	$\langle x^2,zwy\rangle$
180	240	$\mathbf{Z}_5\times D_6$	30/3	5	4	12	$\langle zx^2,wy\rangle$
		$\mathbf{Z}_3\times D_{10}$	30/2	6	4	20	$\langle z^2,x^2,z^3wy\rangle$
192	256	$\mathbf{Z}_2\times[16/8]$	32/10	2	4	8	$\langle z,w,x,y\rangle$
		$\mathbf{Z}_2\times[16/11]$	32/13	2	8	8	$\langle z,x,wy\rangle$
		$(\mathbf{Z}_2\times\mathbf{Z}_8)\rtimes\mathbf{Z}_2$	32/17	2	8	8	$\langle zw,x,y\rangle$
		$\mathbf{Z}_{16}\rtimes\mathbf{Z}_2$	32/22	2	16	8	$\langle xz,wy\rangle$
		$(\mathbf{Z}_2\times\mathbf{Z}_8)\rtimes\mathbf{Z}_2$	32/26	2	4	16	$\langle w,zx,y\rangle$
		...	32/32	2	8	16	$\langle zx,wy\rangle$
		$(\mathbf{Z}_4\times\mathbf{Z}_4)\rtimes\mathbf{Z}_2$	32/31	2	8	16	$\langle wx,zy\rangle$
		D_{32}	32/49	2	4	32	$\langle zx,wy\rangle$
		$\mathbf{Z}_{16}\rtimes\mathbf{Z}_2$	32/50	2	4	32	$\langle wx,zy\rangle$
		$\mathbf{Z}_4\rtimes\mathbf{Z}_8$	32/15	4	4	8	$\langle zw,x,y\rangle$
		$\mathbf{Z}_4\times D_8$	32/14	4	4	8	$\langle zw,zx,y\rangle$
		$\mathbf{Z}_8\rtimes\mathbf{Z}_4$	32/19	4	8	8	$\langle zx,wy\rangle$
		$(\mathbf{Z}_2\times\mathbf{Z}_8)\rtimes\mathbf{Z}_2$	32/27	4	4	16	$\langle zx,wy\rangle$
		$\mathbf{Z}_8\rtimes\mathbf{Z}_4$	32/29	4	4	16	$\langle x,zwy\rangle$
		...	32/28	4	4	16	$\langle zwx,y\rangle$
		$\mathbf{Z}_8\rtimes\mathbf{Z}_4$	32/30	4	4	16	$\langle wx,zy\rangle$
		$\mathbf{Z}_4\rtimes\mathbf{Z}_8$	32/21	8	1	8	$\langle x,zy\rangle$
216	288	$\mathbf{Z}_3\times Q_{12}$	36/6	2	3	12	$\langle w,zx,y\rangle$
		$\mathbf{Z}_6\times D_6$	36/5	2	12	12	$\langle zx,wy\rangle$
		D_{36}	36/8	2	4	36	$\langle zx^2,wy\rangle$
		Q_{36}	36/12	4	4	36	$\langle x^2,zwy\rangle$
240	320	$\mathbf{Z}_5\times Q_8$	40/5	2	5	8	$\langle w,x,zy\rangle$
		$\mathbf{Z}_4\times D_{10}$	40/6	2	4	20	$\langle w,x,zy\rangle$
		$\mathbf{Z}_2\times D_{20}$	40/7	2	4	20	$\langle z,x,wy\rangle$
		$\mathbf{Z}_5\rtimes\mathbf{Z}_8$	40/12	2	8	20	$\langle x^2,zwy\rangle$
		$\mathbf{Z}_5\rtimes D_8$	40/11	2	4	40	$\langle wx,zy\rangle$
		D_{40}	40/10	2	4	40	$\langle zx,wy\rangle$
		$\mathbf{Z}_2\times Q_{20}$	40/8	4	4	20	$\langle x,zwy\rangle$
		$\mathbf{Z}_5\times D_8$	40/4	5	4	8	$\langle z,wx,y\rangle$
252	336	$\mathbf{Z}_7\times D_6$	42/2	7	4	12	$\langle z,x^2,wy\rangle$
		$\mathbf{Z}_3\times D_{14}$	42/3	6	4	28	$\langle z^2,x^2,z^3wy\rangle$
264	352	D_{44}	44/3	2	4	44	$\langle zx^2,wy\rangle$
		Q_{44}	44/4	4	4	44	$\langle x^2,zwy\rangle$
288	384	$\mathbf{Z}_6\times Q_8$	48/9	2	3	8	$\langle z,w,x,y\rangle$
		$\mathbf{Z}_3\times[16/8]$	48/11	2	12	8	$\langle w,x,zy\rangle$

TABLE I. (*Continued.*)

$\lvert\chi\rvert_{\min}$		Φ		Subgroup of $\mathbf{Z}_m \times \mathbf{Z}_n$ ($\times$ or $\circ$) Q_p			Canonical generators
		$\mathbf{Z}_6 \times D_8$	48/8	2	12	8	$\langle z,w^4,x,w^3y\rangle$
		$\mathbf{Z}_4 \times D_{12}$	48/7	2	4	12	$\langle z,w,x,y\rangle$
		$\mathbf{Z}_8 \times D_6$	48/6	2	8	12	$\langle zw,x,y\rangle$
		$\mathbf{Z}_2 \times [24/14]$	48/19	2	8	12	$\langle z,x^2,wy\rangle$
		$\mathbf{Z}_3 \rtimes \mathbf{Z}_{16}$	48/26	2	16	12	$\langle x^2,zwy\rangle$
		$\mathbf{Z}_3 \times Q_{16}$	48/17	2	3	16	$\langle w,x,zy\rangle$
		$D_{24} \rtimes \mathbf{Z}_2$	48/24	2	4	24	$\langle w,x,zy\rangle$
		$\mathbf{Z}_2 \times D_{24}$	48/18	2	4	24	$\langle z,x,wy\rangle$
		$\mathbf{Z}_2 \times [24/15]$	48/20	2	4	24	$\langle z,wx,y\rangle$
		$\mathbf{Z}_3 \rtimes [16/10]$	48/28	2	8	24	$\langle wx,zy\rangle$
		$\mathbf{Z}_3 \rtimes [16/13]$	48/29	2	8	24	$\langle x,zwy\rangle$
		$\mathbf{Z}_2 \times \tilde{T}_{24}$	48/21	*	*	*	$\langle z,x,y,wv\rangle$
		D_{48}	48/22	2	4	48	$\langle zx,wy\rangle$
		$\mathbf{Z}_{24} \rtimes \mathbf{Z}_2$	48/23	2	4	48	$\langle wx,zy\rangle$
		$\mathbf{Z}_3 \times [16/10]$	48/13	4	3	8	$\langle w,x,zy\rangle$
		$\mathbf{Z}_3 \times [16/9]$	48/12	4	12	8	$\langle w^4,zx,w^3y\rangle$
		$\mathbf{Z}_4 \times Q_{12}$	48/10	4	4	12	$\langle zw,x,y\rangle$
		$\mathbf{Z}_{12} \rtimes \mathbf{Z}_4$	48/25	4	4	24	$\langle x,zwy\rangle$
		$\mathbf{Z}_3 \rtimes [16/9]$	48/27	4	4	24	$\langle wx,zy\rangle$
		$\mathbf{Z}_3 \times [16/11]$	48/14	6	8	8	$\langle z^2,z^3x,wy\rangle$
		$\mathbf{Z}_3 \times [16/13]$	48/16	6	4	16	$\langle z^2,wx,z^3y\rangle$
		$\mathbf{Z}_3 \times D_{16}$	48/15	6	4	16	$\langle z^2,z^3,x,wy\rangle$
300	400	$\mathbf{Z}_5 \times D_{10}$	50/3	2	20	20	$\langle x,zwy\rangle$
312	416	D_{52}	52/3	2	4	52	$\langle zx^2,wy\rangle$
		Q_{52}	52/4	4	4	52	$\langle x^2,zwy\rangle$
324	432	$\mathbf{Z}_9 \times D_6$	54/4	9	4	12	$\langle z,x^2,wy\rangle$
		$\mathbf{Z}_3 \times \mathbf{Z}_3 \times D_6$	54/5	6	12	12	$\langle z^2,x^2,z^3wy\rangle$
		$\mathbf{Z}_3 \times D_{18}$	54/6	6	4	36	$\langle z^2,x^2,z^3wy\rangle$
336	448	$\mathbf{Z}_7 \times Q_8$	56/5	2	7	8	$\langle w,x,zy\rangle$
		$\mathbf{Z}_4 \times D_{14}$	56/6	2	4	28	$\langle w,x,zy\rangle$
		$\mathbf{Z}_2 \times \mathbf{Z}_2 \times D_{14}$	56/7	2	4	28	$\langle z,x,wy\rangle$
		$\mathbf{Z}_7 \rtimes \mathbf{Z}_8$	56/10	2	8	28	$\langle x^2,zwy\rangle$
		D_{56}	56/9	2	4	56	$\langle zx,wy\rangle$
		$\mathbf{Z}_7 \rtimes D_8$	56/11	2	4	56	$\langle wx,zy\rangle$
		$\mathbf{Z}_2 \times Q_{28}$	56/8	4	4	28	$\langle x,zwy\rangle$
		$\mathbf{Z}_7 \times D_8$	56/4	7	4	8	$\langle z,wx,y\rangle$
360	480	$\mathbf{Z}_5 \times Q_{12}$	60/5	2	5	12	$\langle w,x,zy\rangle$
		$\mathbf{Z}_3 \times Q_{20}$	60/6	2	3	20	$\langle w,x,zy\rangle$
		D_{60}	60/7	2	4	60	$\langle zx^2,wy\rangle$
		$\mathbf{Z}_5 \times D_{12}$	60/4	5	4	12	$\langle z,x^2,wy\rangle$
		$\mathbf{Z}_6 \times D_{10}$	60/3	6	4	20	$\langle z,x^2,wy\rangle$
		Q_{60}	60/8	4	4	60	$\langle x^2,zwy\rangle$

easy to see that the subgroup $\langle w,z,zy\rangle$, generated by w, x, and zy, is isomorphic to the group $\mathbf{Z}_4 \circ Q_8$ of order 16, that is, to $\langle w,x,y\rangle$ itself. But what is that group? Clearly, it is also generated by w, wx, and wy, where $(wx)^2 = w^2x^2 = x^4 = 1$. (This last is true because $y^{-1}xy = x^{-1}$ implies $y^{-1}x^2y = x^{-2}$, whence $x^2 = x^{-2}$.) Similarly, $(wy)^2 = 1$ and $w^4 = x^4 = 1$, while $(wy)^{-1}wxwy = w^2(wx)^{-1}$, so our group has the form $\langle a,b,c\rangle$ with $a^2 = b^2 = c^4 = 1$, $b^{-1}ab = c^2a^{-1}$. This is the semidirect product $(\mathbf{Z}_2 \times \mathbf{Z}_4) \rtimes \mathbf{Z}_2$, type 16/8 in the notation of Ref. 14. Again, consider $\langle z,wx,y\rangle$. A short calculation shows that $x^{-1}yx = y^{-1}$, and so we have $(wx)^{-1}y\,wx = y^{-1}$, where $(wx)^2 = y^4 = 1$. Thus wx and y generate the dihedral group D_8, and so $\langle z,wx,y\rangle = \mathbf{Z}_2 \times D_8$. Similarly, $\langle wx,zy\rangle = D_8$. This is the smallest finite holonomy group capable of breaking E_6 to H_5. We leave it to the reader to show that all other admissible subgroups of $\mathbf{Z}_2 \times (\mathbf{Z}_4 \circ Q_8)$ are isomorphic to one of these three. *Note* that $\langle zw,x,y\rangle$, $\langle x,zwy\rangle$, and so on are *not* admissible, because in this case $n = 2m$, and we saw in the previous section that zw does not have the correct centralizer when $n = 2m$ or $m = 2n$. It is important to eliminate these special cases.

B. The semidirect product $\mathbf{Z}_8 \rtimes \mathbf{Z}_2$, type 16/13

Take $\mathbf{Z}_2 \times (\mathbf{Z}_4 \circ Q_{16})$, with $z^2 = 1$, $w^2 = x^4 = y^2$, $y^{-1}xy = x^{-1}$. The group $\mathbf{Z}_4 \circ Q_{16}$ is also generated by x,

wx^2, wy, where $x^8 = 1 = (wx^2)^2 = (wy)^2$, and we have $(wy)^{-1}wx^2wy = wx^{-2} = w^2(wx^2)^{-1} = x^4(wx^2)^{-1}$, so $\mathbf{Z}_4 \circ Q_{16}$ is isomorphic to $(\mathbf{Z}_2 \times \mathbf{Z}_8) \rtimes \mathbf{Z}_2$, type 32/26. The subgroup of $\mathbf{Z}_2 \times (\mathbf{Z}_4 \circ Q_{16})$ generated by wx, zy, is also generated by $zwxy$, wx. Noting that $(xy)^2 = xyxy = xy^2y^{-1}xy = xy^2x^{-1} = y^2$, we have $(zwxy)^2 = 1$, $(wx)^8 = 1$, $(zwxy)^{-1}\, wx(zwxy) = wx^{-1} = w^2(wx)^{-1} = x^4(wx)^{-1} = (wx)^4(wx)^{-1} = (wx)^3$. Hence, $\langle wx,zy\rangle$ is $\mathbf{Z}_8 \rtimes \mathbf{Z}_2$, type 16/13.

C. The semidirect product $\mathbf{Z}_3 \rtimes D_8$

Consider the $\langle wx,zy\rangle$ subgroup of $\mathbf{Z}_2 \times (\mathbf{Z}_4 \circ Q_{24})$. We have $w^2 = x^6 = y^2$, $y^{-1}xy = x^{-1}$, and so $(wx)^6 = 1$. Clearly, wx can be replaced by $(wx)^2$ and $(wx)^3$ together, where $(wx)^2 = x^8$ is of order 3 and $(wx)^3$ is of order 2. Now $y^4 = 1$, and $\langle y,(wx)^3\rangle$ is D_8 since $[(wx)^3]^{-1}y(wx)^3 = (wx)^3y(wx)^3 = w^6x^3yx^3 = w^2y$ (because $w^4 = 1$ and $xyx = y$, so $x^3yx^3 = y$), and then $w^2y = y^{-1}$ since $w^2 = y^2$ and $y^3 = y^{-1}$. This D_8 acts on the $\mathbf{Z}_3$ generated by x^8 according to $y^{-1}x^8y = x^{-8}$, and so the group is $\mathbf{Z}_3 \rtimes D_8$, type 24/15.

D. The group 32/32

Not all of our holonomy groups can be expressed as a semidirect product of cyclic or dihedral groups. An interesting counterexample is provided by the $\langle zx,wy\rangle$ subgroup of $\mathbf{Z}_2 \times \mathbf{Z}_8 \circ Q_{16}$, with $z^2 = 1$, $w^4 = x^4 = y^2$, $y^{-1}xy = x^{-1}$. We have $(zx)^8 = (wy)^8 = 1$, but the group is of order 32 because $(zx)^4 = x^4 = w^4 = (wy)^4$, and $(wy)^{-1}zxwy = zx^{-1} = (zx)^{-1}$. This is the group 32/32, which is not a semidirect product. On the other hand, $\langle wx,zy\rangle$ is also generated by wx, $w^2(=(zwxy)^{-2})$ and w^2zy, which satisfy $(wx)^4 = (w^2)^4 = (w^2zy)^2 = 1$, $(w^2zy)^{-1}wx(w^2zy) = w^2(wx)^{-1}$, and so this subgroup is a semidirect product, $(\mathbf{Z}_4 \times \mathbf{Z}_4) \rtimes \mathbf{Z}_2$, type 32/31.

E. The quaternionic group Q_{36}

The reader may find it surprising that our list contains Q_{12}, Q_{20}, Q_{28}, Q_{36}, and so on, but not Q_8, Q_{16}, etc. The reason is simple: All of the quaternionic groups Q_{8n+4} are semidirect products $\mathbf{Z}_{2n+1} \rtimes \mathbf{Z}_4$, while the others are not. It is easy to verify, for example, that Q_{36}, defined by $x^9 = y^2$, $y^{-1}xy = x^{-1}$, is also generated by x^2 and y, where $(x^2)^9 = y^4 = 1$, $y^{-1}(x^2)y = (x^2)^{-1}$. Hence, the subgroup of $\mathbf{Z}_4 \times (\mathbf{Z}_4 \circ Q_{36})$ given by $\langle x^2,zwy\rangle$ is also isomorphic to $\mathbf{Z}_9 \rtimes \mathbf{Z}_4 = Q_{36}$. The corresponding statements for Q_{32} or Q_{40} would not be valid.

F. The group $\mathbf{Z}_2 \times \tilde{T}_{24}$

As there are only three binary polyhedral groups, they are of particular interest, and it is perhaps unfortunate that they appear in our list only once, and even then only at order 48. The tetrahedral group T_{12} is $D_4 \rtimes \mathbf{Z}_3$, and, corresponding, $\tilde{T}_{24}$ is $Q_8 \rtimes \mathbf{Z}_3$, with three generators x,y,v satisfying $x^2 = y^2$, $y^{-1}xy = x^{-1}$, $v^3 = 1$, $v^{-1}xv = y$, $v^{-1}yv = xy$. Taking $\mathbf{Z}_2 \times \mathbf{Z}_3 \times \tilde{T}_{24}$ (with $z^2 = w^3 = 1$), one sees easily that the subgroup $\langle z,x,y,wv\rangle$ is isomorphic to $\mathbf{Z}_2 \times \tilde{T}_{24}$. The binary polyhedral groups do not appear again until one reaches order 96. (The interesting group $\mathbf{Z}_4 \circ \tilde{T}_{24}$ is of order 48, but it occurs as the $\langle zw,x,y,v\rangle$ subgroup of $\mathbf{Z}_2 \times (\mathbf{Z}_4 \circ \tilde{T}_{24})$, which is forbidden because $n = 2m$.)

G. The group $\mathbf{Z}_5 \times D_{10}$

This group is remarkable because it is one of only nine groups in the table with order not equal to a multiple of 4. It occurs as the $\langle x^2,zwy\rangle$ subgroup of $\mathbf{Z}_2 \times \mathbf{Z}_{20} \circ Q_{20}$, a group of order 400. We have $x^{10} = (zwy)^{10} = 1$, so zwy can be replaced by $(zwy)^2$ and $(zwy)^5$. The former is w^{12} and generates $\mathbf{Z}_5$, while the latter is zw^5y, of order 2. Since $(zw^5y)^{-1}x^2(zw^5y) = x^{-2}$, we have the group $\mathbf{Z}_5 \times D_{10}$, of order 50.

This concludes our discussion of the derivation of Table I. We now consider an application.

VI. FUNDAMENTAL GROUPS OF CALABI-YAU SPACES ON WHICH E_6 BREAKS TO H_5

Table I tells us that *if*, on a given manifold M with fundamental group π, we can construct a vacuum gauge configuration with holonomy group (say) $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$, then we can break E_6 to H_5 on that manifold. We know that a *necessary* condition for this is that π should have a normal subgroup N with π/N isomorphic to the holonomy group. But—although this is universally taken for granted—it is very far from obvious that this condition is *sufficient*. Given, say, a manifold[4] with fundamental group $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$, one might expect to find that, in many cases, a principal E_6 bundle with holonomy group $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$ simply cannot be constructed on that manifold. It is a remarkable and regrettable fact that this cannot happen on physically interesting manifolds.

Theorem 9: Let M be a connected paracompact manifold with dimension greater than 1, let G be a connected Lie group, and let F be a finite subgroup of G. Then there exists a principal G bundle over M, with a connection having holonomy group isomorphic to F, if and only if the fundamental group π (of M) admits a normal subgroup N with π/N isomorphic to F.

The proof is a special case of one of the results of Ref. 11. It depends on the difficult Hano–Ozeki–Nomizu theorem,[8] and hence is anything but trivial. The theorem shows that on a manifold with fundamental group $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$, it is indeed possible to construct an E_6 bundle, together with a connection with holonomy group $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$; and since $(\mathbf{Z}_4 \rtimes \mathbf{Z}_4)/\mathbf{Z}_2 = D_8$, it is also possible to construct (another) E_6 bundle with a connection having holonomy group D_8. This justifies the assumptions of Ref. 4.

The theorem also shows us how to solve the second part of the Hosotani inverse problem: Now that we know precisely which Φ can break E_6 to H_5, what are the possibilities for π? We must find all π with $\pi/N = \Phi$. Of course, there are infinitely many such π for each given Φ, but again we are mainly interested in the groups of lowest order. That means that N should be as small as possible, and so we can ask, for example, for all groups π satisfying $\pi/\mathbf{Z}_1 = D_8$, $\pi/\mathbf{Z}_2 = D_8$, $\pi/\mathbf{Z}_3 = D_8$, and so on.

Now this is a very well-known problem in group theory: The groups π satisfying $\pi/N = \Phi$ are called extensions of Φ

(or of N, but the former is more appropriate here). The problem of finding extensions can be formulated in terms of cohomology theory[16] and is well understood in principle. There can, however, be few areas of mathematics in which practice diverges more completely from principle than here; he who wishes actually to compute the relevant extensions will find the study of the corresponding cohomology rings as futile as it is amusing. Instead we shall use a combination of elementary group theory and force. Table II presents a complete list of all groups of order less than 48 with π/N isomorphic to some group in Table I. (The number 48 was chosen for convenience; it would be tedious, but by no means impossible, to extend this to 60.) The full proof is, once again, far too long to be given here, and so we merely sketch the main ideas.

Notice first that we can always choose N to be trivial, and so $\pi = \Phi$ is one candidate for every Φ in Table I. Clearly, for $|\Phi| \geqslant 24$, these are the only extensions with order less than 48, and that is why Table II stops at order 20 (that is, to avoid needless repetition, not because these are not valid extensions).

Second, the case $N = \mathbf{Z}_2$ is fairly easy to handle, because it is not difficult to show that if $\mathbf{Z}_2$ is normal, then it must be a subgroup of the center. To find all π with $\pi/\mathbf{Z}_2 = D_{12}$, for example, one merely needs to examine the tables of groups of order 24 in Ref. 14, and factor by $\mathbf{Z}_2$ subgroups of the centers. For the groups of order 40 with $\pi/\mathbf{Z}_2 = D_{20}$, we proceed differently. Since the extensions are central, we take D_{20} (generated by a,b, with $a^{10} = b^2 = 1$, $b^{-1}ab = a^{-1}$) and introduce a new generator c that commutes with a and b and satisfies $c^2 = 1$. Then we put $a^{10} = c^{\alpha}$, $b^2 = c^{\beta}$, $b^{-1}ab = c^{\gamma}a^{-1}$, where α, β, γ are equal to 0 or 1 and must

TABLE II. All candidates of order less than 48, for fundamental groups of manifolds on which E_6 breaks to a subgroup of rank five. When $|\Phi| > 20$, the groups π with $|\pi| < 48$ are all of the form $\pi = \Phi$, and so may be read from Table I.

			$\lvert\chi\rvert$	
Φ	π	N	$g=3$	$g=4$
D_8	D_8	$\mathbf{Z}_1$	48	64
D_8	16/6,16/9,16/10,16/12,16/13,16/14	$\mathbf{Z}_2$	96	128
D_8	24/7,24/10,24/15	$\mathbf{Z}_3$	144	196
D_8	32/8,32/11,32/12,32/23,32/24, 32/25,32/27,32/28,32/29,32/30,32/36	$\mathbf{Z}_2 \times \mathbf{Z}_2$	196	256
D_8	32/14,32/21,32/26,32/31,32/32,32/34, 32/35,32/39,32/49,32/50,32/51	$\mathbf{Z}_4$	196	256
D_8	40/4,40/10,40/11	$\mathbf{Z}_5$	240	320
D_{12}	D_{12}	$\mathbf{Z}_1$	72	96
D_{12}	24/4,24/6,24/9,24/10,24/11,24/15	$\mathbf{Z}_2$	144	192
D_{12}	36/5,36/8,36/9,36/10,36/14	$\mathbf{Z}_3$	216	288
Q_{12}	Q_{12}	$\mathbf{Z}_1$	72	96
Q_{12}	24/6,24/14	$\mathbf{Z}_2$	144	192
Q_{12}	36/6,36/11,36/12	$\mathbf{Z}_3$	216	288
16/6	16/6	$\mathbf{Z}_1$	96	128
16/6	32/8,32/11,32/12,32/14,32/23,32/24, 32/25,32/26,32/33,32/34,32/35,32/36, 32/37,32/38,32/39,32/44,32/45	$\mathbf{Z}_2$	192	256
16/8	16/8	$\mathbf{Z}_1$	96	128
16/8	32/10,32/14,32/15,32/16,32/36,32/37 32/38,32/39,32/40,32/41	$\mathbf{Z}_2$	192	256
16/9	16/9	$\mathbf{Z}_1$	96	128
16/9	32/11,32/18,32/20,32/27,32/28,32/31, 32/46,32/47,32/48	$\mathbf{Z}_2$	192	256
16/10	16/10	$\mathbf{Z}_1$	96	128
16/10	32/12,32/18,32/21,32/29,32/30,32/32	$\mathbf{Z}_2$	192	256
16/11	16/11	$\mathbf{Z}_1$	96	128
16/11	32/13,32/19,32/20,32/21	$\mathbf{Z}_2$	192	256
16/12	16/12	$\mathbf{Z}_1$	96	128
16/12	32/23,32/27,32/29,32/49,32/50,32/51	$\mathbf{Z}_2$	192	256
16/13	16/13	$\mathbf{Z}_1$	96	128
16/13	32/24,32/27,32/28,32/30	$\mathbf{Z}_2$	192	256
$\mathbf{Z}_3 \times D_6$	$\mathbf{Z}_3 \times D_6$	$\mathbf{Z}_1$	108	144
$\mathbf{Z}_3 \times D_6$	36/5,36/6	$\mathbf{Z}_2$	216	288
D_{20}	D_{20}	$\mathbf{Z}_1$	120	160
D_{20}	40/6,40/7,40/8,40/10,40/11,40/14	$\mathbf{Z}_2$	240	320
Q_{20}	Q_{20}	$\mathbf{Z}_1$	120	160
Q_{20}	40/8,40/12	$\mathbf{Z}_2$	240	320

be determined. Clearly, $(\alpha,\beta,\gamma) = (0,0,0)$ is just $\mathbf{Z}_2 \times D_{20}$, while (1,1,0) gives $a^{10} = b^2$, $b^{-1}ab = a^{-1}$ which is Q_{40}; less trivially, (0,1,1) is the group with relations $a^{10} = 1 = b^4$, $b^{-1}ab = b^2a^{-1}$, which is also generated by a^2, a^5, b. These satisfy $(a^5)^{-1}ba^5 = a^5ba^5 = b^{-1}$ (because $aba = b^3 = b^{-1}$) as well as $b^{-1}a^2b = a^{-2}$. This is the group $\mathbf{Z}_5 \rtimes D_8$. This procedure yields all of the desired extensions.

For the groups of order 32 with $\pi/\mathbf{Z}_2$ isomorphic to a group of order 16, this method is too unwieldy. But in this case (and also for $\pi/\mathbf{Z}_2 = D_8$) one can use the tables of Ref. 13, but in reverse. Those tables give the "first quotient signals" of every group of order 32. Thus, if one wishes to find, for example, all $\mathbf{Z}_2$ extensions of D_{16} (the group Γ_{3a1} in the notation of Ref. 13), then he needs only to run through the tables and identify all groups of order 32 with first quotient signal $3a1$.

The case $N = \mathbf{Z}_3$ is more difficult because such extensions need not be central. The extensions of D_8 of order 24 are easily found using the Sylow theorems:[15] Every group of order 24 has a $\mathbf{Z}_3$ subgroup that is normal only if it is unique. The tables of Ref. 14 then quickly settle the matter. For the groups of order 36, the easiest procedure is simply to list all of them and compute. (The ten non-Abelian groups of order 36 are $\mathbf{Z}_6 \times D_6$, $D_6 \times D_6$, $\mathbf{Z}_3 \times Q_{12}$, $\mathbf{Z}_3 \times T_{12}$, $\mathbf{Z}_2 \times D_{18}$, $\mathbf{Z}_2 \times [(\mathbf{Z}_3 \times \mathbf{Z}_3) \rtimes \mathbf{Z}_2]$, $(\mathbf{Z}_3 \times \mathbf{Z}_3) \rtimes \mathbf{Z}_4$, $(\mathbf{Z}_2 \times \mathbf{Z}_2) \rtimes \mathbf{Z}_9$, Q_{36}, and $D_6 \rtimes D_6$.)

The cases $N = \mathbf{Z}_2 \times \mathbf{Z}_2$ and $N = \mathbf{Z}_4$, $\Phi = D_8$ are the most unpleasant of all. Take the full list of groups of order 32, with generators and relations given in Refs. 13 or 14, and compute. (That is, factor out all $\mathbf{Z}_2 \times \mathbf{Z}_2$ and $\mathbf{Z}_4$ subgroups.)

The final case, $N = \mathbf{Z}_5$, $\Phi = D_8$, is easy because of the following theorem:[15] If $\pi/N = \Phi$ and the integers $|N|$ and $|\Phi|$ are relatively prime, then π must be a semidirect product of N and Φ. By definition, semidirect products $\mathbf{Z}_5 \rtimes D_8$ correspond to homomorphisms from D_8 to the automorphism group of $\mathbf{Z}_5$, which is $\mathbf{Z}_4$. There are no homomorphisms from D_8 onto $\mathbf{Z}_4$, but there are two onto its $\mathbf{Z}_2$ subgroup [corresponding to $D_8/\mathbf{Z}_4$ and $D_8/(\mathbf{Z}_2 \times \mathbf{Z}_2)$] and so we obtain $\mathbf{Z}_5 \rtimes D_8$, a second group isomorphic to D_{40}, and of course $\mathbf{Z}_5 \times D_8$. This concludes our discussion of the derivation of Table II.

The table is arranged as follows. Each group of order less than 48 in Table I is listed under Φ. The second column gives a complete list of all fundamental groups π of manifolds on which it is possible to construct an E_6 bundle with a connection having a holonomy group isomorphic to Φ. The next column gives the normal subgroup N such that $\pi/N = \Phi$. The last column lists the corresponding values of χ for g = three or four generations. All of the π are listed according to their index numbers assigned either in Ref. 14 or in the Appendix.

VII. CONCLUSIONS

The original purpose of this work was to put the (slightly mysterious) finite group theory of Ref. 4 into perspective and to survey all possible groups that can arise in this context. Tables I and II give the complete solution for the lowest values of $|\chi|$. The fundamental group of the manifold considered in Ref. 4 is $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$, type 16/10. This group occurs in Table II (in the column labeled π) twice, corresponding to holonomy group $\Phi = D_8$ or $\Phi = 16/10$ itself. On this manifold, then, it is possible to break E_6 to H_5 in two distinct ways; this is in agreement with Ref. 4.

The main result of this paper is the fact that a complete, explicit solution of the "inverse problem" is possible in some cases, including $G = E_6$, $H = H_5$. But can one use this solution to guide a search for further examples of Calabi–Yau manifolds on which E_6 can be broken to a low-energy group of rank five? We consider that this may well be possible, along the following general lines.

First, the tables already strongly restrict the universal covering manifold $\tilde{M}$ by giving the possible values of $|\chi|$. The values less than 200 are (for three generations) 48, 72, 96, 108, 120, 144, 168, 180, and 192. This already considerably reduces, for example, the list given in Ref. 17.

Secondly, Ref. 4 shows that we have some direct control over Φ, as follows. Both Φ and the residual gauge symmetry $C(\Phi)$ are subgroups of the original gauge group G. The subset $\Phi \cdot C(\Phi)$ consisting of products of elements drawn from Φ and $C(\Phi)$ is in fact a *subgroup* of G. It is isomorphic to $[\Phi \times C(\Phi)]/Z\Phi$, where $Z\Phi$ is the center of Φ. Thus $[\Phi \times C(\Phi)]/Z\Phi$ [and *not* $\Phi \times C(\Phi)$ as is often said] is a subgroup of G. Irreducible representations of G will decompose into irreducible representations of $[\Phi \times C(\Phi)]/Z\Phi$, and not of the low-energy group $C(\Phi)$ alone. Thus, in general, Φ itself has an effect on the low-energy multiplets, and so it may be possible to use phenomenology directly to constrain Φ. That is precisely what is done, most ingeniously, by the authors of Ref. 4. In that case, $\pi = \mathbf{Z}_4 \rtimes \mathbf{Z}_4$, $G = E_6$, $\Phi = \mathbf{Z}_4 \rtimes \mathbf{Z}_4$ or D_8, and $[\Phi \times C(\Phi)]/Z\Phi = [(\mathbf{Z}_4 \rtimes \mathbf{Z}_4) \times H_5]/\mathbf{Z}_2$ or $[D_8 \times H_5]/\mathbf{Z}_2$ since both $\mathbf{Z}_4 \rtimes \mathbf{Z}_4$ and D_8 have centers isomorphic to $\mathbf{Z}_2$, and a careful analysis of the behavior of various multiplets under the action of these groups permits a direct deduction that $\Phi = D_8$ is to be preferred to $\Phi = \mathbf{Z}_4 \rtimes \mathbf{Z}_4$. It should be possible to generalize these arguments so that, independently of other considerations, certain candidates for Φ can be pronounced more (physically) interesting than others. Then candidates for π could be read from Table II.

Finally, of course, one must verify that π acts freely on $\tilde{M}$. That would eliminate many—or, in some cases, all—candidates for π. The remainder will yield spaces $\tilde{M}/\pi$ on which it is certainly possible to break E_6 to H_5 by means of the Hosotani mechanism. (Notice, however, that except for considerations involving the Euler characteristic, we have not hitherto used the assumption that π acts freely.) It would undoubtedly be preferable to have at least some necessary condition for π to act freely on simply connected Calabi–Yau manifolds, since that might permit us to eliminate many entries in the tables. At the level of generality (deliberately) maintained in this work—where $\tilde{M}$ can be *any* compact manifold, not necessarily even a Calabi–Yau space—there is little hope of finding such criteria; yet it might be possible for more specialized classes.

In short, then, the possession of a complete list of finite holonomy groups capable of breaking E_6 to a low-energy

group of rank five does suggest a program for finding examples of the appropriate Calabi–Yau manifolds. We hope to pursue this in a more specific context elsewhere.

ACKNOWLEDGMENTS

The author wishes to thank Professor J. Goldstone, whose hospitality at the Center for Theoretical Physics made it possible to complete the lengthy computations involved in this work.

This work is supported in part by funds provided by the U.S. Department of Energy (D.O.E) under contract # DE-AC02-76ER03069.

APPENDIX: SELECTED GROUPS OF ORDER >32

Reference 14 lists all groups of order less than 33. Here we extend the list and give definitions, in terms of generators and relations, of some groups of higher order. Note that these lists are not complete (except for order 36): we have included only those groups, of each order, which appear in Tables I and II. Furthermore, we have not always included product groups, since the structure of these is obvious.

Group	Index No.	Generators	Nontrivial relations
$\mathbf{Z}_6 \times D_6$	36/5	$a^3 = b^2 = c^6 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_3 \times Q_{12}$	36/6	$a^6 = b^4 = c^3 = 1$	$b^{-1}ab = a^{-1}, a^3 = b^2$
$\mathbf{Z}_3 \times T_{12}$	36/7	$a^2 = b^2 = c^3 = d^3 = 1$	$c^{-1}ac = b, c^{-1}bc = ab$
$\mathbf{Z}_2 \times D_{18}$	36/8	$a^9 = b^2 = c^2 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_2 \times [(\mathbf{Z}_3 \times \mathbf{Z}_3) \rtimes \mathbf{Z}_2]$	36/9	$a^3 = b^3 = c^2 = d^2 = 1$	$c^{-1}ac = a^{-1}, c^{-1}bc = b^{-1}$
$D_6 \times D_6$	36/10	$a^3 = b^2 = c^3 = d^2 = 1$	$b^{-1}ab = a^{-1}, d^{-1}cd = c^{-1}$, $d^{-1}ad = a^{-1}, c^{-1}bc = ba$
$(\mathbf{Z}_3 \times \mathbf{Z}_3) \rtimes \mathbf{Z}_4$	36/11	$a^3 = b^3 = c^4 = 1$	$c^{-1}ac = a^{-1}, c^{-1}bc = b^{-1}$
Q_{36}	36/12	$a^{18} = b^4 = 1$	$b^{-1}ab = a^{-1}, a^9 = b^2$
$(\mathbf{Z}_2 \times \mathbf{Z}_2) \rtimes \mathbf{Z}_9$	36/13	$a^2 = b^2 = c^9 = 1$	$c^{-1}ac = b, c^{-1}bc = ab$
$D_6 \times D_6$	36/14	$a^3 = b^2 = c^3 = d^2 = 1$	$b^{-1}ab = a^{-1}, d^{-1}cd = c^{-1}$
$\mathbf{Z}_5 \times D_8$	40/4	$a^4 = b^2 = c^5 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_4 \times D_{10}$	40/6	$a^5 = b^2 = c^4 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_2 \times D_{20}$	40/7	$a^{10} = b^2 = c^2 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_2 \times Q_{20}$	40/8	$a^{10} = b^4 = c^2 = 1$	$b^{-1}ab = a^{-1}, a^5 = b^2$
D_{40}	40/10	$a^{20} = b^2 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_5 \rtimes D_8$	40/11	$a^4 = b^2 = c^5 = 1$	$b^{-1}ab = a^{-1}, a^{-1}ca = c^{-1}$
$\mathbf{Z}_5 \rtimes \mathbf{Z}_8$	40/12	$a^5 = b^8 = 1$	$b^{-1}ab = a^{-1}$
Q_{40}	40/14	$a^{20} = b^4 = 1$	$b^{-1}ab = a^{-1}, a^{10} = b^2$
$\mathbf{Z}_{24} \rtimes \mathbf{Z}_2$	48/23	$a^{24} = b^2 = 1$	$b^{-1}ab = a^{11}$
$D_{24} \rtimes \mathbf{Z}_2$	48/24	$a^{12} = b^2 = c^2 = 1$	$b^{-1}ab = a^{-1}, c^{-1}bc = a^6b$
$\mathbf{Z}_{12} \rtimes \mathbf{Z}_4$	48/25	$a^{12} = b^4 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_3 \rtimes \mathbf{Z}_{16}$	48/26	$a^3 = b^{16} = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_3 \rtimes [16/9]$	48/27	$a^4 = b^2 = c^2 = d^3 = 1$	$b^{-1}ab = ca^{-1}, a^{-1}da = d^{-1}$
$\mathbf{Z}_3 \rtimes [16/10]$	48/28	$a^4 = b^4 = c^3 = 1$	$b^{-1}ab = a^{-1}, a^{-1}ca = c^{-1}$
$\mathbf{Z}_3 \rtimes [16/13]$	48/29	$a^8 = b^2 = c^3 = 1$	$b^{-1}ab = a^5, a^{-1}ca = c^{-1}$
$\mathbf{Z}_7 \rtimes \mathbf{Z}_8$	56/10	$a^7 = b^8 = 1$	$b^{-1}ab = a^{-1}$
$\mathbf{Z}_7 \rtimes D_8$	56/11	$a^4 = b^2 = c^7 = 1$	$b^{-1}ab = a^{-1}, a^{-1}ca = c^{-1}$

[1] P. Candelas, G. T. Horowitz, A. Strominger, and E. Witten, Nucl. Phys. B **258**, 46 (1985).
[2] J. Preskill, Nucl. Phys. B **323**, 141 (1989).
[3] Y. Hosotani, Phys. Lett. B **126**, 309 (1983).
[4] J. Ellis, K. Enqvist, S. Kalara, D. V. Nanopoulos, and K. A. Olive, Nucl. Phys. B **306**, 445 (1988).
[5] A. Strominger and E. Witten, Commun. Math. Phys. **101**, 341 (1985).
[6] B. Booss and D. D. Bleecker, *The Atiyah–Singer Index Formula and Gauge-Theoretic Physics* (Springer, New York, 1985).
[7] A. E. Fischer, Commun. Math. Phys. **113**, 231 (1987).
[8] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. I.
[9] A. L. Besse, *Einstein Manifolds* (Springer, Berlin, 1987).
[10] J. A. Wolf, *Spaces of Constant Curvature* (McGraw–Hill, New York, 1967).
[11] B. McInnes, J. Phys. A **22**, 2309 (1989).
[12] M. A. Armstrong, *Groups and Symmetry* (Springer, New York, 1988).
[13] M. Hall and J. K. Senior, *The Groups of Order 2 " (n≤6)* (Macmillan, New York, 1964).
[14] A. D. Thomas and G. V. Wood, *Group Tables* (Shiva, Orpington, 1980).
[15] M. Suzuki, *Group Theory* (Springer, Berlin, 1982), Vol. I.
[16] K. S. Brown, *Cohomology of Groups* (Springer, New York, 1982).
[17] P. Candelas, A. M. Dale, C. A. Lütken, and R. Schimmrigk, Nucl. Phys. B **298**, 493 (1988).

A new approach to global stability analysis of one-dimensional continuous dissipative systems

E. Santamato,[a] G. Abbate,[b] and P. Maddalena[a]
Dipartimento di Scienze Fisiche, Università di Napoli, Pad. 20, Mostra d'Oltremare, 80125 Napoli, Italy

(Received 14 November 1989; accepted for publication 9 May 1990)

It is proved by a direct construction that symmetry-breaking instabilities in one-dimensional dissipative continuous systems can be studied in terms of a suitable function G whose minima correspond to stable steady states of the system, while other extremal points correspond to unstable states. The existence of such a function is proved for a large class of continuous dissipative physicochemical systems in one spatial dimension. The function G is exact, in the sense that it is obtained by taking into account all nonlinear terms in the evolution equations of the system. Since in some respect this function has properties similar to the Ginzburg–Landau function widely used in second-order phase transitions, it is called the nonlinear Ginzburg–Landau function (NLGLF) of the system. The NLGLF may be useful for studying the global stability under symmetry-breaking conditions as well as the character (first and second order) of the transitions between different steady states. The construction of the NLGLF usually requires simple numerical or analytical calculations. A specific example, taken from the physics of liquid crystals, has been worked out analytically in the present paper.

I. INTRODUCTION

Symmetry-breaking instabilities have been known from a very long time in hydrodynamics,[1] but only in more recent times they have been extensively studied in spatially inhomogeneous physicochemical systems,[2] In the last decade, much effort was devoted to the study of symmetry-breaking instabilities in nonlinear optics of intense laser beams in Kerr media[3] and in liquid crystalline media.[4]

Symmetry breaking may occur only when the system is driven far from equilibrium by external constraints, some of which are under the control of the experimenter. Variation of the control parameters modifies the motion of the system. Over a certain range of values of the control parameters, the trajectories of the motion undergo quantitative modifications, whereas at critical values of the control parameters the system trajectories undergo qualitative changes. Specifically, we will consider trajectories of the perturbed motion so that when a critical value of the control parameters is attained, by varying the system constraints, there can occur an unstable transition between two steady states of the system, having different spatial organization.

The systems exhibiting spatial symmetry breaking are usually described by a set of nonlinear partial differential equations in space and time coordinates. Several mathematical methods have been exploited to study the occurrence of instability in such system as the fixed point or the maximum principle method,[5] but certainly the most useful and widely used ones are Liapounov's first and second methods. Liapounov's first method consists of linearizing the given set of differential equations around the given reference solution, which may correspond to the equilibrium state or to any other steady state of the system. Boundary conditions are then exploited to expand the solution of the linearized problem in a series of spatial modes. The time dependence of each mode is assumed of the form $\exp(\lambda t)$, λ being the Liapounov exponent of the mode. The resulting eigenvalue problem is finally solved to obtain the set of Liapounov's exponents. The reference state will be locally asymptotically stable if and only if all Liapounov's exponents have a negative real part. This method, although yielding both necessary and sufficient conditions for asymptotic stability, leads to very long calculations, since it must be repeated for each steady-state solution of the full nonlinear problem. Moreover, this method yields criteria for local stability only, i.e., for stability with respect to very small perturbations. The regions of stability of the system (i.e., the range of allowable perturbation amplitudes for which the reference state remains stable) cannot be determined by Liapounov's first method.

A not local sufficiency criterion for stability is provided by the second (or direct) Liapounov's method. This method was in great vogue some decades ago, especially because it was hoped that it would provide a general nonlinear thermodynamic criterion for stability of far from equilibrium dissipative systems.[2] The direct method was first formulated to study the stability of finite-dimensional systems (described by a set of total differential equations) and then adapted to treat continuous media.[6] In the finite-dimensional case, the method consists of looking for a special function, the Liapounov function, of the thermodynamic variables of the system that will act as a bowl with the reference steady state at its lowest point and such that the solution of the equations of motion always runs down to this lowest point. In the continuum case, the Liapounov function must be replaced by an appropriate Liapounov's functional of the thermodynamic variables and their spatial derivatives. The direct method has the disadvantage of providing only sufficient criteria for stability. It should be noted, however, that when physicists formulate complex problems, they generally start with an ac-

[a] Also at Centro Interuniversitario di Elettronica Quantistica e Plasmi I-80125, Napoli, Italy.
[b] Also at Centro Interuniversitario di Struttura della Materia I-80125, Napoli, Italy.

 0022-2488/90/092105-11$03.00

tion, or a free energy, or a Liapounov functional, because they know that their problem has a conserved quantity or at least an equilibrium state, and that general differential equations will not have integrating factors that lead to this. This phenomenological point of view will be adopted in this paper too, so that we will assume the existence of a Liapounov functional far from the beginning [see Eq. (1), below].

As we noted before, the passage from discrete to continuum systems implies the replacement of the Liapounov function with a functional. Since a functional is an object much more complicated that a function, this passage is not without cost, as we will see from the following example.

Let us suppose, for instance, that the plot of the Liapounov function for a finite-dimensional system is as shown in Fig. 1. Here, q denotes the set of thermodynamical coordinates of the system. Then, the minima of $G(q)$ correspond to the stable and the maxima to the unstable steady states of the system. Moreover, since $G(q)$ accounts for all the system nonlinearities, the distance between two successive maxima yields the stability region for the state M. If external control parameters are present, we obtain a family of Liapounov functions $G(q;\alpha)$, α denoting the set of control parameters. The study of $G(q;\alpha)$ in function of the control parameters permits us to find the critical values of I at which instability occurs. Moreover, from this study, we can also deduce the character of the transition (first or second order[7]) at the critical point, as shown in Fig. 2. From the figure, the similarity between the Liapounov function in the finite-dimensional systems and the Ginzburg–Landau function used to describe phase transitions[8] is evident, the q's playing the role of order parameters. It should be noted, however, that the Ginzburg–Landau function is usually obtained after a power expansion in the order parameter, whereas $G(q;\alpha)$ accounts for all nonlinearities of the system. We may say, therefore, that, in the finite-dimensional case, the Liapounov function also provides a nonlinear Ginzburg–Landau function (NLGLF) for the system.

The situation is quite different in the case of continuous

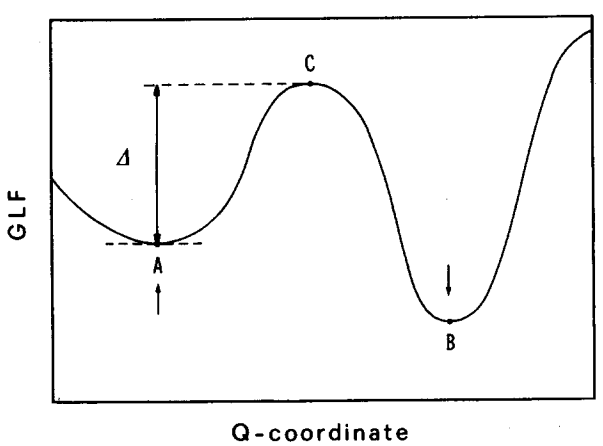


FIG. 1. A typical NLGLF for a multisteady-state system. The minima (indicated by arrows) correspond to stable steady-states, the maxima to unstable steady states. Here Δ is the energy barrier between the stable states A and B. The energy Δ must be provided to the system by external sources to induce the transition $A \to B$.

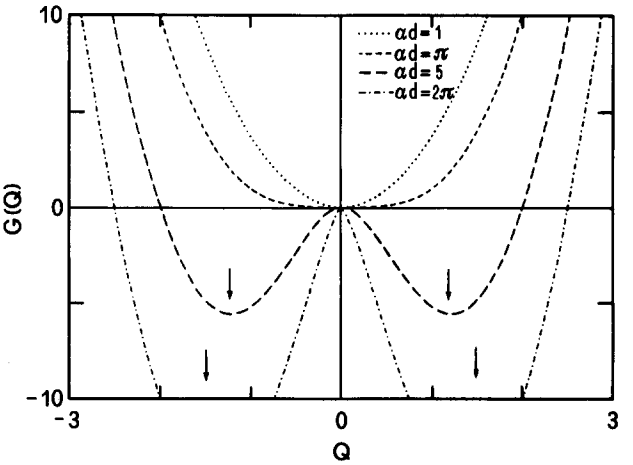


FIG. 2. The NLGLF $G(Q;\alpha)$ of the system considered in Sec. VIII for various values of αd (d is the sample thickness). The equilibrium state $Q=0$ becomes unstable for $\alpha d > \pi$. The positions of the minima of $G(Q)$ are indicated by the arrows. For $\alpha d = 2\pi$, $G(Q)$ has a cusp at $Q=0$.

systems. The thermodynamical coordinates $q = q(z,t)$ are now functions of both time t and space coordinate z. Physically acceptable steady states of the system are described by some curves $\bar{C}$: $q = \bar{q}(z)$ in the space X of the thermodynamic coordinates. The Liapounov function is replaced now by a suitable functional $F[C]$, defined on the set of continuous curves C: $q = q(z)$ in X, having class C_1. (The physical meaning of $F[C]$ depends on the particular problem envisaged; in many cases it may be identified with the total free energy of the system.) The main property of the Liapounov functional $F[C]$ is that *stable* steady states $\bar{q}(z)$ of the systems afford a *minimum* value to $F[C]$. In this work, the existence of such an $F[C]$ is taken for granted.

Although the functional $F[C]$ itself may be of some utility in studying the stability of the system steady states [see Ref. 6], one of the most useful properties of the Liapounov function is lost: namely, the property that just looking to the plot of this function permits us to get immediate global information on all steady states of the system as well as on all regions of stability.

The main purpose of this work is to show that for a large class of nonlinear, one-dimensional, space-dependent dissipative systems, a global stability analysis of steady states can be performed by replacing the functional $F[C]$ by an ordinary function $G(q)$ of the thermodynamic coordinates, having, in essence, the same properties of the nonlinear Ginzburg–Landau function employed in the case of discrete systems. In particular, *minima* of $G(q)$ are put in one-to-one correspondence with *asymptotically stable* steady states of the system. The NLGLF $G(q)$ is obviously intimately related to the functional $F[C]$ and, as we will see below, we may always arrange the system so that, at the point $q = \bar{q}$, corresponding to the steady state $\bar{q}(z)$ of the system, the function $G(q)$ takes the same value as the function $F[C]$ evaluated along $\bar{q}(z)$. It should be stressed, however, that the existence of an NLGLF $G(q)$ also for continuous systems is a nontrivial result pursued in this paper.

In Sec. II the main properties of the simple dissipative

systems considered in this paper are revised, the total free-energy and the dissipation functionals are introduced, and their general relationships with the equations of motion, governing the decay of the system toward equilibrium, are established. The NLGL function $G(q;\alpha)$ is introduced in Sec. III, which contains the original part of the paper. The remarkable properties of this function are outlined in this section, while Sec. IV is devoted to their rigorous demonstration. The explicit construction of the NLGLF $G(q;\alpha)$ is made in Sec. V with particular emphasis on its suitability for numerical computation. Sections VI and VII deal with the particular cases of fixed boundary conditions and of a multivalued function $G(q;\alpha)$. Finally, in Sec. VIII a significant example, taken from the physics of liquid crystals, is worked out analytically. Two appendices are devoted to the most technical aspects of the theory.

II. SIMPLE DISSIPATIVE SYSTEMS

Consider a continuous dissipative system described by a set of internal generalized thermodynamical coordinates $q^i(z,t)$ $(i=1,2,...,N)$ expressed as functions of one space coordinate z and time. Equilibrium steady states are described by curves C: $q^i = q^i(z)$ in the N-dimensional space spanned by the thermodynamical coordinates. In general, we allow for multiple steady states.

Stable equilibrium steady states are assumed to afford local minima to a suitable functional $F[C]$, having the general form

$$F[C] = (C)\int_0^d L(q,q',z)dz + f_0(q_0) - f_d(q_d), \quad (1)$$

where $L(q,q',z)$ is a function of class C_2 in its arguments, q' denoting the tangent vector to the curve C: $q'(z) = dq^i/dz$. The arbitrary functions $f_0(q)$ and $f_d(q)$ lead to an explicit dependence of F from the endpoints $q_0 = q^i(0)$ and $q_d = q^i(d)$ of C. The functions f_0 and f_d in Eqs. (1) are usually referred to as the *surface potentials*.[9] These potentials are phenomenologically introduced to model the interfacial interactions at the sample walls. The gradients of the surface potentials with respect to the thermodynamical coordinates q^i are then interpreted as the surface forces governing boundary conditions at $z=0$ and $z=d$.

Interaction with externally applied fields is included in F. In the following, we will always refer to F as the total free-energy functional and to $L(q,q',z)$ as the corresponding density of free energy. The function $L(q,q',z)$ is supposed to satisfy Legendre's sufficiency condition for minima:

$$\det\left[\frac{\partial^2 L}{\partial q'^i\,\partial q'^j}\right] > 0. \quad (2)$$

Relaxation to equilibrium may be described in terms of a dissipation functional that we assume having the form

$$D(t) = \int_0^d \frac{1}{2}\left[R_{ij}\frac{\partial q^i}{\partial t}\frac{\partial q^j}{\partial t}\right]dz. \quad (3)$$

The dissipation functional $D(t)$ is supposed to have been evaluated along the actual path $q^i = q^i(z,t)$, followed by the system in its decay toward equilibrium. The constant matrix R_{ij} is assumed to be positive definite, so that $D(t)\geqslant 0$ and dissipation vanishes in steady states, where fluxes vanish as well. Unless otherwise specified, summation over repeated indices is always intended. The factor $\frac{1}{2}$ in Eq. (3) has been inserted for further convenience. This factor scales $D(t)$ so that, in most practical cases, $2D(t)$ is the total power dissipated in the system during relaxation.

The existence of the dissipation functional (3) guarantees that *minima* of $F[C]$ correspond to *asymptotically stable* steady states of the system. Power balance requires, in fact, that the total free energy lost by the system in the unit time should equate the power dissipated by internal fluxes, viz.,

$$\frac{dF}{dt} = -2D(t) \leqslant 0. \quad (4)$$

In Eq. (4) the functional $F[C]$ must be evaluated along the real path $q(z,t)$ followed by the system during relaxation. Now, let us suppose that the time-independent functions $\bar{q}^i(z)$ afford a local *minimum* value to the free-energy functional[1] and let $\bar{F}$ be the value of F taken along $\bar{q}^i(z)$. Then, any other functions $q^i(z,t)$ will render the free energy larger than $\bar{F}$, provided, of course, that $q^i(z,t)$ is close enough (in some topological sense) to the reference steady state $\bar{q}^i(z)$. Then, we have $F(t) - \bar{F} \geqslant 0$ and, from Eq. (16), $d(F(t) - \bar{F})/dt = dF/dt \leqslant 0$. From this, we see that the functional $(F - \bar{F})$ behaves as a Liapounov functional for the system, which guarantees the asymptotic stability of the reference steady state $\bar{q}^i(z)$.

From the power balance equation (4), we may easily obtain the differential equation governing the decay to equilibrium (or to any other steady state of the system). Insertion of definitions (1) and (3) into the power balance equation (4) yields, in fact,

$$\int_d^0 \left[R_{ij}\frac{\partial q^j}{\partial t} - E_i(q,q',q'',z)\right]\left(\frac{\partial q^i}{\partial t}\right)dz + \left[\frac{\partial L}{\partial q'^i} - \frac{\partial f_d}{\partial q^i}\right]_{z=d} - \left[\frac{\partial L}{\partial q'^i} - \frac{\partial f_0}{\partial q^i}\right]_{z=0} = 0, \quad (5)$$

with

$$E_i(q,q',q'',z) = \frac{\partial}{\partial z}\left(\frac{\partial L}{\partial q'^i}\right) - \frac{\partial L}{\partial q^i}. \quad (6)$$

Since Eq. (5) must hold for arbitrary $q^i(z,t)$ at any fixed time t, the time evolution of the system is governed by the parabolic partial differential equations

$$E_i(q,q',q'',z) = R_{ij}\frac{\partial q^j}{\partial t}, \quad (7)$$

supplemented with the boundary conditions

$$\frac{\partial L}{\partial q'^i} = \frac{\partial f_0}{\partial q^i}, \quad \text{at } z=0$$

and

$$\frac{\partial L}{\partial q'^i} = \frac{\partial f_d}{\partial q^i}, \quad \text{at } z=d. \quad (8)$$

In view of condition (2), Eqs. (8) can be solved with respect to q'^i, yielding the boundary conditions in the standard form $q'^i = g^i_{0,d}(q)$ at $z=0$ and $z=d$, respectively [$g^i{}_0(q)$ and $g^i{}_d(q)$ are functions of the internal coordinates].

Beside boundary conditions (8), steady states of the system obey the set of ordinary differential equations

$$E_i(q,q',q'') = 0. \tag{9}$$

As shown by Eqs. (6), Eqs. (9) are the Euler–Lagrange equations associated with $F[C]$ and, therefore provide necessary (but in general not sufficient) conditions for affording a minimum value to the total free-energy functional.

Equations (7) are usually interpreted as the balance between the bulk thermodynamical forces on the left and the bulk dissipative forces on the right. The boundary conditions (8), instead, express the balance between surface thermodynamical forces on the left and externally imposed surface constraints on the right. Notice that, in the present model, dissipation at the surfaces is neglected.

In the following, we will refer to a system obeying a set of evolution equations of the form (7), with boundary conditions (8), as to a *simple* (one-dimensional) dissipative system. Only simple dissipative systems will be considered in this paper. It should be noted, however, that many dissipative systems of practical interest may be considered as simple. They include heat conduction, chemical reactions in the presence of diffusion, reorientation in liquid crystals, optical bistability in inhomogeneous media, etc.

III. THE NONLINEAR GINZBURG–LANDAU FUNCTION FOR SIMPLE CONTINUOUS SYSTEMS

In this section we prove the existence of an NLGLF $G(q)$ for simple systems by direct construction. It is obvious that the possibility of having an NLGLF for these systems is strongly related to the existence of the free-energy functional (1). In order to construct our NLGLF, we use Carathéodory's method of equivalent integrals.[10] Let $\bar{z}$ be an arbitrarily fixed point between 0 and d. Then, the $(n+1)$-dimensional space R_{n+1} formed by the coordinates q_i and z is divided in two regions G_0 and G_d by the hyperplane $z = \bar{z} = \text{const}$. Now, let $S_0(q,z)$ and $S_d(q,z)$ be two functions defined in G_0 and G_d, respectively, obeying the boundary conditions

$$\begin{aligned} S_0(q,0) &= f_0(q), \\ S_d(q,d) &= f_d(q), \end{aligned} \tag{10}$$

for any q.

Following Carathéodory, we want to see if it is possible to choose $S_0(q,z)$ and $S_d(q,z)$ so that they also obey the differential inequalities

$$\begin{aligned} dS_0(q,q',z) - L(q,q',z)dz &\leqslant 0, \\ dS_d(q,q',z) - L(q,q',z)dz &\leqslant 0, \end{aligned} \tag{11}$$

for any given line elements (q,q',z) in G_0 and G_d, respectively. How to construct explicitly the functions S_0 and S_d will be shown in the next section. For the moment, let us suppose that such functions $S_0(q,z)$ and $S_d(q,z)$ have been found. Then, if $q^i = g^i(z)$ is an arbitrary curve C joining the hyperplanes $z = 0$ and $z = d$ in R_{n+1}, we may integrate the first of Eqs. (11) along C between the points $P_0 = (q(0),0)$ and $\bar{P} = (\bar{q} = q(\bar{z}),\bar{z})$, obtaining

$$_{(C)}\int_{\bar{z}}^{0} L(q,q',z)dz \geqslant S_0(\bar{q},\bar{z}) - S_0(q(0),0). \tag{12a}$$

Analogously, we can integrate the second of Eqs. (11) along C, between the points $\bar{P}(\bar{q},\bar{z})$ and $P_d = (q(d),d)$, obtaining

$$_{(C)}\int_{\bar{z}}^{d} L(q,q',z)dz \geqslant S_d(q(d),d) - S_d(\bar{q},\bar{z}). \tag{12b}$$

By adding Eqs. (12a) and (12b) together and considering also the boundary conditions (10), we see that the free-energy functional (1) evaluated along the *arbitrary* curve C (of class C_1) obeys the inequality

$$F[C] \geqslant S_0(\bar{q},\bar{z}) - S_d(\bar{q},\bar{z}), \tag{13}$$

for any fixed point $\bar{z}$ between 0 and d. We remark, however, that the coordinates $\bar{q}^i$ depend on the curve C, since they are given by $\bar{q}^i = g^i(\bar{z})$.

It is now almost evident that an NLGLF for our system could be defined as

$$G(Q) = S_0(Q,\bar{z}) - S_d(Q,\bar{z}). \tag{14}$$

In fact, let Q^{*i} be a point where $G(Q)$ has a minimum value G^*. Then, inequality (13) shows that all curves C joining the hyperplanes $z = 0$ and $z = d$ and intersecting the hyperplane $z = \bar{z}$ in a point $\bar{P}$ other than $P^* = (Q^*,\bar{z})$ afford larger values to the free-energy integral than the set of curves passing through P^* [if the minimum of $G(Q)$ is local, $\bar{P}$ must be in a suitable neighborhood of P^* on the hyperplane $z = \bar{z}$]. Moreover, as will be shown in the next section by explicit construction, we can choose the functions S_0 and S_d so that, among the last curves, there is one curve, C^* say, for which *the equality* holds in Eq. (13). This curve affords, of course, a minimum value to the free-energy integral with respect all varied curves between $z = 0$ and $z = d$ intersecting the hyperplane $z = \bar{z}$ in a neighborhood of P^*. The curve C^* must then necessarily obey all requirements imposed by the variational problem $\delta F = 0$. In particular, C^* must obey the Euler–Lagrange equations (9) as well as the boundary conditions (8) so that C^* represents a steady state of the system. Moreover, since C^* affords a minimum value to F, this state is asymptotically stable. We have established, in this way, a correspondence between asymptotically stable states of the system and minima of the ordinary function $G(Q)$.

We may also make this correspondence more explicit. First, we fix our attention on a given point $\bar{P} = (Q,\bar{z})$ on the hyperplane $z = \bar{z}$ and consider the set of all curves passing through $\bar{P}$. Among these curves there will be one curve, $\bar{C}$ say, that minimizes the functional F. Along $\bar{C}$, equality holds in Eq. (13), i.e., $F[\bar{C}] = G(Q)$. Under a very general hypothesis (see Sec. VIII), this establishes a one-to-one correspondence between curves in R_{n+1} and values of the NLGLF $G(Q)$. In general, the curve $\bar{C}$ corresponding to $G(Q)$ *does not* represent a possible steady state of the system, since $\bar{C}$ minimizes F only with respect to all curve passing through $\bar{P}$ and not with respect all curves between $z = 0$ and $z = d$. Only the curves C^* corresponding to points Q^{*i} affording *extremal values* G^* to $G(Q)$, in fact, can be associated to steady states of the system and, in particular, the curves corresponding to minima of $G(Q)$ to steady states that are asymptotically stable.

The NLGLF $G(Q)$ as defined in Eq. (14) has many remarkable properties:

(i) $G(Q;\alpha)$ is an ordinary function of the coordinate Q and of the external parameters α;

(ii) the extremal points of $G(Q;\alpha)$ for fixed α correspond uniquely to the steady-state solutions of Eqs. (7), governing the evolution of the system;

(iii) the *minima* of $G(Q;\alpha)$ for fixed α correspond to *stable* steady states, other extremal points to unstable states.

(iv) the difference of the GLF between two consecutive minima and maxima yields the amount of energy Δ that must be provided to the system for inducing a transition between the two stable states separated by the barrier Δ.

This is shown in Fig. 1. All these properties of the NLGLF $G(Q)$ will be proved in the next section.

IV. PROPERTIES OF THE NONLINEAR GINZBURG–LANDAU FUNCTION

The curve $\bar{C}$ for which equality holds in Eq. (13) can be obtained by adjusting its slope so to minimize the left-hand sides of Eqs. (11) at any point (q,z) of R_{n+1}. For the sake of brevity, let us denote with $S(q,z)$ either of the functions $S_0(q,z)$ or $S_d(q,z)$ and with G the corresponding regions G_0 and G_d of R_{n+1} where these functions are defined, respectively. Then, inequalities (11) can be rewritten both as $dS - L\,dz \leqslant 0$ or, equivalently,

$$\frac{\partial S}{\partial z} + q'^i \frac{\partial S}{\partial q^i} - L(q,q',z) \leqslant 0. \tag{15}$$

The proper slope of the curve $\bar{C}$ is found by minimizing the left-hand side of Eq. (15) with respect to q'^i for fixed (q,z). This yields the canonical relation

$$p^i \equiv \frac{\partial L}{\partial q'^i}(q',q,z) = \frac{\partial S}{\partial q^i(q,z)}, \tag{16}$$

as well as Legendre's necessary conditions (2).

In view of the last conditions, Eqs. (16) can be solved with respect to q'^i, yielding

$$q'^i = \frac{dq^i}{dz} = \Psi^i(q,z). \tag{17}$$

This is a set of ordinary differential equations whose solutions $q^i = q^i(z)$ yield the minimizing curves $\bar{C}$ in G. Obviously, since S represents any one of the functions S_0 or S_d, we have two different sets of differential equations of the form (17) in each of the regions G_0 and G_d, respectively. Evaluating the canonical relations (16) at $z = 0$ and $z = d$ and comparing with Eqs. (10), we see that all the solutions of Eqs. (17) obey also the boundary conditions (8). More precisely, through any point $P_0(q_0,0)$ of the hyperplane $z = 0$ in R_{n+1} passes one solution $\bar{C}_0$ of Eqs. (17) [with $S(q,z)$ replaced by $S_0(q,z)$] obeying the first of boundary conditions (8) at P_0. Similarly, through any point $P_d(q_d,d)$ on the hyperplane $z = d$ passes one solution $\bar{C}_d$ of Eqs. (17) [with $S(q,z)$ replaced by $S_d(q,z)$] obeying the second of boundary conditions (8) at P_d. Equations (17) define, therefore, two congruences K_0 and K_d of solution curves in the regions G_0 and G_d of R_{n+1}. In the following, we will suppose that the congruences K_0 and K_d cover their respective regions G_0 and G_d simply, i.e., that one and only one member of the congruences passes through any point of the regions.

Since the regions G_0 and G_d are divided by the hyperplane $z = \bar{z}$, through any point $\bar{P}$ on this plane will pass one curve of K_0 and one curve of K_d, intersecting at $\bar{P}$. In this way, we may put in one-to-one correspondence the points $\bar{P}$ of the hyperplane $z = \bar{z}$ with *a couple* of curves, namely, the members of K_0 and K_0 meeting at $\bar{P}$ (see Fig. 3). This correspondence will have a fundamental role in the present theory.

Up until now the function $S(q,t)$ is completely arbitrary. We can specify it by imposing that along the minimizing curve $\bar{C}$ equality must hold in Eqs. (11). Then, inserting Eqs. (16) into Eqs. (11) (with the equality sign) yields

$$\frac{\partial S}{\partial z} + q'^i \frac{\partial L}{\partial q'^i} - L = 0 \tag{18}$$

or, equivalently,

$$\frac{\partial S}{\partial z} + H(\nabla S,q,z) = 0, \tag{19}$$

where $\nabla S = \partial S/\partial q^i$ and $H(p,q,z)$ is the Hamiltonian associated with $L(q,q',z)$, expressed as a function of the canonical coordinates and momenta. We see from Eq. (19) that $S(q,z)$ (and hence both S_0 and S_d) are solutions of the Hamilton–Jacobi equation associated with H. The boundary conditions (10) may be considered as Cauchy data for Eq. (19). These data uniquely determine the functions $S_0(q,z)$ and $S_d(q,z)$. From Eq. (14), we see that the NLGLF is given by the difference of the two solutions of the Hamilton–Jacobi equation (19), having Cauchy data (10), evaluated at $z = \bar{z}$. The NLGLF is, therefore, not unique, different choices of the point $\bar{z}$ yielding different NLGLF. If the free energy of the system also depends on a set of external parameters α as well, then S_0 and S_d [and hence also $G(Q)$] will depend on the parameters α. Although not explicitly written, this dependence on the external parameters will be always intended.

Once the appropriate solutions $S_0(q,z)$ and $S_d(q,z)$ of

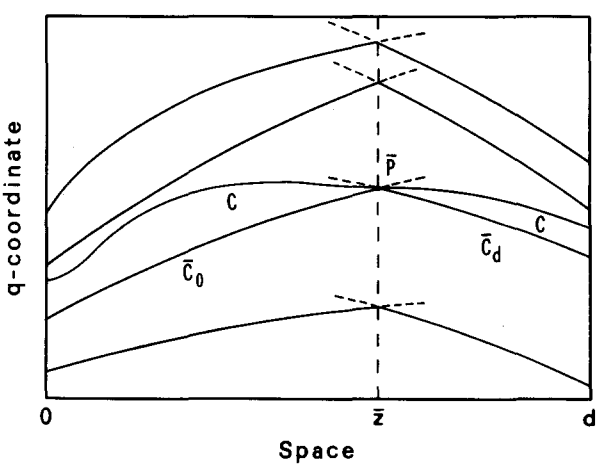


FIG. 3. Only two curves of the families of extremals K_0 and K_d intersect at the generic point $\bar{P}$ on the hyperplane $z = \bar{z}$; C is an arbitrary curve between $z = 0$ and $z = d$ passing through $\bar{P}$.

the Hamilton–Jacobi equation are found, the associated congruences K_0 and K_d can be determined by integrating Eqs. (17).

The curves of the congruences K_0 and K_d provide a partial solution to the problem of finding a minima to the free-energy functional F. In fact, let $\overline{P}(Q,\bar{z})$ be a fixed point on the hyperplane $z=\bar{z}$ and let $\overline{C}_0$ and $\overline{C}_d$ be the two members of K_0 and K_d meeting at $\overline{P}$. Let $\overline{C}=\overline{C}_0\oplus\overline{C}_d$ be the curve obtained by joining $\overline{C}_0$ and $\overline{C}_d$. Then, if C is an arbitrary curve between $z=0$ and $z=d$ passing through the fixed point $\overline{P}$, we have, by construction, the fundamental inequality

$$F[C]\geqslant F[\overline{C}]=S_0(Q,\bar{z})-S_d(Q,\bar{z})\equiv G(Q) \tag{20}$$

so that $\overline{C}=\overline{C}_0\oplus\overline{C}_d$ affords a minimum value to F with respect all varied curves between $z=0$ and $z=d$, passing through the fixed point $\overline{P}$. From this we can infer that each of the curves $\overline{C}_0$ belonging to K_0 and $\overline{C}_d$ belonging to K_d is an extremal curve of the functional F; i.e., they obey the Euler–Lagrange equations (9). This could also be proved by a direct calculation based on the Hamilton–Jacobi equation (Appendix B).

We are now in a position to prove the main properties (i)–(iv) of the GLF $G(Q)$ defined by Eq. (14) [see also Eq. (20)]. Property (i) is evident by construction. In order to prove property (ii), let us denote with $Q^i=Q^{i\star}$ an extremum point of $G(Q)$. Then, $\partial G/\partial Q_i=0$ at $Q^i=Q^{i\star}$ and, from definition (16) and Eqs. (19), we find that the canonical momenta $p_i=\partial L/\partial q'^i$ along the two extremal arcs $\overline{C}_0$ and $\overline{C}_d$ coincide at the intersection point $P^\star=(Q^\star,\bar{z})$ on the hyperplane $z=\bar{z}$. In other words, for $Q^i=Q^{i\star}$, the curve $\overline{C}=\overline{C}_0\oplus\overline{C}_d$ reduces to a curve $C^\star$, connecting the hyperplanes $z=0$ and $z=d$ and satisfying the Erdmann–Weierstrass condition that $\partial L/\partial q'^i$ is continuous at the corner $P^\star=(Q^\star,\bar{z})$. As is well known from the Hamilton–Jacobi theory of variational calculus, this implies that the whole curve $C^\star$ is an extremal curve for the free-energy functional F.[11] Since, by construction, $C^\star$ also satisfies the boundary conditions (8) at $z=0$ and $z=d$, it represents a possible steady state of the system. Conversely, in order that a curve $\overline{C}=\overline{C}_0\oplus\overline{C}_d$, made of two extremal arcs joining at $\overline{P}(Q,\bar{z})$, may be an extremal curve for the functional F, the Erdmann–Weierstrass condition must be met at the corner $\overline{P}$. Then, Eqs. (16) and (19) imply $\partial G/\partial Q^i=0$ at $\overline{P}$, so that Q^i is an extremum point of the GLF $G(Q)$. We have therefore proved the property (ii) of the NLGLF $G(Q)$.

It should now be almost obvious that extremal points of $G(Q)$ other than minima must not correspond to minima of the free-energy functional F. In fact, if $Q^{i\star}$ is an extremal but not a minimum point of $G(Q)$, some curve $\overline{C}=\overline{C}_0\oplus\overline{C}_d$ [intersecting the hyperplane $z=\bar{z}$ at some point $\overline{P}$ other than $P^\star(Q^\star,\bar{z})$] will exist for which $F[\overline{C}]=G(Q)<G(Q^\star)=F[C^\star]$. The steady-state curve $C^\star$ thus cannot afford a minimum to the total free-energy F. We conclude that only minima of the GLF $G(Q)$ may correspond to curves $C^\star$, minimizing the free-energy functional, which proves property (iii) of our NLGLF. Finally, property (iv) is easily deduced from Eq. (20) with $\overline{C}$ replaced by $C^\star$.

V. CONSTRUCTION OF THE GINZBURG–LANDAU FUNCTION

The GLF $G(Q)$ defined in Eqs. (14) and (20) can be constructed by observing that, along each curve of the previously defined congruences K_0 and K_d, we have equality in Eqs. (11) [and, hence, also in Eqs. (12a) and (12b)], once the arbitrary curve C is substituted by a curve $\overline{C}_0$ belonging to K_0 or by a curve $\overline{C}_d$ belonging to K_d, respectively. Then, integration of Eq. (12a) yields

$$S_0(Q,\bar{z})=f_0(q_0(Q,\bar{z}))+\int_{(\overline{C}_0)}^{\bar{z}}L\,dz, \tag{21a}$$

where $\overline{C}_0$ is the member of K_0 starting at the point $P_0=(q_0,0)$ at $z=0$. In Eq. (21a), Q^i denote the coordinates of the point $\overline{P}=(Q,\bar{z})$, where $\overline{C}_0$ intersects the fixed hyperplane $z=\bar{z}$. In deriving Eq. (21a), the first of boundary conditions (8) has been used. Notice that in Eq. (21a), all quantities and, in particular, the initial coordinates $q_0{}^i$ of the curve $\overline{C}$ must be expressed as functions of the coordinates Q^i at $z=\bar{z}$.

Similarly, integration of Eq. (12b) yields

$$S_d(Q,\bar{z})=f_d(q_d(Q,\bar{z}))-\int_{(\overline{C}_d)}^{d}L\,dz, \tag{21b}$$

where $\overline{C}_d$ is the member of K_d passing through the point $P_d(q_d,d)$ on the hyperplane $z=d$. All quantities and, in particular, the coordinates q_d^i must be expressed as functions of the coordinates Q^i of the point $\overline{P}$, where $\overline{C}_d$ intersect the hyperplane $z=\bar{z}$. Once the congruences K_0 and K_d are given, the functions $S_0(Q,\bar{z})$ and $S_d(Q,\bar{z})$ can be evaluated from Eqs. (21a) and (21b). As shown in these equations, the calculation involves the computation of the integral of $L(q,q',z)dz$ along the members of K_0 and K_d, which can be easily carried out by standard numerical methods.

The construction of the congruences K_0 and K_d is just as easy. In fact, K_0 and K_d consist of curves that (a) obey the Euler–Lagrange equations (9) associated with $L(q,q',z)$ and (b) obey the boundary conditions (8) at $z=0$ and $z=d$, respectively. The congruence K_0 can therefore be obtained by integrating Hamilton's canonical equations

$$\begin{aligned} p_i' &= -\frac{\partial H(p,q)}{\partial q^i}, \\ q'^i &= \frac{\partial H(p,q)}{\partial p_i}, \end{aligned} \tag{22}$$

with initial conditions $q^i(0)=q_0^i$ and $p_i(0)=\partial f_0/\partial q^i(q_0)$. For any given value of the initial coordinates q_0^i, we have an integral curve $q^i=q^i(z;q_0)$ of Eqs. (22), i.e., a member $\overline{C}_0$ of K_0, the q_0's being the N parameters of the congruence. As shown in Eq. (21a), in order to evaluate the function $S_0(Q,\bar{z})$, we must take as parameters of K_0 the values $Q^i=q^i(\bar{z};q_0)$ assumed by $\overline{C}_0$ at $z=\bar{z}$. The switching to the new parameters Q^i is possible only if the Jacobian determinant

$$J_0(z)=\frac{\partial(q^1,q^2,\ldots,q^N)}{\partial(q_0^1,q_0^2,\ldots,q_0^N)} \tag{23}$$

does not vanish for $0\leqslant z\leqslant\bar{z}$, i.e., if the congruence K_0 given by $q^i=q^i(z;q_0)$ covers the region G_0 of R_{n+1} simply. This con-

dition is fundamental in our construction and, as it will be shown later, its failure is fatal for the stability of the system.

The congruence K_d can be constructed in an analogous way: The only difference is that Hamilton's equations must be integrated backward and the relevant Jacobian $J_d(z)$ is between the coordinates Q^i at $z=\bar{z}$ and the coordinates q_d^i at $z=d$. [For the particular case of fixed boundary conditions, the appropriate Jacobian is given by Eq. (25) below.]

It is worth noting that the construction of the integral curves of Hamilton's equations belonging to K_0 and K_d involves only conditions at one point ($z=0$ or $z=d$) at once and, therefore, it can be carried out with simple and fast numerical techniques.

The hyperplane $z=\bar{z}$ can be chosen arbitrarily, and we can exploit this fact to improve the accuracy of the numerical routine. Although the optimal choice of $\bar{z}$ may depend strongly on the actual problem under study, a rule of thumb may be to take $\bar{z}$ so that the product of the Jacobians $J_0(\bar{z})J_d(\bar{z})$, evaluated at $z=\bar{z}$, is as large as possible (in absolute value). If the Jacobians are large, in fact, the integration of Hamilton's equations is less stiff and the accuracy of the inversion routine needed to pass from the initial (or final) coordinates to the Q^i is improved.

For spatially centrosymmetric systems, the calculations may be simplified. In this case, in fact, symmetry requires $f_0(q) = -f_d(q)$ as well as the invariance of the free-energy density $L(q,q',z)$ under the change $z \to d-z$. Then, if we choose $\bar{z}=\frac{1}{2}d$ and perform the variable change $z \to d-z$ in the integral in Eq. (21b), we obtain $S_d(Q,\frac{1}{2}d) = -S_0(Q,\frac{1}{2}d)$ and, hence, $G(Q) = 2S_0(Q,\frac{1}{2}d)$. We see, therefore, that, in order to construct the GLF for symmetric systems, only the congruence K_0 is needed. Moreover, the choice $\bar{z}=\frac{1}{2}d$ automatically maximizes the product J_0J_d. We used this simplification in working out the example of Sec. VIII.

VI. THE CASE OF FIXED BOUNDARY CONDITIONS

Our NLGLF was derived for systems obeying the time-independent boundary conditions (8). Although these conditions are quite general and are met by many physicochemical systems, an important class of systems is apparently excluded: the systems having fixed internal coordinates $q^i(0)=q_0^i$, $q^i(d)=q_d^i$ at boundaries $z=0$ and $z=d$. In the case of chemical reactions, this corresponds to a fixed concentration of reagents at the reservoir boundaries; in the case of liquid crystals, this corresponds to strong anchoring at the walls. The case of fixed boundary conditions corresponds to infinitely large surface potentials f_0 and f_d and, therefore, it must be handled with care. Since the values of q^i at boundaries are now fixed, the relevant free-energy functional is only $\int_0^d L\,dz$, the surface contribution (although infinite) being a constant. The congruences K_0 and K_d must be formed by extremal curves satisfying the boundary conditions $q^i(0)=q_0^i$ and $q^i(d)=q_d^i$, respectively, with arbitrarily given q_0's and q_d's. These congruences can be obtained by integrating Hamilton's Eq. (22), with the above boundary conditions for the coordinates, by setting $p_i(0)=p_{i0}$ and $p_i(d)=p_{id}$, in the two cases. The $2N$ momenta p_0 and p_d now play the role of parameters of the congruences K_0 and K_d, respectively. The construction of the NLGLF $G(Q)$ now proceeds as in the previous section, with Eqs. (21a) and (21b) replaced by

$$S_0(Q,\bar{z}) = \int_{(\bar{C}_0)}^{\bar{z}} L\,dz, \tag{24a}$$

$$S_d(Q,\bar{z}) = -\int_{(\bar{C}_d)}^{d} L\,dz, \tag{24b}$$

and the Jacobian determinant replaced by

$$\tilde{J}_0(z) = \frac{\partial(q^1,q^2,\ldots,q^N)}{\partial(p_{10},p_{20},\ldots,P_{N0})}. \tag{25}$$

We notice that, in the present case, the functions $S_0(Q,\bar{z})$ and $S_d(Q,\bar{z})$ can be expressed in terms of the well-known Hamilton's *principal function* $W(q,z|q_0,z_0)$[12] as $S_0(Q,\bar{z}) = W(Q,\bar{z}|q_0,0)$ and $S_d(Q,\bar{z}) = -W(q_d,d|Q,\bar{z})$.

VII. THE CASE OF MULTIVALUED GINZBURG–LANDAU FUNCTION

In deriving the NLGLF $G(Q)$, we have supposed that the Jacobian determinants (23) and (24) [or (25)] do not vanish. If this conditions is not fulfilled, the curves of the congruences K_0 and K_d do not cover the regions G_0 and G_d of R_{n+1} simply; i.e., they have an envelope (caustic) or a focus in G_0 and in G_d, respectively. In this case, the functions $S_0(Q,\bar{z})$ and $S_d(Q,\bar{z})$ and, hence, $G(Q)$ are multivalued. This status of affairs is depicted in Fig. 4 for the example of Sec. VIII. It is worth nothing that, even in the pathological case of the presence of caustics, the construction of our NLGLF can be carried out without modification. The only difference will be that the final $G(Q)$ will turn out to be multivalued. Now, it can be shown that Q values corresponding to a multivalued $G(Q)$ cannot correspond to stable steady states of the system, even if they afford a local minimum to $G(Q)$. This happens, for instance, for the *unstable* state $Q=0$ in Fig. 5. The demonstration that a multivalued

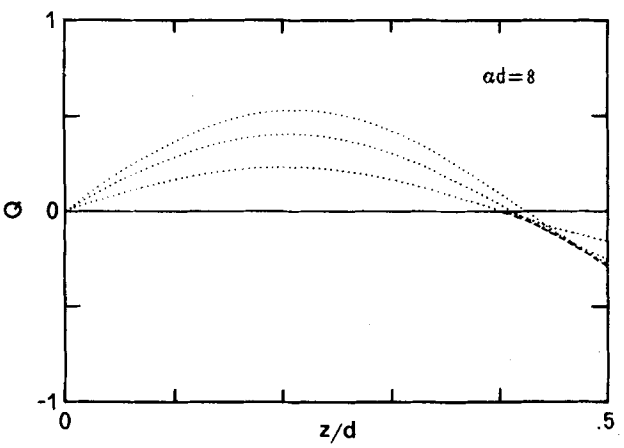


FIG. 4. The family of extremals K_0 for the example in Sec. VIII, for $\alpha d = 8$. The family has an *envelope* (caustic) in the region G_0, corresponding to $0 \leqslant z \leqslant \bar{z} = \frac{1}{2}d$. In this case, as shown in Fig. 5, the NLGLF $G(Q;\alpha)$ is multivalued.

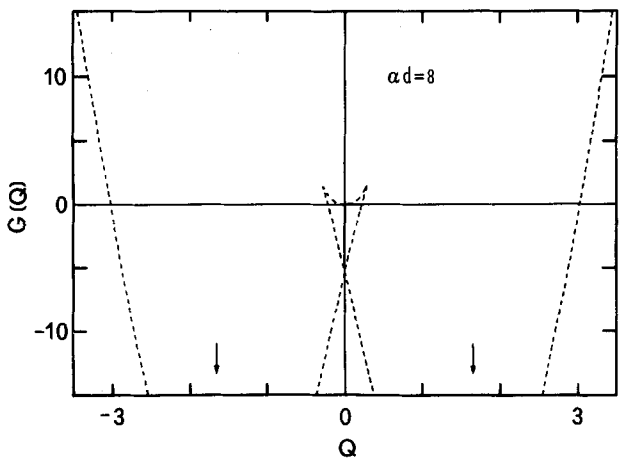


FIG. 5. The NLGLF $G(Q;\alpha)$ of the example in Sec. 8 for $\alpha d = 8$ ($> 2\pi$). The state $Q = 0$ is unstable because G is multivalued at $Q = 0$. The only stable states are those indicated by the arrows.

$G(Q)$ is incompatible with the stability of the system is long and requires a detailed analysis of Jacobi's equations for the second variation of the free-energy functional F. For this reason, it will be reported in Appendix A.

VIII. A GINZBURG–LANDAU FUNCTION FOR THE "PARABOLIC" SINE-GORDON EQUATION

As a nontrivial example that can be worked out analytically, we will construct explicitly, in this section, the nonlinear Ginzburg–Landau function associated to the parabolic version of the sine–Gordon equation:

$$-a\frac{\partial q}{\partial t}+\frac{\partial^2 q}{\partial z^2}+\alpha^2 \sin q \cos q = 0. \tag{26}$$

This equation has been taken from the physics of liquid crystals.[13] It describes the molecular reorientation induced by a constant magnetic field in a nematic liquid crystal in the twist geometry. The coordinate q is the director twist angle, the quantity a is Leslie's viscosity coefficient, and finally the parameter α is proportional to the intensity of the externally applied magnetic field.

For the sake of simplicity, we assume the time-independent boundary conditions[14]

$$q_i(0,t) = q_i(d,t) = 0, \tag{27}$$

d being the sample thickness along the z axis.

The steady-state solutions of Eq. (26) obey the pendulum equation

$$q'' + \alpha^2 \sin q \cos q = 0, \tag{28}$$

with $q'' = d^2q/dz^2$ and $0 \leqslant q \leqslant \frac{1}{2}\pi$. The trivial solution $q(z) \equiv 0$ obeys boundary conditions (27) and corresponds to equilibrium.

The usual way to investigate the stability of the equilibrium state is to write down the solution of Eq. (26) [with boundary conditions (27)] as a Fourier sum $q(z,t) = \Sigma_n c_n(t) \sin(n\pi z/d)$ and then to substitute into Eq. (26). At this point, a small perturbation is assumed [$|q(z,t)| \ll 1$] and the last term on the left in Eq. (26) is expanded in a power series of q. Usually, only linear terms in q are retained in Eq. (3) and only the fundamental Fourier component of $q(z,t)$ is considered.[15] Although this approximated approach yields the right value for the critical value α_c of the external parameter α where the state $q = 0$ becomes unstable, it cannot be directly extended either to other (spatially dependent) steady states or to the cases where hard (i.e., first-order) transitions may occur.

In order to construct the NLGLF in the present case, we first notice that the steady-state equation (28) [obtained by zeroing the time derivative in Eq. (26)] is the Euler–Lagrange equation associated to the action integral

$$F = \int_0^d \frac{1}{2}(q'^2 - \alpha^2 \sin^2 q)dz. \tag{29}$$

Then, our NLGLF can be constructed in the following steps.

Step (1): Hamilton's equations associated with the free-energy integral (29), viz.,

$$q' = \frac{\partial H}{\partial p} = p,$$

$$p' = -\frac{\partial H}{\partial q} = -\alpha^2 \sin q \cos q,$$

$$H = \tfrac{1}{2}p^2 + \alpha^2 \sin^2 q,$$

are solved with initial conditions $q(0) = 0$ [see Eqs. (27)] and $p(0) = p_0$, obtaining

$$\sin[q(z;m)] = \begin{cases} \pm m^{1/2} sn(\alpha z|m) & (0 \leqslant m \leqslant 1), \\ \pm sn(m^{1/2}\alpha z|m^{-1}) & (m > 1), \end{cases} \tag{30}$$

where $sn(u|m)$ is the Jacobi elliptic sine function with parameter m.[16] This solution depends parametrically on p_0, since m is related to p_0 by $m = (p_0/\alpha)^2$. The upper and lower signs in Eq. (30) correspond to positive and negative values of p_0, respectively.

Equation (30) represents a one-parameter family of curves starting at the point $q = 0, z = 0$ in the (q,z) plane, m (or p_0) being the parameter of the family. Each curve of the family is an extremal curve for the functional F, since it satisfies the Euler–Lagrange equation (28). We notice, however, that the members of the family (30) *do not* correspond to steady states of the system, since they fail to meet the condition $q(d;m) = 0$ at the boundary $z = d$. In the most general case, Hamilton's equations cannot be solved analytically and numerical integration is made necessary. We notice, however, that we are always faced with a set of first-order total differential equations with *one-end-point* initial data, that can be solved with standard Runge–Kutta integration routines.

For further convenience, we denote with Q the value assumed by $q(z;m)$ at the point $z = \frac{1}{2}d$. Q is given by

$$Q = \begin{cases} \pm \sin^{-1}[m^{1/2} sn(\frac{1}{2}\alpha d|m)] & (0 \leqslant m \leqslant 1), \\ \pm \sin^{-1}[sn(m^{1/2}\frac{1}{2}\alpha d|m^{-1})] & (m > 1). \end{cases} \tag{31}$$

As explained in Sec. VII, the choice $z = \frac{1}{2}d$ is due to the symmetry of Eq. (28) and of the boundary conditions (27) with respect to the change $z \to d - z$. We notice that Eq. (31) defines a correspondence between the values of Q and the

curves of the family (30). In the most general case, this correspondence is not one-to-one, since two or more curves of the family (30) may be associated to the same Q value (see Sec. VIII).

Step (2): The free-energy integral (29) is then evaluated along each member of the family (30) from $z=0$ to $z=\frac{1}{2}d$. Again, the choice $z=\frac{1}{2}d$ for the integration upper limit is due to the particular symmetry of the problem. The result is an ordinary function $F(m;\alpha)$ of the parameter m of the family (30) and of the external parameter α. In the general case, the integration of the free energy must be performed numerically. In the present example, however, the integration can be carried out analytically, yielding

$$F(m;\alpha)=\begin{cases}2\alpha\left[E(\frac{1}{2}\alpha d\,|m)-\frac{1}{4}\alpha d(2-m)\right] & (0\leqslant m\leqslant 1),\\ 2\alpha\left[m^{1/2}E(\frac{1}{2}m^{1/2}\alpha d\,|m^{-1})-\frac{1}{4}\alpha d m\right] & (m>1),\end{cases} \tag{32}$$

where $E(u|m)$ is the incomplete elliptic integral of the second kind.

Step (3): The parameter m is eliminated between Eqs. (31) and (32), obtaining F as a function of Q and α: $F=F(Q;\alpha)$. Only positive values of Q can be retained in Eq. (31) since, in the present case, $F(Q;\alpha)$ is symmetric: $F(Q;\alpha)=F(-Q;\alpha)$. Then, the function $G(Q;\alpha)=2F(Q;\alpha)$ is the NLGLF of our problem.

The choice of the factor 2 in front of F guarantees that the numerical values $G=G_e$, corresponding to extremal points of $G(Q;\alpha)$, coincide with the values of the free-energy functional F, given by Eq. (29), evaluated along the corresponding steady states of the system. In this way, $G(Q;\alpha)$ fulfills also property (d) above.

The NLGLF $G(Q;\alpha)$ is reported in Fig. 2 for different values of α. For $\alpha d<\pi$, $G(Q;\alpha)$ has only one minimum at $Q=0$. This minimum corresponds to the equilibrium state $q\equiv 0$, which is therefore stable. For $\pi<\alpha d<2\pi$, the equilibrium state becomes unstable (local maximum of G) and two new stable states (denoted with the arrows) appear. These states are located symmetrically with respect to the equilibrium state $Q=0$. The steady-state solutions of Eq. (26) corresponding to these new stable states are found immediately from the previously mentioned correspondence between Q values and curves of the family (30). One can easily verify that these solutions satisfy also to the second of boundary conditions (27). When α reaches the critical value $\alpha_c=\pi/d$, the stable and unstable states coalesce at $Q=0$. The corresponding critical G curve is also reported in Fig. 2. This critical curve has a minimum at $Q=0$, indicating that the critical state is itself stable. The transition is henceforth soft (second order).

In Fig. 5 is reported the NLGLF $G(Q;\alpha)$ for a value of αd larger than 2π. We notice that now the function $G(Q;\alpha)$ is multivalued. As will be proved in Appendix A, a *multivalued* NLGLF always corresponds to *unstable* states. The state $Q=0$ is therefore unstable, despite the fact that it corresponds to a local minimum of G. The only stable states are the ones denoted by the arrows in Fig. 5.

We remark again that the NLGLF $G(Q;\alpha)$ has been constructed by retaining all nonlinear terms in Eq. (26) and, therefore, we may call it an *exact* NLGLF for the system. The curves of the family (30) corresponding to extremal points of $G(Q;\alpha)$ are, in fact, exact steady-state solutions of Eq. (26), satisfying all of boundary conditions (27).

IX. CONCLUSIONS

We have proved by direct construction that a Ginzburg–Landau function exists for a large class of physicochemical continuous systems, depending on one spatial coordinate and time. Our construction leads to an exact NLGLF, in the sense that all nonlinear terms in the equations of motion are retained, without approximations. A series expansion of $G(Q)$ around some reference state $\overline{Q}$ is, therefore, optional.

Although, in general, the evaluation of $G(Q)$ may require numerical integration, the integration problem is always well posed, so that standard routines can be exploited. Moreover, the construction method, presented in this paper, does not break out when caustics or other pathological behaviors are encountered in the fields of extremals used in the calculations.

Having an exact NLGLF provides a *global* stability criterion, which gives all stable states at once by simply looking at the minima of an ordinary function. Furthermore, if the NLGLF is known (or calculated) as a function of an external parameter, transitions between steady states can be studied in detail. The NLGLF, in fact, gives qualitative information about the character of the transition (soft or hard) as well as quantitative information about the critical points and the energy barriers between adjacent stable states.

ACKNOWLEDGMENTS

We acknowledge Professor G. Vilasi of the University of Salerno, Italy, for critical reading of the manuscript.

This work was supported by M.P.I. (Ministero della Pubblica Instruzione), Italy.

APPENDIX A

In this Appendix we show that the vanishing of the Jacobian (23) [or (25)] and, hence, the occurrence of a multivalued $G(Q)$, prevents the stability of the system, even if all other conditions [as Legendre's condition (2) or the state corresponding to a minimum of $G(Q)$] are satisfied.

We consider here only the Jacobian J_0 associated to the congruence K_0. A completely analogous proof can be done for J_d and the congruence K_d.

Since all curves $q^i = q^i(z;q_0)$ of K_0 are extremal curves of $L(q,q',z)$, the N^2 functions

$$\eta_j^i = \frac{\partial q^i}{\partial q_0^j}(z;q_0). \tag{A1}$$

obey the Jacobi equations

$$\frac{d}{dz}\left(\frac{\partial \Omega}{\partial \eta_j'^i}\right) - \frac{\partial \Omega}{\partial \eta_j^i} = 0, \tag{A2}$$

where

$$2\Omega(\eta,\eta',z) = \left(\frac{\partial^2 L}{\partial q'^i\,\partial q'^j}\right)\eta'^i\eta'^j + 2\left(\frac{\partial^2 L}{\partial q^i\,\partial q'^j}\right)\eta^i\eta'^j + \left(\frac{\partial^2 L}{\partial q^i\,\partial q^j}\right)\eta^i\eta^j. \tag{A3}$$

Jacobi's equations (A2) are linear and homogeneous in η^i, η'^i, and η''^i. They are, in fact, the linearized version of the Euler–Lagrange equations (9) around the extremal $\bar{C}_0$ of K_0, having parameters q_0^i. It is almost obvious that if η^i and η'^i are both zero at a point $z = z^*$, then also η''^i and all higher derivatives of η^i vanish at z^*, so that $\eta \equiv 0$ in a finite neighborhood of z^*.

Since all curves of K_0 obey the first of boundary conditions (8), differentiating this condition with respect to q_0^j yields

$$\left(\frac{\partial^2 f_0}{\partial q^i\,\partial q^j} + \frac{\partial^2 L}{\partial q'^i\,\partial q^j}\right)\eta_k^j + \left(\frac{\partial^2 L}{\partial q'^i\,\partial q'^j}\right)\eta_k'^j = 0, \tag{A4}$$

at $z = 0$. Moreover, differentiating the identities $q^i(0;q_0) = q_0^i$ yields

$$\eta_j^i = \delta_j^i \tag{A5}$$

at $z = 0$. The boundary conditions (A4) and (A5) at $z = 0$ are to be added to the Jacobi equations (A2). The functions $\eta_j^i(z)$ being linearly independent, any solution ξ^i of the Jacobi equations (A2), obeying the boundary conditions (A4), can be expressed as $\xi^i = \alpha^j\eta_j^i$, with nonzero coefficients α^j (summation over repeated indices is intended).

Now, let us suppose that the Jacobian J_0, given by Eq. (23), vanishes at some point z^* in the interval $[0,\bar{z}]$. Then, we can find numbers α^j that are not zero so that the N functions ξ^i, defined as

$$\xi^i(z) = \alpha^j\eta_j^i(z), \tag{A6}$$

are zero at $z = z^*$. In view of the linearity of Jacobi's equations (A2) as well as of the boundary conditions (A4) and (A5), $\xi^i(z)$ is also a solution of Eq. (A2), satisfying the boundary conditions (A4) and $\xi^i(0) = \alpha^i$. The vanishing of J_0 at $z = z^*$ implies, therefore, the existence of a solution of the Jacobi equation (A2), satisfying the boundary conditions (A4) and vanishing at $z = z^*$.

On the other hand, if it is known that a nonzero solution $\xi^i(z)$ of the Jacobi equation exists obeying the boundary conditions (A4) and vanishing at $z = z^*$, we may pose $\xi^i = \alpha^j\eta_j^i$ and this expression vanishes at $z = z^*$, implying $J_0 = 0$ at $z = z^*$.

At this point, we need only to prove that the existence of a solution $\xi^i(z)$ of Jacobi's equation, obeying the boundary conditions (A4) and vanishing at some point z^* in the interval $[0,\bar{z}]$, implies the instability of the systems.

Consider a steady state of the system, represented by the curve $\bar{C}$: $q^i = q^i(z)$ between $z = 0$ and $z = d$. Consider also the second variation of the total free-energy functional F around the reference curve $\bar{C}$:

$$\delta^2 F = \left[\left(\frac{\partial^2 f_0}{\partial q^i\,\partial q^j}\right)\eta^i\eta^j\right]_{z=0} - \left[\left(\frac{\partial^2 f_d}{\partial q^i\,\partial q^j}\right)\eta^i\eta^j\right]_{z=d} = \int_0^d 2\Omega(\eta,\eta',z)\,dz, \tag{A7}$$

where $\eta^i(z)$ denotes a small nonzero perturbation of the extremal $\bar{C}$. Then, a necessary condition for the curve $\bar{C}$ to correspond to a stable state is that $\bar{C}$ affords a minimum to F, which implies $\delta F = 0$ and also $\delta^2 F \geqslant 0$.

A partial integration in Eq. (A7) yields

$$\delta^2 F = \left\{\left[\left(\frac{\partial^2 f_0}{\partial q^i\,\partial q^j} - \frac{\partial^2 L}{\partial q'^i\,\partial q^j}\right)\eta^j - \left(\frac{\partial^2 L}{\partial q'^i\,\partial q'^j}\right)\eta'^j\right]\eta^i\right\}_{z=0} - \left\{\left[\left(\frac{\partial^2 f_d}{\partial q^i\,\partial q^j} - \frac{\partial^2 L}{\partial q'^i\,\partial q^j}\right)\eta^j - \left(\frac{\partial^2 L}{\partial q'^i\,\partial q'^j}\right)\eta'^j\right]\eta^i\right\}_{z=d} - \int_0^d \left[\frac{d}{dz}\left(\frac{\partial \Omega}{\partial \eta'^i}\right) - \frac{\partial \Omega}{\partial \eta^i}\right]\eta^i\,dz. \tag{A8}$$

This equation shows that the Jacobi equations (A2) are also the Euler–Lagrange equations for the second variation of F. They provide, therefore, a necessary condition to η^i for affording a minimum to $\delta^2 F$.

At this point, following Bliss,[17] we take as perturbation η^i

$$\eta^i(z) = \begin{cases} \xi^i(z), & \text{for } 0 \leqslant z \leqslant z^*, \\ 0, & \text{for } z \geqslant z^*. \end{cases} \tag{A9}$$

We notice that η^i, given by Eq. (A9), is made of two solutions of Jacobi's equations connected at the corner point $z = z^*$. Moreover, by construction, η^i (as ξ^i) satisfies the boundary condition (A4).

Inserting Eq. (A9) into Eq. (A8), we find $\delta^2 F = 0$ along η^i. The last term in Eq. (A8), in fact, vanishes because η^i is made up of solutions of Jacobi's equations, the second term vanishes since η^i and η'^i are zero at $z = d > z^*$, and, finally, the first term is zero by virtue of boundary conditions (A4).

Now, let us suppose that the reference extremal $\bar{C}$ affords a minimum to the free-energy functional F. Then, as a matter of necessity, $\delta^2 F \geqslant 0$ so that the curve η^i, given by Eq. (A9), should correspond to a minimum of $\delta^2 F$. But, if this were the case, η^i should also satisfy the Erdmann–Weierstrass conditions at the corner $z = z^*$; i.e., we should have $\partial\Omega/\partial\eta'^i(\eta,\eta',z)$ continuous at $z = z^*$. From Eq. (A2) we get

$$\frac{\partial \Omega}{\partial \eta'^i} = \left(\frac{\partial^2 L}{\partial q'^i\,\partial q'^j}\right)\eta'^j + \left(\frac{\partial^2 L}{\partial q'^i\,\partial q^j}\right)\eta^j. \tag{A10}$$

Then, since η^i vanishes at $z = z^*$, we find

$$\left(\frac{\partial^2 L}{\partial q'^i \partial q'^j}\right)\eta'^j = 0 \tag{A11}$$

at $z = z^*$ and, by virtue of Legendre's conditions (2), also, $\eta'^i = 0$. But the vanishing of both η^i and η'^i at the same point $z = z^*$ implies $\eta^i \equiv 0$ in a neighborhood of z^*, contrary to what was stated in Eq. (A9). The occurrence of this contradiction proves that our starting hypothesis, namely, that the reference extremal $\bar{C}$ affords a minimum to F, cannot be true. The steady state corresponding to $\bar{C}$ is therefore unstable. Since the construction of the curve (A9) depends crucially on the existence of a solution ξ^i of the Jacobi equations obeying conditions (A4), which, in turn, is obtained under the condition of the vanishing of the Jacobian determinant J_0, we have proved our theorem.

APPENDIX B

Let $\bar{C}$ be a curve belonging either to K_0 or K_d. Then, along $\bar{C}$, the canonical relation (16) holds, with $S(q,z)$ obeying the Hamilton–Jacobi equation (19). Taking the z derivative of Eq. (16) along $\bar{C}$, we obtain

$$\frac{d}{dz}\left(\frac{\partial L}{\partial q'^i}\right) = \frac{\partial^2 S}{\partial q^i \partial z} + q'^j \frac{\partial^2 S}{\partial q^j \partial q^i}. \tag{B1}$$

Insertion of the Hamilton–Jacobi equation (19) into Eq. (B1) yields

$$\frac{d}{dz}\left(\frac{\partial L}{\partial q'^i}\right) = -\frac{\partial H(\nabla S,q,z)}{\partial q^i} + q'^j \frac{\partial^2 S}{\partial q^j \partial q^i}. \tag{B2}$$

The q derivative of H in this equation is made by keeping z constant. We have, therefore, the chain rule

$$\frac{\partial H(\nabla S,q,z)}{\partial q^i} = \left(\frac{\partial H}{\partial q^i}\right)_{p,z} + \left(\frac{\partial H}{\partial p_j}\right)_{q,z}\left(\frac{\partial^2 S}{\partial q^j \partial q^i}\right). \tag{B3}$$

Now, as is well known, the definition itself of H, viz., $H = p_i q'^i - L$, implies the following identities (Ref. 10, p. 117):

$$\left(\frac{\partial H}{\partial p_i}\right)_{q,z} \equiv q'^i \tag{B4}$$

and

$$\left(\frac{\partial H}{\partial q^i}\right)_{p,z} \equiv -\left(\frac{\partial L}{\partial q^i}\right)_{q',z}. \tag{B5}$$

Then, insertion of Eq. (B4) into Eq. (B3) yields

$$\frac{\partial H(\nabla S,q,z)}{\partial q^i} = \left(\frac{\partial H}{\partial q^i}\right)_{p,z} + q'^j\left(\frac{\partial^2 S}{\partial q^j \partial q^i}\right) \tag{B6}$$

and insertion of Eq. (B6) into Eq. (B2) yields

$$\frac{d}{dz}\left(\frac{\partial L}{\partial q'^i}\right) = -\left(\frac{\partial H}{\partial q^i}\right)_{p,z} \tag{B7}$$

or, in virtue of the identity (B5),

$$\frac{d}{dz}\left(\frac{\partial L}{\partial q'^i}\right) - \left(\frac{\partial L}{\partial q^i}\right)_{q',z} = 0. \tag{B8}$$

Relation (B8) explicitly shows that along $\bar{C}$ the Euler–Lagrange equations are fulfilled.

[1] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics* (Pergamon, Oxford, 1959).

[2] See, for instance, the well-known texts on nonequilibrium thermodynamics: S. R. de Groot and P. Mazur, *Nonequilibrium Thermodynamics* (North-Holland, Amsterdam, 1962); D. D. Fitts, *Nonequilibrium Thermodynamics* (McGraw-Hill, New York, 1962); C. Truesdell, *Rational Thermodynamics* (McGraw-Hill, New York, 1969). For more specific reference to symmetry-breaking instabilities, see P. Glansdorff and I. Prigogine, *Thermodynamic Theory of Structure, Stability and Fluctuations* (Wiley, London, 1971) and references therein. See also the book by B. H. Lavenda, *Thermodynamics of Irreversible Processes* (McMillan, London, 1978).

[3] For a review, see E. Abraham and S. D. Smith, Rep. Prog. Phys. **45**, 815 (1982); L. A. Lugiato, "Theory of Optical Bistability" in *Progress in Optics*, edited by E. Wolf (North-Holland, Amsterdam, 1984), Vol. XXI.

[4] For a review, see Y. R. Shen, Philos. Trans. R. Soc. London Ser. A **313**, 327 (1984); N. V. Tabiryan, A. V. Sukhov, and B. Ya. Zel'dovich, Mol. Cryst. Liq. Cryst. **136**, 1 (1986).

[5] See, for instance, R. Aris, Chem. Eng. Sci. **24**, 149 (1969).

[6] T. K. Fowler, J. Math. Phys. **4**, 559 (1963). See, also, J. Wei, Chem. Eng. Sci. **20**, 729 (1965) and the survey by M. Fjeld in "On the Stability of Distributed Parameter Systems," Institutt for Reguleringsteknikk, Norges Tekniske Høgskole, Trondheim (1967).

[7] A transition between stationary states is *first order* if it occurs in a discontinuous way as the external control parameters are slowly changed. The transition is *second order* if the discontinuity is restricted to the first- (and higher-) order derivatives of the transition parameter.

[8] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon, Oxford, 1958), Sec. 138.

[9] In the case of liquid crystals, the surface potentials at the walls are assumed to be of the form $f(q) = \frac{1}{2}\Sigma_n A_{2n} \sin^{2n} q$, where A_{2n} represents the anchoring strength [P. Sheng, Phys. Rev. Lett. **37**, 1059 (1976); Phys. Rev. A **26**, 1610 (1982)].

[10] C. Carathéodory, "Variationsrechnung und partielle Differentialgleichungen erster Ordnung," Teubner, Leipzig and Berlin, 1935); Zentralbl. Math. Grenzgebiete **11**, 356 (1935). This method is often referred to as "der Königsweg" (the royal road) of Carathéodory.

[11] H. Rund, *The Hamilton–Jacobi Theory in the Calculus of Variations* (Krieger, Huntington, NY, 1973), pp. 43, 44. The corner condition used in the text is the first of the Erdmann–Weierstrass conditions. The second of the Erdmann–Weierstrass conditions is automatically satisfied in our case, since we assumed the Hamiltonian $H(p,q,z)$ continuous in all its arguments.

[12] W. R. Hamilton, *Mathematical Papers, Vol. I. Geometrical Optics* (Cambridge U.P., Cambridge 1931); Zentralbl. Math. Grenzgebiete **2**, 85 (1931).

[13] S. I. Ben-Abraham, Phys. Rev. A **14**, 1251 (1976).

[14] In the case of pure twist distortion in planarly aligned nematic liquid crystal films, these boundary conditions correspond to strong anchoring at the sample walls.

[15] See, for instance, S. Chandrasekhar, *Liquid Crystals* (Cambridge U.P., Cambridge, 1977), p. 150. A different but equivalent method is to insert the Fourier expansion for $q(z)$ in the free-energy functional and retain only terms up to q^2 [P. G. De Gennes, *The Physics of Liquid Crystals* (Clarendon, Oxford, 1974), pp. 90, 91.

[16] We use the notation of the *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1965).

[17] G. A. Bliss, Trans. Am. Math. Soc. **17**, 195 (1916).

Treatment of higher-order Lagrangians via the construction of dynamically equivalent first-order Lagrangians

Piotr W. Hebda
Department of Physics and Astronomy, University of Georgia, Athens, Georgia 30602

(Received 22 December 1989; accepted for publication 2 May 1990)

For a given, in general, singular Lagrangian containing higher-order time derivatives, a dynamically equivalent Lagrangian with only first-order time derivatives is constructed. A Hamiltonian structure for this first-order Lagrangian is then found with the use of the Dirac theory of constraints. It is shown that in the case of a nonsingular higher-order Lagrangian, the Ostrogradsky dynamics is derived in this way. Further, it is shown that ambiguities characteristic of higher-order Lagrangian systems do not appear when using this construction. In particular, it is shown that the addition of a total time derivative term to the higher-order Lagrangian can only induce a time-independent canonical transformation, even in the case of a singular Lagrangian.

I. INTRODUCTION

A Hamiltonian formalism for Lagrangians with higher-order time derivatives was first given by Ostrogradsky[1] and later independently by Borneas.[2] Since then, many general aspects of the formalism have been studied, for example, the extension of Poisson brackets,[3] Noether's theorem,[4] and Hamilton–Jacobi theory[5] to systems with higher-order Lagrangians. Recently, a dynamical formalism for systems with singular[6] higher-order Lagrangians has been proposed.[7] Some serious ambiguities have been shown to arise due to the freedom one has in choosing the higher-order Lagrangian for a given physical system. This results in obtaining different quantum results for systems that are classically identical.[8]

The approach we present here is completely different. For a given Nth-order Lagrangian (N refers to the highest time derivative in the Lagrangian), we construct a dynamically equivalent first-order Lagrangian. This construction can always be realized. Starting with this first-order Lagrangian, a Hamiltonian formalism is derived. In this derivation, a modified version of Dirac's theory[9] for construction of Hamiltonians for constrained systems is used. This is necessary because our first-order Lagrangians are always singular. However, the extra constraints, which are an unavoidable consequence of the construction, turn out to produce the Poisson brackets used in the Ostrogradsky formalism.[3]

This paper is a part of a larger project,[10] which aims to present a consistent approach to systems with holonomic and nonholonomic constraints, by treating their Lagrange multipliers as independent variables and then reducing them out together with their canonical momenta by the proper use of the Dirac brackets.[9] Here we present only that part of the picture sufficient for dealing with Lagrangians of a higher order.

In Sec. II, we present the dynamics of singular first-order Lagrangian systems in the form it will be used in this paper. The method we use to derive the constraints of the system is slightly different from the one used by Dirac[9] and, to the best of our knowledge, original. This way of obtaining constraints is crucial for the results of Sec. V, where we deal with the question of the uniqueness of our formalism. In order to make the discussion self contained, we also define Dirac brackets[9] at the end of this section.

In Sec. III, we present the construction of the first-order Lagrangian for a given Nth-order one, and we show directly the equivalence of the Euler–Lagrange equations of motion for these two systems.

In Sec. IV, we study the Hamiltonian obtained from our first-order Lagrangian, using a variation (Sec. II) of the Dirac method[9] for constructing the dynamics of singular first-order Lagrangians (the original Dirac method could be used as well, giving the same results). It turns out that the well-known Ostrogradsky canonical momenta and the Hamiltonian are derived this way, and the Poisson brackets used in the Ostrogradsky formalism are the Dirac brackets[9] of our first-order theory.

In Sec. V, we study the uniqueness of the formalism. It is well known that if we decide to use higher-order Lagrangians, we face the problem of having different Lagrangians, often of a different order in time, which give equivalent classical equations of motion, but appear to produce different Hamiltonian systems, and consequently appear to produce different quantum results. For example the problem posted by Hayes and Jankowski,[8] in which the addition of complete time derivative term in the form $[-(m/2)(d/dt)(\dot{x}x)]$ to the usual Lagrangian of a harmonic oscillator, apparently produces inequivalent quantum results while leaving the classical solutions unchanged. Two ways of dealing with this ambiguity have been proposed. One by Hayes,[11] which used the equations of motion to change the form of the Hamiltonian (however, we still have no criteria to choose the Hamiltonian obtained this way over the one obtained in the first attempt, other than our previous knowledge of the properties of the harmonic oscillator). The other way was given by Ryan,[12] who proposed the method of reducing the Lagrangian to the lowest possible order. This method, although self-consistent still gives no answer to the question of why we have to use these reduced Lagrangians, why one Lagrangian appears to be "better" then the other.

 0022-2488/90/092116-10$03.00

In Sec. V, we give the solution to this problem, showing directly that despite which Nth-order Lagrangian we choose among the available ones, if we use our construction of the first-order Lagrangian and the Dirac method of dealing with constraints, the coordinates and momenta remaining in the Hamiltonian formalism, the form of the Hamiltonian itself and the dynamical brackets of the formalism (the Dirac's brackets) will remain unchanged, or will differ at most by a time-independent canonical transformation.

It should be pointed out here, that some results concerning the ambiguity problem were discussed in the recent paper by Saito *et al.*[7] In particular the Hayes–Jankowski problem could be solved within the scope of their paper. Please see Sec. VI for a short discussion of their important paper.

Also in Sec. VI, we briefly discuss advantages of our approach, and we list some other possible applications of the methods we use in this paper.

In this paper, we assume everywhere that we are dealing with so called singular N th-order Lagrangians.[6] In the case of nonsingular N th-order Lagrangian, we can still proceed in the same way (with possible simplification of some steps).

It may be worth mentioning that if our construction is used for a first-order Lagrangian, the Lagrangian remains unchanged.

II. PRELIMINARIES—CONSTRAINED DYNAMICS

The problem of constructing the dynamics of a system with a singular Lagrangian was first seriously studied by Dirac[9] and there now exists an extensive literature on the subject.[13,14] The method we present in this section, although equivalent to the original Dirac method, uses Euler–Lagrange equations rather than the Hamiltonian to obtain the secondary constraints. The Hamiltonian is defined at the end of the process. Because this way of obtaining constraints is necessary for the considerations of Sec. V, and a description of it in this form does not exist in the literature, we describe it in some detail.

The summation convention will be used throughout this section. The range of indices i and j will be from 1 to n.

Assume we have a system described by coordinates $q=(q_1,...,q_n)$ and a time-independent, first-order, singular Lagrangian:

$$L=L(\dot{q},q). \tag{2.1}$$

Let us define the vector q^X in the coordinate space by

$$q^X=(0,...,0,q_{X+1},q_{X+2},...,q_n),$$

where X can be $0,...,n$ (the upper indices will describe the vectors, while the lower ones the components). Note $q^0=q$, and we will often write q for q^0.

If we define the canonical momenta in the usual way as

$$p_i=\frac{\partial L}{\partial \dot{q}_i}, \tag{2.2a}$$

then the Euler–Lagrange equations of motion can be written as

$$\dot{p}_i=\frac{\partial L}{\partial q_i}. \tag{2.2b}$$

The first step in looking for the constraints of the system is to solve equations (2.2a) for the $\dot{q}$'s (the velocities). The singularity of the Lagrangian means that the Hessian matrix

$$W_{ij}=\frac{\partial^2 L}{\partial \dot{q}_i\,\partial \dot{q}_j}, \tag{2.3}$$

is singular, and has rank $R(1)$, $0\leqslant R(1)<n$. This implies that Eqs. (2.2a) can only be solved for $R(1)$ of the velocities and that there exist $[n-R(1)]$ independent relations among the q's and p's. These relations are called primary constraints and they are direct consequences of the definitions of momenta in the singular case. [All algebraic relations among q's and p's, which are consequences of Eqs. (2.2) are called constraints, those which are consequence of only equations (2.2a) are primary, all others are secondary.] Without loss of generality, we can assume that Eqs. (2.2a) can be solved for the first $R(1)$ of the velocities. Equations (2.2) can then be written as

$$\dot{p}_i=\frac{\partial L}{\partial q_i}(q,p,\dot{q}^{R(1)}), \tag{2.4a}$$

$$\dot{q}_\alpha=f^1_\alpha(q,p,\dot{q}^{R(1)}),\quad 1\leqslant\alpha\leqslant R(1), \tag{2.4b}$$

$$\varphi_a(q,p)=p_a-\frac{\partial L}{\partial \dot{q}_a}(q,p)=0,\quad R(1)<a\leqslant n, \tag{2.4c}$$

where φ_a are primary constraints. The consequence of the existence of the constraints (2.4c) is that their time derivatives have to be equal to zero. This condition along with the use of Eqs. (2.4a) and (2.4b) gives $[n-R(1)]$ equations in the form

$$\frac{d\varphi_a}{dt}(q,p,\dot{q}^{R(1)})=0,\quad R(1)<a\leqslant n. \tag{2.5}$$

The second step is to solve equations (2.5) for the velocities. Similarly, as in the previous step, if the matrix

$$M^1_{aa'}=\frac{\partial(d\varphi_a/dt)}{\partial \dot{q}_{a'}},\quad R(1)<a,a'\leqslant n \tag{2.6}$$

is of rank $R(2)-R(1)$, $R(1)\leqslant R(2)\leqslant n$, then equations (2.6) can be solved for $R(2)-R(1)$ velocities, and there may exist a number of independent relations among the q's and p's. Some of these relations can be dependent on the prior constraints. Those which are not, are secondary constraints. Equations (2.4) can be now written as

$$\dot{p}_i=\frac{\partial L}{\partial q_i}(q,p,\dot{q}^{R(2)}), \tag{2.7a}$$

$$\dot{q}_\alpha=f^2_\alpha(q,p,\dot{q}^{R(2)}),\quad 1\leqslant\alpha\leqslant R(2), \tag{2.7b}$$

$$\varphi_a(q,p)=0,\quad R(1)<a\leqslant n, \tag{2.7c}$$

$$\varphi_{b_1}(q,p)=0, \tag{2.7d}$$

where φ_a are the primary, and φ_{b_1} are the secondary constraints obtained in this step of the process.

Now, just as we did for primary constraints, we have to study the consequences of time derivatives of secondary constraints (2.7d) being zero. This process can give some new expressions for $\dot{q}$'s, as well as new constraints (all of them would be secondary). This has to be repeated as long as we obtain new constraints [the number of steps will be finite, because we cannot have more than $2n$ independent constraints in the phase space (q,p)]. After the final step, which we call k th, we will obtain

$$\dot{p}_i = \frac{\partial L}{\partial q_i}(q,p,\dot{q}^{R(k)}), \tag{2.8a}$$

$$\dot{q}_\alpha = f^k_\alpha(q,p,\dot{q}^{R(k)}), \quad 1 \leqslant \alpha \leqslant R(k), \tag{2.8b}$$

$$\varphi_a(q,p) = 0, \quad R(1) < a \leqslant n, \tag{2.8c}$$

$$\varphi_b(q,p) = 0, \tag{2.8d}$$

where φ_a denotes primary constraints, and φ_b denotes all secondary constraints obtained in the process. The nonzero components of $\dot{q}^{R(k)}$ (unsolved velocities) are free in the formalism [they can be set equal to arbitrary functions of time and still the constraints (2.8c), (2.8d) will be satisfied and Eqs. (2.8a), (2.8b) will have solutions]. The different solutions obtained for differently fixed $\dot{q}^{R(k)}$ time dependence have to be physically equivalent, which means that we deal with a theory with gauge (e.g., Ref. 14). The arbitrariness can be removed by imposing a gauge. This means we impose new constraints

$$G_c(q,p) = 0, \quad 1 \leqslant c \leqslant n - R(k), \tag{2.9}$$

which have the property that their consequence

$$\frac{dG_c}{dt}(q,p,\dot{q}^{R(k)}) = 0, \tag{2.10}$$

can be solved for all velocities in $\dot{q}^{R(k)}$ (for example $G_c = q_{R(k)+c} + \text{const}_c$). So, finally, we will obtain the following system of first-order equations which are equivalent (in the fixed gauge) to the Euler–Lagrange equations (2.2) and are solved for all $\dot{q}_i, \dot{p}_i$:

$$\dot{p}_i = \frac{\partial L}{\partial q_i}(q,p), \tag{2.11a}$$

$$\dot{q}_i = f_i(q,p), \tag{2.11b}$$

$$\varphi_a(q,p) = 0, \tag{2.11c}$$

$$\varphi_b(q,p) = 0, \tag{2.11d}$$

$$G_c(q,p) = 0. \tag{2.11e}$$

Now, we can define a Hamiltonian $H(q,p)$. In order to do this let us first define

$$H_0 = [p_i\dot{q}_i - L(\dot{q},q)]\,|_{\dot{q}_\alpha = f^1_\alpha \quad 0 < \alpha \leqslant R(1)}, \tag{2.12}$$

where the p_i's are defined by Eqs. (2.2b), and f^1_α is taken from Eqs. (2.4b). It can be shown (e.g., Ref. 13) that H_0 defined this way does not depend on $\dot{q}_a$, $R(1) < a \leqslant n$ [velocities unsolved for in the definitions of momenta (2.2a)]. The Hamiltonian can be defined as

$$H(q,p) = H_0 + \lambda_a\varphi_a, \tag{2.13}$$

where λ_a are Lagrange multipliers, and φ_a are the primary constraints.

It can be shown (e.g., Ref. 13) that Hamilton's equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \tag{2.14a}$$

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}, \tag{2.14b}$$

$$\varphi_a(q,p) = 0, \tag{2.14c}$$

are equivalent to the Euler–Lagrange equations (2.2) of the singular Lagrangian (2.1). Consequently, they are also equivalent to Eqs. (2.11) (in the fixed gauge).

Because of the existence of Eqs. (2.14a) and (2.14b), we can write Hamilton's equations in the form:

$$\dot{F} = \{F,H\}, \tag{2.14d}$$

where $\{\,,\}$ is the usual Poisson bracket.

The Hamiltonian (2.13) is not exactly in a form convenient to use in this paper, so let us rewrite it as

$$H = H_0 + \lambda_a\varphi_a + \lambda_b\varphi_b + \lambda_c G_c, \tag{2.15}$$

where λ_b and λ_c are equal to zero.

Now we can make H more convenient to use. It can be easily proven that if in any expression $B(q,p)$ we use constraints to replace something, then the result is the same expression plus some linear combination of constraints. Different equations that give $\dot{q}_i$ can differ only by the constraints (otherwise they would not give the same time evolution of q_i). So if instead of $\dot{q}_i$ given by Eqs. (2.4b), we use in the H_0 in the definition of the Hamiltonian (2.15) $\dot{q}_i = f_i$ given by Eq. (2.11b), this will only change the Lagrange multipliers there. So the Hamiltonian can be written as

$$H = [p_i\dot{q}_i - L]\,|_{\dot{q}_i = f_i} + \lambda_a\varphi_a + \lambda_b\varphi_b + \lambda_c G_c, \tag{2.16}$$

where f_i is taken from Eq. (2.11b). [It is interesting to notice that we obtained the Hamiltonian in the extended form,[9,13] which nevertheless is the same function of q's and p's as the Hamiltonian we begin with, given by Eqs. (2.13).] Equation (2.16) is the form of the Hamiltonian we will use most often in this paper.

Please note that, similarly, if we use any constraint or gauge in H, we can compensate for it by redefining the Lagrange multipliers, and still obtain the same Hamiltonian.

The Lagrange multipliers in Eq. (2.16) can be established at the end of the process by the consistency conditions.[15]

Let us now define the Dirac brackets.[9] Assume we have some set of constraints χ_u, $u = 1,...,U$ (not necessarily all constraints in the theory), and the constraints matrix defined as

$$X_{uw} = \{\chi_u,\chi_w\}, \tag{2.17}$$

is invertible. ($\{\,,\}$ is the usual Poisson bracket.) We can define the Dirac brackets with respect to constraints set χ_u as

$$\{F,G\}_\chi = \{F,G\} - \{F,\chi_u\}c_{uw}\{\chi_w,G\}, \tag{2.18}$$

where c_{uw} is the matrix inverse to X_{uw}.

The Dirac bracket has all general properties of the Poisson bracket, and also satisfies

$$\{F,H\}_\chi = \{F,H\} + a_k\varphi_k, \tag{2.19}$$

$$\{F,\chi_u\}_\chi = 0, \quad u = 1,...,U, \tag{2.20}$$

where φ_k denotes all constraints and gauges in the formalism, and F is an arbitrary function of q's and p's.

From Eq. (2.19) we conclude that Dirac bracket can be used instead of Poisson bracket in the Hamilton's equations of motion:

$$\dot{F} = \{F,H\}_\chi. \tag{2.21}$$

From Eq. (2.20), we conclude that if we decide to use the

Dirac brackets defined with respect to constraints χ_w, instead of Poisson ones, we can use the constraints χ_w freely in the Hamiltonian, and we do not have to bother about their Lagrange multipliers anymore.

At the end of this section, we want to make a technical remark that expends upon the example given by Dirac.[9] If in the formalism we have a constraints of the form

$$\chi_{(1)w} = p_w = 0, \quad w = 1,...,U, \tag{2.22a}$$

$$\chi_{(2)w} = q_w - g_w(q^U,p^U) = 0, \quad w = 1,...,U, \tag{2.22b}$$

where g_w are some functions, and $q^U = (0,...,0,q_{U+1},...,q_n), p^U = (0,...,0,p_{U+1},...,p_n)$, then the Dirac bracket calculated for these constraints gives

$$\{q_d,p_e\}_\chi = \delta_{de}, \quad d,e = U+1,...,n. \tag{2.23}$$

Also, the q_w, p_w, $w = 1,...,U$, can be completely eliminated from the Hamiltonian H by the use of constraints (2.22). This means that q_w and p_w are auxiliary variables without dynamics of their own, and Eqs. (2.22) can be looked at as merely their definition in terms of the "real" variables q_d and p_d, $d = U+1,...,n$.

III. THE FIRST-ORDER LAGRANGIAN EQUIVALENT TO A GIVEN NTH-ORDER LAGRANGIAN

In this section, we present, for a given Nth-order Lagrangian, a method of construction of a first-order one. Then we show the equivalence of their Euler–Lagrange equations of motion.

The summation convention will be used throughout this section. The range of indices are

$$i,j = 1,...,n, \quad k = 0,...,N-2, \quad m = 1,...,N-2.$$

Assume we have the system described by the coordinates $q = (q_1,...,q_n)$ and the Nth-order Lagrangian,

$$L = \mathscr{L}\left(\overset{(N)}{q}, \overset{(N-1)}{q}, \overset{(N-2)}{q}, ..., \overset{(1)}{q}, q\right), \tag{3.1}$$

where

$$\overset{(s)}{q} = \frac{d^s}{dt^s} q.$$

Then, the usual procedure of the calculus of variations (e.g., Ref. 16) gives the Euler–Lagrange equations of motion in the form

$$\sum_{s=0}^{N} (-1)^s \frac{d^s}{dt^s}\left(\frac{\partial L}{\partial \overset{(s)}{q_i}}\right) = 0. \tag{3.2}$$

We will now show that the same system can be described by a first-order Lagrangian of the form (a variation of this form could also be used[17])

$$L' = \mathscr{L}(\dot{\tilde{q}}^{N-1,0},\tilde{q}^{N-1,0},\tilde{q}^{N-2,0},\tilde{q}^{N-3,0},...,\tilde{q}^{1,0},\tilde{q}^{0,0}) + \mu_{ki}(\dot{q}_{ki} - q_{k+1,i}). \tag{3.3}$$

We use the notation which is a natural generalization of the notation used in Sec. II, for the Nth-order case,

$$\tilde{q} = (q_{N-1,1},...,q_{N-1,n},...,q_{01},...,q_{0n});$$

$$\tilde{q}^{Y,X} = (0,...,0,q_{Y,X+1},...,q_{Y,n},0,...,0),$$

$$Y = 0,...,N-1, \quad X = 0,...,n; \quad \dot{q}_{ki} = \frac{d}{dt} q_{ki}.$$

The coordinates q_{0i} in Eq. (3.3) are the same that appear as q_i in Eq. (3.1), $q_{k+1,i}$ and μ_{ki} $k = 0,...,N-2$ are some other coordinates (their interpretation will be given later). It is crucial that we treat all q's and μ's on the same footing, as independent coordinates.

In Eq. (3.3), $\mathscr{L}$, by definition, is obtained from $\mathscr{L}$ in Eq. (3.1) by substituting $\dot{q}_{N-1,i}$ in the place of $\overset{(N)}{q_i}$, and q_{Yi} in the place of $\overset{(Y)}{q_i}$.

Following the standard procedure for the first-order Lagrangians, we obtain Euler–Lagrange equations for the Lagrangian (3.3) in the form

$$\frac{\partial \mathscr{L}}{\partial q_{0i}} - \dot{\mu}_{0i} = 0, \tag{3.4a}$$

$$\frac{\partial \mathscr{L}}{\partial q_{mi}} - \mu_{m-1,i} - \dot{\mu}_{mi} = 0, \tag{3.4b}$$

$$\frac{\partial \mathscr{L}}{\partial q_{N-1,i}} - \mu_{N-2,i} - \frac{d}{dt}\left(\frac{\partial \mathscr{L}}{\partial \dot{q}_{N-1,i}}\right) = 0, \tag{3.4c}$$

$$\dot{q}_{ki} - q_{k+1,i} = 0. \tag{3.4d}$$

We can eliminate $\dot{\mu}$'s and $\dot{q}_{ki}$'s from Eqs. (3.4) and rewrite it as

$$\sum_{s=0}^{N-1} (-1)^s \frac{d^s}{dt^s}\left(\frac{\partial \mathscr{L}}{\partial q_{si}}\right) + (-1)^N \frac{d^N}{dt^N}\left(\frac{\partial \mathscr{L}}{\partial \dot{q}_{N-1,i}}\right) = 0, \tag{3.5a}$$

$$q_{k+1,i} = \overset{(k+1)}{q_{0i}}, \tag{3.5b}$$

$$\mu_{ki} = \sum_{s=0}^{N-k-2} (-1)^s \frac{d^S}{dt^S}\left(\frac{\partial \mathscr{L}}{\partial q_{k+s+1,i}}\right) + (-1)^{N-k-1} \frac{d^{N-k-1}}{dt^{N-k-1}}\left(\frac{\partial \mathscr{L}}{\partial \dot{q}_{N-1,i}}\right) \tag{3.5c}$$

(no summation over k).

Equations (3.5a) with the help of Eqs. (3.5b) are equivalent to the Euler–Lagrange equations (3.2) obtained from the Nth-order Lagrangian (3.1). Equations (3.5b) give the interpretation of q_{li} $l = 1,...,N-1$. These equations show that q_{li} is equal on the solution to the lth time derivative of q_{0i}, while still being an independent coordinate.

Equations (3.5c) contain no dynamics on their own. They just express μ's as some functions of other variables. As we will see in the next section, these equations will become constraints in the phase space, and taking the proper Dirac bracket will eliminate μ's and their canonical momenta completely from the Hamiltonian formalism.

For the purpose of the identification, we will call Lagrangian (3.3) the equivalent first-order Lagrangian (EFL).

IV. A HAMILTONIAN FORMALISM FOR THE EQUIVALENT FIRST-ORDER LAGRANGIAN

In this section, we will derive a Hamiltonian formalism for the first-order Lagrangian (3.3), using the procedure described in Sec. II, and we will compare it with the Ostrogradsky results for the Nth-order Lagrangian (3.1).

We assume here the general case of, what is called in the literature, singular Nth-order Lagrangian.[6] In the nonsingular case, the same procedure can still be used, with some steps trivialized or absent.

The summation convention will be used through this section. The range of indices will be:

$$i,j=1,...,n, \quad k=0,...,N-2, \quad m=1,...,N-2,$$
$$s=0,...,N-1, \quad r=1,...,N-1.$$

We will use the notation

$$\tilde{q}=(q_{N-1,1},...,q_{N-1,n},...,q_{01},...,q_{0n}),$$
$$\tilde{p}=(p_{N-1,1},...,p_{N-1,n},...,p_{01},...,p_{0n}),$$
$$\tilde{q}^{Y,X}=(0,...,0,q_{Y,X+1},...,q_{Y,n},0,...,0),$$

where $Y=0,...,N-1$, $X=0,...,n$.

Let us now consider an Hamiltonian structure of the first-order Lagrangian (3.3). The Euler–Lagrange equations written in terms of canonical momenta are [see Eqs. (2.2)]

$$\dot{p}_{ri}=\frac{\partial \mathscr{L}}{\partial q_{ri}}-\mu_{r-1,i}, \tag{4.1a}$$

$$\dot{p}_{0i}=\frac{\partial \mathscr{L}}{\partial q_{0i}}, \tag{4.1b}$$

$$\dot{p}_{\mu_{ki}}=\dot{q}_{ki}-q_{k+1,i}, \tag{4.1c}$$

$$p_{N-1,i}=\frac{\partial L'}{\partial \dot{q}_{N-1,i}}=\frac{\partial \mathscr{L}}{\partial \dot{q}_{N-1,i}}, \tag{4.1d}$$

$$p_{ki}=\frac{\partial L'}{\partial \dot{q}_{ki}}=\mu_{ki}, \tag{4.1e}$$

$$p_{\mu_{ki}}=\frac{\partial L'}{\partial \dot{\mu}_{ki}}=0. \tag{4.1f}$$

[Equations (4.1d)–(4.1f) are the usual definitions of canonical momenta.] Equations (4.1e) are primary constraints. The time derivative of them with the help of Eqs. (4.1a) gives

$$\dot{\mu}_{mi}=\frac{\partial \mathscr{L}}{\partial q_{mi}}-\mu_{m-1,i}, \tag{4.2a}$$

$$\dot{\mu}_{0i}=\frac{\partial \mathscr{L}}{\partial q_{0i}}, \tag{4.2b}$$

which are not constraints.

Equations (4.1f) are primary constraints. The time derivative of it with the help of Eqs. (4.1c) gives

$$\dot{q}_{ki}=q_{k+1,i}, \tag{4.3}$$

which are not constraints.

If the restricted Hessian matrix defined as

$$W'_{ij}=\frac{\partial^2 \mathscr{L}}{\partial \dot{q}_{N-1,i}\,\partial \dot{q}_{N-1,j}},$$

has the rank R, then we can solve Eqs. (4.1d) for R of the velocities $\dot{q}_{N-1,i}$, and there exists $n-R$ relations among q's and p's (see Sec. II). So we obtain

$$\dot{q}_{N-1,\alpha}=f^1_{N-1,\alpha}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \quad 0\leqslant\alpha\leqslant R, \tag{4.4a}$$

$$\varphi_a(\tilde{q},\tilde{p})=0, \quad R<a\leqslant n, \tag{4.4b}$$

and Eqs. (4.4b) are primary constraints. Using Eqs. (4.3) and (4.4) we can write Eqs. (4.1) as

$$\dot{p}_{ri}=\frac{\partial \mathscr{L}}{\partial q_{ri}}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R})-p_{r-1,i}, \tag{4.5a}$$

$$\dot{p}_{0i}=\frac{\partial \mathscr{L}}{\partial q_{0i}}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \tag{4.5b}$$

$$\dot{q}_{ki}=q_{k+1,i}, \tag{4.5c}$$

$$\dot{q}_{N-1,\alpha}=f^1_{N-1,\alpha}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \quad 0\leqslant\alpha\leqslant R, \tag{4.5d}$$

$$\varphi_a(\tilde{q},\tilde{p})=0, \quad R<a\leqslant n, \tag{4.5e}$$

$$p_{ki}=\mu_{ki}, \tag{4.5f}$$

$$p_{\mu_{ki}}=0. \tag{4.5g}$$

Continuing with the method described in Sec. II, we can find constraints secondary to constraints (4.5e):

$$\varphi_b(\tilde{q},\tilde{p})=0, \tag{4.6}$$

and fix a gauge

$$G_c(\tilde{q},\tilde{p})=0. \tag{4.7}$$

Then, all $\dot{q}_{N-1,i}$ can be expressed as

$$\dot{q}_{N-1,i}=f_{N-1,i}(\tilde{q},\tilde{p}). \tag{4.8}$$

The specific form of the secondary constraints and gauge will depend on the specific form of the given Lagrangian. Please note that to find Eqs. (4.6)–(4.8), we only have to use Eqs. (4.5a)–(4.5e).

A Hamiltonian is defined as in (2.16) by

$$H=\left[p_{si}\dot{q}_{si}+p_{\mu_{ki}}\dot{\mu}_{ki}-\mathscr{L}-\mu_{ki}(\dot{q}_{ki}-q_{k+1,i})\right]_A$$
$$+\lambda_a\varphi_a+\lambda_b\varphi_b+\lambda_cG_c$$
$$+\lambda_{ki}p_{\mu_{ki}}+\lambda'_{ki}(p_{ki}-\mu_{ki}). \tag{4.9}$$

By writing A in the formula, we denote the fact that inside of the brackets we use equations (4.2), (4.3), and (4.8) to express time derivatives of the coordinates.

We can now use constraints (4.5f) and (4.5g) to define Dirac brackets. According to the remark at the end of Sec. II, we will obtain

$$\{q_{si},p_{tj}\}_D=\delta_{st}\delta_{ij}, \quad s,t=0,...,N-1. \tag{4.10}$$

If we use the Dirac brackets (4.10) in the dynamics, we can eliminate μ's and p_μ's completely from the Hamiltonian (4.9) obtaining

$$H=p_{N-1,i}f_{N-1,i}+p_{ki}q_{k+1,i}-\mathscr{L}\Big|_{\substack{\dot{q}_{ki}=q_{k+1,i}\\ \dot{q}_{N-1,i}=f_{N-1,i}}}$$
$$+\lambda_a\varphi_a+\lambda_b\varphi_b+\lambda_cG_c. \tag{4.11}$$

It is now interesting to compare our results with the Ostrogradsky method[1] of constructing a Hamiltonian formalism. Let us do this for a nonsingular[6] Nth-order Lagrangian. In our formalism, this means that $R=n$ in formula (4.4) and that the constraints φ_a,φ_b, and G_c are absent. It is easy to see that, in this case, the Hamiltonian (4.11) is exactly equal to the Hamiltonian given by Ostrogradsky. The Dirac brackets

we obtained in Eq. (4.10) are exactly the Poisson brackets used in the Ostrogradsky formalism.[3] From Eqs. (4.1a), (4.1d), and (4.1e), we can also obtain

$$p_{N-1,i} = \frac{\partial \mathscr{L}}{\partial \dot{q}_{N-1,i}}, \tag{4.12a}$$

$$p_{ki} = \frac{\partial \mathscr{L}}{\partial q_{k+1,i}} - \dot{p}_{k+1,i}. \tag{4.12b}$$

Equations (4.12b) are equations of motion, not the definitions of canonical momenta as in Ostrogradsky formalism. We can conclude that our EFL construction together with the Dirac constraints theory provides us with a derivation of the Ostrogradsky formalism.[18]

The more general, nonsingular case is included in a natural way in our first-order formalism, so our construction provides us with the generalization of the Ostrogradsky formalism to the singular case. In this case, we would use all constraints and gauges to define Dirac brackets.[14] It is interesting to notice that, in the singular case, Eqs. (4.12) (Ostrogradsky definitions of canonical momenta) still hold.

V. UNIQUENESS OF THE DYNAMICAL STRUCTURE

In contrast to the usual situation of using first-order Lagrangians, when we decide to use the higher-order ones, we have much more freedom in choosing among different Lagrangians with classically equivalent equations of motion. In particular, it is possible to treat a given Lagrangian as being of a different order in time, or to add the complete time derivative term to it. In this section, we want to show that despite this fact, the Hamiltonian structure we obtain is always the same (up to a canonical transformation). More exactly, we will show that no matter how we change the Lagrangian, the coordinates and momenta remaining in the Hamiltonian formalism, the form of the Hamiltonian itself and the Dirac brackets of the theory will remain unchanged, or will only differ by time-independent canonical transformation. We assume the general case of singular Nth-order Lagrangian.[6] In the nonsingular case, the same proof is still valid, with some steps trivialized or absent.

The summation convention will be used through this section. The range of indices will be:

$$i,j = 1,...,n, \quad k = 0,...,N-2, \quad m = 1,...,N-2,$$
$$s = 0,...,N-1, \quad r = 1,...,N-1.$$

As in Sec.IV, we will use the notation

$$\tilde{q} = (q_{N-1,1},...,q_{N-1,n},...,q_{01},...,q_{0n}),$$
$$\tilde{p} = (p_{N-1,1},...,p_{N-1,n},...,p_{01},...,p_{0n}),$$
$$\tilde{q}^{Y,X} = (0,...,0,q_{Y,X+1},...,q_{Y,n},0,...,0),$$
$$Y = 0,...,N-1, \quad X = 0,...,n.$$

A. Changing the order of the Lagrangian

The first problem we encounter is how to choose the order of the Lagrangian. Let us consider what will happen in a case when a Lagrangian is given as the Nth-order in time, but we insist in treating it as being of some higher order. In this case, our construction will give, instead of the Lagrangian (3.3), the first-order Lagrangian in the form

$$L'' = \mathscr{L}(\tilde{q}^{N,0},\tilde{q}^{N-1,0},...,\tilde{q}^{0,0}) + \mu_{ui}(\dot{q}_{ui} - q_{u+1,i}). \tag{5.1}$$

where $u = 0,...,M$, $M > N-2$.

The Euler–Lagrange equations expressed in terms of canonical momenta (2.2) will then have the form

$$\dot{p}_{ui} = -\mu_{u-1,i}, \quad u = N+1,...,M+1, \tag{5.2a}$$

$$\dot{p}_{ui} = \frac{\partial \mathscr{L}}{\partial q_{ui}} - \mu_{u-1,i}, \quad u = 1,...,N, \tag{5.2b}$$

$$\dot{p}_{0i} = \frac{\partial \mathscr{L}}{\partial q_{0i}}, \tag{5.2c}$$

$$\dot{p}_{\mu_{ui}} = \dot{q}_{ui} - q_{u+1,i}, \quad u = 0,...,M, \tag{5.2d}$$

where

$$p_{M+1,i} = \frac{\partial L''}{\partial \dot{q}_{M+1,i}} = 0, \tag{5.2e}$$

$$p_{ui} = \frac{\partial L''}{\partial \dot{q}_{ui}} = \mu_{ui}, \quad u = 0,...,M, \tag{5.2f}$$

$$p_{\mu_{ui}} = \frac{\partial L''}{\partial \dot{\mu}_{ui}} = 0, \quad u = 0,...,M. \tag{5.2g}$$

Equations (5.2g) and (5.2f) are primary constraints. The time derivative of it with the help of Eqs. (5.2a)–(5.2d) gives

$$\dot{q}_{ui} = q_{u+1,i}, \quad u = 0,...,M, \tag{5.3}$$

$$\dot{\mu}_{ui} = -\mu_{u-1,i}, \quad u = N+1,...,M, \tag{5.4a}$$

$$\dot{\mu}_{ui} = \frac{\partial \mathscr{L}}{\partial q_{ui}} - \mu_{u-1,i}, \quad u = 1,...,N, \tag{5.4b}$$

$$\dot{\mu}_{0i} = \frac{\partial \mathscr{L}}{\partial q_{0i}}, \tag{5.4c}$$

which are not constraints.

Equations (5.2e) are primary constraints. The time derivative of it with the help of Eq. (5.2a) will give the set of secondary constraints

$$\alpha_{ui} = 0, \quad u = N,...,M, \tag{5.5a}$$

$$p_{ui} = 0, \quad u = N,...,M. \tag{5.5b}$$

The time derivative of constraints (5.5b) for $u = N$ with the help of Eqs. (5.2b) and (5.2f) will give

$$p_{N-1,i} = \frac{\partial \mathscr{L}}{\partial q_{Ni}}, \tag{5.6}$$

which are secondary constraints. Because Eqs. (5.6) have the same functional form as Eq. (4.1d), we can solve them for q_{Ni} in similar way and obtain [see Eqs. (4.4)]

$$q_{N\alpha} = f^1_{N-1,\alpha}(\tilde{q},\tilde{p},\tilde{q}^{N,R}), \quad 0 \leqslant \alpha \leqslant R, \tag{5.7a}$$

$$\varphi_a(\tilde{q},\tilde{p}) = 0, \quad R < a \leqslant n, \tag{5.7b}$$

which are secondary constraints equivalent to constraints (5.6).

Using Eqs. (5.3) and (5.7a), we can also write the expression for $\dot{q}_{N-1,i}$:

$$\dot{q}_{N-1,\alpha} = f^1_{N-1,\alpha}(\tilde{q},\tilde{p},\tilde{q}^{N-1,R}), \quad 0 \leqslant \alpha \leqslant R. \tag{5.8}$$

Using this last equation, we can write some of the equations that we have already obtained, in the form:

$$\dot{p}_{ri} = \frac{\partial \mathcal{L}}{\partial q_{ri}}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}) - p_{r-1,i}, \tag{5.9a}$$

$$\dot{p}_{0i} = \frac{\partial \mathcal{L}}{\partial q_{0i}}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \tag{5.9b}$$

$$\dot{q}_{ki} = q_{k+1,i}, \tag{5.9c}$$

$$\dot{q}_{N-1,\alpha} = f^{1}_{N-1,\alpha}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \quad 0 \leqslant \alpha \leqslant R, \tag{5.9d}$$

$$\varphi_a(\tilde{q},\tilde{p}) = 0, \quad R < a \leqslant n. \tag{5.9e}$$

It is easy to notice that Eqs. (5.9) are identical to Eqs. (4.5a)–(4.5e), which were used to find secondary constraints (4.6), the gauge (4.7) and the expression (4.8). We can do exactly the same operations this time and obtain

$$\varphi_b(\tilde{q},\tilde{p}) = 0, \tag{5.10}$$

$$G_c(\tilde{q},\tilde{p}) = 0, \tag{5.11}$$

$$\dot{q}_{N-1,i} = f_{N-1,i}(\tilde{q},\tilde{p}), \tag{5.12}$$

exactly in the same form as in Sec. IV [we are able to fix the gauge before considering other equations, because the Euler–Lagrange equations of motion are equivalent for the Lagrangians (3.3) and (5.1), so the same velocities can be fixed as the same functions of time, and the same gauges can be imposed].

Equations (5.12) together with Eqs. (5.3) will give

$$q_{Ni} = f_{N-1,i}(\tilde{q},\tilde{p}), \tag{5.13}$$

which are secondary constraints [actually it is a kind of mixture of constraints and gauge, some of them are equivalent to constraints (5.7a) in the fixed gauge]. Taking many times its time derivative, with the help of Eqs. (5.3), (5.12), (5.9a), and (5.9b), we obtain secondary constraints

$$q_{ui} = f_{u-1,i}(\tilde{q},\tilde{p}), \quad u = N+1,\ldots,M+1, \tag{5.14}$$

and, in the last step, the expressions

$$\dot{q}_{M+1,i} = f_{M+1,i}(\tilde{q},\tilde{p}), \tag{5.15}$$

which are not constraints. This finishes the process of obtaining the constraints of the system.

We can now write the Hamiltonian [see (2.16)] in the form

$$H = \left(\sum_{u=0}^{M+1} p_{ui}\dot{q}_{ui} + \sum_{u=0}^{M} p_{\mu_{ui}}\dot{\mu}_{ui} - \mathcal{L} - \sum_{u=0}^{M} \mu_{ui}(\dot{q}_{ui} - q_{u+1,i}) \right) \Bigg|_{\substack{\dot{q}_{M+1,i} = f_{M+1,i} \\ \dot{q}_{N-1,i} = f_{N-1,i} \\ \dot{q}_{ui} = q_{u+1,i} \quad u = 0,\ldots,N-2,N,\ldots,M}}$$

$$+ \lambda^{1}_{i} p_{M+1,i} + \sum_{u=N}^{M} \lambda^{2}_{ui} p_{ui}$$

$$+ \sum_{u=0}^{M} \lambda^{3}_{ui} p_{\mu_{ui}} + \lambda^{4}_{si}(p_{si} - \mu_{si}) + \sum_{u=N}^{M} \lambda^{5}_{ui}\mu_{ui} + \sum_{u=N}^{M+1} \lambda^{6}_{ui}\{q_{ui} - f_{u-1,i}(\tilde{q},\tilde{p})\} + \lambda_a\varphi_a + \lambda_b\varphi_b + \lambda_c G_c. \tag{5.16}$$

The constraints (5.2f), (5.2g), (5.5), (5.2e), (5.12), and (5.13) can be used to define the Dirac brackets (2.18). According to the remark at the end of Sec. II, we obtain

$$\{q_{si},p_{tj}\}_D = \delta_{st}\delta_{ij}, \quad s,t = 0,\ldots,N-1, \tag{5.17}$$

which is exactly the formula (4.10) obtained in Sec. IV. If we use the constraints in the Hamiltonian (5.16) we obtain

$$H = p_{N-1,i} f_{N-1,i} + p_{ki} q_{k+1,i} - \mathcal{L}\Big|_{\substack{\dot{q}_{ki} = q_{k+1,i} \\ \dot{q}_{N-1,i} = f_{N-1,i}}}$$

$$+ \lambda_a\varphi_a + \lambda_b\varphi_b + \lambda_c G_c, \tag{5.18}$$

which is exactly the Hamiltonian (4.11). So the dynamics of the Lagrangians (3.3) and (5.1) are identical.

B. Leaving more derivatives in $\mathcal{L}$

Another change we can make in choosing the Lagrangian, is to leave not only the time derivative of $\tilde{q}^{N-1,0}$ in $\mathcal{L}$, but some other derivatives too. We can use the Lagrangian

$$L''' = \mathcal{L}'(\dot{\tilde{q}}^{N-1,0},\tilde{q}^{N-1,0},\dot{\tilde{q}}^{N-2,0},\tilde{q}^{N-2,0} \times \dot{\tilde{q}}^{N-3,0},\tilde{q}^{N-3,0},\ldots,\dot{\tilde{q}}^{0,0}\tilde{q}^{0,0}) + \mu_{ki}(\dot{q}_{ki} - q_{k+1,i}), \tag{5.19}$$

instead of Lagrangian (3.3), as long as

$$\mathcal{L}'|_{\dot{q}_{ki} = q_{k+1,i}} = \mathcal{L}, \tag{5.20}$$

where $\mathcal{L}$ is that from definition (3.3).

Euler–Lagrange equations for the Lagrangian L''' are

$$\dot{p}_{ri} = \frac{\partial \mathcal{L}'}{\partial q_{ri}} - \mu_{r-1,i}, \tag{5.21a}$$

$$\dot{p}_{0i} = \frac{\partial \mathcal{L}'}{\partial q_{0i}}, \tag{5.21b}$$

$$\dot{p}_{\mu_{ki}} = \dot{q}_{ki} - q_{k+1,i}, \tag{5.21c}$$

$$P_{N-1,i} = \frac{\partial L'''}{\partial \dot{q}_{N-1,i}} = \frac{\partial \mathcal{L}'}{\partial \dot{q}_{N-1,i}}, \tag{5.21d}$$

$$p_{ki} = \frac{\partial L'''}{\partial \dot{q}_{ki}} = \frac{\partial \mathcal{L}'}{\partial \dot{q}_{ki}} + \mu_{ki}, \tag{5.21e}$$

$$p_{\mu_{ki}} = \frac{\partial L'''}{\partial \mu_{ki}} = 0. \tag{5.21f}$$

Using $\dot{q}_{ki} = q_{k+1,i}$, which comes as a consequence of Eqs. (5.21f), (5.21c), and (5.21e), (5.20a), and the fact that from Eq. (5.20), we have

$$\frac{\partial \mathcal{L}'}{\partial q_{ri}} + \frac{\partial \mathcal{L}'}{\partial \dot{q}_{r-1,i}} = \frac{\partial \mathcal{L}}{\partial q_{ri}}, \tag{5.22a}$$

$$\frac{\partial \mathscr{L}'}{\partial q_{0i}} = \frac{\partial \mathscr{L}}{\partial q_{0i}}, \tag{5.22b}$$

$$P_{N-1,i} = \frac{\partial \mathscr{L}'}{\partial \dot{q}_{N-1,i}} = \frac{\partial \mathscr{L}}{\partial \dot{q}_{N-1,i}}, \tag{5.22c}$$

Eqs. (5.21d) can be written [see (2.4)] as

$$\dot{q}_{N-1,\alpha} = f^1_{N-1,\alpha}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \quad 0 \leqslant \alpha \leqslant R, \tag{5.22c'}$$

$$\varphi_a(\tilde{q},\tilde{p}) = 0, \quad R < a \leqslant n, \tag{5.22c''}$$

where Eqs. (5.22c'), (5.22c'') are of the same form as Eqs. (4.4).

Now we can write (5.21) as

$$\dot{p}_{ri} = \frac{\partial \mathscr{L}}{\partial q_{ri}}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}) - p_{r-1,i}, \tag{5.23a}$$

$$\dot{p}_{0i} = \frac{\partial \mathscr{L}}{\partial q_{0i}}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \tag{5.23b}$$

$$\dot{q}_{ki} = q_{k+1,i}, \tag{5.23c}$$

$$\dot{q}_{N-1,\alpha} = f^1_{N-1,\alpha}(\tilde{q},\tilde{p},\dot{\tilde{q}}^{N-1,R}), \quad 0 \leqslant \alpha \leqslant R, \tag{5.23d}$$

$$\varphi_a(\tilde{q},\tilde{p}) = 0, \quad R < a \leqslant n, \tag{5.23e}$$

$$p_{ki} = \frac{\partial \mathscr{L}'}{\partial \dot{q}_{ki}} - \mu_{ki}, \tag{5.23f}$$

$$p_{\mu_{ki}} = 0. \tag{5.23g}$$

Equations (5.23a)–(5.23e) are identical to Eqs. (4.5a)–(4.5e). So we can obtain constraints, gauge, and expression for $\dot{q}_{N-1,i}$ in exactly the same form as we did in Sec. IV:

$$\varphi_b(\tilde{q},\tilde{p}) = 0, \tag{5.24a}$$

$$G_c(\tilde{q},\tilde{p}) = 0, \tag{5.24b}$$

$$\dot{q}_{N-1,i} = f_{N-1,i}(\tilde{q},\tilde{p}). \tag{5.24c}$$

[We are able to fix the gauge before considering other equations, because the Euler–Lagrange equations of motion are equivalent for the Lagrangians (3.3) and (5.1), so the same velocities can be fixed as the same arbitrary functions of time, and the same gauges can be fixed.]

After the use of Eqs. (5.24c) and (5.23c), the (5.23f) becomes a secondary constraint.

The Hamiltonian is defined as in (2.16) by

$$\begin{aligned} H''' = (&p_{N-1,i}\dot{q}_{N-1,i} + p_{ki}\dot{q}_{ki} + p_{\mu_{ki}}\dot{\mu}_{ki} - \mathscr{L}' \\ &- \mu_{ki}(\dot{q}_{ki} - q_{k+1,i}))_{\substack{\dot{q}_{N-1,i} = f_{N-1,i} \\ \dot{q}_{ki} = q_{k+1,i}}} + \lambda_a \varphi_a \\ &+ \lambda_b \varphi_b + \lambda_c G_c + \lambda_{ki} p_{\mu_{ki}} \\ &+ \lambda'_{ki}\left(p_{ki} - \frac{\partial \mathscr{L}'}{\partial \dot{q}_{ki}} - \mu_{ki}\right). \end{aligned} \tag{5.25}$$

We can use constraints (5.23f) and (5.23g) to calculate the Dirac brackets. According to the remark at the end of Sec.II, we will obtain

$$\{q_{si}, p_{tj}\}_D = \delta_{st}\delta_{ij}, \quad s,t = 0,\ldots,N-1, \tag{5.26}$$

which is exactly Eq. (4.10). If we use the constraints and Eq. (5.20) in the Hamiltonian (5.25), we will obtain

$$\begin{aligned} H''' = &p_{N-1,i} f_{N-1,i} + p_{ki} q_{k+1,i} - \mathscr{L}\Big|_{\substack{\dot{q}_{ki} = q_{k+1,i} \\ \dot{q}_{N-1,i} = f_{N-1,i}}} \\ &+ \lambda_a \varphi_a + \lambda_b \varphi_b + \lambda_c G_c, \end{aligned} \tag{5.27}$$

which is exactly the Hamiltonian (4.11). So, we have obtained the dynamics identical to the one introduced in Sec. IV.

C. Adding a $(d/dt)F$ term to the Lagrangian

It is a well-known fact, in the first-order formalism, that we can add a complete time derivative to the Lagrangian, and it does not change the dynamics of the system. We now want to show that this is true, in general, for the Lagrangian of Nth order in time (even in the singular case). Let us assume we have the Lagrangian

$$L = \mathscr{L}\left(\overset{(N)}{q}, \overset{(N-1)}{q}, \ldots, \overset{(1)}{q}, q\right), \tag{5.28}$$

and we add a time derivative term obtaining

$$L_{\text{add}} = \mathscr{L}\left(\overset{(N)}{q}, \overset{(N-1)}{q}, \ldots, \overset{(1)}{q}, q\right) + \frac{d}{dt}F, \tag{5.29}$$

where F is a function of q and its time derivatives of arbitrary order.

Using the first-order formalism (Secs. III, IV, V A, and V B), we employ, instead of the Lagrangian (5.28), the Lagrangian

$$\begin{aligned} L' = &\mathscr{L}(\dot{\tilde{q}}^{N-1,0}, \tilde{q}^{N-1,0}, \ldots, \tilde{q}^{0,0}) \\ &+ \sum_{u=0,\ldots,M} \mu_{ui}(\dot{q}_{ui} - q_{u+1,i}). \end{aligned} \tag{5.30}$$

Similarly, instead of the Lagrangian (5.29), we can use

$$L'_{\text{add}} = L' + \sum_{u=0,\ldots,M} \dot{q}_{ui} \frac{\partial F}{\partial q_{ui}}, \tag{5.31}$$

where M was taken big enough so that F can be expressed without using time derivatives. From now on let us use the unified variables x_β, for denoting all q's and α's. Then, we have

$$L'_{\text{add}}(\dot{x},x) = L'(\dot{x},x) + \dot{x}_\beta \frac{\partial F}{\partial x_\beta}. \tag{5.32}$$

Let us consider L' first. Canonical momenta are given by

$$p_\beta = \frac{\partial L'}{\partial \dot{x}_\beta}. \tag{5.33}$$

Equation (5.33), if not solvable for all $\dot{x}_\beta$, will give primary constraints [see (2.4)]

$$\varphi_a = p_a - \frac{\partial L'}{\partial \dot{x}_a} = 0. \tag{5.34}$$

The Hamiltonian is equal [here we use Eq. (2.14) as definition, because it is more convenient]:

$$H' = p_\beta \dot{x}_\beta - \mathscr{L} + \lambda_a \varphi_a. \tag{5.35}$$

The other Lagrangian will give

$$p'_\beta = \frac{\partial L'_{\text{add}}}{\partial \dot{x}_\beta} = p_\beta + \frac{\partial F}{\partial x_\beta}, \tag{5.36}$$

which will give primary constraints

$$\varphi'_a = p'_a - \frac{\partial F}{\partial x_a} - \frac{\partial L}{\partial \dot{x}_a} = \varphi_a = 0, \tag{5.37}$$

and the Hamiltonian

$$H_{\text{add}} = p'_\beta \dot{x}_\beta - L'_{\text{add}} + \lambda_a \varphi_a = p_\beta \dot{x}_\beta - L' + \lambda_a \varphi_a = H, \tag{5.38}$$

so the Hamiltonians are equal. The change of variables from (x_β, p_β) to (x_β, p'_β) is canonical because

$$\{x_\beta, p'_{\beta'}\} = \left\{x_\beta, p_{\beta'} - \frac{\partial F}{\partial x_{\beta'}}\right\} = \{x_\beta, p_{\beta'}\} = \delta_{\beta\beta'}, \tag{5.39}$$

where $\{\,,\}$ denotes the Poisson brackets with respect to (x_β, p_β). So, adding the complete time derivative to the Lagrangian can only induce a canonical transformation in the phase space.

VI. FINAL REMARKS

In this paper, we have shown that it is always possible, for a given Lagrangian of N th-order in time, to construct a first-order Lagrangian with equivalent equations of motion, and that by the use of the Dirac theory of constraints we can find a Hamiltonian structure for this Lagrangian. There are advantages to our approach. One is the fact that the theory of higher-order Lagrangians can easily be incorporated in the traditional formalism, and thus we do not have to construct a new one. Once our Lagrangian is constructed, we can employ all well-known methods of investigating the usual, first-order systems (e.g., Noether theorem, Hamilton–Jacobi method, canonical transformations, and so on). We were also able to show, using this first-order formalism, that the ambiguities, usually connected with higher-order Lagrangians, do not appear in our approach.

We would like to say now a few words about an important paper by Saito *et al.*,[7] which contains a version of "Hamiltonization" of singular higher-order Lagrangians. We would like to briefly discuss differences and similarities between their approach and ours. This discussion is necessary to show that our work is not just an independent proof of their results, but a different, self consistent approach to the problem. We will restrict our comments to Sec. II and the parts of Sec. I in their paper, which concern the construction of Hamiltonian formalism.

The first important result we want to discuss, is the structure of constraints proposed in their paper. In the Ostrogradsky formalism, if for an s and i, the definition of a canonical momentum

$$p_i^{(s-1)} = \frac{\partial L}{\partial \overset{(s)}{q_i}} - \dot{p}_i^{(s)} \quad (s = 1,\ldots,N-1),$$

does not contain $\overset{(N+k)}{q}$ $k \geqslant 0$, then it is a constraint. Such and only such constraints are considered by Saito *et al.* as being "contained in the Ostrogradsky formalism." However, the Ostrogradsky formalism can contain a variety of other constraints. For example, let us consider the Lagrangian

$$L = \tfrac{1}{2}(\overset{(3)}{x_1})^2 + \overset{(2)}{x_1}\overset{(2)}{x_2}, \tag{6.1}$$

where $\overset{(s)}{x_i} = d^s x_i / dt^s$.

The Ostrogradsky transformation will produce, among the others, the canonical momenta

$$p_1^{(2)} = \overset{(3)}{x_1}, \tag{6.2a}$$

$$p_2^{(0)} = -\overset{(3)}{x_1}. \tag{6.2b}$$

So we obtain the constraint

$$p_1^{(2)} = -p_2^{(0)}, \tag{6.3}$$

despite the fact, that both definitions (6.2) contain $\overset{(3)}{x_1}$.

In other words, any equation containing only $\overset{(s)}{q_i}$ and $\overset{(s)}{p_i}$ $s = 0,\ldots,N-1$ which can be obtained from the Ostrogradsky definition of momenta is a constraint resulting from the Ostrogradsky transformation, and has to appear in the formalism. Constraint (6.3) according to the definition of Saito *et al.* is not "contained in the Ostrogradsky transformation." This is not only a question of terminology, because for the consistency of their formalism, it is necessary for them to prove that all constraints of this kind will be obtained as secondary to the ones they consider as primary. However they fail to show this for constraints other than the ones they call "contained in the Ostrogradsky formalism." Their proof is tailored to a specific form of the constraints and does not seem to be easily adjustable to the possible variety of all constraints produced by the Ostrogradsky formalism. It is worth noting that in our method, thanks to the first-order formalism we use, we avoid the problem completely.

Also, the proof given by Saito *et al.* for the very strong claim that adding a complete time derivative term to the higher-order nonsingular Lagrangian has no physical effect does not justify the conclusion. They only show that the Hamiltonian will remain unchanged, the proper number of variables will be eliminated by constraints and the Euler–Lagrange equations of the new system will be equivalent. However, this may be not enough for the formalisms to be equivalent. Also, their counting of constraints, which is crucial for the proof, is inaccurate. For example, consider the Lagrangian

$$L = \frac{1}{2}\left[\overset{(1)}{x}\right]^2 + \overset{(3)}{x}, \tag{6.4}$$

which is obtained from the usual Lagrangian of a free, one dimensional particle, by the addition of the total time derivative term $\overset{(3)}{x}$. Using the method proposed by Saito *et al.*, we obtain four second-class[9] constraints, instead of two second-class constraints and one first-class[9] constraint as predicted by these authors in their proof. It is worth noting that in our paper we show explicitly that adding a complete time derivative to the higher-order Lagrangian can produce at most a canonical transformation of the formalism, even in the more general case of a singular Lagrangian.

Despite the criticism, we think that the paper by Saito *et al.* is a very important one, and we fully agree with the general idea, that singular higher-order Lagrangian systems should be treated by the use of a combination of Ostrogradsky transformation and Dirac formalism. However, we also think that such a system should be treated as a first-order one, in the way we presented here.

It may be interesting to say a few words about the appli-

cability of the methods we employed here, for some problems which are outside the scope of this paper. It is easy to notice that in this paper we look at the N th-order Lagrangian as on a first-order system with nonholonomic constraints in the form $\dot{q}_{ki} = q_{k+1,i}$ imposed on it. These constraints become a natural part of the formalism (they become equations of motion), when we add them to the Lagrangian and treat their Lagrange multipliers as independent coordinates. The Dirac theory is then used to obtain a Hamiltonian formalism. The same approach can be used for other systems with different holonomic or nonholonomic constraints imposed. Also, it can be used to construct a Lagrangian and a Hamiltonian for a given system of differential and algebraic equations, when the usual Lagrangian is not known. In this case, a higher-order Lagrangian can be defined as a linear combination of the equations, multiplied by Lagrange multipliers, which are then treated as independent coordinates (later we can reduce the order of the Lagrangian as shown in this paper). However, in this method not all Lagrange multipliers will be reduced out by the Dirac method, so the obtained system describes an embedding of the original dynamics into that of a larger dynamical system.

The method we present in this paper can be also used in the case of higher-order Lagrangian systems with infinite number of degrees of freedom. In this case, we introduce independent coordinates and Lagrange multipliers not for time derivatives only, but for other higher-order derivatives as well.

Note added in proof: (1) Lagrangians similar to (3.3) were used by B. Kupershmidt[19] in Hamilton–Cartan formalism of classical field theory. (2) The idea mentioned in Sec. VI, of constructing the Lagrangian for a given differential equation, by multiplying the equation by a Lagrange multiplier, and then treating the multiplier as independent variable, was first given by Bateman.[20] However, the Dirac theory of constraints was not known at the time, so he was not able to derive a Hamiltonian from his Lagrangian, at least not in the general case.

ACKNOWLEDGMENT

The author is deeply grateful to Professor Robert L. Anderson for very valuable discussions and comments, and constant encouragement to pursue this work.

[1] M. Ostrogradsky, Mem. Acad. St. Petersburg **6** (4), 385 (1850); E. T. Whittaker, *Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge U.P., Cambridge, 1959), 4th ed., p. 266.

[2] M. Borneas, Am. J. Phys. **27**, 265 (1959); Nuovo Cimento **16** (5), 806 (1960).

[3] J. G. Koestler and J. A. Smith, Am. J. Phys. **33**, 140 (1965).

[4] G. C. Constantelos, Nuovo Cimento **21B** (2), 279 (1974).

[5] G. C. Constantelos, Nuovo Cimento **84B** (1), 91 (1984).

[6] The N th-order Lagrangian is called singular if the equations

$$p_i^{(N-1)} = \frac{\partial L}{\partial q_i^{(N)}}$$

cannot be solved for all $q_i^{(N)}$, otherwise it is called nonsingular. This definition is somewhat arbitrary, because we can always take N greater than the order of time derivatives actually appearing in L, and the Lagrangian becomes singular. The results of Sec. V show that this choice does not change the final Hamiltonian structure. So the words "singular" or "nonsingular" used in this paper refers also to the way we treat a given Lagrangian, not only to the Lagrangian itself.

[7] Y. Saito, R. Sugano, T. Ohta, and T. Kimura, J. Math. Phys. **30** 1122 (1989).

[8] C. F. Hayes and J. M. Jankowski, Nuovo Cimento **58B**, (2), 494 (1968).

[9] P. A. M. Dirac, *Lectures on Quantum Mechanics* (Belfer Graduate School, Yeshiva University, New York, 1964); Can. J. Math. **2**, 129 (1950).

[10] P. W. Hebda, Ph.D. dissertation, Univ. of Georgia, in preparation.

[11] C. F. Hayes, J. Math. Phys. **10**, 1555 (1969).

[12] C. Ryan, J. Math. Phys. **13**, 283 (1972).

[13] E. C. G. Sudarshan and M. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974).

[14] A. J. Hanson, T. Regge, and C. Teitelboim, *Constrained Hamiltonian Systems* (Accad. Nazionale dei Lincei; Scienze Mathematiche e loro Applicazioni, 1976), Vol. 22.

[15] The Lagrange multipliers are given uniquely (up to a linear combination of constraints and gauges) for all primary constraints and all secondary constraints of the second class[9] by the condition that the time derivatives of the constraints and gauges are zero. If there are secondary constraints of the first class[9] in the formalism, then the Lagrangian coefficients for them can be established (again up to a linear combination of the constraints and gauges) by the condition that the Hamiltonian equations are equivalent for the Hamiltonians (2.13) and (2.16). It is obvious that such a condition gives unique (up to a linear combination of constraints and gauges) Lagrange multipliers. Otherwise, the time evolution of an observable, which is "noncommuting" with secondary constraints of the first class, would be unique with Hamiltonian (2.13) and free with (2.16).

[16] I. M. Gelfand and S. V. Fomin, *Calculus of Variations* (Prentice-Hall, New York, 1963).

[17] It is possible to use the Lagrangian given by

$$L' = \mathscr{L} + \Sigma_i \Sigma_{k=0}^{N_i-1} \mu_{ki}(\dot{q}_{ki} - q_{k+1,i}),$$

where N_i is the highest order of time derivative of q_i in L, and q_{ki} is substituted for $q_i^{(k)}$ $k = 0,\ldots,N_i - 1$ in $\mathscr{L}$. In the formula (3.3), we use instead the form with N common for all q_i, just for simplicity. The results of Sec. V show that this does not change the final Hamiltonian structure obtained from the formalism.

[18] The Ostrogradsky formalism can be introduced also in a more general case of a weaker definition of singularity, where we ask for the possibility of expressing all the highest time derivatives of q's actually appearing in higher-order Lagrangian, by the Ostrogradsky canonical momenta. In this case, it is possible to have $n > R$ in (4.4) and our procedure will give some φ_a and φ_b constraints, and we again obtain Ostrogradsky results. Also, in this case, we can use the Lagrangian mentioned in Ref. 17 instead of Lagrangian (3.3), which would yield no φ_a, φ_b constraints.

[19] B. Kupershmidt, Lecture Notes in Mathematics, Vol. 775, 162 (1980).

[20] H. Bateman, Phys. Rev. **38**, 815 (1931).

Lagrangian and Hamiltonian many-time equations

Luca Lusanna
Sezione I.N.F.N. di Firenze, Largo E. Fermi 2, Arcetri, 50125 Firenze, Italy

(Received 13 December 1989; accepted for publication 11 April 1990)

The Lagrangian and Hamiltonian many-time equations are derived for a finite-dimensional system with an arbitray number of primary and secondary first-class constraints. Assuming that all the secondary constraints are generators of gauge transformations, the general form of the Lagrangian gauge algebra is given.

I. INTRODUCTION

In two recent papers[1,2] the general structure of singular Lagrangians and of their associated phase-space (Dirac–Bergmann theory of Hamiltonian constraints) and velocity-space descriptions have been studied in the finite-dimensional case by using the second Noether theorem. In Ref. 1 and then in Ref. 3 it has been pointed out that in the case of only first-class constraints (only true gauge transformations at the Lagrangian level) one can reformulate the phase space Hamilton–Dirac equations by means of the so-called "many-time approach".[4] In it the arbitrary Dirac multipliers are replaced with an equal number of independent "times" and the canonical variables are thought as functions of the ordinary time (when a nonvanishing canonical Hamiltonian exists) and of these times. Then one considers a system of as many pairs of Hamilton equations as the total number of times: one pair has the canonical Hamiltonian as Hamiltonian, while the Hamiltonians of the other pairs are either the first-class constraints themselves, if they satisfy an Abelian Poisson algebra, or one of their Abelianized forms. When the original constraints satisfy a nonabelian Lie–Poisson algbra it is possible to write the many-time Hamilton equations in a form that uses the original constraints as Hamiltonians, if one uses the left-invariant vector fields dual of the Maurer–Cartan one-forms on the group manifold associated to the Lie–Poisson algebra. In the definition of the physical system under consideration, there must be contained the information of which is the Lie group whose Lie algebra is realized as the Lie–Poisson algebra of the first-class constraints. The integrability conditions of the many-time Hamilton equations are just the statement that the constraints are first-class and that the canonical Hamiltonian is a first-class quantity (a Dirac observable).

What was not clarified in the previous papers is the Lagrangian counterpart of the many-time Hamilton equations. That is, which are the equations to be added to the original Euler–Lagrange equations and which are the integrability conditions for the resulting set of equations. A partial answer to these problems is already contained in the generalized Lie equations of Batalin and Vilkovisky[5] and in the hypothesis of existence and closure of the gauge algebra of the gauge transformations.[5,6]

In this paper we shall clarify this matter following the treatment of Refs. 1 and 2.

In Sec. II the previous works are reviewed, while in Sec. III the Hamiltonian many-time equations are analyzed. In Sec. IV we study the Lagrangian gauge algebra and its consequences. In Sec. V, the resulting Lagrangian many-time equations are studied. After the Conclusions (Sec. VI), an Appendix contains some commutators of vector fields.

II. GENERALITIES

Let us consider a system described by the coordinates q^i, $i=1,...,N$, and by a singular Lagrangian $L(q,\dot{q})$. Let the Hessian matrix $A_{ij}=\partial^2 L/\partial\dot{q}^i\partial\dot{q}^j$ have n null eigenvalues with associated null eigenvectors ${}_A\xi_0(q,\dot{q})$, $A=1,...,n$, $A_{ij}\,{}_A\xi^j_0=0$. Let the Lagrangian be quasiinvariant under the following n sets of gauge transformations:

$$\delta_A q^i=\sum_0^{J_A}{}_j\,{}^{(j)}\epsilon^A(t)\,{}_A\xi^i_{J_A-j}(q,\dot{q}),\qquad {}^{(j)}\epsilon^A(t)=\frac{d^j\epsilon^A(t)}{dt^j},\tag{1}$$

$$\delta_A L=\delta_A q^i L_i+\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{q}^i}\delta_A q^i\right)\equiv\frac{d}{dt}F_A(q,\dot{q},\epsilon^A(t))$$

$$F_a(q,\dot{q},\epsilon^A(t))=\sum_0^{J_A}{}_j\,{}^{(j)}\epsilon^A(t)\,{}_AF_{J_A-j}(q,\dot{q}).\tag{2}$$

Here $\epsilon^A(t)$ are n arbitrary functions of time and J_A are n integers that can be identified with the method of Ref. 2. The second Noether theorem[1,2] implies the existence of the following n Noether identities:

$$\frac{dG_A(q,\dot{q},\epsilon^A(t))}{dt}\equiv-\delta_A q^i L_i\doteq 0,$$

$$G_A(q,\dot{q},\epsilon^A(t))=\frac{\partial L}{\partial\dot{q}^i}\delta_A q^i-F_A$$

$$=\sum_0^{J_A}{}_j\,{}^{(j)}\epsilon^A(t)\,{}_AG_{J_A-j}(q,\dot{q}).\tag{3}$$

Here,

$$L_i=\frac{\partial L}{\partial q^i}-\frac{d}{dt}\frac{\partial L}{\partial\dot{q}^i}=-(A_{ij}\ddot{q}^j-\alpha_i)\doteq 0$$

are the Euler–Lagrange equations

$$\left(\alpha_i=\frac{\partial L}{\partial q^i}-\frac{\partial^2 L}{\partial\dot{q}^i\partial q^j}\dot{q}^j\right)$$

and $\doteq$ means evaluated on their solutions. One has for every A ${}_AG_0(q,\dot{q})\equiv 0$ and ${}_AG_{J_A-j}(q,\dot{q})\doteq 0$, $J=0,1,...,J_A-1$. Since the ${}_AG_{J_A-j}(q,\dot{q})$ are projectable to phase space, one gets the following situation ($p_i=\partial L/\partial\dot{q}^i$): for every A to the identically vanishing ${}_AG_0$ there corresponds a primary constraint

 0022-2488/90/092126-10$03.00

$$\bar{\phi}_A(q,p) = {}_A\bar{\phi}_0(q,p) \approx 0 \left({}_A G_0(q,\dot{q}) = \bar{\phi}_A \left(q, \frac{\partial L}{\partial \dot{q}} \right) \equiv 0 \right)$$

and then one has the secondary, tertiary..., constraints ${}_A G_{J_A - j}(q,\dot{q}) = {}_A\bar{\phi}_{J_A - j}(q,p) \approx 0$. Moreover, one has $\partial\bar{\phi}_A / \partial p_i = {}_A\xi^i_0$.

In Refs. 1 and 2 the theory was developed by choosing an orthonormal set of null eigenvectors ${}_A\xi^i_0(q,\dot{q})$: In general, this can be done only locally and has the consequence that the corresponding functional form of the primary $[\bar{\phi}_A$ with ${}_A\xi^i_0(q,\dot{q}) = \partial\bar{\phi}_A(q,p)/\partial p_i]$ and secondary $({}_A\bar{\phi}_{J_A - j})$ constraints implies that they satisfy (locally in phase space) an abelian Poisson algebra $(\{{}_A\bar{\phi}_{J_A - j}, {}_B\bar{\phi}_{J_B - k}\} = 0)$. Let us assume that there exist a unique global (in general not orthonormal) form ${}_A\hat{\xi}^i_0(q,\dot{q})$ of the null eigenvectors, such that the associated constraints ${}_A\hat{\bar{\phi}}_{J_A - j}(q,p)$ satisfy the following phase-space gauge algebra:

$$\left\{ {}_A\hat{\bar{\phi}}_{J_A - j}, {}_B\hat{\bar{\phi}}_{J_B - k} \right\} = \sum_c \sum_{0\,h}^{J_c} \hat{\bar{C}}^{(c|J_c - h)}_{(A|J_A - j)(B|J_B - k)}(q,p)\, {}_C\hat{\bar{\phi}}_{J_C - h}. \tag{4}$$

One has a true Lie algebra when the structure functions $\hat{\bar{C}}$'s are constants and a quasi-Lie algebra when the Batalin's conditions[7] are satisfied. Equation (4) will be taken as defining the global gauge algebra of the system and the equations ${}_A\hat{\bar{\phi}}_{J_A - j} \approx 0$ will globally define the final constraint manifold $\bar{\gamma}_F$ of phase space ($\bar{\gamma}$ is the constraint manifold defined only by the primary constraints $\hat{\bar{\phi}}_A$). Every other choice of the null eigenvectors will generate the same number of independent constraints: these new constraints ${}_A\bar{\phi}_{J_A - j}$ will be functions of the ${}_A\hat{\bar{\phi}}_{J_A - j}$'s, the equations ${}_A\phi_{J_A - j} \approx 0$ will define only a local domain of $\bar{\gamma}_F$ (except when ${}_A\bar{\phi}_{J_A - j} = {}_A\bar{K}_{J_A - j}\, {}_A\hat{\bar{\phi}}_{J_A - j}$ with the functions ${}_A\bar{K}_{J_A - j}$ never vanishing near $\bar{\gamma}$) and the counterpart of Eq. (4) will have different structure functions. Behind this freedom there is the theory of functions groups.[8] In particular, if one chooses an orthonormal basis of ${}_A\xi^i_0$ one gets a possible local Abelianization of the $\hat{\bar{\phi}}$'s (and again the demonstration of the existence of local abelianization can be formulated in terms of functions group[8] at least in the finite-dimensional case and is connected to the BRS approach[9]).

When a canonical Hamiltonian $\bar{H}_c = p_i q^i - L$ exists, the Dirac Hamiltonian has the following form:

$$\bar{H}_D = \bar{H}_c + \sum_A \lambda^A(t)\hat{\bar{\phi}}_A, \tag{5}$$

where the $\lambda^A(t)$'s are the arbitrary Dirac multipliers describing the gauge freedom associated to the primary constraints. As shown in Refs. 1 and 2, one has $\lambda^A(t) \doteq g^A(q,\dot{q})$, where the g^A's are a special functional form of the n velocity functions not projectable to phase space, due to the noninvertibility of the equations $p_i = \partial L/\partial \dot{q}^i$ when $\det(A_{ij}) = 0$.

Equation (4) only defines the structure functions of a real gauge algebra, when the singular Lagrangian L is such that the following equation holds:

$$\bar{H}_c = \bar{H}^F_c + \sum_A \sum_{0\,j}^{J_A - 1} {}^A\bar{d}^{J_A - j}(q,p)\, {}_A\hat{\bar{\phi}}_{J_A - j}. \tag{6}$$

This means that all the secondary first-class constraints ${}_A\hat{\bar{\phi}}_{J_A - j}, j \neq J_A$, are generators of phase-space gauge transformations, so that the Lagrangian gauge transformations are recovered in the following form:

$$\begin{aligned} \delta_A q^i &= \sum_{0\,j}^{J_A} {}^{(j)}\epsilon^A(t)\, {}_A\hat{\xi}^i_{J_A - j}(q,\dot{q}) \\ &= \sum_{0\,j}^{J_A} {}^{(j)}\epsilon^A(t)\{q^i, {}_A\hat{\bar{\phi}}_{J_A - j}\}|_{p = \partial L/\partial \dot{q}} \\ &= \{q^i, \hat{\bar{G}}_A\}|_{p = \partial L/\partial \dot{q}} = \overline{\delta_A q^i}|_{p = \partial L/\partial \dot{q}}. \end{aligned} \tag{7}$$

Equation (6) has to be added as a requirement on L: In the general case, Dirac[10] was not able to deduce it and assumed that all the secondary first-class constraints had to be added by hand to $\bar{H}_D$ with new arbitrary multipliers. Actually, ${}^A d^{J_A - j}(q,\dot{q}) = {}^A\bar{d}^{J_A - j}(q,p)|_{p = \partial L/\partial \dot{q}}$ are functions of $q^i, \dot{q}^i$ which remain undetermined from the Euler–Lagrange equations, because their solutions imply that these ${}^A\bar{d}^{J_A - j}$ are functions of the arbitrary velocity functions $g^A(q,\dot{q}) \doteq \lambda^A(t)$.

Moreover, in Eq. (6) the final canonical Hamiltonian $\bar{H}^F_c$ is a first-class quantity:[10]

$$\{\bar{H}^F_c, {}_A\hat{\bar{\phi}}_{J_A - j}\} = \sum_c \sum_{0\,h}^{J_c} \bar{C}^{(C|J_c - h)}_{0(A|J_A - j)}(q,p)\, {}_c\hat{\bar{\phi}}_{J_c - h}. \tag{8}$$

The extended Dirac Hamiltonian $\bar{H}_E$ can be obtained from Eqs. (5) and (6) by replacing the ${}^A\bar{d}^{J_A - j}(q,p)$ with arbitrary multipliers ${}^A\lambda^{J_A - j}(t)$. In this way, one gets a generalization of the gauge transformations, whose Lagrangian counterpart is to replace Eqs. (1) $(\delta_A q^i = \Sigma_{j=0}^{J_A} {}^{(j)}\epsilon^A(t)\, {}_A\hat{\xi}^i_{J_A - j}(q,\dot{q}))$ with the following ones:

$$\tilde{\delta}_A q^i = \sum_{0\,j}^{J_A} {}^A\eta^j(t)\, {}_A\hat{\xi}^i_{J_A - j}(q,\dot{q}), \tag{9}$$

with the ${}^A\eta^j(t)$ independent arbitrary functions of t.

As shown in Ref. 2 this implies that the quasiinvariance of Eqs. (2) is replaced by a weak quasiinvariance (quasiinvariance modulo the Euler–Lagrange equations and their consequences independent from the accelerations; this is the generic case when second-class constraints are present[1,2]):

$$\begin{aligned} \tilde{\delta}L &\equiv \frac{d}{dt}\left(\sum_{0\,j}^{J_A} {}^A\eta^j(t)\, {}_A F_{J_A - j}(q,\dot{q}) \right) \\ &\quad - \sum_{0\,j}^{J_A - 1} \left({}^A\eta^{j+1}(t) - {}^A\dot{\eta}^j(t) \right) {}_A\hat{G}_{J_A - j}(q,\dot{q}) \\ &\doteq \frac{d}{dt}\left(\sum_{0\,j}^{J_A} {}^A\eta^j(t)\, {}_A F_{J_A - j}(q,\dot{q}) \right). \end{aligned} \tag{10}$$

Equations (2) are recovered when ${}^A\eta^j(t) = {}^{(j)}\epsilon^A(t)$.

Another piece of information, which will be crucial for what follows, is given by the transformation properties[1,2] of the Euler–Lagrange equations under the gauge transformations of Eqs. (2):

$$\begin{aligned} \delta_A L_i &= J_i(\delta_A q) \\ &\equiv -\frac{\partial \delta_A q^k}{\partial q^i} L_k - \frac{d}{dt}\left(\frac{\partial \hat{\bar{G}}_A}{\partial q^i} + R_{ij}\delta_A q^j + A_{ij}\delta_A \dot{q}^j \right) \doteq 0, \end{aligned} \tag{11}$$

where the Noether identities, Eqs. (3), have been used. In Eqs. (11), R_{ij} is given by

$$R_{ij} = \frac{\partial^2 L}{\partial \dot{q}^i \partial q^j} - \frac{\partial^2 L}{\partial \dot{q}^j \partial q^i} \tag{12}$$

and $J_i(\delta q) \doteq 0$ are the Jacobi equations,[11] obtained from the second variation of the action $S = \int dt\, L$ and which vanish, when restricted to the extremals (solution of $L_i \doteq 0$), if $\delta q^i|_{L_i \doteq 0}$ are deviations between two neighboring extremals (i.e., they are Jacobi fields). Therefore, the infinitesimal gauge transformations $\delta_A q^i$ are Jacobi fields, as they should be by their definition itself, only when L is such that the following conditions are satisfied:

$$\frac{\partial \hat{G}_A}{\partial q^i} + R_{ij}\delta_A q^j + A_{ij}\delta_A \dot{q}^j \doteq 0. \tag{13}$$

While Eqs. (4), (6), and (8) plus the knowledge of the phase-space higher-order structure functions (see Ref. 9) include all the information about the phase-space gauge algebra, Eqs. (13) are a prerequisite for the existence of the Lagrangian gauge algebra, which will be discussed in Sec. IV.

Coming back to phase space, the Hamilton–Dirac equations read

$$\frac{d\overline{A}(q,p)}{dt} \doteq \{\overline{A},\overline{H}_D\} \approx \{\overline{A},\overline{H}^F_c\} + \sum_A \lambda^A(t)\{\overline{A},\hat{\phi}_A\} + \sum_A \sum_0^{J_A-1} {}_j{}^A\bar{d}^{J_A-j}(q,p)\{\overline{A},{}_A\hat{\phi}_{J_A-j}\}, \tag{14}$$

while the extended ones are

$$\frac{d\overline{A}(q,p)}{dt} \doteq \{\overline{A},\overline{H}_E\} \approx \{\overline{A},\overline{H}^F_c\} + \sum_A \sum_0^{J_A} {}_j{}^A\lambda^{J_A-j}(t)\{\overline{A},{}_A\hat{\phi}_{J_A-j}\}. \tag{15}$$

Here, ${}^A\lambda^0(t) = \lambda^A(t)$ and $\overline{A}(q,p)$ is a function on phase space.

Either Eqs. (14) or (15) cannot be solved without fixing the arbitrary multipliers λ's: This is a pregauge fixing, in the sense that only the restricted class of gauge-fixing constraints whose time constancy implies just these λ's are allowed, when one wishes to evaluate the Dirac brackets to get the symplectic structure of the reduced phase space. Instead, in the next section the many-time Hamilton equations will be introduced, avoiding this pregauge-fixing condition.

III. MANY-TIME HAMILTON EQUATIONS

Following Refs. 1 and 3 the many-time Hamilton equations will now be reviewed using the extended Hamilton–Dirac equations as a starting point.

When the structure functions of Eqs. (4) and (8) vanish (Abelian case), besides the ordinary time $t = \tau$, one introduces as many times ${}^A\tau^{J_A-j}$ as first-class constraints by means of the equations $d\,{}^A\tau^{J_A-j} = {}^A\lambda^{J_A-j}(\tau)d\tau$. The coordinates $q^i(t), p_i(t)$ are now considered as functions of all the times: $q^i = q^i(\tau, {}^A\tau^{J_A-j})$, $p_i = p_i(\tau, {}^A\tau^{J_A-j})$ (using the same symbols for the functions for the sake of simplicity) and standard Poisson brackets among them are defined at equal value of all the times. Equation (15) is then replaced by the following system of coupled many-time Hamilton equations:

$$\frac{\partial \overline{A}(q,p)}{\partial \tau} \doteq \{\overline{A},\overline{H}^F_c\} \equiv \overline{X}^F_c \overline{A}, \qquad \overline{X}^F_c = \{\cdot,\overline{H}^F_c\},$$
$$\frac{\partial \overline{A}(q,p)}{\partial\, {}^A\tau^{J_A-j}} \doteq \{\overline{A},{}_A\hat{\phi}_{J_A-j}\} \equiv {}_A\hat{X}_{J_A-j}\overline{A}, \qquad {}_A\hat{X}_{J_A-j} = \{\cdot,{}_A\hat{\phi}_{J_A-j}\}, \tag{16}$$
$$A = 1,\ldots,n, \qquad j = 0,1,\ldots,J_A.$$

Their integrability conditions are just Eqs. (4) and (8):

$$\{{}_A\hat{\phi}_{J_A-j},{}_B\hat{\phi}_{J_B-k}\} = \{\overline{H}^F_c,{}_A\hat{\phi}_{J_A-j}\} = 0.$$

The final constraint manifold $\bar{\gamma}_F$ is foliated with leaves, called gauge orbits. The vector fields ${}_A\hat{X}_{J_A-j}$, restricted to $\bar{\gamma}_F$, are tangent to the gauge orbits, while the vector field $\overline{X}^F_C$ connects the gauge orbits among themselves. The gauge orbits may be connected, simply connected, union of disconnected parts and so on: all these global topological properties must be assumed as given. Indeed they are hidden, at the Lagrangian level, in the global structure of the configuration space (of which the q^i are local coordinates) and in the global properties of the gauge transformations under which the singular Lagrangian is quasiinvariant. Once these properties are known, one can face the problem of reconstructing the gauge orbits from the knowledge of the gauge algebra of Eq. (4), which, in terms of the vector fields ${}_A\hat{X}_{J_A-j}$, becomes

$$[{}_A\hat{X}_{J_A-j},{}_B\hat{X}_{J_B-k}] = -\sum_c \sum_0^{J_c} \hat{C}^{(C|J_c-h)}_{(A|J_A-j)(B|J_B-k)}(q,p)\,{}_c\hat{X}_{J_c-h}. \tag{17}$$

In the Abelian case, when the gauge orbits are connected, simply connected the second half of Eqs. (16) are just the Lie equations of the Abelian Lie algebra: given a set of initial data on $\bar{\gamma}_F$, their integration allows to reconstruct the gauge orbit through that point. The gauge orbit turns out to be diffeomorphic to the group manifold of the corresponding Abelian covering Lie group, acting as a transformation group on $\bar{\gamma}_F$.

When the structure functions of Eqs. (4) and (8) are constant, they are the structure constants of a Lie algebra g [actually from Eq. (17) they are defined with a minus sign]. The gauge orbits are diffeomorphic to the group manifold of the corresponding covering Lie group G minus the part generated by $\overline{H}^F_c$. Let us consider[3] the Maurer–Cartan left-invariant one-forms $\vartheta^a = A^a_b(\tau^c)d\tau^c$ and their dual left-invariant vector fields $Y_a = B^b_a(\tau^c)(\partial/\partial\tau^b)$ on the group manifold of G. Here, τ^a is a set of coordinates on the group manifold, with the index a running over all the values $(A|J_A - j)$, $A = 1,\ldots,n; j = 0,1,\ldots,J_A$ and $\tau^0 = t$; $A^a_b B^b_c = \delta^a_c$ is implied by $i_{Ya}\vartheta^b = \delta^b_a$. If E is the identity in G, with coordinates $\tau^a = 0$, we have $Y_{a|_E} = e_a$, where e_a are the generators of g, and $\vartheta^a|_E = e^a$, with e^a the generators of the dual g^* of g (g and g^* are identified with $T_E G$ and $T_E G^*$, respectively). We have

$$[Y_a, Y_b] = -C^c_{ab} Y_c, \tag{18}$$
$$d\vartheta^a = \tfrac{1}{2} C^a_{bc}\vartheta^b \wedge \vartheta^c, \tag{19}$$

$$A^{\ e}_{a}\frac{\partial A^{\ d}_{b}}{\partial\tau^{e}}-A^{\ e}_{b}\frac{\partial A^{\ d}_{a}}{\partial\tau^{e}}=C^{c}_{ab}A^{\ d}_{c}. \tag{20}$$

Both Eq. (18) and (19) are called Maurer–Cartan equations.

To assign the Dirac multipliers $\lambda^a(t)$ is equivalent to assign some set of functions $\tau^A(\tau^0=t)$, i.e., a one-parameter subgroup, and the connection is given by

$$\lambda^{a}(t)=A^{\ a}_{b}(\tau^{c}(t))\frac{d\tau^{b}(t)}{dt}$$

$$\Rightarrow\lambda^{a}(t)dt=\vartheta^{a}|_{\tau^{c}=\tau^{c}(t)}. \tag{21}$$

In the Abelian case one recovers $d\tau^a=\lambda^a(t)dt$.

The many-time Hamilton equations are now

$$\left.\begin{aligned}Y_0\overline{A}(q,p)&\doteq\{\overline{A},\overline{H}^{F}_{c}=\hat{\phi}_0\}\\ Y_a\overline{A}(q,p)&\doteq\{\overline{A},\hat{\phi}_a\}\end{aligned}\right\}\Rightarrow\frac{\partial\overline{A}(q,p)}{\partial\tau^{a}}\doteq B^{\ b}_{a}(\tau^{c})\{\overline{A},\hat{\phi}_b\} \tag{22}$$

and their integrability conditions are just Eqs. (4), (8), and (18). Again these equations are Lie equations and the second set allows the reconstruction of the gauge orbits.

In the generic case of structure functions satisfying the conditions for generating a Batalin quasigroup,[7] one recovers Eqs. (22) with $\overline{B}^{\ b}_{a}(\tau^c,q,p)$ and Eq. (20) (and therefore the integrability conditions) are only satisfied when Eqs. (22) are restricted to $\bar{\gamma}_F$.

In the actual calculations the closed form of the $B^{\ b}_{a}(\tau^c)$ may be either not known or complicated. The only way to use Eqs. (22) is to find an Abelianization $\bar{\phi}_a$ *of the* $\hat{\phi}_a$'s and to use Eqs. (16). In general, this cannot be done globally on $\bar{\gamma}_F$ and one has to find different Abelianizations to cover all γ_F. In the Abelianized procedure one locally find Abelianization factors $\overline{B}^{\ b}_{a}(q,p)$ such that

$$\bar{\phi}_a=\overline{B}^{\ b}_{a}(q,p)\hat{\phi}_b\doteq B^{\ b}_{a}(\tau^{c})\hat{\phi}_b, \tag{23}$$

i.e., on the solutions of the Hamilton equations locally the $\overline{B}^{\ a}_{b}(q,p)$ become the $B^{\ b}_{a}(\tau^c)$. When reparametrization invariance ($H_c\equiv 0$) is present, the abelianization is the only possible approach at present: see the application to the Nambu string in Ref. 12.

IV. LAGRANGIAN GAUGE ALGEBRA

In this section, the Lagrangian counterpart of the vector field ${}_A\tilde{X}_{J_A-j}$ will be considered and the gauge algebra of Eq. (4) will be reformulated.

Since, in general, the Noether gauge transformations $\delta_A q^i$ depend not only on the q^i but also on the velocities $\dot{q}^i$, the gauge transformations of the velocities, $\delta_A\dot{q}^i=(d/dt)\delta_A q^i$, depend on the accelerations. As said in Refs. 1 and 2 this implies the necessity of formulating the theory in the infinite jet bundle with local coordinates $\{t,q^i,\overset{(1)}{q}{}^i,\overset{(2)}{q}{}^i,\ldots\}$. Its points are the equivalence classes $[\{t,c^i(t)\}]$ of all the curves $\{t,c^i,(t)\}$, which at the time t pass through the point $\{q^i\}$ and have there a point of tangency of infinite order: $d^k c^i_1(t)/dt^k=d^k c^i_2(t)/dt^k$ for every k. This implies $(d\overset{(k)}{q}{}^i-\overset{(k+1)}{q}{}^i dt)_{[\{t,c^i(t)\}]}=0$. The vector fields X_A, generating the gauge transformations, have to be expressed as Lie–Bäcklund vector fields:

$$X_A=\sum_0^\infty{}_k\frac{d^k\delta_A q_i}{dt^k}\frac{\partial}{\partial\overset{(k)}{q}{}^i},\qquad \overset{(0)}{q}{}^i=q^i,\ \overset{(1)}{q}{}^i=\dot{q}^i. \tag{24}$$

A Lie–Bäcklund transformation[13] is a tangent transformation of infinite order preserving the tangency of infinite order of two curves. Only when $\delta_A q^i=\delta_A q^i(q)$ we have an ordinary point transformation extended to the derivatives of arbitrary order and the space $\{t,q^i,\dot{q}^i\}$ closes upon itself under these transformations.

Before studying the X_A, let us consider the Lie–Bäcklund vector fields associated to the generalized gauge transformations of Eq. (9) (in what follows we will replace $\hat{\xi}^i$ with ξ^i)

$$\tilde{X}_A=\sum_0^\infty{}_k\frac{d^k\tilde{\delta}_A q^i}{dt^k}\frac{\partial}{\partial\overset{(k)}{q}{}^i}$$

$$=\sum_0^\infty{}_k\sum_0^{J_A}{}_j\sum_0^k{}_h\binom{k}{h}{}_A\overset{(h)}{\eta}{}^j(t)\,{}_A\overset{(k-h)}{\xi}{}^i_{J_A-j}(q,\dot{q})\frac{\partial}{\partial\overset{(k)}{q}{}^i}$$

$$=\sum_0^{J_A}{}_j\sum_0^\infty{}_h\,{}_A\overset{(h)}{\eta}{}^j(t)\,{}_A\tilde{X}_{J_A-j,h}, \tag{25}$$

$${}_A\tilde{X}_{J_A-j,h}=\sum_h^\infty{}_k\binom{k}{h}{}_A\overset{(k-h)}{\xi}{}^i_{J_A-j}(q,\dot{q})\frac{\partial}{\partial\overset{(k)}{q}{}^i}$$

$$={}_A\xi^i_{J_A-j}\frac{\partial}{\partial\overset{(h)}{q}{}^i}+\sum_{h+1}^\infty{}_k\binom{k}{h}{}_A\overset{(k-h)}{\xi}{}^i_{J_A-j}\frac{\partial}{\partial\overset{(k)}{q}{}^i}. \tag{26}$$

A relevant vector field is ${}_A\tilde{X}_{0,2}$:

$${}_A\tilde{X}_{0,2}={}_A\xi^i_0\frac{\partial}{\partial\ddot{q}^i}+\sum_3^\infty{}_k\binom{k}{h}{}_A\overset{(k-2)}{\xi}{}^i_0\frac{\partial}{\partial\overset{(k)}{q}{}^i}.$$

Indeed it is the generator of dynamical symmetries of the Euler–Lagrange equations, which are not Noether symmetries of the Lagrangian:

$${}_A\tilde{X}_{0,2}L_i\equiv 0. \tag{27}$$

In terms of the vector fields ${}_A\tilde{X}_{J_A-j,h}$ we get

$$X_A=\tilde{X}_A|_{A\eta^j(t)=\overset{(j)}{\epsilon}{}^A(t)}=\sum_0^{J_A}{}_j\sum_0^\infty{}_k\,\overset{(k+j)}{\epsilon}{}^A(t)\,{}_A\tilde{X}_{J_A-j,k}$$

$$=\sum_0^{J_A}{}_h\,\overset{(h)}{\epsilon}{}^A(t)\sum_0^h{}_j\,{}_A\tilde{X}_{J_A-j,h-j}$$

$$+\sum_{J_A+1}^\infty{}_h\,\overset{(h)}{\epsilon}{}^A(t)\sum_0^{J_A}{}_j\,{}_A\tilde{X}_{J_A-j,h-j}$$

$$=\sum_0^\infty{}_h\,\overset{(h)}{\epsilon}{}^A(t)\,{}_AX_h, \tag{28}$$

$${}_AX_h=\begin{cases}\sum_0^h{}_j\,{}_A\tilde{X}_{J_A-j,h-j}, & h\leqslant J_A,\\ \sum_0^{J_A}{}_j\,{}_A\tilde{X}_{J_A-j,h-j}, & h>J_A.\end{cases} \tag{29}$$

We shall be interested in the following ${}_AX_h$:

$$
{}_A X_0 = {}_A\tilde{X}_{J_A,0} = {}_A\xi^i_{J_A}\frac{\partial}{\partial q^i} + {}_A\dot{\xi}^i_{J_A}\frac{\partial}{\partial \dot{q}^i} + {}_A\ddot{\xi}^i_{J_A}\frac{\partial}{\partial \ddot{q}^i} + \sum_3^\infty {}_h\, {}_A\overset{(h)}{\xi}{}^i_{J_A}\frac{\partial}{\partial \overset{(h)}{q}{}^i},
$$

$$
{}_A X_1 = {}_A\tilde{X}_{J_A-1,0} + {}_A\tilde{X}_{J_A-1} = {}_A\xi^i_{J_A-1}\frac{\partial}{\partial q^i} + ({}_A\xi^i_{J_A} + {}_A\dot{\xi}^i_{J_A-1})\frac{\partial}{\partial \dot{q}^i} + (2{}_A\dot{\xi}^i_{J_A} + {}_A\ddot{\xi}^i_{J_A-1})\frac{\partial}{\partial \ddot{q}^i}
$$
$$
+ \sum_3^\infty {}_h\left(h\,{}_A\overset{(h-1)}{\xi}{}^i_{J_A} + {}_A\overset{(h)}{\xi}{}^i_{J_A-1}\right)\frac{\partial}{\partial \overset{(h)}{q}{}^i},
$$

$$
\vdots \tag{30}
$$

$$
{}_A X_k = \sum_0^k {}_j\, {}_A\tilde{X}_{J_A-j,k-j} = {}_A\xi^i_{J_A-k}\frac{\partial}{\partial q^i} + ({}_A\xi^i_{J_A-k+1} + {}_A\dot{\xi}^i_{J_A-k})\frac{\partial}{\partial \dot{q}^i}
$$
$$
+ ({}_A\xi^i_{J_A-k+2} + 2{}_A\dot{\xi}^i_{J_A-k+1} + {}_A\ddot{\xi}^i_{J_A-k})\frac{\partial}{\partial \ddot{q}^i} + \left(\sum_3^k {}_h \sum_{k-h}^k {}_j + \sum_{k+1}^\infty {}_h \sum_0^k {}_j\right)\binom{h}{k-j}\, {}_A\overset{(h-k+j)}{\xi}{}^i_{J_A-j}\frac{\partial}{\partial \overset{(h)}{q}{}^i},
$$
$$
2 \leqslant k \leqslant J_A,
$$

$$
\vdots
$$

$$
{}_A X_{J_A+1} = \sum_0^{J_A} {}_j\, {}_A\tilde{X}_{J_A-j,J_A+1-j} = Z_A + ({}_A\xi^i_1 + 2{}_A\dot{\xi}^i_0)\frac{\partial}{\partial \ddot{q}^i}
$$
$$
+ \left(\sum_3^{J_A+1} {}_h \sum_{J_A+1-h}^{J_A} {}_j + \sum_{J_A+2}^\infty {}_h \sum_0^{J_A} {}_j\right)\binom{h}{J_A+1-j}\, {}_A\overset{(h+j-J_A+1)}{\xi}{}^i_{J_A-j}\frac{\partial}{\partial \overset{(h)}{q}{}^i},
$$

$$
{}_A X_{J_A+2} = \sum_0^{J_A} {}_j\, {}_A\tilde{X}_{J_A-j,J_A+2-j} = {}_A\tilde{X}_{0,2} + \left(\sum_3^{J_A+2} {}_h \sum_{J_A+2-h}^{J_A} {}_j + \sum_{J_A+3}^\infty {}_h \sum_0^{J_A} {}_j\right)\binom{h}{J_A+2-j}\, {}_A\overset{(h+j-J_A-2)}{\xi}{}^i_{J_A-j}\frac{\partial}{\partial \overset{(h)}{q}{}^i},
$$

with

$$
Z_A = {}_A\xi^i_0\frac{\partial}{\partial \dot{q}^i}, \tag{31}
$$

because only they act on the subspace $\{q^i,\dot{q}^i,\ddot{q}^i\}$ needed for the Euler–Lagrange equations. The vector fields Z_A play an important role in the velocity space approach,[1] where they are connected to extra gauge freedoms of that formalism, and project to zero in phase space $[Z_A \to {}_A\xi^i_0 A_{ij}(\partial/\partial p_j) = 0]$.

The following notation will be used to indicate the vector fields of Eqs. (30), when truncated to their action on the $\{q^i,\dot{q}^i\}$ and $\{q^i,\dot{q}^i,\ddot{q}^i\}$ subspaces, respectively:

$$
{}_A T'_{J_A} = {}_A X_0^{T_1} = {}_A\tilde{X}^{T_1}_{J_A,0} = {}_A\xi^i_{J_A}\frac{\partial}{\partial q^i} + {}_A\dot{\xi}^i_{J_A}\frac{\partial}{\partial \dot{q}^i},
$$
$$
{}_A T'_{J_A-j} = {}_A X_j^{T_1} = {}_A\tilde{X}^{T_1}_{J_A-j,0} + {}_A\tilde{X}^{T_1}_{J_A-j+1,1} = {}_A\xi^i_{J_A-j}\frac{\partial}{\partial q^i} + ({}_A\xi^i_{J_A-j+1} + {}_A\dot{\xi}^i_{J_A-j})\frac{\partial}{\partial \dot{q}^i}, \quad j = 1,\ldots,J_A, \tag{32}
$$

$$
Z_A = {}_A X^{T_1}_{J_A+1} = {}_A\tilde{X}^{T_1}_{0,1} = {}_A\xi^i_0\frac{\partial}{\partial \dot{q}^i},
$$
$$
{}_A T_{J_A} = {}_A X_0^{T_2} = {}_A\tilde{X}^{T_2}_{J_A,0} = {}_A T'_{J_A} + {}_A\ddot{\xi}^i_{J_A}\frac{\partial}{\partial \ddot{q}^i},
$$
$$
{}_A T_{J_A-1} = {}_A X_1^{T_2} = {}_A\tilde{X}^{T_2}_{J_A-1,0} + {}_A\tilde{X}^{Y_2}_{J_A,1} = {}_A T'_{J_A-1} + (2{}_A\dot{\xi}^i_{J_A} + {}_A\ddot{\xi}^i_{J_A-1})\frac{\partial}{\partial \ddot{q}^i},
$$
$$
{}_A T_{J_A-j} = {}_A X_j^{T_2} = {}_A\tilde{X}^{T_2}_{J_A-j,0} + {}_A\tilde{X}^{T_2}_{J_A-j+1,1} + {}_A\tilde{X}^{T_2}_{J_A-j+2,2} = {}_A T'_{J_A-j} + ({}_A\xi^i_{J_A-j+2} + 2{}_A\dot{\xi}^i_{J_A-j+1} + {}_A\ddot{\xi}^i_{J_A-j})\frac{\partial}{\partial \ddot{q}^i},
$$
$$
Z'_A = {}_A X^{T_2}_{J_A+1} = {}_A\tilde{X}^{T_2}_{0,1} + {}_A\tilde{X}^{T_2}_{1,2} = Z_A + ({}_A\xi^i_1 + 2{}_A\dot{\xi}^i_0)\frac{\partial}{\partial \ddot{q}^i},
$$
$$
{}_A\tilde{X}^{T_2}_{0,2} = {}_A\xi^i_0\frac{\partial}{\partial \ddot{q}^i}, \quad j = 2,\ldots,J_A \tag{33}
$$

When Eq. (13) holds Eqs. (32) are projectable to phase space as shown in Ref. 1. Indeed from Eqs. (28) and (7) one has (remembering that $Z_A \to 0$)

$$X_A^{T_1} = \sum_0^{J_A} {}_h \overset{(h)}{\epsilon}{}^A(t)\, {}_A X_h^{T_1} = \delta_A q^i \frac{\partial}{\partial q^i} + \delta_A \dot{q}^i \frac{\partial}{\partial \dot{q}^i}$$
$$\mapsto \frac{\partial \hat{\bar{G}}_A}{\partial p_i}\frac{\partial}{\partial q^i}\bigg|_p + \left(\frac{\partial^2 L}{\partial \dot{q}^i \partial q^j}\frac{\partial \hat{\bar{G}}_A}{\partial p_j} + A_{ij}\delta_A \dot{q}^j\right)\frac{\partial}{\partial p_i}$$
$$= \{\cdot, \hat{\bar{G}}_A\} + \left(\frac{\partial G_A}{\partial q^i} + R_{ij}\delta_A q^j + A_{ij}\delta_A \dot{q}^j\right)\frac{\partial}{\partial p_i}$$
$$\doteq \sum_0^{J_A} {}_h \overset{(h)}{\epsilon}{}^A(t)\, {}_A \hat{\bar{X}}_{J_A - h}, \tag{34}$$

so that

$${}_A T'_{J_A - j} \mapsto {}_A \hat{\bar{X}}_{J_A - j} = \{\cdot, {}_A \hat{\bar{\phi}}_{J_A - j}\}, \quad j = 0,1,\ldots,J_A$$
$$Z_A \mapsto 0. \tag{35}$$

With Eqs. (32), Eq. (11) may be rewritten as $\delta_A L_i = X_A L_i = J_i(\delta_A q) \doteq 0$ so that we get from them

$${}_A T_{J_A - j} L_i \doteq 0, \quad j = 0,1,\ldots,J_A$$
$$Z'_A L_i \doteq 0, \quad {}_A \tilde{X}^{T_2}_{0,2} L_i \equiv 0, \tag{36}$$

where the last equation is Eq. (27). Equations (33) carry, at the Lagrangian level, the same information of Eq. (8), i.e., that the time evolution is compatible with the gauge transformations, at least for infinitesimal deplacements of both kinds. As it is clear from Eq. (11), Eq. (13), considered as the consistency relations for the definition of the infinitesimal gauge transformations, lie behind Eqs. (33).

For the commutator of two Lie–Bäcklund vector field of Eq. (24) we get[13]

$$[X_A, X_B] = \sum_0^{\infty} {}_k \frac{d^k \delta_{[A,B]} q^i}{dt^k} \frac{\partial}{\partial \overset{(k)}{q}{}^i}, \tag{37}$$

$$\delta_{[A,B]} q^i = X_A^{T_1}(\delta_B q^i) - X_B^{T_1}(\delta_A q^i)$$
$$= \delta_A q^k \frac{\partial \delta_B q^i}{\partial q^k} - \delta_B q^k \frac{\partial \delta_A q^i}{\partial q^k}$$
$$+ \delta_A \dot{q}^k \frac{\partial \delta_B q^i}{\partial \dot{q}^k} - \delta_B \dot{q}^k \frac{\partial \delta_A q^i}{\partial \dot{q}^k}, \tag{38}$$

where use has been made of the following result:

$$\left[X_A, \frac{d}{dt}\right] = 0. \tag{39}$$

By using the Noether identities of Eqs. (2), it has been shown in Ref. 1 that we get

$$\delta_{[A,B]} L \equiv \frac{dF_{[A,B]}}{dt} + J_i(\delta_B q)\delta_A q^i - J_i(\delta_A q)\delta_B q^i, \tag{40}$$

with

$$F_{[A,B]} = X_B(G_A) - X_A(G_B) + \frac{\partial L}{\partial \dot{q}^i}\delta_{[A,B]} q^i$$
$$= -G_{[A,B]} + \frac{\partial L}{\partial \dot{q}^i}\delta_{[A,B]} q^i. \tag{41}$$

One gets the new Noether identities

$$\frac{dG_{[A,B]}}{dt} \equiv -\delta_{[A,B]} q^i L_i + \delta_A q^i J_i(\delta_B q)$$
$$- \delta_B q^i J_i(\delta_A q) \doteq 0, \tag{42}$$

where $G_{[A,B]}$ projects in phase space to $\{\bar{G}_A, \bar{G}_B\}$. Therefore, the phase-space gauge algebra hypothesis, Eqs. (4), implies that $\delta_{[A,B]} q^i$ must be an infinitesimal gauge transformation. This implies $J_i(\delta_{[A,B]} q^i) \doteq 0$ and in Ref. 1 it is shown that this is the condition for getting a quasiinvariance of L under the transformations $\delta_{[D,[A,B]]} q^i$.

Equation (38) shows that, when the infinitesimal gauge transformations are velocity dependent, $\delta_{[A,B]} q^i$ depends upon the accelerations. Therefore, it may be decomposed in a part that depends only on $q^i, \dot{q}^i$ and a part proportional to the Euler–Lagrange equations L_i. The $q^i, \dot{q}^i$ dependent part must moreover be proportional to a certain subset of the ${}_c \xi^i_{J_c - w}(q,\dot{q})$, so that we have a generalized infinitesimal gauge transformation in accord with Eq. (40) and the discussion following them. Therefore, we get

$$\delta_{[A,B]} q^i$$
$$= \sum_0^{J_A} {}_j \overset{(j)}{\epsilon}{}^A(t) \sum_0^{J_B} {}_k \overset{(k)}{\zeta}{}^B(t)$$
$$\times \sum_c \sum_0^{J_c} {}_w C^{(c|J_c - w)}_{(A|J_A - j)(B|J_B - k)}(q,\dot{q})$$
$$\times {}_c \xi^i_{J_c - w}(q,\dot{q})$$
$$+ U^{ij}_{[A,B]}(q,\dot{q}) L_j \tag{43}$$

so that, when the gauge transformations are velocity dependent, we have what is called an "open gauge algebra."[14] The structure functions of Eq. (43) projects on phase space to the structure functions of Eq. (4).

We can now state which are the most general conditions on L for the existence of an open Lagrangian gauge algebra, when we restrict ourselves to the relevant $(q^i, \dot{q}^i \ddot{q}^i)$ subspace. The Lagrangian must be quasiinvariant under as many sets of gauge transformations $\delta_A q^i$ as null eigenvalues of its Hessian matrix, and the functions ${}_A \xi^i_{J_A - j}(q,\dot{q})$ must be such that the vector fields of Eqs. (33) satisfy the following algebra:

$$[{}_A T_{J_A - j}, {}_B T_{J_B - k}] = -\sum_c \sum_0^{J_c} {}_w C^{(c|J_c - w)}_{(A|J_A - j)(B|J_B - k)}(q,\dot{q})$$
$$\times {}_c T_{J_c - w} + \sum_c D^c_{(A|J_A - j)(B|J_B - k)}(q,\dot{q})$$
$$\times Z'_c + \sum_c E^c_{(A|J_A - j)(B|J_B - k)}(q,\dot{q})$$
$$\times {}_A \tilde{X}^{T_2}_{0,2} + U_{(A|J_A - j)(B|J_B - k)}, \tag{44}$$

where the vector field $U_{(A|J_A - j)(B|J_B - j)}$ vanishes by using $L_i \doteq 0$ and its time derivatives. By using the results of the Appendix truncated to the $(q^i, \dot{q}^i, \ddot{q}^i)$ subspace, the remaining part of the algebra must be

$$[{}_A T_{J_A - j}, Z'_B] = \sum_c D^c_{1(A|J_A - j)B}(q,\dot{q}) Z'_c$$

$$+\sum_c E^c_{1(A|J_A-j)B}(q,\dot q)_c\tilde X^{T_2}_{0,2}$$
$$+U_{1(A|J_A-j)B},$$
$$[{}_AT_{J_A-j},{}_B\tilde X^{T_2}_{0,2}]=\sum_c D^c_{2(A|J_A-j)B}(q,\dot q)Z'_c$$
$$+\sum_c E^c_{2(A|J_A-j)B}(q,\dot q)_c\tilde X^{T_2}_{0,2}$$
$$+U_{2(A|J_A-j)B},$$
$$[Z'_A,Z'_B]=\sum_c D^c_{3AB}(q,\dot q)Z'_c$$
$$+\sum_c E^c_{3AB}(q,\dot q)_c\tilde X^{T_2}_{0,2}+U_{3AB},$$
$$[Z'_A,{}_B\tilde X^{T_2}_{0,2}]=\sum_c E^c_{4AB}(q,\dot q)_c\tilde X^{T_2}_{0,2}+U_{4AB},$$
$$[{}_A\tilde X^{T_2}_{0,2},{}_B\tilde X^{T_2}_{0,2}]=0, \tag{45}$$

with all the U's satisfying $U\doteq 0$.

When restricted to the $(q^i,\dot q^i)$ subspace, Eqs. (44) and (45) become

$$[{}_AT'_{J_A-j},{}_BT'_{J_B-k}]$$
$$=-\sum_c\sum_0^{J_c}{}_w C^{(c|J_c-w)}_{(A|J_A-j)(B|J_B-k)}(q,\dot q)_cT'_{J_c-w}$$
$$+\sum_c D^c_{(A|J_A-j)(B|J_B-k)}(q,\dot q)Z_c$$
$$+U'_{(A|J_A-j)(B|J_B-k)}, \tag{46}$$
$$[{}_AT'_{J_A-j},Z_B]=\sum_c D^c_{1(A|J_A-j)B}(q,\dot q)Z_c$$
$$+U'_{1(A|J_A-j)B},$$
$$[Z_A,Z_B]=\sum_c D^c_{3AB}(q,\dot q)Z_c+U'_{3AB}.$$

While the first of Eqs. (46) projects to Eq. (17) in phase space, the other two project to zero. Ths explains why there can be no term in ${}_cT_{J_c-w}$ in the first of Eqs. (45). The terms in Z_c in the first of Eqs. (46) is a slight generalization of the known open gauge algebras, which cannot be excluded *a priori*. While the ${}_AT'_{J_A-j}$ are the vector fields generating the Noether transformations of Eqs. (1) under which L is quasiinvariant and the ${}_AT_{J_A-j}$ and $Z'_A,{}_A\tilde X^{T_2}_{0,2}$ are the generators of the dynamical symmetries of the Euler–Lagrange equations, see Eqs. (36), Z_A is not the generator of a Noether gauge transformation. From Eqs. (32), (28), (24), and the Noether identities (2), we get

$${}_AT'_{J_A}L\equiv{}_A\dot F_{J_A},\quad \delta_Aq^i=\sum_0^{J_A}{}_j{}^{(j)}\epsilon^A(t){}_AT'_{J_A-j}q^i,$$
$${}_AT'_{J_A-j}L\equiv{}_A\dot F_{J_A-j}+{}_AF_{J_A-j+1},\quad j=1,\dots,J_A, \tag{47}$$
$$Z_AL\equiv{}_AF_0,$$

so that $Z_AL\equiv 0$ only when ${}_AF_0\equiv 0$. Instead, in the velocity space[1] the counterparts of the Z_A are generators of extra gauge transformations of that first-order formalism, in which there is no room for the dynamical symmetries of the Euler–Lagrange equations generated by the ${}_A\tilde X^{T_2}_{0,2}$, see Eqs. (27) and (36). While at the Lagrangian level ${}_A\tilde X^{T_2}_{0,2}$ is responsible for the arbitrariness of the accelerations in the directions of the null eigenvectors of the Hessian matrix, the velocity space counterpart of Z_A is responsible for an analogous arbitrariness.

Both in the phase space and Lagrangian formalism the commutator of any order of the gauge transformations must again be a gauge transformation and the iterations of Eqs. (4) and (43), respectively, must be consistent among themselves. This is necessary to reconstruct perturbatively the gauge part of the trajectories (i.e., the gauge orbits) in large: according to the Frobenius theorem, the vector fields ${}_A\bar X_{J_A-j}$ and ${}_AT_{J_A-j}$ (plus Z'_A and ${}_A\tilde X^{T_2}_{0,2}$) must form an involutive distribution (in the Lagrangian case with an open gauge algebra this is true modulo $L_i\doteq 0$).

The final piece of information about the gauge algebra are the higher-order structure functions, which exist when the structure functions $\bar C$ and C are not constant. In the phase-space approach they are evaluated starting from the Jacobi identity for the Poisson brackets of the constraints as shown in Ref. 9. In the Lagrangian formalism there are already the extra structure functions $U^{ij}_{[A,B]}(q,\dot q)$ of Eq. (43). Then, as shown in Ref. 6, one again begins the research of the higher structure functions from the Jacobi identity for the commutator of three ${}_AT'_{J_A-j}$, following a pattern similar to the one of phase space.

V. LAGRANGIAN MANY-TIME EQUATIONS

From Eqs. (28) and (39) we get

$$0=\left[X_A,\frac{d}{dt}\right]$$
$$=\epsilon^A(t)\left[{}_AX_0,\frac{d}{dt}\right]$$
$$+\sum_1^\infty{}_h{}^{(h)}\epsilon^A(t)\left(\left[{}_AX_h,\frac{d}{dt}\right]-{}_AX_{h-1}\right)$$
$$\Rightarrow\begin{cases}\left[{}_AX_0,\dfrac{d}{dt}\right]=0,\\ \left[{}_AX_h,\dfrac{d}{dt}\right]={}_AX_{h-1},\quad h>0.\end{cases} \tag{48}$$

For the vector fields of Eqs. (33), this implies by using Eqs. (30), (33), and (26) that,

$$\left[{}_AX_0,\frac{d}{dt}\right]^{T_2}=\left[{}_A\tilde X_{J_A,0},\frac{d}{dt}\right]^{T_2}=0,$$
$$\left[{}_AX_1,\frac{d}{dt}\right]^{T_2}=\left[{}_A\tilde X_{J_A-1,0}+{}_A\tilde X_{J_A,1},\frac{d}{dt}\right]^{T_2}={}_AT_{J_A},$$
$$\left[{}_AX_j,\frac{d}{dt}\right]^{T_2}$$
$$=\left[\sum_{j-2}^{j}{}_k{}_A\tilde X_{J_A-k,j-k}+\sum_0^{j-3}{}_k{}_A\tilde X_{J_A-k,j-k},\frac{d}{dt}\right]^{T_2}$$
$$={}_AT_{J_A-j+1},\quad j=2,\dots,J_A, \tag{49}$$

$$\left[{}_A X_{J_A+1}, \frac{d}{dt}\right]^{T_2}$$
$$= \left[{}_A \tilde{X}_{0,1} + {}_A \tilde{X}_{1,2} + \sum_0^{J_A-2} {}_k\, {}_A \tilde{X}_{J_A-k,J_A+1-k}, \frac{d}{dt}\right]^{T_2}$$
$$= {}_A T_0,$$
$$\left[{}_A X_{J_A+2}, \frac{d}{dt}\right]^{T_2}$$
$$= \left[{}_A \tilde{X}_{0,2} + \sum_0^{J_A-1} {}_k\, {}_A \tilde{X}_{J_A-k,J_A+2-k}, \frac{d}{dt}\right]^{T_2}$$
$$= Z'_A,$$
$$\left[{}_A X_{J_A+3}, \frac{d}{dt}\right]^{T_2} = {}_A \tilde{X}_{0,2}^{T_2}.$$

The restriction of Eqs. (49) to the subspace $(q^i, \dot{q}^i)$ is

$$\left[{}_A T_{J_A}, \frac{d}{dt}\right]^{T_1} = 0$$
$$\left[{}_A T_{J_A-1}, \frac{d}{dt}\right]^{T_1} = {}_A T_{J_A}^{T_1} = {}_A T'_{J_A},$$
$$\left[{}_A T_{J_A-j}, \frac{d}{dt}\right]^{T_1} = {}_A T_{J_A-j+1}^{T_1} = {}_A T'_{J_A-j+1}, \quad j = 2,\ldots,J_A, \tag{50}$$
$$\left[Z'_A, \frac{d}{dt}\right]^{T_1} = {}_A T_0^{T_1} = {}_A T'_0,$$
$$\left[{}_A \tilde{X}_{0,2}^{T_2}, \frac{d}{dt}\right]^{T_1} = Z_A'^{T_1} = Z_A.$$

To study the Lagrangian many-time equations, let us begin as in phase space with the abelian case, which now means the vanishing of the structure functions C in Eqs. (44). If the functions $q^i(t)$ are considered as functions $q^i(t, {}^A\tau^{J_A-j})$ (for the sake of simplicity we use the same symbol for the new functions) with as many extra "times" ${}^A\tau^{J_A-j}$ as vector fields ${}_A T_{J_A-j}$ [we are working in the $(q^i, \dot{q}^i, \ddot{q}^i)$ subspace], the equations of motion are

$$L_i \doteq 0,$$
$$\frac{\partial q^i}{\partial\, {}^A\tau^{J_A-j}} \doteq {}_A T_{J_A-j} q^i = {}_A \xi^i_{J_A-j}(q_i \dot{q}), \quad \begin{matrix} A = 1,\ldots,n, \\ j = 0,1,\ldots,J_A, \end{matrix} \tag{51}$$

where the second set are the Lagrangian many-time equations. Now $\dot{q}^i$ and $\ddot{q}^i$ are interpreted as $(\partial/\partial t)q^i$ and $(\partial^2/\partial t^2)q^i$, respectively.

Equations (51) are integrable due to Eqs. (44) and (36) by using $L_i \doteq 0$, i.e., they are integrable on the solutions of the ordinary Euler–Lagrange equations.

Equations (49) and (50) are used to check that Eqs. (51), defined in terms of the vector fields ${}_A T_{J_A-j}$, are consistent

$$\frac{\partial^2 q^i}{\partial\, {}^A\tau^{J_A-j}\, \partial t} \doteq \frac{\partial}{\partial t}\, {}_A T_{J_A-j} q^i$$
$$= {}_A T_{J_A-j} \dot{q}^i - {}_A T_{J_A-j+1} q^i$$
$$= {}_A \dot{\xi}^i_{J_A-j} + {}_A \xi^i_{J_A-j+1} - {}_A \xi^i_{J_A-j+1}$$
$$= {}_A \dot{\xi}^i_{J_A-j}, \tag{52}$$

$$\frac{\partial^3 q^i}{\partial\, {}^A\tau^{J_A-j}\, \partial^2 t} \doteq \frac{\partial^2}{\partial t^2}\, {}_A T_{J_A-j} q^i$$
$$= \frac{\partial}{\partial t}({}_A T_{J_A-j} \dot{q}^i - {}_A T_{J_A-j+1} q^i)$$
$$= {}_A T_{J_A-j} \ddot{q}^i - 2\, {}_A T_{J_A-j+1} \dot{q}^i + {}_A T_{J_A-j+2} q^i$$
$$= {}_A \ddot{\xi}^i_{J_A-j} + 2\, {}_A \dot{\xi}^i_{J_A-j+1} + {}_A \xi^i_{J_A-j+2}$$
$$- 2({}_A \dot{\xi}^i_{J_A-j+1} + {}_A \xi^i_{J_A-j+2})$$
$$+ {}_A \xi^i_{J_A-j+2} = {}_A \ddot{\xi}^i_{J_A-j},$$

where Eqs. (30) have been used.

When the structure functions C of Eqs. (44), are the structure constants of a Lie algebra, by using the vector fields Y_a of Eqs. (18) (with the value $a = 0$ excluded) we have the following set of integrable equations:

$$L_i \doteq 0,$$
$$Y_a q^i \doteq T_a q^i \Rightarrow \frac{\partial q^i}{\partial \tau^a} \doteq B^b_a(\tau^c) T_b q^i. \tag{53}$$

In the generic case of structure functions C in Eqs. (44), one could translate at the Lagrangian level the conditions for the Batalin quasigroup and obtain Eqs. (53), but with a $B^b_a = B^b_a(\tau^c, q, \dot{q})$.

Actually, as in phase space, one looks for an Abelianization (with respect to the structure functions C) of Eqs. (44), i.e., for new vector fields $\breve{T}_a = R^b_a(q, \dot{q}) T_b \doteq B^b_a T_b$ which allow the local use of Eqs. (51). Since the Abelianization procedure amounts to make a local choice for the gauge variables, we rediscover the generalized Lie equations of Batalin–Vilkovisky[5] as an Abelianization of the many-time Lagrangian equations.

VI. CONCLUSIONS

We have developed the general structure of the many-time Lagrangian and Hamiltonian equations for a finite-dimensional system with an arbitrary number of first-class constraints, when all the possible existing secondary constraints are generators of gauge transformations. This requires the existence of a gauge algebra, which, in general, is open at the Lagrangian level.

When second-class constraints are present, nothing changes of the previous considerations. Indeed the second-class constraints $\bar{\psi}_k \approx 0$ are preserved in time by construction, and from Eqs. (22) we get that $Y_a \bar{\psi}_k \doteq \{\bar{\psi}_k, \hat{\phi}_a\} \approx 0$ due to the definition of first-class constraints. Therefore, the second-class constraints are preserved also in the extra "times" τ^a.

We have also shown the existence of the vector field Z'_A and ${}_A \tilde{X}_{0,2}$, connected to dynamical symmetries of the Euler–Lagrange equations. Both of them are not present in the phase space approach, while in the velocity space one there are extra gauge freedoms generated by vector fields $\tilde{Z}_A$, which are the counterpart of the Z'_A, but which introduce

the same kind of arbitrariness which the ${}_A\tilde{X}_{0,2}$ introduce at the Lagrangian level.

APPENDIX: SOME COMMUTATORS

By taking into account that ${}_A\xi^i_{J_A-j} = {}_A\xi^i_{J_A-j}(q,\dot{q})$, we get the following expressions for the commutator of two vector fields ${}_A\tilde{X}_{J_A-j,k}$:

$$
\begin{aligned}
&[{}_A\tilde{X}_{J_A-j,r},{}_B\tilde{X}_{J_B-k,s}] \\
&\quad=\sum_{r+s-1}^{\infty}{}^{v}\left[\binom{v}{s}\sum_{r}^{v-s+1}{}_{u}\binom{u}{r}_A{}^{(u-r)n}_{\quad J_A-j}\frac{\partial_B{}^{(v-s)i}_{\quad J_B-k}}{\partial^{(u)}q^n}\right. \\
&\qquad\left.-\binom{v}{r}\sum_{s}^{v-r+1}{}_{u}\binom{u}{s}_B{}^{(u-s)n}_{\quad J_B-k}\frac{\partial_A{}^{(v-r)i}_{\quad J_A-j}}{\partial^{(u)}q^n}\right]\frac{\partial}{\partial^{(v)}q^i}.
\end{aligned}
$$

Then we get

$$
\begin{aligned}
&[{}_AT_{J_A-j},{}_BT_{J_B-k}] \\
&\quad=[{}_A\tilde{X}_{J_A-j,0}+{}_A\tilde{X}_{J_A-j+1,1}+{}_A\tilde{X}_{J_A-j+2,2},{}_B\tilde{X}_{J_B-k,0}+{}_B\tilde{X}_{J_B-k+1,1}+{}_B\tilde{X}_{J_B-k+2,2}]^{T_2} \\
&\quad=\left[\sum_0^1 u\left({}_A\xi^{(u)n}_{\quad J_A-j}\frac{\partial_B\xi^i_{J_B-k}}{\partial^{(u)}q^n}-{}_B\xi^{(u)n}_{\quad J_B-k}\frac{\partial_A\xi^i_{J_A-j}}{\partial^{(u)}q^n}\right)+{}_A\xi^n_{J_A-j+1}\frac{\partial_B\xi^i_{J_B-k}}{\partial\dot{q}^n}-{}_B\xi^n_{J_B-k+1}\frac{\partial_A\xi^i_{J_A-j}}{\partial\dot{q}^n}\right]\frac{\partial}{\partial q^i} \\
&\qquad+\left[\sum_0^2 u\left({}_A\xi^{(u)n}_{\quad J_A-j}\frac{\partial_B\dot{\xi}^i_{J_B-k}}{\partial^{(u)}q^n}-{}_B\xi^{(u)n}_{\quad J_B-k}\frac{\partial_A\dot{\xi}^i_{J_A-j}}{\partial^{(u)}q^n}\right)+\sum_0^1 u\left({}_A\xi^{(u)n}_{\quad J_A-j}\frac{\partial_B\xi^i_{J_B-k+1}}{\partial^{(u)}q^n}-{}_B\xi^{(u)n}_{\quad J_B-k}\frac{\partial_A\xi^i_{J_A-j+1}}{\partial^{(u)}q^n}\right)\right. \\
&\qquad+\sum_1^2 u\left({}_A\xi^{(u-1)n}_{\quad J_A-j+1}\frac{\partial_B\dot{\xi}^i_{J_B-k}}{\partial^{(u)}q^n}-{}_B\xi^{(u-1)n}_{\quad J_B-k+1}\frac{\partial_A\dot{\xi}^i_{J_A-j}}{\partial^{(u)}q^n}\right)+{}_A\xi^n_{J_A-j+1}\frac{\partial_B\xi^i_{J_B-k+1}}{\partial\dot{q}^n}-{}_B\xi^n_{J_B-k+1}\frac{\partial_A\xi^i_{J_A-j+1}}{\partial\dot{q}^n} \\
&\qquad\left.+{}_A\xi^n_{J_A-j+2}\frac{\partial_B\dot{\xi}^i_{J_B-k}}{\partial\ddot{q}^n}-{}_B\xi^n_{J_B-k+2}\frac{\partial_A\dot{\xi}^i_{J_A-j}}{\partial\ddot{q}^n}\right]\frac{\partial}{\partial\dot{q}^i}+\left[\sum_0^3 u\left({}_A\xi^{(u)n}_{\quad J_A-j}\frac{\partial_B\ddot{\xi}^i_{J_B-k}}{\partial^{(u)}q^n}-{}_B\xi^{(u)n}_{\quad J_B-k}\frac{\partial_A\ddot{\xi}^i_{J_A-j}}{\partial^{(u)}q^n}\right)\right. \\
&\qquad+2\sum_0^2 u\left({}_A\xi^{(u)n}_{\quad J_A-j}\frac{\partial_B\dot{\xi}^i_{J_B-k+1}}{\partial^{(u)}q^n}-{}_B\xi^{(u)n}_{\quad J_B-k}\frac{\partial_A\dot{\xi}^i_{J_A-j+1}}{\partial^{(u)}q^n}\right) \\
&\qquad+\sum_1^2 u\left({}_A\xi^{(u-1)n}_{\quad J_A-j+1}\frac{\partial_B\ddot{\xi}^i_{J_B-k}}{\partial^{(u)}q^n}-{}_B\xi^{(u-1)n}_{\quad J_B-k+1}\frac{\partial_A\ddot{\xi}^i_{J_A-j}}{\partial^{(u)}q^n}\right) \\
&\qquad+\sum_0^1 u\left({}_A\xi^{(u)n}_{\quad J_A-j}\frac{\partial_B\xi^i_{J_B-k+2}}{\partial^{(u)}q^n}-{}_B\xi^{(u)n}_{\quad J_B-k}\frac{\partial_A\xi^i_{J_A-j+2}}{\partial^{(u)}q^n}\right) \\
&\qquad+2\sum_1^2 u\left({}_A\xi^{(u-1)n}_{\quad J_A-j+1}\frac{\partial_B\dot{\xi}^i_{J_B-k+1}}{\partial^{(u)}q^n}-{}_B\xi^{(u-1)n}_{\quad J_B-k+1}\frac{\partial_A\dot{\xi}^i_{J_A-j+1}}{\partial^{(u)}q^n}\right) \\
&\qquad+\sum_2^3 u\binom{u}{2}\left({}_A\xi^{(u-2)n}_{\quad J_A-j+2}\frac{\partial_B\ddot{\xi}^i_{J_B-k}}{\partial^{(u)}q^n}-{}_B\xi^{(u-2)n}_{\quad J_B-k+2}\frac{\partial_A\ddot{\xi}^i_{J_A-j}}{\partial^{(u)}q^n}\right) \\
&\qquad+2\left({}_A\xi^n_{J_A-j+2}\frac{\partial_B\dot{\xi}^i_{J_B-k+1}}{\partial\ddot{q}^n}-{}_B\xi^n_{J_B-k+2}\frac{\partial_A\dot{\xi}^i_{J_A-j+1}}{\partial\ddot{q}^n}\right) \\
&\qquad\left.+{}_A\xi^n_{J_A-j+1}\frac{\partial_B\xi^i_{J_B-k+2}}{\partial\dot{q}^n}-{}_B\xi^n_{J_B-k+1}\frac{\partial_A\xi^i_{J_A-j+2}}{\partial\dot{q}^n}\right]\frac{\partial}{\partial\ddot{q}^i}.
\end{aligned}
$$

[1]L. Lusanna, Phys. Rep. **185**, 1 (1990).
[2]L. Lusanna, "The Second Noether Theorem as the Basis of the Theory of Singular Lagrangians and Hamiltonian Contraints," to appear in Riv. Nuovo Cimento.
[3]L. Lusanna, J. Math. Phys. **31**, 428 (1990).
[4]L. Lusanna, in Proc. IV M. Grossmann Meeting, Rome 1985, edited by R. Ruffini (Elsevier, Amsterdam, 1986), p. 1163; G. Longhi, "Multitime Approach to Nonrelativistic and Relativistic Quantum Mechanics," talk given at the Encuentros Espanolos, St. Ander, 1984; G. Longhi, L. Lusanna, and J. M. Pons, J. Math. Phys. **30**, 1893 (1989).
[5]I. A. Batalin and G. A. Vilkovisky, Nucl. Phys. B **234**, 106 (1984).
[6]I. A. Batalin and G. A. Vilkovisky, J. Math. Phys. **26**, 172 (1985).
[7]I. A. Batalin, J. Math. Phys. **22**, 1837 (1981).
[8]J. A. Schouten and M. van der Kulk, *Pfaff's Problem and its Generalizations* (Claredon, Oxford, 1949); R. O. Fulp and J. A. Marlin, Pacific J. Math. **67**, 373 (1976); Rep. Math. Phys. **18**, 295 (1980); L. P. Eisenhart, *Continuous Groups of Transformations* (Princeton U.P., Princeton, 1933); G. Marmo, in *Proceedings of the IUTAM-ISIMM Symposium on Modern Developments in Analytical Mechanics*, Vol. I, Torino 1982, edited by S. Benenti, M. Francaviglia, and A. Lichnerowicz, Supplemento al n. 117 (1983) degli Atti della Accademia delle Scienze di Torino, Classe di Scienze Fisiche, Matematiche e Naturali, Torino, 1983.
[9]M. Henneaux, Phys. Rep. **126**, 1 (1985).
[10]P. A. M. Dirac, *Lectures on Quantum Mechanics*, Belfer Graduate School of Science, Monographs Series (Yeshiva University, New York, 1964).
[11]R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953), Vol. 1; L. S. Schulman, *Techniques and Applications of Path Integrations* (Wiley, New York, 1981).
[12]F. Colomo, G. Longhi, and L. Lusanna, "Classical Solutions of the Many-time Functional Equations of Motion of the Mambu String," to appear in Int. J. Mod. Phys. A; Mod. Phys. Lett. A **5**, 17 (1990); Firenze University preprint, 1989.
[13]N. H. Ibragimov, Sov. Math. Dokl. **17**, 1242 (1976); R. L. Anderson and N. H. Ibragimov, *Lie–Bäcklund Transformations in Applications* (SIAM, Philadelphia, 1979); F. Guil Guerrero and L. Martinez Alonso, J. Phys. A **13**, 689 (1980).
[14]D. Z. Freedman and P. van Nieuwenhiuzen, Phys. Rev. D **14**, 912 (1976).

Balance laws and centro velocity in dissipative systems

E. van Groesen
Department of Applied Mathematics, University of Twente, Twente, The Netherlands

F. Mainardi
Department of Physics, University of Bologna, Bologna, Italy

(Received 13 November 1989; accepted for publication 4 April 1990)

Starting with a density that is conserved for a dynamical system when dissipation is ignored, a local conservation law is derived for which the total flux (integrated over the spatial domain) is unique. When dissipation is incorporated, the conservation law becomes a balance law. The contribution due to dissipation in this balance law is split in a unique way in a part that is proportional to the density and in a divergence expression that adds to the original (conservative) flux density; the total additional flux is uniquely defined. It is shown that these total fluxes appear in the expression for the centro velocity, i.e., in the velocity of the center of gravity of the density, which shows that this velocity can be defined in a unique way (in contrast to a local velocity). Applications to the Korteweg–de Vries–Burgers equations and to the incompressible Navier–Stokes equations are given.

I. INTRODUCTION

This paper is concerned with some basic observations about balance laws for continuous systems and some consequences. Although the methods and results are rather straightforward, we are not aware of any direct treatment of these matters in the literature.

The starting point is a distinguished density E of a certain continuous system, an expression in the state variable and its derivatives. The system is assumed to be "dissipative" in the sense that the integrated quantity $\langle E \rangle$ (where $\langle\ \rangle$ denotes integration over the fixed spatial domain) will not be conserved during the evolution. However, it is assumed that the dissipation can be recognized explicitly, and that when it is ignored, $\langle E \rangle$ is a constant of the motion of the resulting "conservative" system. Although this is not essential for the following, it is helpful to think of E as an energylike quantity.

The aim of this paper is to investigate in which way the "dissipation" can be understood in its effect on E. In Sec. II it is shown that when dissipation is ignored and $\langle E \rangle$ is conserved, E satisfies a local conservation law for which the total flux $\langle \mathbf{F} \rangle$ can be uniquely defined (the flux itself is unique only in the class of curl-free functions). In the presence of dissipation, E satisfies a local balance law. In Sec. III it is shown that the contribution of dissipation can be split in a unique way in a part that is proportional to E, with the dissipation rate of $\langle E \rangle$ as factor of proportionality, and in a part that changes the original flux density $\mathbf{F}$ of E with a certain amount Θ, for which $\langle \Theta \rangle$ is unique. In Sec. IV the resulting formulation of the balance law is interpreted as a conservation law for a modified energy density that depends explicitly on time. Moreover, it is shown that the additional flux due to dissipation will also appear in the expression for the centro velocity of E, i.e., the velocity of its center of gravity. Then, as in the nondissipative case, this centro velocity equals the energy-flux velocity, but now the flux consists of the sum of the flux of the nondissipative system and the flux due to dissipation. Since in general the total flux $\langle \Theta \rangle$ due to dissipation does not vanish, the resulting expression for the centro velocity is different in the conservative and in the dissipative case. In Sec. V some examples from fluid dynamics are considered; 1-D wave equations like the Korteweg–de Vries–Burgers equation, and the inviscid Navier–Stokes equations. The final section contains some conclusions and remarks.

It may be stressed that the results in this paper are quite general and applicable to nonlinear equations. Most results in the literature about propagation velocity are for linear equations, and then often deal with harmonic waves and relate the velocity to the group or phase velocity. For the energy-flux velocity see Refs. 1–8 for systems without dissipation, and Refs. 9 and 10 for systems with dissipation. For the centro velocity in systems without dissipation see Refs. 11 and 12, the latter in particular also for nonlinear equations, and for dissipative systems see Refs. 13 and 14.

II. UNIQUENESS OF THE TOTAL FLUX IN LOCAL CONSERVATION LAWS

Let the state of a system be described by some vector function $\mathbf{u}(\mathbf{x},t)$, where $\mathbf{x}$ belongs to a spatial domain $\Omega \subset \mathbb{R}^n$. The domain Ω is assumed to be given here, possibly the whole space. The boundary of Ω is denoted by $\partial\Omega$, and the normal to the boundary by $\mathbf{n}$. It is assumed, without loss of generality that the evolution is described by an evolution equation (generally a partial differential equation) that is of first order in time. All densities to be considered below are then expressions in $\mathbf{u}$ and its spatial derivatives.

To motivate the following, assume that the evolution equation for $\mathbf{u}$ can be thought of to consist of a conservative part $\mathbf{K}_1$ and a part that may account for dissipation (or production) $\mathbf{K}_2$. To recognize the effect of dissipation in the following, a parameter ν is introduced and the equation for $\mathbf{u}$ is written as

$$\partial_t \mathbf{u} = \mathbf{K}_1(\mathbf{u}) + \nu \mathbf{K}_2(\mathbf{u}). \tag{2.1a}$$

Unfortunately, both notions of "conservative" and "dissipa-

 0022-2488/90/092136-05$03.00

tive" are difficult to define in general. What is meant here is that Eq. (2.1a) with $\nu = 0$,

$$\partial_t \mathbf{u} = \mathbf{K}_1(\mathbf{u}), \tag{2.1b}$$

has some conserved density E, while E is generally not conserved for (2.1a) when $\nu \neq 0$. In this section we consider the conservative case before dealing with the dissipative equation in the next section.

To be more specific, let E be this density and assume that it is positive definite in the following sense:

$$\langle E(\mathbf{u}) \rangle \equiv \int_\Omega E(\mathbf{u}) \geqslant 0, \quad \text{if } \mathbf{u} \neq 0.$$

Here, E will be referred to as the energy density and $\langle E \rangle$ as the total energy. The statement that E is a conserved density of (2.1b) is defined to mean that the total energy is conserved:

$$\partial_t \langle E(\mathbf{u}) \rangle = 0. \tag{2.2}$$

The first result states that then E satisfies a local conservation law.

Proposition 2.1: For the density E satisfying (2.2), there is a *local conservation law* of the form

$$\partial_t E(\mathbf{u}) + \operatorname{div} \mathbf{F}(\mathbf{u}) = 0, \tag{2.3}$$

for some *flux density* $\mathbf{F}$ that can be chosen to satisfy the boundary condition

$$\mathbf{F}(\mathbf{u}) \cdot \mathbf{n} = 0, \quad \text{on } \partial\Omega. \tag{2.4}$$

With this boundary condition, the flux is unique, possibly up to the addition of a function $\mathbf{f}$ satisfying

$$\operatorname{div} \mathbf{f} = 0, \quad \text{in } \Omega \text{ and } \mathbf{f} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega. \tag{2.5}$$

For all $\mathbf{F}$ satisfying the boundary condition (2.4) the total flux is uniquely defined and is given by

$$\langle \mathbf{F} \rangle = \langle \mathbf{x} \cdot \partial_t E \rangle, \tag{2.6}$$

where $\partial_t E$ is calculated for solutions of (2.1b).

Before proving this result, some remarks are in order.

Remark 2.1: Concerning uniqueness it can be said that $\mathbf{F}$ satisfying Eqs. (2.3) and (2.4) is unique up to elements from the kernel of the divergence operation satisfying the homogeneous boundary conditions. Since any function can be written as the sum of a solenoidal function and some gradient function, $\mathbf{F}$ could be made unique by requiring curl $\mathbf{F} = 0$. The actual construction of $\mathbf{F}$ will be performed in this way in the proof below. Note that if $n = 1$, the only function that satisfies (2.5) is the zero function, so that F is unique then. If $n = 2$, for any scalar function ψ that is constant on $\partial\Omega$, the function $\mathbf{f} = (-\psi_y, \psi_x)$ satisfies (2.5). In the same way, if $n = 3$, any function $\mathbf{f} = \operatorname{curl} \mathbf{a}$, with $\mathbf{curla} \cdot \mathbf{n} = 0$ on $\partial\Omega$, satisfies (2.5).

Remark 2.2: If one just starts with a local conservation law like (2.3), say

$$\partial_t E(\mathbf{u}) + \operatorname{div} \mathbf{F}^*(\mathbf{u}) = 0, \tag{2.7}$$

for some flux density $\mathbf{F}^*$, then upon integrating over the domain Ω and using Gauss' theorem, one obtains

$$\partial_t \langle E(\mathbf{u}) \rangle + \int_{\partial\Omega} \mathbf{F}^*(\mathbf{u}) \cdot \mathbf{n} = 0. \tag{2.8}$$

Conservation of total energy, i.e., (2.2), is recovered if $\int_{\partial\Omega} \mathbf{F}^*(\mathbf{u}) \cdot \mathbf{n} = 0$. Note, however, that unless the flux density $\mathbf{F}^*$ satisfies the pointwise condition (2.4), the total flux $\langle \mathbf{F}^* \rangle$ will differ in general from that given by (2.6) [see formula (2.10) below].

Remark 2.3: In order to specify the phrase in (2.6), suppose that E is differentiable and let $D_{\mathbf{u}} E(\mathbf{u})$ denote its Frechet derivative. That is, for any function $\mathbf{v}$, and ϵ real,

$$D_{\mathbf{u}} E(\mathbf{u}) \cdot \mathbf{v} = \frac{d}{d\epsilon} E(\mathbf{u} + \epsilon \mathbf{v}) \Big|_{\epsilon = 0},$$

or, equivalently, a first-order Taylor expansion reads

$$E(\mathbf{u} + \epsilon \mathbf{v}) = E(\mathbf{u}) + \epsilon D_{\mathbf{u}} E(\mathbf{u}) \cdot \mathbf{v} + O(\epsilon^2).$$

Then $\partial_t E(\mathbf{u})$ in (2.6) can be written like $\partial_t E(\mathbf{u}) = D_{\mathbf{u}} E(\mathbf{u}) \cdot \partial_t \mathbf{u}$, and inserting the evolution equation (2.1b), we obtain

$$\partial_t E(\mathbf{u}) = D_{\mathbf{u}} E(\mathbf{u}) \cdot \mathbf{K}_1(\mathbf{u}). \tag{2.9}$$

This is the expression that is meant in (2.6).

Proof of Proposition 2.1: The proof of this proposition uses the following standard result from potential theory.

Lemma: For a given function g on Ω, consider the following Poisson equation for a scalar function θ satisfying homogeneous Neumann boundary conditions:

$$-\nabla^2 \theta = g(x), \quad \text{in } \Omega \text{ and } \nabla\theta \cdot \mathbf{n} = 0 \text{ on } \partial\Omega.$$

This problem has a solution θ if and only if $\langle g \rangle = \int_\Omega g \, dx = 0$, and this solution θ is uniquely determined up to an additive constant.

Using this lemma, the proof of the proposition is immediate: Take for the function g the expression $\partial_t E(\mathbf{u})$, i.e., $D_{\mathbf{u}} E(\mathbf{u}) \cdot \mathbf{K}_1(\mathbf{u})$ according to (2.9). Because of the requirement (2.2), the solvability condition is satisfied and a solution θ is obtained. Then define $\mathbf{F}$ to be $\mathbf{F} = -\nabla\theta$. Since θ is defined uniquely up to a constant, $\mathbf{F}$ is uniquely defined from θ and satisfies curl $\mathbf{F} = 0$. Moreover, the boundary conditions for θ imply that $\mathbf{F} \cdot \mathbf{n} = 0$. The uniqueness up to functions $\mathbf{f}$ that satisfy (2.5) is clear. The expression for the total flux follows easily from the following formula that is obtained by applying Gauss' theorem:

$$\int_\Omega \mathbf{x} \operatorname{div} \mathbf{F} = -\int_\Omega \mathbf{F} + \int_{\partial\Omega} \mathbf{x}(\mathbf{F} \cdot \mathbf{n}). \tag{2.10}$$

This completes the proof of the proposition.

III. UNIQUE DECOMPOSITION OF THE BALANCE LAW

In this section we will take a density E and flux F that satisfy (2.3) if u evolves according to (2.1b). Then if u evolves according to (2.1a), the density E will satisfy an expression like

$$\partial_t E(\mathbf{u}) + \operatorname{div} \mathbf{F}(\mathbf{u}) = -\nu S(\mathbf{u}), \tag{3.1}$$

where S is some scalar density due to the addition of the term $\nu \mathbf{K}_2(\mathbf{u})$ to the equation. (It can be interpreted as a "loss" or "production" term, but no sign restrictions will be imposed here.) Using the Frechet derivative as in Remark 2.4, $S(\mathbf{u})$ reads

$$S(\mathbf{u}) = -D_{\mathbf{u}} E(\mathbf{u}) \cdot \mathbf{K}_2(\mathbf{u}).$$

We will refer to (3.1) as a *local balance law* for E. The aim is now to rewrite (3.1) in such a way that the contribution

from S that is responsible for a direct decrease (or increase) in E is separated from the contribution that adds to the flux F. To that end, we start with the *global* expression corresponding to (3.1). Integrating (3.1) over Ω one obtains

$$\partial_t \langle E(\mathbf{u}) \rangle = -\nu \langle S(\mathbf{u}) \rangle. \tag{3.2}$$

Introducing the instantaneous dissipation rate α as the time-dependent functional

$$\alpha(\mathbf{u}) = \langle S(\mathbf{u}) \rangle / \langle E(\mathbf{u}) \rangle, \tag{3.3}$$

(3.2) can be written like

$$\partial_t \langle E(\mathbf{u}) \rangle = -\nu \alpha(\mathbf{u}) \langle E(\mathbf{u}) \rangle. \tag{3.4}$$

The definition of α implies that

$$\langle S - \alpha E \rangle = 0, \tag{3.5}$$

which makes it possible to split S in the desired way.

Proposition 3.1: The density S can be decomposed like

$$S = \alpha E + \operatorname{div} \boldsymbol{\Theta}, \tag{3.6}$$

where $\boldsymbol{\Theta}(\mathbf{u})$ is a loss flux density that can be chosen to satisfy the boundary condition

$$\boldsymbol{\Theta} \cdot \mathbf{n} = 0, \quad \text{on } \partial\Omega. \tag{3.7}$$

Just as in proposition 2.1, $\boldsymbol{\Theta}$ is unique up to the addition of a function $\mathbf{f}$ that satisfies (2.5); the total loss flux is unique and given by

$$\langle \boldsymbol{\Theta} \rangle = \langle \mathbf{x} \cdot (S - \alpha E) \rangle. \tag{3.8}$$

Proof: This proposition follows in the same way as proposition 2.1:

$\boldsymbol{\Theta} = -\nabla\theta$, where θ satisfies $-\nabla^2\theta = S - \alpha E$, in Ω and $\nabla\theta \cdot \mathbf{n} = 0$ on $\partial\Omega$.

Substitution of (3.6) into (3.1) leads to the following proposition.

Proposition 3.2: The balance law (3.1) can be formulated as,

$$\partial_t E(\mathbf{u}) + \operatorname{div}[\mathbf{F}(\mathbf{u}) + \nu \boldsymbol{\Theta}(\mathbf{u})] = -\nu \alpha(\mathbf{u}) E(\mathbf{u}), \tag{3.9}$$

with

$$[\mathbf{F}(\mathbf{u}) + \nu \boldsymbol{\Theta}(\mathbf{u})] \cdot \mathbf{n} = 0, \quad \text{on } \partial\Omega, \tag{3.10}$$

where F, $\boldsymbol{\Theta}$, and α are defined as before, and where the total flux $\langle \mathbf{F}(\mathbf{u}) + \nu \boldsymbol{\Theta}(\mathbf{u}) \rangle$ is uniquely defined.

This result shows in an explicit way that the "loss density" S is split in a part that takes account for the change in $\langle E \rangle$ according to (3.4) and a part that is added to the flux F of the conservative system.

IV. INTERPRETATION AND CONSEQUENCE FOR THE CENTRO VELOCITY

The dissipation rate α as it appears in the balance law (3.9) acts like a uniform damping factor for E. This can be seen quite clearly in the following way. Introduce the primitive of α that will be a functional β that depends on the complete evolution of $\mathbf{u}$ from the initial time up to time t:

$$\beta(\mathbf{u}, t) := \int_0^t \alpha(\mathbf{u}(s)) ds. \tag{4.1}$$

Multiplying (3.9) by $e^{\nu\beta}$ the result can be written like a local conservation law for a density E^* that depends explicitly on time:

$$\partial_t E^*(\mathbf{u}) + \operatorname{div} \mathbf{G}^*(\mathbf{u}) = 0, \tag{4.2}$$

where

$$E^* = e^{\nu\beta} E, \quad \text{and} \quad \mathbf{G}^* = e^{\nu\beta}[\mathbf{F} + \nu\boldsymbol{\Theta}]. \tag{4.3}$$

[To see this, note that $\alpha(\mathbf{u})$ is a *functional* of $\mathbf{u}$, so the term $e^{\nu\beta}$ is at each t just a number, not a function of x, and can be interchanged with the divergence operation.] Since $\mathbf{G}^*$ satisfies the boundary condition $\mathbf{G}^* \cdot \mathbf{n} = 0$ on $\partial\Omega$, according to (3.10), it follows that E^* is really a conserved density:

$$\partial_t \langle E^*(\mathbf{u}) \rangle = 0. \tag{4.4}$$

An important consequence of the decomposition (3.9) concerns the centro velocity of the density E. This is defined as the velocity $\mathbf{V}$ of the center of gravity of E:

$$\mathbf{V} = \partial_t \mathbf{X}, \quad \text{where } \langle (\mathbf{x} - \mathbf{X}) E \rangle = 0. \tag{4.5}$$

To calculate $\mathbf{V}$, start with

$$0 = \partial_t \langle (\mathbf{x} - \mathbf{X}) E \rangle = \langle (\mathbf{x} - \mathbf{X}) \partial_t E \rangle - \mathbf{V} \langle E \rangle.$$

Then substitute the expression (3.9) for $\partial_t E$. Write $\mathbf{G} = [\mathbf{F} + \nu\boldsymbol{\Theta}]$ and apply (3.10) to see that $\mathbf{G} \cdot \mathbf{n} = 0$ on $\partial\Omega$. Consequently, the boundary integral in the following partial integration vanishes [cf. (2.10)]:

$$\int_\Omega (\mathbf{x} - \mathbf{X}) \operatorname{div} \mathbf{G} = -\int_\Omega \mathbf{G} + \int_{\partial\Omega} (\mathbf{x} - \mathbf{X}) \mathbf{G} \cdot \mathbf{n}. \tag{4.6}$$

Using all these properties we may deduce that

$$\begin{aligned} \langle (\mathbf{x} - \mathbf{X}) \partial_t E \rangle &= \langle (\mathbf{x} - \mathbf{X}) \{ -\operatorname{div}[\mathbf{F} + \nu\boldsymbol{\Theta}] - \nu\alpha E \} \rangle \\ &= \langle \mathbf{F} + \nu\boldsymbol{\Theta} \rangle - \nu\alpha \langle (\mathbf{x} - \mathbf{X}) E \rangle \\ &= \langle \mathbf{F} + \nu\boldsymbol{\Theta} \rangle. \end{aligned}$$

Thus it follows that

$$\mathbf{V} = \frac{\langle \mathbf{F} + \nu\boldsymbol{\Theta} \rangle}{\langle E \rangle} = \frac{\langle \mathbf{F} \rangle}{\langle E \rangle} + \nu \frac{\langle \boldsymbol{\Theta} \rangle}{\langle E \rangle}. \tag{4.7}$$

This expression for $\mathbf{V}$ shows that, just as in the case when there is no dissipation ($\nu = 0$), the expression for the centro velocity is the energy-flux velocity, but now the total flux incorporates a term $\langle \boldsymbol{\Theta} \rangle$ due to dissipation.

The expression (4.7) with $\nu = 0$;

$$\mathbf{V} = \langle \mathbf{F} \rangle / \langle E \rangle, \tag{4.8}$$

differs from (4.7) since $\langle \boldsymbol{\Theta} \rangle \neq 0$ in general. Only when the dissipation is *uniform*, i.e., when

$$S(\mathbf{u}) = \alpha(\mathbf{u}) E(\mathbf{u}), \tag{4.9}$$

for some functional α, $\boldsymbol{\Theta}$ vanishes identically and (4.7) and (4.8) coincide.

In the literature it is customary to take (4.8) as the energy velocity even when dissipation is present (see Refs. 9 and 10).

V. EXAMPLES FROM FLUID DYNAMICS

Consider as a first example the *Korteweg–de Vries* (KdV) equation to which some dissipation is added. This equation for a scalar function $u(x,t)$ of one space variable x reads

$$\partial_t u - \partial_x [u_{xx} - 3u^2] = -\nu D(u). \tag{5.1}$$

For the equation with $\nu = 0$, a conserved density is $E = \frac{1}{2}u^2$ so that

$$\partial_t E + \partial_x F = 0, \quad \text{with } F = \tfrac{1}{2}u_x^2 - uu_{xx} - 2u^3. \tag{5.2}$$

In this case of one-space dimension, let Ω be the whole real line. Functions u are considered that vanish, together with all their derivatives sufficiently fast at infinity. Then the flux density F as given in (5.2) satisfies the pointwise boundary condition (2.4), and so this flux density is unique, according to Remark 2.1. With this F the centro velocity is given by (4.8). Uniform damping is encountered for e.g.,

$$D(u) = \tfrac{1}{2}\alpha(u)u, \quad \text{with some functional } \alpha. \tag{5.3}$$

Then the loss-density S in the balance law (3.1) reads $S(u) = \frac{1}{2}\alpha(u)\, u^2 = \alpha(u)E(u)$, so α is the dissipation rate. A simple example is the case $\alpha(u) = \alpha$, with α a constant.

The *Korteweg–de Vries–Burgers* equation is of the form (5.1) with the viscous dissipation given by

$$D(u) = -u_{xx}. \tag{5.4}$$

Then the loss density in (1) is $S(u) = -uu_{xx}$. Since $\langle S(u)\rangle = \langle -uu_{xx}\rangle = \langle u_x^2\rangle$, the dissipation rate becomes

$$\alpha(u) = 2\langle u_x{}^2\rangle/\langle u^2\rangle. \tag{5.5}$$

Therefore, since $S(u) = -uu_{xx} = \alpha E + \partial_x\theta$, θ must satisfy the equation

$$\partial_x\theta = -uu_{xx} - \langle u_x^2\rangle/\langle u^2\rangle \cdot u^2. \tag{5.6}$$

The total flux $\langle\theta\rangle$, of interest for the centro velocity can be written as

$$\langle\theta\rangle = \langle -x\partial_x\theta\rangle. \tag{5.7}$$

From this expression, which follows from an integration by parts of its right-hand side upon use of the boundary conditions on θ, it is seen that $\langle\theta\rangle$ does not vanish in general since arbitrary solutions will not be symmetric. This shows the necessity of the additional term in (4.7) compared to (4.8). See also the remarks in the next section.

In Ref. 14 the result (4.7) is related to an expression derived by Vainshtein, and the velocity is investigated in great detail for linear wave equations.

As another example we consider the incompressible *Navier–Stokes* equations. Incompressible fluid flow is described by a velocity field $\mathbf{v}$ satisfying $\operatorname{div}\mathbf{v} = 0$, and pressure p by

$$\partial_t\mathbf{v} + (\mathbf{v}\cdot\nabla)v + \nabla p = \nu\nabla^2\mathbf{v}, \tag{5.8}$$

or, equivalently,

$$\partial_t\mathbf{v} + (curl\,\mathbf{v})\times\mathbf{v} + \nabla\left[p + \tfrac{1}{2}|\mathbf{v}|^2\right] = \nu\nabla^2 v,$$

where ν is the kinematic viscosity. For the kinetic energy density $E = \frac{1}{2}|\mathbf{v}|^2$ the balance law reads

$$\partial_t E + \operatorname{div}[(p+E)\mathbf{v}] = \nu\mathbf{v}\cdot\nabla^2\mathbf{v}. \tag{5.9}$$

The flux $\mathbf{F} = (p+E)\,\mathbf{v}$ satisfies the pointwise no-flux condition whenever the velocity field satisfies $\mathbf{v}\cdot\mathbf{n} = 0$ on $\partial\Omega$. From (5.9) it follows that

$$\partial_t\langle E\rangle = -\nu\alpha\langle E\rangle, \quad \text{with } \alpha = 2(\langle|\nabla v|^2\rangle/\langle|v|^2\rangle). \tag{5.10}$$

For inviscid fluids the centro velocity of E is given by the expression

$$\mathbf{V}_E = \langle(p+E)\mathbf{v}\rangle/\langle E\rangle. \tag{5.11}$$

For viscous fluids an additional flux $\langle\Theta\rangle$ has to be added; it is the three-dimensional analog of the viscous contribution in the KdV–Burgers equation.

For plane flows there is another interesting density: the *enstrophy density*. Expressed in terms of the component ω of the vorticity vector curl $\mathbf{v}$ perpendicular to the plane, it is given by $W = \frac{1}{2}\omega^2$, and it is conserved for inviscid fluids. In terms of ω Eq. (5.8) takes the form known as the vorticity balance equation,

$$\partial_t\omega + \mathbf{v}\cdot\nabla\omega = \nu\nabla^2\omega, \tag{5.12}$$

and the local balance law for W is easily shown to be

$$\partial_t W + \operatorname{div}[W\mathbf{v}] = \nu\omega\nabla^2\omega. \tag{5.13}$$

For the total enstrophy we thus obtain

$$\partial_t\langle W\rangle = -\nu\gamma\langle W\rangle, \quad \text{with } \gamma = 2\langle|\nabla\omega|^2\rangle/\langle\omega^2\rangle \tag{5.14}$$

and the centro velocity for W for inviscid fluids is given by

$$\mathbf{V}_W = \langle W\mathbf{v}\rangle/\langle W\rangle, \tag{5.15}$$

while for viscous fluids an additional term should be added.

In general the dissipation rate $\alpha(\mathbf{v})$ of the energy density will not coincide with the dissipation rate $\gamma(\mathbf{v})$ of the enstrophy density (They coincide only for so-called planar Taylor vortices, see Ref. 15). This fact shows that viscosity acts different on different densities. This *selective dissipation* is well known and is responsible for the self-organization process in Navier–Stokes equations (see Ref. 15).

VI. CONCLUDING REMARKS

The motivation for and results of this paper can be explained by starting with some balance law for a certain continuous system. If E is a specified density that has a clear physical meaning (the energy, say), a balance law for E is of the form (3.1):

$$\partial_t E(\mathbf{u}) + \operatorname{div}\mathbf{F}(\mathbf{u}) = -\nu S(\mathbf{u}).$$

This relation does not define F or S in a unique way. In fact, even if S is the result of adding dissipation to the the governing equation (symbolized by the factor ν), so that E satisfies the local conservation law (2.3),

$$\partial_t E(\mathbf{u}) + \operatorname{div}\mathbf{F}(\mathbf{u}) = 0,$$

when dissipation is ignored, the flux density $\mathbf{F}$ is not uniquely defined since any function curl $\mathbf{a}$ can be added to $\mathbf{F}$ without changing (2.3). This nonuniqueness is quite cumbersome if one looks for a physical meaning of the flux $\mathbf{F}$ or quantities expressed in $\mathbf{F}$. A particular example is the local energy velocity. Upon integrating (2.3) over arbitrary subdomains, it is quite natural to define a *local* energy velocity $\mathbf{v}_E$ by

$$E\cdot\mathbf{v}_E = \mathbf{F}.$$

However, the velocity defined in this way clearly changes when a nonvanishing function curl $\mathbf{a}$ is added to $\mathbf{F}$. The same problem is encountered if dissipation is present.

This paper shows that this problem can be partly solved if it is known that, provided dissipation is ignored, the total energy $\langle E\rangle$ is conserved. Then, although $\mathbf{F}$ is not unique, the

total flux $\langle \mathbf{F} \rangle$ can be uniquely defined by taking pointwise no-flux boundary conditions for $\mathbf{F}$ (the conservation of $\langle E \rangle$ implies that the integration over the boundary of the flux component normal to the boundary vanishes). In the particular case of one-dimensional problems, the scalar function F is defined uniquely.

The uniqueness of $\langle \mathbf{F} \rangle$ turns out to be of vital importance if one considers instead of the local energy velocity, an averagedlike velocity. In particular, in conservative systems, the velocity of the center of gravity—the centro velocity—of the energy is shown to be the energy-flux velocity (4.8), i.e., the quotient of $\langle \mathbf{F} \rangle$ and $\langle \mathbf{E} \rangle$, and is therefore uniquely defined. The same has been shown to be true if dissipation is added to the equations. Then the contribution S has been split into a part proportional to E and a divergence term. In the centro velocity appears the integrated expression of this divergence term; this term has also been chosen in a unique way, leading to a unique expression for the centro velocity.

Concerning the use of the centro velocity as a physical quantity to measure propagation speed, the following remarks are in order. First of all, by its definition, the centro velocity has some physical meaning. It will be clear, however, that when dissipation is the dominating feature, and not just acts as a perturbation of a conserved system, this concept, although well defined, will have little practical importance.

On the other hand, the centro velocity can be defined for (highly) nonlinear equations, and for all kind of solutions (provided the boundary conditions are satisfied). This is different for two other velocity concepts that are often used, the group and phase velocity, which are only well suited for linear (or weakly nonlinear) equations and for (quasi)monochromatic solutions.

A final remark is about the effect of dissipation on the centro velocity, i.e., about the contribution of the term $\langle \mathbf{\Theta} \rangle$ in (4.7);

$$\mathbf{V} = \frac{\langle \mathbf{F} + \nu \mathbf{\Theta} \rangle}{\langle E \rangle} = \frac{\langle \mathbf{F} \rangle}{\langle E \rangle} + \nu \frac{\langle \mathbf{\Theta} \rangle}{\langle E \rangle}.$$

In a somewhat different way, this has been investigated for equations like the uniformly damped KdV, and the KdV–Burgers equation (5.1) in a recent paper.[16] To summarize these results here, it must first be noted that the KdV equation itself (so with $\nu = 0$) has a family of travelling waves (cnoidal waves) that can be parameterized with the total energy $e \equiv \langle E \rangle = \langle \frac{1}{2} u^2 \rangle$ and propagate with a certain velocity $\lambda = \lambda(e)$ depending on e. We denote such a wave by $U(e;x - \lambda t)$. Taking such a waveform with a specific value of e as an initial condition for the dissipative KdV equation, the resulting decaying evolution was written like

$$u(x,t) = U(e(t); x - \phi(t)), \tag{6.1}$$

for some functions $e(t)$ and $\phi(t)$, so as if the evolution follows an adiabatic path along the family of travelling waves. Equations for the functions $e(t)$ and $\phi(t)$ were derived. The equation for e in lowest order of ν turns out to be an ordinary differential equation for e itself, and can be solved (numerically for the KdV–Burgers equation; for the uniformly damped case, e decreases exponentially). In the same order, the position of the wave $\phi(t)$ follows from the equation

$$\partial_t \phi(t) = \lambda(e(t)), \tag{6.2}$$

which can be integrated once $e(t)$ has been determined. Numerical calculations of these equations, and a comparison with numerical calculations of the initial value problem of the continuous dissipative KdV equation itself showed that these simple, lowest-order equations are in fact very accurate. This holds true even when highly nonlinear effects are dominant initially, and up to rather large values of ν.

Of relevance for the present paper is particularly the expression (6.2) that depicts the instantaneous propagation speed of the decaying wave as the speed of the exact travelling wave to which it is compared at that time. Since for a function u given by (6.1) the energy-flux velocity equals the propagation speed $\lambda(e)$:

$$\langle F \rangle / \langle E \rangle = \lambda(e),$$
$$\text{for } u = U(e;x - \phi), \quad \text{any value of } \phi, \tag{6.3}$$

these results show that the contribution $\langle \theta \rangle$ is negligible in this case. Another way to see this is to note that the cnoidal wave $U(e;x - \phi)$ is an even function about $x = \phi$. This causes the expression in (5.6) to be even, and therefore the total flux $\langle \theta \rangle$ given by (5.7) to be zero. Of course, the actual solution will not be even, but the numerical calculations show that the uneveness has only small effect on the propagation speed.

ACKNOWLEDGMENTS

The comments of K. Hütter (Darmstadt) were greatly appreciated and considerably improved the presentation of the results in this paper. EVG wishes to thank C.N.R. Comitato Matematiche, for providing the opportunity to visit the University of Bologna.

[1] L. J. F. Broer, Appl. Sci. Res. Sect. A **2**, 329 (1951).
[2] M. A. Biot, Phys. Rev. **105**, 1129 (1957).
[3] G. B. Whitham, Commun. Pure Appl. Math. **14**, 675 (1961).
[4] M. J. Lighthill, J. Inst. Math. Appl. **1**, 1 (1965).
[5] G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974), Chap. 11.
[6] M. Hayes, Proc. R. Soc. London Ser. A **354**, 533 (1977).
[7] L. J. F. Broer and L. A. Peletier, Appl. Sci. Res. Sect. A **17**, 65 (1967).
[8] J. de Graaf and L. J. F. Broer, Rep. Math. Phys. **3**, 109 (1972).
[9] L. Brillouin, *Wave Propagation and Group Velocity* (Academic, New York, 1960).
[10] F. Mainardi, Wave Motion **9**, 201 (1987).
[11] J. W. Wehausen and E. V. Laitone, in *Handbuch der Physik*, edited by S. Flügge (Springer, Berlin 1960), Vol. 9, pp. 446–778.
[12] E. van Groesen, J. Math. Phys. **21**, 1646 (1980).
[13] L. A. Vainshtein, Sov. Phys. Tech. Phys. **2**, 2420 (1957).
[14] E. van Groesen and F. Mainardi, Wave Motion **11**, 201 (1989).
[15] E. van Groesen, Physica A **148**, 312 (1988).
[16] E. van Groesen, F. P. H. van Beckum, and T. P. Valkering, Z. Angew. Math. Phys. **41** (1990).

Stability analysis of the Yagle–Levy multidimensional inverse scattering algorithm

Margaret Cheney
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12180

(Received 27 February 1990; accepted for publication 9 May 1990)

A layer-stripping algorithm suggested by Yagle and Levy for solving a multidimensional inverse scattering problem is analyzed. It is shown that the unfiltered version of the algorithm is unstable; for the filtered version, it is proven that the reconstruction depends continuously on the data.

I. INTRODUCTION

In this paper we consider a two-dimensional inverse scattering problem for

$$[\nabla^2 - \partial_t^2 - V(x,z)]P(x,z,t) = 0. \tag{1.1}$$

Here, the potential V is assumed to have compact support contained in the half-space $z>0$, which we refer to as the downward direction. We also assume that V is smooth and that $\nabla^2 - V$ has no bound states.[1]

We perform the following experiment. We send in the plane wave $\delta(t-z)$ and measure on the plane $z=0$ the quantities $P(x,0,t)$ and $\partial_z P(x,0,t)$. From this information we want to reconstruct $V(x,z)$.

In 1986, Yagle and Levy[2] proposed a layer-stripping algorithm for solving the corresponding three-dimensional inverse scattering problem. This algorithm, which is discussed in Sec. II, is reminiscent of the procedure one would use to solve a boundary value problem for a hyperbolic equation, a problem that is well known[3] to be ill-posed in the sense that the solution does not depend continuously on the data. This similarity suggests that the Yagle–Levy algorithm is unstable. The present paper confirms this diagnosis by exhibiting data, which, when run through the layer-stripping algorithm, gives arbitrarily wild results at the very first step.

Yagle and Levy themselves sensed this difficulty with the algorithm, and suggested a remedy. This remedy involves filtering, namely, introducing a cutoff in the Fourier domain. This paper shows that this remedy does indeed result in an algorithm that is stable in a specific norm.

II. THE YAGLE–LEVY METHOD

The Yagle–Levy idea is to write (1.1) as

$$(\partial_z + \partial_t)(\partial_z - \partial_t)P = [V - \partial_x^2]P. \tag{2.1}$$

Thus the differential operator of (1.1) is decomposed into an upgoing part $\partial_z + \partial_t$, a downgoing part $\partial_z - \partial_t$, and a lateral part ∂_x^2. We define

$$Q(x,z,t) = (\partial_z - \partial_t)P(x,z,t). \tag{2.2}$$

Because our incident wave is $\delta(t-z)$ and V is assumed to be smooth, P and Q have the form

$$P(x,z,t) = \delta(t-z) + p(x,z,t)H(t-z), \tag{2.3a}$$

$$Q(x,z,t) = q(x,z,t)H(t-z), \tag{2.3b}$$

where H is the Heaviside function that is one for positive arguments and zero for negative arguments. The coefficients p and q then satisfy

$$(\partial_z + \partial_t)p = q, \tag{2.4a}$$

$$(\partial_z - \partial_t)q = (V - \partial_x^2)p, \tag{2.4b}$$

$$V(x,z) = -2q(x,z,t=z). \tag{2.4c}$$

Roughly, the Yagle–Levy reconstruction method is to discretize the z and t derivatives and recursively solve (2.4), beginning with $p(x,0,t)$ and $q(x,0,t)$ and recovering V at each step from (2.4c).

Yagle and Levy recognized that this procedure is likely to be unstable because of the second-order lateral derivative in (2.4b). Their remedy involves working in the (transverse) Fourier domain. In particular, we define the x-Fourier transform

$$\hat{p}(k,z,t) = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{-ikx}p(x,z,t)dx,$$

with similar definitions for $\hat{q}$ and $\hat{V}$. Then (2.4) transforms into

$$(\partial_z + \partial_t)\hat{p}(k,z,t) = \hat{q}(k,z,t), \tag{2.5a}$$

$$(\partial_z - \partial_t)\hat{q}(k,z,t) = k^2\hat{p}(k,z,t) + \int \hat{V}(k-h,z)\hat{p}(h,z,t)dh, \tag{2.5b}$$

$$\hat{V}(k,z) = -2\hat{q}(k,z,t=z). \tag{2.5c}$$

A finite difference approximation of (2.5) yields

$$\hat{p}(k,z+\Delta,t+\Delta) = \hat{p}(k,z,t) + \Delta\hat{q}(k,z,t), \tag{2.6a}$$

$$\hat{q}(k,z+\Delta,t-\Delta) = \hat{q}(k,z,t) + \Delta k^2\hat{p}(k,z,t) + \Delta\int \hat{V}(k-h,z)\hat{p}(h,z,t)dh, \tag{2.6b}$$

$$\hat{V}(k,z+\Delta) = -2\hat{q}(k,z+\Delta,t=z+\Delta). \tag{2.6c}$$

Recall that we are given $\hat{p}(k,0,t)$ and $\hat{q}(k,0,t)$ for all k and t. From (2.6c) we first find $V(k,0)$ for all k. We then use (2.6a) and (2.6b) to compute $\hat{p}(k,\Delta,t)$ and $\hat{q}(k,\Delta,t)$ for all k and t. We obtain $\hat{V}(k,\Delta)$ from (2.6c); then we repeat the process to obtain $\hat{p}$, $\hat{q}$, and $\hat{V}$ for $z=2\Delta$, etc.

A difficulty arises because of the factor k^2 on the right side of (2.6b): At each step in the algorithm, the lateral high-

 0022-2488/90/092141-04$03.00

frequency components grow. In order to regularize the problem, Yagle and Levy suggested replacing the multiplier k^2 in (2.5b) by

$$H_L(k) = \begin{cases} k^2, & \text{for } |k| < L, \\ 0, & \text{for } |k| \geq L. \end{cases} \tag{2.7}$$

This filter prevents the algorithm (2.6) from amplifying the lateral high frequencies. In the next section, we will see the effect it has on stability of the algorithm.

III. STABILITY ANALYSIS OF THE YAGLE–LEVY METHOD

A. Instability example

In this section we show that the inversion method based on (2.5) without a filter such as (2.7) is indeed unstable. We do this by considering the example with data

$$p(x,0,t) = n^{-1} \exp(inx), \tag{3.1a}$$

$$q(x,0,t) = 0. \tag{3.1b}$$

In the Fourier domain, these data are

$$\hat{p}(k,0,t) = n^{-1}\delta(n-k) \tag{3.2a}$$

$$\hat{q}(k,0,t) = 0. \tag{3.2b}$$

The first step of the algorithm (2.6) applied to (3.2) first yields $V(k,0) = 0$ and then

$$\hat{p}(k,z=\Delta,t+\Delta) = n^{-1}\delta(n-k), \tag{3.3a}$$

$$\hat{q}(k,z=\Delta,t-\Delta) = \Delta k^2 n^{-1}\delta(n-k), \tag{3.3b}$$

$$\hat{V}(k,z=\Delta) = -2k^2 \Delta n^{-1}\delta(n-k). \tag{3.3c}$$

Taking the inverse Fourier transform of (3.3c), we find

$$\begin{aligned} V(x,z=\Delta) &= -2\Delta n^{-1} \int_{-\infty}^{\infty} e^{ikx} k^2 \delta(n-k)\, dk \\ &= -2\Delta n \exp(inx). \end{aligned} \tag{3.4}$$

Thus we see that by taking n large, we can obtain at the very first step a reconstruction $V(x,\Delta)$ that is as large as we please. For large n, however, the data (3.1) become arbitrarily close to zero in the sup norm. A reconstruction corresponding to $p = q = 0$, on the other hand, is identically zero. Thus we have exhibited two sets of data, namely, (3.1) and $p = q = 0$, that in the sup norm are arbitrarily close but which give reconstructions that diverge by as much as we please.

B. The same example with the filtered method

The instability exhibited in Sec. III A is due, as Yagle and Levy suggest, to the factor of k^2 in (2.6b). If we replace this k^2 by H_L defined by (2.7), then we obtain at the first step

$$\hat{q}(k,z=\Delta,t+\Delta) = \Delta H_L n^{-1}\delta(n-k), \tag{3.5a}$$

$$\hat{V}(k,z=\Delta) = -2\Delta H_L n^{-1}\delta(n-k). \tag{3.5b}$$

The inverse Fourier transform of (3.5b) is

$$\begin{aligned} V_L(x,z=\Delta) &= -2\Delta n^{-1} \int_{-L}^{L} e^{-kx} k^2 \delta(n-k)\, dk \\ &= -2\Delta n^{-1} \begin{cases} n^2 \exp(inx), & \text{if } -L \leqslant n \leqslant L, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3.6}$$

We see that the magnitude of $V_L(x,\Delta)$ can be no larger than $2\Delta L$.

C. Continuous dependence result for filtered method

Here we assume that we start with two sets of data, namely, $\{p_1,q_1\}$ and $\{p_2,q_2\}$. When fed into the filtered algorithm (2.6), they lead to reconstructions V_1 and V_2. If the data sets are close in some norm, are the reconstructions also close? The answer is yes if we measure closeness of the data in the following norm:

$$\|f\|_{1,\infty} = \sup_{t>0} \|f(\cdot,t)\|_1, \tag{3.7}$$

where $\|\cdot\|_1$ denotes the usual L^1 norm.

For the continuous dependence result, we need the following lemma concerning the possible growth of the reconstruction with depth.

Lemma: Suppose the data satisfy $\|\hat{p}(\cdot,0,\cdot)\|_{1,\infty} \leqslant M$ and $\|\hat{q}(\cdot,0,\cdot)\|_{1,\infty} \leqslant M$. Then the filtered reconstruction satisfies

$$\|\hat{p}(\cdot,n\Delta,\cdot)\|_{1,\infty} \leqslant f_p^n(\Delta,L^2,M), \tag{3.8a}$$

$$\|\hat{q}(\cdot,n\Delta,\cdot)\|_{1,\infty} \leqslant f_q^n(\Delta,L^2,M), \tag{3.8b}$$

$$\|\hat{V}(\cdot,n\Delta)\|_1 \leqslant 2f_q^n(\Delta,L^2,M), \tag{3.8c}$$

where f_p^n and f_q^n are polynomials in Δ, L^2, and M. These polynomials can be determined recursively by the following algorithm:

$$f_p^1 = M(1+\Delta), \tag{3.9a}$$

$$f_q^1 = M + \Delta L^2 M + 2M^2\Delta, \tag{3.9b}$$

$$f_p^{n+1} = f_p^n + \Delta f_q^n, \tag{3.10a}$$

$$f_q^{n+1} = f_q^n + \Delta L^2 f_p^n + 2\Delta f_q^n f_p^n. \tag{3.10b}$$

Proof: We proceed by induction. The zeroth step is to obtain $V(x,0)$ from the data:

$$\hat{V}(k,0) = -2\hat{q}(k,0,0). \tag{3.11}$$

The first step is

$$\hat{p}(k,\Delta,t+\Delta) = \hat{p}(k,0,t) + \Delta\hat{q}(k,0,t), \tag{3.12a}$$

$$\begin{aligned} \hat{q}(k,\Delta,t-\Delta) &= \hat{q}(k,0,t) + \Delta H_L \hat{p}(k,0,t) \\ &\quad + \Delta \int \hat{V}(k-h,0)\hat{p}(h,0,t)dh, \end{aligned} \tag{3.12b}$$

$$\hat{V}(k,\Delta) = -2\hat{q}(k,\Delta,\Delta). \tag{3.12c}$$

We easily estimate (3.12a) in the L^1-L^∞ norm:

$$\|\hat{p}(\cdot,\Delta,\cdot)\|_{1,\infty} \leqslant M(1+\Delta). \tag{3.13a}$$

To estimate (3.12b), we bound the second term on the right side by

$$\sup_{t>0} \Delta \int_{-L}^{L} k^2 |\hat{p}(k,0,t)|dk \leqslant \Delta L^2 M.$$

In the third term on the right side of (3.12b), we use (3.11) and Young's inequality,[4]

$$\sup_{t>0} \Delta \| -2\hat{q}(\cdot,0,0) * \hat{p}(\cdot,0,t) \|_1 \leqslant 2M^2\Delta.$$

Thus we estimate (3.12b) by

$$\|\hat{q}(\cdot,\Delta,\cdot)\|_{1,\infty} \leqslant M + \Delta L^2 M + 2M^2\Delta. \tag{3.13b}$$

From (3.12c) and (3.13b) we have

$$\|\hat{V}(\cdot,\Delta)\|_1 \leqslant 2(2\Delta M^2 + M(1+\Delta L^2)). \tag{3.13c}$$

The $(n+1)$st step of the algorithm is

$$\hat{p}(k,(n+1)\Delta,t+\Delta) = \hat{p}(k,n\Delta,t) + \Delta\hat{q}(k,n\Delta,t), \tag{3.14a}$$

$$\hat{q}(k,(n+1)\Delta,t-\Delta) = \hat{q}(k,n\Delta,t) + \Delta H_L\, \hat{p}(k,n\Delta,t) + \Delta \int \hat{V}(k-h,n\Delta)\hat{p}(h,n\Delta,t)\,dh, \tag{3.14b}$$

$$\hat{V}(k,(n+1)\Delta) = -2xq(k,(n+1)\Delta,(n+1)\Delta). \tag{3.14c}$$

If we use the same procedure for estimating the $(n+1)$st step as we did for the first step, we see that the results of the Lemma follow immediately. Q.E.D.

Corollary 1: If, as previously, the data satisfy $\|\hat{p}\|_{1,\infty} \leqslant M$, and $\|\hat{q}\|_{1,\infty} \leqslant M$, then the filtered reconstruction V satisfies

$$\|V(\cdot,n\Delta)\|_\infty \leqslant 2f_q^n(\Delta,L^2,M). \tag{3.15}$$

Proof: We merely take the inverse Fourier transform of $\hat{V}$:

$$\|V(\cdot,n\Delta)\|_\infty = \sup_x \left| \int e^{ikx}\hat{V}(k,n\Delta)\,dk \right| \leqslant \|\hat{V}(\cdot,n\Delta)\|_1.$$

Q.E.D. (3.16)

Theorem: Suppose the data satisfy

$$\|\hat{p}_i(\cdot,0,\cdot)\|_{1,\infty} \leqslant M \quad \text{and} \quad \|\hat{q}_i(\cdot,0,\cdot)\|_{1,\infty} \leqslant M,$$

for $i = 1,2$.

Then

$$\|(\hat{p}_1 - \hat{p}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty} \leqslant g_{p,p}^n(\Delta,L^2,M)\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty} + g_{p,q}^n(\Delta,L^2,M)\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty}, \tag{3.17a}$$

$$\|(\hat{q}_1 - \hat{q}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty} \leqslant g_{q,p}^n(\Delta,L^2,M)\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty} + g_{q,q}^n(\Delta,L^2,M)\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty}, \tag{3.17b}$$

$$\|(\hat{V}_1 - \hat{V}_2)(\cdot,n\Delta)\|_1 \leqslant 2g_{q,p}^n(\Delta,L^2,M)\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty} + 2g_{q,q}^n(\Delta,L^2,M)\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty}, \tag{3.17c}$$

where the g's are polynomials in Δ, L^2, and M that can be determined recursively from the f's of the Lemma by the following algorithm:

$$g_{p,p}^1 = 1, \tag{3.18a}$$

$$g_{p,q}^1 = \Delta, \tag{3.18b}$$

$$g_{q,p}^1 = \Delta(L^2 + 2M), \tag{3.19a}$$

$$g_{q,q}^1 = 1 + 2\Delta M, \tag{3.19b}$$

$$g_{p,p}^{n+1} = g_{p,p}^n + \Delta g_{q,p}^n, \tag{3.20a}$$

$$g_{p,q}^{n+1} = g_{p,q}^n + \Delta g_{q,q}^n, \tag{3.20b}$$

$$g_{q,p}^{n+1} = g_{q,p}^n + \Delta L^2 g_{p,p}^n + 2\Delta f_p^n g_{q,p}^n + 2\Delta f_q^n g_{p,p}^n, \tag{3.21a}$$

$$g_{q,q}^{n+1} = g_{q,q}^n + \Delta L^2 g_{p,q}^n + 2\Delta f_p^n g_{q,q}^n + 2\Delta f_q^n g_{p,q}^n. \tag{3.21b}$$

Proof: The proof is similar to that of the Lemma. At step zero we have

$$\|(\hat{V}_1 - \hat{V}_2)(\cdot,0)\|_\infty \leqslant 2\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty}. \tag{3.22}$$

At the first step, we write out (3.12) for $\{p_1,q_1\}$ and $\{p_2,q_2\}$ and subtract. From the resulting equation for $p_1 - p_2$, we have

$$\|(\hat{p}_1 - \hat{p}_2)(\cdot,\Delta,\cdot)\|_{1,\infty} \leqslant \|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty} + \Delta\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty}. \tag{3.23a}$$

In the equation for $q_1 - q_2$, we add and subtract the term $\hat{V}_2 * \hat{p}_1$. The estimate that results is

$$\begin{aligned}\|(\hat{q}_1 - \hat{q}_2)(\cdot,\Delta,\cdot)\|_{1,\infty} &\leqslant \|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty} \\ &\quad + \Delta L^2\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty} \\ &\quad + 2\Delta M\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty} \\ &\quad + 2\Delta M\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty}.\end{aligned} \tag{3.23b}$$

Naturally, $\hat{V}_1 - \hat{V}_2$ is bounded by twice the right-hand side of (3.23b), which we simplify to

$$\|(\hat{q}_1 - \hat{q}_2)(\cdot,\Delta,\cdot)\|_{1,\infty} \leqslant (1 + 2\Delta M)\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty} + \Delta(L^2 + 2M)\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty}. \tag{3.24}$$

At the $n+1$st step, we have

$$\|(\hat{p}_1 - \hat{p}_2)(\cdot,(n+1)\Delta,\cdot)\|_{1,\infty} \leqslant \|(\hat{p}_1 - \hat{p}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty} + \Delta\|(\hat{q}_1 - \hat{q}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty}, \tag{3.25a}$$

$$\begin{aligned}\|(\hat{q}_1 - \hat{q}_2)(\cdot,(n+1)\Delta,\cdot)\|_{1,\infty} &\leqslant \|(\hat{q}_1 - \hat{q}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty} \\ &\quad + \Delta L^2\|(\hat{p}_1 - \hat{p}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty} \\ &\quad + 2\Delta f_p^n(\Delta,L^2,M)\|(\hat{q}_1 - \hat{q}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty} \\ &\quad + 2\Delta f_q^n(\Delta,L^2,M)\|(\hat{p}_1 - \hat{p}_2)(\cdot,n\Delta,\cdot)\|_{1,\infty}.\end{aligned} \tag{3.25b}$$

From (3.25) the results of the Theorem follow. Q.E.D.

Corollary 2: If the data satisfy the hypotheses of the Theorem, then

$$\|(\hat{V}_1 - \hat{V}_2)(\cdot,n\Delta)\|_\infty \leq 2g^n_{q,p}\|(\hat{p}_1 - \hat{p}_2)(\cdot,0,\cdot)\|_{1,\infty} + 2g^n_{q,q}\|(\hat{q}_1 - \hat{q}_2)(\cdot,0,\cdot)\|_{1,\infty}. \quad (3.26)$$

Proof: We merely Fourier transform (3.17c), as in the proof of Corollary 1.

Remark: Although the example in Sec. III A does not satisfy the hypotheses of the Theorem, it can be easily modified so that it does. In particular, the data

$$p(x,0,t) = \begin{cases} n^{-1}\exp(inx), & \text{for } |x| < a, \\ 0, & \text{for } |x| \geq a, \end{cases} \quad (3.27a)$$

$$q(x,0,t) = 0, \quad (3.27b)$$

when Fourier transformed, have finite L^1-L^∞ norms. Again the computations for the first step of algorithm (2.6) can be carried out analytically and the result is a reconstruction that grows with n.

IV. REFORMULATION OF THE INVERSE PROBLEM

One might wonder whether the data $p(x,0,t)$ and $q(x,0,t)$ can be specified arbitrarily: perhaps they must satisfy consistency conditions. The following reformulation of the inverse problem suggests that they can be specified arbitrarily.

We recombine Eqs. (2.4) to read

$$(\nabla^2 - \partial_t^2)p(x,z,t) = [\,-2(\partial_z + \partial_t)p(x,z,t=z)\,]p(x,z,t).$$

Thus the inverse problem is recast as a nonlinear, nonlocal partial differential equation. The boundary data we specify are $p(x,0,t)$ and $q(x,0,t) = (\partial_z + \partial_t)p(x,0,t)$. Since $\partial_t p(x,0,t)$ can be found from $p(x,0,t)$, we could instead specify $p(x,0,t)$ and $\partial_z p(x,0,t)$.

The study of equations such as (4.1) might be a profitable approach to understanding inverse problems. For example, perhaps a local existence and uniqueness theorem could be obtained by Cauchy–Kovalevsky techniques. This we leave for future work.

ACKNOWLEDGMENTS

This paper would not have been written without the assistance of Ann Sink, with whom the author spent many enjoyable hours talking about this problem. The author is also grateful to Russ Caflisch and Stephanos Venakides for some helpful comments.

The work was partially supported by ONR Grant No. N00014-89-J-1129 and by funds from Duke University under Young Investigator Grant No. N00014-85-K-0224 from the Office of Naval Research.

[1] R. G. Newton, "Remarks on the relation between the Schrödinger equation and the plasma wave equation," Phys. Rev. A **31**, 3305 (1988).

[2] A. E. Yagle and B. Levy, "Layer-stripping solutions of multidimensional inverse scattering problems," J. Math. Phys. **27**, 1701 (1986).

[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1962), Vol. II.

[4] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975), Vol. II.

The zero curvature formulation of the sKdV equations

Ashok Das and Shibaji Roy
Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627

(Received 28 November 1989; accepted for publication 18 April 1990)

The fermionic extensions of the KdV equation are derived from the zero curvature condition associated with the superalgebra OSp(2|1). This derivation clarifies why there are only two such extensions possible and why only one of them is supersymmetric. A Lenard type of derivation of these equations is also presented.

I. INTRODUCTION

The classical integrable models[1-7] have been studied quite intensively in the past. These consist of both lattice as well as continuum models with the right number of conserved quantities in involution (for continuum models, there is an infinite number of them) to determine the flow exactly. Each of the conserved quantities can be thought of as a Hamiltonian generating its own flow and all these flows would commute. In other words, with every integrable system is associated a hierarchy of equations that are also integrable.

Although much is known about the classical integrable models, the study of the quantum systems is relatively new and has already led to interesting areas of research such as quantum algebras and quantum groups.[8-10] More recently, there has also been a lot of interest in the integrable models from the point of view of conformal field theories and string theories. For example, it is observed[11] that the dynamical variable of the Korteweg–de Vries (KdV) equation can be appropriately related to a stress tensor so that the second Poisson bracket structure associated with this system coincides with the Virasoro algebra. From the string theory point of view, it is, of course, the supersymmetric theories that are more interesting, and these are associated with the graded Virasoro algebra. A natural question, therefore, is whether there exist fermionic extensions of the KdV equation that are also integrable. This question has already been studied[12-22] in some detail and it appears that there are two fermionic extensions of the KdV equation associated with the graded Virasoro algebra, one of which is manifestly supersymmetric whereas the other is not. This is indeed quite surprising.

To understand this in more detail, we have chosen to analyze the problem from a different point of view. Let us note that the continuum integrable models can be obtained from the zero curvature condition associated with some Lie algebra. Thus, for example, the KdV equation can be obtained from the zero curvature condition[7,23,24] associated with the Lie algebra of SL(2,R), which is a subalgebra of the Virasoro algebra. To obtain the fermionic extensions of the KdV equation, it is, therefore, natural to study the zero curvature condition associated with the simplest grading of the SL(2,R) algebra, namely, the OSp(2|1) algebra. Our analysis shows that every equation in the KdV hierarchy can be independently supersymmetrized and may have a hierarchy of its own. The lowest fermionic equation in such a hierarchy will be supersymmetric, but the higher ones need not be. This explains why there are two fermionic extensions of the KdV equation. In particular, it points out that the nonsupersymmetric extension of the KdV equation is really the second equation in the hierarchy of the chiral superparticle equation.

This paper is organized as follows. In Sec. II, we briefly review the zero curvature formulation of the KdV equation based on the Lie algebra SL(2,R). In Sec. III, we study the zero curvature condition associated with the superalgebra OSp(2|1) and show how the two fermionic extensions of the KdV equation [(super KdV) equations] can be obtained from it. We also clarify various properties of these equations. In Sec. IV, we present a Lenard type of derivation of these equations with some concluding comments in Sec. V.

II. THE ZERO CURVATURE FORMULATION OF THE KdV EQUATION

In this section, we recapitulate very briefly how the KdV equation is obtained from the zero curvature condition[7,23,24] associated with the Lie algebra SL(2,R). Let us recall that the SL(2,R) algebra consists of three generators T_a, $a = 1,2,3$, satisfying

$$[T_a, T_b] = i f^c_{ab} T_c, \tag{2.1}$$

where

$$f^1_{23} = -f^1_{32} = 2,$$
$$f^2_{12} = -f^2_{21} = f^3_{31} = -f^3_{13} = -1, \tag{2.2}$$

with all other structure constants vanishing. (One can identify the T_a's with the generators of the Virasoro algebra as $T_1 = L_0$, $T_2 = L_1$, $T_3 = L_{-1}$.)

Let us next assume that $A_\mu(x)$ define $(1+1)$-dimensional gauge fields belonging to this algebra. The zero curvature condition (also known as the Cartan–Maurer equation) in this case can be written as

$$F^a_{\mu\nu} = \partial_\mu A^{\ a}_{\nu} - \partial_\nu A^{\ a}_{\mu} - f^a_{bc} A^{\ b}_{\mu} A^{\ c}_{\nu} = 0, \quad a = 1,2,3, \tag{2.3}$$

where $\mu,\nu = 0,1$ and the structure constants, f^a_{bc}, are the ones defined in Eq. (2.2). Let us also note here that

$$\partial_0 = \frac{\partial}{\partial t} \quad \text{and} \quad \partial_1 = \frac{\partial}{\partial x}. \tag{2.4}$$

Furthermore, let us choose

$$A^{\ 1}_{1} = 2\sqrt{\lambda}, \quad A^{\ 3}_{1} = -1. \tag{2.5}$$

The parameter λ is known as the spectral parameter and is associated with the eigenvalues of a linear problem.

 0022-2488/90/092145-05$03.00

With this choice of the variables, it is clear that the equations for $a=3$ and $a=1$ lead, respectively, to the following constraint equations:

$$A^1_0 = A^3_{0,x} - 2\sqrt{\lambda}\,A^3_0,$$
$$A^2_0 = \tfrac{1}{2}A^3_{0,xx} - \sqrt{\lambda}\,A^3_{0,x} - A^3_0 A^2_1. \tag{2.6}$$

In Eq. (2.6) and in what follows, a variable with a subscript x or t merely represents a derivative with respect to that variable. The dynamical equation is now obtained from $a=2$ and with the constraints in Eq. (2.6), it takes the form

$$A^2_{1,t} = \tfrac{1}{2}A^3_{0,xxx} - 2A^2_1 A^3_{0,x} - A^2_{1,x}A^3_0 - 2\lambda A^3_{0,x}. \tag{2.7}$$

If we now identify the dynamical variable of the KdV equation with A^2_1, namely, if

$$A^2_1 = u(x,t), \tag{2.8}$$

and define

$$A^3_0 = A(\lambda,u), \tag{2.9}$$

then Eq. (2.7) takes the form

$$u_t = \frac{1}{2}A_{xxx} - 2uA_x - u_x A - 2\lambda A_x$$
$$= \frac{1}{2}\left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A - 2\lambda\frac{\partial}{\partial x}A. \tag{2.10}$$

Since the variable $u(x,t)$ is independent of the spectral parameter λ, we can make a power series expansion

$$A(\lambda,u) = 2\sum_{j=0}^{n}(4\lambda)^{n-j}A_j(u), \tag{2.11}$$

with

$$A_0 = \tfrac{1}{2}.$$

Substituting this expansion into Eq. (2.10), we obtain

$$\left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A_j = \frac{\partial}{\partial x}A_{j+1},$$
$$j = 0,1,2,\ldots,n-1 \tag{2.12}$$

and

$$u_t = \left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A_n. \tag{2.13}$$

We recognize Eq. (2.12) to be identical to the recursion relation between the conserved quantities of the KdV equation. Thus we can identify

$$A_j = \delta H_j/\delta u, \tag{2.14}$$

where the H_j's define the conserved quantities (Hamiltonians) of the KdV equation. It then follows that Eq. (2.13) defines the nth equation of the KdV hierarchy. In particular, note that if we choose

$$A(\lambda,u) = (4\lambda - 2u), \tag{2.15}$$

then Eq. (2.10) would give

$$u_t = -u_{xxx} + 6uu_x, \tag{2.16}$$

which is the KdV equation.[25] Let us note here that the simplest way we can obtain just the nth equation in the hierarchy is by setting $\lambda = 0$ in Eq. (2.10) and identifying appropriately

$$A = A_n = \delta H_n/\delta u. \tag{2.17}$$

Let us also note that if we had chosen

$$A^1_1 = 2\sqrt{\lambda}, \quad A^2_1 = A^3_1 = -iv,$$
$$A^1_0 = -8\sqrt{\lambda}\,(\lambda - \tfrac{1}{2}v^2), \tag{2.18}$$

where $v(x,t)$ is the dynamical variable, then from Eq. (2.3) we would have obtained

$$v_t = -v_{xxx} + 6v^2 v_x. \tag{2.19}$$

This is nothing other than the modified KdV (MKdV) equation, and is related to the KdV equation through the Riccati relation

$$u = v^2 + v_x. \tag{2.20}$$

Thus the MKdV equation can also be obtained from the zero curvature condition associated with the SL(2,R) algebra.

III. THE ZERO CURVATURE FORMULATION OF THE sKdV EQUATIONS

Let us now study the zero curvature condition associated with the superalgebra OSp(2|1). This algebra is obtained through a grading of the SL(2,R) algebra. Therefore, in addition to the three generators of SL(2,R), we also have two fermionic generators T_α, $\alpha = 4,5$, such that the algebra takes the form

$$[T_a,T_b] = if^c_{ab}T_c,$$
$$[T_a,T_\alpha] = if^\beta_{a\alpha}T_\beta, \tag{3.1}$$
$$[T_\alpha,T_\beta]_+ = f^a_{\alpha\beta}T_a,$$

where $a,b,c = 1,2,3$ and $\alpha,\beta = 4,5$. Here $+$ denotes the anticommutator and f^c_{ab}'s are the structure constants defined in Eq. (2.2) while

$$f^4_{14} = -f^4_{41} = f^5_{51} = -f^5_{15} = -\tfrac{1}{2},$$
$$f^4_{25} = -f^4_{52} = f^5_{43} = -f^5_{34} = 1, \tag{3.2}$$
$$f^1_{45} = f^2_{44} = f^3_{55} = 2.$$

(Note here that the additional fermionic generators can be identified with those of the superconformal algebra as $T_4 = G_{1/2}$ and $T_5 = G_{-1/2}$.) The zero curvature condition, in the present case, would take the form

$$F^I_{\mu\nu} = \partial_\mu A^I_\nu - \partial_\nu A^I_\mu - f^I_{JK}A^J_\mu A^K_\nu = 0, \tag{3.3}$$

where $I = \{a,\alpha\}$ takes values 1,2,...,5 and f^I_{JK} are the structure constants of the OSp(2|1) algebra. Note again that $\mu,\nu = 0,1$ and A^4_μ and A^5_μ are fermionic in nature.

Let us next choose

$$A^1_1 = 2\sqrt{\lambda}, \quad A^3_1 = -1, \quad A^5_1 = 0. \tag{3.4}$$

Once again λ is the spectral parameter, as we will see in Sec. IV. Note that this set of conditions is consistent with Eq. (2.5) in that when the fermionic variables are set to zero, this set reduces to the earlier set for the pure bosonic theory. With this choice of the gauge fields, it is clear that the equations for $I = 3,1$, and 5 lead, respectively, to the following constraints:

$$A^1_0 = A^3_{0,x} - 2\sqrt{\lambda}\,A^3_0,$$
$$A^2_0 = \tfrac{1}{2}A^3_{0,xx} - \sqrt{\lambda}\,A^3_{0,x} - A^3_0 A^2_1 + A^5_0 A^4_1, \tag{3.5}$$
$$A^4_0 = A^5_{0,x} - \sqrt{\lambda}\,A^5_0 - A^3_0 A^4_1.$$

Note that when the fermionic variables are set to zero, these equations reduce to those of Eq. (2.6). The bosonic and the fermionic dynamical equations are now obtained from the $I=2$ and $I=4$ components of Eq. (3.3) to be

$$A^{2}_{1,t} = \tfrac{1}{2}A^{3}_{0,xxx} - 2A^{3}_{0,x}A^{2}_{1} - A^{2}_{1,x}A^{3}_{0} - 2\lambda A^{3}_{0,x} + 3A^{5}_{0,x}A^{4}_{1} + A^{5}_{0}A^{4}_{1,x}$$
$$A^{4}_{1,t} = A^{5}_{0,xx} - A^{5}_{0}A^{2}_{1} - \tfrac{3}{2}A^{3}_{0,x}A^{4}_{1} - A^{3}_{0}A^{4}_{1,x} - \lambda A^{5}_{0}. \tag{3.6}$$

If we now identify

$$A^{2}_{1} = u(x,t), \quad A^{4}_{1} = i\phi(x,t),$$
$$A^{3}_{0} = A(\lambda,u,\phi), \quad A^{5}_{0} = -(i/2)\alpha(\lambda,u,\phi), \tag{3.7}$$

where u and ϕ are the bosonic and the fermionic dynamical variables, respectively, then Eq. (3.6) gives

$$u_t = \frac{1}{2}\left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A - 2\lambda\frac{\partial A}{\partial x} + \frac{3}{2}\frac{\partial\alpha}{\partial x}\phi + \frac{1}{2}\alpha\phi x, \tag{3.8}$$
$$\phi_t = -\frac{1}{2}\alpha_{xx} + \frac{1}{2}u\alpha - \frac{3}{2}A_x\phi - A\phi_x + \frac{\lambda}{2}\alpha.$$

Furthermore, since u and ϕ are independent of the spectral parameter λ, we can make the power series expansions of the form

$$A(\lambda,u,\phi) = 2\sum_{j=0}^{n}(4\lambda)^{n-j}A_j(u,\phi),$$
$$\alpha(\lambda,u,\phi) = 2\sum_{j=0}^{n}(4\lambda)^{n-j}G_j(u,\phi), \tag{3.9}$$

with

$$A_0 = \tfrac{1}{2}, \quad G_0 = 0.$$

Substituting these into Eq. (3.8), we obtain

$$\frac{\partial A_{j+1}}{\partial x} = \left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A_j + 3\frac{\partial G_j}{\partial x}\phi + G_{j\phi},$$
$$G_{j+1} = 4\left(\frac{\partial^2}{\partial x^2} - u\right)G_j + 12\frac{\partial A_j}{\partial x}\phi + 8A_j\phi_x,$$
$$j = 0,1,2,\ldots,n-1, \tag{3.10}$$

and

$$u_t = \left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A_n + 3\frac{\partial G_n}{\partial x}\phi + G_n\phi_x,$$
$$\phi_t = -\left(\frac{\partial^2}{\partial x^2} - u\right)G_n - 3\frac{\partial A_n}{\partial x}\phi - 2A_n\phi. \tag{3.11}$$

Thus we see once again a hierarchial structure, with Eqs. (3.10) defining some form of recursion relations, with Eqs. (3.11) giving the nth equation of the hierarchy.

Let us next start with the zeroth order equation

$$A^{(0)} = 2A_0 = 1, \quad \alpha^{(0)} = 2G_0 = 0. \tag{3.12}$$

The dynamical equations are easily obtained from Eq. (3.8) or (3.11) to be

$$u_t^{(0)} = -u_x, \quad \phi_t^{(0)} = -\phi_x. \tag{3.13}$$

These equations represent the superchiral waves or the chiral superparticles and are obviously the supersymmetrization of the lowest order equation in the KdV hierarchy. The supersymmetry transformations under which Eq. (3.13) is invariant are given by

$$\delta u = \epsilon\phi_x, \quad \delta\phi = \epsilon u, \tag{3.14}$$

where ϵ is a constant anticommuting parameter.

We can now construct A_1 and G_1 from the recursion relations Eq. (3.10) and they turn out to be

$$A_1 = -u, \quad G_1 = 4\phi_x. \tag{3.15}$$

Thus, the next equation in the hierarchy is obtained from Eq. (3.11) to be

$$u_t^{(1)} = -u_{xxx} + 6uu_x + 12\phi_{xx}\phi,$$
$$\phi_t^{(1)} = -4\phi_{xxx} + 6u\phi_x + 3u_x\phi. \tag{3.16}$$

We readily recognize this to be one of the fermionic extensions of the KdV equation[12] (with the scaling $\phi\to\frac{1}{2}\phi$). This equation, however, is no longer invariant under the supersymmetry transformations of Eq. (3.14) even though the lower member of the hierarchy is. Similarly, we can construct

$$A_2 = -u_{xx} + 3u^2 + 12\phi_x\phi,$$
$$G_2 = 16\phi_{xxx} - 12u_x\phi - 24u\phi_x, \tag{3.17}$$

so that the next equation in the hierarchy would be [from Eq. (3.11)]

$$u_t^{(2)} = -u_{xxxxx} + 10u_{xxx}u + 20u_{xx}u_x - 30u_xu^2 - 120u_x\phi_x\phi - 120u\phi_{xx}\phi + 40\phi_{xxx}\phi_x + 60\phi_{xxxx}\phi, \tag{3.18}$$
$$\phi_t^{(2)} = -16\phi_{xxxxx} + 40u\phi_{xxx} + 60u_x\phi_{xx} + 50u_{xx}\phi_x - 30u^2\phi_x - 30uu_x\phi + 15u_{xxx}\phi,$$

and so on. Once again this equation does not possess the supersymmetry of Eq. (3.14).

Let us now analyze the Poisson bracket structures of this hierarchy. The lowest order equation, (3.13), is Hamiltonian with

$$H^{(0)} = -\frac{1}{2}\int dx(u^2 + \phi_x\phi) \tag{3.19}$$

and

$$\{u(x),u(y)\}_1 = \frac{\partial}{\partial x}\delta(x-y),$$
$$\{u(x),\phi(y)\}_1 = 0, \tag{3.20}$$
$$\{\phi(x),\phi(y)\}_1 = -\delta(x-y).$$

It can be seen readily that the second equation of the hierarchy, namely, Eq. (3.16), is Hamiltonian with

$$H^{(1)} = \int dx\left(u^3 + \frac{1}{2}u_x^2 + 12u\phi_x\phi - 8\phi_{xxx}\phi\right) \tag{3.21}$$

and the Poisson bracket relations of Eq. (3.20). It is also Hamiltonian with the Poisson bracket relations

$$\{u(x),u(y)\}_2 = \left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)\delta(x-y),$$

$$\{u(x),\phi(y)\}_2 = -\left(\frac{\partial}{\partial x}\phi(x) + 2\phi(x)\frac{\partial}{\partial x}\right)\delta(x-y),$$

$$\{\phi(x),u(y)\}_2 = -\left(2\frac{\partial}{\partial x}\phi(x) + \phi(x)\frac{\partial}{\partial x}\right)\delta(x-y),$$

$$\{\phi(x),\phi(y)\}_2 = 4\left(\frac{\partial^2}{\partial x^2} - u(x)\right)\delta(x-y), \tag{3.22}$$

and $H^{(0)}$ playing the role of the Hamiltonian. This brings out the bi-Hamiltonian structure of the system.

Let us next understand the lack of supersymmetry for the higher equations of the hierarchy in some detail. If we go back to Eq. (3.8), we note that except for the manifestly λ-dependent terms, these equations are invariant under the supersymmetry transformations

$$\delta u = \epsilon\phi_x, \quad \delta\phi = \epsilon u, \quad \delta A = -\epsilon\alpha, \quad \delta\alpha = -\epsilon A_x. \tag{3.23}$$

It is the manifestly λ-dependent terms that, however, give rise to the recursion relations. Thus we conclude that the recursion relations in these hierarchies break supersymmetry and, consequently, the higher order equations in a hierarchy will not be supersymmetric. This explains why one of the fermionic extensions of the KdV equation is not manifestly supersymmetric.

Let us note, however, that every equation in the KdV hierarchy can be independently supersymmetrized. Let us recall that any equation in the hierarchy can be obtained by setting $\lambda = 0$ and choosing the dynamical variables A and α appropriately. If $\lambda = 0$, the dynamical equations take the form

$$u_t = \frac{1}{2}\left(\frac{\partial^3}{\partial x^3} - 2\left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\right)A + \frac{3}{2}\alpha_x\phi + \frac{1}{2}\alpha\phi_x,$$

$$\phi_t = -\frac{1}{2}\alpha_{xx} + \frac{1}{2}u\alpha - \frac{3}{2}A_x\phi - A\phi_x. \tag{3.24}$$

As we have noted earlier, these equations are invariant under the transformations of equation (3.22). Thus with appropriate choices for A and α, we can supersymmetrize every equation of the KdV hierarchy independently. As an example, let us choose [consistent with equation (3.22)]

$$A = -2u, \quad \alpha = 2\phi_x. \tag{3.25}$$

This gives rise to the dynamical equations

$$u_t = -u_{xxx} + 6uu_x + 3\phi_{xx}\phi,$$
$$\phi_t = -\phi_{xxx} + 3(u\phi)_x, \tag{3.26}$$

which are nothing other than the second fermionic extension of the KdV equation,[13,15] and by construction these are supersymmetric. It is clear now that this method would lead to a unique supersymmetrization of every equation in the KdV hierarchy. However, there may be more than one fermionic extension of a given equation in the hierarchy and only one of them would be manifestly supersymmetric.

IV. A LENARD TYPE OF DERIVATION

Let us next give a derivation of the sKdV equations following from the linear problem. Let us consider the coupled linear equations

$$\psi_{xx} - (u+\lambda)\psi - i\phi\Phi = 0, \quad \Phi_x - i\phi\psi = 0, \tag{4.1}$$

where ψ and Φ are, respectively, bosonic and fermionic, depending on the coordinates x and t. The variables u and ϕ will be identified with the dynamical variables of the sKdV equations later. Note that λ is the spectral parameter assumed to be independent of x and t.

Let us assume that ψ is real and normalized so that

$$\int dx\,\psi^2 = 1. \tag{4.2}$$

It then follows that

$$\lambda = \int dx(\psi\psi_{xx} - u\psi^2 - i\phi\Phi\psi). \tag{4.3}$$

Since

$$\lambda_t = 0, \tag{4.4}$$

we obtain

$$\int dx(2\psi\psi_{xxt} - u_t\psi^2 - 2u\psi\psi_t - i\phi_t\Phi\psi - i\phi\Phi_t\psi - i\phi\Phi\psi_t) = 0. \tag{4.5}$$

Using Eq. (4.1), this can be simplified to give

$$\int dx(u_t\psi^2 + 2i\phi_t\psi\Phi) = 0. \tag{4.6}$$

It is clear now that for Eq. (4.6) to be true, the integrand must be a total derivative. Writing

$$u_t\psi^2 + 2i\phi_t\psi\Phi = \frac{\partial}{\partial x}P, \tag{4.7}$$

we see that the most general form for P can be written as

$$P = A\psi_x^2 + B\psi\psi_x + C\psi^2 + i\alpha\psi_x\Phi + i\beta\psi\Phi, \tag{4.8}$$

where A, B, and C are bosonic and α, β are fermionic functions of u and ϕ and their x derivatives. Any higher derivative of ψ and Φ in Eq. (4.8) can, of course, be reduced through the use of Eq. (4.1). If we now require

$$\frac{\partial P}{\partial x} = u_t\psi^2 + 2i\phi_t\psi\Phi,$$

we obtain relations between the different functions in (4.8), namely

$$B = -A_x,$$
$$C = \tfrac{1}{2}A_{xx} - A(u+\lambda) + \tfrac{1}{2}\alpha\phi, \tag{4.9}$$
$$\beta = -2A\phi - \alpha_x.$$

Furthermore, the dynamical equations for u and ϕ then turn out to be

$$u_t = \tfrac{1}{2}A_{xxx} - 2uA_x - Au_x - 2\lambda A_x + \tfrac{3}{2}\alpha_x\phi + \tfrac{1}{2}\alpha\phi_x,$$
$$\phi_t = -\tfrac{1}{2}\alpha_{xx} + \tfrac{1}{2}u\alpha - \tfrac{3}{2}A_x\phi - A\phi_x + \tfrac{1}{2}\lambda\alpha. \tag{4.10}$$

We recognize these to be identical to Eq. (3.8) and, therefore, the entire discussion of Sec. III can now be carried over. This, therefore, gives a Lenard type of derivation[26] of the sKdV equations starting from the linear problem.

V. CONCLUSION

We have derived both of the sKdV equations from the zero curvature condition associated with the superalgebra $OSp(2|1)$. We have shown that every member of the KdV hierarchy can be independently supersymmetrized and may have a hierarchy of its own. The higher members of the hierarchy, however, need not be supersymmetric. This explains why there are two fermionic extensions of the KdV equation. In particular, we have shown that one of the extensions is really nonsupersymmetric because it corresponds to a higher equation of the chiral superparticle hierarchy. We have also presented a Lenard type of derivation of the sKdV equations starting from the linear problem.

ACKNOWLEDGMENTS

We would like to thank Professor P. Mathieu for pointing out an error in an earlier version of the paper.

This work was supported in part by the U.S. Department of Energy under Contract No. DE-AC02-76ER13065.

[1] G. B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974).
[2] G. L. Lamb, Jr., *Elements of Soliton Theory* (Wiley, New York, 1980).
[3] M. A. Olshanetsky and A. M. Perelomov, Phys. Rep. **71**, 315 (1981).
[4] P. G. Drazin, *Solitons* (Cambridge U.P., London, 1983).
[5] A. C. Newell, *Solitons in Mathematical Physics* (SIAM, Philadelphia, 1985).
[6] L. D. Fadeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer-Veriag, Berlin, 1987).
[7] A. Das, *Integrable Models* (World Scientific, Singapore 1989).
[8] L. D. Fadeev, N. Yu. Reshetikhin, and L. A. Takhtajan, "Quantization of Lie groups and Lie algebras," LOMI preprint E-14-87 (1987).
[9] V. G. Drinfeld, *Quantum Groups*, Proc. Intl. Congr. Math. (Berkeley, 1986).
[10] Y. I. Manin, *Quantum Groups and Noncommutative Geometry* (Centre de Recherches Mathematiques, Montreal, 1989).
[11] J. L. Gervais, Phys. Lett. B **160**, 277 and 279 (1985).
[12] B. A. Kupershmidt, Phys. Lett. A **102**, 213 (1984).
[13] Y. Manin and A. O. Radul, Commun. Math. Phys. **98**, 65 (1985).
[14] M. Gurses and O. Oguz, Lett. Math. Phys. **11**, 235 (1986).
[15] M. Chainchian and P. Kulish, Phys. Lett. B **183**, 169 (1987).
[16] P. Mathieu, Phys. Lett. B **203**, 287 (1988).
[17] A. Bilal and J. L. Gervais, Phys. Lett. B **211**, 95 (1988).
[18] P. Mathieu, J. Math. Phys. **29**, 2499 (1988).
[19] P. Mathieu, Université Laval preprint, talk presented at the CRM workshop on "Hamiltonian Systems, Transformation Groups and Spectral Transform Method," Montreal (1989), edited by J. Harnad.
[20] P. Mathieu, Université Laval preprint, to appear in *Integrable and Superintegrable System*," edited by B. A. Kupershmidt (World Scientific, Singapore, 1989).
[21] S. Nam, Int. J. Mod. Phys. A **4**, 4083 (1989).
[22] P. I. Holod and S. Z. Pakuliak, preprint ITP-89-18E (unpublished).
[23] M. Crampin, F. A. E. Pirani, and D. C. Robinson, Lett. Math. Phys. **2**, 15 (1977).
[24] S. S. Chern and C. K. Peng, Manuscripta Math. **28**, 207 (1979).
[25] D. J. Korteweg and G. de Vries, Philos. Mag. **39**, 422 (1895).
[26] A. Lenard (unpublished), as reported in C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Commun. Pure Appl. Math. **27**, 97 (1974).

Bäcklund transformations for the isospectral and nonisospectral AKNS hierarchies

C. Tian and Y. J. Zhang
Department of Mathematics, University of Science and Technology of China, Hefei, Anhui, People's Republic of China

(Received 15 June 1989; accepted for publication 18 April 1990)

By converting the usual Lax pairs for the isospectral and nonisospectral AKNS hierarchies into Lax pairs in Riccati forms, a unified explicit form of Bäcklund transformations for these hierarchies of nonlinear evolution equations can be obtained. In the isospectral case it is an auto-Bäcklund transformation; however, in the nonisospectral case it is not an auto-Bäcklund transformation.

I. INTRODUCTION

It is well known that the Bäcklund transformation is a powerful means in the construction of solutions for nonlinear evolution equations.[1-6] In recent years, it is noticed that by using the Darboux matrix method, a unified explicit form of auto-Bäcklund transformations can be obtained for some hierarchies of isospectral nonlinear evolution equations, such as isospectral KdV, MKdV, sine–Gordon, and AKNS hierarchy.[7-11] The approach to the study is to construct the Darboux matrix first, and then to prove the gauge equivalence of the related Lax pairs. However, the demonstration of the t part is quite difficult, and it is also hard to employ this method to obtain the Bäcklund transformations for hierarchies of nonisospectral nonlinear evolution equations.

Recently, we obtained a unified explicit form of Bäcklund transformations for both the isospectral and nonisospectral KdV and MKdV hierarchies.[12,13] The method we used is to convert the usual Lax pairs for these hierarchies of equations to Lax pairs in Riccati form, and then get use of some obvious invariabilities of the t parts of the Lax pairs. The advantage of our approach is that it not only enables us to get the Bäcklund transformations for both the isospectral and nonisospectral hierarchies of nonlinear evolution equations in a unified way, but also makes the procedure much more simple and clear.

In this paper, we apply our method to the isospectral and nonisospectral AKNS hierarchies. We obtain a unified explicit form of Bäcklund transformations for these hierarchies of equations. In the isospectral case it is an auto-Bäcklund transformation; however, in the nonisospectral case it is not an auto-Bäcklund transformation. For clearness, we first consider the isospectral AKNS hierarchy in Sec. II, and then consider the nonisospectral AKNS hierarchy in Sec. III.

II. ISOSPECTRAL AKNS HIERARCHY

For convenience we always assume in this and the next section that $q(x,t)$ and $r(x,t)$ are smooth functions of x and t, and their any-order derivatives with respect to x vanish rapidly when $x \to -\infty$.

Consider the isospectral AKNS hierarchy

$$\begin{pmatrix} q \\ -r \end{pmatrix}_t = -\Phi^{n+1}\begin{pmatrix} q \\ r \end{pmatrix} = -\Phi^n \begin{pmatrix} q_x \\ -r_x \end{pmatrix}, \quad n = 0,1,2,..., \tag{2.1}$$

where

$$\Phi = \begin{pmatrix} D - 2qD^{-1}r & 2qD^{-1}q \\ -2rD^{-1}r & -D + 2rD^{-1}q \end{pmatrix}.$$

Equation (2.1) has the following Lax pair:[14]

$$\begin{pmatrix} \varphi_{ix} \\ \varphi_{2x} \end{pmatrix} = \begin{pmatrix} \eta & q \\ r & -\eta \end{pmatrix}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \qquad \begin{pmatrix} \varphi_{1t} \\ \varphi_{2t} \end{pmatrix} = \begin{pmatrix} A_n & B_n \\ C_n & -A_n \end{pmatrix}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \tag{2.2}$$

where

$$A_n = D^{-1}(qC_n - rB_n) - 2^n\eta^{n+1}, \tag{2.3}$$

$$\begin{pmatrix} B_n \\ C_n \end{pmatrix} = -\sum_{j=0}^{n} 2^j \eta^j \Phi^{n-j}\begin{pmatrix} q \\ r \end{pmatrix}, \tag{2.4}$$

and η is the spectral parameter, $\eta_t = 0$. Let

$$\begin{pmatrix} \varphi_{11}(x,t,\eta) & \varphi_{12}(x,t,\eta) \\ \varphi_{21}(x,t,\eta) & \varphi_{22}(x,t,\eta) \end{pmatrix}$$

be any fundamental solution matrix for Eq. (2.2), and define

$$\zeta_j = \frac{\mu_j\varphi_{21}(x,t,\eta_j) + \nu_j\varphi_{22}(x,t,\eta_j)}{\mu_j\varphi_{11}(x,t,\eta_j) + \nu_j\varphi_{12}(x,t,\eta_j)}, \quad j = 1,2,$$

where $\eta_i \neq \eta_j$ when $i \neq j$, and μ_j, ν_j are any constants with $|\mu_j| + |\nu_j| \neq 0$. From (2.2) we know that ζ_j satisfy the following Riccati equations:

$$\zeta_{jx} = r - 2\eta_j\zeta_j - q\zeta_j^2, \tag{2.5a}$$

$$\zeta_{jt} = C - 2A\zeta_j - B\zeta_j^2. \tag{2.5b}$$

Define

$$T = \begin{pmatrix} \dfrac{D + 2\eta_1 + 2q\zeta_1}{\zeta_2^2 - \zeta_1^2} & -\dfrac{D + 2\eta_2 + 2q\zeta_2}{\zeta_2^2 - \zeta_1^2} \\ -\dfrac{\zeta_2^2(D + 2\eta_1 + 2q\zeta_1)}{\zeta_2^2 - \zeta_1^2} & \dfrac{\zeta_1^2(D + 2\eta_2 + 2q\zeta_2)}{\zeta_2^2 - \zeta_1^2} \end{pmatrix}.$$

Then, from (2.5) we have

$$\begin{pmatrix} q_x \\ -r_x \end{pmatrix} = T\begin{pmatrix} \zeta_{1x} \\ \zeta_{2x} \end{pmatrix}, \tag{2.6}$$

 0022-2488/90/092150-05$03.00

$$\begin{pmatrix} q_t \\ -r_t \end{pmatrix} = T\begin{pmatrix} \xi_{1t} \\ \xi_{2t} \end{pmatrix} + \begin{pmatrix} \dfrac{2\eta_t(\xi_1-\xi_2)}{\xi_2^2-\xi_1^2} \\ \dfrac{2\eta_t(\xi_1^2\xi_2-\xi_1\xi_2^2)}{\xi_2^2-\xi_1^2} \end{pmatrix}. \quad (2.7)$$

Define

$$\tilde{\Phi} = \begin{pmatrix} \tilde{\Phi}_{11}(\eta_1,\eta_2,\xi_1,\xi_2) & \tilde{\Phi}_{12}(\eta_1,\eta_2,\xi_1,\xi_2) \\ \tilde{\Phi}_{21}(\eta_1,\eta_2,\xi_1,\xi_2) & \tilde{\Phi}_{22}(\eta_1,\eta_2,\xi_1,\xi_2) \end{pmatrix},$$

where

$$\begin{aligned}\tilde{\Phi}_{11}(\eta_1,\eta_2,\xi_1,\xi_2) &= -\frac{\xi_1^2+\xi_2^2}{\xi_2^2-\xi_1^2}D + \frac{2\xi_1\xi_2^2\xi_{1x}-2\xi_1^3\xi_{1x}}{(\xi_2^2-\xi_1^2)^2} \\ &\quad + \xi_1 D^{-1}G, \\ \tilde{\Phi}_{21}(\eta_1,\eta_2,\xi_1,\xi_2) &= -[2\xi_2^2/(\xi_2^2-\xi_1^2)]D + (\xi_2^2-\xi_1^2)^{-2} \\ &\quad \times[(8\xi_1\xi_2^3-4\xi_1^2\xi_2^2-4\xi_2^4)(\eta_1+\eta_2) \\ &\quad -4\xi_1\xi_2^2\xi_{1x}-2\xi_2^3\xi_{2x}-2\xi_1^2\xi_2\xi_{2x} \\ &\quad +4\xi_1\xi_2^2\xi_{2x}+4\xi_2^3\xi_{1x}]+\xi_2 D^{-1}G,\end{aligned}$$

where

$$\begin{aligned} G &= 2\left\{-\left(\frac{r}{\xi_2^2-\xi_1^2}\right)_x - \left(\frac{q\xi_2^2}{\xi_2^2-\xi_1^2}\right)_x + \frac{2r(\eta_1+q\xi_1)+2q\xi_2^2(\eta_1+q\xi_1)}{\xi_2^2-\xi_1^2}\right\} \\ &= 2(\xi_2^2-\xi_1^2)^{-3}\{(\eta_1+\eta_2)(-2\xi_2^4-12\xi_1^2\xi_2^2+8\xi_1\xi_2^3+8\xi_1^3\xi_2-2\xi_1^4)\xi_{2x}+8(\eta_1^2+\eta_2^2)(\xi_1^3\xi_2^2+\xi_1\xi_2^4) \\ &\quad -4\eta_1\eta_2(6\xi_1^2\xi_2^3+\xi_1^4\xi_2+\xi_2^5)-2\xi_2^2(\xi_2^2-\xi_1^2)\xi_{1xx}+(\xi_1^2+\xi_2^2)(\xi_2^2-\xi_1^2)\xi_{2xx}+4(\xi_2^3+\xi_1^2\xi_2)\xi_{1x}\xi_{2x} \\ &\quad +(-2\xi_2^3+2\xi_1^3+2\xi_1\xi_2^2-6\xi_1^2\xi_2)\xi_{2x}^2-4\xi_1\xi_2^2\xi_{1x}^2\}, \\ \tilde{\Phi}_{12}(\eta_1,\eta_2,\xi_1,\xi_2) &= \tilde{\Phi}_{21}(\eta_2,\eta_1,\xi_2,\xi_1), \quad \tilde{\Phi}_{22}(\eta_1,\eta_2,\xi_1,\xi_2) = \tilde{\Phi}_{11}(\eta_2,\eta_1,\xi_2,\xi_1). \end{aligned}$$

It is a straightforward calculation to verify that

$$\Phi T = T\tilde{\Phi}, \quad (2.8)$$

and from the definition of $\tilde{\Phi}$ and G we have the following lemma.

Lemma 2.1: If we take the elements of $\tilde{\Phi}$ to be polynomials of η_1 and η_2, then they are symmetric polynomials.

Lemma 2.2: Let $r(x,t)$, $q(x,t)$, and ξ_1, ξ_2 be related by Eq. (2.5a), then we have

$$\begin{pmatrix} C_n - 2A_n\xi_1 - B_n\xi_1^2 \\ C_n - 2A_n\xi_2 - B_n\xi_2^2 \end{pmatrix} = -\tilde{\Phi}^n\begin{pmatrix} \xi_{1x} \\ \xi_{2x} \end{pmatrix}.$$

Proof: See Appendix A.

By using the identity (2.8) we get

$$\Phi^n\begin{pmatrix} q_x \\ -r_x \end{pmatrix} = \Phi^n T\begin{pmatrix} \xi_{1x} \\ \xi_{2x} \end{pmatrix} = T\tilde{\Phi}^n\begin{pmatrix} \xi_{1x} \\ \xi_{2x} \end{pmatrix}, \quad (2.9)$$

so from (2.7) with $\eta_{t=0}$, we have

$$\begin{pmatrix} q_t \\ -r_t \end{pmatrix} + \Phi^n\begin{pmatrix} q_x \\ -r_x \end{pmatrix} = T\left\{\begin{pmatrix} \xi_{1t} \\ \xi_{2t} \end{pmatrix} + \tilde{\Phi}^n\begin{pmatrix} \xi_{1x} \\ \xi_{2x} \end{pmatrix}\right\}. \quad (2.10)$$

From the above identity we obtain the following lemma.

Lemma 2.3: If ξ_1,ξ_2 satisfy Eqs. (2.5b), then q and r, which are defined by (2.5a), satisfy Eq. (2.1).

Now we are prepared to obtain the auto-Bäcklund transformations for Eq. (2.1).

Theorem 2.1: Let q,r satisfy Eq. (2.1). Choose ξ_i $(i=1,2)$ appropriately, such that

$$\lim_{x\to-\infty}\xi_1 = 0, \quad \lim_{x\to-\infty}|\xi_2| = \infty$$

$$(\text{or } \lim_{x\to-\infty}|\xi_1| = \infty, \quad \lim_{x\to-\infty}\xi_2 = 0).$$

Then $\bar{q},\bar{r}$ defined by

$$\bar{q} = q + \frac{2(\eta_2-\eta_1)}{\xi_2-\xi_1}, \quad \bar{r} = r + \frac{2(\eta_2-\eta_1)\xi_1\xi_2}{\xi_2-\xi_1}, \quad (2.11)$$

also satisfy Eq. (2.1). We now prove this theorem in detail.

Proof: Denote Φ, T, $\tilde{\Phi}$ in formula (2.8) by $\Phi(q,r)$, $T(\xi_1,\xi_2,\eta_1,\eta_2)$, $\tilde{\Phi}(\xi_1,\xi_2,\eta_1,\eta_2)$, respectively. Motivated by Lemma 2.1, we define $\bar{q},\bar{r}$ by

$$\xi_{1x} = \bar{r} - 2\eta_2\xi_1 - \bar{q}\xi_1^2, \quad (2.12a)$$

$$\xi_{2x} = \bar{r} - 2\eta_1\xi_2 - \bar{q}\xi_2^2. \quad (2.12b)$$

From (2.5a) we know that $\bar{q},\bar{r}$ is given by (2.12). From the boundary conditions of ξ_1 and ξ_2, we know that $\bar{q}$ and $\bar{r}$ have the same boundary conditions as that of q and r, so from (2.8) and (2.12) we have

$$\begin{aligned}&\Phi(\bar{q},\bar{r})T(\xi_1,\xi_2,\eta_2,\eta_1) \\ &\quad = T(\xi_1,\xi_2,\eta_2,\eta_1)\tilde{\Phi}(\xi_1,\xi_2,\eta_2,\eta_1), \end{aligned} \quad (2.13)$$

so from Lemma 2.1, Lemma 2.2, and (2.5b) we have

$$\begin{aligned}&\begin{pmatrix} \bar{q}_t \\ -\bar{r}_t \end{pmatrix} + \Phi^n(\bar{q},\bar{r})\begin{pmatrix} \bar{q}_x \\ \bar{r}_x \end{pmatrix} \\ &\quad = T(\xi_1,\xi_2,\eta_2,\eta_1)\left\{\begin{pmatrix} \xi_{1t} \\ \xi_{2t} \end{pmatrix} + \tilde{\Phi}^n(\xi_1,\xi_2,\eta_2,\eta_1)\begin{pmatrix} \xi_{1x} \\ \xi_{2x} \end{pmatrix}\right\} \\ &\quad = T(\xi_1,\xi_2,\eta_2,\eta_1)\left\{\begin{pmatrix} \xi_{1t} \\ \xi_{2t} \end{pmatrix} + \tilde{\Phi}^n(\xi_1,\xi_2,\eta_1,\eta_2)\begin{pmatrix} \xi_{1x} \\ \xi_{2x} \end{pmatrix}\right\} \\ &\quad = 0, \end{aligned}$$

which proves the theorem.

The auto-Bäcklund transformation (2.11) coincides with the auto-Bäcklund transformation which was presented in Ref. 9. In Ref. 10 three kinds of auto-Bäcklund trans-

formations were given, (2.11) corresponds to the third one, the first and second one can also be obtained by using our method. However, the main purpose of this paper is to employ our approach to obtain the Bäcklund transformations for the nonisospectral AKNS hierarchy. This will be done in the next section.

III. NONISOSPECTRAL AKNS HIERARCHY

Consider the nonisospectral AKNS hierarchy

$$\begin{pmatrix} q_t \\ -r_t \end{pmatrix} = -\Phi^{n+1}\begin{pmatrix} xq \\ xr \end{pmatrix} = -\Phi^n\begin{pmatrix} xq_x+q \\ -xr_x-r \end{pmatrix}. \quad (3.1)$$

Equation (3.1) has the following Lax pair:

$$\begin{pmatrix} \varphi_{1x} \\ \varphi_{2x} \end{pmatrix} = \begin{pmatrix} \eta & q \\ r & -\eta \end{pmatrix}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \begin{pmatrix} \varphi_{1t} \\ \varphi_{2t} \end{pmatrix} = \begin{pmatrix} A_n & B_n \\ C_n & -A_n \end{pmatrix}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \quad (3.2)$$

$$A_n = D^{-1}(qC-rB) - 2^n x\eta^{n+1}, \quad (3.3)$$

$$\begin{pmatrix} B_n \\ C_n \end{pmatrix} = -\sum_{j=0}^{n} 2^j\eta^j\Phi^{n-j}\begin{pmatrix} xq \\ xr \end{pmatrix}, \quad (3.4)$$

where the spectral parameter η satisfies $\eta_t = -2^n\eta^{n+1}$.

Let $\begin{pmatrix} \varphi_{11} & \varphi_{21} \\ \varphi_{12} & \varphi_{22} \end{pmatrix}$ be any fundamental solution of Eq. (3.2), define $\zeta_j, j=1,2$ as in Sec. II, then we have

$$\zeta_{jx} = r - 2\eta_j\zeta_j - q\zeta_j^2, \quad (3.5a)$$

$$\zeta_{jt} = C - 2A\zeta_j - B\zeta_j^2, \quad (3.5b)$$

where A_n, B_n, C_n are defined by (3.3) and (3.4).

Define

$$a_j = \frac{2^j\eta_1^j\zeta_1 - 2^j\eta_2^j\zeta_2}{\zeta_2^2-\zeta_1^2}, \quad b_j = \frac{2^j\eta_2^j\zeta_1^2\zeta_2 - 2^j\eta_1^j\zeta_1\zeta_2^2}{\zeta_2^2-\zeta_1^2}, \quad (3.6)$$

$$E_j = \begin{pmatrix} a_j\zeta_1^2 - b_j + 2\zeta_1 D^{-1}(b_jq - a_jr) \\ a_j\zeta_2^2 - b_j + 2\zeta_2 D^{-1}(b_jq - a_jr) \end{pmatrix}(-1), \quad (3.7)$$

then from (3.5a) and (3.5b) we have

$$\begin{pmatrix} q_t \\ -r_t \end{pmatrix} = T\begin{pmatrix} \zeta_{1t} \\ \zeta_{2t} \end{pmatrix} - \begin{pmatrix} a_{n+1} \\ b_{n+1} \end{pmatrix}, \quad (3.8)$$

$$\begin{pmatrix} q+xq_x \\ -r-xr_x \end{pmatrix} = T\begin{pmatrix} x\zeta_{1x} \\ x\zeta_{2x} \end{pmatrix} + \begin{pmatrix} a_1 \\ b_1 \end{pmatrix}. \quad (3.9)$$

Lemma 3.1: $\Phi\begin{pmatrix} a_j \\ b_j \end{pmatrix} = TE_j + \begin{pmatrix} a_{j+1} \\ b_{j+1} \end{pmatrix}$, $j=1,2,\ldots$.

Proof: By direct calculation.

From Lemma 3.1 we have

$$\Phi^k\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} = \sum_{j=0}^{k-1} T\tilde{\Phi}E_{k-j} + \begin{pmatrix} a_{k+1} \\ b_{k+1} \end{pmatrix}, \quad k=1,2,\ldots. \quad (3.10)$$

Lemma 3.2: Assume ζ_1 and ζ_2 have the same boundary conditions as in Theorem 2.1, using (3.5a), we represent the q,r in (3.7) by ζ_j and η_j ($=1,2$). Then if we denote this E_j by $E_j(\zeta_1,\zeta_2,\eta_1,\eta_2)$, we have

$$E_j(\zeta_1,\zeta_2,\eta_1,\eta_2) - E_j(\zeta_1,\zeta_2,\eta_2,\eta_1) = 2^j(\eta_2^j-\eta_1^j)\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}.$$

Proof: Denote $E_j = \begin{pmatrix} \zeta_1 M_j \\ \zeta_2 N_j \end{pmatrix}$. By a direct calculation we have

$$M_{jx}(\zeta_1,\zeta_2,\eta_1,\eta_2) - M_{jx}(\zeta_1,\zeta_2,\eta_2,\eta_1) = 0.$$

From the boundary conditions of ζ_j ($j=1,2$) we have

$$\lim_{x\to-\infty}[M_j(\zeta_1,\zeta_2,\eta_1,\eta_2) - M_j(\zeta_1,\zeta_2,\eta_2,\eta_1)] = 2^j(\eta_2^j-\eta_1^j),$$

so we get

$$M_j(\zeta_1,\zeta_2,\eta_1,\eta_2) - M_j(\zeta_1,\zeta_2,\eta_2,\eta_1) = 2^j(\eta_2^j-\eta_1^j).$$

Similarly, we have

$$N_j(\zeta_1,\zeta_2,\eta_1,\eta_2) - N_j(\zeta_1,\zeta_2,\eta_2,\eta_1) = 2^j(\eta_2^j-\eta_1^j),$$

which proves the lemma.

Lemma 3.3: If q,r and ζ_1,ζ_2 are related by (3.5a), then

$$\begin{pmatrix} C_n - 2A_n\zeta_1 - B_n\zeta_1^2 \\ C_n - 2A_n\zeta_2 - B_n\zeta_2^2 \end{pmatrix} = -\tilde{\Phi}^n\begin{pmatrix} x\zeta_{1x} \\ x\zeta_{2x} \end{pmatrix} - \sum_{j=0}^{n-1}\tilde{\Phi}^jE_{n-j}.$$

Proof: See Appendix B.

Lemma 3.4: If ζ_1 and ζ_2 satisfy Eq. (3.5b), and q,r are defined by (3.5a), then q,r satisfy Eq. (3.1).

Proof: From (2.8) and (3.9) we have

$$\begin{aligned} &-\Phi^{n+1}\begin{pmatrix} xq \\ xr \end{pmatrix} \\ &= -T\tilde{\Phi}^n\begin{pmatrix} x\zeta_{1x} \\ x\zeta_{2x} \end{pmatrix} - \Phi^n\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \\ &= -T\tilde{\Phi}^n\begin{pmatrix} x\zeta_1 \\ x\zeta_2 \end{pmatrix} - \sum_{j=0}^{n-1} T\tilde{\Phi}^jE_{n-j} - \begin{pmatrix} a_{n+1} \\ b_{n+1} \end{pmatrix}. \end{aligned}$$

Using (3.8) we have

$$\begin{pmatrix} q_t \\ -r_t \end{pmatrix} + \Phi^{n+1}\begin{pmatrix} xq \\ xr \end{pmatrix} = T\left[\begin{pmatrix} \zeta_{1t} \\ \zeta_{2t} \end{pmatrix} + \tilde{\Phi}^n\begin{pmatrix} x\zeta_{1x} \\ x\zeta_{2x} \end{pmatrix} + \sum_{j=0}^{n-1}\tilde{\Phi}^jE_{n-j}\right], \quad (3.11)$$

which proves the lemma immediately by using Lemma 3.3.

Now we are prepared to obtain the Bäcklund transformations for the nonisospectral AKNS hierarchy.

Theorem 3.1: Let q,r satisfy Eq. (3.1), ζ_1,ζ_2 have the same boundary conditions as in Theorem 2.1. Define

$$\bar{q} = q + \frac{2(\eta_2-\eta_1)}{\zeta_2-\zeta_1}, \quad \bar{r} = r + \frac{2(\eta_1-\eta_1)\zeta_1\zeta_2}{\zeta_2-\zeta_1},$$

then we have

(a) When $\lim_{x\to-\infty}\zeta_1=0$, $\lim_{x\to-\infty}|\zeta_2|=\infty$, $\bar q,\bar r$ satisfy

$$\begin{pmatrix}\bar q_t\\ -\bar r_t\end{pmatrix}+\Phi^{n+1}(\bar q,\bar r)\begin{pmatrix}x\bar q\\ x\bar r\end{pmatrix}+\sum_{j=1}^{\eta}2^j(\eta_1^j-\eta_2^j)\Phi^{n-j}(\bar q,\bar r)\begin{pmatrix}\bar q\\ \bar r\end{pmatrix}=0;$$

(b) When $\lim_{x\to-\infty}\zeta_2=0$, $\lim_{x\to-\infty}|\zeta_1|=\infty$, $\bar q,\bar r$ satisfy

$$\begin{pmatrix}\bar q_t\\ -\bar r_t\end{pmatrix}+\Phi^{n+1}(\bar q,\bar r)\begin{pmatrix}x\bar q\\ x\bar r\end{pmatrix}+\sum_{j=1}^{n}2^j(\eta_2^j-\eta_1^j)\Phi^{n-j}(\bar q,\bar r)\begin{pmatrix}\bar q\\ \bar r\end{pmatrix}=0.$$

Proof: We only prove case (a), the proof of case (b) is similar. Define $\bar q,\bar r$ as in the proof of Theorem 2.1, then by using Lemmas 3.1–3.3 and relation (2.8) we have

$$\begin{aligned}
&\begin{pmatrix}\bar q_t\\ -\bar r_t\end{pmatrix}+\Phi^{n+1}(\bar q,\bar r)\begin{pmatrix}x\bar q\\ x\bar r\end{pmatrix}\\
&=T(\zeta_1,\zeta_2,\eta_2,\eta_1)\left\{\begin{pmatrix}\zeta_{1t}\\ \zeta_{2t}\end{pmatrix}+\Phi^n(\bar q,\bar r)T(\zeta_1,\zeta_2,\eta_2,\eta_1)\begin{pmatrix}x\zeta_{1x}\\ x\zeta_{2x}\end{pmatrix}+\Phi^n(\bar q,\bar r)\begin{pmatrix}a_1(\zeta_1,\zeta_2,\eta_2,\eta_1)\\ b_1(\zeta_1,\zeta_2,\eta_2,\eta_1)\end{pmatrix}+\begin{pmatrix}a_{n+1}(\zeta_1,\zeta_2,\eta_2,\eta_1)\\ b_{n+1}(\zeta_1,\zeta_2,\eta_2,\eta_1)\end{pmatrix}\right\}\\
&=T(\zeta_1,\zeta_2,\eta_2,\eta_1)\left\{\begin{pmatrix}\zeta_{1t}\\ \zeta_{2t}\end{pmatrix}+\tilde\Phi^n(\zeta_1,\zeta_2,\eta_2,\eta_1)\begin{pmatrix}x\zeta_{1x}\\ x\zeta_2\end{pmatrix}+\sum_{j=1}^{n}\tilde\Phi^j(\zeta_1,\zeta_2,\eta_2,\eta_1)E_{n-j}(\zeta_1,\zeta_2,\eta_2,\eta_1)\right\}\\
&=T(\zeta_1,\zeta_2,\eta_2,\eta_1)\left\{\begin{pmatrix}\zeta_{1t}\\ \zeta_{2t}\end{pmatrix}+\tilde\Phi^n(\zeta_1,\zeta_2,\eta_1,\eta_2)\begin{pmatrix}x\zeta_{1x}\\ x\zeta_{2x}\end{pmatrix}+\sum_{j=1}^{\eta}\tilde\Phi^j(\zeta_1,\zeta_2,\eta_1,\eta_2)E_{n-j}(\zeta_1,\zeta_2,\eta_1,\eta_2)\right\}\\
&\quad+T(\zeta_1,\zeta_2,\eta_2,\eta_1)\left\{\begin{pmatrix}\zeta_{1t}\\ \zeta_{2t}\end{pmatrix}+\tilde\Phi^n(\zeta_1,\zeta_2,\eta_2,\eta_1)2^j(\eta_1^j-\eta_2^j)\begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix}\right\}\\
&=\sum_{j=1}^{n}2^j(\eta_2^j-\eta_1^j)\Phi^{n-j}(\bar q,\bar r)\begin{pmatrix}\bar q\\ \bar r\end{pmatrix},
\end{aligned}$$

which proves case (a). The theorem is proved.

We notice that the Lax pair of Eqs. (3.12a) and (3.12b) can be obtained from the Lax pair of Eq. (3.1) and Lax pairs of the equations of the isospectral hierarchy. So we can get solutions of Eq. (3.1) by using Theorem 3.1.

Remark: We call the following equations the modified isospectral AKNS hierarchy:

$$\begin{pmatrix}\zeta_{1t}\\ \zeta_{2t}\end{pmatrix}=-\tilde\Phi^n\begin{pmatrix}\zeta_{1x}\\ \zeta_{2x}\end{pmatrix}.$$

Since $\tilde\Phi$ is a hereditary symmetry, we can obtain two infinite sets of symmetries for this hierarchy of equations, and we can also consider the Lie algebraic structure of these sets of symmetries. Similarly, we can consider the modified noniso-spectral AKNS hierarchy

$$\begin{pmatrix}\zeta_{1t}\\ \zeta_{2t}\end{pmatrix}=-\tilde\Phi^n\begin{pmatrix}x\zeta_{1x}\\ x\zeta_{2x}\end{pmatrix}+\sum_{j=0}^{n-1}\tilde\Phi^jE_{n-j}.$$

We will discuss this aspect in another paper.

ACKNOWLEDGMENT

This work was supported by the National Science Fund.

APPENDIX A: THE PROOF OF LEMMA 2.2

Denote

$$U=\begin{pmatrix}q\\ r\end{pmatrix},\quad V=\begin{pmatrix}\zeta_{1x}\\ \zeta_{2x}\end{pmatrix},$$

$$R_j=\begin{pmatrix}2^{j+1}\eta_1^j\zeta_1\\ 2^{j+1}\eta_2^j\zeta_2\end{pmatrix},\quad H_j=\begin{pmatrix}2^j\eta_1^j\zeta_1^2 & -2^j\eta_1^j\\ 2^j\eta_2^j\zeta_2^2 & -2^j\eta_2^j\end{pmatrix},$$

$$J_j=\begin{pmatrix}2^j\eta_1^j & 0\\ 0 & 2^j\eta_2^j\end{pmatrix}.$$

First, it is not hard to prove that

$$\begin{pmatrix}\zeta_1^2 & -1\\ \zeta_2^2 & -1\end{pmatrix}T+\begin{pmatrix}2\zeta_1D^{-1}(-r,q)T\\ 2\zeta_2D^{-1}(-r,q)T\end{pmatrix}=-\tilde\Phi+\begin{pmatrix}2\eta_1 & 0\\ 0 & 2\eta_2\end{pmatrix}.\tag{A1}$$

Then by using (2.3)–(2.5a), (2.6), (2.8), and (A1) we have

$$\begin{aligned}
&\begin{pmatrix}C_n(\eta_1)-2A_n(\eta_1)\zeta_1-B_n(\eta_1)\zeta_1^2\\ C_n(\eta_2)-2A_n(\eta_2)\zeta_2-B_n(\eta_2)\zeta_2^2\end{pmatrix}\\
&=\sum_{j=0}^{n}H_j\Phi^{n-j}U+\sum_{j=0}^{n}R_jD^{-1}(-r,q)\Phi^{n-j}U+\tfrac12R_{n+r}\\
&=\sum_{j=0}^{n-1}\{H_jT+R_jD^{-1}(-r,q)T\}\tilde\Phi^{n-j-1}V\\
&\quad+H_nU+\tfrac12R_{n+1}\\
&=\sum_{j=0}^{n-1}J_j\tilde\Phi^{n-j}V+\sum_{j=0}^{n-1}J_{j+1}\tilde\Phi^{n-j-1}V+H_nU+\tfrac12R_{n+1}\\
&=\tilde\Phi^nV+\{J_nV+H_nU+\tfrac12R_{n+1}\}\\
&=-\tilde\Phi^nV,
\end{aligned}$$

which proves the lemma.

APPENDIX B: THE PROOF OF LEMMA 3.3

Define U, V, R_j, H_j, J_j as in Appendix A. From (3.9), Lemma 3.1, (2.8), and Appendix A we have [here, we also define $F_j=\binom{b_j}{a_j}$]

$$
\begin{aligned}
&\begin{pmatrix} C_n(\eta_1) - 2A_n(\eta_1)\zeta_1 - B_n(\eta_1)\zeta_1^2 \\ C_n(\eta_2) - 2A_n(\eta_2)\zeta_2 - B_n(\eta_2)\zeta_2^2 \end{pmatrix} + \tilde{\Phi}^n \begin{pmatrix} x\zeta_{1x} \\ x\zeta_{2x} \end{pmatrix} \\
&= \sum_{j=0}^{n-2} H_j \Phi^{n-j-1} F_1 + H_{n-1} F_1 + \sum_{j=0}^{n-2} R_j D^{-1}(-r,q) \Phi^{n-j-1} F_1 + R_{n-1} D^{-1}(-r,q) F_1 \\
&= \sum_{j=0}^{n-2} [H_j + R_j D^{-1}(-r,q)] \left\{ \sum_{l=0}^{n-j-2} T \tilde{\Phi}^l E_{n-j-l-1} + F_{n-j} \right\} + H_{n-1} F_1 + R_{n-1} D^{-1}(-r,q) F_1 \\
&= \sum_{j=0}^{n-2} [H_j T + R_j D^{-1}(-r,q) T] \sum_{l=0}^{n-j-2} \tilde{\Phi}^l E_{n-j-l-1} + S \\
&= \sum_{j=0}^{n-2} [-J_j \tilde{\Phi} + J_{j+1}] \sum_{l=0}^{n-j-2} \tilde{\Phi}^l E_{n-j-l-1} + S \\
&= -\sum_{j=0}^{n-2} J_j \sum_{l=0}^{h-j-2} \tilde{\Phi}^{l+1} E_{n-j-l-1} + \sum_{j=0}^{n-2} J_{j+1} \sum_{l=0}^{n-j-2} \tilde{\Phi}^l E_{n-j-l-1} + S \\
&= -\sum_{j=0}^{n-2} J_j \sum_{l=1}^{n-j-1} \tilde{\Phi}^l E_{n-j-l} + \sum_{j=1}^{n-1} J_j \sum_{l=0}^{n-j-1} \tilde{\Phi}^l E_{n-j-l} + S \\
&= \sum_{j=0}^{n-2} J_j E_{n-j} - \sum_{j=0}^{n-2} J_j \sum_{l=0}^{n-j-1} \tilde{\Phi}^l E_{n-j-l} + \sum_{j=1}^{n-1} J_j \sum_{l=0}^{n-j-1} \tilde{\Phi}^l E_{n-j-l} + S \\
&= -\sum_{l=0}^{n-1} \tilde{\Phi}^l E_{n-l} + J_{n-1} E_1 + S + \sum_{j=0}^{n-2} J_j E_{n-j} \\
&= -\sum_{l=0}^{n-1} \tilde{\Phi}^l E_{n-l} + J_{n-1} E_1 + \sum_{j=0}^{n-2} J_j E_{n-j} + \sum_{j=0}^{n-1} [H_j + R_j D^{-1}(-r,q)] F_{n-j} \\
&= -\sum_{l=0}^{n-1} \tilde{\Phi}^l E_{n-l} + \sum_{j=0}^{n-1} J_j E_{n-j} - \sum_{j=0}^{n-1} J_j E_{n-j} \\
&= -\sum_{l=0}^{n-1} \tilde{\Phi}^l E_{n-l}.
\end{aligned}
$$

Above we have used the definition

$$
S = \sum_{j=0}^{n-2} [H_j + R_j D^{-1}(-r,q)] F_{n-j} + H_{n-1} F_1 + R_{n-1} D^{-1}(-r,q) F_1.
$$

The lemma was proved.

[1] R. M. Miura *et al.*, *Bäcklund Transformations*, Lecture Notes in Mathematics, Vol. 515 (Springer-Verlag, Berlin, 1974).
[2] C. Rogers and W. R. Shadwick, *Bäcklund Transformation Applications* (Academic, New York, 1982).
[3] Y. S. Li, "Gauge transformation, Bäcklund transformation and nonlinear superposition laws," to be published in Adv. Math.
[4] H. D. Wahlquist and F. Estabrook, "Prolongation structure of nonlinear evolution equations," J. Math. Phys. **10**, 1 (1973).
[5] M. Wadati, H. Sanuki, and K. Konno, "Relationship among inverse method, Bäcklund transformation and infinite number of conservation laws," Prog. Theor. Phys. **53**, 417 (1975).
[6] C. Tian, "Bäcklund transformation of nonlinear evolution equations," Acta Math. Appl. Sinica **2**, 89 (1985).
[7] C. H. Gu and H. S. Hu, "A unified explicit form of Bäcklund transformation for generalized hierarchies of KdV equations," Lett. Math. Phys. **11**, 325 (1986).
[8] C. H. Gu, "On Bäcklund transformation for the generalized hierarchies of compound MKdV-SG equations," Lett. Math. Phys. **12**, 31 (1986).
[9] C. H. Gu and Z. X. Zhou, "On the Darboux matrices of Bäcklund transformations for the AKNS systems," Lett. Math. Phys. **13**, 179 (1987).
[10] Y. S. Li, X. S. Gu, and M. R. Zou, "Three kinds of Darboux transformation for the evolution equations which connect with AKNS eigenvalue problem," Acta Math. Sinica **3**, 143 (1987).
[11] X. S. Gu, "L-degree Darboux transformation for the evolution equations which are associated with AKNS eigenvalue problem and its reductions," Ann. Diff. Eqs. **3**, 13 (1987).
[12] C. Tian and Y. J. Zhang, "Auto-Bäcklund transformations for the isospectral and non-isospectral KdV hierarchies," preprint (1989).
[13] C. Tian and Y. J. Zhang, "Auto-Bäcklund transformations for the isospectral and non-isospectral MKdV hierarchies," preprint (1989).
[14] Y. S. Li, Sci. Sinica A **25**, 385 (1982).

Classical, linear, electromagnetic impedance theory with infinite integrable discontinuities

Brian DeFacio
Department of Physics and Astronomy, University of Missouri at Columbia, Columbia, Missouri 65211

(Received 17 March 1989; accepted for publication 18 April 1990)

The impedance theory is formulated for classical, linear electromagnetic scattering from a compact obstacle with a finite number of nonintersecting boundaries. The boundaries are allowed to support infinite, integrable discontinuities in electromagnetic response and the compact regions can depend on space and time. The direct scattering problem is discussed, generalizing recent results by Sabatier and collaborators for the scalar impedance acoustic problem to classical electromagnetism. A chain of Maxwell scattering equations are derived for the direct scattering problem. Two kinds of ambiguities of electromagnetism at a fixed angle of incidence are found to arise, one from discontinuities in electromagnetic material properties, and the other is from time dispersion. Cases are mentioned when parts of the scattering medium are allowed to have time-dependent motions. This is in contrast to the case of scalar acoustics where ambiguities are intrinsic to certain infinite families of values of Young's modulii.

I. INTRODUCTION

Recently, Sabatier and his collaborators have made a breakthrough in direct and inverse scattering problems for the scalar impedance equations.[1-5] Both one-dimensional and three-dimensional results for acoustic waves have been obtained, which significantly extend earlier results.[6,7] The impedance problem is based on the linear pde:

$$[\alpha^2(\mathbf{x})]^{-1}\nabla\cdot\{\alpha^2(\mathbf{x})(\nabla\Psi)\} + \{\omega^2/c^2(\mathbf{x}) + V(\mathbf{x})\}\Psi = 0, \quad (1)$$

for $\mathbf{x}\in R^3$, $\alpha^2\in C^2(R^3)\setminus\{U_iS_i\}$ with the S_i's N-orientable, nonintersecting surfaces with outward normals $\hat{n}_i(\mathbf{x},t)$, $V:R^3\to R^1$ is a multiplication function for which a scattering theory is defined such as the Rollnik class, ∇ is the del operator, and Ψ is the condensation or variation in pressure from the mean value. The boundary conditions are that Ψ and traction $\alpha_i\Psi_i$ (i labels the interface) are continuous at the interface S_i. The impedance is allowed to be singular on the surfaces S_i with the integrable singularities. The scattering data from each $\mathbf{x}\in S_i$ are given by

$$\{t_i(\mathbf{x})\}^{-1} = \frac{1}{2}\left\{\frac{\alpha(\mathbf{x}+)}{\alpha(\mathbf{x}-)} + \frac{\alpha(\mathbf{x}-)}{\alpha(\mathbf{x}+)}\right\}, \quad (2a)$$

$$\frac{r_i(x)}{t_i(\mathbf{x})} = \frac{1}{2}\left\{\frac{\alpha(\mathbf{x}+)}{\alpha(\mathbf{x}-)} - \frac{\alpha(\mathbf{x}-)}{\alpha(\mathbf{x}+)}\right\}, \quad (2b)$$

and

$$\frac{s_i(\mathbf{x})}{t_i(\mathbf{x})} = \frac{1}{2}\hat{n}\cdot\left\{\frac{\nabla\alpha(\mathbf{x}-)}{\alpha(\mathbf{x}+)} - \frac{\nabla\alpha(\mathbf{x}+)}{\alpha(\mathbf{x}-)}\right\}. \quad (2c)$$

The interpretation of these equations is that r_i and t_i are reflection and transmission factors and s_i is the slope factor for the ith interface S_i. Obviously, for each i

$$r_i^2 + t_i^2 = 1.$$

In Ref. 2 it is shown that in the special case of one dimension (1D), when at some interface j, $r_j = 0$, $t_j = 1$, then the 1-D version of Eq. (1) reduces to the Schrödinger equation:

$$\left[\frac{d^2}{dx^2} + k^2 - V - 2\sum_i^N s_i\delta(x-x_i)\right]\Psi(k,x) = 0. \quad (3)$$

This special case allows an interpretation of the slope factors s_i as the magnitude of the discontinuity on the ith surface. Whenever r_i and t_j are both nonzero, Eq. (1) does not reduce to any Schrödinger equation. In general, the slope data are singular functions required by the boundary conditions and could be obtained from any compatible boundary condition (usually continuous traction or slip *bc* in acoustics). Following Coston,[8] the jumps will be restricted being no more singular than the Dirac delta distribution so that the media differ by a Heaviside function together with a finite number of point sources. All of the smooth terms are included in the potential $V(\mathbf{x})$. The integrable discontinuities could arise in at least four different ways; as discontinuities of material parameters, or in linearizations of an electromagnetic shock wave,[9] or a distributional metric,[10] structure, or from a surface electronic state.[11] Density $\rho(\mathbf{x})$ or Young's modulus tensors $E_{ij}(\mathbf{x})$; premittivity ϵ, conductivity, or permeability μ are the examples of material parameters in acoustics and electromagnetism, respectively. An impedance theory explicitly studies the discontinuities in material parameters.

The results on scalar impedance, which have been obtained by Sabatier and co-workers,[1-7] will be generalized to classical linear, macroscopic electromagnetism.[10-25] This analysis should be valid for times that correspond to frequencies $\nu < 10^{11}$ Hz, i.e., below the infrared frequency region. The standard ambiguities of Sabatier *et al.* will be identified for electromagnetism.[11-25] The case of a material discontinuity is barely treated in the present work. The very different physics of temporal dispersion is studied in more detail. They are natural for sharp time pulses that will cover large frequency windows due to the uncertainty principle.

Krueger[17] has introduced jump discontinuities, into one-dimensional electromagnetic inversion and elasticity,

 0022-2488/90/092155-10$03.00

and Costen[8] has given a thorough discussion of three-dimensional electromagnetic jumps with more general discontinuities. The impedance structure given here briefly discusses moving, deforming surfaces with variable pointwise mean and curvature.

The electromagnetic generalization of scalar impedance problems treats spatial discontinuities in the constitutive functions, and also time dispersion in the dielectric response function $\epsilon(t,\mathbf{x})$ and the permeability function $\mu(t,\mathbf{x})$ must be considered at higher frequencies. Their resonant response frequency is not equal to the frequency of the electromagnetic field in general and any phase lags in response are important. Parts of the scattering obstacle may be moving spatially or changing in time. It is useful to define effective current densities of electric polarization $\mathbf{J}_p$ and magnetization $\mathbf{J}_M$ in terms of the polarization $\mathbf{P}$ and magnetization $\mathbf{M}$ as

$$\mathbf{J}_p(\mathbf{x},t) = \left(\frac{\partial \mathbf{P}}{\partial t}\right)(\mathbf{x},t) \tag{4a}$$

and

$$\mathbf{J}_M(\mathbf{x},t) = (\nabla \times \mathbf{M})(\mathbf{x},t), \tag{4b}$$

in matter. In nonmagnetic objects $\mathbf{J}_M = \mathbf{0}$ and in conductors the effective polarization current density is numerically much smaller than the conduction or free current density $\mathbf{J}_f$ at low frequencies. The fact that spatial dispersion automatically follows from nonlocality has been discussed by Eringen.[18] The most general linear local case will be formulated here for completeness in the time domain.

In biophysics, in geophysics, and in nondestructive evaluation, a continuously varying medium $V(\mathbf{x})$ populated by various surfaces of discontinuity S_i, is a canonical model. The fact that the propagation of energy and information are changed in a qualitative way which *cannot* be treated in the standard manner[1-7] may add to the value of this investigation. As Sabatier has stated,[4] this involves the propagation of the singularities of a pde that is a basic aspect of its solution. Indeed, the difference in his acoustic case and the present electromagnetic study occurs because of the transverse vector nature of electromagnetism and the higher frequencies considered.

In order to maximize the number of readers, the presentation is in terms of local coordinates and in (awful) SI units. It is assumed that the objects $(\mathbf{x},t)$, $(\mathbf{k},w)$, etc. are expressed on some smooth coordinate chart of an atlas.

II. ASSUMPTIONS AND NOTATION

A classical, linear electromagnetic wave[12-25] in matter satisfies Maxwell's equations in three-vector form where

$$\nabla \times \mathbf{B} = \mu \mathbf{J}_T + \epsilon\mu \frac{\partial \mathbf{E}}{\partial t}, \tag{5a}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{5b}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{5c}$$

$$\nabla \cdot \mathbf{E} = \rho/\epsilon, \tag{5d}$$

where $\mathbf{E}$ is the electric field, $\mathbf{B}$ is the magnetic field, $\mathbf{D}$ is the electric displacement vector, H is the magnetic displacement, and $\mathbf{J}_T, \rho$ are the total current density and total charge density sources, respectively. The linear, constitutive equations are

$$\mathbf{D} = \epsilon \mathbf{E} = \epsilon_0 \mathbf{E} + \mathbf{P}, \tag{6a}$$

$$\mathbf{B} = \mu \mathbf{H} = \mu_0 \mathbf{H} + \mu_0 \mathbf{M}, \tag{6b}$$

$$J_f = \sigma \mathbf{E}, \tag{6c}$$

and

$$\mathbf{J}_T = \mathbf{J}_f + \mathbf{J}_M + \mathbf{J}_p, \tag{6d}$$

where ϵ, μ, σ are the symmetric, electric permittivity, magnetic permeability, and conductivity tensors, and $\mathbf{P}$, $\mathbf{M}$ are the macroscopic electric and magnetic polarization vectors. For simplicity of exposition, the linear, inhomogeneous isotropic stationary case will be presented. The physical fundamentals of these constitutive equations are sketched in the Appendix. The scattering regions discussed in Sec. IV will assume that the response functions ϵ and μ are $C^2(R^3 \times [0,T])$, except upon the surfaces of discontinuity between regions. The Lorentz force law describes the action of the electromagnetic field on a point particle of mass m, charge e, and instantaneous velocity $\mathbf{v}$.

Consider a single moving, deforming surface S between regions 1 and 2 are parametrized by the $C^2(R^3 \times [0,T)]$ function ϕ according to

$$\phi(\mathbf{x},t) = 0. \tag{7a}$$

The unit outward normal to the surface $\hat{n}(\mathbf{x},t)$ is given by

$$\hat{n}(\mathbf{x},t) = (1/|\nabla\phi|)(\nabla\phi), \tag{7b}$$

the speed of a point on the surface along $\hat{n}$ is

$$u_n = \hat{n} \cdot \mathbf{v}_s = \frac{1}{|\nabla\phi|}\left(\frac{\partial \phi}{\partial t}\right), \tag{7c}$$

where $\mathbf{v}_s$ is the velocity at the point and the pointwise mena curvature, positively concave in the direction $-\hat{n}$, is given by

$$\kappa = -\tfrac{1}{2}\nabla \cdot \hat{n}. \tag{7d}$$

This comes directly from Sec. 278 of Truesdell and Toupin's classical article in Ref. 15 and Costen in Ref. 8.

The following notation was required to allow general relations between the surface S and the electromagnetic wave.

(1) The square bracket symbol $[\cdot]$ denotes the jump in a field quantity across the surface S, including $[\mathbf{E}]$, $[\mathbf{B}]$, $[\mathbf{D}]$, $[\mathbf{H}]$, $[\mathbf{J}]$, and $[\rho]$. The general expressions for these vector fields will be presented later, as taken from Ref. 8.

(2) The component of a vector in the surface S will have subscript "s" added, i.e., $\mathbf{E}_s$, $\mathbf{B}_s$, $\mathbf{D}_s$, $\mathbf{H}_s$, $\mathbf{J}_s$, and ρ_s. The first four of these terms arise from the point charges in the materials interface.

(3) The unit vectors $(\hat{t}_1, \hat{t}_2, \hat{n})$ form an orthonormal basis for R^3 with origin in S. A generic tangential unit vector will be written as $\hat{t}$ and the outward normal derivative to S will be written as

$$\frac{\partial}{\partial n} = \hat{n} \cdot \nabla.$$

For two or more surfaces it is necessary that the surfaces

either nest or are approximately parallel without intersections.

The case of a jump discontinuity is well known, and Jackson[11] correctly states that it is easy to generalize the jump conditions but only gives partial results. Truesdell and Toupin[15] develop the framework and give the most general jump relations for a discontinuity in their Sec. 278 as

$$[\mathbf{E}] = f\hat{n} - u_n \mathbf{k}_{TT}, \tag{8a}$$

$$[\mathbf{B}] = \mathbf{k}_{TT} \times \hat{n}, \tag{8b}$$

where $f\hat{n}$ and $\mathbf{k}_{TT}$ are "arbitrary vector fields on the material surface of discontinuity." In order to define a scattering theory, Hölder continuity conditions[25] as specified in the next section are required. Moreover, to apply Eqs. (8a) and (8b), it is necessary to know the physical and geometrical content of the fields f and $\mathbf{k}_{TT}$ that Costen[8] has given. The scattered electromagnetic wave from simple jump discontinuities between fixed media was given by Ström[16] in a lovely expository paper. This paper generalizes his result in two respects: (1) A moving deforming boundary between media is treated instead of his fixed boundaries, and (2) spatially inhomogeneous media are allowed in the present analysis. This was given in Ref. 8 and in addition, the discontinuity equations for the macroscopic fields **H** and **D** are needed for the interplay of the geometry of the surface in motion, surface charges, and surface currents. There is one major difference between the impedance theory presented here and the work by Costen.[8] Since the impedance structure follows from the nature of the discontinuity, his interpolating field that was defined everywhere is not useful here.

The interpretation of the various terms include the following.

The term

$$\left(\frac{\partial}{\partial t} - u_n \frac{\partial}{\partial n}\right)(\cdot)$$

represents the time rate of change of $(\cdot)$ moving with the moving surface along its normal. The term $2\kappa u_m(\cdot)$ represents a local increase (or decrease) of $(\cdot)$ due to expansion (or contraction) of the interface at the point, a term $u_n[(\cdot)]$ represents a "garden-plow effect" of piling up (or pushing away) discontinuity of $(\cdot)$.

The discontinuity equation

$$\hat{n}\times[\mathbf{H}] + \hat{n}(\hat{n}(\nabla\times\mathbf{H}_s)) - ((\hat{n}\times\mathbf{H}_s)\cdot\nabla)\hat{n} - \hat{n}\times\nabla(\hat{n}\cdot\mathbf{H}_s) = \mathbf{J}_s + u_n[\mathbf{D}] + 2\kappa u_n \mathbf{D}_s - \left(\frac{\partial}{\partial t} + u_n \frac{\partial}{\partial n}\right)\mathbf{D}_s, \tag{9a}$$

was obtained for Eq. (5a) by Costen. He applied a similar analysis to Eq. (5b) and this gave

$$\hat{n}\times[\mathbf{E}] + \hat{n}(\hat{n}\cdot(\nabla\times\mathbf{E}_s)) - ((\hat{n}\times\mathbf{E}_s)\cdot\nabla)\hat{n} - \hat{n}\times\nabla(\hat{n}\cdot\mathbf{E}_s) = u_n[\mathbf{B}] + 2\kappa u_n \mathbf{B}_s - \left(\frac{\partial}{\partial t} + u_n \frac{\partial}{\partial n}\right)\mathbf{B}_s, \tag{9b}$$

for the $\nabla\times\mathbf{E}$ Maxwell equation. Using a pillbox integration region the divergence Maxwell equations have the discontinuity equations

$$\hat{n}\cdot[\mathbf{B}] + \hat{n}\cdot(\nabla\times(\hat{n}\times\mathbf{B}_s)) = 0, \tag{9c}$$

$$\hat{n}\cdot[\mathbf{D}] + \hat{n}\cdot(\nabla\times(\hat{n}\times\mathbf{D}_s)) = 4\pi[\rho]. \tag{9d}$$

Interpretation: For sufficiently differentiable generic vectors $\mathbf{A},\mathbf{A}_s$: (1) $((\hat{n}\times\mathbf{A})\cdot\nabla)\hat{n}$ is a contribution from the twist and curvature of the material surface along the direction $(\hat{n}\times\mathbf{A})$; (2) $\hat{n}\times(\nabla(\hat{n}\cdot\mathbf{A}_s))$ is the contribution of the normal part of the line integrals. This occurs with the present degree of singularity in the boundary in addition to the usual tangential contribution $\hat{n}\times\mathbf{A}_s$; (3) $\hat{n}\cdot(\nabla\times(\hat{n}\times\mathbf{D}_s)) = \nabla_2\cdot(\mathbf{D}_s - \hat{n}D_n)$, where ∇_2 is a two-dimensional divergence and $\mathbf{D}_s - \hat{n}D_n$ is a two-component part of **D** in the surface, $\phi = 0$.

Remarks: Using Eqs. (9a)–(9d) it is now possible to identify the fields f and $\mathbf{k}_{TT}$ of Eqs. (8a) and (8b).

(2) Many of the terms such as $\mathbf{D}_s$, $\mathbf{H}_s$, $\mathbf{J}_s$, etc., have their support in the material surface $\phi(\mathbf{x},t) = 0$ and are identically zero elsewhere.

(3) From Eqs. (9a)–(9d) it is clear that there are kinematic relations between the tangential and normal components of the jumps in vector fields for general moving material interfaces. In the case of an electromagnetic shock wave,[9] these terms couple the shock to the vorticity of the field.[8]

(4) The general discontinuity equations given in Eqs. (9a)–(9d) show the source of growth/decay/scattering due to interactions with material boundaries.

III. DIRECT SCATTERING FOR ELECTROMAGNETIC IMPEDANCE THEORY, A CHAIN OF MAXWELL EQUATIONS

The geometry assumed in the remainder of this paper is shown in Figs. 1 and 2. A schematic of the scattering geometry is given in Fig. 1 where the source is contained in a two-

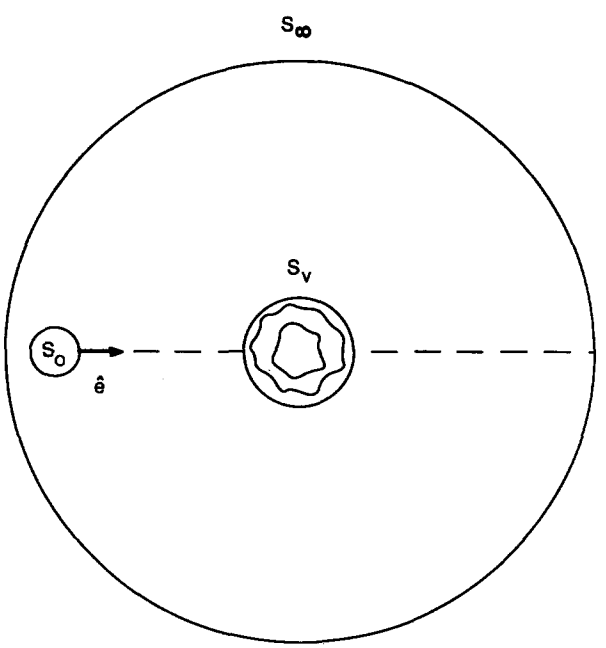


FIG. 1. A schematic of the geometry for the scattering process. The source is contained in the sphere S_0 and launches a wave or pulse in the fixed direction $\hat{e}$, the obstacle is in the sphere S_v and need not be spherically symmetric. The detector is free to move anywhere on the large sphere S_∞ centered at an origin in the obstacle. The scattering region is $S_c = S_\infty \setminus (S_0 \cup S_v)$.

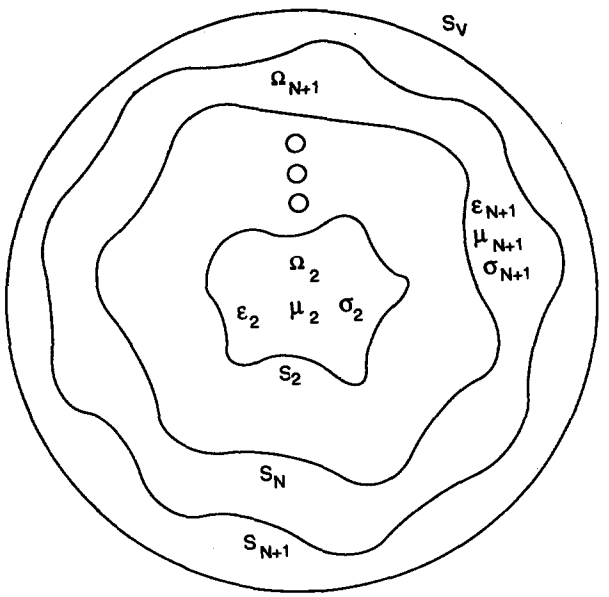


FIG. 2. The contents of the sphere S_v sketched in more detail, for each $i = 2,3,...,N+1$ each Ω_i contains material with variable electromagnetic parameters and with boundary $\partial\Omega_i = S_i$. The hypotheses on the media and their boundaries are given in the text as H1–H4.

sphere S_0; the scattering region, which need not be spherically symmetric, has its support properly contained in the two-sphere S_v and the detector can range anywhere over the surface of the arbitrarily large two-sphere S_∞. The "scattering region" $\Omega_{sc} = S_\infty \setminus (S_0 \cup S_v)$ contains a linear, isotropic, homogeneous dielectric material with constitutive parameters ϵ_1, μ_1. The scattering region contains N-smooth, orientable, nonintersecting surfaces indexed from 2 to $(N+1)$ with boundaries $S_i = \partial\Omega_i$ as shown in Fig. 2. The constitutive parameters in the ith region, Ω_i, $1 \leqslant i \leqslant N$ are written as $\alpha_i = \kappa_{ei}, \mu_i$ and vary spatially and satisfy hypothesis H1–H4 below. The interfaces $S_i = \partial\Omega_i$ can be parametrized some by C^2 functions $\phi_i(\mathbf{x},t) = 0$. The total surface is $S_{N+1} = U_{i=1}^{N+1} S_i$ and the total scattering volume is $\Omega_{N+1} = U_i^{N+1} Q_i$.

Let $\mathbf{A}$ denote $\mathbf{E}$ and $\mathbf{H}$ generically and in each Ω_i let α_i denote (κ_{ei}, μ_i) generically. Following Colton and Kress,[24] a hypothesis that will assure the existence of unique, complete scattering operators for the direct scattering problem in the frequency domain are given by H1–H4. The standard notation of Ref. 24 is used, where $C^{0,\alpha}(\Omega_N \setminus S_N)$, $C^{1,\alpha}(\Omega_N \setminus S_N)$ are the sets of functions uniformly Hölder continuous with index $0 < \alpha \leqslant 1$ and functions whose first derivations exist with respect to all variables and are uniformly Hölder continuous with index $0 < \alpha \leqslant 1$ in $\Omega_N \setminus S_N$, respectively. The *smooth tangential fields* on a surface S_i are given by

$$\Im(S_i) = \{\mathbf{A} | \hat{n} \cdot \mathbf{A} = 0, \mathbf{A} \in C(S_i)\}, \tag{10a}$$

where $\mathbf{A}: S_i \to C^3$ (C^1 complex numbers) and the *uniformly Hölder continuous vector fields* on a surface S_i are given by

$$\Im^{0,\alpha}(S_i) = \{\mathbf{A} | \mathbf{A} \in \Im(S_i), \mathbf{A} \in C^{0,\alpha}(S_i)\}, \tag{10b}$$

for $0 < \alpha \leqslant 1$. Using these spaces,[24] the conditions on the electromagnetic fields and the constitutive functions are assumed to satisfy the following hypothesis.

H1: The $\mathbf{A}$'s and α_i's have finite limits in $\Omega_N \setminus S_N$ and

$$\left\{\nabla \times \mathbf{A}, \nabla \cdot (\alpha_i \mathbf{A}), \left(\frac{\partial \alpha_i}{\partial t}\right)\right\} \in C^{0,\alpha}(\Omega_N \setminus S_N), \quad 0 < \alpha \leqslant 1.$$

H2: The surface sources $[\rho_i]$, $[\mathbf{J}_i] \in C^{0,\alpha}(S_i)$, with $0 < \alpha \leqslant 1$, for each S_i.

H3: Each $\alpha_i \in C^{0,\alpha}(S_i)$ in both tangential directions $\{\hat{t}_i, \hat{t}_2\}$ and $\alpha_i \in C^{1,\alpha}(S_i)$ with $0 < \alpha \leqslant 1$.

H4: The product of $\{G^+(\mathbf{x},\mathbf{y}) - G_0(\mathbf{x},\mathbf{y})\}$ and $(\hat{n} \cdot \nabla \times \mathbf{A}) | S_i, \mathbf{x}, \mathbf{y} \in S_i$, is a compact map from $\Im(S_i)$ into $\Im^{0,\alpha}(S_i)$.

Remarks: (1) There are no surface fields $\{\mathbf{E}_s, \mathbf{B}_s, \mathbf{D}_s, \mathbf{H}_s\}$ in Ref. 24 but if the above hypothesis is applied to more general media α_i with point surfaces the arguments are valid.

(2) These conditions can be gathered using Eqs. (9a)–(9d) into the surface operators of Colton and Kress, where H4 is related to their operator $\mathbf{N}$–$\mathbf{N}_0$. That operator is proven to be compact in Theorem 2.33 of Ref. 24.

(3) Time domain scattering theory is well known.[26–28,12,14] The outgoing solution $\mathbf{u}$ to a vector-valued wave equation subject to $t = 0$ initial conditions $\{\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x})\}$ has energy $\mathscr{E}$ in an exterior volume Ω_{sc} given by

$$\mathscr{E}(\mathbf{u}, \Omega_{sc}, 0) = \int_{\Omega_{sc}} \{|(\nabla \cdot \mathbf{f})(\mathbf{x})|^2 + |\mathbf{g}(\mathbf{x})|^2\} d^3\mathbf{x}. \tag{11a}$$

The conservation of energy principle is that for conservative systems for each $t \in R^1$

$$\mathscr{E}(\mathbf{u}, \Omega_{sc}, 0) = \mathscr{E}(\mathbf{u}, \Omega_{sc}, t) = c_1, \tag{11b}$$

where the constant c_1 can be finite or infinite. The physically interesting cases have finite energy $c_1 < \infty$ and propagate outward from the region S_v toward S_∞ in Fig. 1. A precise statement of outward propagation is that

$$\lim_{t \to \infty} \mathscr{E}(\mathbf{u}, K \cap \Omega_{sc}, t) = 0, \tag{11c}$$

for any compact set K, i.e., at long enough times the energy propagates out of any bounded set K. When the conditions in H1–H4 are placed upon the object inside S_v the behavior of the field $\mathbf{u}$ at large times is

$$\mathbf{u}(\mathbf{x},t)_{t \to \infty} \simeq [\mathbf{u}_0(\mathbf{x},t) + \mathbf{u}_{sc}(\mathbf{x},t)], \tag{11d}$$

where $\mathbf{u}_0$ is the incident wave and

$$u_{sc}(\mathbf{x},t) = R(t - |x|/c, \hat{x})/|\mathbf{x}| \tag{11e}$$

is the scattered wave with $R(\cdot,\cdot)$ as the unit impulse response. The unit impulse response includes both the time domain scattering amplitude and the outgoing wave. Uniqueness for the direct scattering problem follows from the Sommerfield radiation conditions in Eqs. (11d) and (11e). The existence is proven by either establishing unitary translation groups for $\mathbf{f}, \mathbf{g} \in L^2$ as in Ref. 26 or by placing conditions on the coefficients and proving local compactness. In the present discussion, the focus is on nonuniqueness, or am-

biguities, caused by electromagnetic media that do not satisfy the conditions. Both surfaces of discontinuity and temporal dispersion of the medium will be shown by yield ambiguities.

If the stationary nonconducting material in the region Ω_{sc} is linear, isotropic homogeneous then the curl of Eqs. (5a) and (6b) yields the vector-valued wave equations

$$\left\{\left(\Delta - \frac{1}{c_1^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{E}\right\}(\mathbf{x},t) = \mathbf{0} \tag{12a}$$

and

$$\left\{\left(\Delta - \frac{1}{c_1^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{H}\right\}(\mathbf{x},t) = \mathbf{0} \tag{12b}$$

where $c_1^2 = c^2/\mu_1\epsilon_1$ is the speed of the light wave in the medium and $\Delta = \nabla\cdot\nabla$ is the Laplacian. The scalar fundamental solution that satisfies outgoing boundary conditions and the equation,

$$\left(\Delta - \frac{1}{c_1^2}\frac{\partial^2}{\partial t^2}\right)G^+ = \delta(R)\delta(t-t'), \tag{12c}$$

and is given by

$$G^+(R,t-t') = -[\delta(t'-t-R/c_1)/4\pi R\,], \tag{12d}$$

where $R = |\mathbf{x}-\mathbf{x}|$. It is important to use this object instead of a dyad to get the direct scattering terms to show all discontinuities as Ström showed in Ref. 16, albeit where his presentation was in the frequency domain. Inside the scattering regions $\Omega_k \subset S_v$ the media depend upon $(\mathbf{x},t)$. The curl of Eqs. (5a) and (5b) for such media becomes

$$\Delta\mathbf{E} - \mu_\kappa\epsilon_\kappa\frac{\partial^2\mathbf{E}}{\partial t^2} = \nabla(\nabla\cdot\mathbf{E}) + \frac{\partial}{\partial t}\mathbf{J}_T \tag{13a}$$

and

$$\Delta\mathbf{H} - \mu_\kappa\epsilon_\kappa\frac{\partial^2\mathbf{H}}{\partial t^2} = \nabla(\nabla\cdot\mathbf{H}) + \nabla\times\mathbf{J}_T, \tag{13b}$$

where the total current density $\mathbf{J}_T$ was defined as

$$\mathbf{J}_T = \mathbf{J}_f + \mathbf{J}_p + \mathbf{J}_M.$$

The integral equations for scattering are derived from Eqs. (13a) and (13b) by multiplying them by $G^+(R,t-t')$ and subtracting $\mathbf{E}$ or $\mathbf{H}$ times Eq. (12d) and then integrating over the volume Ω_{sc} and time. Applying the radiation conditions, the terms with integrand given by $G^+\{\Box_{c_1}\mathbf{E}\} - \mathbf{E}\{\Box_{c_1}G^+\}$ and a similar term in $\mathbf{H}$ vanish at the limits. The integration over the surface of S_0 yields an incident wave $(\mathbf{E}_0,\mathbf{H}_0)$, analogous to the $\mathbf{u}_0$ term in Eq. (11d), from the source. All of this yields

$$\mathbf{E}(\mathbf{x},t) = \mathbf{E}_0 - \frac{1}{4\pi}\int_{\overline{\Omega}_i} d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla(\nabla\cdot\mathbf{E}) - \dot{\mathbf{J}}_T\}_{\text{ret}} \tag{14a}$$

and

$$\mathbf{H} = \mathbf{H}_0 - \frac{1}{4\pi}\int_{\overline{\Omega}_i} d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla(\nabla\cdot\mathbf{H}) - \nabla\times\mathbf{J}_T\}_{\text{ret}}, \tag{14b}$$

where $\dot{\mathbf{J}}_T = (\partial/\partial t)(\mathbf{J}_T)$ and "ret" means that in the term inside the curly bracket t' is evaluated at $t' = t - R/c$. If the crude, numerical approximations

$$\mathbf{D}(\mathbf{x},t) \approx \epsilon(\mathbf{x},t)\mathbf{E}(\mathbf{x},t), \tag{15a}$$

$$\mathbf{B}(\mathbf{x},t) \approx \mu(\mathbf{x},t)\mathbf{H}(\mathbf{x},t), \tag{15b}$$

which are discussed in the Appendix are used, Eqs. (14a) and (14b) can be rewritten as

$$\mathbf{E} \approx \mathbf{E}_0 - \frac{1}{4\pi}\int_{\overline{\Omega}_i} d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|} \times\left\{\nabla(\nabla\cdot\mathbf{E}) - \dot{\mathbf{J}}_f - \frac{\partial^2\epsilon}{\partial t^2}\mathbf{E} - 2\left(\frac{\partial\epsilon}{\partial t}\right)\left(\frac{\partial\mathbf{E}}{\partial t}\right)\right\}_{\text{ret}} \tag{16a}$$

and

$$\mathbf{H} \approx \mathbf{H}_0 - \frac{1}{4\pi}\int_{\overline{\Omega}_i} d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|} \times\left\{\nabla(\nabla\cdot\mathbf{H}) - \nabla\times\mathbf{J}_f - \left(\frac{\partial^2\mu}{\partial t^2}\right)\mathbf{H} + 2\left(\frac{\partial\mu}{\partial t}\right)\left(\frac{\partial\mathbf{H}}{\partial t}\right)\right\}_{\text{ret}}. \tag{16b}$$

The volume form of the integral equations for scattering for the $(N+1)$ media shown in Fig. 2 can be obtained by summing over the different scattering regions in Eq. (14). Under the assumptions made here these equations are

$$\mathbf{E} = \mathbf{E}_{\text{inc}} - \sum_{i=2}^{N+1}\left\{\frac{1}{4\pi}\int_{\overline{\Omega}_i} d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla(\nabla\cdot\mathbf{E}) - \dot{\mathbf{J}}_T\}_{\text{ret}}\right\} \tag{17a}$$

and

$$\mathbf{H} = \mathbf{H}_0 - \sum_{i=2}^{N+1}\left\{\frac{1}{4\pi}\int_{\overline{\Omega}_i} d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|} \times\{\nabla(\nabla\cdot H) - \nabla\times\mathbf{J}_T\}_{\text{ret}}\right\}. \tag{17b}$$

These volume equations over closed subvolumes $\overline{\Omega}_i$ have well-defined scattering operators uniformly in the direction of $\hat{x}$ as $|\mathbf{x}| \to \infty$ whenever the N-scattering media satisfy H1–H4. They are valid even when the regions $\Omega_2 - \Omega_{N+1}$ are either nonspherically symmetric or inhomogeneous. For homogeneous media there is a beautiful "T-matrix method" of Waterman[29] and many others, see, e.g., Refs. 16, 29, 30, which has made a number of new scattering calculations possible. In the case of a conductor, the electric field vanishes in the interior. However, the magnetic field need not vanish there, but its volume terms will be very small at frequencies at and above the radar region.

Still, Eqs. (17a) and (17b) *hide* two important structural features. One aspect is the discontinuities at the N interfaces. Using the hypothesis H1–H4 it is necessary to recognize that the $\mathbf{E}$ and $\mathbf{H}$ fields in the volume integrals over Ω_i now represent fields due to the continuous dependence of $(\nabla\times\mathbf{J}_T,\mathbf{J}_T)$ in the presence of the discontinuous interfaces S_i, i.e., all of the multiple scattering from all interfaces also included in these fields. The discontinuities will be written out explicitly. The second hidden aspect is that the different terms have completely different time—or frequency—dependence. Workers in geophysics and radar scattering have a useful classification of the type of terms that Sabatier[5] in-

troduced into impedance theory in Ref. 5. The description of these terms is the following: (1) The pure volume terms will be called *diffuse reflectors*; (2) surface terms involving discontinuities in the material parameters are called *soft reflectors*; and (3) surface terms involving gradients of discontinuities in the material parameters and fields are called *hard reflectors*. At short enough times (or high enough frequencies) the hard reflectors are called *specular* points in the scattering and dominate the signal. Similarly, the other terms will dominate low- and mid-range scattering processes:

$$\mathbf{E}=\mathbf{E}_0+\frac{1}{4\pi}\sum_{i=2}^{N+1}\int_{\Omega_i}d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\dot{\mathbf{J}}_T\}_{\text{ret}}-\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla'[\epsilon]E_n)\}_{\text{ret}}$$
$$+\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\left\{\frac{\partial}{\partial t}[\mathbf{J}_{fi}]\right\}_{\text{ret}}-\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\left\{\frac{\partial^2}{\partial t^2}[\mathbf{P}_i]\right\}_{\text{ret}}$$
$$-\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\left\{\frac{\partial}{\partial t}(\nabla'x[\mathbf{M}_i])\right\}_{\text{ret}} \tag{18a}$$

and

$$\mathbf{H}=\mathbf{H}_0+\frac{1}{4\pi}\sum_{i=2}^{N+1}\int_{\omega_i}d^3\mathbf{x}'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla'\times\mathbf{J}_T\}_{\text{ret}}-\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla'([\mu]^{-1}\cdot B_n)\}_{\text{ret}}$$
$$-\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla'\times[\mathbf{J}_{fi}]\}_{\text{ret}}-\frac{1}{4\pi}\sum^{N+1}\oint_{S_i}d^2S_i\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\left\{\nabla'\times\left(\frac{\partial}{\partial t}[\mathbf{P}_i]\right)\right\}_{\text{ret}}$$
$$-\frac{1}{4\pi}\sum_{i=1}^{N+1}\oint_{S_i}d^2S_i'\,\frac{1}{|\mathbf{x}-\mathbf{x}'|}\{\nabla'\times(\nabla'\times[\mathbf{M}_i])\}_{\text{ret}}. \tag{18b}$$

Now, Eqs. (17a) and (17b) have been rewritten to expose the two kinds of scattering processes in Eqs. (18a) and (18b). The three volume terms on the right-hand side are the diffuse scatterers, the next term is a soft reflector and the remaining two terms are hard reflectors. At radar and higher frequencies, the volume scattering terms will be negligible for conductors, zero for **E**, and numerically very small for **H**. The Fourier transforms of these equations have Fourier transforms which agree with Ström in Ref. 16. Equations (18a) and (18b) are a chain of Maxwell scattering equations linked by the surfaces S_i.

At least for small motions, the normal components of the terms in Eqs. (9a)–(9d) generalize Eqs. (18a) and (18b) to include moving surfaces.[31] The Truesdell–Toupin equations, here Eqs. (8a) and (8b) remain valid but Costen's expressions give the explicit form of $\mathbf{k}_{TT}$ and f. This means that the tangential fields of hypothesis H4 and any evanescent waves are more complicated and more interesting. In Eqs. (18a) and (18b) this simply adds new surface terms that are *links* in the chain of Maxwell scattering equations presented there.

Various experiments in the physics and chemistry of solids and liquids show that from the UHF ($\sim 10^8$ Hz) to microwave (3×10^{11} Hz) frequencies, the molecular dipoles of the matter respond strongly to electromagnetic waves. For the generalized dielectric response of Eq. (6) ionic and electronic response continues to (and well past) the optical region. On the other hand, the magnetic permeability μ for most materials becomes and remains constant (one in Gaussian units) at optical frequencies and above.[31] It has been known for a long time[31] that many dielectric solids have an approximate Lorentzian permittivity $\epsilon(\omega)\approx[(\omega^2-\omega_0^2)+i\omega\gamma]^{-1}$. This follows directly from the Fourier transform of complex exponential time dependence of the polarization or magnetization. In small electron beams and plasmas, very different time-dispersion relations for $\epsilon(\omega)$ and $\mu(\omega)$ are well known.[32] Thus there is a frequency window from UHF to midinfrared where time dispersion is physically important and classical electromagnetism is valid. In this case the polarization **P** and the magnetization **M** vary with t or ω.

The existence and uniqueness of far-field scattering operators follow from the conditions on the polarization and magnetization vectors and the compactness of the surface operators and follows from mimicking the proofs of Lax and Phillips[26] for the direct scattering problem and Stefanov[27,28] for the inverse problem.

The next section will apply these results to electromagnetic impedance ambiguities. Both geometrical effects due to spatial discontinuities, analogous to Sabatier's acoustic results in Refs. 1–5; and a new kind of ambiguity due to dispersion experimentally known at UHF, microwave frequencies and above, are exhibited.

IV. AMBIGUITIES IN INVERSE ELECTROMAGNETIC IMPEDANCE THEORY

Two natural questions that motivated this study are whether *ambiguities* exist in electromagnetism and, if so,

what is their structure? The transverse structure of Maxwell's equations is very different from the longitudinal acoustics for which Sabatier has presented a thorough, complete study. The same geometrical situation as the previous section will be considered where an obstacle with variable μ_2, and ϵ_2 is embedded in a constant host material μ_1,ϵ_1.

Two independent scaling functions $\epsilon_s(\cdot)$ and $\mu_s(\cdot)$ are required for electromagnetic impedance theory since there are two fields (**E**,**B**). The scaling relations between pairs of electromagnetic fields (**E**,**B**) and ($\mathbf{E}^{\sim}$,$\mathbf{B}^{\sim}$) is given by

$$\mathbf{E}^{\sim} = \epsilon_s \mathbf{E}, \tag{19a}$$

$$\mathbf{B}^{\sim} = \mathbf{B}/\mu_s, \tag{19b}$$

$$\mathbf{J}_{\tilde{T}} = \mathbf{J}_T/\mu_s, \tag{19c}$$

and

$$\rho^{\sim} = \epsilon_s \rho. \tag{19d}$$

There are then three types of ambiguities depending upon whether the scaling functions depend on **x** only, t only, or both **x** and t. Physically, these three cases correspond to spatial boundaries, temporal dispersion, and combinations of boundaries and dispersion, respectively. All are taken at a fixed angle of incidence.

It is easiest to discuss electromagnetic ambiguities if three simple calculations are presented. The first is for the exterior region Ω_{sc} where the curl of the two curl Maxwell equations can be written as

$$\Delta \mathbf{E} - \epsilon_1 \mu_1 \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mu_1 \left(\frac{\partial \mathbf{J}_T}{\partial t} \right) + \nabla(\nabla \cdot \mathbf{E}) \tag{20a}$$

and

$$\Delta \mathbf{B} - \epsilon_1 \mu_1 \frac{\partial^2 \mathbf{B}}{\partial t^2} = -\mu_1 (\nabla \times \mathbf{J}_2). \tag{20b}$$

In an interior region with electric and magnetic response functions $\epsilon_2(\cdot)$ and $\mu_2(\cdot)$ there are three cases. The first is spatially varying media $\epsilon_2(\mathbf{x})$ and $\mu_2(\mathbf{x})$ in $c^2(R^3 \setminus S_2)$ with possible integrable discontinuities on the boundary $S_2 = \partial\Omega_2$. The scaling functions will also be taken to depend on **x** only. This was given for unscaled fields in Eqs. (17a) and (17b) and the integral equations for scattering were presented in Eqs. (18a) and (18b). The idea is the one advanced by Sabatier in Ref. 1, except that two scaling functions ϵ_s and μ_s are required. First substitute the scaled fields and sources into Maxwell's equations and take the curl of each side. Rearrange this result into wave equations and the pointwise condition $\epsilon_s \mu_s = 1$ will be required for equal phase velocities. The equality of the remaining terms then give the families of ambiguities. Maxwell's equations for the scaled fields in the interior region yield the wave equations

$$\Delta \mathbf{E} - \frac{\epsilon_2 \mu_2}{\epsilon_s \mu_s} \frac{\partial^2 \mathbf{E}}{\partial t^2} = \frac{\rho}{\epsilon_s \epsilon_2} (\nabla \epsilon_s) + \frac{1}{\epsilon_s} \nabla \frac{\rho}{\epsilon_s} + \frac{\mu_2}{\epsilon_s \mu_s} \dot{\mathbf{J}}_T + \frac{1}{\epsilon_s} \nabla((\nabla \epsilon_s) \cdot \mathbf{E}) - \frac{1}{\epsilon_s} (\Delta \epsilon_s)\mathbf{E} - \frac{1}{\epsilon_s \mu_s^2} (\nabla \mu_s) \times (\nabla \times \mathbf{E}) \tag{21a}$$

and

$$\Delta \mathbf{B} - \epsilon_2 \mu_2 \epsilon_2 \mu_s \frac{\partial^2 \mathbf{B}}{\partial t^2} = +\left(\frac{\mu_s}{\epsilon_2} - \mu_2 \right) \nabla \times \mathbf{J}_T - \mu_s \nabla \left(\frac{\mu_2}{\mu_s} \right) \times \mathbf{J}_T + \frac{4}{\mu_s^2} (\nabla \mu_s)^2 \mathbf{B} + \frac{2}{\mu_s} (\Delta \mu_s)\mathbf{B} - \frac{1}{\mu_s} (\nabla \mu_s) \times \mathbf{B} - \frac{\mu_s}{\epsilon_2 \mu_s} \nabla(\epsilon_2 \mu_s \mu_s) \times (\nabla \times \mathbf{B}), \tag{21b}$$

for the case where the scaling functions depend only upon **x**. The second case occurs when the physical response functions (ϵ_2, μ_2) and the scaling functions (ϵ_s, μ_s) depend only on time. This would describe an isotropic bulk medium with strong dispersion. The wave equations in this medium are

$$\Delta \mathbf{E} - \frac{\epsilon_2 \mu_s}{\epsilon_s \mu_s} \frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla \left(\frac{\rho}{\epsilon_2} \right) + \frac{1}{\epsilon_s \mu_s} \frac{\partial}{\partial t} (\mu_s \mathbf{J}_T) - \frac{\mu_2}{\epsilon_s \mu_s^2} \left(\frac{d \mu_s}{dt} \right) \mathbf{J}_T + \left\{ \frac{1}{\epsilon_s \mu_s} \frac{\partial}{\partial t} (\epsilon_2 \mu_2) - \frac{\epsilon_2 \mu_2}{\epsilon_s \mu_s^2} \left(\frac{d \mu_s}{dt} \right) \right\} \left(\frac{\partial \mathbf{E}}{\partial t} \right) \tag{22a}$$

and

$$\Delta \mathbf{B} - \epsilon_2 \mu_2 \epsilon_s \mu_s \left(\frac{\partial^2 \mathbf{B}}{\partial t^2} \right) = -\mu_2 \nabla \times \mathbf{J}_T - \epsilon_s \mu_2 \mu_s \left(\frac{\partial \epsilon_s}{dt} \right) \left(\frac{\partial \mathbf{B}}{\partial t} \right). \tag{22b}$$

The wave equations for media with space and time dependence contain a new complication, the **E** and **B** fields are coupled. This coupling occurs through nonzero electric and magnetic pressure terms $\mathbf{p}_e$ and $\mathbf{p}_m$:

$$\mathbf{p}_e = \nabla \left(\frac{\partial \epsilon_s}{\partial t} \right) \tag{23a}$$

and

$$\mathbf{p}_m = \nabla \left(\frac{\partial \mu_s}{\partial t} \right). \tag{23b}$$

These wave equations are given by

$$\Delta \mathbf{E} - \frac{\epsilon_2 \mu_2}{\epsilon_s \mu_s} \left(\frac{\partial^2 \mathbf{E}}{\partial t^2} \right) = \frac{1}{\epsilon_s \mu_s} \frac{\partial}{\partial t} (\mu_2 \mathbf{J}_T) + \nabla \left(\frac{\rho}{\epsilon_s \epsilon_2} \right) - \frac{1}{\epsilon_s \mu_s} \frac{\partial}{\partial t} (\epsilon_2 \mu_s) \times \left(\frac{\partial \mathbf{E}}{\partial t} \right) - \frac{1}{\epsilon_s} (\nabla \epsilon_s) \times \frac{\partial \mathbf{E}}{\partial t} - \frac{1}{\epsilon_s \mu_s^2} \nabla \left(\frac{\partial \mu_s}{\partial t} \right) \times \mathbf{B} \tag{24a}$$

and

$$\Delta \mathbf{B} - \epsilon_2 \mu_s \epsilon_s \mu_s \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\mu_s \nabla \times (\mu_s \mathbf{J}_T) - \frac{\mu_s}{\epsilon_2} \nabla(\epsilon_2 \epsilon_s \mu_s) \times \mathbf{J}_T$$

$$-\epsilon_2\mu_2\mu_s\left(\frac{\partial\epsilon_s}{\partial t}\right)\left(\frac{\partial\mathbf{B}}{\partial t}\right)+\frac{2}{\mu_s}\left\{\frac{2}{\mu_s}(\nabla\mu_s)^2-(\Delta\mu_s)\right\}\mathbf{B}$$
$$-\frac{\mu_s}{\epsilon_2\mu_2}\nabla(\epsilon_2\epsilon_s\mu_s)\times(\nabla\times\mathbf{B})$$
$$-\mu_s\nabla\left(\epsilon_2\mu_s\left(\frac{\partial\epsilon_s}{\partial t}\right)\right)\times\mathbf{E}. \tag{24b}$$

The pressure coupling terms mentioned are the last term in each equation.

Lemma 1: In time-dependent, linear, stationary media that vary spatially, $\{\epsilon_2(\mathbf{x}),\mu_2(\mathbf{x})\}$, the scaling functions $\{\epsilon_s,\mu_s\}$ depend only upon $\mathbf{x}$ and all families of sources $\{\mathbf{J}_T,\rho\}$ and $\{\mathbf{J}_{\bar{T}},\rho^{\sim}\}$ that satisfy hypothesis H1–H4 together with the equations

$$\mu_2\left(\frac{\partial\mathbf{J}_T}{\partial t}\right)+\nabla\left(\frac{\rho}{\epsilon_2}\right)$$
$$=\mu_2\left(\frac{\partial\mathbf{J}_{\bar{T}}}{\partial t}\right)+\nabla\left(\frac{\rho^{\sim}}{\epsilon_2}\right)+\rho(\nabla\epsilon_s)+\frac{1}{\epsilon_s}\nabla((\nabla\epsilon_s)\cdot\mathbf{E})$$
$$-\frac{1}{\epsilon_s}(\Delta\epsilon_s)\mathbf{E}-\frac{1}{\mu_s}(\nabla\mu_s)\times(\nabla\times\mathbf{E}) \tag{25a}$$

and

$$\mu_2(\nabla\times\mathbf{J}_T)=\left(\frac{\mu_s}{\epsilon_2-\mu_2}\right)(\nabla\times\mathbf{J}_{\bar{T}})$$
$$+\frac{2}{\mu_s}\left[\frac{1}{\mu_s}(\nabla\mu_s)^2+(\Delta\mu_s)\right]\mathbf{B}$$
$$-\frac{1}{\mu_s}(\nabla\mu_s)\cdot\mathbf{B}-\mu_s\nabla(\epsilon_2\mu_2\mu_s)\times(\nabla\times\mathbf{B}), \tag{25b}$$

with

$$\epsilon_s(\mathbf{x})\mu_s(\mathbf{x})=1 \tag{25c}$$

continuous tangential fields (H3), response functions that satisfy $\alpha_i(\mathbf{x}^-)=\lambda_i\alpha_i(x_i^+)$ for each $\mathbf{x}\in\text{supp}\,(S)$, S the surface of discontinuity are *ambiguous.*

Proof: Straightforward calculation using Eqs. (19a)–(19d) in Eqs. (5a)–(5d).

Lemma 2: In a spatially homogeneous, time-dependent linear, stationary obstacle and all families of sources $\{\mathbf{J}_T,\rho\}$ and $\{\mathbf{J}_{\bar{T}},\rho^{\sim}\}$ which satisfy hypothesis H1–H4 in a medium with constitutive tensor $\{\epsilon_2,\mu_s\}$ and scaling functions that depend only upon time, if $\epsilon_s\mu_s=1$ pointwise and the equations

$$\nabla\left(\frac{\rho}{\epsilon_2}\right)+\frac{\partial}{\partial t}(\mu_2\mathbf{J}_T)$$
$$=\nabla\left(\frac{\rho^{\sim}}{\epsilon_2}\right)+\frac{\partial}{\partial t}(\mu_2\mathbf{J}_{\bar{T}})-\frac{\mu_2}{\mu_s}\left(\frac{d\mu_s}{dt}\right)\mathbf{J}_{\bar{T}}$$
$$+\left\{\frac{\partial}{\partial t}(\epsilon_2\mu_2)-\epsilon_2\mu_2\left(\frac{d\mu_s}{dt}\right)\right\}\left(\frac{\partial\mathbf{E}}{\partial t}\right), \tag{26}$$

are *ambiguous.*

Proof: Straightforward calculation using Eqs. (19a)–(19d) in Eqs. (5a)–(5d).

Lemma 3: In time-dependent, spatially inhomogeneous, stationary obstacle $\{\epsilon_2(\mathbf{x},t),\mu_2(\mathbf{x},t)\}$ the scaling functions $\{\epsilon_s\mu_s\}$ depend upon $\mathbf{x}$ and t and all families of sources $\{\mathbf{J}_T,\rho\}$ and $\{\mathbf{J}_{\bar{T}},\rho\sim\}$ that satisfy hypothesis H1–H4 together with the equations $\epsilon_s\mu_s=1$ pointwise and

$$\frac{\partial}{\partial t}(\mu_2\mathbf{J}_T)-\nabla\left(\frac{\rho}{\epsilon_2}\right)-\left\{\frac{\partial}{\partial t}(\epsilon_2\mu_2)\right\}\left(\frac{\partial\mathbf{E}}{\partial t}\right)$$
$$=\frac{\partial}{\partial t}(\mu_2\mathbf{J}_{\bar{T}})-\nabla\left(\frac{\rho^{\sim}}{\epsilon_2}\right)-\left\{\frac{\partial}{\partial t}(\epsilon_2\mu_2)\right\}\left(\frac{\partial\mathbf{E}}{\partial t}\right)$$
$$-\frac{1}{\epsilon_2}(\nabla\epsilon_s)\times(\nabla\times\mathbf{E})-\frac{1}{\mu_s}\nabla\left(\frac{\partial\mu_s}{\partial t}\right)\times\mathbf{B} \tag{27a}$$

and

$$\nabla\times(\mu_2\mathbf{J}_T)+\frac{1}{\epsilon_2}\nabla(\epsilon_2\mu_2)\times\mathbf{J}_T+\frac{1}{\epsilon_2\mu_2}\nabla(\epsilon_2\mu_2)\times(\nabla\times\mathbf{B})$$
$$=\mu_s\nabla\times(\mu_2\mathbf{J}_{\bar{T}})+\frac{\mu_s}{\epsilon_2}\nabla(\epsilon_2\epsilon_s\mu_2)\times\mathbf{J}_{\bar{T}}$$
$$+\frac{2}{\mu_s}\left\{\frac{2}{\mu_s}(\nabla\mu_s)^2-(\Delta\mu_s)\right\}\epsilon_2\mu_2\mu_s\left(\frac{\partial\epsilon_s}{\partial t}\right)\left(\frac{\partial\mathbf{B}}{\partial t}\right)$$
$$+\frac{\mu_s}{\mu_2\epsilon_2}\nabla(\epsilon_2\epsilon_s\mu_2)\times(\nabla\times\mathbf{B})+\mu_s\nabla\left(\epsilon_2\mu_2\frac{\partial\epsilon_s}{\partial t}\right)\times\mathbf{E} \tag{27b}$$

are *ambiguous.*

Proof: Straightforward calculation using Eqs. (19a)–(19d) in Eqs. (5a)–(5d).

From these three lemmas, the *Maxwell equivalence* of different media that give identical near-field or far-field scattering follow. One *standard equivalence* is the special case of Lemma 1 when (ϵ_s,μ_s) are nowhere vanishing with

$$\epsilon_s\mu_s=1, \tag{28}$$

nontrivially, pointwise, and when Eqs. (25a) and (25b) hold. Additional cases are given in Lemmas 2 and 3.

The above ambiguities are exact at fixed angle of incidence and occur in principle. When the data are bandlimited and noisy, as discussed for acoustics in Ref. 2, the following *homogenization* can occur. Many weak reflectors or fewer strong reflectors may approximately mimic one another or a diffuse reflector. Only in the limit as noise goes to zero and the filter goes away (i.e., the data become complete) do these possibilities become distinguishable. At the same time, the *homogenization* can be used to simplify an inverse problem as follows. Uniqueness is already lost when there are ambiguities so one need seek only a simple member of the ambiguous family. The problem is that some mathematical structure such as the one-dimensional Darboux transform in Ref. 3 is needed to know when two sets of sources belong to a single family.

V. CONCLUSIONS AND FUTURE STUDIES

Two classes of ambiguities for classical, linear, electromagnetic scattering at fixed angle of incidence have been given. One in Lemma 1 involves sharp interfaces and is a direct generalization of Sabatier's impedance studies.[1-7] The second given in Lemma 2 involves time fluctuations and

time dispersion and becomes important at UHF, microwave, and infrared frequencies. In general, these are *Maxwell equivalence classes* in the special case of Eqs. (28a) and (28b) and they reduce to *electromagnetic standard equivalence classes*. These latter ambiguities can occur under the stated conditions in light scattering from molecules, liquids, and solids.

In addition, chains of Maxwell's equations were presented in Eqs. (18a) and (18b) and Eqs. (28a) and (28b). The more general boundary conditions in Eqs. (19a)–(19d), give a *new* chain of Maxwell equations with the additional terms from the normal component of the new surface fields $\mathbf{E}_s$ and $\mathbf{H}_s$ to be studied at fixed incident angle.

By inspecting Refs. 1–5c it is clear that much remains to be done before electromagnetic problems will be as well understood as acoustics problems. In order of importance, these include (1) a path integral to semiclassically couple various waves at boundaries; (2) the ambiguities reexpressed in terms of the boundary operators of Ref. 24 and the roles of the single layer and double layer potentials established; (3) a full-space Green's distribution, and (4) a perturbation theory for Eqs. (18a) and (18b). Of these, only number (1) is completed.[33]

ACKNOWLEDGMENTS

This work was supported by NATO Grant 095/87, AFOSR 89-0311, the French CNRS at the Laboratory for Physical Mathematics at USTL in Montpellier during the summer 1987, and the University of Missouri at Columbia.

Useful conversation with Professor P. C. Sabatier at USTL, who introduced the author to his work are gratefully acknowledged. Professor R. J. Krueger first focused the author's interest on discontinuities about a decade ago. This small paper is dedicated to the memory of Dr. Bob Krueger, a friend and a scholar, who will be missed. Professor W.-M. Boerner is thanked for providing the author copies of the nice works of Professor E. M. Kennaugh on polarized radar scattering at Ohio State University between 1946–1977 and Professor C. Truesdell is thanked for a useful letter and a copy of Ref. 23. Dr. R. Albanese, Dr. P. Rudolph, and Professor K. Oughnun are thanked for raising the questions that led to a correction in Sec. III and the Appendix.

APPENDIX

The decision to study time dispersion requires a more careful discussion of the structure of the constitutive relations than the usual superficial treatments. This is required by Physics, and even Biology. Thus it is necessary to take into account the general treatment in Refs. 15 and 20.

In linear electrodynamics of isotropic media the *material equation* gives the relation between the electric displacement $\mathbf{D}_t$ and the electric intensity $\mathbf{E}$ as

$$\mathbf{D}(\mathbf{r},t)=\int_{-\infty}^{t} dt' \int d^3\mathbf{r}'\epsilon(t,t',\mathbf{r},\mathbf{r})\mathbf{E}(\mathbf{r},t'), \tag{A1}$$

where the kernel ϵ is a complex-valued dielectric response gives the physical meaning of the medium, the constitutive equations. The principle of causality is satisfied by Eq. (A1) because the electric displacement $\mathbf{D}$ at position $\mathbf{r}$ and time t depends on the dielectric response of the medium ϵ as a kernel and the electric field $\mathbf{E}$ at all times $t' \leqslant t$. This is the classical causality which states that the effect in $\mathbf{D}(t)$ must follow from the cause $\mathbf{E}(t')$ only at earlier times t'. *Stability* is another general principle that all materials must satisfy, *objectivity* is a fundamental principle that could be neglected for the isotropic media in a single frame considered. In anisotropic media or multiple frames it is a powerful restriction. If the medium is spatially homogeneous, the dielectric response kernel depends only on $\mathbf{r}-\mathbf{r}'$ and time independent media depend only on $t-t'$. The dielectric response is a complex-valued function with real and imaginary parts ϵ' and ϵ'',

$$\epsilon=\epsilon'+i\epsilon''. \tag{A2}$$

If $\epsilon(t-t',\cdot)$ is in $L^2(R^3\times R^1)$, its Fourier transform $\hat{\epsilon}(\omega,\cdot)$ automatically exists so assume this for the rest of this appendix. The causality of a spatially homogeneous medium that is temporally inhomogeneous provides a sharp cutoff for all $t'>t$. This guarantees that the Fourier transform $\hat{\epsilon}(\omega)$ of $\epsilon(t-t')$ is half-plane analytic, and that the real and imaginary parts of this Fourier transform,

$$\epsilon(\omega)=\hat{\epsilon}'(\omega)+i\hat{\epsilon}''(\omega), \tag{A3}$$

are a Hilbert transform pair. These Hilbert transforms are called the Kramers–Kronig relations of the dielectric response function in physics. The neglect of spatial inhomogeneity is a good approximation for nonconducting scatterers at the frequencies considered here because the size of the regions in which $\mathbf{E}$ determines $\mathbf{D}$ depends on if b is of order one Å and the wavelength λ is greater than 10^{-2} m so that (b/λ) is negligible.

There is an interesting subtlety when the dielectric response $\hat{\epsilon}(\omega,\mathbf{k})$ has non-negligible $\mathbf{k}$ dependence which Pines and Nozieres[34] first pointed out and which Dolgov, Krirzhnits, and Maksimov[35] have reviewed. The subtle point is that the causality principle cannot imply the Hilbert transform relations between the real and imaginary parts of $\hat{\epsilon}(\omega,\mathbf{k})$. This is because $\hat{\epsilon}$ is the response to the *total field* which is not finite, in general. The *external field* can be controlled and hence the real and imaginary parts of $[\epsilon(\omega,\mathbf{k})]^{-1}$ satisfy Hilbert transform relations. On the other hand both $\epsilon(\omega)$ and $[\epsilon(\omega)]^{-1}$ satisfy Kramers–Kronig relations, which may be the source of this widespread error. The convenient relation

$$\mathbf{D}(\mathbf{r},t)\simeq\epsilon(t,\mathbf{r})\mathbf{E}(\mathbf{r},t), \tag{A4}$$

was given between Eqs. (1.4) and (1.5) in an unnumbered equation in Ref. 20 and is stated there to have "only a symbolic significance" and to hold "only in exceptional cases (and then only approximately)." Stated another way, dispersion always has memory.[36]

[1]P. C. Sabatier, Inv. Probs. **3**, 296 (1987); **4**, L1 (1988).
[2]P. C. Sabatier and B. Dolveck-Gulipard, J. Math. Phys. **29**, 861 (1988).
[3]A. Degasperis and P. C. Sabatier, Inv. Probls. **3**, 73 (1987).
[4]P. C. Sabatier, in *Nonlinear Evolutions*, edited by J. León (World Scientific, Singapore, 1988).
[5]P. C. Sabatier, J. Math. Phys. **30**, 2585 (1989).

[6] A. Martin and P. C. Sabatier, J. Math. Phys. **18**, 1623 (1977).
[7] P. C. Sabatier, C. R. Acad. Sci. Paris Ser. B **278**, 603 (1974).
[8] R. C. Costen, J. Math. Phys. **22**, 1377 (1981).
[9] G. Boillat, J. Math. Phys. **11**, 941 (1970).
[10] P. E. Parker, J. Math. Phys. **20**, 1423 (1979).
[11] (a) J. A. Applebaum and D. R. Hamann, Phys. Rev. B **6**, 2166 (1972), seem to have published the first fully self-consistent calculation of an electronic surface state. Their work was based on the ideas in the paper; (b) W. Kohn and L. J. Sham, Phys. Rev. A **140**, 1133 (1965).
[12] D. S. Jones, *Methods in Electromagnetic Wave Propagation, Vols. I, II* (Oxford U. P., Oxford, 1987).
[13] J. D. Jackson, *Electromagnetism 2/E* (Wiley, New York, 1975).
[14] M. Kline and I. W. Kay, *Electromagnetic Theory and Geometrical Optics* (Krieger, Huntington, NY, 1979).
[15] C. Truesdell and R. Toupin, Handbuch Phys. **III/I**, 491 (1960).
[16] S. Ström, Am. J. Phys. **43**, 1060 (1975).
[17] R. J. Krueger, J. Math. Phys. **23**, 396 (1982); Q. Appl. Math. **34**, 129 (1976).
[18] A. C. Eringen, J. Math. Phys. **25**, 717 (1984).
[19] S. R. de Groot and L. G. Suttorp, *Foundations of Electrodynamics* (North-Holland, Amsterdam, 1972).
[20] V. M. Agranovich and V. L. Ginzburg, "Crystal Optics with Spatial Dispersion," in *Progress in Optics Vol. IV*, edited by E. Wolf (North-Holland, Amsterdam, 1971), pp. 235–280.
[21] F. N. H. Robinson, *Macroscopic Electromagnetism* (Pergamon, Oxford, 1973).
[22] A. M. Portis, *Electromagnetic Fields, Sources, and Media* (McGraw-Hill, New York, 1978).
[23] R. P. Kanwal and C. Truesdell, Phys. Fluids **5**, 368 (1962).
[24] D. Colton and R. Kress, *Integral Equations Methods in Scattering Theory* (Wiley, New York, 1983).
[25] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, translated by J. B. Sykes and J. S. Bell (Pergamon, Oxford, 1960), pp. 251–253. Note that the conclusion on p. 253 is wrong for magnetic semiconductors (discovered well after this book was written) but remains correct for other materials.
[26] P. D. Lax and R. S. Phillips, *Scattering Theory* (Academic, New York, 1967).
[27] P. D. Stefanov, Inv. Probs. **4**, 921 (1988).
[28] P. D. Stefanov, Inv. Probs. **4**, 913 (1988).
[29] P. C. Waterman, Phys. Rev. D **3**, 825 (1971).
[30] G. Kristensson, J. Appl. Phys. **51**, 3486 (1980).
[31] P. Penfield, Jr. and H. A. Haus, *Electrodynamics of Moving Media* (MIT, Cambridge, MA, 1967).
[33] B. DeFacio and O. Brander, Rend del Circ Mat Di Palermo **17**, (Series II) 185 (1987); in *Nonlinear Evolutions*, edited by J. León (World Scientific, Singapore, 1988), pp. 750–766 and in preparation.
[34] D. Pines and P. Noziéres, *The Theory of Quantum Liquids* (Benjamin, New York, 1966).
[35] O. V. Dolgov, D. A. Kirzhnits, and E. G. Maksimov, Rev. Mod. Phys. **53**, 81 (1981).
[36] J. Corones and A. Karlsson, Inv. Probs. **4**, 643 (1988).

On the modification of the Stark states of hydrogen by a weak ac electric field

Viorica Florescu and Suzana Patrascu
Faculty of Physics, University of Bucharest, P.O. Box 5211, Bucharest-Magurele, 76900, Romania

(Received 29 September 1989; accepted for publication 11 April 1990)

A derivation is presented for the compact integral representation that describes the first-order perturbed wavefunction of an electron in a fixed Coulomb field and in a harmonic uniform electric field, for the case of an initial stationary Stark state with the symmetry axis along the electric field.

I. INTRODUCTION

This paper considers the effect of a weak harmonic uniform electric field on the Stark states of an electron in the Coulomb field. The Stark states, found by solving the Schrödinger equation in parabolic coordinates, were obtained by Schrödinger at the very beginning of quantum mechanics.[1] Today these states can be selected and directly used in experiments: Bayfield and Pinnaduwage[2] produced and investigated for the first time the extreme Stark states with principal quantum number $n = 60$ and electric quantum number $n_e = -59$.

We present here our work concerning the first-order perturbed wavefunction of a hydrogenic atom (with fixed nucleus) in an ac electric field. The initial state of the electron is a Stark state described by the parabolic quantum numbers n_1, n_2, and m. The electric field is switched on adiabatically.

Independent calculations for this problem have been performed recently by Marian.[3] Our results for the perturbed parabolic states agree with those in Ref. 3, but the calculations differ essentially in their details and provide different insights into the underlying theory. In our paper we shall stress the similarities and the differences of the two calculations. Both calculations are based on the results obtained recently[4] for the linear modification of a bound "spherical" state of hydrogen (arbitrary values for the quantum numbers nlm).

In Sec. II, based on Ref. 4, we write the general equations for the first-order perturbed spherical wavefunctions in a way that is convenient for the construction of the perturbed parabolic states. Section III presents the main steps of our derivation. We also point out two new relations [Eq. (21)] for the coefficients connecting spherical and parabolic states of an electron in the Coulomb field. In Sec. IV we comment on some properties of the linear response of the parabolic states and on its application in the study of some radiative processes.

II. THE FIRST-ORDER PERTURBED SPHERICAL STATES

We consider an electron in an initial stationary state of energy E_n,

$$\psi_{\rm in}^{(0)}(\mathbf{r},t) = u_{\rm in}(\mathbf{r})\exp(-(i/\hbar)E_n t), \tag{1}$$

where $u_{\rm in}(\mathbf{r})$ is uniquely characterized by a set of quantum numbers that will be specified later. A harmonic uniform electric field,

$$\mathscr{E} = \mathscr{E}_0 \cos\omega t, \tag{2}$$

is switched on adiabatically. We describe the electric field by the potentials

$$\mathbf{A} = -(\mathscr{E}_0/\omega)\sin\omega t, \quad \phi = 0, \tag{3}$$

so, in the first order, the perturbation is

$$H' = (e/m_e)\mathbf{A}\cdot\mathbf{P},$$

with e the elementary charge ($e>0$), m_e the electron mass, and $\mathbf{P}$ the momentum operator. The calculation of the response of the hydrogen atom to a harmonic uniform electric field (2) is equivalent to the calculation of the dipole approximation of the response to a monochromatic electromagnetic plane wave.

The first-order modification of the wavefunction, which has the simple time-dependence

$$\psi_{\rm in}^{(1)}(\mathbf{r},t) = [f_{\rm in}^{+}(\omega,\mathbf{r})\exp(i\omega t) - f_{\rm in}^{-}(\omega,\mathbf{r}) \times\exp(-i\omega t)]\exp(-(i/\hbar)E_n t), \tag{4}$$

was studied for the first time by Podolsky.[5] Closed-form expressions for Podolsky's functions $f^{\pm}$ for the ground state case were given much later by Luban and co-workers.[6] Less known is the unpublished work of Johansson.[7]

The functions $f_{\rm in}^{\pm}$ for a given initial state of energy E_n can both be obtained from a single function, denoted $F_{\rm in}(\Omega,\mathbf{r})$, taken for different values of the parameter Ω,

$$f_{\rm in}^{\pm}(\omega,\mathbf{r}) = F_{\rm in}(E_n \mp \hbar\omega,\mathbf{r}). \tag{5}$$

The function $F_{\rm in}$ is the projection along the electric field of a vector field $\mathbf{w}_{\rm in}$:

$$F_{\rm in}(\Omega;\mathbf{r}) = -(ie/2m_e\omega)\mathscr{E}_0\cdot\mathbf{w}_{\rm in}(\Omega;\mathbf{r}). \tag{6}$$

The vector $\mathbf{w}_{\rm in}$ is defined as

$$\mathbf{w}_{\rm in}(\Omega;\mathbf{r}) \equiv \int G(\mathbf{r},\mathbf{r}';\Omega)\mathbf{P}' u_{\rm in}(\mathbf{r}')d\mathbf{r}', \tag{7}$$

with G the Coulomb Green's function.

If the initial state (1) is characterized by the quantum numbers nlm, the vector $\mathbf{w}_{nlm}$ in (7) can be expressed in terms of two vector spherical harmonics (see Ref. 4):

 0022-2488/90/092165-05$03.00

$$\mathbf{w}_{nlm}(\Omega;\mathbf{r}) = \frac{im_e}{\hbar}\left[-\left(\frac{l+1}{2l+1}\right)^{1/2}\mathscr{B}_{n,l,l+1}(\Omega;r)\mathbf{V}_{l+1,l,m}(\theta,\varphi) + \left(\frac{l}{2l+1}\right)^{1/2}\mathscr{B}_{n,l,l-1}(\Omega;r)\mathbf{V}_{l-1,l,m}(\theta,\varphi)\right]. \quad (8)$$

The second term exists only for $l>0$. The expressions of the radial functions $\mathscr{B}_{n,l,l\pm1}$ are given in Eqs. (17)–(20) of Ref. 4 as integral representations. They were also expressed in terms of Humbert functions.

Here we present a derivation for the function $F_{\rm in}$ in (6) for the case of an initial Stark state characterized by the quantum numbers n_1, n_2, and m with respect to a quantization axis taken along the direction of the applied electric field (2). First, using Eqs. (6) and (8) here, together with Eqs. (17) and (20) of Ref. 4, we transcribe F_{nlm} as follows:

$$F_{nlm}(\Omega;\mathbf{r}) = \frac{e\mathscr{E}_0}{2\hbar\omega}(2\kappa_n)^{1/2}\frac{i\exp(i\pi\tau)}{2\sin\pi\tau} \times\int_1^{(0+)} t^{-\tau}\left[N_-^2\left(\frac{n+1}{n}\right)^2 g_{nlm}^{(-)}(t,\Omega;\mathbf{r}) - N_+^2\left(\frac{n-1}{n}\right)^2 g_{nlm}^{(+)}(t,\Omega;\mathbf{r})\right] \times\exp\left(-\frac{n+\tau-(n-\tau)t}{N_+}\frac{Xr}{\hbar}+\frac{y_n}{2}\right)dt, \quad (9)$$

with the notations

$$X\equiv(-2m_e\Omega)^{1/2},\quad \mathrm{Re}\,X>0,$$
$$\tau\equiv\alpha Z m_e c/X. \quad (10)$$

Also

$$N_\pm\equiv n\pm\tau+(n\mp\tau)t,$$
$$y_n\equiv(8n^2\kappa_n t/N_+N_-)r, \quad (11)$$

α is the fine structure constant, c the velocity of light, Z the nuclear charge, and the constant κ_n is defined in (A2). The functions $g_{nlm}^{(\mp)}$ appearing in the integral (9) are expressed in the following in terms of hydrogenic functions in which the usual radial variable $2\kappa_n r$ is replaced by y_n, defined in Eq. (11), regardless of the value of the principal quantum number $(n\pm1)$ of these functions. In order to distinguish from the usual functions u_{nlm}, we use the tilde sign. The explicit expression of the functions $g_{nlm}^{(\mp)}$ is

$$g_{nlm}^{(\mp)}(t,\Omega;\mathbf{r}) = \alpha_{nlm}^{(\mp)}\tilde{u}_{n\pm1,l+1,m} + \beta_{nlm}^{(\mp)}\tilde{u}_{n\pm1,l-1,m}, \quad (12)$$

where

$$\alpha_{nlm}^{(-)}\equiv\left(\frac{(l+1)^2-m^2}{2l+3}(n+l+1)(n+l+2)\right)^{1/2},$$
$$\beta_{nlm}^{(-)}\equiv\left(\frac{l^2-m^2}{2l+3}(n-l)(n-l+1)\right)^{1/2}, \quad (13)$$

while $\alpha_{nlm}^{(+)}$ and $\beta_{nlm}^{(+)}$ are given by the same expressions with n replaced by $-n$.

The essential step for our calculation is the use of hydrogenic functions in Eq. (12). The starting point in Ref. 3 for the problem we discuss here is Eq. (B3), which has a certain elegance: it expresses the vector $\mathbf{w}_{nlm}$ through the action of an operator $\mathscr{P}_{n\tau}$ [given in Eq. (B4) of Ref. 3] applied to a single modified hydrogenic energy eigenfunctions inside integral representation. But the operator $\mathscr{P}_{n\tau}$ itself is not too simple, and in the final stage of the calculation it has to be applied effectively, which implies long calculations. In our procedure, we work with explicit expressions at each stage, for each component of the vector $\mathbf{w}_{n_1n_2m}$. We present here the case of the component of $\mathbf{w}_{n_1n_2m}$ along the symmetry axis of the initial states, but the other components can be treated in a similar way; our results are in accord with Ref. 3.

III. THE FIRST-ORDER PERTURBED STARK STATES

As shown by Schrödinger,[1] for the electron in an attractive Coulomb field, one can construct stationary states by separating the variables in the parabolic coordinates ξ, η, and φ:

$$\xi\equiv r+z,\quad \eta\equiv r-z. \quad (14)$$

Because the final results will contain, under a integral sign, these energy eigenfunctions, denoted here by $\varphi_{n_1n_2m}(\mathbf{r})$, they are reproduced in Appendix A. A function $\varphi_{n_1n_2m}$ corresponds to the Bohr energy E_n with

$$n=n_1+n_2+|m|+1. \quad (15)$$

The relation between the spherical energy eigenfunctions u_{nlm} and the parabolic energy eigenfunctions $\varphi_{n_1n_2m}$ is

$$\varphi_{n_1n_2m}(\mathbf{r}) = \sum_{l=|m|}^{n-1} A_{nlm}^{n_1n_2}u_{nlm}(\mathbf{r}). \quad (16)$$

We use the notation of Tarter[8] for the coefficients in (16). For a given n, they form a quadratic orthogonal matrix. Some details regarding the $A_{nlm}^{n_1n_2}$ can be found in Appendix A, together with their explicit expression in terms of a generalized hypergeometric function ${}_3F_2$ of variable 1. This expression was directly used in our calculation.

In contrast to the case of the (nlm) states, the charge distribution of the Stark states is not symmetric with respect to the plane $x0y$. It keeps only the symmetry around the quantization axis z. As the electric quantum number defined as

$$n_e\equiv n_1-n_2 \quad (17)$$

increases, the charge distribution becomes more eccentric. For $n_e<0$ it is oriented along the z axis, for $n_e>0$ in the opposite direction.

In the case of an initial state (1) characterized by the energy eigenfunctions $\varphi_{n_1n_2m}$ given by (A1), with the symmetry axis along the applied ac field (2), the corresponding function F in (5) will be denoted by $F_{n_1n_2m}$. From Eqs. (6), (7), and (16), one gets the connection

$$F_{n_1n_2m}(\Omega;\mathbf{r}) = \sum_{l=|m|}^{n-1} A_{nlm}^{n_1n_2}F_{nlm}(\Omega;\mathbf{r}). \quad (18)$$

For F_{nlm} we shall use the expression (9).

Equations (18) and (9) show that $F_{n_1n_2m}$ will have the same structure as F_{nlm}, with $g_{nlm}^{(\mp)}(t,\Omega;\mathbf{r})$ in (9) replaced by

$$g^{(\mp)}_{n_1n_2m}(t,\Omega;\mathbf{r}) = \sum_{l=|m|}^{n-1} A^{n_1n_2}_{nlm} g^{(\mp)}_{nlm}(t,\Omega;\mathbf{r}). \tag{19}$$

Replacing (12) in (19) and ordering the terms after the hydrogenic functions, we obtain

$$g^{(\mp)}_{n_1n_2m}(t,\Omega;\mathbf{r}) = \sum_{l=|m|}^{n} (\alpha^{(\mp)}_{n(l-1)m} A^{n_1n_2}_{n(l-1)m} + \beta^{(\mp)}_{n(l+1)m} A^{n_1n_2}_{n(l+1)m}) \tilde{u}_{(n\pm1)lm}. \tag{20}$$

Because of the structure of our starting point for F_{nlm} [Eqs. (9) and (12)], we cannot directly use Eq. (16) in order to perform the summation of the corrections to the spherical bound states. In contrast to Ref. 3, we are forced to prove some identities for the coefficients $A^{n_1n_2}_{nlm}$, and only after this can we exploit the basic equation (16). These identities, which to our knowledge are new, can be written compactly as

$$\alpha^{(\mp)}_{n(l-1)m} A^{n_1n_2}_{n(l-1)m} + \beta^{(\mp)}_{n(l+1)m} A^{n_1n_2}_{n(l+1)m} = \gamma^{(\mp)}_{n_1n_2m} A^{(n_1\pm1)n_2}_{(n\pm1)lm} + \delta^{(\mp)}_{n_1n_2m} A^{n_1(n_2\pm1)}_{(n\pm1)lm}, \tag{21}$$

with

$$\gamma^{(-)}_{n_1n_2m} = -[(n_1+1)(n_1+|m|+1)]^{1/2},$$

$$\delta^{(-)}_{n_1n_2m} = [(n_2+1)(n_2+|m|+1)]^{1/2}, \tag{22}$$

and

$$\gamma^{(+)}_{n_1n_2m} = \gamma^{(-)}_{(n_1-1)n_2m}, \quad \delta^{(+)}_{n_1n_2m} = \delta^{(-)}_{n_1(n_2-1)m}. \tag{23}$$

The derivation of the identities (21) is sketched in Appendix B.

Now, as mentioned before, with Eq. (21) replaced in (20), the summation can be performed directly using (16), and this way the calculation is completed. The result for the functions (20) is

$$g^{(\mp)}_{n_1n_2m}(t,\Omega;\mathbf{r}) = \gamma^{(\mp)}_{n_1n_2m} \tilde{\varphi}_{(n_1\pm1)n_2m} + \delta^{(\mp)}_{n_1n_2m} \tilde{\varphi}_{n_1(n_2\pm1)m}, \tag{24}$$

where the tilde sign is used in order to note that the usual variables $\kappa_n\xi$ and $\kappa_n\eta$ of the functions $\varphi_{n_1n_2m}$ in (A1) are replaced, according to (7) and (11), by

$$\tilde{\xi} \equiv \frac{4n^2t}{N_+N_-}\kappa_n\xi, \quad \tilde{\eta} \equiv \frac{4n^2t}{N_+N_-}\kappa_n\eta, \tag{25}$$

regardless of the value of the principal quantum number $(n+1$ or $n-1)$ with which the functions are associated.

Consequently, our final result for the function $F_{n_1n_2m}$ in Eq. (6) is the integral representation

$$F_{n_1n_2m}(\Omega;\mathbf{r}) = \frac{e\mathscr{E}_0}{2\hbar\omega}(2\kappa_n)^{1/2}\frac{i\exp(i\pi\tau)}{2\sin\pi\tau}\int_1^{(0+)} t^{-\tau}\left[N_-^2\left(\frac{n+1}{n}\right)^2(\gamma^{(-)}_{n_1n_2m}\tilde{\varphi}_{(n_1+1)n_2m} + \delta^{(-)}_{n_1n_2m}\tilde{\varphi}_{n_1(n_2+1)m}) - N_+^2\left(\frac{n-1}{n}\right)^2(\gamma^{(+)}_{n_1n_2m}\tilde{\varphi}_{(n_1-1)n_2m} + \delta^{(+)}_{n_1n_2m}\tilde{\varphi}_{n_1(n_2-1)m})\right]\exp\left(-\frac{n+\tau-(n-\tau)t}{N_+}\frac{Xr}{\hbar} + \frac{y_n}{2}\right)dt. \tag{26}$$

The result agrees with Eq. (37) of Ref. 3 [the case $\mu = 0$, corresponding to the component of the vector $\mathbf{w}_{n_1n_2m}$ in Eq. (7) along the quantization axis]. For further calculations, we prefer to give the final result in the form (26). The main feature of our result is the possibility of expressing the function under the integral sign in (26) in terms of parabolic hydrogenic functions depending on the variables (25). The integration variable t appears in a way that shows that the functions $F_{n_1n_2m}$ themselves do not have a simple structure in parabolic coordinates, in contrast to the energy eigenfunctions in a static electric field.[9] Some properties of the functions $F_{n_1n_2m}$ are described in Sec. IV.

IV. PROPERTIES OF THE FUNCTIONS $F_{n_1n_2m}$

Not all the properties of the functions $F_{n_1n_2m}$ and, consequently, of the linear modification (4) of the wavefunction, are directly accessible from Eq. (26). As in the case of the functions F_{nlm} corresponding to the spherical states, the functions $F_{n_1n_2m}$ can be expressed in terms of a finite number of Humbert functions ϕ_1. Such an expansion is described in Ref. 3. The expression of the functions $F_{n_1n_2m}$ in terms of Humbert functions is useful in a numerical evaluation of the linear response or of some functions of it having physical significance, like induced charge density or electric field in the interior of the atom. These quantities deserve further investigation. The expression also makes possible an analytic investigation of the behavior of the functions $F_{n_1n_2m}$ for low frequencies ω. For $\omega \to 0$, the quantity τ in Eq. (10) approaches the value of the principal quantum number n. The two variables of the Humbert functions go to zero in this limit. In order to get the correct answer, some attention has to be paid to the behavior of the parameters of these functions, too. The result is

$$F_{n_1n_2m}(E_n \mp \hbar\omega;\mathbf{r}) = \frac{e}{2\hbar\omega}\left(r\cos\theta - \frac{3}{2}nn_1\frac{a_0}{Z}\right)\varphi_{n_1n_2m}(\mathbf{r}) \pm \tfrac{1}{2}\chi_{n_1n_2m}(\mathbf{r}) + \mathscr{O}(\omega), \tag{27}$$

where $\chi_{n_1n_2m}(\mathbf{r})$ is the first-order modification of the Stark states in a static field, as given by Eq. (69) of Omidvar,[9] and a_0 is the Bohr radius.

Equation (27) can be predicted without using the explicit expression of the functions $F_{n_1n_2m}$, starting from the general identity[10]

$$(\hbar/im_e)\mathbf{w}_{\text{in}}(\Omega;\mathbf{r}) = \mathbf{r}u_{\text{in}}(\mathbf{r}) + (\Omega - E_n)\mathbf{v}_{\text{in}}(\Omega;\mathbf{r}),$$

where $\mathbf{v}_{\text{in}}(\Omega;\mathbf{r})$ is defined by an expression similar to (7), with the momentum operator $\mathbf{P}'$ replaced by the position operator $\mathbf{r}'$. After isolating from $\mathbf{v}_{\text{in}}$ the contribution of the states with energy E_n, the limit $\omega \to 0$ can be taken directly. Alternatively, as a check, the first term in Eq. (27) was extracted from (26) by integration by part. More details con-

cerning the static limit of the linear response (4) are given in Ref. 3, together with the analytic expressions for the vectors $\mathbf{v}_{n_1n_2m}$.

Some application of the functions $\mathbf{w}_{n_1n_2m}$ arises in the study of two-photon processes. General analytic results for bound–bound transitions have been presented recently.[11] We have studied independently[12] the particular transitions in which one of the states is an extreme Stark state ($n_1 = n - 1$). For the ground state case, we have directly used the connection with our results for $1s \rightarrow ns,nd$ transitions,[13] in order to predict numerically the behavior of the transition amplitudes from the ground state to Stark states. The relevant equations, together with their numerical consequences, will be published subsequently.

ACKNOWLEDGMENT

One of the authors (V. F.) gratefully acknowledges useful discussions with Professor R. H. Pratt from the University of Pittsburgh.

APPENDIX A: THE STARK STATES AND THE COEFFICIENTS $A^{n_1n_2}_{nlm}$

The explicit expression of the Stark (or parabolic) energy eigenfunctions for an electron in the field of a fixed nucleus of charge Z is[14]

$$\begin{aligned}\varphi_{n_1n_2m}(\mathbf{r}) = N_{n_1n_2m}(\kappa_n^2\xi\eta)^{|m|/2}\exp(im\varphi)\\ \times {}_1F_1(-n_1,|m|+1;\kappa_n\xi)\\ \times {}_1F_1(-n_2,|m|+1;\kappa_n\eta)\\ \times\exp[-\tfrac{1}{2}\kappa_n(\xi+\eta)],\end{aligned} \tag{A1}$$

with

$$\begin{aligned}N_{n_1n_2m} &= \frac{2^{1/2}}{n^2}\left(\frac{Z}{a_0}\right)^{3/2}\frac{1}{(|m|!)^2}\\ &\times\left(\frac{(n_1+|m|)!(n_2+|m|)!}{n_1!\,n_2!}\right)^{1/2},\end{aligned}$$

and

$$\kappa_n \equiv (Z/n)a_0. \tag{A2}$$

The coefficients $A^{n_1n_2}_{nlm}$ in Eq. (16) connecting the parabolic and the spherical energy eigenstates have been expressed by Park[15] as particular Clebsch–Gordan coefficients. For a fixed n the matrix A diagonalizes the electron energy $H' = e\mathscr{E}z$ in a dc electric field directed along the z axis. This leads to the basic property[9]

$$\begin{aligned}c_{n(l+1)m}A^{n_1n_2}_{n(l+1)m} + c_{nlm}A^{n_1n_2}_{n(l-1)m}\\ = -n_eA^{n_1n_2}_{nlm},\end{aligned} \tag{A3}$$

with

$$c_{nlm} \equiv [(n^2-l^2)(l^2-m^2)/(4l^2-1)]^{1/2},$$

and n_e defined in (17).

According to Eq. (22) of Ref. 8, the coefficient $A^{n_1n_2}_{nlm}$ has the analytic expression

$$\begin{aligned}A^{n_1n_2}_{nlm} = (-1)^{l-|m|}\frac{(n-|m|-1)!}{|m|!}\\ \times\left(\frac{(2l+1)(l+|m|)!(n_1+|m|)!(n_2+|m|)!}{(n+l)!}\right)^{1/2}\\ \times {}_3F_2\left(\begin{matrix}l+|m|+1,-l+|m|,-n_2;\\ |m|+1,-n+|m|+1;\end{matrix}1\right),\end{aligned} \tag{A4}$$

where ${}_3F_2$ is the hypergeometric generalized function

$${}_3F_2\left(\begin{matrix}\alpha,\beta_1,\beta_2;\\ \gamma_1,\gamma_2;\end{matrix}x\right) \equiv \sum_{k=0}^{\infty}\frac{\alpha_k(\beta_1)_k(\beta_2)_k}{(\gamma_1)_k(\gamma_2)_k 1_k}x^k,$$

with α_k Pochhammer's symbol.

The coefficients $A^{n_1n_2}_{nlm}$ depend only on $|m|$. The interchange of n_1 and n_2 leads to

$$A^{n_2n_1}_{nlm} = (-1)^{l-m}A^{n_1n_2}_{nlm}. \tag{A5}$$

The values of the coefficients for low values of n can be found in several papers.[16,9,8]

APPENDIX B: RECURRENCE RELATIONS FOR THE FUNCTION ${}_3F_2$

The key relations (21) are based on four recurrence relations for the functions ${}_3F_2$ of variable 1 that express the coefficients $A^{n_1n_2}_{nlm}$ according to (A4). These are Eqs. (B3)–(B8) whose derivation is now sketched.

We denote

$$\begin{aligned}h_0 &\equiv {}_3F_2\left(\begin{matrix}\alpha_1,\alpha_2,\alpha_3;\\ \beta_1,\beta_2;\end{matrix}1\right),\\ h_1 &\equiv {}_3F_2\left(\begin{matrix}\alpha_1\mp1,\alpha_2\pm1,\alpha_3;\\ \beta_1,\beta_2;\end{matrix}1\right),\\ h_2 &\equiv {}_3F_2\left(\begin{matrix}\alpha_1,\alpha_2,\alpha_3\mp1;\\ \beta_1,\beta_2\mp1;\end{matrix}1\right),\\ h_3 &\equiv {}_3F_2\left(\begin{matrix}\alpha_1,\alpha_2,\alpha_3;\\ \beta_1,\beta_2\mp1;\end{matrix}1\right),\\ h_4 &\equiv {}_3F_2\left(\begin{matrix}\alpha_1,\alpha_2,\alpha_3;\\ \beta_1,\beta_2;\end{matrix}1\right),\\ h_5 &\equiv {}_3F_2\left(\begin{matrix}\alpha_1,\alpha_2,\alpha_3;\\ \beta_1,\beta_2\mp1;\end{matrix}1\right).\end{aligned} \tag{B1}$$

Using Eqs. (14), (15), (19), and (21) in Chapter 5 of Rainville's book,[17] we obtain the recurrence relation

$$\begin{aligned}\alpha_2(\alpha_1-\beta_1)(\alpha_1-\beta_2)h_1 = (\alpha_2-\beta_2+1)[(\alpha_1-1)(\alpha_1-\beta_1)+(\alpha_1-\alpha_2-1)(\alpha_3-\beta_2+1)h_0]\\ -(\alpha_1-\alpha_2-1)(\beta_2-1)(\alpha_1+\alpha_2+\alpha_3-\beta_1-\beta_2+1)h_5.\end{aligned} \tag{B2}$$

Interchanging α_1 and α_2 in (B2) we obtain another recurrence relation connecting h_2, h_0, and h_5. We may also prove that

$$(\beta_2-1)(\alpha_3-\beta_1)h_3 = -(\alpha_1+\alpha_2+\alpha_3-\beta_1-\beta_2+1)(\beta_2-1)h_5 + (\alpha_1-\beta_2+1)(\alpha_2-\beta_2+1)h_0. \tag{B3}$$

The case we are interested in corresponds to

$$\alpha_1 = l + |m| + 1, \quad \alpha_2 = -l + |m|, \quad \alpha_3 = -n_2,$$
$$\beta_1 = |m| + 1, \quad \beta_2 = -n_1 - n_2. \tag{B4}$$

We notice, as a check, that the elimination of the function h_5 between (B2) and its analogous with $\alpha_1 \leftrightarrow \alpha_2$ gives a relation between h_1, h_2, and h_0, which leads directly to (A3). The elimination of the function h_0 gives

$$(n+l)(n+l+1)(l+|m|)h_1 = (n-|m|)[(2n+l-|m|-2n_2)(n+|m|+1)h_3 - (n-n_2)(l+|m|+2n_2+2)h_5]. \tag{B5}$$

Using (B3) we get

$$(n-l)(n-l-1)(l+|m|+1)h_2 = (n-|m|)[(2n-l-|m|-2n_2-1) \times (n_2+|m|+1)h_3] - (n-n_2)(l-|m|-2n_2-1)h_5]. \tag{B6}$$

The parameters in the functions h in (B5) and (B6) are given by (B4). The first equation (21) follows directly from (A4), (B5), and (B6). Similar techniques give

$$(|m|+1-n)(l-|m|)h_1 = (l+|m|)n_2 h_4 + n_1(2n-l+|m|-2)h_6, \tag{B7}$$

$$(|m|+1-n)(l+|m|+1)h_2 = (l+|m|+1)n_2 h_4 - n_1(2n+l+|m|-1)h_6. \tag{B8}$$

Here, again, the parameters in the ${}_3F_2$ functions are given by (B4). The second equation (21) follows directly from (A4), (B7), and (B8).

[1] E. Schrödinger, Ann. Physik (Leipzig) **80**, 437 (1926).
[2] J. E. Bayfield and L. A. Pinnaduwage, Phys. Rev. Lett. **54**, 313 (1985).
[3] T. Marian, Phys. Rev. A **39**, 3803 (1989).
[4] V. Florescu and T. Marian, Phys. Rev. A **34**, 4641 (1986).
[5] B. Podolsky, Proc. Natl. Acad. Sci. USA **14**, 253 (1928).
[6] M. Luban, B. Nudler, and I. Freund, Phys. Lett. A **47**, 447 (1974); M. Luban and B. Nudler-Blum, J. Math. Phys. **18**, 1871 (1977).
[7] H. Johansson, *An Investigation into the Scattering of Radiation by Hydrogenlike Atoms*, disertation (Almqvist and Wiksells, Uppsala, 1942).
[8] C. B. Tarter, J. Math. Phys. **11**, 3192 (1970).
[9] K. Omidvar, Phys. Rev. **153**, 121 (1967).
[10] V. Florescu and T. Marian, Central Institute of Physics, Bucharest, Report No. FT-245 (unpublished).
[11] T. Marian, Phys. Rev. A **39**, 3816 (1989).
[12] V. Florescu and S. Patrascu (unpublished).
[13] V. Florescu, S. Patrascu, and O. Stoican, Phys. Rev. A **36**, 2155 (1987).
[14] H. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms* (Springer, Berlin, 1957).
[15] D. Park, Z. Phys. **159**, 153 (1960).
[16] V. Rojansky, Phys. Rev. **33**, 1 (1929).
[17] E. D. Rainville, *Special Functions* (MacMillan, New York, 1960).

Properties of Leach–Flessas–Gorringe polynomials

D. L. Pursey
Department of Physics, Iowa State University, Ames, Iowa 50011

(Received 11 January 1990; accepted for publication 2 May 1990)

A generating function is obtained for the polynomials recently introduced by Leach, Flessas, and Gorringe [J. Math. Phys. **30**, 406 (1989)], and is then used to relate the Leach–Flessas–Gorringe (or LFG) polynomials to Hermite polynomials. The generating function is also used to express a number of integrals involving the LFG polynomials as finite sums of parabolic cylinder functions.

I. INTRODUCTION

Several papers dealing with the sextic anharmonic oscillator have been published in recent years.[1,2] Particularly interesting is the technique developed by Leach, Flessas, and Gorringe.[2] These authors considered the one-dimensional anharmonic oscillator with Hamiltonian

$$H = -\frac{1}{2}\frac{d^2}{dx^2} + \frac{1}{2}[(ax^2+b)^2x^2 - kax^2], \tag{1}$$

and sought solutions of the form

$$\psi(x) = \sum_{n=0}^{\infty} c_n f_n(x) \exp\left(-\frac{1}{4}ax^4 - \frac{1}{2}bx^2\right), \tag{2}$$

for even parity states, and

$$\psi(x) = \sum_{n=0}^{\infty} c_n g_n(x) \exp\left(-\frac{1}{4}ax^4 - \frac{1}{2}bx^2\right), \tag{3}$$

for odd parity states, where $f_n(x)$ and $g_n(x)$ are polynomials defined by

$$f_n(x) = \frac{1}{2^n n!} e^{(1/2)ax^4 + bx^2}\left(-\frac{1}{x}\frac{d}{dx}\right)^n \times e^{-(1/2)ax^4 - bx^2}, \tag{4}$$

and

$$g_n(x) = x f_n(x). \tag{5}$$

With this ansatz, the eigenvalue problem reduces to the determination of the eigenvalues of an infinite tridiagonal matrix. Furthermore, if $k = 4N-1$ $(4N+1)$, the lowest N even (odd) parity eigenvalues are the eigenvalues of a finite $N \times N$ tridiagonal matrix.

The purpose of this paper is to explore some of the properties of the Leach–Flessas–Gorringe (or LFG) polynomials $f_n(x)$ and $g_n(x)$. In the next section, I shall find a generating function for the even LFG polynomials $f_n(x)$ and relate them to the Hermite polynomials $H_n[(a/2)^{1/2}x^2 + b(2a)^{-1/2}]$.

Normalization of the Leach–Flessas–Gorringe states involves sums over the integrals

$$I_{m,n} \equiv \int_{-\infty}^{\infty} dx\, f_m(x) f_n(x) e^{-(1/2)ax^4 - bx^2} \tag{6}$$

and

$$J_{m,n} \equiv \int_{-\infty}^{\infty} dx\, g_m(x) g_n(x) e^{-(1/2)ax^4 - bx^2}, \tag{7}$$

while transition matrix elements involve the integrals

$$I_{m,n,p} \equiv \int_{-\infty}^{\infty} dx\, x^{2p} f_m(x) f_n(x) e^{-(1/2)ax^4 - bx^2} \tag{8}$$

and

$$J_{m,n,p} \equiv \int_{-\infty}^{\infty} dx\, x^{2p} g_m(x) g_n(x) e^{-(1/2)ax^4 - bx^2}. \tag{9}$$

Clearly,

$$J_{m,n} \equiv I_{m,n,1}, \quad J_{m,n,p} \equiv I_{m,n,p+1}. \tag{10}$$

In Sec. III, I shall derive an explicit expression for $I_{m,n}$ as a finite sum of parabolic cylinder functions $D_{p-1/2}(b/a^{1/2})$. The integrals $I_{m,n,p}$, $J_{m,n}$, and $J_{m,n,p}$, are then readily expressed in terms of $I_{m,n}$ by first expanding $x^{2p}f_n(x)$ as a linear combination of even LFG polynomials, and then using Eq. (6).

In the course of the mathematical development, I use standard results found in Abramowitz and Stegun,[3] in Gradshteyn and Ryzhik,[4] and in Spanier and Oldham.[5] I refer to these by the initials AS, GR, and SO, respectively, followed by the formula number in the quoted reference.

II. GENERATING FUNCTION AND RELATION TO HERMITE POLYNOMIALS

In order to find a generating function for the even LFG polynomials, I first note that if $u = (a/2)^{1/2}x^2$ then

$$(2x)^{-1}\frac{d}{dx} = \left(\frac{2}{a}\right)^{1/2}\frac{d}{du}.$$

For convenience, I also define $\sigma = b/(2a)^{1/2}$. From Eq. (4),

$$\sum_{n=0}^{\infty} f_n(x)\left(\frac{2}{a}\right)^{n/2} t^n = e^{u^2+2\sigma u}\left(-t\frac{d}{du}\right)^n e^{-u^2-2\sigma u} = e^{u^2+2\sigma u} e^{-(u-t)^2-2\sigma(u-t)}. \tag{11}$$

Hence, a generating function for the even LFG polynomials is

$$S \equiv \sum_{n=0}^{\infty} f_n(x)\left(\frac{2}{a}\right)^{n/2} t^n = e^{-t^2+2(u+\sigma)t} = e^{-t^2+2[(a/2)^{1/2}x^2 + b/(2a)^{1/2}]t}. \tag{12}$$

From Ref. 3 (AS22.9.17), S is also

$$S = \sum_{n=0}^{\infty}\frac{t^n}{n!}H_n(u+\sigma) = \sum_{n=0}^{\infty}\frac{t^n}{n!}H_n\left[\left(\frac{a}{2}\right)^{1/2}x^2 + \frac{b}{(2a)^{1/2}}\right]. \tag{13}$$

 0022-2488/90/092170-02$03.00

Hence

$$f_n(x) = \left(\frac{a}{2}\right)^{n/2} \frac{1}{n!} H_n\left[\left(\frac{a}{2}\right)^{1/2} x^2 + b(2a)^{-1/2}\right]. \quad (14)$$

III. OVERLAP INTEGRALS

In this section, I obtain an explicit expression for the overlap integrals

$$I_{m,n} \equiv \int_{-\infty}^{\infty} dx\, f_m(x) f_n(x) e^{-(1/2)ax^4 - bx^2}. \quad (6)$$

This is most conveniently achieved by developing a generating function I for the $I_{m,n}$ using the generating function S of Eq. (12). I then use the recurrence relations for the LFG polynomials to obtain recursive formulas for $I_{m,n,p}$, $J_{m,n}$, and $J_{m,n,p}$, which allow these integrals to be expressed in terms of the $I_{m,n}$.

For convenience, I define the scaled variable $y = (a/2)^{1/4}x$. Then

$$I \equiv \sum_{m,n=0}^{\infty} \left(\frac{2}{a}\right)^{(1/2)(m+n)} s^m t^n I_{m,n} \quad (15)$$

$$= e^{-(s^2+t^2)+2\sigma(s+t)} \left(\frac{2}{a}\right)^{1/4} \times 2\int_0^{\infty} dy\, e^{-y^4 - 2(\sigma - s - t)y^2}. \quad (16)$$

The integral in Eq. (16) is evaluated using, Ref. 4, GR3.469.1, to yield

$$I = (1/2a)^{1/4} e^{2st + \sigma^2 - 1/2(\sigma - s - t)^2} (\sigma - s - t)^{1/2} \times K_{1/4}[\tfrac{1}{2}(\sigma - s - t)^2]. \quad (17)$$

The Bessel function is expressed as a parabolic cylinder function using, Ref. 5, SO46:4:5, with the result that

$$I = (\pi^2/a)^{1/4} e^{2st + \sigma^2 - (1/2)(\sigma - s - t)^2} \times D_{-1/2}[2^{1/2}(\sigma - s - t)]. \quad (18)$$

From SO46:5:2 it follows that

$$I = (\pi^2/a)^{1/4} e^{2st + (1/2)\sigma^2} \times \sum_{p=0}^{\infty} \frac{2^{p/2}(s+t)^p}{p!} D_{p-1/2}(2^{1/2}\sigma). \quad (19)$$

This may be compared with Eq. (15) to obtain

$$I_{m,n} = (\pi^2/a)^{1/4} a^{(m+n)/2} e^{\sigma^2/2} \times \sum_{p=0}^{\infty} \frac{1}{p!(m-p)!(n-p)!} D_{m+n-2p-1/2}(2^{1/2}\sigma), \quad (20)$$

where the summation automatically terminates at the lesser of m and n. In terms of the original parameters a and b, the overlap integrals are expressed by

$$I_{m,n} = (\pi^2/a)^{1/4} a^{(m+n)/2} e^{b^2/4a} \times \sum_{p=0}^{\infty} \frac{1}{p!(m-p)!(n-p)!} D_{m+n-2p-1/2}\left(\frac{b}{a^{1/2}}\right). \quad (21)$$

In order to find recursion relations for the other integrals of interest I expand $x^2 f_n(x)$ in terms of LFG polynomials, using the recurrence relation Eq. (3.3) of Ref. 2. This may be rewritten as

$$x^2 f_n(x) = f_{n-1}(x) - \frac{b}{a} f_n(x) + \frac{n+1}{a} f_{n+1}(x). \quad (22)$$

[The same result may be obtained using Eq. (14) together with AS22.7.13.] Hence

$$I_{m,n,p} = I_{m,n-1,p-1} - \frac{b}{a} I_{m,n,p-1} + \frac{n+1}{a} I_{m,n+1,p-1}, \quad (23a)$$

$$= I_{m-1,n,p-1} - \frac{b}{a} I_{m,n,p-1} + \frac{m+1}{a} I_{m+1,n,p-1}, \quad (23b)$$

where $I_{m,n,p} = 0$ if either $m < 0$ or $n < 0$. By repeated iterations of Eq. (23a) or Eq. (23b), any of the integrals $I_{m,n,p}$, $J_{m,n}$, or $J_{m,n,p}$ may be expressed as a linear combination of integrals of the form $I_{m,n} \equiv I_{m,n,0}$, evaluated in Eq. (21) above.

ACKNOWLEDGMENT

I am indebted to Dr. B. C. Carlson for helpful comments and for drawing my attention to Ref. 5.

[1] V. Singh, S. N. Biswas, and K. Dutta, Phys. Rev. D **18**, 1901 (1978); P. G. L. Leach, Physica D **17**, 331 (1985); M. H. Blecher and P. G. L. Leach, J. Phys. A **20**, 5923 (1987); A. K. Dutta and R. S. Willey, J. Math. Phys. **29**, 892 (1988).

[2] P. G. L. Leach, G. P. Flessas, and V. M. Gorringe, J. Math. Phys. **30**, 406 (1989).

[3] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972), 9th ed.

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

[5] J. Spanier and K. B. Oldham, *An Atlas of Functions* (Hemisphere, New York, 1987).

Inverse scattering problem for the 3-D Schrödinger equation and Wiener–Hopf factorization of the scattering operator

Tuncay Aktosun
Department of Mathematics, Southern Methodist University, Dallas, Texas 75275
Cornelis van der Mee
Department of Physics and Astronomy, Free University, Amsterdam, The Netherlands

(Received 13 November 1989; accepted for publication 14 March 1990)

Sufficient conditions are given for the existence of a Wiener–Hopf factorization of the scattering operator for the 3-D Schrödinger equation with a potential having no spherical symmetry. A consequence of this factorization is the solution of a related Riemann–Hilbert problem, thus providing a solution of the 3-D inverse scattering problem.

I. INTRODUCTION

Consider the Schrödinger equation in three dimensions

$$\Delta\psi(k,x,\theta) + k^2\psi(k,x,\theta) = V(x)\psi(k,x,\theta), \qquad (1.1)$$

where Δ is the Laplacian, $x\in\mathbf{R}^3$ is the space coordinate, $\theta\in S^2$ is a unit vector in $\mathbf{R}^3$, and $k^2\in\mathbf{R}$ is energy. The potential $V(x)$ is assumed to decrease to zero sufficiently fast as $|x|\to\infty$. However, we do not assume any spherical symmetry on the potential. As $|x|\to\infty$, the wave function $\psi(k,x,\theta)$ behaves as

$$\psi(k,x,\theta) = e^{ik\theta\cdot x} + \frac{e^{ik|x|}}{|x|}A\left(k,\frac{x}{|x|},\theta\right) + o\left(\frac{1}{|x|}\right), \qquad (1.2)$$

where $A(k,\theta,\theta')$ is the scattering amplitude. The scattering operator $S(k,\theta,\theta')$ is then defined by

$$S(k,\theta,\theta') = \delta(\theta-\theta') - (k/2\pi i)A(k,\theta,\theta'), \qquad (1.3)$$

where δ is the Dirac delta distribution on S^2. In operator notation (1.3) is written as

$$S(k) = \mathbf{I} - (k/2\pi i)A(k),$$

where the operators are defined on $L^2(S^2)$, the Hilbert space of complex-valued, square-integrable functions on the unit sphere S^2 in $\mathbf{R}^3$ with the usual inner product $\langle\cdot,\cdot\rangle$.

The direct scattering problem is to obtain $S(k,\theta,\theta')$ when $V(x)$ is given. The inverse scattering problem, however, is to recover $V(x)$ when $S(k,\theta,\theta')$ is known. Since the main source of information about molecular, atomic, and subatomic particles consists of collision experiments, solving the inverse scattering problem is equivalent to determining the forces between particles from scattering data.

For one-dimensional and radial Schrödinger equations, the inverse scattering problem is fairly well understood (at least for certain classes of potentials).[1] In higher dimensions, however, the situation is quite different. The solution methods developed in higher dimensions include the Newton–Marchenko method,[2–4] the Gel'fand–Levitan method,[2–5] the $\bar\partial$ method,[6–9] the generalized Jost–Kohn method,[10–13] and a method that uses the Green's function of Faddeev.[14–16] There are still many open problems in multidimensional inverse scattering, and the methods developed are still far from being complete. A comprehensive review of the methods and related open problems in multidimensional inverse scattering can be found in Newton's recent book[17] or in Ref. 1.

The principal idea behind both the Newton–Marchenko and Gel'fand–Levitan methods is to formulate the inverse scattering problem as a Riemann–Hilbert boundary value problem, to transform this Riemann–Hilbert problem into a nonhomogeneous integral equation where the kernel and the nonhomogeneous term contain the Fourier transform of the scattering data, and to obtain the potential from the solution of the resulting integral equation. In this paper we present a solution of the 3-D inverse scattering problem by establishing a Wiener–Hopf factorization for the scattering operator and thus solving the corresponding Riemann–Hilbert problem. The usual theory of Wiener–Hopf factorization, however, deals with scalar functions and square matrix functions. Here, we need the Wiener–Hopf factorization of an operator function in an infinite-dimensional setting, and for this we draw on some results by Gohberg and Leiterer.[18]

The present paper is organized as follows. In Sec. II we define the class of potentials (which we will name the Newton class) for which corresponding scattering operators have a Wiener–Hopf factorization. In Sec. III we give some estimates on the scattering amplitude and its derivative and establish the Hölder continuity of the scattering operator. In Sec. IV we define the Wiener–Hopf factorization for operator-valued functions and prove its existence for scattering operators corresponding to potentials in the Newton class. In Sec. V we solve a related Riemann–Hilbert problem using the Wiener–Hopf factorization of the scattering operator. In Sec. VI the solution of the inverse scattering problem is given. Also in this section, for potentials in the Newton class having no bound states, we give the necessary and sufficient conditions for the existence and uniqueness of the Jost operator in terms of the partial indices of the scattering operator. In Sec. VII we summarize the main results of the paper and give the conclusion.

II. ESTIMATES ON THE SCATTERING OPERATOR

We first identify the class of potentials for which all of the results in this paper are valid. Except for the third condition given in the following definition, these conditions are standard assumptions on the potential.[17] The second condi-

 0022-2488/90/092172-09$03.00

tion is much weaker than the usual assumptions.[17] The third condition is needed only twice: first to establish a uniform operator bound for the derivative of the scattering amplitude, and second to use an interpolation argument. Note that all four conditions used below are only sufficient conditions and might possibly be weakened.

Definition 2.1: A potential $V(x)$ is said to belong to the **Newton class** if $V(x)$ is real valued and measurable and satisfies

(i) $\exists a,b>0$ such that

$$\int_{\mathbf{R}^3} dx|V(x)|\left(\frac{|x|+|y|+a}{|x-y|}\right)^2 \leqslant b, \quad \forall y\in\mathbf{R}^3. \tag{2.1}$$

(ii) $\exists c>0$, $s>\frac{1}{2}$ such that $\forall x\in\mathbf{R}^3$

$$|V(x)|\leqslant c/(1+|x|^2)^s. \tag{2.2}$$

(iii) $\exists \gamma>0$ and $\beta\in(0,1]$ such that

$$\int_{\mathbf{R}^3} dx|x|^\beta|V(x)|\leqslant\gamma. \tag{2.3}$$

(iv) The point $k=0$ is not an exceptional point.[19] This condition is satisfied if at zero energy there are neither bound states nor half-bound states.

Remark 2.2: If $V(x)$ satisfies (2.1), we have

$$\int_{\mathbf{R}^3} dx|V(x)|<\infty,$$

$$\int_{\mathbf{R}^3} dx\frac{|V(x)|}{|x-y|}<\infty, \quad \forall y\in\mathbf{R}^3,$$

$$\int_{\mathbf{R}^3} dx\frac{|V(x)|}{|x-y|^2}<\infty, \quad \forall y\in\mathbf{R}^3,$$

$$\int\int_{\mathbf{R}^3\times\mathbf{R}^3} dx\,dy\frac{|V(x)V(y)|}{|x-y|}<\infty,$$

$$\|V\|_R=\left(\int\int_{\mathbf{R}^3\times\mathbf{R}^3} dx\,dy\frac{|V(x)V(y)|}{|x-y|^2}\right)^{1/2}<\infty.$$

The last integral defines the Rollnik norm of the potential. The real potentials with a finite Rollnik norm make up the Rollnik class. The number of bound states n_B for potentials in the Rollnik class is finite[20,21] and $n_B\leqslant\|V\|_R^2/(16\pi^2)$.

Remark 2.3: In (2.2), whenever $s>\frac{3}{4}$, the potential $V\in L^2(\mathbf{R}^3)$. If $s>\frac{1}{2}$, there are no nonzero real exceptional points and hence no positive-energy bound states.[22]

The kernel of the scattering operator $A(k)$ has the representation

$$A(k,\theta,\theta')=-\frac{1}{4\pi}\int_{\mathbf{R}^3} dx\,V(x)e^{-ik\theta\cdot x}\psi(k,x,\theta'), \tag{2.4}$$

where $\psi(k,x,\theta)$ is the solution of the Schrödinger equation. The 3-D Lippmann–Schwinger equation corresponding to the Schrödinger equation satisfying (1.2) is given by

$$\psi(k,x,\theta)=e^{ik\theta\cdot x}-\frac{1}{4\pi}\int_{\mathbf{R}^3} dy\frac{e^{ik|x-y|}}{|x-y|}V(x)\psi(k,y,\theta). \tag{2.5}$$

Iterating (2.5) three times, we obtain

$$\psi_1(k,x,\theta)=e^{ik\theta\cdot x},$$

$$\psi_j(k,x,\theta)=-\frac{1}{4\pi}\int_{\mathbf{R}^3} dy\frac{e^{ik|x-y|}}{|x-y|}V(y)\psi_{j-1}(k,y,\theta), \quad j=2,3,$$

$$\psi_4(k,x,\theta)=\psi(k,x,\theta)-\sum_{j=1}^{3}\psi_j(k,x,\theta).$$

Then we can write (2.4) as[17]

$$A(k,\theta,\theta')=-\frac{1}{4\pi}\sum_{j=1}^{4}A_j(k,\theta,\theta'), \tag{2.6}$$

where

$$A_j(k,\theta,\theta')=\int_{\mathbf{R}^3} dx\,V(x)e^{-ik\theta\cdot x}\psi_j(k,x,\theta'), \quad j=1,2,3,4. \tag{2.7}$$

Proposition 2.4: If the potential $V(x)$ satisfies the first and fourth conditions in the Newton class, the corresponding scattering amplitude $A(k)$ is a continuous operator function in $k\in\mathbf{R}$ on $L^2(S^2)$.

Proof: From (2.7) we obtain the estimates

$$|A_1(k,\theta,\theta')|\leqslant\int_{\mathbf{R}^3} dx|V(x)|,$$

$$|A_2(k,\theta,\theta')|\leqslant\frac{1}{4\pi}\int\int_{\mathbf{R}^3\times\mathbf{R}^3} dx\,dy\frac{|V(x)V(y)|}{|x-y|},$$

$$|A_3(k,\theta,\theta')|\leqslant\frac{1}{(4\pi)^2}\int_{\mathbf{R}^3} dx|V(x)|\times\left\{\int_{\mathbf{R}^3} dy\frac{|V(y)|}{|x-y|}\left[\int_{\mathbf{R}^3} dz\frac{|V(z)|}{|x-z|}\right]\right\},$$

and hence, using Remark 2.2 and Lebesgue's dominated convergence theorem, (2.1) is sufficient to conclude that $A_j(k)$ is continuous for $j=1,2,3$ in the operator norm on $L^2(S^2)$. The continuity of $A_4(k)$ follows[17] under the sufficient condition (2.1) and the fourth condition in the Newton class. ■

The next result is due to Weder.[23] A proof convenient to our present problem is provided by Newton.[17]

Proposition 2.5: If the potential $V(x)$ satisfies (2.1) and (2.2) with $s>1$, and the fourth condition in the definition of the Newton class, $\exists C>0$ such that $\|kA(k)\|\leqslant C/(1+|k|)$ for all $k\in\mathbf{R}$, where the norm is the operator norm on $L^2(S^2)$.

The following proposition generalizes Proposition 2.5 under a much weaker condition.

Proposition 2.6: If the potential $V(x)$ satisfies (2.1) and (2.2) with $\frac{1}{2}<s<1$, and the fourth condition in the definition of the Newton class, $\exists E>0$ such that $\|kA(k)\|\leqslant E/(1+|k|)^{2-2s}$ for all $k\in\mathbf{R}$, where the norm is the operator norm on $L^2(S^2)$.

Proof: When $|k|\leqslant 1$, using (1.3) and the unitarity of $S(k)$ we obtain

$$\|kA(k)\|\leqslant 2\pi(\|S(k)\|+1)\leqslant 4\pi\cdot 2^{2-2s}/(1+|k|)^{2-2s}.$$

When $|k|\geqslant 1$, we proceed as follows. According to the lemma due to Vega,[24] $\forall g\in L^2(S^2)$, we have

$$\left[\int_{\mathbf{R}^3} dx|(\sigma^\dagger(-1)g)(x)|^2(1+|x|^2)^{-s}\right]^{1/2}\leqslant c\|g\|, \quad \forall s>\tfrac{1}{2},$$

where c is a constant and

$$(\sigma^{\dagger}(k)g)(x)=\int_{S^2}d\theta\, e^{-ik\theta\cdot x}g(\theta).$$

Replacing x by kx in Vega's lemma and using $|k|\geqslant 1$, we obtain

$$\int_{\mathbf{R}^3}dx|(\sigma^{\dagger}(k)g)(x)|^2(1+|x|^2)^{-s}$$
$$\leqslant c^2\|g\|^2/|k|^{3-2s}\leqslant 2^{3-2s}c^2\|g\|^2/(1+|k|)^{3-2s}. \quad (2.8)$$

Next, we apply the representation for the scattering amplitude[17,23]

$$A(k)=-(1/4\pi)\sigma(k)V^{1/2}[\mathbf{I}-L(k)]^{-1}|V|^{1/2}\sigma^{\dagger}(k), \quad (2.9)$$

where $\sigma(k)$ is the adjoint of $\sigma^{\dagger}(k)$, V is the potential $V(x)$, $V^{1/2}=\operatorname{sgn}(V)|V|^{1/2}$, and $L(k)$ is the operator whose kernel

$$L(k,x,y)=-(1/4\pi)|V(x)|^{1/2}e^{ik|x-y|}V(y)^{1/2}/|x-y|,$$

is closely related to the kernel of the Lippmann–Schwinger equation (2.5). It is known that $[\mathbf{I}-L(k)]^{-1}$ is uniformly bounded in k in operator norm.[17] Hence, if $V(x)$ satisfies (2.2) with $s>\frac{1}{2}$, using (2.8) and (2.9) we obtain $\|kA(k)\|\leqslant E/(1+|k|)^{2-2s}$, for some constant E. ∎

The choice $\beta=1$ in the next three propositions may seem to be a step backward at first; however, using the interpolation in Proposition 2.10, the results of Propositions 2.8 and 2.9 will be strengthened to include $\beta\in(0,1]$. The next proposition gives the uniform boundedness of the derivative of the scattering amplitude.

Proposition 2.7: If the potential $V(x)$ satisfies (2.1) and (2.3) with $\beta=1$, and the fourth condition in the definition of the Newton class, $\exists B>0$ such that $\|dA(k)/dk\|\leqslant B$ for all $k\in\mathbf{R}$, where the norm is the operator norm on $L^2(S^2)$.

Proof: From (2.7) we obtain by direct computation

$$\left|\frac{\partial A_1(k,\theta,\theta')}{\partial k}\right|\leqslant\int_{\mathbf{R}^3}dx|xV(x)|,$$

$$\left|\frac{\partial A_2(k,\theta,\theta')}{\partial k}\right|\leqslant\frac{1}{2\pi}\int_{\mathbf{R}^3}dy|V(y)|\times\left[\int_{\mathbf{R}^3}dx\left(\frac{|x|+|y|}{|x-y|}\right)^2|V(x)|\right],$$

$$\left|\frac{\partial A_3(k,\theta,\theta')}{\partial k}\right|\leqslant 2\int_{\mathbf{R}^3}dy|V(y)|\left\{\int_{\mathbf{R}^3}dz|V(z)|\times\int_{\mathbf{R}^3}dx\frac{|V(x)|}{|x-z|}\right\}+2\int dy|V(y)|\times\left\{\int_{\mathbf{R}^3}dx|V(x)|\left[\int_{\mathbf{R}^3}dz\frac{|zV(z)|}{|z-x|}\right]\right\},$$

and hence, using Lebesgue's dominated-convergence theorem, the first and third conditions in the Newton class are sufficient for the differentiability of $A_j(k)$ with respect to k in the operator norm on $L^2(S^2)$ and the uniform boundedness of its derivative for $k\in\mathbf{R}$, for $j=1,2,3$. The uniform boundedness $\|dA_4(k)/dk\|\leqslant B_4$ has already been established[17] using the first and fourth conditions in the definition of the Newton class. Note that in the above proof, the only place where we used (2.3) was the bound on $\|dA_1(k)/dk\|$. ∎

The Möbius transformation $k\mapsto\xi=(k-i)/(k+i)$ maps the extended real axis $\mathbf{R}_{\infty}$ onto the unit circle $\mathbf{T}$, the upper-half complex plane $\mathbf{C}^+$ onto the unit disk $\mathbf{T}^+$, and the lower-half plane $\mathbf{C}^-$ onto the exterior of the unit disk $\mathbf{T}^-$, where ∞ is considered to be a point of $\mathbf{T}^-$. Let $\tilde{S}(\xi)=S(k)$ under this transformation, and let us adopt this notation throughout the paper.

Let Γ be a Borel set in the complex plane $\mathbf{C}$. Consider an operator-valued function $W:\Gamma\to\mathcal{L}(L^2(S^2))$, where $\mathcal{L}(L^2(S^2))$ is the space of bounded linear operators acting on $L^2(S^2)$. Then the quantity $|||W|||_{\alpha}$, which is given as

$$|||W|||_{\alpha}=\sup_{t\in\Gamma}\|W(t)\|+\sup_{t_1\neq t_2\in\Gamma}\frac{\|W(t_1)-W(t_2)\|}{|t_1-t_2|^{\alpha}},$$

where $\|\cdot\|$ is the operator norm on $L^2(S^2)$ and $\alpha\in(0,1]$, defines a complete norm on the Banach space $\mathcal{H}_{\alpha}[\Gamma;\mathcal{L}(L^2(S^2))]$ of Hölder-continuous operator functions[18,25] with exponent α.

Proposition 2.8: The $\tilde{S}(\xi)$ is Hölder continuous on the unit circle $\mathbf{T}$ with exponent $\frac{1}{4}$ if the potential $V(x)$ is in the Newton class with $s>1$ in (2.2) and $\beta=1$ in (2.3).

Proof: We have to show that $\exists M>0$ such that $\|\tilde{S}(\xi_1)-\tilde{S}(\xi_2)\|\leqslant M|\xi_1-\xi_2|^{1/4}$ for all $\xi_1,\xi_2\in\mathbf{T}$. Using

$$\|\tilde{S}(\xi_1)-\tilde{S}(\xi_2)\|=(1/2\pi)\|k_1A(k_1)-k_2A(k_2)\|$$

and

$$\xi_1-\xi_2=2i(k_1-k_2)/(k_1+i)(k_2+i),$$

we have

$$\frac{\|\tilde{S}(\xi_1)-\tilde{S}(\xi_2)\|}{|\xi_1-\xi_2|^{\epsilon}}=\frac{1}{2\pi}\frac{1}{2^{\epsilon}}(k_1^2+1)^{\epsilon/2}(k_2^2+1)^{\epsilon/2}\times\|k_2A(k_2)-k_1A(k_1)\|/|k_1-k_2|^{\epsilon}.$$

Because of the symmetry in k_1 and k_2, it is sufficient to show that $\lambda(k,\delta)$ is bounded by a constant independent of k and δ for all $\delta>0$ and $-\infty<k<\infty$, where

$$\lambda(k,\delta)=(k^2+1)^{\epsilon/2}[(k+\delta)^2+1]^{\epsilon/2}\times\|(k+\delta)A(k+\delta)-kA(k)\|(1/\delta^{\epsilon}).$$

In our proof we will use Propositions 2.5 and 2.7 and the constants C and B given there.

When $|k|\leqslant 1\leqslant\delta$, using $k^2+1\leqslant 2$, $(k+\delta)^2+1\leqslant 5\delta^2$, and

$$\|(k+\delta)A(k+\delta)-kA(k)\|\leqslant\|(k+\delta)A(k+\delta)\|+\|kA(k)\|\leqslant 2C,$$

we obtain $\lambda(k,\delta)\leqslant 2\cdot 10^{\epsilon/2}C$.

When $|k|\leqslant 1$, $\delta\leqslant 1$, using $k^2+1\leqslant 2$, $(k+\delta)^2+1\leqslant 5$,

$$\|(k+\delta)A(k+\delta)-kA(k)\|^{1-\epsilon}\leqslant(2C)^{1-\epsilon},$$

and

$$\|(k+\delta)A(k+\delta)-kA(k)\|^{\epsilon}\leqslant\left(\|A(k+\delta)\|+2(|k|+\delta)\max_{k\in\mathbf{R}}\left|\left|\frac{dA(k)}{dk}\right|\right|\right)^{\epsilon}\delta^{\epsilon}\leqslant(C+4B)^{\epsilon}\delta^{\epsilon},$$

we obtain $\lambda(k,\delta)\leqslant 10^{\epsilon/2}(2C)^{1-\epsilon}(C+4B)^{\epsilon}$.

When $1\leqslant|k|\leqslant\delta$, using $k^2+1\leqslant 2k^2$, $(k+\delta)^2+1\leqslant 5\delta^2$, and

$$\|(k+\delta)A(k+\delta)-kA(k)\|\leqslant 2C/(1+|k|),$$

we obtain $\lambda(k,\delta)\leqslant 2C\cdot 10^{\epsilon/2}[|k|^{\epsilon}/(1+|k|)]$.

When $1\leqslant\delta\leqslant|k|$, using

$$1/\delta^{\epsilon}\leqslant 1,\ k^2+1\leqslant 2k^2,(k+\delta)^2+1\leqslant 5k^2,$$

and

$$\|(k+\delta)A(k+\delta)-kA(k)\|\leqslant 2C/(1+|k|),$$

we obtain $\lambda(k,\delta)\leqslant 2C\cdot 10^{\epsilon/2}[|k|^{2\epsilon}/(1+|k|)]$.

When $\delta\leqslant 1\leqslant|k|$, using $k^2+1\leqslant 2k^2,(k+\delta)^2+1\leqslant 5k^2$,

$$\|(k+\delta)A(k+\delta)-kA(k)\|^{1-\epsilon}\leqslant(2C)^{1-\epsilon}/(1+|k|)^{1-\epsilon},$$

and

$$\begin{aligned}\|(k+\delta)A(k+\delta)-kA(k)\|^{\epsilon}&\leqslant\delta^{\epsilon}\left(\|A(k+\delta)\|+2(|k|+\delta)\max_{k\in\mathbf{R}}\left|\left|\frac{dA(k)}{dk}\right|\right|\right)^{\epsilon}\\&\leqslant\delta^{\epsilon}(C+4B)^{\epsilon}|k|^{\epsilon},\end{aligned}$$

we obtain

$$\lambda(k,\delta)\leqslant(2C)^{1-\epsilon}\cdot 10^{\epsilon/2}(C+4B)^{\epsilon}[|k|^{3\epsilon}/(1+|k|)^{1-\epsilon}].$$

Hence, whenever $0\leqslant\epsilon\leqslant\frac{1}{4}$, we have $\lambda(k,\delta)\leqslant M$, where M is a constant independent of k and δ. ■

Under weaker assumptions on the potential, we can modify Proposition 2.8 to obtain the following result.

Proposition 2.9: The $\tilde{S}(\xi)$ is Hölder continuous with exponent $2(1-s)/(5-2s)$ if the potential $V(x)$ belongs to the Newton class with some $s\in(\frac{1}{2},1)$ in (2.2) and $\beta=1$ in (2.3).

Proof: The only place in the proof of Proposition 2.8 where we have used Proposition 2.5 are the three cases $1\leqslant|k|\leqslant\delta$, $1\leqslant\delta\leqslant|k|$, and $\delta\leqslant 1\leqslant|k|$. In these three cases, we must use the result in Proposition 2.6 instead of the result in Proposition 2.5. This is accomplished by replacing $2C/(1+|k|)$ by $2E/(1+|k|)^{2-2s}$ in the proof of Proposition 2.8. We have the following.

Using

$$\|(k+\delta)A(k+\delta)-kA(k)\|\leqslant 2E/(1+|k|)^{2-2s},$$

we obtain $\lambda(k,\delta)\leqslant 2E\cdot 10^{\epsilon/2}[|k|^{\epsilon}/(1+|k|)^{2-2s}]$ when $1\leqslant|k|\leqslant\delta$, and

$$\lambda(k,\delta)\leqslant 2E\cdot 10^{\epsilon/2}[|k|^{2\epsilon}/(1+|k|)^{2-2s}],$$

when $1\leqslant\delta\leqslant|k|$.

When $\delta\leqslant 1\leqslant|k|$, we use

$$\begin{aligned}\|(k+\delta)A(k+\delta)-kA(k)\|^{\epsilon}&\leqslant\delta^{\epsilon}\left[\|A(k+\delta)\|+2(|k|+\delta)\max_{k\in\mathbf{R}}\left|\left|\frac{dA(k)}{dk}\right|\right|\right]^{\epsilon}\\&\leqslant\delta^{\epsilon}(E+4B)^{\epsilon}|k|^{\epsilon},\end{aligned}$$

and

$$\|(k+\delta)A(k+\delta)-kA(k)\|^{1-\epsilon}\leqslant\frac{(2E)^{1-\epsilon}}{(1+|k|)^{(1-\epsilon)(2-2s)}},$$

to obtain

$$\lambda(k,\delta)\leqslant(2E)^{1-\epsilon}\cdot 10^{\epsilon/2}(E+4B)^{\epsilon}\times|k|^{3\epsilon}/(1+|k|)^{(1-\epsilon)(2-2s)}.$$

Hence, whenever $0\leqslant\epsilon\leqslant 2(1-s)/(5-2s)$, we have $\lambda(k,\delta)\leqslant M$, where M is a constant independent of k and δ. ■

Proposition 2.10: The $A_1(k)$ defined in (2.7) is Hölder continuous with exponent β whenever the potential $V(x)$ belongs to the Newton class with the constant β in (2.3).

Proof: From (2.7) we have

$$A_1(k,\theta,\theta')=\int_{\mathbf{R}^3}dx\ V(x)e^{ik(\theta-\theta')\cdot x}.$$

Consider the operator $\mathscr{K}:V(x)\mapsto\langle A_1(k)f,g\rangle$, for some fixed $f,g\in L^2(S^2)$; i.e., consider

$$(\mathscr{K}V)(k)=\int\int\int_{\mathbf{R}^3\times S^2\times S^2}dx\,d\theta\,d\theta'\,V(x)\times e^{ik(\theta-\theta')\cdot x}f(\theta)\,\overline{g(\theta')},$$

where the bar denotes complex conjugation. The operator $\mathscr{K}$ is linear from $L^1(\mathbf{R}^3;dx)$, the space of Lebesgue integrable functions with respect to measure dx, into $\mathscr{H}_0$, the Banach space of bounded continuous functions on $\mathbf{R}$. The same operator $\mathscr{K}$ maps $L^1(\mathbf{R}^3;(1+|x|)dx)$ into $\mathscr{H}_1$, the Banach space of bounded Hölder-continuous functions on $\mathbf{R}$ with exponent 1. An application of an interpolation theorem presented by Krein *et al.* (Theorems III.3.5 and III.3.6 of Ref. 25) leads to the result that $\mathscr{K}$ maps $L^1(\mathbf{R}^3;(1+|x|)^{\beta}dx)$ into $\mathscr{H}^0_{\beta}$, where

$$\mathscr{H}^0_{\beta}=\{h\in\mathscr{H}_{\beta}:|h(k_1)-h(k_2)|=o(|k_1-k_2|^{\beta})\ \text{as}\ |k_1-k_2|\to 0\}.$$

Since this result is true uniformly in f,g on bounded subsets of $L^2(S^2)$, $A_1(k)$ belongs to $\mathscr{H}^0_{\beta}[\mathbf{R};\mathscr{L}(L^2(S^2))]$. Hence, $A_1(k)$ is Hölder continuous with exponent β whenever the potential $V(x)$ belongs to the Newton class where β is the constant in (2.3). Note that, strictly speaking, in order to apply Krein's result, one must restrict the function $(\mathscr{K}V)(k)$ to $k\in I$, where $I\subset\mathbf{R}$ is a compact interval, and observe that all the norm bounds are independent of I to pass to the case where $(\mathscr{K}V)(k)$ is considered for all $k\in\mathbf{R}$, which is the case here. ■

Using Proposition 2.10, we improve the results of Propositions 2.8 and 2.9 to obtain the following result that will be used in Sec. V.

Theorem 2.11: If the potential $V(x)$ belongs to the Newton class with some $\beta\in(0,1]$ in (2.3), then on Möbius transformation the corresponding scattering operator $\tilde{S}(\xi)$ belongs to $\mathscr{H}_{\mu}[\mathbf{T};\mathscr{L}(L^2(S^2))]$, where $\mu=\beta/2(1+\beta)$ if $s>1$ in (2.2) and $\mu=\beta(1-s)/(\beta-s+\frac{3}{2})$ if $s\in(\frac{1}{2},1)$ in (2.2). Here, $\mathscr{H}_{\mu}[\mathbf{T};\mathscr{L}(L^2(S^2))]$ is the Banach space of Hölder-continuous operator functions on the unit circle $\mathbf{T}$ with exponent μ.

Proof: Using (1.3) and (2.6) we have

$$S(k_1)-S(k_2)=\frac{1}{8\pi^2 i}\sum_{j=1}^{4}[k_1A_j(k_1)-k_2A_j(k_2)].$$

As mentioned at the end of the proof of Proposition 2.7, the only place where we used (2.3) was in the uniform boundedness of $\|dA_1(k)/dk\|$. Therefore, from the proof of Proposition 2.8, we obtain that for $j=2,3,4$, the operator $kA_j(k)$ is

Hölder continuous of exponent $\frac{1}{4}$ even for $\beta=0$ in (2.3). Hence, to prove the theorem, it is enough to redo the proof of Proposition 2.8 only for $A_1(k)$ and only in two cases; namely, when $|k|\leqslant 1, \delta\leqslant 1$ and when $\delta\leqslant 1\leqslant|k|$; i.e., it is enough to show that

$$\lambda_1(k,\delta)=(k^2+1)^{\epsilon/2}[(k+\delta)^2+1]^{\epsilon/2}\times\|(k+\delta)A_1(k+\delta)-kA_1(k)\|(1/\delta^{\epsilon})\leqslant M,$$

for $\epsilon\leqslant\beta/2(1+\beta)$ if $s>1$ in (2.2) and $\epsilon\leqslant\beta(1-s)/(\beta-s+\frac{3}{2})$ if $s\in(\frac{1}{2},1)$ in (2.2), where M is a constant independent of k and δ. Note that, whenever the potential $V(x)$ satisfies (2.3), from Proposition 2.10 we have $\|A_1(k+\delta)-A_1(k)\|\leqslant N\delta^{\beta}$, where N is a constant independent of k and δ. We will do the case when $s>1$ in (2.2) first.

When $\delta\leqslant 1, |k|\leqslant 1$, using $k^2+1\leqslant 2, (k+\delta)^2+1\leqslant 5$,

$$\begin{aligned}&\|(k+\delta)A_1(k+\delta)-kA_1(k)\|\\&\quad\leqslant|k|\cdot\|A_1(k+\delta)-A_1(k)\|+\delta\|A_1(k+\delta)\|\\&\quad\leqslant|k|N\delta^{\beta}+\delta C\leqslant(N+C)\delta^{\beta},\end{aligned}$$

we have $\lambda_1(k,\delta)\leqslant 10^{\epsilon/2}(N+C)\delta^{\beta-\epsilon}$.

When $\delta\leqslant 1\leqslant|k|$, using $k^2+1\leqslant 2k^2, (k+\delta)^2+1\leqslant 5k^2$,

$$\begin{aligned}&\|(k+\delta)A_1(k+\delta)-kA_1(k)\|^{1-\epsilon/\beta}\\&\quad\leqslant(2C/(1+|k|))^{1-\epsilon/\beta},\end{aligned}$$

and

$$\begin{aligned}\|(k+\delta)A_1(k+\delta)-kA_1(k)\|^{\epsilon/\beta}&\leqslant(2|k|N\delta^{\beta})^{\epsilon/\beta}\\&\quad+(2\delta C)^{\epsilon/\beta}\leqslant(2|k|\delta)^{\epsilon/\beta}(N^{\epsilon/\beta}+C^{\epsilon/\beta}),\end{aligned}$$

we have

$$\lambda_1(k,\delta)\leqslant 2\cdot 10^{1-\epsilon/\beta}(N^{\epsilon/\beta}+C^{\epsilon/\beta})\times\delta^{\epsilon/\beta-\epsilon}|k|^{2\epsilon+\epsilon/\beta}/(1+|k|)^{1-\epsilon/\beta}.$$

Thus, whenever $\epsilon\leqslant\beta/2(1+\beta)$, $\lambda_1(k,\delta)$ is bounded by a constant independent of k and δ, and the proof for $s>1$ is complete.

If $\frac{1}{2}<s<1$ in (2.2), we basically have the same proof with only two minor modifications, which amount to replacing the denominator $(1+|k|)$ by $(1+|k|)^{2-2s}$ and the constant C by E above. As a result, we obtain the sufficient condition

$$2\epsilon+\epsilon/\beta\leqslant 2(1-\epsilon/\beta)(1-s),$$

for the uniform boundedness of $\lambda_1(k,\delta)$. Hence, we must have

$$0<\epsilon\leqslant\beta(1-s)/(\beta-s+\tfrac{3}{2}),$$

which completes the proof. ■

III. RIEMANN–HILBERT PROBLEM

In the Schrödinger equation, k appears as k^2 and hence $\psi(-k,x,\theta)$ is a solution whenever $\psi(k,x,\theta)$ is. These two solutions are related to each other as[2]

$$\psi(k,x,\theta)=\int_{S^2}d\theta'\,S(k,-\theta,\theta')\psi(-k,x,\theta'). \quad (3.1)$$

Define

$$f(k,x,\theta)=e^{-ik\theta\cdot x}\psi(k,x,\theta). \quad (3.2)$$

If the potential satisfies (2.1) and if there are no bound states, for fixed x and θ, the function $f(k,x,\theta)$ has an analytic extension in k to $\mathbf{C}^+$ and $f(k,x,\theta)=1+O(1/|k|)$ as $|k|\to\infty$ there.[2] Similarly, $f(-k,x,\theta)$ has an analytic extension in k to $\mathbf{C}^-$. Hence, using (3.1), we obtain the Riemann–Hilbert problem

$$f(k,x,\theta)=\int_{S^2}d\theta'\,e^{-ik\theta\cdot x}S(k,-\theta,\theta')\times e^{-ik\theta'\cdot x}f(-k,x,\theta'). \quad (3.3)$$

Analogously, in the absence of bound states, we have the associated operator Riemann–Hilbert problem

$$F(k,x,\theta,\theta')=\int_{S^2}d\theta''\,e^{-ik\theta\cdot x}S(k,-\theta,\theta'')\times e^{-ik\theta''\cdot x}F(-k,x,-\theta'',-\theta'), \quad (3.4)$$

where, for fixed x,θ,θ', the operator $F(k,x,\theta,\theta')$ has an analytic extension in k to $\mathbf{C}^+$ and $F(k,x,\theta,\theta')=\delta(\theta-\theta')+O(1/|k|)$ as $|k|\to\infty$ there. Similarly, $F(-k,x,\theta,\theta')$ has an analytic extension in k to $\mathbf{C}^-$.

For fixed x,θ, and θ', let $\mathbf{X}(k,x,\theta,\theta')$ denote both the analytic extension in k of $F(k,x,\theta,\theta')-\delta(\theta-\theta')$ to $\mathbf{C}^+$ and the analytic extension of $F(-k,x,-\theta,-\theta')-\delta(\theta-\theta')$ to $\mathbf{C}^-$. Then $\mathbf{X}(k,x,\theta,\theta')$ is a sectionally analytic operator-valued function of k in the complex plane with a jump on the real axis. For $k\in\mathbf{R}$, define

$$\begin{aligned}\mathbf{X}_+(k,x,\theta,\theta')&=\lim_{\epsilon\to 0^+}\mathbf{X}(k+i\epsilon,x,\theta,\theta')\\&=F(k,x,\theta,\theta')-\delta(\theta-\theta'), \quad (3.5)\end{aligned}$$

$$\begin{aligned}\mathbf{X}_-(k,x,\theta,\theta')&=\lim_{\epsilon\to 0^+}\mathbf{X}(k-i\epsilon,x,\theta,\theta')\\&=F(-k,x,-\theta,-\theta')-\delta(\theta-\theta'), \quad (3.6)\end{aligned}$$

and

$$G(k,x,\theta,\theta')=e^{-ik\theta\cdot x}S(k,-\theta,-\theta')e^{ik\theta'\cdot x}. \quad (3.7)$$

Then, in operator notation, we can write (3.4) as

$$\mathbf{X}_+(k)=G(k)\mathbf{X}_-(k)+[G(k)-\mathbf{I}], \quad (3.8)$$

where we suppress the x dependence; note that x enters (3.8) only as a parameter. The operators $\mathbf{X}_+(k)$, $\mathbf{X}_-(k)$, $G(k)$, and $\mathbf{I}$ all act on $L^2(S^2)$. Let $\hat{1}$ be the constant function on this space defined as $\hat{1}(\theta)=1$, $\forall\theta\in S^2$. Let us define

$$X_+(k)=\mathbf{X}_+(k)\hat{1}=f(k,x,\theta)-1, \quad (3.9)$$

$$X_-(k)=\mathbf{X}_-(k)\hat{1}, \quad (3.10)$$

where $f(k,x,\theta)$ is as in (3.2). Then we can write (3.3) in vector form as

$$X_+(k)=G(k)X_-(k)+[G(k)-\mathbf{I}]\hat{1}. \quad (3.11)$$

If there are bound states, the extension of $f(k,x,\theta)$ in k to $\mathbf{C}^+$ becomes meromorphic with simple poles on the imaginary axis. A pole at $k=i\gamma$ corresponds to a bound state of the Hamiltonian with energy $-\gamma^2$. It is possible to remove these simple poles from the Riemann–Hilbert problem by a reduction method.[4] Assume there is a bound state corresponding to a pole at $k=i\gamma$. Using a suitable orthogonal projection $\mathbf{B}$, we form the rational function

$$\Pi(k) = \mathbf{I} - \mathbf{B} + [(k + i\gamma)/(k - i\gamma)]\mathbf{B}.$$

For the operators $\mathbf{X}_+(k)$, $\mathbf{X}_-(k)$, and $G(k)$, we then define the corresponding reduced operators

$$\mathbf{X}_+^{\text{red}}(k) = \Pi(k)^{-1}\mathbf{X}_+(k) + [\Pi(k)^{-1} - \mathbf{I}], \quad (3.12)$$

$$\mathbf{X}_-^{\text{red}}(k) = \Pi(k)\mathbf{X}_-(k) + [\Pi(k) - \mathbf{I}], \quad (3.13)$$

$$G^{\text{red}}(k) = \Pi(k)^{-1}G(k)\Pi(k).$$

Thus we have

$$X_+^{\text{red}}(k) = \Pi(k)^{-1}X_+(k) + [\Pi(k)^{-1} - \mathbf{I}]\hat{\mathbf{1}}, \quad (3.14)$$

$$X_-^{\text{red}}(k) = \Pi(k)X_-(k) + [\Pi(k) - \mathbf{I}]\hat{\mathbf{1}}. \quad (3.15)$$

As a result, $\mathbf{X}_+^{\text{red}}(k)$ and $X_+^{\text{red}}(k)$ do not have a pole at $k = i\gamma$, and $\mathbf{X}_-^{\text{red}}(k)$ and $X_-^{\text{red}}(k)$ do not have a pole at $k = -i\gamma$. If there is more than one bound state, this procedure must be repeated to remove the finitely many poles corresponding to the bound states; the details can be found in Ref. 4. This eventually leads to the operator Riemann–Hilbert problem

$$\mathbf{X}_+^{\text{red}}(k) = G^{\text{red}}(k)\mathbf{X}_-^{\text{red}}(k) + [G^{\text{red}}(k) - \mathbf{I}], \quad (3.16)$$

and the vector Riemann–Hilbert problem

$$X_+^{\text{red}}(k) = G^{\text{red}}(k)X_-^{\text{red}}(k) + [G^{\text{red}}(k) - \mathbf{I}]\hat{\mathbf{1}}. \quad (3.17)$$

Once the reduced Riemann–Hilbert problems (3.16) and (3.17) are solved, the solutions of the original Riemann–Hilbert problems (3.8) and (3.11) can be obtained using (3.12), (3.13), (3.14), and (3.15). Hence, in the following sections we will give the solutions of both the operator and vector Riemann–Hilbert problems assuming that $X_+(k)$ and $X_-(k)$, and similarly $\mathbf{X}_+(k)$ and $\mathbf{X}_-(k)$, have analytic extensions to $\mathbf{C}^+$ and $\mathbf{C}^-$, respectively.

IV. WIENER–HOPF FACTORIZATION OF THE SCATTERING MATRIX

The usual theory for the existence of Wiener–Hopf factorizations deals either with scalar functions[26] or with matrix functions.[27–30] In our case we study the Wiener–Hopf factorization of operator-valued functions. Hence, we must study Wiener–Hopf factorization in an infinite-dimensional setting[31] and use results on the existence of Wiener–Hopf factorizations of operator functions.[18,32,33]

By a **(left) Wiener–Hopf factorization** of an operator-valued function $G:\mathbf{R}_\infty \to \mathscr{L}(L^2(S^2))$, we mean a representation of $G(k)$ in the form

$$G(k) = G_+(k)D(k)G_-(k), \quad k\in\mathbf{R}_\infty, \quad (4.1)$$

with

$$D(k) = P_0 + \sum_{j=1}^{m}\left(\frac{k-i}{k+i}\right)^{\rho_j}P_j,$$

where

(i) $G_+(k)$ is continuous in $\mathbf{C}^+$ in the operator norm on $\mathscr{L}(L^2(S^2))$ and boundedly invertible there. Similarly, $G_-(k)$ is continuous in $\mathbf{C}^-$ in the operator norm and boundedly invertible there;

(ii) $G_+(k)$ is analytic in $\mathbf{C}^+$ and $G_-(k)$ is analytic in $\mathbf{C}^-$; and

(iii) $G_+(\infty) = G_-(\infty) = \mathbf{I}$.

The projections $P_1,...,P_m$ are finite in number, are mutually disjoint, and have rank one, while $P_0 = \mathbf{I} - \Sigma_{j=1}^m P_j$. The **(left) partial indices** $\rho_1,...,\rho_m$ are nonzero integers. In the absence of partial indices, we have $D(k) = \mathbf{I}$, in which case the Wiener–Hopf factorization is called **(left) canonical**. The partial indices depend neither on the choice of the factors $G_+(k)$ and $G_-(k)$ nor on the choice of the projections $P_1,...,P_m$. If the factorization is (left) canonical, the factors $G_+(k)$ and $G_-(k)$ are unique, as one sees by applying Liouville's theorem.

In the same way, we define a right Wiener–Hopf factorization, right partial indices, and a right canonical factorization by interchanging $G_+(k)$ and $G_-(k)$ in (4.1). The right indices may be different, both in number and in value, from the left indices, but the sum of the right indices coincides with the sum of the left indices. This sum is called the **sum index** of $G(k)$.

By using the Möbius transformation defined above Proposition 2.8, we can define the left and right Wiener–Hopf factorizations of operator functions on the unit circle $\mathbf{T}$ in the complex plane. The left and right partial indices are invariant under this Möbius transformation.

Remark 4.1: If $G(k)$ has a left Wiener–Hopf factorization of the form (4.1) with left partial indices $\rho_1,...,\rho_m$, then taking the inverses of both sides of (4.1) converts it into a right Wiener–Hopf factorization of $G(k)^{-1}$ with right partial indices $-\rho_1,...,-\rho_m$. On the other hand, if we consider the right Wiener–Hopf factorization

$$G(k) = \hat{G}_-(k)\hat{D}(k)\hat{G}_+(k), \quad k\in\mathbf{R}_\infty, \quad (4.2)$$

with

$$\hat{D}(k) = \hat{P}_0 + \sum_{j=1}^{\hat{m}}\left(\frac{k-i}{k+i}\right)^{\hat{\rho}_j}\hat{P}_j,$$

and take the adjoints on both sides of (4.1) with k replaced by its complex conjugate $\bar{k}$, we convert it into a right Wiener–Hopf factorization of $G(\bar{k})^\dagger$ with right partial indices $-\hat{\rho}_1,...,-\hat{\rho}_{\hat{m}}$. Hence, if $G(k)$ is unitary for every real k, which is the case in inverse scattering theory, the sets of left and right partial indices of $G(k)$ necessarily coincide. Moreover, the projections and factors appearing in (4.1) and (4.2) are related by

$$\hat{P}_j = (P_j)^\dagger \text{ for } j = 1,...,m; \text{ and } G_\pm(k)^{-1} = G_\mp(\bar{k})^\dagger.$$

In the remainder of this section we will only consider left Wiener–Hopf factorizations, though our results can also be derived for right Wiener–Hopf factorizations.

Theorem 4.2: If the potential $V(x)$ is in the Newton class, the operator function $G(k)$ defined in (3.7) has a left Wiener–Hopf factorization.

Proof: According to Theorem 6.1 (or 6.2) of Ref. 18, it is sufficient to show the following:

(i) $G(k)$ is boundedly invertible for every $k\in\mathbf{R}_\infty$;

(ii) $G(k)$ is a compact perturbation of the identity for every $k\in\mathbf{R}_\infty$; and

(iii) $\tilde{G}(\xi)\in\mathscr{H}_\alpha[\mathbf{T};\mathscr{L}(L^2(S^2))]$ for some $\alpha\in(0,1)$, where $\tilde{G}(\xi)$ is the Möbius transform of $G(k)$, as explained above in Proposition 2.8.

Under these conditions there exists a left Wiener–Hopf

factorization of $\tilde{G}(\xi)$ with respect to the unit circle $\mathbf{T}$ that is given by

$$\tilde{G}(\xi) = \tilde{G}_+(\xi)\tilde{D}(\xi)\tilde{G}_-(\xi),$$

where

$$\tilde{D}(\xi) = P_0 + \sum_{j=1}^{m} \xi^{\rho_j} P_j,$$

$\tilde{G}_+(\xi) \in \mathscr{H}_\alpha[\mathbf{T}^+; \mathscr{L}(L^2(S^2))]$ and is invertible there, $\tilde{G}_-(\xi) \in \mathscr{H}_\alpha[\mathbf{T}^-; \mathscr{L}(L^2(S^2))]$ and is invertible there, and $\tilde{G}_+(\xi)$ and $\tilde{G}_-(\xi)$ are analytic in $\mathbf{T}^+$ and in $\mathbf{T}^-$, respectively. The inverse of the Möbius transformation given above Proposition 2.8 then yields a left Wiener–Hopf factorization for $G(k)$ of the type (4.1) where the Möbius transformed factors $\tilde{G}_+(\xi)$ and $\tilde{G}_-(\xi)$ as well as their inverses are Hölder continuous of exponent α in the operator norm in $\mathbf{T}^+$ and $\mathbf{T}^-$, respectively.

First, note that we can use (3.7) to write

$$G(k) = U(k)QS(k)QU(k)^\dagger, \tag{4.3}$$

where

$$(U(k)f)(\theta) = e^{-ik\theta\cdot x} f(\theta), \quad (Qf)(\theta) = f(-\theta),$$

so that $G(k)$ is unitarily equivalent to $S(k)$. Hence, $G(k)$ is boundedly invertible for every $k \in \mathbf{R}_\infty$.

Next, since $A(k,\theta,\theta')$ is bounded and continuous in all three variables, it is Hilbert–Schmidt and hence compact as an operator on $L^2(S^2)$ for every real k. As a result,

$$\mathbf{I} - G(k) = -(k/2\pi i)U(k)QA(k)QU(k)^\dagger$$

is compact for every real k, and thus $G(k)$ is a compact perturbation of the identity.

Moreover, using (4.3) as well as the unitarity of $U(k)$ and Q, we have the estimate

$$\begin{aligned}\|G(k_1) - G(k_2)\| \leqslant{}& \|U(k_1) - U(k_2)\| \cdot \|S(k_1)\| \\ &+ \|S(k_1) - S(k_2)\| + \|S(k_2)\| \\ &\cdot \|U(k_1)^\dagger - U(k_2)^\dagger\|.\end{aligned}$$

Because $U(k)$ has a k derivative whose operator norm is uniformly bounded in k for every x, it is Lipschitz continuous in the operator norm with a Lipschitz constant independent of $k \in \mathbf{R}$. Further, according to Theorem 2.11 we have $\tilde{S}(\xi) \in \mathscr{H}_\mu[\mathbf{T}; \mathscr{L}(L^2(S^2))]$ for some $\mu \in (0,1)$. Hence, $\tilde{G}(\xi) \in \mathscr{H}_\mu[\mathbf{T}; \mathscr{L}(L^2(S^2))]$ for some positive μ.

Thus all three conditions needed to apply the above mentioned Gohberg–Leiterer result are satisfied, and the proof is complete. ■

Remark 4.3: Using the symmetry relation $G(-k) = QG(k)^{-1}Q$, we can prove that it is possible to choose $G_+(k)$ and $G_-(k)$ in (4.1) such that

$$G_\pm(-k) = QG_\mp(k)^{-1}Q. \tag{4.4}$$

Indeed, from (4.1) and using $D(-k) = D(k)^{-1}$ we have

$$\begin{aligned}G(k)^{-1} &= G_-(k)^{-1}D(k)^{-1}G_+(k)^{-1} \\ &= QG_+(-k)D(k)^{-1}G_-(-k)Q,\end{aligned}$$

so that

$$\begin{aligned}&G_+(k)^{-1}QG_-(-k)D(k) \\ &\quad = D(k)G_-(k)^{-1}QG_+(-k).\end{aligned} \tag{4.5}$$

If the factorization (4.1) is canonical, i.e., if $D(k) \equiv \mathbf{I}$, Liouville's theorem gives (4.4) directly from (4.5). If (4.1) is not a canonical factorization, we obtain

$$\begin{aligned}&((k+i)/(k-i))^{\rho_r - \rho_s} P_r G_+(k)^{-1} QG_-(-k)^{-1} P_s \\ &\quad = P_r G_-(k) QG_+(-k) P_s,\end{aligned}$$

where P_r and P_s are two of the projections appearing in $D(k)$ with $r, s \in \{0,1,2,\ldots,m\}$. If $\rho_r \leqslant \rho_s$, both sides of the last equation are equal to $P_r Q P_s$, due to Liouville's theorem. If $\rho_r > \rho_s$, however, we have

$$\begin{aligned}&P_r G_+(k)^{-1} QG_-(-k)^{-1} P_s \\ &\quad = \left[\phi_{rs}(k)/(k+i)^{\rho_r - \rho_s}\right] P_r Q P_s,\end{aligned} \tag{4.6}$$

$$\begin{aligned}&P_r G_-(k) QG_+(-k) P_s \\ &\quad = \left[\phi_{rs}(k)/(k-i)^{\rho_r - \rho_s}\right] P_r Q P_s,\end{aligned} \tag{4.7}$$

where $\phi_{rs}(k)$ is a polynomial of degree $(\rho_r - \rho_s)$ with leading coefficient 1. Using the procedure in Ref. 31, we can multiply $G_\pm(k)$ by suitable rational functions and therefore change our original factorization (4.1) in such a way that both sides of (4.6) and (4.7) reduce to $P_r Q P_s$. Then, using $\Sigma_{r=0}^m P_r = \Sigma_{s=0}^m P_s = \mathbf{I}$, we find (4.4) for this modified factorization.

V. SOLUTION OF THE RIEMANN–HILBERT PROBLEM

In Sec. IV we have derived the existence of a Wiener–Hopf factorization of the operator function relevant to the Riemann–Hilbert problems (3.8) and (3.11). This result was obtained under the assumption that the potential $V(x)$ belongs to the Newton class. In this section we will use the factorization (4.1) to obtain the solutions of the Riemann–Hilbert problems (3.8) and (3.11). During the process the variable x enters as a dummy variable, which may affect the partial indices and the factors in (4.1) and hence the unique solvability properties of (3.8) and (3.11) and the explicit form of their solutions, but does not affect the way in which the solution itself is obtained. Therefore, to simplify our notation we suppress the x dependence of all vectors, operators, and partial indices.

Starting from the Wiener–Hopf factorization (4.1) of $G(k)$, we define

$$D_+(k) = P_0 + \sum_{\rho_j > 0} \left(\frac{k-i}{k+i}\right)^{\rho_j} P_j + \sum_{\rho_j < 0} P_j$$

and

$$D_-(k) = P_0 + \sum_{\rho_j > 0} P_j + \sum_{\rho_j < 0} \left(\frac{k-i}{k+i}\right)^{\rho_j} P_j,$$

where $P_1, \ldots, P_m$ are the mutually disjoint, rank one projections appearing in the diagonal factor $D(k)$ and $P_0 = \mathbf{I} - \Sigma_{j=1}^m P_j$.

Using (4.1), let us write (3.11) in the form

$$\begin{aligned}X_+(k) = {}& G_+(k)D_+(k)D_-(k)G_-(k)X_-(k) \\ &+ [G_+(k)D_+(k)D_-(k)G_-(k) - \mathbf{I}]\hat{1},\end{aligned} \tag{5.1}$$

where $\hat{1}$ is the function in $L^2(S^2)$ as defined above (3.9). Then

$$D_+(k)^{-1}G_+(k)^{-1}X_+(k) = D_-(k)G_-(k)X_-(k) + [D_-(k)G_-(k) - D_+(k)^{-1}G_+(k)^{-1}]\hat{1}. \quad (5.2)$$

Premultiplying both sides by P_0 yields

$$P_0G_+(k)^{-1}X_+(k) + P_0G_+(k)^{-1}\hat{1} = P_0G_-(k)X_-(k) + P_0G_-(k)\hat{1}. \quad (5.3)$$

The left-hand side of (5.3) is analytic in $\mathbf{C}^+$, the right-hand side is analytic in $\mathbf{C}^-$, and both sides tend to $P_0\hat{1}$ as $k\to\infty$ from the appropriate half-plane. Hence, by Liouville's theorem,

$$P_0G_+(k)^{-1}X_+(k) = P_0[\mathbf{I} - G_+(k)^{-1}]\hat{1} \quad (5.4)$$

and

$$P_0G_-(k)X_-(k) = P_0[\mathbf{I} - G_-(k)]\hat{1}. \quad (5.5)$$

Similarly, premultiplying both sides of (5.2) by $(k-i)^{\rho_j}P_j$ with $\rho_j > 0$ and using Liouville's theorem, we obtain

$$P_jG_+(k)^{-1}X_+(k) = P_j[\mathbf{I} - G_+(k)^{-1}]\hat{1} + [\varphi_j(k)/(k+i)^{\rho_j}]\pi_j \quad (5.6)$$

and

$$P_jG_-(k)X_-(k) = -P_jG_-(k)\hat{1} + [\varphi_j(k)/(k-i)^{\rho_j}] \times \pi_j + [(k+i)/(k-i)]^{\rho_j}P_j\hat{1}. \quad (5.7)$$

Here, π_j is a fixed nonzero vector in the range of P_j, and $\varphi_j(k)$ is an arbitrary polynomial of degree less than ρ_j. Next, premultiplication of both sides of (5.2) by P_j with $\rho_j < 0$ and yet another application of Liouville's theorem yield

$$P_jG_+(k)^{-1}X_+(k) = P_j[\mathbf{I} - G_+(k)^{-1}]\hat{1} \quad (5.8)$$

and

$$P_jG_-(k)X_-(k) = -P_jG_-(k)\hat{1} + [(k+i)/(k-i)]^{\rho_j}P_j\hat{1}, \quad (5.9)$$

provided the second term on the right-hand side of (5.9) is analytic at $k = -i$. Because $\rho_j < 0$, the latter happens if and only if $P_j\hat{1} = 0$.

Finally, adding (5.4), (5.6), and (5.8) together as well as (5.5), (5.7), and (5.9), and using $P_0 + \Sigma_{\rho_j>0}P_j + \Sigma_{\rho_j<0}P_j = \mathbf{I}$, we obtain

$$X_+(k) = [G_+(k) - \mathbf{I}]\hat{1} + G_+(k)\sum_{\rho_j>0}\frac{\varphi_j(k)}{(k+i)^{\rho_j}}\pi_j \quad (5.10)$$

and

$$X_-(k) = [G_-(k)^{-1} - \mathbf{I}]\hat{1} + G_-(k)^{-1} \times \sum_{\rho_j>0}\frac{\varphi_j(k)\pi_j + [(k+i)^{\rho_j} - (k-i)^{\rho_j}]P_j\hat{1}}{(k-i)^{\rho_j}}, \quad (5.11)$$

provided $P_j\hat{1} = 0$ whenever $\rho_j < 0$. Hence, if these $(-\Sigma_{\rho_j<0}\rho_j)$ linear constraints on P_j for $\rho_j < 0$ are satisfied, there is a $(\Sigma_{\rho_j>0}\rho_j)$ parameter family of solutions to (3.11), and these solutions are given by (5.10) and (5.11).

We can summarize the above results as follows.

Theorem 5.1: Let $V(x)$ be a potential in the Newton class. Then the Riemann–Hilbert problem (3.11) has a solution, if and only if $P_j\hat{1} = 0$ whenever $\rho_j < 0$. In that case the solutions are given by (5.10) and (5.11), where $\varphi_j(k)$ is an arbitrary polynomial of degree less than ρ_j associated with each $\rho_j > 0$.

The solution of the operator Riemann–Hilbert problem (3.8) is obtained in the same way as the vector Riemann–Hilbert problem (3.11) is solved using (5.1) through (5.11). The solution of (3.8) is given by

$$\mathbf{X}_+(k) = G_+(k) - \mathbf{I} + G_+(k)\sum_{\rho_j>0}\frac{\varphi_j(k)}{(k+i)^{\rho_j}}P_j \quad (5.12)$$

and

$$\mathbf{X}_-(k) = G_-(k)^{-1} - \mathbf{I} + G_-(k)^{-1} \times \sum_{\rho_j>0}\frac{\varphi_j(k) + [(k+i)^{\rho_j} - (k-i)^{\rho_j}]}{(k-i)^{\rho_j}}P_j, \quad (5.13)$$

provided there are no negative partial indices. If there are any negative partial indices, the solution does not exist. Due to the presence of $\varphi_j(k)$ in (5.12) and (5.13), the solution is not unique unless there are no positive partial indices.

Note that, when there are no bound states, for $x = 0$, the operator $[\mathbf{I} + \mathbf{X}_+(k)]$ becomes related to the 3-D Jost operator used in the 3-D Gel'fand–Levitan inversion method.[2–4] Hence, we obtain the following result.

Corollary 5.2: If the potential $V(x)$ belongs to the Newton class with no bound states, the Jost operator exists if and only if there are no partial indices of the scattering operator. In that case the Jost operator is given by

$$J(k) = QS_+(k)Q, \quad (5.14)$$

where $S_+(k)$ is the operator that is given by $G_+(k)$ evaluated at $x = 0$.

VI. SOLUTION OF THE INVERSE PROBLEM

Once the Riemann–Hilbert problem posed in (3.11) is solved by the Wiener–Hopf factorization method given in Sec. V, we obtain $f(k,x,\theta)$ given in (3.2) using (3.9). If there are no bound states, from the Schrödinger equation (1.1) we then obtain the potential as

$$V(x) = \frac{(\Delta + 2ik\theta\cdot\nabla)X_+(k,x,\theta)}{1 + X_+(k,x,\theta)}. \quad (6.1)$$

Note that the right-hand side of this equation contains θ and k whereas these two variables are absent from the left-hand side. Hence, the solution of the Riemann–Hilbert problem will lead to a potential only if the right-hand side of (6.1) is independent of θ and k. Below, we show that if the so-called miracle condition[2] occurs, the right-hand side of (6.1) is independent of θ and k and becomes equal to a potential function of x.

Let the Fourier transform of $X_+(k,x,\theta)$ be given by

$$\eta(\alpha,x,\theta) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}dk\,X_+(k,x,\theta)e^{-ik\alpha}. \quad (6.2)$$

Since $X_+(k,x,\theta) = O(1/|k|)$ as $k \to \pm\infty$ and is analytic in k in $\mathbf{C}^+$, we have $\eta(\alpha,x,\theta) = 0$ for $\alpha < 0$. The function $\eta(\alpha,x,\theta)$ plays a major role in the 3-D Newton–Marchenko inversion theory.[2] In case the Riemann–Hilbert problem (3.11) has a unique solution, $\eta(\alpha,x,\theta)$ satisfies the equation[2]

$$\left[\Delta - 2\frac{\partial}{\partial\alpha}\theta\cdot\nabla - V(x)\right]\eta(\alpha,x,\theta) = 0, \tag{6.3}$$

where the potential is obtained as

$$V(x) = -2\theta\cdot\nabla \lim_{\alpha\to 0^+} \eta(\alpha,x,\theta), \tag{6.4}$$

provided the right-hand side of (6.4) is independent of θ. The θ independence of the right-hand side of (6.4) is known as the "miracle" condition of Newton.[2]

From (6.2) we have

$$ikX_+(k,x,\theta) = -\lim_{\alpha\to 0^+}\eta(\alpha,x,\theta) - \int_0^\infty d\alpha\, e^{ik\alpha}\frac{\partial}{\partial\alpha}\eta(\alpha,x,\theta). \tag{6.5}$$

Hence, using (6.2), (6.4), and (6.5), we obtain

$$[\Delta + 2ik\theta\cdot\nabla - V(x)]X_+(k,x,\theta) = V(x) + \int_0^\infty d\alpha\, e^{ik\alpha}\left[\Delta - 2\frac{\partial}{\partial\alpha}\theta\cdot\nabla - V(x)\right]\eta(\alpha,x,\theta).$$

Thus (6.1) is equivalent to (6.3) and (6.4) in the absence of bound states.

If there are any bound states, the above procedure can be modified to prove that the potential $V(x)$ is obtained from (6.1) if and only if (6.3) and (6.4) hold true.[34]

VII. CONCLUSION

In this paper we have established the following results. If the potential $V(x)$ belongs to the Newton class defined in Sec. II, the corresponding scattering operator has a Wiener–Hopf factorization. The related Riemann–Hilbert problem (3.11) can be solved by using these factors. The related operator Riemann–Hilbert problem (3.8) is also solvable by using the Wiener–Hopf factors. A consequence of this is the following. For potentials in the Newton class with no bound states, the Jost operator (as defined in Ref. 2) exists if and only if the corresponding scattering operator does not have any partial indices. If and only if Newton's miracle condition is satisfied, the solution of the Riemann–Hilbert problem leads to a potential.

The physical interpretation of the partial indices of the scattering operator is an open problem. It is known that the total index is related to the total number of bound states of the potential, but the relationship of each partial index to the bound states or to any physical parameters is presently not known.

A simple condition[34] that guarantees the unique solvability of the Riemann–Hilbert problems (3.8) and (3.11) is given by $\max_{k\in\mathbf{R}}\|S(k) - \mathbf{I}\| < 1$, where the norm is the operator norm on $L^2(S^2)$. When this happens, the scattering operator $S(k)$ has neither positive nor negative partial indices.

The results presented in this paper remain valid for any real, measurable potential $V(x)$ on $\mathbf{R}^n$ with $n \geqslant 2$ without real exceptional points that lead to a scattering operator $S(k)$ such that $S(k) - \mathbf{I}$ is compact for all $k\in\mathbf{R}$ and that $\tilde{S}(\xi) = S(i(1+\xi)/(1-\xi))$ is Hölder continuous in $\xi\in\mathbf{T}$.

ACKNOWLEDGMENTS

The authors are indebted to Roger Newton for his comments.

The research leading to this article was supported in part by the National Science Foundation under Grants DMS 8823102 and DMS 9001903.

[1] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1989), 2nd ed.
[2] R. G. Newton, J. Math. Phys. **21**, 1698 (1980); **22**, 631 (1981); **23**, 693 (1982).
[3] R. G. Newton, J. Math. Phys. **22**, 2191 (1981); **23**, 693 (1982).
[4] R. G. Newton, J. Math. Phys. **23**, 2257 (1982).
[5] R. G. Newton, *Scattering Theory in Mathematical Physics*, edited by J. A. Lavita and J.-P. Marchand (Reidel, Dordrecht, The Netherlands, 1974), p. 193.
[6] A. I. Nachman and M. J. Ablowitz, Stud. Appl. Math. **71**, 243 (1984).
[7] R. Beals and R. R. Coifman, Proc. Symp. Pure Math. **43**, 45 (1985).
[8] R. Beals and R. R. Coifman, Physica D **18**, 242 (1986).
[9] R. G. Novikov and G. M. Henkin, Sov. Math. Dokl. **35**, 153 (1987) [Dokl. Akad. Nauk SSSR **292**, 814 (1987) (Russian)].
[10] R. T. Prosser, J. Math. Phys. **10**, 1819 (1969).
[11] R. T. Prosser, J. Math. Phys. **17**, 1775 (1976).
[12] R. T. Prosser, J. Math. Phys. **21**, 2648 (1980).
[13] R. T. Prosser, J. Math. Phys. **23**, 2127 (1982).
[14] L. D. Faddeev, Sov. Phys. Dokl. **10**, 1033 (1965) [Dokl. Akad. Nauk SSSR **165**, 514 (1965) (Russian)].
[15] L. D. Faddeev, J. Sov. Math. **5**, 334 (1976) [Itogi Nauki Tekh. **3**, 93 (1974) (Russian)].
[16] R. G. Newton, Inverse Problems **1**, 371 (1985).
[17] R. G. Newton, *Inverse Schrödinger Scattering in Three Dimensions* (Springer, New York, 1989).
[18] I. C. Gohberg and J. Leiterer, Math. Nachrichten **55**, 33 (1973) (Russian).
[19] R. G. Newton, J. Math. Phys. **18**, 1348 (1977).
[20] M. S. Birman, Am. Math. Soc. Transl. **53**, 23 (1966) [Mat. Sb. **55**, 125 (1961) (Russian)].
[21] J. Schwinger, Proc. Nat. Acad. Sci. U.S.A. **47**, 122 (1961).
[22] T. Kato, Commun. Pure Appl. Math. **12**, 403 (1959).
[23] R. Weder, Inverse Problems **6**, 267 (1990).
[24] L. Vega, Proc. Am. Math. Soc. **102**, 874 (1988).
[25] S. G. Krein, Yu. I. Petunin, and E. M. Semenov, *Interpolation of Linear Operators* [Transl. Math. Monographs, Vol. 54 (Am. Math. Soc., Providence, 1981)] [Nauka, Moscow, 1978 (Russian)].
[26] N. I. Muskhelishvili, *Singular Integral Equations* (Noordhoff, Groningen, The Netherlands, 1953) [Nauka, Moscow, 1946 (Russian)].
[27] I. C. Gohberg and M. G. Krein, Am. Math. Soc. Transl. **14**, 217 (1960) [Uspekhi Mat. Nauk **13** (2), 3 (1959) (Russian)].
[28] N. P. Vekua, *Systems of Singular Integral Equations* (Noordhoff, Groningen, The Netherlands, 1967) [Nauka, Moscow, 1950 (Russian)].
[29] I. C. Gohberg and I. A. Feldman, *Convolution Equations and Projection Methods for their Solution* [Transl. Math. Monographs, Vol. 41 (Am. Math. Soc., Providence, 1974)] [Nauka, Moscow, 1971 (Russian)].
[30] K. Clancey and I. Gohberg, *Factorization of Matrix Functions and Singular Integral Operators* (Birkhäuser OT **3**, Boston, 1981).
[31] I. C. Gohberg, Am. Math. Soc. Transl. **49**, 130 (1966) [Izvestiya Akad. Nauk SSSR, Ser. Matem. **28**, 1055 (1964) (Russian)].
[32] I. C. Gohberg and J. Leiterer, Math. Nachrichten **54**, 41 (1972) (Russian).
[33] I. C. Gohberg and J. Leiterer, Sov. Math. Dokl. **14**, 425 (1973) [Dokl. Akad. Nauk SSSR **209**, 529 (1973) (Russian)].
[34] T. Aktosun and C. van der Mee, SIAM J. Math. Anal. (to be published).

Scattering for step-periodic potentials in one dimension

Thomas M. Roberts[a)]
Mathematical Physics Program, Indiana University, Bloomington, Indiana 47405

(Received 29 September 1989, accepted for publication 9 May 1990)

Quantum scattering is developed for impurities in potentials that tend to a periodic function in one direction and a constant in the other. Two new technical results are obtained for Hill's equation. Analytic, asymptotic, and spectral properties are established for solutions of the Schrödinger equation for step-periodic potentials, with and without impurity. The properties have all been used in Marchenko–Newton inverse scattering. Results apply feasibly to electron, photon, and phonon propagation in layered media.

I. INTRODUCTION

This scattering paper contributes to a novel line of inquiry in inverse scattering. The goal is inverse scattering for Schrödinger operators whose spectra are interesting.

The one-dimensional Schrödinger equation

$$[-\partial_x^2 + p(x) + q(x) - \lambda]f = 0 \tag{1.1}$$

represents a localized impurity q, which vanishes as $x \to \pm\infty$, in a background p. The seminal inverse problem, which has been studied voluminously,[1,2] has vacuous background $p \equiv 0$.

The background operator for the seminal problem is $-\partial_x^2$. Its spectrum $[0, +\infty)$ and continuous spectrum are identical. The localized impurity q does not affect the continuous spectrum, but is responsible for the point spectrum of the full operator, whose potential is $p+q$. Data are measured over the spectrum of the full operator.

The next more complicated Schrödinger operator[3] $-\partial_x^2 + p + q$ has a localized impurity q in a Heaviside step background $p = |V_0|H(x)$, where $|V_0|$ is constant, $H(x<0) \equiv 0$, and $H(x \geqslant 0) \equiv 1$. The background operator $-\partial_x^2 + p$ has no point spectrum, though the impurity can introduce bound state eigenvalues at negative energies λ. The continuous spectra of background $(-\partial_x^2 + p)$ and full $(-\partial_x^2 + p + q)$ operators are identical. The continuous spectrum $[0, +\infty)$ has multiplicity[4] one in $[0, |V_0|)$ and multiplicity two in $[|V_0|, +\infty)$. Data on the multiplicity-two part consist of four scattering coefficients, corresponding to reflection and transmission of waves incident from left and right. Data for multiplicity-one regions have fewer scattering coefficients, which correspond to total reflection of waves incident from $x = -\infty$. Also measured are eigenvalues, which form the point spectrum, and physical data about normalization of L_2 eigenfunctions.

We see that spectra support data. As spectra become more interesting, so does the nature of data.

Periodic backgrounds were studied next. Firsova[5] and Newton[6] solved independently, and with different methods, inverse problems for localized impurities q in periodic backgrounds p, for which the Schrödinger equation is $(-\partial_x^2 + p + q - \lambda)f = 0$ with $p(x+1) = p(x)$. The spectrum, and hence the data, has interesting structure. The continuous spectrum has multiplicity two and consists of infinitely many intervals, called bands, each with finite, nonzero length. Intervals between bands are called gaps. Bands and gaps alternate as one traces toward $+\infty$ along the axis of real λ. There is one more gap, which is a ray of real energies λ extending toward $-\infty$. The point spectrum of the periodic-background operator is empty, but for a large class of impurities[7] the point spectrum of $-\partial_x^2 + p + q$ is *unbounded* and each gap of sufficiently high energy has at least one bound state.

This paper is about localized impurities q in step-periodic backgrounds $p(x) = p(x+1)H(x)$, which are periodic for $x > 0$ and constant for $x < 0$. The spectrum of $-\partial_x^2 + p$ combines features of periodicity and step. The step-periodic case has bands and gaps, a nonempty point spectrum, and a continuous spectrum with intervals of multiplicity two and of multiplicity one. This paper develops theorems on analyticity and asymptotics used[8] in Marchenko–Newton inverse scattering.

Physical applications further motivate the step-periodic problem. Widely known techniques, which are reviewed in Ref. 9, show electromagnetic (photon) and acoustic (phonon) propagation in one dimension are modeled exactly by the Schrödinger equation. Schrödinger models of periodically layered media have periodic potentials, so bands and gaps are expected. There is experimental evidence[10] for bands and gaps in photon and phonon spectra of periodically layered materials, establishing that the Schrödinger model plausibly is applicable.

The infinitude of bound states[5-7] for impure periodic potentials *without* steps has two inconvenient practical consequences, which are significant because the Schrödinger model plausibly is applicable. One inconvenience is numerical: An infinite process[5,6] is needed in inverse scattering to remove from data for $p+q$ the effect of infinitely many bound states. A more basic objection is that materials less than infinitely thick—for which step-periodic potentials are more realistic[11] models—should have only finitely many bound states. Theorems 3.1 and 4.1 show that pure and impure step-periodic potentials, unlike many impure periodic potentials, have finitely many bound states. Finiteness is an advantage.

This paper's main goal is to find a functional[12] S matrix, called $\hat{S}$, such that $\hat{S} - \mathbf{I}$ has a Fourier transform. The trans-

[a)] Current address: Applied Mathematical Sciences, Ames Laboratory, United States Department of Energy, Ames, Iowa 50011.

 0022-2488/90/092181-11$03.00

form is used[8] in inverse scattering, as are the other results in this paper. Other results include equivalence of bound states and zeros of Wronskians, and theorems on analytic and asymptotic properties of Wronskians, Jost solutions, scattering solutions, regular solutions, and the Jost matrix.

I know of three papers about step-periodic or similar potentials. Pavlov and Smirnov[11] show that one reflection coefficient for a step-periodic potential (without impurity) is approximated well by a reflection coefficient for a large finite number of repeated layers. Hinton, Klaus, and Shaw[13] use an interesting generalization of the Weyl function m [see (2.3)] to study half-bound states for a Schrödinger equation whose background p is periodic on the half-line $[0,+\infty)$, whose impurity q is supported on $[0,+\infty)$, and whose wave functions obey the boundary condition $f(0)=0$. Gesztesy[14] studies the spectrum and scattering coefficients for two adjoined semi-infinite crystals.

II. PERIODIC POTENTIALS

The Schrödinger equation $[-\partial_x^2+p(x)+q(x)-\lambda]f=0$, for a localized impurity q embedded in a background p that is constant on the left $[p(x)H(-x)\equiv\lambda_0 H(-x)]$ and periodic on the right $[p(x)H(x)\equiv p(x+1)H(x)]$, will be studied in stages. This section is the first stage, whose Schrödinger equation $[-\partial_x^2+p_a(x)-\lambda]f=0$ has a potential $p_a(x+1)\equiv p_a(x)$ that is a periodic extension of p: $p_a H(x)\equiv pH(x)$. The Schrödinger equation with a periodic potential is called Hill's equation. There is a voluminous literature[15] on Hill's equation, from which excerpts form the bulk of this section. The section ends with two new lemmas on Hill's equation.

The following abbreviations are used throughout: $\mathbf{Z}\equiv\{\text{integers}\}$, $\mathbf{R}\equiv\{\text{real numbers}\}$, $\mathbf{C}\equiv\{\text{complex numbers}\}$, $\mathbf{C}^+\equiv\{z\in\mathbf{C}\colon \operatorname{Im} z>0\}$, "on $\mathbf{R}_x$" means $\forall x\in\mathbf{R}$, and $[a,b]_\lambda\equiv\{\lambda\colon \lambda\in[a,b]\}$. The symbol $\ni$ means *such that* and "e.p." means *except in pathological cases* and alludes to sets of measure zero. The phrase $\lambda\in\mathbf{C}_\sigma^+$, in context of the definition $\sigma\equiv\sqrt{\lambda}$, means $\exists\sigma\in\mathbf{C}^+\ni\lambda=\sigma^2$. The symbol ^{T} means transpose, * means complex conjugate, $\hat{1}\equiv(1,1)^T$, $f'(\lambda,x)\equiv\partial_x f$, and the Wronskian $W[f,g]\equiv fg'-f'g$. The phrases s analytic, s meromorphic, and s entire mean analytic, meromorphic, and entire as functions of s. An s simple pole has the form $(s-s_0)^{-1}$.

Hill's equation has regular solutions $y=(y_1,y_2)^T$ defined on $\mathbf{C}_\sigma\times\mathbf{R}_x$ by boundary conditions

$$y(x=0)\equiv(1,0)^T,\ y'(x=0)\equiv(0,1)^T,$$
$$s\equiv\begin{cases}\varsigma, & \lambda_0<0\\ \sigma, & \lambda_0>0\end{cases}\equiv\begin{cases}\sqrt{\lambda-\lambda_0}, & \lambda_0<0,\\ \sqrt{\lambda}, & \lambda_0>0,\end{cases}\tag{2.1}$$

where λ_0 is defined in this section's first sentence. The zero of energy (λ) is defined, following custom, as the lowest energy that is still in a band. The regular solutions y have analytic continuations which are σ entire with asymptotics

$$\begin{aligned}y_1&=\cos\sigma x+O(\sigma^{-1}e^{|\nu x|}),\\ y_2&=\sigma^{-1}\sin\sigma x+O(\lambda^{-1}e^{|\nu x|})\end{aligned}\tag{2.2}$$

as $|\sigma|\to\infty$ with $\operatorname{Im}\sigma=\nu$ fixed. Weyl functions $m=(m_1,m_2)^T$, quasimomentum κ, and other x independent quantities $\epsilon=(\epsilon_1,\epsilon_2)^T$ and $\zeta=(\zeta_1,\zeta_2)^T$ are defined on $\mathbf{C}_\sigma$ by

$$\begin{aligned}&\epsilon\equiv y(x=1),\ \zeta\equiv y'(1),\\ &\kappa\equiv\cos^{-1}[(\epsilon_1+\zeta_2)/2],\\ &m_j\equiv(e^{\pm i\kappa}-\epsilon_1)/\epsilon_2=\zeta_1/(\epsilon_1-e^{\mp i\kappa}).\end{aligned}\tag{2.3}$$

From among branches[16] of $\cos^{-1}$, use the one that satisfies $\cos^{-1}1\equiv 0$. Then the well-established fact that $(\epsilon_1+\zeta_2)/2=1$ at $\lambda=0$ implies $\kappa=0$ there.

The map κ separates $\mathbf{R}_\lambda$ into infinitely many bands $\mathscr{B}$ and gaps $\mathscr{G}$, illustrated in Fig. 1. (Let $\mathscr{B}\mathscr{G}\equiv\mathscr{B}\cup\mathscr{G}$.) Each band is an interval $[n\pi,(n+1)\pi]\subset\mathbf{R}_\kappa$ with $n\in\mathbf{Z}$. Bands in $\mathbf{C}_\lambda$ and $\mathbf{C}_\sigma$ also are closed intervals of nonzero length. In $\mathbf{C}_\lambda$, the gaps are $(-\infty,0)$ and the open real intervals between bands. Gaps are symmetric across $\mathbf{R}_\kappa$ and their lengths go to zero as $|\operatorname{Re}\kappa|\to\infty$. Lengths of gaps in $\mathbf{C}_\lambda$ and $\mathbf{C}_\sigma$ also go to zero as $\lambda\uparrow+\infty$.

The map $\sqrt{}$ from $\mathbf{C}_\lambda$ onto $\mathbf{C}_\sigma^+\cup\mathbf{R}_\sigma^+$ is trivial. The map κ: $\mathbf{C}_\sigma\to\mathbf{C}_\kappa$ sends one imaginary axis onto the other, $\mathscr{B}_\sigma$ onto $\mathbf{R}_\kappa$, and $\mathbf{C}_\sigma^\pm$ into $\mathbf{C}_\kappa^\pm$, respectively. The map κ is one-to-one in $\mathscr{B}$ and $[0,+i\infty]_\kappa$, and in gaps is described by paths in Fig. 1. The map is analytic (without poles) in $\mathbf{C}_\sigma^{\text{cut}}\equiv\mathbf{C}_\sigma\setminus(\mathscr{G}\cap\mathbf{R}_\sigma)$, its symmetry properties follow directly from Schwarz's reflection principle, and it has an inverse in $\mathbf{C}_\kappa^{\text{cut}}$.

Boundary values (2.1) and asymptotics (2.2) show that y is a periodic-potential analog of trigonometric solutions of the Schrödinger equation with potential zero. Quasimomentum κ will be used to describe a periodic-case analog of exponential solutions.

Bloch solutions $\beta=(\beta_1,\beta_2)^T$ are defined on $\mathbf{C}_\kappa\times\mathbf{R}_x$ as

$$\beta\equiv y_1\hat{1}+(m_1,m_2)^T y_2.\tag{2.4}$$

The solutions have simple x dependence:[17] $\exists\xi(\kappa,x)=(\xi_1,\xi_2)^T\in L_\infty(\mathscr{B}\times\mathbf{R}_x)\ni$

$$\beta_1=e^{i\kappa x}\xi_1,\ \beta_2=e^{-i\kappa x}\xi_2,\ \xi(x+1)=\xi(x)\tag{2.5}$$

in $\mathbf{C}_\kappa\times\mathbf{R}_x$. The solutions have useful properties:[18]

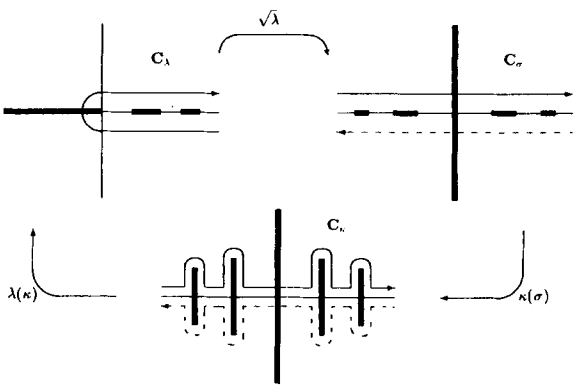


FIG. 1. The spectrum for a step-periodic potential. The thicker intervals, lines, and ray represent gaps. Bands are complements in R of gaps. The solid path in $\mathbf{C}_\lambda$ maps onto solid and dashed curves in $\mathbf{C}_\sigma$ and $\mathbf{C}_\kappa$.

$$\beta(x=0)=\hat{1},\quad \beta'(0)=m,$$
$$m\colon \mathscr{G}\to\mathbf{R},\quad \beta\colon \mathscr{G}\times\mathbf{R}_x\to\mathbf{R},$$
$$y\colon \{\sigma,i\sigma\colon \sigma\in\mathbf{R}\}\times\mathbf{R}_x\to\mathbf{R},$$
$$W_a\equiv W[\beta_1,\beta_2]=\beta_1\beta_2'-\beta_1'\beta_2=-2i(\sin\kappa)/\epsilon_2, \quad (2.6)$$

and $\xi=\hat{1}+O(\kappa^{-1})$ as $|\kappa|\uparrow\infty$ in $\mathbf{C}_\kappa^{\rm cut}$. Bloch solutions and ξ are κ analytic in $\mathbf{C}_\kappa^{\rm cut}$, each zero of ϵ_2 is in a finite-length gap, and there is precisely one zero in each gap of finite length.[19] Also, β and ξ have finite boundary values on $\mathscr{B}\mathscr{G}$, except for σ simple[20] poles where $\epsilon_2=0$.

The operator $-\partial_x^2+p_a$ is self-adjoint, so its L_∞ states are at real λ. Equation (2.5) and Fig. 1 show $\beta\in L_\infty$ when $\kappa\in\mathscr{B}$. In the interior of $\mathscr{G}$, $\operatorname{Im}\kappa>0$; so β_1 blows up and $\beta_2\to 0$ exponentially as $x\downarrow-\infty$, and β_2 blows up and $\beta_1\to 0$ exponentially as $x\uparrow+\infty$.

For each $\kappa\in\mathscr{B}$, Bloch waves in (2.5) represent exponential traveling waves modulated by periodic functions ξ. The wave β_1 travels toward $x=+\infty$ and β_2 travels toward $x=-\infty$.

The following lemmas are used later. Proofs are appended.

Lemma 2.1: At each λ for which $\epsilon_2(\lambda)=0$, $\exists n(\lambda)\in\mathbf{Z}\ni\kappa(\lambda)=n\pi+i\operatorname{Im}\kappa$ and $(-1)^n\partial_\lambda\epsilon_2>0$.

Lemma 2.2: $(\epsilon_2\operatorname{sign}\kappa)\sin\kappa>0$ in $\operatorname{Int}\mathscr{B}$.

III. STEP-PERIODIC POTENTIALS

This section is about the Schrödinger equation with step-periodic potential

$$(-\partial_x^2+p-\lambda)f=0,$$
$$p_b(x)\equiv p\equiv\lambda_0H(-x)+p_aH(x), \quad (3.1)$$

for any constant $\lambda_0\in\mathbf{R}$ and real valued, periodic p_a without impurity. This section has definitions and straightforward facts about spectrum. A theorem about bound states is proven. The section ends with statements on analyticity and asymptotics.

Eight solutions of (3.1) will be used later. They are modified regular solutions $Y=(Y_1,Y_2)^T$, modified exponential solutions $\hat{\psi}=(\hat{\psi}_1,\hat{\psi}_2)^T$, modified Bloch solutions $B=(B_1,B_2)^T$, and Jost solutions $\psi=(\psi_1,\psi_2)^T$:

$$\begin{aligned}
Y_1&\equiv \cos\varsigma x H(-x)+y_1H(x),\\
Y_2&\equiv\varsigma^{-1}\sin\varsigma x H(-x)+y_2H(x),\\
\hat{\psi}_1&\equiv e^{i\varsigma x}H(-x)+(y_1+i\varsigma y_2)H(x),\\
\hat{\psi}_2&\equiv e^{-i\varsigma x}H(-x)+(y_1-i\varsigma y_2)H(x),\\
B&\equiv(\hat{1}\cos\varsigma x+m\varsigma^{-1}\sin\varsigma x)H(-x)+\beta H(x),\\
\psi_1&\equiv B_1,\quad \psi_2\equiv e^{-i\varsigma x}H(-x)+(y_1-i\varsigma y_2)H(x).
\end{aligned} \quad (3.2)$$

Boundary values are $Y(x=0)=(1,0)^T$, $Y'(0)=(0,1)^T$, $\hat{\psi}(0)=\hat{1}$, $\hat{\psi}'(0)=(i\varsigma,-i\varsigma)^T$, $\psi(0)=\hat{1}$, and $\psi'(0)=(m_1,-i\varsigma)^T$. Jost solutions can be written as traveling waves

$$\begin{aligned}
\psi_1&=(T_1^0)^{-1}(e^{i\varsigma x}+R_1^0e^{-i\varsigma x})H(-x)+\beta_1(x)H(x),\\
\psi_2&=e^{-i\varsigma x}H(-x)+(T_2^0)^{-1}(R_2^0\beta_1+\beta_2)H(x),\\
T_1^0&\equiv -2i\varsigma/W,\quad T_2^0\equiv W_a/W,\\
R_1^0&\equiv(m_1-i\varsigma)/W,\quad R_2^0\equiv(m_2+i\varsigma)/W,\\
W&\equiv W_b\equiv W[\psi_1,\psi_2]=-(m_1+i\varsigma).
\end{aligned} \quad (3.3)$$

The Jost solution ψ_1 represents a wave incident from $x=-\infty$ that scatters from p; ψ_2 is incident from $x=+\infty$. The traveling-wave interpretation makes sense in some, but not all, of the spectrum.

The whole spectrum is described in Fig. 2 and Table I, whose contents are justified in the appendix. In the figure and table, subscripts on $\mathscr{B}$ and $\mathscr{G}$ denote multiplicity of the continuous spectrum. Also, an energy λ is said to be above threshold iff $\lambda>\lambda_0$, where λ_0 is defined in the first paragraph of Sec. II. The table shows, for example, that $\mathscr{B}_2$ and the bands above threshold are identical and have $\varsigma,\kappa\in\mathbf{R}$.

The motive for defining s [see (2.1)] can be understood now: This paper develops theorems used[8] in inverse scattering with the Marchenko–Newton equation. The equation comes from Fourier transformation of scattering data on the continuous spectrum $\mathscr{B}\cup\mathscr{G}_1$. The transform involves integration over a variable that should be real in $\mathscr{B}\cup\mathscr{G}_1$ when $\lambda_0<0$ and when $\lambda_0>0$. Variables κ, σ, and ς are not always

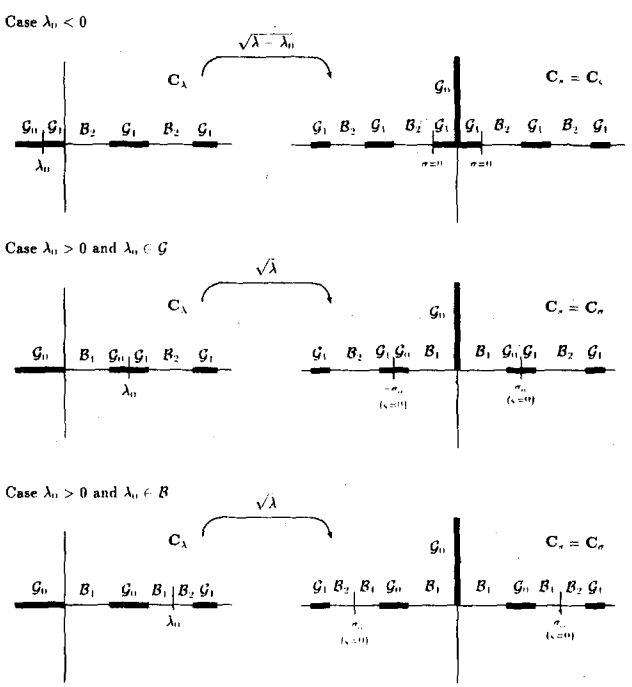


FIG. 2. The threshold λ_0 affects multiplicity. One consequence: Theorems 3.1 and 4.1 allow bound states in $\operatorname{Int}\mathscr{G}_0$ only. Hence, if $\lambda_0<0$ then no bound state is in $\mathbf{R}_x$; but if $\lambda_0>0$ then bound states in $\mathbf{R}_x$ are possible.

TABLE I. The spectrum of $-\partial_x^2+p$. The continuous spectrum is $\mathscr{B}_2\cup\mathscr{B}_1\cup\mathscr{G}_1$. The point spectrum is $\mathscr{G}_0\cap\{\lambda:W(\lambda)=0\}=\{\text{bound state energies}\}$. The threshold λ_0 is the value of $p(x)$ for $x<0$. An energy λ is below threshold iff $\lambda<\lambda_0$.

	$\mathscr{B}$ and ($\kappa\in\mathbf{R}$)	$\mathscr{G}$ ap ($\kappa=n\pi+i\operatorname{Im}\kappa$, $\operatorname{Im}\kappa>0$)
Above threshold ($\varsigma\in\mathbf{R}$)	$\mathscr{B}_2$	$\mathscr{G}_1$
Below threshold ($\varsigma=i\lvert\varsigma\rvert$)	$\mathscr{B}_1$	$\mathscr{G}_0$

real in $\mathscr{B} \cup \mathscr{G}_1$; s and λ are. I use s because sx is dimensionless. The Fourier transform with respect to s unites diverse types of scattering data.

The caption of Table I says every bound state energy is a zero of the Wronskian W of Jost solutions ψ_1 and ψ_2. The converse is true:

Theorem 3.1: The following are true e.p.[21] Equation (3.1) has a bound state at energy λ iff $W(\lambda) = 0$. The zeros of W in $\mathbf{C}_s^+ \cup \mathbf{R}_s$ are in Int $\mathscr{G}_0$, are finite in number, and are s simple.

Proof: The proof has three parts. The first part is about Int $\mathscr{G}_0$ and the equivalence of bound states and zeros of W.

If W were to have a zero in $\mathbf{C}_s^+ \setminus$ [imaginary axis] then (A2) would be a solution of (3.1). The solution would be in $L_2(-\infty, \infty)$ and $-\partial_x^2 + p$ would have a nonreal eigenvalue, contradicting self-adjointness. Therefore, zeros of W are in $\mathscr{B}\mathscr{G} \equiv \mathscr{B} \cup \mathscr{G}$.

In $\mathscr{B}_2$: If Im W, which equals $-\text{Im}[i\varsigma + (e^{i\kappa} - \epsilon_1)/\epsilon_2]$, vanishes then Table I implies $(\sin\kappa)/(\varsigma\epsilon_2) = -1$, which would contradict Lemma 2.2 because Figs. 1 and 2 show sign κ = sign ς in $\mathscr{B}_2$. So W has no zero in Int $\mathscr{B}_2$.

In $\mathscr{B}_1$, in $\mathscr{G}_1$, and at points $\kappa = n\pi$ that are above threshold, Table I shows that either m_1 or $i\varsigma$ is real and the other is not.[22] Therefore, the zeros of W are in the closure of $\mathscr{G}_0$.

At endpoints of $\mathscr{G}_0$-type gaps, either $\varsigma = 0$ or $\kappa = n\pi$: Use (2.3) to examine what $m_1 + i\varsigma = 0$ would imply in those two cases. If $\varsigma = 0$, then $W = 0$ and (2.3) imply $\epsilon_1(\varsigma = 0) = e^{i\kappa} = \exp\{i\cos^{-1}[(\epsilon_1 + \zeta_2)/2]\}$, which occurs at $\varsigma = 0$ only pathologically because the σ entire functions ϵ_1 and ζ_2, defined in (2.3), are independent of the threshold (λ_0) at which $\varsigma = 0$. If $\kappa = n\pi$, then $W = 0$ and (2.3) imply $\epsilon_1(\kappa = n\pi) = (-1)^n - |\varsigma|\epsilon_2$, which happens only pathologically. Thus, at *each* of the finitely many endpoints of $\mathscr{G}_0$ it is pathological if $W = 0$. It is pathological also if $W = 0$ at *any* of the finitely many endpoints. So the zeros of W in $\mathbf{C}_s^+ \cup \mathbf{R}_s$ are in Int $\mathscr{G}_0$, e.p.

Equivalence of bound state energies and zeros of W follows from the previous sentence, Table I, and sentences surrounding (A2).

The proof's first part is complete. The second part is about finiteness.

Functions κ and $e^{i\kappa}$ are s analytic in an open set containing the gap $(0, +i\infty)_s$. Equations (2.1) and (2.3) then show $W = -(m_1 + i\varsigma)$ is s analytic in an open set that contains $(0, +i\infty)_s$ but not necessarily $s = 0$. Thus, if there are infinitely many zeros of W in $(0, +i\infty)_s$ then they accumulate at $s = 0$ or at $s = +i\infty$. If they accumulate at $s = 0$, then $W(s = 0) = 0$ means either $W(\varsigma = 0) = 0$ (if $\lambda_0 < 0$) or $W(\kappa = 0) = 0$ (if $\lambda_0 > 0$): each case is shown three paragraphs above to be pathological. If zeros of $-m_1 + i\varsigma$ accumulate at $s = +i\infty$ then (3.4) is contradicted. [Equation (3.4) is written after the present proof, but the straightforward proof of (3.4) is independent of Theorem 3.1.] So there are only finitely many zeros of W in $(0, +i\infty)_s$, e.p.

Let f be $e^{i\kappa}$ restricted to $\mathbf{C}_s^+ \cup \mathbf{R}_s$. Then f is analytic in $\mathbf{C}_s^+$ because κ is. Each finite-length $\mathscr{G}_0$ gap is in $\mathbf{R}_s$. (See Fig. 2.) The function f is real on each of those gaps because Re $\kappa = n\pi$ on all gaps. Schwarz's reflection principle shows f has an analytic continuation across the gaps[23] so W does too. There are only finitely many gaps in $\mathscr{G}_0$, so if W has infinitely many zeros then they accumulate at a gap's endpoint and $W = 0$ at that endpoint, which happens only pathologically. So W has finitely many zeros, e.p.

The proof's second part is complete. It remains to show that the zeros of W are s simple. Simpleness follows from an argument that parallels Ref. 24, with details in Ref. 8. ■

Statements on analyticity and asymptotics complete the section.

It is noted after (2.6) that β is κ analytic in $\mathbf{C}_\kappa^{\text{cut}}$. To show β is s analytic in $\mathbf{C}_s^+$ one need only note the preimage of $\mathbf{C}_s^+$ is in $\mathbf{C}_\kappa^{\text{cut}}$ and prove $\exists \partial_s \kappa$ in the preimage. Use (2.3) to evaluate $\partial_\sigma \cos\kappa$ and obtain $\partial_\sigma \kappa = -(\sin\kappa)^{-1}\partial_\sigma[(\epsilon_1 + \zeta_2)/2]$, which exists in $\mathbf{C}_\kappa^{\text{cut}}$ because ϵ and ζ are σ entire. So $\exists \partial_s \kappa$ when $\lambda_0 < 0$. For $\lambda_0 > 0$ note $\varsigma^2 = \sigma^2 - \lambda_0$ implies $\partial_\varsigma \sigma = \varsigma/\sigma$, which exists in $\mathbf{C}_s^+$ because $\mathbf{R}^+$ is the branch cut of $\sqrt{\ }$. (See Fig. 2.) Therefore β is at worst s meromorphic in $\mathbf{C}_s^+$. Equations (2.3) and (2.4) show that the only possible poles of β occur when $\epsilon_2 = 0$. Lines following (2.6) imply the only possible poles of β are in $\mathscr{G} \subset \mathbf{R}_s$. So β is s analytic in $\mathbf{C}_s^+$. Equation (3.2) shows that B, ψ, and Y are s analytic in $\mathbf{C}_s^+$ and that $\epsilon_2\psi_1$, ψ_2, $\epsilon_2 B_1$, B_2, and Y are s continuous (without poles) on $\mathbf{C}_s^+ \cup \mathbf{R}_s$.

Reference 15 has statements about asymptotics in terms of κ and σ. It is easy to introduce $\varsigma = \sqrt{\sigma^2 - \lambda_0}$ into the statements and thereby introduce s. As $|s| \uparrow \infty$ in $\mathbf{C}_s^+ \cup \mathbf{R}_s$,

$$\kappa = s + O(s^{-1}) = \sigma + O(\sigma^{-1}). \tag{3.4a}$$

As $|s| \uparrow \infty$ with Im $s = \tau \geqslant 0$ fixed:

$$e^{\pm i\varsigma x}, \quad e^{\pm i\sigma x}, \quad e^{\pm i\kappa x} = e^{\pm isx} + O(s^{-1}e^{\mp \tau x}),$$

$$y_1 = \cos sx + O(s^{-1}e^{|\tau x|}), \qquad \epsilon_1 = \cos s + O(s^{-1}e^{\tau}),$$

$$y_2 = s^{-1}\sin sx + O(s^{-2}e^{|\tau x|}), \qquad \epsilon_2 = s^{-1}\sin s + O(s^{-2}e^{\tau}),$$

$$y_1' = -s\sin sx + O(e^{|\tau x|}), \qquad \zeta_1 = -s\sin s + O(e^{\tau}),$$

$$y_2' = \cos sx + O(s^{-1}e^{|\tau x|}), \qquad \zeta_2 = \cos s + O(s^{-1}e^{\tau}),$$

$$\epsilon_2 m_1 = e^{i\kappa} - \epsilon_1 = i\sin s + O(s^{-1}e^{\tau}). \tag{3.4b}$$

Reference 25 has a result for s asymptotics of ψ which will be generalized to account for poles of β on the unbounded sequence of energies at which $\epsilon_2 = 0$. Equations (2.3) and (2.4) imply $\epsilon_2\beta = \epsilon_2 y_1 \hat{1} + [X(\kappa, x = 1) - \epsilon_1]y_2\hat{1}$, where $X(\kappa, x) \equiv \text{diag}\{e^{i\kappa x}, e^{-i\kappa x}\}$. Use (3.4) to show $\epsilon_2\beta = [1 + O(s^{-1})]s^{-1}(\sin s)X(s,x)\hat{1}$ on $\mathbf{R}_x$ as $|s| \uparrow \infty$ in $\mathbf{C}_s^+ \cup \mathbf{R}_s$, where $X(s,x) = \text{diag}\{e^{isx}, e^{-isx}\}$. Similarly,

$$\epsilon_2\psi_1 = [1 + O(s^{-1})]s^{-1}(\sin s)e^{isx},$$

$$\psi_2 = [1 + O(s^{-1})]e^{-isx} \tag{3.5}$$

on $\mathbf{R}_x$ as $|s| \uparrow \infty$ in $\mathbf{C}_s^+ \cup \mathbf{R}_s$. Equations (3.2) and (3.4) imply $\psi_2' = -ise^{-isx}[1 + O(s^{-1})]$. Then $-\epsilon_2(m_1 + i\varsigma) = \epsilon_2 W = \epsilon_2\psi_1\psi_2' - \epsilon_2\psi_1'\psi_2$ and (3.4) and (3.5) yield

$$\epsilon_2\psi_1' = i(\sin s)e^{isx}[1 + O(s^{-1})],$$

$$\psi_2' = -ise^{-isx}[1 + O(s^{-1})] \tag{3.6}$$

on $\mathbf{R}_x$ as $|s| \uparrow \infty$ in $\mathbf{C}_s^+ \cup \mathbf{R}_s$.

IV. STEP-PERIODIC POTENTIALS WITH IMPURITY

A. Jost solutions

Lemma 4.1 defines Jost solutions Ψ, Lemma 4.2 gives s asymptotics of Ψ, and a discussion of x asymptotics follows. The lemmas' proofs are appended.

Lemma 4.1: Let $(1+|x|)q\in L_1(\mathbf{R}_x)$, $q\in L_2(\mathbf{R}_x)$, and $g(x',x)\equiv[\psi_1(x')\psi_2(x)-\psi_1(x)\psi_2(x')]/W$. Then $\exists!\Psi=(\Psi_1,\Psi_2)^T\ni$

$$\Psi_1=\psi_1-\int_x^\infty dx'\,g(x',x)q(x')\Psi_1(x'),$$
$$\Psi_2=\psi_2-\int_{-\infty}^x dx'\,g(x,x')q(x')\Psi_2(x') \tag{4.1}$$

in $\mathbf{C}_s^+\cup\mathbf{R}_s\times\mathbf{R}_x$. Moreover, $\epsilon_2\Psi_1$ and Ψ_2 are s continuous (without poles) on $\mathbf{C}_s^+\cup\mathbf{R}_s\times\mathbf{R}_x$, are s analytic on $\mathbf{C}_s^+\times\mathbf{R}_x$, and are solutions of the full Schrödinger equation (1.1).

Lemma 4.2: We have $\Psi=[1+O(s^{-1})]\psi$ and $\Psi'=[1+O(s^{-1})]\psi'$ on $\mathbf{R}_x$ as $|s|\uparrow\infty$ in $\mathbf{C}_s^+\cup\mathbf{R}_s$.

Spacial asymptotics

$$\begin{aligned}
&\Psi_1\to(1/T_1)e^{i\varsigma x}+(R_1/T_1)e^{-i\varsigma x} && \text{as } x\downarrow-\infty,\\
&\Psi_1\to\beta_1 && \text{as } x\uparrow+\infty,\\
&\Psi_2\to e^{-i\varsigma x} && \text{as } x\downarrow-\infty,\\
&\Psi_2\to(R_2/T_2)\beta_1+(1/T_2)\beta_2 && \text{as } x\uparrow+\infty
\end{aligned} \tag{4.2}$$

define transmission (T_i) and reflection (R_i) coefficients wherever (4.2) makes sense. Asymptotics (4.2) do not define R_1/T_1 below threshold (where $\varsigma=i|\varsigma|$) because $e^{-i\varsigma x}$ vanishes as $x\downarrow-\infty$ and $e^{i\varsigma x}$ does not. However, $1/T_1$ is defined below threshold as well as above. Similarly, R_2/T_2 is not defined in gaps but $1/T_2$ is defined on $\mathscr{B}\mathscr{G}$. Thus, T_i and R_i are defined on $\mathscr{B}\mathscr{G}$, except R_1 is not defined in $\mathscr{B}_1$, R_2 is not defined in $\mathscr{G}_1$, and R_1 and R_2 are not defined in $\mathscr{G}_0$.

Asymptotics (4.2) show $-\partial_x^2+p+q$ has the same spectral structure as $-\partial_x^2+p$. That is, $-\partial_x^2+p+q$ has eigenvalues in $\mathscr{G}_0$ only, has $\mathscr{G}_1\cup\mathscr{B}_1$ as its continuous spectrum of multiplicity one, and has $\mathscr{B}_2$ as its continuous spectrum of multiplicity two. The proof follows from (4.2) in the same way that Table I follows from (3.3).

Evaluate x independent Wronskians $W[f,g]=fg'-f'g$ and currents $j[f]\equiv f^*f'-ff^{*\prime}$ to get relations among T_i and R_i. Then use (4.2) to evaluate

$$W[\Psi_1(x\downarrow-\infty),\Psi_2(x\downarrow-\infty)]=W[\Psi_1(x\uparrow+\infty),\Psi_2(x\uparrow+\infty)]$$

and obtain

$$W_c\equiv W[\Psi_1,\Psi_2]=-2i\varsigma/T_1=W_a/T_2,$$
$$W_a=-2i(\sin\kappa)/\epsilon_2. \tag{4.3}$$

Similar evaluations of $W[\Psi_1,\Psi_2^*]$, $j[\Psi_1]$, and $j[\Psi_2]$ show that

$$-2i\varsigma T_2=W_aT_1,\quad -2i\varsigma T_2^*R_1+W_aT_1R_2^*=0,$$
$$[W_a/(-2i\varsigma)]|T_1|^2+|R_1|^2=1=(-2i\varsigma/W_a)|T_2|^2+|R_2|^2 \tag{4.4}$$

in $\mathscr{B}_2$;

$$|R_2|=1,\quad R_2=T_2/T_2^*=T_1/T_1^*; \tag{4.5}$$

in $\mathscr{B}_1$; and

$$|R_1|=1,\quad R_1=T_1/T_1^*=-T_2/T_2^* \tag{4.6}$$

in $\mathscr{G}_1$. Equations (4.4)–(4.6) are analogs of S matrix unitarity in the $p\equiv0$ case.

Equations (3.3) and (4.3) relate transmission coefficients for p and $p+q$:

$$\frac{T_1}{T_1^0}=\frac{T_2}{T_2^0}=\frac{W}{W_c},\quad T_1=-\frac{2i\varsigma}{W_c},\quad T_2=\frac{W_a}{W_c}. \tag{4.7}$$

Equation (4.1) implies

$$\Psi_1\to\psi_1\left[1+W^{-1}\int_{-\infty}^\infty dx'\,\psi_2q\Psi_1\right]-\psi_2W^{-1}\int_{-\infty}^\infty dx'\,\psi_1q\Psi_1,$$

as $x\downarrow-\infty$. Then

$$\frac{W_c}{W}=1+W^{-1}\int_{-\infty}^\infty dx'\,\psi_1q\Psi_2=1+W^{-1}\int_{-\infty}^\infty dx'\,\psi_2q\Psi_1 \tag{4.8}$$

is obtained by substituting (3.3), equating coefficients with (4.2), and using (4.7) to get the first equality. The second equality comes from similar treatment of $\Psi_2(x\uparrow+\infty)$.

Collecting scattering data in a matrix that has a Fourier transform is this paper's main goal. The matrix must have good s asymptotics if its transform is to exist. A straightforward collection of T_i and R_i, as is used in the seminal $p\equiv0$ case, does not have good asymptotics, as (4.4)–(4.6) and the patchwork state of definitions [(4.2)ff.] show. The ratio W_c/W of Wronskians, which is related to scattering coefficients by (4.7), simplifies s asymptotics greatly. The ratio has a role, as crucial as that of s, in unifying diverse data.

Theorem 4.1: The following are true e.p. The quotient W_c/W is s meromorphic in $\mathbf{C}_s^+$ and s continuous (with poles) on $\mathbf{C}_s^+\cup\mathbf{R}_s$. The poles in $\mathbf{C}_s^+\cup\mathbf{R}_s$ of W_c/W are the zeros of W and are s simple. The zeros in $\mathbf{C}_s^+\cup\mathbf{R}_s$ of W_c are in Int $\mathscr{G}_0$, are finite in number, and are s simple. Equation (1.1) has a bound state of energy λ iff $W_c(\lambda)=0$. Also $W_c/W=1+O(s^{-1})$ as $|s|\uparrow\infty$ in $\mathbf{C}_s^+\cup\mathbf{R}_s$.

Proof: The proof has four parts. The first is about analyticity, continuity, and poles.[26]

Meromorphism and continuity follow from (4.8) and Lemma 4.1. The quotient W_c/W has a pole at each zero of W, except in the pathological case that one of the finitely many zeros of the q-dependent $W_c=W[\Psi_1,\Psi_2]$ coincides with one of the finitely many zeros of q-independent $W=W[\psi_1,\psi_2]$.[27] That W_c/W has no other pole[28] follows from (4.3) and Lemma 4.1. Theorem 3.1 shows poles of W_c/W are s simple.

The proof's first part is finished. The second part is about Int $\mathscr{G}_0$ and bound states.

If W_c were to have a zero in $\mathbf{C}_s^+\setminus\mathscr{B}\mathscr{G}$ then $\Psi_1\propto\Psi_2$ at an energy $\lambda\notin\mathbf{R}$. Asymptotics $\Psi_1(+\infty)$ and $\Psi_2(-\infty)$ from (4.2) would then imply $\Psi\in L_2(\mathbf{R}_x)$, contradicting self-adjointness. So zeros in $\mathbf{C}_s^+\cup\mathbf{R}_s$ of W_c are in $\mathscr{B}\mathscr{G}$.

Equations (4.7) and (3.3) would contradict (4.4) at a hypothetical zero of W_c in $\mathscr{B}_2$. At a hypothetical zero in $\mathscr{B}_1$: $\Psi_1 \propto \Psi_2$, so $+\infty$ asymptotics in (4.2) would imply first that $1/T_2 = 0$, then $R_2 = 0$, contradicting (4.5). At a hypothetical zero in $\mathscr{G}_1$: $\Psi_1 \propto \Psi_2$, so $-\infty$ asymptotics in (4.2) would imply first that $1/T_1 = 0$, then $R_1 = 0$, contradicting (4.6). It is pathological if any zero of q-dependent W_c coincides with any endpoint of the finitely many gaps in $\mathscr{G}_0$.[29] So the zeros of W_c are in Int $\mathscr{G}_0$, e.p.

Self-adjointness restricts bound state energies to $\mathscr{B}\mathscr{G}$ and (4.2) yields a tighter restriction to $\mathscr{G}_0$. If a bound state energy were in $\mathscr{G}_0$ at a point where $W_c \neq 0$ then both $1/T_1$ and $1/T_2$ would be nonzero [by (4.3), e.p.] and Ψ_1 would not be proportional to Ψ_2. Therefore, the bound state would be a nonzero linear combination of Ψ_1 and Ψ_2 with no Ψ_1 component, because of $-\infty$ asymptotics in (4.2) and no Ψ_2 component, because of $+\infty$ asymptotics in (4.2)—a clear contradiction. So bound state $\Rightarrow W_c = 0$. Conversely, at a zero of W_c: $\Psi_1 \propto \Psi_2$ and $s \in \mathscr{G}_0$, so x asymptotics in (4.2) imply $\Psi_1 \in L_2(\mathbf{R}_x)$ is a bound state. Therefore, bound state $\Leftrightarrow W_c = 0$.

The proof's second part is finished. The third part is about asymptotics and finiteness.

Lemma 4.2 and (4.3) yield $W_c/W = 1 + O(s^{-1})$. That statement and the fact that the (finitely many) zeros of W do not accumulate at $s = \infty$ imply zeros of W_c also do not accumulate at $s = \infty$.

Table I, (2.3), and (3.2) imply ψ: $\mathscr{G}_0 \to \mathbf{R}$ and W: $\mathscr{G}_0 \to \mathbf{R}$. Therefore, (4.1) implies Ψ: $\mathscr{G}_0 \to \mathbf{R}$ and (4.3) implies W_c: $\mathscr{G}_0 \to \mathbf{R}$. So W_c is real valued on Int $\mathscr{G}_0$, which is finitely many segments in $\mathbf{C}_s^+ \cup \mathbf{R}_s$. Schwarz's reflection principle shows W_c has an s analytic continuation across each segment of Int $\mathscr{G}_0$. It follows that zeros of W_c can accumulate only at the finitely many endpoints of Int $\mathscr{G}_0$, excluding $s = +i\infty$.

It has just been established that there are, at most, countably many zeros of W_c—a fact which completes the proof in the previous part that no zero of W_c is an endpoint of $\mathscr{G}_0$, e.p. Continuity implies zeros cannot accumulate at endpoints, e.p. So zeros in $\mathbf{C}_s^+ \cup \mathbf{R}_s$ of W_c accumulate nowhere and are finite in number, e.p. Such finiteness was used in the proof's first part, now complete.

The proof's first three parts are complete. It remains to show zeros of W_c are s simple.

Apply the technique in Ref. 24 to $(-\partial_x^2 + p + q - \lambda)\Psi = 0$ and $W_c = W[\Psi_1, \Psi_2]$ to show $\exists a \neq 0 \ni -\partial_s W_c(W_c = 0) = 2sa\int_{-\infty}^{\infty} dx'\, \Psi_1^2$. It was just shown that $W_c = 0 \Rightarrow s \neq 0$ e.p. [because $s = 0$ is an endpoint of the gap $(0, +i\infty)_s$] and that $\Psi(W_c = 0)$ is real valued. So $\partial_s W_c(W_c = 0) \neq 0$ e.p. ∎

B. Scattering solutions

Scattering solutions $\Phi = (\Phi_1, \Phi_2)^T$ are defined with $g^+(x,x') \equiv W^{-1}\psi_1[\max(x,x')]\psi_2[\min(x,x')]$ as

$$\Phi = \psi - \int_{-\infty}^{\infty} dx'\, g^+(x,x') q \Phi$$
$$= \psi - W^{-1}\psi_1 \int_{-\infty}^{x} dx'\, \psi_2 q \Phi - W^{-1}\psi_2 \int_{x}^{+\infty} dx'\, \psi_1 q \Phi. \tag{4.9}$$

The Fredholm method will show $\exists!$ solution Φ to (4.9). It will be shown also that Φ solves the full Schrödinger equation (1.1).

The Fredholm method applies to (4.9) only at those s for which $\int_{-\infty}^{\infty} dx |-g^+(x,x)q(x)| < \infty$. Equation (3.3) shows

$$|-g^+(x,x)|$$
$$= |[m_1 + i\varsigma + (i\varsigma - m_1)e^{-2i\varsigma x}]/(2\varsigma W)|H(-x)$$
$$+ |\xi_1[W\xi_2 + (m_2 + i\varsigma)\xi_1 e^{2i\kappa x}]/(W_a W)|H(x).$$

Therefore, if $(1+x)q \in L_1(\mathbf{R}_x)$ then the Fredholm method applies in $\mathbf{C}_s^+ \cup \mathbf{R}_s$, except[30] possibly at s poles of $g^+(x,x)$. Now $W_a = -2i(\sin\kappa)/\epsilon_2$, $W = -(m_1 + i\varsigma)$, $m_j = (e^{\pm i\kappa} - \epsilon_1)/\epsilon_2$, and the equation for $|-g^+(x,x)|$ show $g^+(x,x)$ can have poles only at energies for which $\varsigma = 0$, $\kappa = n\pi$, $\epsilon_2 = 0$, or $W = 0$.[31] Equation (3.2) and the definition of g^+ show that poles of g^+ at $\epsilon_2 = 0$, $\varsigma = 0$, and $\kappa = n\pi$ are removable and that g^+ has a nonremovable pole wherever $W = 0$. So the Fredholm method applies to (4.9) in $\mathbf{C}_s^+ \cup \mathbf{R}_s$ except at zeros of W, which are described in Theorem 3.1.

The Fredholm determinant of the operator $[1 + (-g^+ q)]$ in (4.9) is computed by reproducing line-by-line Appendix A of Ref. 32. The result is

$$\det_{\text{Fredholm}}[1 + (-g^+ q)] = \frac{T_1^0}{T_1} = \frac{T_2^0}{T_2} = \frac{W_c}{W}. \tag{4.10}$$

So (4.9) defines Φ in $\mathbf{C}_s^+ \cup \mathbf{R}_s$ except possibly at the (at most) finitely many points in $\mathscr{G}_0$ where $W = 0$. Also, Φ is s meromorphic in $\mathbf{C}_s^+$ with poles at the zeros of W_c.

Definition (4.9) shows that Φ solves the full Schrödinger equation (1.1). It follows that Φ_1 is a linear combination of Ψ_1 and Ψ_2, and Φ_2 is too. The particular linear combinations will be determined.

Take limits as $x \to \pm\infty$ of (4.9) to get

$$\Phi_1 \overset{x\downarrow -\infty}{\to} \psi_1 - W^{-1}\psi_2 \int_{-\infty}^{\infty} dx'\, \psi_1 q \Phi_1,$$
$$\Phi_1 \overset{x\uparrow +\infty}{\to} \psi_1 \left(1 - W^{-1}\int_{-\infty}^{\infty} dx'\, \psi_2 q \Phi_1\right),$$
$$\Phi_2 \overset{x\downarrow -\infty}{\to} \psi_2 \left(1 - W^{-1}\int_{-\infty}^{\infty} dx'\, \psi_1 q \Phi_2\right),$$
$$\Phi_2 \overset{x\uparrow +\infty}{\to} \psi_2 - W^{-1}\psi_1 \int_{-\infty}^{\infty} dx'\, \psi_2 q \Phi_2.$$

The $+\infty$ limits of Φ_1 and Ψ_1 [see (4.2)], the $-\infty$ limits of Φ_2 and Ψ_2, and (3.2) imply $\Phi_1 \propto \Psi_1$ and $\Phi_2 \propto \Psi_2$, with x independent constants of proportionality. Use (3.3) to match $e^{i\varsigma x}$ coefficients of Φ_1 and Ψ_1 as $x\downarrow -\infty$ and to match β_2 coefficients of Φ_2 and Ψ_2 as $x\uparrow +\infty$. The result is $T_1^0\Phi_1 = T_1\Psi_1$ and $T_2^0\Phi_2 = T_2\Psi_2$. It follows from (4.10) that

$$\Phi = W\Psi/W_c. \tag{4.11}$$

Define a 2×2 matrix S by

$$\Phi^* = S\Phi.$$

Use (4.10), (4.11), (4.2), and what is known about β^* and

ς^* in each part of the spectrum to find $S(s)$ in terms of T_i, R_i, T_i^0, R_i^0, and their complex conjugates. Use (3.3) to show $T_1^0/T_1^{0*} = -T_2^0/T_2^{0*} = R_1^0$ in $\mathscr{G}_1$, $T_1^0/T_1^{0*} = -T_2^0/T_2^{0*} = -R_2^0$ in $\mathscr{B}_1$, and $T_1^0/T_1^{0*} = T_2^0/T_2^{0*} = -W^*/W$ in $\mathscr{B}_2$. It follows that

$$S(s) = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & s \in \mathscr{G}_0 \\ R_1^0 \begin{pmatrix} R_1 & 0 \\ -T_1 & -1 \end{pmatrix}^*, & s \in \mathscr{G}_1 \\ R_2^0 \begin{pmatrix} -1 & -T_2 \\ 0 & R_2 \end{pmatrix}^*, & s \in \mathscr{B}_1 \\ \dfrac{T_1^0}{T_1^{0*}} \begin{pmatrix} R_1 & T_2 \\ T_1 & R_2 \end{pmatrix}^*, & s \in \mathscr{B}_2 \end{cases} . \tag{4.12}$$

The matrix S is called a functional[12] S matrix.

C. Modified solutions

This paper develops results needed for deriving a Marchenko–Newton equation, which involves Fourier transforms of scattering coefficients. It would be convenient if S had a Fourier transform. However, $S_{12} = 0$ in $\mathscr{G}_1$ so $S_{12} \to 0$ as $s \to \pm\infty$ from within $\mathscr{G}_1$, but it can be shown that S_{12} has different s asymptotics from within $\mathscr{B}_2$. So S does not have a Fourier transform; and worse, there is no s independent matrix S_0 such that $S - S_0$ has a Fourier transform. The matrix S is an inconvenient collection of data.

A matrix $\hat{S}$, which has an s limit, will be defined. The paper ends with a proof that $\hat{S}(s) - \mathbf{I}$ has a Fourier transform in the L_2 sense.

Functions ψ and $\hat{\psi}$ are solutions of the same Schrödinger equation, so

$$\psi = \hat{M}\hat{\psi}$$

defines the 2×2 matrix $\hat{M}$ on $\mathscr{B}\,\mathscr{G}$. Use (3.2) to compare boundary values of ψ and $\hat{\psi}$ at $x = 0$ and obtain

$$\hat{M} = (2i\varsigma)^{-1} \begin{pmatrix} i\varsigma + m_1 & i\varsigma - m_1 \\ 0 & 2i\varsigma \end{pmatrix},$$

$$\hat{M}^{-1} = (m_1 + i\varsigma)^{-1} \begin{pmatrix} 2i\varsigma & m_1 - i\varsigma \\ 0 & m_1 + i\varsigma \end{pmatrix}. \tag{4.13}$$

Equation (3.3) yields $\det \hat{M} = (m_1 + i\varsigma)/(2i\varsigma) = W/(-2i\varsigma)$. Note that $\hat{M}^{-1}$ has an s meromorphic continuation into $\mathbf{C}_s^+ \cup \mathbf{R}_s$.

Define

$$\hat{\Phi} \equiv \hat{M}^{-1}\Phi \tag{4.14}$$

on $\mathbf{C}_s^+ \cup \mathbf{R}_s \times \mathbf{R}_x$ and see from (4.9) that $\hat{\Phi} = \hat{\psi} - \int_{-\infty}^{\infty} dx'\, g^+(x,x') q \hat{\Phi}$. Wronskian algebra (Ref. 33, p. 2157) (see Appendix) and (4.11) and (4.14) imply

$$W[\hat{\Phi}_1, \hat{\Phi}_2] = -2i\varsigma W/W_c. \tag{4.15}$$

Define a new functional S matrix on $\mathbf{R}_s$ by

$$\hat{\Phi}^* = \hat{S} Q \hat{\Phi},$$

where $Q = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Equation (4.14) implies

$$\hat{S}Q\hat{\Phi} = \hat{\Phi}^* = \hat{M}^{*-1}\Phi^* = \hat{M}^{*-1}S\Phi = \hat{M}^{*-1}S\hat{M}\hat{\Phi}$$

on $\mathbf{R}_s \times \mathbf{R}_x$. So

$$\hat{S} = \hat{M}^{*-1} S \hat{M} Q \tag{4.16}$$

on $\mathbf{R}_s$.

The vector $v \equiv (Y_1 + isY_2,\ Y_1 - isY_2)^T$—whose boundary values are $v(0) = \hat{1}$ and $v'(0) = is(1,-1)^T$—is s analytic on $\mathbf{C}_s^+ \times \mathbf{R}_x$ and s continuous (without poles) on $\mathbf{C}_s^+ \cup \mathbf{R}_s \times \mathbf{R}_x$ because Y is. The Volterra equation $\Upsilon(x \geqslant 0) = v - \int_0^x dx'\, g(x,x') q \Upsilon$ has a kernel gq which is λ entire, as shown by (A8). The Fredholm method and s analyticity of v imply $\exists! \Upsilon(x \geqslant 0)$ that solves the integral equation, which Υ has the same s analytic and s continuous properties as v. Similarly, $\exists! \Upsilon$ that solves

$$\Upsilon(x) = v - H(-x) \int_x^0 dx'\, g(x',x) q \Upsilon - H(x) \int_0^x dx'\, g(x,x') q \Upsilon$$

on $\mathbf{C}_s^+ \cup \mathbf{R}_s \times \mathbf{R}_x$. The regular solution Υ is s analytic on $\mathbf{C}_s^+ \times \mathbf{R}_x$, is s continuous (without poles) on $\mathbf{C}_s^+ \cup \mathbf{R}_s \times \mathbf{R}_x$, and solves the Schrödinger equation (1.1). The regular solution satisfies $\Upsilon(0) = \hat{1}$, $\Upsilon'(0) = is(1,-1)^T$, and $W[\Upsilon_1, \Upsilon_2] = -2is$—so that $\Upsilon_1^* = \Upsilon_2$ and $\Upsilon^* = Q\Upsilon$ on $\mathbf{R}_s \times \mathbf{R}_x$.

Define a Jost matrix $\hat{J}$ by $\Upsilon = \hat{J}\hat{\Phi}$. The equation is one of three that are the basis for Newton's inverse scattering method:

$$\begin{aligned} \hat{\Phi}^* &= \hat{S} Q \hat{\Phi} \quad \text{on} \quad \mathscr{B}\,\mathscr{G} \times \mathbf{R}_x, \\ \Upsilon &= \hat{J}\hat{\Phi} \quad \text{on} \quad \mathbf{C}_s^+ \cup \mathbf{R}_s \times \mathbf{R}_x, \\ \Upsilon^* &= Q\Upsilon \quad \text{on} \quad \mathbf{R}_s \times \mathbf{R}_x. \end{aligned} \tag{4.17}$$

Wronskian algebra (see the Appendix) and relations among Υ, $\hat{\Phi}$, Φ, and Ψ yield $\hat{J} = -W[\Upsilon, \Psi] P \hat{M}/W = W_c W[\Upsilon, \hat{\Phi}] P/(2i\varsigma W)$, where $P = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Use (4.15) to show[34]

$$\det \hat{J} = s W_c/(\varsigma W), \tag{4.18}$$

$$\begin{aligned} \hat{J} &= W^{-1} \begin{pmatrix} [\Psi_2' - is\Psi_2] & [is\Psi_1 - \Psi_1'] \\ [\Psi_2' + is\Psi_2] & -[is\Psi_1 + \Psi_1'] \end{pmatrix}_{x=0} \cdot \hat{M} \\ &= W_c (2i\varsigma W)^{-1} \begin{pmatrix} [\hat{\Phi}_2' - is\hat{\Phi}_2] & [is\hat{\Phi}_1 - \hat{\Phi}_1'] \\ [\hat{\Phi}_2' + is\hat{\Phi}_2] & -[is\hat{\Phi}_1 + \hat{\Phi}_1'] \end{pmatrix}_{x=0}. \end{aligned} \tag{4.19}$$

Some terminology for the following theorem: A vector or matrix has a pole iff *any* of its entries has a pole and it has a zero iff *each* of its entries is zero. An x dependent vector has an s zero (s pole) iff the vector is zero (∞) for almost every x.

Proofs of Theorems 4.2 and 4.3 are appended.

Theorem 4.2: The following are true e.p. Functions in Table II have poles and zeros only where indicated and are otherwise s analytic on $\mathbf{C}_s^+$ and s continuous on $\mathbf{C}_s^+ \cup \mathbf{R}_s$. The matrix $\hat{S}$ is s continuous (without poles) on $\mathbf{R}_s$. All tabulated poles and zeros are s simple.

Theorem 4.3: As $|s| \uparrow \infty$ in $\mathbf{C}_s^+ \cup \mathbf{R}_s$: $\hat{\Phi} = [1 + O(s^{-1})] X(s,x) \hat{1}$ and $\hat{\Phi}' = [1 + O(s^{-1})] \times X'(s,x) \hat{1}$ on $\mathbf{R}_x^-$, and $\hat{J} = \mathbf{I} + O(s^{-1})$. Also, $\hat{S} = \mathbf{I} + O(s^{-1})$ as $s \to \pm\infty$ in $\mathbf{R}_s$.

This paper's main goal is to find a collection of scatter-

ing coefficients that has a Fourier transform. Theorems 4.2 and 4.3 show $\hat{S}-\mathbf{I}$ has a Fourier transform in the L_2 sense. The matrix $\hat{S}$ is related to scattering data by $\hat{S}=\hat{M}^{*-1}S\hat{M}Q$, with definitions in (4.12), (4.2), and (4.13). Also obtained, for step-periodic potentials with impurity, are matrix equations (4.17) and asymptotic and analytic properties (Theorems 4.2 and 4.3) used in Marchenko–Newton inverse scattering.

V. CONCLUSION

The full Schrödinger operator $-\partial_x^2+p+q$ for a localized impurity q in a step-periodic background p has a richly structured spectrum sketched in Figs. 1 and 2. Scattering coefficients R_i and T_i are defined in (4.2) in terms of asymptotics of scattering solutions. Coefficients are collected in a functional[12] S matrix called $\hat{S}$, defined by (4.16), (4.13), and(4.12). The matrix $\hat{S}$ unifies scattering data in the sense that $\hat{S}(s)-\mathbf{I}$ has a Fourier transform, which is used[8] in inverse scattering.

There are two key steps in unifying scattering data. First is definition in (2.1) of a variable s that is real valued on the continuous spectrum. Second is identification in Theorem 4.1 of an s meromorphic function—a ratio of Wronskians—that is related to scattering data and has simple s asymptotics.

This paper defines other functions—scattering solutions, regular solutions, and a Jost matrix—also needed for Marchenko–Newton inverse scattering. The basic matrix equations are $\hat{\Phi}^*=\hat{S}Q\hat{\Phi}$, $\Upsilon=\hat{J}\hat{\Phi}$, and $\Upsilon^*=Q\Upsilon$, from which follow

$$(\hat{J}^{-1}-\mathbf{I})=(\hat{S}^*-\mathbf{I})+Q(\hat{J}^{-1}-\mathbf{I})^*Q + (\hat{S}^*-\mathbf{I})Q(\hat{J}^{-1}-\mathbf{I})^*Q.$$

Fourier transformation of the latter equation leads to a Marchenko–Newton equation for inverse scattering. Derivation of the Marchenko–Newton equation uses[8] analytic and asymptotic properties established in this paper's theorems and in Table II.

ACKNOWLEDGMENTS

Roger G. Newton was advisor for the thesis from which this paper is adapted. He devoted 200 contact hours to supervising my readings and research. I had useful discussions with Sung H. Yoon. The thesis was written at Indiana University and was adapted for publication at Ames Laboratory.

Research was supported for 1 year by an unsolicited fellowship from Conoco. Adaptation was supported fully by the Applied Mathematical Sciences subprogram of the Office of Energy Research, United States Department of Energy under contract W-7405-ENG-82 and by the Office of Naval Research under contract N0014-83-K-0038.

APPENDIX: PROOFS AND DISCUSSION

Wronskian algebra:[33] Let f and g be vector solutions of the same Schrödinger equation and A and B be x independent 2×2 matrices. Let $W[f,g]$ be the 2×2 matrix whose ij component is $W[f_i,g_j]$. Then $W[Af,Bg]=AW[f,g]B^T$ and $W[f,f]=W[f_1,f_2]P$, where $P\equiv\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)$. Also, $f=Ag\Rightarrow W[f_1,f_2]=W[g_1,g_2]\det A$ and $A=-W[f,g]P/W[g_1,g_2]$.

Proof of Lemma 2.1: Levitan and Sargsjan[19] showed

$$\partial_\lambda y_2(\lambda,x)=\int_0^x dx'\,y_2(\lambda,x')[\,y_1(\lambda,x)y_2(\lambda,x')-y_1(\lambda,x')y_2(\lambda,x)\,].$$

Set $x=1$, use the definition (2.3) of ϵ, and set $\lambda\ni\epsilon_2(\lambda)=0$ to show $\partial_\lambda\epsilon_2=\epsilon_1\int_0^1dx'[\,y_2(x')\,]^2$ when $\epsilon_2=0$.

Levitan and Sargsjan[19] proved the zeros in $\mathbf{C}_\lambda$ of ϵ_2 are in $\mathcal{G}$ (and that there is precisely one zero in each finite-length gap), so $\sigma\in\mathbf{R}$. Uniqueness of the solution of the Volterra equation[35]

$$y_2(x)=\sigma^{-1}\sin\sigma x+\sigma^{-1}\int_0^x dx'\sin\sigma(x-x')p_a(x')y_2(x') \qquad \text{(A1)}$$

on $\mathbf{C}_\sigma\times[0,1]_x$, together with $p_a=p_a^*$, implies y_2 maps $\mathcal{G}\times[0,1]_x\to\mathbf{R}$. Therefore, $\partial_\lambda\epsilon_2=\epsilon_1\int_0^1dx'|\,y_2(x')\,|^2$ when $\epsilon_2=0$.

Newton[19] proved $\epsilon_2=0\Rightarrow\epsilon_1=e^{i\kappa}$ or $\epsilon_1=e^{-i\kappa}$. But $\epsilon_2=0\Rightarrow\kappa\in\mathcal{G}\Rightarrow\kappa=n\pi+i\,\mathrm{Im}\,\kappa\Rightarrow\epsilon_1=(-1)^n\times\exp(\mp\,\mathrm{Im}\,\kappa)\Rightarrow\partial_\lambda\epsilon_2(\epsilon_2=0)=(-1)^n$

TABLE II. Poles and zeros. A function f is ∞, $*$, or 0 iff f has a pole, is pole- and zero-free, or has a zero, respectively. Also, $\hat{S}$ is s continuous (without poles) on $\mathbf{R}_s$.

	$s=0$ $(\lambda_0<0)$	$\sigma=0$ $(\lambda_0<0)$	$s=0$ $(\lambda_0>0)$	$\varsigma=0$ $(\lambda_0>0)$	$W=0$	$W_c=0$	$\kappa=n\pi$ $(n\neq0)$	$m_1=\infty$
$\hat{M}$	∞	$*$	$*$	∞	$*$	$*$	$*$	∞
$\hat{M}^{-1}$	$*$	$*$	$*$	$*$	∞	$*$	$*$	$*$
$\hat{\psi}$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$
$\hat{\Phi}$	$*$	$*$	$*$	$*$	$*$	∞	$*$	$*$
$\hat{J}$	∞	$*$	$*$	∞	∞	$*$	$*$	$*$
$\det\hat{J}$	$*$	$*$	0	∞	∞	0	$*$	$*$
$\hat{J}^{-1}$	∞	$*$	∞	$*$	$*$	∞	$*$	$*$

$\times \exp(\mp \operatorname{Im}\kappa)\int_0^1 dx'|y_2(x')|^2$. The last integral is not zero; if it were then y_2 would not satisfy (A1) $\forall x\in[0,1]$. ■

Proof of Lemma 2.2: Equation (A1) implies y_2 maps $\mathscr{B}\mathscr{G}\times[0,1]_x\to\mathbf{R}$. So $\epsilon_2=y_2(1)$ is real on $\mathscr{B}\mathscr{G}$ and $\exists$ sign ϵ_2.

The literature[15] shows that the zeros in $\mathbf{C}_\lambda$ of ϵ_2 are in $\mathscr{G}$ and that there is precisely one zero in each finite-length gap. Lemma 2.1 shows the zeros are simple. Therefore, sign ϵ_2 is constant in each band and ϵ_2 has opposite signs in neighboring bands in $\mathbf{R}_\lambda$. Also, Lemma 2.1 shows $\partial_\lambda\epsilon_2<0$ at the zero of ϵ_2 in the lowest-energy finite-length gap. So $\epsilon_2>0$ in the lowest-energy band. These facts and Fig. 1 imply sign $\epsilon_2=(-1)^n\operatorname{sign}\kappa$ in the band $(n\pi,[n+1]\pi)_\kappa$, in which $\operatorname{sign}(\sin\kappa)=(-1)^n$ as well.

Justification for Fig. 2 and Table I: Multiplicities will be determined by x asymptotics of $e^{\pm i\varsigma x}H(-x)$ and $e^{\pm i\kappa x}\xi H(x)$. It is useful to note $\kappa\in\mathbf{R}$ in $\mathscr{B}$, $\operatorname{Im}\kappa>0$ in Int $\mathscr{G}$, $\varsigma\in\mathbf{R}$ above threshold ($\lambda>\lambda_0$), and $\varsigma=i|\varsigma|$ below threshold.

In each band above threshold, $e^{\pm i\varsigma x}H(-x)$ and $\beta H(x)$ are in $L_\infty\setminus L_2$ so there are two linearly independent L_∞ solutions of (3.1). Bands above threshold form $\mathscr{B}_2$.

Each L_∞ solution of (3.1) in a gap above threshold has the form const$\times\beta_1$ for $x>0$. Above threshold, $e^{\pm i\varsigma x}H(-x)\in L_\infty\setminus L_2$. So there is a one-parameter family (const$\times\psi_1$) of $L_\infty\setminus L_2$ solutions in gaps above threshold, which gaps form $\mathscr{G}_1$.

Each L_∞ solution of (3.1) below threshold has the form const$\times e^{-i\varsigma x}$ for $x<0$. In bands, $\beta H(x)\in L_\infty\setminus L_2$ so there is a one-parameter family (const$\times\psi_2$) of $L_\infty\setminus L_2$ solutions in bands below threshold, which bands form $\mathscr{B}_1$. In gaps, $\beta_2 H(x)\notin L_\infty\cup L_2$ but $\beta_1 H(x)\in L_\infty\cap L_2$, so the only L_∞ solutions in gaps below threshold have the form $\text{const}_1\times e^{-i\varsigma x}H(-x)+\text{const}_2\times\beta_1 H(x)\in L_2(\mathbf{R}_x)$: Gaps below threshold form $\mathscr{G}_0$.

The operator $-\partial_x^2+p$ is self-adjoint so its spectrum is in $\mathbf{R}_\lambda$. The operator's spectrum is $\mathscr{G}_1\cup\mathscr{B}$, in which there is no L_2 solution of (3.1). Therefore, the point spectrum of $-\partial_x^2+p$ is in $\mathscr{G}_0$.

The L_2 solutions of (3.1) are bound states. If bound states exist then they are in $\mathscr{G}_0$, and, being solutions of (3.1), they are twice differentiable. Continuity and an argument two paragraphs above show that bound states have the form

$$e^{-i\varsigma x}H(-x)\beta_1 H(x) \tag{A2}$$

times an x independent constant. Continuity of the first derivative shows (A2) is a solution of (3.1) iff $W\equiv-(m_1+i\varsigma)=0$. Therefore the point spectrum of $-\partial_x^2+p$ is $\{\lambda\in\mathscr{G}_0: W(\lambda)=0\}$, which is the set of bound state energies of (3.1).

Table I summarizes the previous six paragraphs. ■

Sketch of proof of Lemma 4.1: Suppose f_i solve $(-\partial_x^2+\text{potential}-\lambda)f_i=0$, f_1 and f_2 are linearly independent, as are f_3 and f_4. Then

$$[f_1(x')f_2(x)-f_1(x)f_2(x')]/W[f_1,f_2] = [f_3(x')f_4(x)-f_3(x)f_4(x')]/W[f_3,f_4]. \tag{A3}$$

Apply (A3) to (4.1), using B of (3.2), to get

$$\epsilon_2\Psi_1(x\geqslant 0)=\epsilon_2\beta_1-W_a^{-1}\int_x^\infty dx'[\beta_1(x')\beta_2(x)-\beta_1(x)\beta_2(x')]q(x')\epsilon_2\Psi_1(x'). \tag{A4}$$

The equation is identical to (4.1) in Ref. 33, from which it follows that $\epsilon_2\Psi_1(x\geqslant 0)$ has properties claimed in the lemma.

It follows from (A4) that

$$\epsilon_2\Psi_1(0)-\epsilon_2\psi_1(0)=-\int_0^\infty dx'\,g(x',x)q\epsilon_2\Psi_1.$$

Then (4.1) implies

$$\epsilon_2\Psi_1(x<0)=\epsilon_2[\Psi_1(0)-\psi_1(0)+\psi_1(x)]-\int_x^0 dx'\,g(x',x)q\epsilon_2\Psi_1.$$

Equation (A3) and Y [see (3.2)] are used to write g conveniently:

$$\epsilon_2\Psi_1(x<0)=\epsilon_2[\Psi_1(0)-\psi_1(0)+\psi_1(x)]-\int_x^0 dx'\varsigma^{-1}[\sin\varsigma(x-x')]q\epsilon_2\Psi_1. \tag{A5}$$

Equation (A5) can be used to prove $\epsilon_2\Psi_1(x<0)$ has properties claimed in the lemma. The proof follows one in Ref. 33 for $\epsilon_2\Psi_1(x\geqslant 0)$, so it is omitted.

The proof for Ψ_2 is similar to that for Ψ_1, but use

$$\Psi_2(x\leqslant 0)=e^{-i\varsigma x}-\varsigma^{-1}\int_{-\infty}^x dx'[\sin\varsigma(x'-x)]q\Psi_2 \tag{A6}$$

in place of (A4) and

$$\Psi_2(x>0)=[\Psi_2(0)-\psi_2(0)+\psi_2(x)]-W_a^{-1}\int_0^x dx'[\beta_1(x')\beta_2(x)-\beta_1(x)\beta_2(x')]q\Psi_2 \tag{A7}$$

in place of (A5). ■

Proof of Lemma 4.2: The proof has two parts. The first is about Ψ.

Compare (A4) and (A6) to similar equations in Ref. 33 in order to show $\Psi_1(x\geqslant 0)$ and $\Psi_2(x\leqslant 0)$ have properties claimed in the lemma.

The proof for $\Psi_2(x>0)$ uses the technique in Ref. 36, for which it need only be shown that $\Sigma_{n=1}^\infty\Psi_2^{(n)}(x)$ converges uniformly on compact subsets of $\mathbf{C}_s^+\cup\mathbf{R}_s$, where

$$\Psi_2^{(n)}(x>0)=(-1)^n\int_0^x dx_1\cdots\int_0^{x_{n-1}}dx_n[g(x,x_1)\times\cdots g(x_{n-1},x_n)][q(x_1)\cdots q(x_n)]\times[\Psi_2(0)-1+\psi_2(x_n)]$$

is the nth iterate of (A7). [The previous equation uses $\psi(0)=\hat{1}$, which follows (3.2).] Toward that end use (A3), (3.4), and Y to show

$$g(x>0,x'>0)=[y_1(x)y_2(x')-y_1(x')y_2(x)] \tag{A8}$$

and to show

$$g(x>0,x'>0)=[1+O(s^{-1})]s^{-1}\sin s(x'-x)$$

and

$$|g(x>0,x'>0)|\leqslant[1+O(|s|^{-1})]\,|s|^{-1}e^{\tau(x-x')}$$

as $|s|\uparrow\infty$ in $\mathbf{C}_s^+\cup\mathbf{R}_s$ with $0\leqslant\mathrm{Im}\,s\leqslant\tau$. It follows from (3.5) and $\Psi_2(0)=1+O(s^{-1})$ that

$$|\Psi_2^{(n)}(x>0)|\leqslant[1+O(|s|^{-1})]\,|s|^{-n}e^{\tau x}\int_0^x dx_1 \times\cdots\int_0^{x_{n-1}}dx_n|q(x_1)\cdots q(x_n)|$$

as $|s|\uparrow\infty$ in $\mathbf{C}_s^+\cup\mathbf{R}_s$ with $0\leqslant\mathrm{Im}\,s\leqslant\tau$. The remainder of the proof for $\Psi_2(x>0)$ is as in Ref. 36, with details in Ref. 8.

The proof of the lemma for $\Psi_1(x<0)$ is similar to that for $\Psi_2(x>0)$.

The proof's first part is complete. The second part is about Ψ'.

Equation (4.1) implies

$$\epsilon_2(\Psi_1'-\psi_1')=-\int_x^\infty dx'[\,y_1(x')y_2'(x) -y_1'(x)y_2(x')\,]q\epsilon_2\Psi_1. \tag{A9}$$

Use (2.2), (3.5), and $\Psi_1=[1+O(s^{-1})]\psi_1$ to obtain

$$|\epsilon_2(\Psi_1'-\psi_1')_{x>0}| \leqslant[1+O(|s|^{-1})]\,|s^{-1}\sin s|e^{-\tau x}\int_0^\infty dx'|q|$$

in the usual s limit, with $\tau=\mathrm{Im}\,s$. Then (3.6) completes the proof for $\Psi_1'(x\geqslant 0)$. For $\Psi_1'(x<0)$ replace $\int_x^\infty$ with $\int_x^0+\int_0^\infty$ in (A9) and treat separately the two integrals.

The proof for Ψ_2' is similar to that for Ψ_1'. ■

Proof of Theorem 4.2: The proof has two parts. The first part is about independence of columns in Table II.

The function ς depends on λ_0 but σ is λ_0 independent, so $\varsigma(\sigma=0)\neq 0$ and $\sigma(\varsigma=0)\neq 0$ e.p. However, either $\sigma(s=0)=0$ or $\varsigma(s=0)=0$, and $s(\sigma=0)=0$ or $s(\varsigma=0)=0$.

Quantities in Table II inherit poles and zeros at $\epsilon_2=0$ only by way of poles of $m_1=(e^{i\kappa}-\epsilon_1)/\epsilon_2=\zeta_1/(\epsilon_1-e^{-i\kappa})$. The latter equality shows that if $\epsilon_2=0$ then either $\epsilon_1=e^{i\kappa}$ or $\epsilon_1=e^{-i\kappa}$. If $\epsilon_2=0$ and $\epsilon_1=e^{i\kappa}$ then m_1 is finite, but if $\epsilon_2=0$ and $\epsilon_1=e^{-i\kappa}$ then m_1 is infinite. Conversely, $\epsilon_2=0$ at every pole of m_1.

The column heading $W_c=0$ is the only q dependent one, so it overlaps no other heading e.p. The heading $W=0$ is the only one which is both p dependent and q independent, so it is distinct e.p. Headings $\kappa=n\pi$ and $\epsilon_2=0$ depend only on the periodic part of p, so they are distinct from all others e.p. However, $\kappa=n\pi$ at gap endpoints and $\epsilon_2=0$ in Int $\mathscr{G}_0$. Therefore, each column heading is independent of other headings e.p.

The proof's first part is complete. The proof's second part is about row entries. The qualifier e.p. is suppressed in the remainder of the proof.

Rows for $\hat{M}$ and $\hat{M}^{-1}$ follow directly from (4.13) and the row for $\hat{\psi}$ follows from (3.2). The row for $\hat{\Phi}$ follows from the row for $\hat{\psi}$ by the Fredholm method, except at $W=0$, where $\hat{\Phi}=(W\hat{M}^{-1})\Psi/W_c\sim *$. The row for $\hat{J}$ follows from (4.18) and the row for $\hat{\Phi}$. The row for $\det\hat{J}$ follows from (4.18). The row for $\hat{J}^{-1}$ follows from the two previous rows.

Wronskian algebra (see beginning of this appendix) and (4.17) yield

$$\hat{S}=W_cW[\hat{\Phi}^*,\hat{\Phi}]PQ/(2i\varsigma W), \tag{A10}$$

showing that any *hypothetical* pole of $\hat{S}$ in $\mathbf{R}_s$ is at $\varsigma=0$, $W=0$, or $W_c=0$. Theorems 3.1 and 4.1 show that zeros of W and W_c are in Int $\mathscr{G}_0$ and Fig. 2 shows $\varsigma=0$ is in the closure of $\mathscr{G}_0$ except when $0<\lambda_0\in\mathscr{B}$. Therefore, if it is proven that $\hat{S}$ has no pole in the closure of $\mathscr{G}_0$, and that $W[\hat{\Phi}^*,\hat{\Phi}]=0$ at $\varsigma=0$ when $0<\lambda_0\in\mathscr{B}$, then the proof is complete.

This paragraph applies only to the closure of $\mathscr{G}_0$. Table I and (2.6) show $\varsigma=i|\varsigma|$ and $m^*=m$. Then (4.13) implies $\hat{M}^*=\hat{M}$. Equations (4.12) and (4.16) imply $\hat{S}=\hat{M}^{*-1}\mathbf{I}\hat{M}Q=Q\sim *$ in the closure of $\mathscr{G}_0$.

It was just shown that $\hat{S}$ has no pole in $\mathbf{R}_s$, except possibly at $\varsigma=0$ when $0<\lambda_0\in\mathscr{B}$. This paragraph applies only to the possible exception. Figure 2 shows such an exceptional point must be an endpoint of $\mathscr{B}_1$, and (4.12) shows $S_{21}=0$ there. Wronskian algebra and $\Phi^*=S\Phi$ imply $W[\Phi_2^*,\Phi_2]=W^2S_{21}/W_c=0$. Moreover, (4.13) and $\hat{\Phi}=\hat{M}^{-1}\Phi$ imply $\hat{\Phi}=\Phi_2\hat{1}$ whenever $\varsigma=0$. It follows that $\hat{\Phi}_1=\hat{\Phi}_2=\Phi_2\propto\Phi_2^*$, which implies $W[\hat{\Phi}^*,\hat{\Phi}]=0$. Then (A10) shows $\hat{S}$ has no pole unless $W=0$ at $\varsigma=0$ when $0<\lambda_0\in\mathscr{B}$. But Theorem 3.1 says W cannot vanish in $\mathscr{B}$. So $\hat{S}$ has no pole at $\varsigma=0$ when $0<\lambda_0\in\mathscr{B}$. ■

Proof of Theorem 4.3: Theorem 4.2 shows $\hat{\Phi}$ has only finitely many poles, so its poles do not affect asymptotics.

Equations (3.3) and (4.13) and $\hat{\Phi}=\hat{M}^{-1}\Phi$ show $\hat{\Phi}_1=-[2i\varsigma\Phi_1+(m_1-i\varsigma)\Phi_2]/W$ and $\hat{\Phi}_2=\Phi_2$. Then (4.11) implies

$$\hat{\Phi}_1=-[2i\varsigma W\Psi_1+(m_1-i\varsigma)\Psi_2]/W_c,\quad \hat{\Phi}_2=W\Psi_2/W_c. \tag{A11}$$

The s asymptotics of $\hat{\Phi}_2$ and $\hat{\Phi}_2'$ then follow from (3.5), (3.6), Lemma 4.2, and Theorem 4.1.

For asymptotics of $\hat{\Phi}_1$ in $\mathbf{R}_s$, use Lemma 4.2 and Theorem 4.1 to obtain $\hat{\Phi}_1=[1+O(s^{-1})][2i\varsigma\psi_1+(m_1-i\varsigma)\psi_2]/W$. Then (3.2), (3.4), and (4.13) imply $\hat{\Phi}_1=[1+O(s^{-1})]\hat{\psi}_1=[1+O(s^{-1})]e^{-isx}$ on $\mathbf{R}_x^-$.

For $|s|,\infty$ in $\mathbf{C}_s^+$, use (4.1) to obtain

$$\left[\frac{2i\varsigma\Psi_1}{W}\right]=\left[\frac{2i\varsigma\psi_1}{W}\right]-\int_x^\infty dx'\,g(x',x)q\left[\frac{2i\varsigma\Psi_1}{W}\right].$$

An argument in the proof of Lemma 4.2 shows $[2i\varsigma\Psi_1/W]=[1+O(|s|^{-1})][2i\varsigma\psi_1/W]$. Use (3.3) and domination of $e^{i\varsigma x}$ over $e^{-i\varsigma x}$ in $\mathbf{C}_s^+\times\mathbf{R}_x^-$ to show $[2i\varsigma\psi_1/W]=[1+O(|s|^{-1})]e^{i\varsigma x}$ on $\mathbf{R}_x^-$. Therefore, $[2i\varsigma\Psi_1/W]=-[1+O(|s|^{-1})]e^{isx}$ on $\mathbf{R}_x^-$. It follows from (3.2) and Lemma 4.2 that $\Psi_2=[1+O(|s|^{-1})]e^{-i\varsigma x}$ in $\mathbf{R}_x^-$. Theorem 4.1 and (A11) then imply $\hat{\Phi}_1=[1+O(|s|^{-1})]e^{isx}$ in $\mathbf{R}_x^-$.

It was just proved that $\hat{\Phi}_1$, $\hat{\Phi}_1'$, and $\hat{\Phi}_2$ have the s asymptotics asserted in the theorem. Those results and (4.15) show $\hat{\Phi}_1'=[1+O(|s|^{-1})]ise^{isx}$ in $\mathbf{R}_x^-$.

The s asymptotics of $\hat{J}$ come from Theorem 4.1, (4.18), and s asymptotics of $\hat{\Phi}$ and $\hat{\Phi}'$. Equation (4.17) is used in the derivation $Q\hat{J}\hat{\Phi}=Q\Upsilon=\Upsilon^*=\hat{J}^*\hat{\Phi}^*=\hat{J}^*\hat{S}Q\hat{\Phi}\Rightarrow Q\hat{J}$

$= \hat{J}^* \hat{S} Q \Rightarrow \hat{S} = \hat{J}^{*-1} Q \hat{J} Q$. The s asymptotics of $\hat{S}$ in $\mathbf{R}_s$ then follow from s asymptotics of $\hat{J}$. ■

[1]R. G. Newton, foreword to *Inverse problems in Quantum Scattering Theory* by K. Chadan and P. C. Sabatier (Springer, New York, 1977), pp. v–xvii; L. D. Faddeev, Usp. Mat. Nauk. **14**, 57 (1959) [Transl. B. Seckler, J. Math. Phys. **4**, 72 (1963)]; L. D. Faddeev, Itogi Nauk. Tekh. Sov. Problemy Mat. **3**, 93 (1974) [J. Sov. Math. **5**, 334 (1976)].

[2]R. G. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982), 2nd ed., esp. pp. 667–670.

[3]V. S. Buslaev and V. L. Fomin, Vestn. Leningr. Un-ta, 56 (1962) [Carol Hutchins, physics librarian for Indiana University, wrote in a personal communication, "It appears that a [translation] of this [paper] does not exist." Results are summarized in English in L. D. Faddeev, J. Sov. Math. **5**, 334 (1976), pp. 367–71]; J. Legendre, Ph.D. thesis, Université des Sciences et Techniques du Languedoc, Académie de Montpellier, France, 1982; E. Ja. Hruslov, Math. Sb. **99**, 141 (1976) [Math. USSR Sbornik **28**, 229 (1976)].

[4]The continuous spectrum has multiplicity N at energies λ for which the Schrödinger equation has N linearly independent solutions in $L\infty \setminus L_2$.

[5]N. E. Firsova, Mat. Zametki **18**, 831 [Math. Notes **18**, 1085 (1975)].

[6]R. G. Newton, J. Math. Phys. **24**, 2152 (1983); J. Math. Phys. **26**, 311 (1985).

[7]If $p(x) = p(x+1)$, $\int_{-\infty}^{\infty} dx(1+|x|)|q| < \infty$, and $\int_{-\infty}^{\infty} dx\, q \neq 0$ then there is one bound state of $-\partial_x^2 + p + q$ in each gap of sufficiently high energy. Counterexamples include reflectionless potentials. The fact was proven by V. A. Zheludev, in *Topics in Mathematical Physics*, edited by M. Sh. Birman (Consultants Bureau, New York, 1986), Vol. 2, pp. 87–101; M. Klaus, Hel. Phys. Acta **55**, 49 (1982); F. S. Rofe-Beketov, *Constructive Theory of Functions: Proceedings of the International Conference on Constructive Theory of Functions*, edited by B. Sendov *et al.* (Bulgarian Academy of Sciences, Sophia, Bulgaria, 1984), pp. 757–766; N. E. Firsova, Teor. Mat. Fiz. **62**, 196 (1985) [Theor. Math. Phys. **62**, 130 (1985)].

[8]T. M. Roberts, Ph.D. thesis, Indiana University, 1987 (University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106); "Inverse scattering for step-periodic potentials in one dimension," to appear in Inverse Problems.

[9]R. G. Newton, Geophys. J. R. Astron. Soc. **65**, 191 (1981), eps. pp. 193–195; [Geophys. J. R. Astron. Soc. **69**, 571 (1982)]; in *Conference on Inverse Scattering: Theory and Applications*, edited by J. B. Bednar, R. Redner, E. Robinson, and A. Weglein (SIAM, Philadelphia, 1983), pp. 1–74, esp. pp. 6–8.

[10]Data usually are measured for one gap or one band only. A typical engineering application is filtering of transmitted signals. Within a finite interval of frequencies, the filter would reject a subinterval (gap) or pass the subinterval (band). The data are evidence for bands and gaps in the sense of consistency with models that use bands and gaps. For phonons, read: V. Narayanamurti, H. L. Störmer, M. A. Chin, A. C. Gossard, and W. Wiegmann, Phys. Rev. Lett. **43**, 2012 (1979); C. Colvard, T. A. Gant, M. V. Klein, R. Merlin, R. Fischer, H. Morkoc, and A. C. Gossard, Phys. Rev. B **31**, 31 (1985); B. Jusserand and D. Paquet, in *Heterojunctions and Semiconductor Superlattices*, edited by G. Allan, G. Bastard, N. Boccara, M. Lannoo, and M. Voos (Springer, New York, 1985), pp. 108–126. For photons, read: D. L. Perry, Appl. Opt. **4**, 987 (1965); A. Yariv and P. Yeh, *Optical Waves in Crystals* (Wiley, New York, 1984), Chap. 6; and E. Ritter, in *Physics of Thin Films*, edited by G. Hass, M. H. Francombe, and R. W. Hoffman (Academic, New York, 1975), Vol. 8, pp. 1–49, esp. top of p. 44.

[11]B. S. Pavlov, and N. V. Smirnov, Vestn. Leningr. Univ., Mat. Meh. Astronom. 71–80, 171 (1977) [Transl. J. R. Schulenberger, Vestn. Leningr. Univ. Math. **10**, 307 (1982)].

[12]R. G. Newton, J. Math. Phys. **25**, 2991 (1984).

[13]D. B. Hinton, M. Klaus, and J. K. Shaw, "On the Titchmarsh-Weyl function for the half-line perturbed periodic Hill's equation," to appear in Q. J. Appl. Math. (Oxford).

[14]F. Gesztesy, Lect. Notes Math. **1218**, 93–122 (1986).

[15]Section II borrows without individual attribution from the following: E. C. Titchmarsh, *Eigenfunction Expansions Associated With Second-Order Differential Equations* (Oxford U.P., London, 1958), Part 2, pp. 290ff; W. Kohn, Phys. Rev. **115**, 809 (1959); B. M. Levitan and I. S. Sargsjan, *Introduction to Spectral Theory* (American Mathematical Society, Providence, 1975); H. P. McKean and P. van Moerbeke, Inventiones Math. **30**, 217 (1975); N. E. Firsova, Zap. Nauch. Sem. Leningrad. Otdel. Mat. Inst. Steklova **51**, 183 (1975) [J. Sov. Math. **11**, 487 (1979)]; Ref. 5; H. Hochstadt, SIAM J. Appl. Math. **31**, 392 (1976); J. E. Avron and B. Simon, Ann. Phys. **110**, 85 (1978); M. Reed and B. Simon, *Methods of Modern Mathematical Physics IV: Analysis of Operators* (Academic, New York, 1978), pp. 279ff; W. Magnus and S. Winkler, *Hill's Equation* (Dover, New York, 1979); R. G. Newton, J. Math. Phys. **24**, 2152 (1983).

[16]L. V. Ahlfors, *Complex Analysis* (McGraw-Hill, New York, 1979), 3rd ed., p. 71.

[17]The literature shows $\xi(\kappa,x)$ is x periodic $\forall \kappa \in \mathscr{B}\mathscr{G}$, but later work requires a broader statement. So note $\forall x \in \mathbf{R}$, $\xi = e^{\mp i\kappa x}\beta = e^{\mp i\kappa x}(y_1 + my_2)$ has a κ-analytic continuation from $\mathscr{B}$ onto $\mathbf{C}_\kappa^{\text{cut}} = \mathbf{C}_\kappa \setminus [\text{finite gaps}]$. So for each x, $\xi(\kappa, x+1) = \xi(\kappa,x)$ in $\mathbf{C}_\kappa^{\text{cut}}$ because $\xi(\kappa,x+1) = \xi(\kappa,x)$ on nonempty intervals ($\forall \kappa \in \mathscr{B}$) in $\mathbf{C}_\kappa^{\text{cut}}$. But ξ is x periodic on the κ cuts too, so ξ is x periodic $\forall \kappa \in \mathbf{C}_\kappa$.

[18]If $(-\partial_x^2 + \text{potential} - \lambda)f_i = 0$ then the Wronskian $W[f_1,f_2] = f_1 f_2' - f_1' f_2$ is x independent.

[19]R. G. Newton, J. Math. Phys. **24**, 2152 (1983), p. 2155; E. Trubowitz, Commun. Pure Appl. Math. XXX, 321 (1977), p. 322; B. M. Levitan and I. S. Sargsjan, Ref. 15, p. 21.

[20]The poles are λ simple because Lemma 2.1 shows $\partial_\lambda \epsilon_2(\epsilon_2 = 0) \neq 0$. The poles are σ simple because $\partial_\sigma = 2\sigma\partial_\lambda$ and because $\sigma(\kappa) = 0$ iff $\kappa = 0$. I regard the possibility $\epsilon_2(\lambda = 0) = 0$ as pathological because ϵ_2 is σ entire.

[21]The abbreviation e.p. means *except in pathological cases.*

[22]The statement is false if $\epsilon_2 = 0$ because $m_1 = \infty$ then. But then $-i\varsigma \neq \infty = m_1(\epsilon_2 = 0)$. So $W(\epsilon_2 = 0) \neq 0$.

[23]The function $e^{i\kappa}$ is analytic across *bands* because κ is. The fact that $e^{i\kappa}$ is not the Schwarz reflection of f across *gaps*, together with the monodromy theorem, implies $\kappa = \kappa(s)$ has branch points at the endpoints of gaps. For monodromy, read John B. Conway, *Functions of One Complex Variable* (Springer, New York, 1978), 2nd ed., pp. 219, 220.

[24]Pages 346, 347 of Ref. 2.

[25]E. C. Titchmarsh, *Eigenfunction Expansions Associated With Second-Order Differential Equations* (Oxford U.P., London, 1958), Pt. 2, pp. 291, 292 and 302, 303. Add to the reference's equation (21.7.2) a proviso that $\text{Im}\sqrt{\lambda}$ is not small.

[26]At two points in this proof I refer to facts established later in the proof.

[27]Finiteness of the zeros of W comes from Theorem 3.1. Finiteness for W_c is shown later in this proof.

[28]It follows that, in (4.8), s simple poles of ψ_1 and Ψ_1 at $\epsilon_2 = 0$ are removed by s simple zeros of $1/W$.

[29]The coincidence is pathological only if $\{\lambda : W_c(\lambda) = 0\}$ has measure zero. The proof 's third part shows the set's measure is zero.

[30]At $\varsigma = 0$: $\int_0^\infty dx| - g^+ q|$ is finite if $q \in L_1$, but (3.2) shows $\int_{-\infty}^0 dx| - g^+ q| = \int_{-\infty}^0 dx|(1 + m_1 x)q/m_1|$. That is why $xq \in L_1$ is assumed.

[31]The points $\varsigma = 0$, $\kappa = n\pi$, and $W = 0$ are distinct, e.p.

[32]R. G. Newton, J. Math. Phys. **21**, 493 (1980). The last integrand in the left column of p. 504 should be multiplied by V.

[33]R. G. Newton, J. Math. Phys. **24**, 2152 (1983).

[34]Compare (4.18) with T. Aktosun, Inverse Problems **3**, L5 (1987).

[35]Levitan and Sargsjan, Ref. 15, p. 5.

[36]Pages 333, 334 of Ref. 2 and p. 2158 of Ref. 33.

A Clifford algebra quantization of Dirac's electron positron field

H. T. Cho, Adel Diek, and R. Kantowski
Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019

(Received 23 June 1989; accepted for publication 9 May 1990)

The quantum field theory of free Dirac particles (four-component massive spin-$\frac{1}{2}$ particles) is "derived" by a Segal quantization procedure. First, details are given on how the spinor space of Dirac is actually a minimal left ideal of the Clifford algebra associated with a Lorentz inner product space $(+,-,-,-)$, and how the homogeneous group actions break the natural two-component quaternion structure to give familiar four-component complex spinors. Second, Wigner's procedure for constructing unitary representations of the Poincaré group is used to construct the appropriately induced infinite-dimensional representation of the inhomogeneous group starting from the above four-dimensional nonunitary representation. Third, and finally, Segal's procedure for quantizing classical Fermion fields is adapted to this infinite-dimensional Hilbert space to obtain the Clifford algebra of annihilation–creation operators for spin-$\frac{1}{2}$ particles. The familiar Fock space appears as a minimal left ideal in this second Clifford algebra.

I. INTRODUCTION

In this paper we show how the familiar quantum field theory of free massive Dirac spin-$\frac{1}{2}$ particles[1,2] can be obtained by two successive Clifford algebra constructions. We refer to this generic field as the electron–positron field for convenience, and have attempted to use notation familiar to the physics community when possible.

In Sec. II we give a short introduction to the standard construction of the Clifford algebra associated with a real vector space possessing a nondegenerate inner product.[3,4] This construction is applied to an infinite-dimensional space in Sec. IV. However, in Sec. II emphasis is placed on a Minkowski inner-product space with signature -2 and its corresponding Clifford algebra of gamma matrices. The technique for generating spinor representations of the associated Clifford algebras is given and applied to the four-dimensional case.[5–8] By carefully including the discrete transformations (parity and time reversal) we are able to show how the presence of projective representations and additional group actions (i.e., phase rotation and charge conjugation) in the Dirac theory destroys the expected two-component quaternion structure of spinors for the $(+,-,-,-)$ metric. We assume the homogeneous group structure to consist of a covering group of the homogeneous Lorentz group and the above additional members. Throughout we use the four-dimensional spinor basis corresponding to rest states having spins oriented along the $\pm z$ axis and possessing ± 1 parity. Our motive is to use an explicit basis that produces the Pauli–Dirac representation of the gamma matrices familiar to all physicists.[1,2] Other representations such as Weyl or Majorana could easily be used.

In Sec. III we use Wigner's procedure for constructing unitary representations of the Poincaré group to construct a representation of the inhomogeneous group obtained by combining the above homogeneous group with Minkowski space translations.[9–14] Two ingredients are critical and both make use of the four-component Dirac representation of Sec. II. The first is a character, or equivalently a one-dimensional unitary representation of the translations (massive for the case considered here), and the second is a unitary representation of the Little group of this character (the invariance group of this character). Both are found in the above four-component representation; e.g., when the homogeneous group is restricted to the Little subgroup, the spin representation becomes unitary.

The reason for constructing this representation and its infinite-dimensional Hilbert space is that in Sec. IV a second Clifford algebra, the operator algebra of the electron–positron theory, is constructed. By a straightforward extension of the general construction outlined in Sec. II, and previously attributed to Segal,[15–21] this complex Clifford algebra is constructed from the infinite-dimensional complex Hilbert space. All the general notions introduced in Sec. II can be applied to this infinite-dimensional Clifford algebra. In particular, a projection operator (the Fock vacuum) is used to generate the space of spinors (the Fock space).[22] This is a new construction differing from a previously introduced Fock space.[18] Prior to Sec. IV the only complex structure present came from the four-component spinor representation of the first Clifford algebra and appeared in the unitary representation of Sec. III; however, in Sec. IV another complex structure in the second Clifford algebra appears.

In this paper we have tried to "draw" a straight line from Minkowski space to the quantum field Ψ, however, as the reader will obviously notice we made several choices (usually among a few possibilities) along the way. Since the Dirac theory is the standard theory for electrons and positrons, we have used it as a guide to make the appropriate choices.

II. THE MINKOWSKI CLIFFORD ALGEBRA AND DIRAC SPINORS

This section serves primarily to establish needed background and notation. However, inclusion of group actions, beyond special Lorentz, is new and allows us to clarify why Dirac spinors are four-component complex and not two-

 0022-2488/90/092192-09$03.00

component quaternion when the signature is -2.[3-8,23-25]

The 16-dimensional real Clifford algebra $R_{1,3}$ can be attached to each point $x \in M^4$ of Minkowski space by selecting a translationally invariant set of basis vectors $e_\mu \cong \partial_\mu$ satisfying $e_\mu \cdot e_\nu = g_{\mu\nu} = \text{diagonal}(1,-1,-1,-1)$ to span the tangent space at each x. In general, the universal Clifford algebra (CA) associated with a real vector space V possessing a nondegenerate symmetric bilinear form $(,)$ is the unique associative algebra:

(i) with identity I,

(ii) generated by a subspace $V^1 \subset \text{CA}$ isomorphic to V,

(iii) which has its algebraic multiplication constrained by

$$v^1 w^1 + w^1 v^1 = 2(v,w)I. \tag{2.1}$$

For M^4 we write $v = r^\mu e_\mu$, and the isomorphic vector images in $V^1 \subset \text{CA}$ as $v^1 = r^\mu \gamma_\mu$. The defining algebraic constraint (iii) familiarly appears as $\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2g_{\mu\nu} I$.

Using orthonormal frames such as these ($e_\mu \leftrightarrow \gamma_\mu$) allows the 16-dimensional real Clifford algebra $R_{1,3}$ to be decomposed into a direct sum of Lorentz scalars, vectors, bivectors, pseudovectors, and pseudoscalars $R_{1,3} = V^0 \oplus V^1 \oplus V^2 \oplus V^3 \oplus V^4$, where

$$\begin{aligned} &V^0 = \{rI\}, \quad V^1 = \{r^\mu \gamma_\mu\}, \quad V^2 = \{r^{\mu\nu} \gamma_\mu \gamma_\nu\}, \\ &V^3 = \{r^{\mu\nu\lambda} \gamma_\mu \gamma_\nu \gamma_\gamma\} = \{r^\mu \gamma_\mu \gamma\}, \\ &V^4 = \{r\gamma\} \quad \text{where } \gamma \equiv \gamma_0 \gamma_1 \gamma_2 \gamma_3, \\ &\text{with } r, r^\mu, r^{\mu\nu}, \text{ and } r^{\mu\nu\lambda} \text{ real}, \\ &\mu < \nu < \lambda, \quad \text{and } \gamma\gamma = -I. \end{aligned} \tag{2.2}$$

In general, the even subalgebra $\text{CA}^+ \equiv \{V^0 \oplus V^2 \oplus V^4 \oplus \cdots\}$ of a Clifford algebra is isomorphic to another Clifford algebra. In the case of $R_{1,3}$ the even subalgebra is isomorphic to $R_{3,0}$ the eight-dimensional real algebra associated with the three-dimensional Euclidean space. The algebra $R_{3,0}$ is isomorphic to the four-dimensional complex Pauli algebra because the center of $R_{3,0}$ is isomorphic to the complex numbers. Then,

$$\begin{aligned} R^+_{1,3} &= \{rI\} \oplus \{r^{\mu\nu} \gamma_\mu \gamma_\nu\} \oplus \{r\gamma\} \\ &= \{rI\} \oplus \{r^i \alpha_i\} \oplus \{r^i \alpha_i \alpha\} \oplus \{r\alpha\} \\ &\cong R_{3,0} = \{rI\} \oplus \{r^i \sigma_i\} \oplus \{r^i \sigma_i \sigma\} \oplus \{r\sigma\} \\ &\cong \text{Pauli} = \{(r + ir')\} \oplus \{(r^i + ir'^i)\sigma_i\}, \end{aligned} \tag{2.3}$$

where $\alpha_i \equiv \gamma_i \gamma_0$, $\alpha \equiv \alpha_1 \alpha_2 \alpha_3 = \gamma$. The center of $R_{3,0}$ is $\{rI\} \oplus \{r'\sigma\} \cong \{(r + ir')I\}$. In (2.3) the σ_i are images of an orthonormal basis ϵ_i, $(\epsilon_i \cdot \epsilon_j = \delta_{ij})$ of a three-dimensional Euclidean space and generate the associated Clifford algebra $R_{3,0}$. The isomorphism to the familiar complex Pauli matrix form is $\sigma_i \leftrightarrow$Pauli matrix σ_i, $\sigma \leftrightarrow$imaginary unit "i." The precise identification of $R^+_{1,3}$ with $R_{3,0}$ requires a choice of observer e_0 and its corresponding γ_0 in V^1. By identifying $\sigma_i \leftrightarrow \alpha_i = \gamma_i \gamma_0$, then $i \leftrightarrow \alpha = \gamma$ and we have the desired even-subalgebra isomorphism. We also have the decomposition of the total Clifford algebra,

$$R_{1,3} \cong R_{3,0} \oplus \gamma_0 R_{3,0}, \tag{2.4}$$

allowing (anti) automorphisms of $R_{3,0}$ to be extended to $R_{1,3}$ by defining their effect on γ_0. Every universal Clifford algebra associated with a real or complex vector space possesses a fundamental antiautomorphic involution called reversion $c \to \tilde{c}$ and a fundamental automorphic involution called inversion $c \to \bar{c}$. The defining properties are

$$\begin{aligned} &c_1 c_2 = \tilde{c}_2 \tilde{c}_1, \quad \tilde{c} = c, \quad c \in V^1, \\ &\overline{c_1 c_2} = \bar{c}_1 \bar{c}_2, \quad \bar{c} = -c, \quad c \in V^1. \end{aligned} \tag{2.5}$$

Vectors are invariant under reversion but are reversed in direction by inversion. For $R_{1,3}$ we denote reversion by tilde as above but for inversion we can write

$$\bar{c} = \gamma_5(c) \equiv \gamma c \gamma^{-1}, \tag{2.6}$$

where we have used the notation γ_5 of the Dirac helicity operator. For $R_{3,0}$ we use $c \to c^\dagger$, the familiar Hermitian conjugation for the Pauli algebra, for which $\alpha_i^\dagger = \alpha_i$, $\alpha^\dagger = (\alpha_1 \alpha_2 \alpha_3)^\dagger = \alpha_3 \alpha_2 \alpha_1 = -\alpha$. Equation (2.4) allows Pauli reversion to be extended to $R_{1,3}$ by requiring $\gamma_0^\dagger = \gamma_0$ where

$$c \to c^\dagger \equiv \gamma_0 \tilde{c} \gamma_0^{-1} \Rightarrow \gamma_i^\dagger = -\gamma_i. \tag{2.7}$$

The Pin group is a subgroup of the multiplicative group of invertible elements in CA that leave the subspace V^1 invariant when acting as inner automorphisms, i.e.,

$$c \in \text{Pin} \Leftrightarrow c V^1 c^{-1} = V^1, \quad \text{and satisfy } \tilde{c}c = \pm I. \tag{2.8}$$

With a choice of observer γ_0 in $R_{1,3}$, the $\text{Pin}_{1.3}$ constraint (2.8) can be written

$$c^\dagger \gamma_0 c = \pm \gamma_0. \tag{2.9}$$

The subgroup connected to the identity is isomorphic to SL(2,C) and is generated by products of rotations and boosts,

$$\begin{aligned} &\text{Rotations} = e^{(\theta^i/2)\gamma\alpha_i}, \quad \sqrt{\theta^i \theta^i} \leqslant 4\pi, \\ &\text{Boosts} = e^{(\zeta^i/2)\alpha_i}, \quad -\infty < \zeta^i < \infty, \\ &\text{SL}(2,\text{C}) \cong \left\{e^{[(\theta^i/2)\gamma + \zeta^i/2]\alpha_i}\right\}. \end{aligned} \tag{2.10}$$

The three other disconnected parts of $\text{Pin}_{1,3}$ are generated by products of the identity component with parity P (multiplication by γ_0) and time reversal T (multiplication by $\gamma_1 \gamma_2 \gamma_3$). The identity and parity components satisfy $\tilde{c}c = +I$ or equivalently $c^\dagger \gamma_0 c = +\gamma_0$ whereas the T and PT components satisfy $\tilde{c}c = -I$ or equivalently $c^\dagger \gamma_0 c = -\gamma_0$. When Pin acts as inner automorphisms on V^1 (called the vector representation) it double covers the invariance group of V's inner product (for $R_{1,3}$ the invariance group is the homogeneous Lorentz group),

$$\gamma_\mu \to (\pm c)\gamma_\mu (\pm c)^{-1} = \gamma_\nu \Lambda^\nu{}_\mu, \tag{2.11}$$

where $\Lambda^\nu{}_\mu$ is a Lorentz matrix. The spin representation of Pin arises by letting Pin act as left multiplications on CA. It is reducible with each invariant subspace giving a spinor space. A CA is decomposed into a direct sum of minimal left ideals, called spinor spaces CA_n, by finding a complete set of mutually annihilating primitive idempotents (projection operators) P_n,[26]

$$\begin{aligned} &P_n P_m = \delta_{n,m} P_n, \quad I = \sum_n P_n \Rightarrow \text{CA} = \sum_n \text{CA}_n, \\ &\text{where } \text{CA}_n \equiv \text{CA}\, P_n. \end{aligned} \tag{2.12}$$

Decomposition of $R_{1,3}$ requires two idempotents. To obtain matrix representations of the γ_μ's familiar to the physics community, we choose an observer (γ_0) and construct a pair of projection operators $P_\pm$ using a unit spatial direction α_3 ($\alpha_3^2 = I$) in the even subalgebra,

$$P_\pm \equiv \tfrac{1}{2}(I \pm \alpha_3), \Rightarrow R_{1,3} = R_{1,3+} \oplus R_{1,3-}. \tag{2.13}$$

It is at this point that clear differences in the fields of $R_{1,3}$, $R_{1,3\pm}$, and the complex numbers required for Dirac theory begin to appear. Both spinor spaces $R_{1,3\pm}$ as subspaces of $R_{1,3}$ form eight-dimensional vector spaces over the reals, but if they are used only as representation spaces for left multiplication by $\mathrm{Pin}_{1,3}$, they form two-dimensional vector spaces over the field of quaternions. However, Dirac theory contains an additional continuous U_1 group action (a right multiplication on $R_{1,3}$), uses a projective representative for time reversal T rather than using left multiplication by $\mathsf{T} = \gamma_1\gamma_2\gamma_3$, and introduces a projective representative for the additional charge conjugation symmetry C, all of which are inconsistent with the quaternion structure of $R_{1,3\pm}$. These new group actions "break" the two-component quaternion structure leaving a four-component complex structure for each spinor space. The remaining complex structure is defined by right multiplication by γ, i.e., multiplication by the unit imaginary "i" of the complex field is defined by $i(c) \equiv c\gamma$. Using the above minimal left ideals ($R_{1,3\pm}$) the U_1 group action on $R_{1,3}$ can be taken as right multiplication by

$$U_1(\phi) = e^{\phi\gamma\alpha_3} \quad (0 \leqslant \phi < 2\pi), \tag{2.14}$$

rotating the phases of the two spinor spaces $R_{1,3\pm}$ oppositely. The quaternion structure "broken" by (2.14) but not by left multiplications (spin transformations) is generated by right multiplications by γ_1, γ_2, and $\gamma_1\gamma_2$. These commute with the projection operators $P_\pm$ leaving $R_{1,3\pm}$ invariant and obviously commute with left multiplications. The full 16-dimensional real Clifford algebra could be represented by 2×2 quaternion matrices using for example w_1 and w_3 from (2.15) below as basis vectors of $R_{1,3+}$.[23,24,27] Because (2.14) does not commute with γ_1 or γ_2, two additional basis vectors must be introduced to represent the U_1 needed in Dirac theory. To obtain the familiar Pauli–Dirac matrix representation for the γ_μ's, we use the following w_Ω basis of $R_{1,3+}$ and to obtain the associated z axis oriented, positive and negative energy spinors $u_{A,\mathbf{p}}$ and $v_{A,\mathbf{p}}$, we use the associated basis z_Ω:

$$\begin{aligned}
&w_1 \equiv (I+\gamma_0)P_+, \quad z_1 \equiv (I+\gamma_0)P_+,\\
&w_2 \equiv (I+\gamma_0)\alpha_1 P_+, \quad z_2 \equiv (I+\gamma_0)\alpha_1 P_+,\\
&w_3 \equiv (I-\gamma_0)P_+, \quad z_3 \equiv (I-\gamma_0)\alpha_1 P_+,\\
&w_4 \equiv (I-\gamma_0)\alpha_1 P_+, \quad z_4 \equiv (I-\gamma_0)P_+.
\end{aligned} \tag{2.15}$$

Using

$$\gamma_\mu w_\Omega \equiv w_\Lambda \Gamma_\mu{}^{\hat{\Lambda}}_{\Omega}, \quad \text{where } (\Omega,\Lambda) = \{1,2,3,4\} \tag{2.16}$$

gives

$$\begin{aligned}
&\gamma_0 \cong \Gamma_0 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \gamma_i \cong \Gamma_i = \begin{pmatrix} 0 & -\sigma_i \\ \sigma_i & 0 \end{pmatrix},\\
&\gamma \cong \Gamma \equiv \Gamma_0\Gamma_1\Gamma_2\Gamma_3 = \begin{pmatrix} 0 & iI \\ iI & 0 \end{pmatrix},
\end{aligned} \tag{2.17}$$

and

$$\gamma_5(w_\Omega) = \gamma w_\Omega \gamma^{-1} \equiv w_\Lambda \Gamma_5{}^{\hat{\Lambda}}_{\Omega}, \Rightarrow \Gamma_5 = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix},$$

where in (2.17) I is the 2×2 identity, σ_i are the standard Pauli matrices, and γ on the right (defining the complex structure) has been replaced by multiplication of the imaginary unit "i" from the complex field. If we were interested in the Dirac wave theory we could and would now go to the above matrix representation; however, since we are interested in the Dirac quantum field theory, which requires a different complex structure (see Sec. IV), we keep the concrete basis picture w_Ω with right multiplication by γ. It should be observed that the representation matrices for the Clifford algebra (2.16) and (2.17) are unaffected by a change of basis $w_\Omega \to w_\Omega e^{\phi\gamma\alpha_3} = w_\Omega e^{\phi\gamma}$. This invariance constitutes the global U_1 invariance of the free electron–positron theory and it along with the following $\star$ innerautomorphism of $R_{1,3}$ are at the core of the projective representatives of time reversal and charge conjugation actions on spinors. Given a basis for $R_{1,3+}$ such as w_Ω, and its complex structure mapping γ, Pauli reversion ‡ of (2.7) can be decomposed into a commuting pair of involutions, * and ^{T}, called complex conjugation and transposition,

$$(\overset{\star}{c})^T = (c^T)^\star = c^\ddagger, \quad (c^T)^T = c, \quad (\overset{\star}{c})^\star = c,$$

constrained by

$$\overset{\star}{w}_\Omega = w_\Omega, \quad \overset{\star}{\gamma} = -\gamma, \quad \Rightarrow \gamma^T = \gamma. \tag{2.18}$$

Complex conjugation * is a pure innerautomorphism, and transposition is * followed by ‡. They are defined in terms of an element $C \in R_{1,3}$ that depends on the spinor basis,

$$\begin{aligned}
(\overset{\star}{c}) &= (C\gamma_0\gamma)c(C\gamma_0\gamma)^{-1} = (C\gamma_0)\gamma_5(c)(C\gamma_0)^{-1}\\
&= C\tilde{\overset{\ddagger}{c}}C^{-1},\\
c^T = \overset{\star}{c}{}^\ddagger &= C\tilde{\tilde{c}}\,C^{-1}, \Rightarrow C\gamma_\mu C^{-1} = -\gamma_\mu^T,
\end{aligned}$$

where

$$C^\ddagger = C^{-1}, \quad \tilde{\tilde{C}} = \pm C. \tag{2.19}$$

For the basis (2.15) we have

$$C = \pm\alpha_2 \Rightarrow \overset{\star}{\gamma}_0 = \gamma_0, \quad \overset{\star}{\gamma}_1 = \gamma_1, \quad \overset{\star}{\gamma}_3 = \gamma_3,$$

and

$$\overset{\star}{\gamma}_2 = -\gamma_2. \tag{2.20}$$

The constraints of (2.18) were placed on * so that the representative matrices (2.17) also satisfy (2.20) where * becomes complex conjugation of components.

The discrete spinor transformations of Dirac theory representing parity P, time reversal T, and charge conjugation C (written in capital bold Roman letters) are:

$$\begin{aligned}
&\mathbf{P}(c) \equiv \pm\gamma_0 c,\\
&\mathbf{T}(c) \equiv \tilde{c}^\ddagger C^{-1} e^{\phi_T\gamma\alpha_3} = -\gamma_0 c\gamma_2 e^{\phi_T\gamma\alpha_3} \quad (C = +\alpha_2),\\
&\mathbf{C}(c) \equiv \gamma c\gamma_0 C^{-1} e^{\phi_C\gamma\alpha_3} = -\gamma c\gamma_2 e^{\phi_C\gamma\alpha_3},
\end{aligned} \tag{2.21}$$

where ϕ_T and ϕ_C are arbitrary real constants. Of these three, only the parity representative could be guessed without prior knowledge of the Dirac wave theory. The choice of " $\pm$ " is left to convention, both generate the same connected part of the group and both have the same effect on vectors. The projective representative of time reversal is antilinear, satisfies $\mathbf{T}^2 = -I$, does not change parity or charge, but flips the spin eigenvalue. The projective spin representative of charge conjugation is antilinear, satisfies $\mathbf{C}^2 = I$, flips the parity and the charge, but not the spin. We give the matrix representatives of the discrete transformations using the z_Ω basis (2.15) rather than the w_Ω basis because we need them in Sec. III:

$$\mathbf{P}(z_\Omega) \equiv z_\Lambda \mathbf{P}^\Lambda_\Omega, \quad \mathbf{T}(z_\Omega) \equiv z_\Lambda \mathbf{T}^\Lambda_\Omega, \quad \mathbf{C}(z_\Omega) \equiv z_\Lambda \mathbf{C}^\Lambda_\Omega,$$

where

$$\mathbf{P} = \pm \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} \sigma_2 & 0 \\ 0 & -\sigma_2 \end{pmatrix},$$
$$\mathbf{C} = \begin{pmatrix} 0 & -\sigma_3 \\ -\sigma_3 & 0 \end{pmatrix}. \tag{2.22}$$

In the above we have described a representation of a group we call the homogeneous group H without discussing the structure of H itself. Our point of view is that H consists of eight disconnected parts, the identity component being $\mathrm{Sl}(2,\mathrm{C}) \times \mathrm{U}_1$ and the other seven given by products of $\mathrm{Sl}(2,\mathrm{C}) \times \mathrm{U}_1$ and one or more of P, T, and C. The structure of this homogeneous electron–positron invariance group is a direct product of $\mathrm{Pin}_{1,3}$ and the gauge group G, which consists of the phase rotations U_1 and charge conjugaton C,

$$H = \mathrm{Pin}_{1,3} \times G. \tag{2.23}$$

The $\mathrm{Pin}_{1,3}$ group consists of the four disconnected parts described below (2.10) and the $G \cong \mathrm{U}_1 \circledS \{I,\mathrm{C}\}$ group consists of two parts, giving eight all together. In G the semidirect product action of C on U_1 is $u \to u^{-1}$. The spin representatives of H are just the transformations of $R_{1,3}$ generated by products of left multiplication by (2.10), right multiplication by (2.14), and (2.22) actions; and leave invariant a Hermitian inner product on $R_{1,3}$ (considered as an eight-dimensional complex representation space with γ on the right defining the complex structure). The Hermitian (Dirac) inner product is the familiar one constructed using Pauli reversion,

$$\langle c_1, c_2 \rangle \equiv (c_1^\dagger \gamma_0 c_2)_{s+ps} = (c_2^\dagger \gamma_0 c_1)^\ddagger_{s+ps} = \langle c_2, c_1 \rangle^\ddagger, \tag{2.24}$$

where $s+ps$ stands for scalar and pseudoscalar (i.e., $V^0 \oplus V^4$) parts only. When † is applied to the unit pseudoscalar γ in (2.2) it changes its sign, i.e., changes i to $-i$ as required of Hermitian inner products. In other words, (2.24) says that to compute the Hermitian inner product of c_1 and c_2 considered as two eight-dimensional complex vectors in $R_{1,3}$, use the 16-dimensional real Clifford algebra multiplication to evaluate $c_1^\dagger \gamma_0 c_2$ and keep only the scalar and pseudoscalar parts, remembering that the unit pseudoscalar γ of the Clifford algebra, when acting on the right, is equivalent to multiplying by the unit imaginary "i" of the complex field. We observe that because $\langle R_{1,3+}, R_{1,3-} \rangle = 0$, the Hermitian inner product provides an inner product on each spinor subspace separately, i.e., on $R_{1,3+}$; it is the familiar Dirac $\bar{\Psi}\Phi$,

$$\langle w_\Omega, w_\Lambda \rangle = \langle z_\Omega, z_\Lambda \rangle = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \tag{2.25}$$

where z_Ω and w_Ω are basis vectors of (2.15). We also observe that when the Boosts are excluded from the homogeneous group, invariance of (2.24) is equivalent to invariance of the Hilbert space inner product

$$\langle\langle c_1, c_2 \rangle\rangle \equiv (c_1^\dagger c_2)_{s+ps} = (c_2^\dagger c_1)^\ddagger_{s+ps} = \langle\langle c_2, c_1 \rangle\rangle^\ddagger, \tag{2.26}$$

which on $R_{1,3+}$ is simply

$$\langle\langle z_\Omega, z_\Lambda \rangle\rangle = \delta_{\Omega\Sigma}. \tag{2.27}$$

With respect to this inner product, the identity and parity component representatives are unitary whereas the time reversal and charge conjugation component representatives are projective and antiunitary.

For a global picture we wish to look at the homogeneous group H as the Lie group of a trivial principal fiber bundle HB over M^4,

$$HB \cong M^4 \times H \to M^4. \tag{2.28}$$

We call this the homogeneous bundle and think of it as an enlargement of the bundle of orthonormal frames for M^4 (whose global gauge group is the homogeneous Lorentz group and whose fibers consist of only four disconnected parts). We take this point of view primarily to avoid double-valued representations. It is sometimes beneficial (but incorrect) to think of HB as the spinor frame bundle. One obvious incorrectness occurs because of the projective action of the T and C components on the spinor frames. The invariance group P, of Dirac's free electron–positron theory is a semidirect product of space-time translations R^4 and the homogeneous group H,

$$P = R^4 \circledS H, \tag{2.29}$$

and can be thought of as the Poincare group enlarged to remove double-valued representations of the Lorentz group as well as to include U_1 and charge conjugation C. In (2.29), H is the isotropy of the origin ($x = 0$). The inhomogeneous group acts as fiber preserving mappings on HB, $(r,h) \in P$ acts on $(x,h') \in HB$ by

$$(x,h') \to (r + h(x), hh'), \tag{2.30}$$

where the h action on x is the expected vector action for $h \in \mathrm{Pin}_{1,3}$ and is ineffectual for $h \in G$. Here, P is seen to act as a group of bundle automorphisms because its action commutes with the H action of HB. The isotropy subgroup $H_x \subset P$ is isomorphic to H and consists of those inhomogeneous transformations that leave $x \in M^4$ invariant,

$$H_x \equiv \{(x - h(x), h) \in P\} \cong H. \tag{2.31}$$

III. WIGNER'S INDUCED REPRESENTATION PROCEDURE

Dirac electron–positron theory contains another representation of the semidirect product (2.29) beyond the four-component spin representation given in Sec. II. In this section we use Wigner's procedure for constructing induced

representations to construct this second representation that is in fact faithful, irreducible, and unitary[9–14] We are necessarily careful to keep track of the gauge group action and the discrete transformations. In Sec. IV we apply Segal's quantization procedure to this infinite-dimensional representation and arrive at free positron–electron quantum field theory.

To construct a unitary representation of the inhomogeneous group $R^4 \circledS H$, Wigner's procedure requires first the selection of a one-dimensional representation of the translations (often called a character);

$$\chi: R^4 \to U_1 \Leftrightarrow \chi(r) = e^{-\gamma p_\mu r^\mu} = e^{-\gamma mcr^0}, \tag{3.1}$$

to which we have already adapted a global frame e_μ (fixes e_0 only). We have selected a character appropriate for a massive particle and, by using the complex structure mapping γ rather than imaginary field unit "i," indicated that the translations will act on the right of the four-component spinor space $R_{1,3+}$. The subgroup $L \subset H$ whose vector action leaves e_0 (and hence χ) invariant is called the Little group and is generated by products of spatial rotations $SU_2 \subset SL(2,C)$, U_1 phase rotations, parity P, and charge conjugation C. The Little group L thus consists of four disconnected parts: the indentity component $SU_2 \times U_1$ and three other components generated by multiplications with P and C. The second needed ingredient in the Wigner construction is an irriducible unitary representation of L. In the electron–positron case this representation is given by the $L \subset H$ actions on $R_{1,3+}$ described in Sec. II. Even though, as a representatin of H, the four-component spin representation is not unitary, as a representation of the subgroup L, it is [see (2.26)]. Under L actions, $R_{1,3+}$ decomposes into the direct sum of two orthogonal two-component Pauli spinor subspaces of opposite parity,

$$\begin{aligned} R_{1,3+} &= {}_+R_{1,3} \oplus {}_-R_{1,3+}, \\ {}_\pm R_{1,3+} &\equiv [(I \pm \gamma_0)/2] R_{1,3+}. \end{aligned} \tag{3.2}$$

Each is invariant under SU_2 and parity P [P of (2.21)] but is exchanged by the action of charge conjugation C [C of (2.21)]. This is easily seen by choosing (z_1,z_2) and (z_3,z_4) of (2.15) as respective pairs of basis vectors.

The next step in the Wigner construction is to define the infinite-dimensional complex vector space $\mathscr{H}$ of functions from $H/L \cong$ Boosts (part connected to coset L only) into the representation space for L, i.e., into $R_{1,3+}$. Time reversal actions are defined on these functions. The Boost actions on translations R^4 pull back to actions on the set of characters and make the Boosts topologically equivalent to the upper mass shell. In particular,

$$e^{(\zeta^i/2)\alpha_i} \to e^{-\gamma mc \Lambda^0_\mu r^\mu} = e^{-\gamma p_\mu r^\mu}, \tag{3.3}$$

where the Boost parameters ζ^i are related to the mass shell point by (2.11),

$$e^{(\zeta^i/2)\alpha_i} \gamma_0 e^{-(\zeta^i/2)\alpha_i} = \gamma_\mu \Lambda^\mu_0 = \gamma_\mu p^\mu / mc. \tag{3.4}$$

Consequently, $\mathscr{H}$ is equivalent to functions from the upper mass shell to $R_{1,3+}$, i.e., four-component spinor valued functions of p^μ, $\psi(p^\mu) = z_\Omega \psi^\Omega(p^\mu)$. Notice that the components $\psi^\Omega(p^\mu)$ appear on the right of the basis vectors z_Ω because the complex structure has been defined as right multiplication by γ.

The induced action $U_{(r,h)}$ of $(r,h) \in R^4 \circledS H$ on $\psi \in \mathscr{H}$ is, for $s \in SL(2,C) \subset H$,

$$[U_{(0,s)}\psi](p^\nu) = e^{-(\zeta'^i/2)\alpha_i} S e^{(\zeta'^i/2)\alpha_i} \psi(\Lambda^{-1\nu}{}_\lambda p^\lambda), \tag{3.5}$$

where S is the four-component spin representative of s [see (2.10)], $\Lambda^{-1\nu}{}_\lambda$ is related to S by (2.11), and ζ'^i is related to $p'^\nu \equiv \Lambda^{-1\nu}{}_\lambda p^\lambda$ by (3.4). The combination

$$W(s,\mathbf{p}) \equiv e^{-(\zeta'^i/2)\alpha_i} S e^{(\zeta'^i/2)\alpha_i} = e^{-\gamma(\omega/2)\cdot\alpha} \tag{3.6}$$

is commonly referred to as a Wigner rotation with $\omega(s,\mathbf{p})$ being the three rotation angles and is represented by a direct sum of two SU_2 rotations. With respect to the z_Ω basis of (2.15) we have the matrix representation,

$$\begin{aligned} &W(s,\mathbf{p}) z_\Omega = z_\Lambda (W)^\Lambda_\Omega, \\ &(W) = \begin{pmatrix} D^{1/2} & 0 \\ 0 & \sigma_1 D^{1/2} \sigma_1 \end{pmatrix}, \\ &D^{1/2} \equiv (e^{-i\omega\cdot\sigma/2}). \end{aligned} \tag{3.7}$$

Here σ are the 2×2 Pauli matrices. For $u(\phi) \in U_1$, the action is

$$[U_{(0,u)}\psi](p^\nu) = \psi(p^\nu) e^{\gamma\phi}. \tag{3.8}$$

The action of parity P is identical to (3.5) [see (2.21)],

$$[U_{(0,P)}\psi](p^0,\mathbf{p}) = \pm \gamma_0 \psi(p^0, -\mathbf{p}). \tag{3.9}$$

The action of time reversal T follows (2.21),

$$\begin{aligned} [A_{(0,T)}\psi](p^0,\mathbf{p}) &\equiv \tilde{\psi}^\dagger(p^0,\mathbf{p}) C^{-1} e^{\phi_T \gamma\alpha_3}, \\ &= -\gamma_0 \psi(p^0, -\mathbf{p}) \gamma_2 e^{\phi_T \gamma\alpha_3}, \end{aligned} \tag{3.10}$$

as does the action of charge conjugation C,

$$\begin{aligned} [A_{(0,C)}\psi](p^\nu) &\equiv \gamma\psi(p^\nu)\gamma_0 C^{-1} e^{\phi_C \gamma\alpha_3} \\ &= -\gamma\psi(p^\nu)\gamma_2 e^{\phi_C \gamma\alpha_3}. \end{aligned} \tag{3.11}$$

Completing the induced representation, we use the character and have for translation $r \in R^4$,

$$\begin{aligned} [U_{(r,I)}\psi](p^\nu) = {}& [(I+\gamma_0)/2]\psi(p^\nu) e^{-\gamma p_\mu r^\mu} \\ &+ [(I-\gamma_0)/2]\psi(p^\nu) e^{+\gamma p_\mu r^\mu}, \end{aligned} \tag{3.12}$$

i.e., the two parity components are phase rotated oppositely by a translation.

The Hilbert space inner product on $\mathscr{H}$ for which (3.5)–(3.9), and (3.12) are unitary and (3.10) and (3.11) are antiunitary is

$$\langle\langle \psi,\phi \rangle\rangle_{\mathscr{H}} \equiv \int dp \langle\langle \psi(p),\phi(p) \rangle\rangle, \tag{3.13}$$

where

$$dp \equiv (2\pi)^{-3} (mc/p^0) d^3p,$$

and where $\langle\langle \psi(p),\phi(p) \rangle\rangle$ is defined in (2.26). The invariant volume element is dp and the integration domain is the entire upper mass shell. To make clear the details of Wigner's induced unitary representation as well as to proceed with Segal quantization in Sec. IV, we introduce the basis functions $a_{A,\mathbf{q}}$ and $c_{A,\mathbf{q}}$:

$$a_{A,\mathbf{q}}(\mathbf{p}) \equiv (2\pi)^3 (q_0/mc)\delta^3(\mathbf{p}-\mathbf{q}) z_A,$$
$$c_{A,\mathbf{q}}(\mathbf{p}) \equiv (2\pi)^3 (q_0/mc)\delta^3(\mathbf{p}-\mathbf{q}) z_{A+2}, \tag{3.14}$$

where $q_0 \equiv +[m^2c^2 + \mathbf{q}\cdot\mathbf{q}]^{1/2}$ and $A = \{1,2\}$. The normalization has been chosen so that

$$\langle\langle a_{A,\mathbf{q}}, a_{B,\mathbf{p}}\rangle\rangle_{\mathscr{H}} = (2\pi)^3 (q_0/mc)\delta_{AB}\delta^3(\mathbf{q}-\mathbf{p}),$$
$$\langle\langle c_{A,\mathbf{q}}, c_{B,\mathbf{p}}\rangle\rangle_{\mathscr{H}} = (2\pi)^3 (q_0/mc)\delta_{AB}\delta^3(\mathbf{q}-\mathbf{p}),$$
$$\langle\langle a_{A,\mathbf{q}}, c_{B,\mathbf{p}}\rangle\rangle_{\mathscr{H}} = 0. \tag{3.15}$$

The above group actions (3.8)–(3.12) on $a_{A,\mathbf{q}}$ and $c_{A,\mathbf{q}}$ are

$$U_{(0,s)} a_{A,\mathbf{q}} = a_{B,\Lambda\mathbf{q}} D^{(1/2)B}{}_A[\omega(s,\Lambda\mathbf{q})],$$
$$U_{(0,s)} c_{A,\mathbf{q}} = c_{D,\Lambda\mathbf{q}} \sigma^E_{1C} D^{(1/2)C}{}_B[\omega(s,\Lambda\mathbf{q})]\sigma^B_{1A}, \tag{3.16}$$

where $j = \frac{1}{2}$ representation matrices $D^{1/2}$ are defined by (3.6) and (3.7):

$$U_{(0,u)} a_{A,\mathbf{q}} = a_{A,\mathbf{q}} e^{+\gamma\phi}, \quad A_{(0,\mathrm{T})} a_{A,\mathbf{q}} = a_{B,-\mathbf{q}} \sigma^B_{2A} e^{+\gamma\phi_T},$$
$$U_{(0,u)} c_{A,\mathbf{q}} = c_{A,\mathbf{q}} e^{+\gamma\phi}, \quad A_{(0,\mathrm{T})} c_{A,\mathbf{q}} = -c_{B,-\mathbf{q}} \sigma^B_{2A} e^{+\gamma\phi_T},$$
$$U_{(0,\mathrm{P})} a_{A,\mathbf{q}} = \pm a_{A,-\mathbf{q}}, \quad A_{(0,\mathrm{C})} a_{A,\mathbf{q}} = -c_{B,-\mathbf{q}} \sigma^B_{3A} e^{+\gamma\phi_C}, \tag{3.17}$$
$$U_{(0,\mathrm{P})} c_{A,\mathbf{q}} = \mp c_{A,-\mathbf{q}}, \quad A_{(0,\mathrm{C})} c_{A,\mathbf{q}} = -a_{B,-\mathbf{q}} \sigma^B_{3A} e^{+\gamma\phi_C},$$

follow from (2.22) and complete the unitary and projective antiunitary actions of the homogeneous group on the basis functions we use for $\mathscr{H}$. Now,

$$U_{(r,I)} a_{A,\mathbf{q}} = a_{A,\mathbf{q}} e^{-\gamma q_\mu r^\mu}, \quad U_{(r,I)} c_{A,\mathbf{q}} = c_{A,\mathbf{q}} e^{+\gamma q_\mu r^\mu}, \tag{3.18}$$

give the unitary actions of the translations.

The action of the isotropy subgroup $H_x \cong H$ of the point $x \in M^4$ on the basis $a_{A,\mathbf{q}}, c_{A,\mathbf{q}}$, see (2.31), can easily be constructed by applying a homogeneous transformation $h \in H \cong H_{x=0}$, e.g., (3.16) and (3.17) followed by (3.18) with $r = x - h(x)$.

IV. SEGAL QUANTIZING WIGNER'S INDUCED REPRESENTATION

In this section we construct the complex Clifford algebra $\mathscr{C}$ associated with the Hilbert space $\mathscr{H}$ in Sec. III. This is the algebra of annihilation and creation operators of Dirac's electron–positron theory. To construct this algebra we follow the procedure described by Shale and Stinespring and frequently called Segal quantization.[15–21,28–30] The procedure starts by identifying the complex space $\mathscr{H}$ with a real Hilbert space $\mathscr{H}_R$ possessing a symmetric inner product, followed by the construction of its associated Clifford algebra $\mathscr{C}_R$ according to the prescription given in (2.1). This real infinite-dimensional Clifford algebra, when complexified, becomes the desired operator algebra $\mathscr{C}$. Because this is the second vector space→Clifford algebra construction required to obtain a quantum field theory of electrons and positrons, we call it second Cliffordization.

In the first step of second Cliffordization, complex vectors $\psi, \phi \in \mathscr{H}$, are mapped, respectively, one-to-one onto real vectors $\psi_R, \phi_R \in \mathscr{H}_R$ in such a manner as to relate real and complex inner products by

$$(\phi_R, \psi_R)_R \equiv (\langle\langle\phi,\psi\rangle\rangle_{\mathscr{H}} + \langle\langle\psi,\phi\rangle\rangle_{\mathscr{H}})/2. \tag{4.1}$$

Multiplying $\psi \in \mathscr{H}$ by the unit imaginary number γ does not give an independent vector $\psi\gamma \in \mathscr{H}$; however, their images in $\mathscr{H}_R, \psi_R$, and $(\psi\gamma)_R$ are not only independent but, according to (4.1) are orthogonal. Multiplication by γ in $\mathscr{H}$ induces a complex structure on $\mathscr{H}_R$, $\gamma_R | \mathscr{H}_R \to \mathscr{H}_R$ defined by

$$\gamma_R[\psi_R] \equiv (\psi\gamma)_R, \tag{4.2}$$

and satisfies the required $\gamma_R\gamma_R = -I$. Every complex linear transformation of $\mathscr{H}$ induces a corresponding real linear transformation $\mathscr{H}_R$ that commutes with this complex structure. In particular the unitary invariance group $\mathbf{U}_{\mathscr{H}}$ of $\langle\langle,\rangle\rangle_{\mathscr{H}}$ is easily seen to be isomorphic to the subgroup of $(,)_R$ invariant orthogonal transformations $\mathbf{Q}_{R,\gamma}$ that commute with γ_R, i.e., every $U_{\mathscr{H}} \in \mathbf{U}_{\mathscr{H}}$ satisfies $U_{\mathscr{H}} \to O_R$ for some O_R provided

$$(Q_R\phi_R, O_R\psi_R)_R = (\psi_R, \phi_R)_R \quad \text{for all } \phi_R, \psi_R \in \mathscr{H}_R$$

and

$$\gamma_R O_R = O_R \gamma_R, \quad \text{i.e., provided } O_R \in \mathbf{O}_{R,\gamma}. \tag{4.3}$$

Notice that $\gamma \leftrightarrow \gamma_R$ belongs to this isomorphism. In terms of the basis functions (3.14) of $\mathscr{H}$, the one-to-one $\mathscr{H} \leftrightarrow \mathscr{H}_R$ mapping appears as

$$a_{A,\mathbf{q}} \leftrightarrow a_{RA,\mathbf{q}}, \quad a_{A,\mathbf{q}}\gamma \leftrightarrow \gamma_R[a_{RA,\mathbf{q}}],$$
$$c_{A,\mathbf{q}} \leftrightarrow c_{RA,\mathbf{q}}, \quad c_{A,\mathbf{q}}\gamma \leftrightarrow \gamma_R[c_{RA,\mathbf{q}}], \tag{4.4}$$

and from (3.15) and (4.1) the symmetric real inner product of basis vectors becomes

$$(a_{RA,\mathbf{q}}, a_{RB,\mathbf{p}})_R = (c_{RA,\mathbf{q}}, c_{RB,\mathbf{p}})_R$$
$$= (2\pi)^3 (q_0/mc)\delta_{AB}\delta^3(\mathbf{q}-\mathbf{p}),$$
$$(\gamma_R[a_{RA,\mathbf{q}}], \gamma_R[a_{RB,\mathbf{p}}])_R = (\gamma_R[c_{RA,\mathbf{q}}], \gamma_R[c_{RB,\mathbf{p}}])_R$$
$$= (2\pi)^3 (q_0/mc)\delta_{AB}\delta^3(\mathbf{q}-\mathbf{p}),$$
$$(a_{RA,\mathbf{q}}, c_{RB,\mathbf{p}})_R = (\gamma_R[a_{RA,\mathbf{q}}], \gamma_R[c_{RB,\mathbf{p}}])_R = \text{etc.} = 0. \tag{4.5}$$

The real vector space $\mathscr{H}_R$ plays the same role as the four-dimensional Lorentz inner product tangent space of M^4 plays in Sec. II, and the above orthonormal basis plays the role of the e_μ. The unitary representation $P \to \mathbf{U}_{\mathscr{H}}$, of the inhomogeneous group (2.29), whose action on the basis vectors (3.14) of $\mathscr{H}$ is given by (3.16) to (3.18) and preserves (3.15), would now appear as an orthogonal representation appropriately transforming the basis vectors (4.4) while preserving (4.5).

Following (2.1) an infinite-dimensional Clifford algebra $\mathscr{C}_R$ can be constructed from $\mathscr{H}_R$. We write the isomorphic image of $\mathscr{H}_R$ in $\mathscr{C}_R$ as $\mathscr{C}^1_R$, and write the image vectors in boldface rather than with a "1" superscript as in (2.1). For example, the $\mathscr{H}_R \leftrightarrow \mathscr{C}^1_R$ mapping of basis vectors is written as

$$a_{RA,\mathbf{q}} \leftrightarrow \mathbf{a}_{A,\mathbf{q}}, \quad \gamma_R[a_{RA,\mathbf{q}}] \leftrightarrow \gamma[\mathbf{a}_{A,\mathbf{q}}],$$
$$c_{RA,\mathbf{q}} \leftrightarrow \mathbf{c}_{A,\mathbf{q}}, \quad \gamma_R[c_{RA,\mathbf{q}}] \leftrightarrow \gamma[\mathbf{c}_{A,\mathbf{q}}]. \tag{4.6}$$

This basis identification is equivalent to $e_\mu \leftrightarrow \gamma_\mu$ for M^4. The Clifford algebra $\mathscr{C}_R$ is generated by all real linear combinations of products of $\mathscr{C}^1_R$ basis vectors, $\{\mathbf{a}_{A,\mathbf{q}}, \mathbf{c}_{A,\mathbf{q}}, \gamma[\mathbf{a}_{A,\mathbf{q}}], \gamma[\mathbf{c}_{A,\mathbf{q}}]\}$, and can be expressed as

$$\mathscr{C}_R = \mathscr{C}^0_R \oplus \mathscr{C}^1_R \oplus \mathscr{C}^2_R \oplus \mathscr{C}^3_R \oplus \cdots, \tag{4.7}$$

where $\mathscr{C}^0_R$ stands for all real multiples of the identity $\mathscr{I}$, $\mathscr{C}^1_R$

for all real linear combinations of basis vectors $\{\mathbf{a}_{A,\mathbf{q}},\mathbf{c}_{A,\mathbf{q}},\gamma[\mathbf{a}_{A,\mathbf{q}}],\gamma[\mathbf{c}_{A,\mathbf{q}}]\}$, $\mathscr{C}^2_R$ for all real linear combinations of products of pairs of basis vectors with no two pairs being equal, e.g., $\mathbf{a}_{A,\mathbf{q}}\mathbf{a}_{B,\mathbf{p}}$ where $A\neq B$ or $\mathbf{q}\neq\mathbf{p}$, $\mathscr{C}^3_R$ for all real linear combinations of products of triples of basis vectors with no two being equal $\cdots$, etc. Linear combinations include integrations over the continuous mass shell indices $\mathbf{q}$, $\mathbf{p}$, etc. The Clifford algebraic multiplication constraint (iii) of (2.1), evaluated using (4.5), appears as

$$\begin{aligned}&\mathbf{a}_{A,\mathbf{q}}\mathbf{a}_{B,\mathbf{p}}+\mathbf{a}_{B,\mathbf{p}}\mathbf{a}_{A,\mathbf{q}}\\&\quad=\gamma[\mathbf{a}_{A,\mathbf{q}}]\gamma[\mathbf{a}_{B,\mathbf{p}}]+\gamma[\mathbf{a}_{B,\mathbf{p}}]\gamma[\mathbf{a}_{A,\mathbf{q}}]\\&\quad=2(2\pi)^3(q_0/mc)\delta_{AB}\delta^3(\mathbf{q}-\mathbf{p})\mathscr{I},\\&\mathbf{a}_{A,\mathbf{q}}\mathbf{c}_{B,\mathbf{p}}+\mathbf{c}_{B,\mathbf{p}}\mathbf{a}_{A,\mathbf{q}}\\&\quad=\gamma[\mathbf{a}_{A,\mathbf{q}}]\gamma[\mathbf{c}_{B,\mathbf{p}}]+\gamma[\mathbf{c}_{B,\mathbf{p}}]\gamma[\mathbf{a}_{A,\mathbf{q}}]\\&\quad=\text{etc.}=0.\end{aligned}\tag{4.8}$$

Because $\mathscr{H}_R$ is isomorphic to $\mathscr{C}^1_R$ every orthogonal transformation of $\mathscr{H}_R$ in $\mathbf{O}_{R,\gamma}$ corresponds to an orthogonal transformation of $\mathscr{C}^1_R$. In particular, the complex structure mapping $\gamma_R|\mathscr{H}_R\to\mathscr{H}_R$ corresponds to the mapping $\gamma|\mathscr{C}^1_R\to\mathscr{C}^1_R$ as indicated in (4.6). Since $\mathscr{C}^1_R$ generates $\mathscr{C}_R$, an orthogonal transformation O_R of $\mathscr{C}^1_R$ can be extended to an algebra automorphism $O\in\mathbf{O}_\gamma\subset\mathrm{Aut}(\mathscr{C}_R)$ of $\mathscr{C}_R$ by simply requiring $O(\mathbf{c}_1\mathbf{c}_2)=O(\mathbf{c}_1)O(\mathbf{c}_2)$ and $O(\mathbf{c}^1)=O_R(\mathbf{c}^1)$ when $\mathbf{c}^1\in\mathscr{C}^1_R$. As an example, the γ mapping extends to all of $\mathscr{C}_R$ as an algebra automorphism; e.g., when γ is applied to the identity $\mathscr{I}$ of $\mathscr{C}_R$, it gives $\mathscr{I}$ back. Consequently, γ does not satisfy the required $\gamma\gamma=-I$ to provide a complex structure for all of $\mathscr{C}_R$. However, $\mathscr{C}_R$ can be complexified by taking complex (rather than real) linear combinations as in (4.7); the resulting complex algebra $\mathscr{C}$ turns out to be the desired electron–positron operator algebra,

$$\mathscr{C}=\mathscr{C}^0\oplus\mathscr{C}^1\oplus\mathscr{C}^2\oplus\mathscr{C}^3\oplus\cdots.\tag{4.9}$$

We will simply denote this complex structure by multiplying by the unit imaginary number "i" and * as complex conjugation. The automorphisms $\mathbf{O}_\gamma$ of $\mathscr{C}_R$ can be extended to $\mathscr{C}$ by simply requiring that they commute with the new complex structure. In particular, γ extends from $\mathscr{C}_R$ to $\mathscr{C}$ by requiring $\gamma i=i\gamma$. Here $\mathscr{C}^1$ represents the space of all one-particle (electron–positron) annihilation and creation operators. Rather than using $\{\mathbf{a}_{B,\mathbf{p}},\mathbf{c}_{B,\mathbf{p}},\gamma[\mathbf{a}_{B,\mathbf{p}}],\gamma[\mathbf{c}_{B,\mathbf{p}}]\}$ as basis vectors, the more conventional set $\{\mathbf{b}_{B\mathbf{p}},\mathbf{b}^\dagger_{B\mathbf{p}},\mathbf{d}_{B\mathbf{p}},\mathbf{d}^\dagger_{B\mathbf{p}}\}$ can be used,

$$\begin{aligned}&\mathbf{b}_{B\mathbf{p}}\equiv\tfrac{1}{2}(\mathbf{a}_{B,\mathbf{p}}-i\gamma[\mathbf{a}_{B,\mathbf{p}}]),\quad \mathbf{b}^\dagger_{B\mathbf{p}}\equiv\tfrac{1}{2}(\mathbf{a}_{B,\mathbf{p}}+i\gamma[\mathbf{a}_{B,\mathbf{p}}]),\\&\mathbf{d}^\dagger_{B\mathbf{p}}\equiv\tfrac{1}{2}(\mathbf{c}_{B,\mathbf{p}}-i\gamma[\mathbf{c}_{B,\mathbf{p}}]),\quad \mathbf{d}_{B\mathbf{p}}\equiv\tfrac{1}{2}(\mathbf{c}_{B,\mathbf{p}}+i\gamma[\mathbf{c}_{B,\mathbf{p}}]).\end{aligned}\tag{4.10}$$

The conjugate linear mapping $^\dagger|\mathscr{C}^1\to\mathscr{C}^1$ defined by (4.10) acts as the identity on $\mathscr{C}^1_R$ commuting with γ but anti commuting with "i" multiplication. Because $\mathscr{C}^1$ generates $\mathscr{C}$, the † mapping can be immediately extended to a unique antiautomorphism of $\mathscr{C}$ by requiring that

$$\begin{aligned}&(\mathbf{c}_1+\mathbf{c}_2)^\dagger=\mathbf{c}_1^\dagger+\mathbf{c}_2^\dagger,\\&(\mathbf{c}_1\mathbf{c}_2)^\dagger=\mathbf{c}_2^\dagger\mathbf{c}_1^\dagger,\quad \mathbf{c}_1\mathbf{c}_2\in\mathscr{C},\\&\mathbf{c}_0^\dagger=(r+\overset{*}{ir'})\mathscr{I},\quad \mathbf{c}_0=(r+ir')\mathscr{I}\in\mathscr{C}^0.\end{aligned}\tag{4.11}$$

The fundamental antiautomorphism † is just the analog of reversion (2.5) for real Clifford algebras. The subset of self-conjugate elements ($\mathbf{c}^\dagger=\mathbf{c}$) is just those identified with the original Hilbert space $\mathscr{H}\cong\mathscr{H}_R$. A complex algebra such as $\mathscr{C}$ with such an involuting, $(\mathbf{c}^\dagger)^\dagger=\mathbf{c}$, antiautomorphism is called a $\star$ algebra and with appropriate norm and completion becomes a $C^\star$ algebra.[31] To make contact with the "CAR" algebra construction approach, one has only to complexify $\mathscr{H}_R$ with imaginary unit "i" by identifying it with $\mathscr{C}^1$. Then $\mathscr{C}$ is the CAR algebra of this complex Hilbert space.

In (4.10) careful attention must be paid to the indices $A,B=\{1,2\}$ being up or down. This index must be raised and lowered with the Pauli spinor metric, e.g., with

$$\begin{aligned}&\epsilon^{AB}=\epsilon_{AB}\equiv i\sigma^A_{2B}=\begin{pmatrix}0&1\\-1&0\end{pmatrix},\\&\epsilon^{AC}\epsilon_{BC}=\delta^A_B.\end{aligned}\tag{4.12}$$

The operators appearing in the Dirac fields $\Psi(x)$ and $\Psi^\dagger(x)$ are with the "A" index "up,"

$$\begin{aligned}&\mathbf{b}^A_{\mathbf{p}}\equiv\epsilon^{AB}\mathbf{b}_{B\mathbf{p}},\quad \mathbf{b}^{A\,\dagger}_{\mathbf{p}}\equiv\epsilon^{AB}\mathbf{b}^\dagger_{B\mathbf{p}},\\&\mathbf{b}_{B\mathbf{p}}=\epsilon_{AB}\mathbf{b}^A_{\mathbf{p}},\quad \mathbf{b}^\dagger_{B\mathbf{p}}=\epsilon_{AB}\mathbf{b}^{A\,\dagger}_{\mathbf{p}},\quad\text{etc.},\end{aligned}\tag{4.13}$$

and from (4.8) satisfy the required anticommutation relations,

$$\begin{aligned}&\mathbf{b}^A_{\mathbf{q}}\mathbf{b}^{B\,\dagger}_{\mathbf{p}}+\mathbf{b}^{B\,\dagger}_{\mathbf{p}}\mathbf{b}^A_{\mathbf{q}}=\mathbf{d}^A_{\mathbf{q}}\mathbf{d}^{B\,\dagger}_{\mathbf{p}}+\mathbf{d}^{B\,\dagger}_{\mathbf{p}}\mathbf{d}^A_{\mathbf{q}}\\&\qquad=(2\pi)^3(q_0/mc)\delta_{AB}\delta^3(\mathbf{q}-\mathbf{p})\mathscr{I},\\&\mathbf{b}^A_{\mathbf{q}}\mathbf{b}^B_{\mathbf{p}}+\mathbf{b}^B_{\mathbf{p}}\mathbf{b}^A_{\mathbf{q}}=\mathbf{d}^A_{\mathbf{q}}\mathbf{d}^B_{\mathbf{p}}+\mathbf{d}^B_{\mathbf{p}}\mathbf{d}^A_{\mathbf{q}}=\text{etc.}=0.\end{aligned}\tag{4.14}$$

The γ action on the conventional basis (4.10) is (using $\gamma\gamma=-I$ on $\mathscr{C}^1$)

$$\begin{aligned}&\gamma[\mathbf{b}_{B\mathbf{p}}]=i\mathbf{b}_{B\mathbf{p}},\quad \gamma[\mathbf{d}^\dagger_{B\mathbf{p}}]=i\mathbf{d}^\dagger_{B\mathbf{p}},\\&\gamma[\mathbf{b}^\dagger_{B\mathbf{p}}]=-i\mathbf{b}^\dagger_{B\mathbf{p}},\quad \gamma[\mathbf{d}_{B\mathbf{p}}]=-i\mathbf{d}_{B\mathbf{p}}.\end{aligned}\tag{4.15}$$

We are now in a position to compute explicitly the extension of the inhomogeneous group's unitary action $\mathbf{U}_{\mathscr{H}}$ (3.16)–(3.18) on $\mathscr{H}$ to an $\mathbf{O}_\gamma$ action on $\mathscr{C}^1$. By applying $U_{(0,s)}$ of (3.16) to (4.10) and using (4.13), the identity component of the homogeneous group acts according to

$$\begin{aligned}O_{(0,s)}\mathbf{b}_{A\mathbf{q}}&=D^{(1/2)}(\omega)^B{}_A\mathbf{b}_{B\Lambda\mathbf{q}},\Leftrightarrow O_{(0,s)}\mathbf{b}^A_{\mathbf{q}}\\&=D^{(1/2)}(-\omega)^A{}_B\mathbf{b}^B_{\Lambda\mathbf{q}},\\O_{(0,s)}\mathbf{d}^\dagger_{A\mathbf{q}}&=\sigma^D_{1C}D^{(1/2)}(\omega)^C{}_B\sigma^B_{1A}\mathbf{d}^\dagger_{D\Lambda\mathbf{q}},\Leftrightarrow O_{(0,s)}\mathbf{d}^{A\,\dagger}_{\mathbf{q}}\\&=\sigma^A_{1B}D^{(1/2)}(-\omega)^B{}_C\sigma^C_{1D}\mathbf{d}^{D\,\dagger}_{\Lambda\mathbf{q}},\end{aligned}$$

etc., where

$$D^{(1/2)}(\omega)^B_A\equiv D^{(1/2)B}{}_A[\omega(s,\Lambda\mathbf{q})]$$

satisfies

$$(\sigma_2D^{1/2}(\omega)\sigma_2)^B{}_A=\overset{*}{D}{}^{1/2}(\omega)^B{}_A=D^{1/2}(-\omega)^A{}_B.\tag{4.16}$$

We have used O's to represent these linear transformations because they are orthogonal on $\mathscr{C}^1_R$ (and would be real unitary on $\mathscr{C}^1$ if given the obvious Hilbert space inner product). The $j=\frac{1}{2}$ unitary representation matrices $D^{1/2}$ are defined by (3.6) and (3.7). The U_1 homogeneous transformations from (3.17) act simply as

$$O_{(0,u)}\mathbf{b}_{\mathbf{q}}^{A}=e^{+i\phi}\mathbf{b}_{\mathbf{q}}^{A},\quad O_{(0,u)}\mathbf{b}_{\mathbf{q}}^{A\dagger}=e^{-i\phi}\mathbf{b}_{\mathbf{q}}^{A\dagger},$$
$$O_{(0,u)}\mathbf{d}_{\mathbf{q}}^{A\dagger}=e^{+i\phi}\mathbf{d}_{\mathbf{q}}^{A\dagger},\quad O_{(0,u)}\mathbf{d}_{\mathbf{q}}^{A}=e^{-i\phi}\mathbf{d}_{\mathbf{q}}^{A}. \tag{4.17}$$

Parity, time reversal, and charge conjugation from (3.17) extend to

$$O_{(0,\mathsf{P})}\mathbf{b}_{\mathbf{q}}^{A}=\pm\mathbf{b}_{-\mathbf{q}}^{A},\quad O_{(0,\mathsf{P})}\mathbf{b}_{\mathbf{q}}^{A\dagger}=\pm\mathbf{b}_{-\mathbf{q}}^{A\dagger},$$
$$O_{(0,\mathsf{P})}\mathbf{d}_{\mathbf{q}}^{A}=\mp\mathbf{d}_{-\mathbf{q}}^{A},\quad O_{(0,\mathsf{P})}\mathbf{d}_{\mathbf{q}}^{A\dagger}=\mp\mathbf{d}_{-\mathbf{q}}^{A\dagger},$$
$$A_{(0,\mathsf{T})}\mathbf{b}_{\mathbf{q}}^{A}=\sigma_{2B}^{A}e^{+i\phi_{\mathsf{T}}}\mathbf{b}_{-\mathbf{q}}^{B},\quad A_{(0,\mathsf{T})}\mathbf{b}_{\mathbf{q}}^{A\dagger}=\overset{*}{\sigma}{}_{2B}^{A}e^{-i\phi_{\mathsf{T}}}\mathbf{b}_{-\mathbf{q}}^{B\dagger}, \tag{4.18}$$
$$A_{(0,\mathsf{T})}\mathbf{d}_{\mathbf{q}}^{A}=\sigma_{2B}^{A}e^{-i\phi_{\mathsf{T}}}\mathbf{d}_{-\mathbf{q}}^{B},\quad A_{(0,\mathsf{T})}\mathbf{d}_{\mathbf{q}}^{A\dagger}=\overset{*}{\sigma}{}_{2B}^{A}e^{+i\phi_{\mathsf{T}}}\mathbf{d}_{-\mathbf{q}}^{B\dagger},$$
$$O_{(0,\mathsf{C})}\mathbf{b}_{\mathbf{q}}^{A}=\sigma_{3B}^{A}e^{-i\phi_{\mathsf{C}}}\mathbf{d}_{\mathbf{q}}^{B},\quad O_{(0,\mathsf{C})}\mathbf{b}_{\mathbf{q}}^{A\dagger}=\sigma_{3B}^{A}e^{+i\phi_{\mathsf{C}}}\mathbf{d}_{\mathbf{q}}^{B\dagger},$$
$$O_{(0,\mathsf{C})}\mathbf{d}_{\mathbf{q}}^{A}=\sigma_{3B}^{A}e^{+i\phi_{\mathsf{C}}}\mathbf{b}_{\mathbf{q}}^{B},\quad O_{(0,\mathsf{C})}\mathbf{d}_{\mathbf{q}}^{A\dagger}=\sigma_{3B}^{A}e^{-i\phi_{\mathsf{C}}}\mathbf{b}_{\mathbf{q}}^{B\dagger},$$

completing the orthgonal and projective antiorthogonal actions of the homogeneous group on the basis for $\mathscr{C}^1$. Notice that P and C actions commute with the new complex structure, "i" multiplication, but the T action has to be taken to anticommute with it. For this reason the action of C in (3.17) has been changed from $A_{(0,\mathsf{C})}\to O_{(0,\mathsf{C})}$ while the T action remains conjugate–linear and written as $A_{(0,\mathsf{T})}$ in (4.18). From (3.18) the translations act on $\mathscr{C}^1$ by

$$O_{(r,I)}\mathbf{b}_{\mathbf{q}}^{A}=e^{-iq_{\mu}r^{\mu}}\mathbf{b}_{\mathbf{q}}^{A},\quad O_{(r,I)}\mathbf{b}_{\mathbf{q}}^{A\dagger}=e^{+iq_{\mu}r^{\mu}}\mathbf{b}_{\mathbf{q}}^{A\dagger},$$
$$O_{(r,I)}\mathbf{d}_{\mathbf{q}}^{A}=e^{-iq_{\mu}r^{\mu}}\mathbf{d}_{\mathbf{q}}^{A},\quad O_{(r,I)}\mathbf{d}_{\mathbf{q}}^{A\dagger}=e^{+iq_{\mu}r^{\mu}}\mathbf{d}_{\mathbf{q}}^{A\dagger}. \tag{4.19}$$

The above linear orthogonal transformations $\mathbf{O}_{\gamma}$ are analogous to the linear Lorentz transformations (Λ_{μ}^{ν}) of (2.11) for the four-dimensional M^4 space. We now look for the equivalent of the $\mathrm{Pin}_{1,3}$ group, i.e., the group $\mathscr{U}$ defined by

$$U\in\mathscr{U}\Leftrightarrow U\in\mathscr{C},$$
$$U^{\dagger}U=\mathscr{I},$$

and

$$U\mathscr{C}^1U^{\dagger}=\mathscr{C}^1. \tag{4.20}$$

The Pin covering of the orthogonal group $\mathscr{U}\to\mathbf{O}_{\gamma}$ is defined analogous to (2.11) by

$$U\mathbf{c}^1U^{\dagger}=O\mathbf{c}^1,\quad \mathbf{c}^1\in\mathscr{C}^1, \tag{4.21}$$

and has a kernel $\cong U_1$, i.e., $e^{i\phi}\mathscr{I}\to I$. The complex Clifford algebra $\mathscr{C}$ can be thought of as a Hilbert space by defining a Hermitian inner product

$$\langle\mathbf{c}_1,\mathbf{c}_2\rangle=(\mathbf{c}_1^{\dagger}\mathbf{c}_2)_{\mathscr{I}}, \tag{4.22}$$

where $(\ )_{\mathscr{I}}$ means the component of $(\ )$ contained in $\mathscr{C}^0\propto\mathscr{I}$. In this way $\mathscr{C}$ can be thought of as a direct sum of Hilbert spaces (4.9), of which $\mathscr{C}^0$ is of dimension 1. The group $\mathscr{U}$ acting as a group of inner automorphisms (vector action) as in (4.20), acts unitarily on each $\mathscr{C}^k$ separately. However, as a spinor action on the left, $\mathbf{c}\to U\mathbf{c}$, $\mathscr{U}$ acts unitarily mixing the $\mathscr{C}^k$. The spin representation of this group is found precisely as we found the four-component Dirac spinors, by finding a primitive idempotent, i.e., a projection operator.[32,33] One choice of the idempotent is

$$\mathscr{P}\equiv\lim_{i\to\mathbf{p}}\prod_{A,i}\mathbf{b}_i^A\mathbf{b}_i^{A\dagger}\mathbf{d}_i^A\mathbf{d}_i^{A\dagger}, \tag{4.23}$$

with the properties that

$$\mathscr{P}^2=\mathscr{P},\quad \mathscr{P}^{\dagger}=\mathscr{P}. \tag{4.24}$$

Note that we have used a discrete label "i" in place of $\mathbf{p}$ and obtain the continuum by a limiting procedure. We assume that this process can be made rigorous;[34] however, for our purpose the reader can take

$$\mathbf{b}_i^A\equiv\frac{1}{\sqrt{V}}\int_{D_i}dp\,\mathbf{b}_{\mathbf{p}}^A, \tag{4.25}$$

where the domains D_i are disjoint, completely cover the upper mass shell, and have invariant volume V. The $\mathbf{b}_i^{A\dagger}$, $\mathbf{d}_i^A$, and $\mathbf{d}_i^{A\dagger}$ are similarly defined and satisfy the expected anticommutation relations, e.g.,

$$\mathbf{b}_i^{A\dagger}\mathbf{b}_j^B+\mathbf{b}_j^B\mathbf{b}_i^{A\dagger}=\delta_{ij}^{AB}\mathscr{I}. \tag{4.26}$$

The important property of $\mathscr{P}$ is that

$$\mathbf{b}_{\mathbf{p}}^A\mathscr{P}=\mathbf{d}_{\mathbf{p}}^A\mathscr{P}=0. \tag{4.27}$$

The Fock vacuum state $|0\rangle$, is defined almost identically to $\mathscr{P}$ except appropriately normalized,

$$|0\rangle\equiv\lim_{i\to\mathbf{p}}\prod_{A,i}2\mathbf{b}_i^A\mathbf{b}_i^{A\dagger}\mathbf{d}_i^A\mathbf{d}_i^{A\dagger}. \tag{4.28}$$

The corresponding minimal left ideal $\mathscr{C}\mathscr{P}$ would be called a spinor space in analogy with (2.12) but is commonly called Fock space. It is generated by multiplying $\mathscr{P}$ or $|0\rangle$ on the left by $\mathscr{C}$ and is spanned by the following basis states:

$$\{|0\rangle,\mathbf{b}_{\mathbf{p}}^{A\dagger}|0\rangle,\mathbf{d}_{\mathbf{p}}^{A\dagger}|0\rangle,\mathbf{b}_{\mathbf{p}}^{A\dagger}\mathbf{b}_{\mathbf{q}}^{B\dagger}|0\rangle,$$
$$\mathbf{d}_{\mathbf{p}}^{A\dagger}\mathbf{d}_{\mathbf{q}}^{B\dagger}|0\rangle,\mathbf{b}_{\mathbf{p}}^{A\dagger}\mathbf{d}_{\mathbf{q}}^{B\dagger}|0\rangle,\cdots\}, \tag{4.29}$$

where either $A\neq B$ or $\mathbf{p}\neq\mathbf{q}$, etc. The normalization of the vacuum state is checked using (4.22),

$$\langle0|0\rangle\equiv(|0\rangle^{\dagger}|0\rangle)_{\mathscr{I}}=\left(\lim_{i\to\mathbf{p}}\prod_{A,i}4\mathbf{b}_i^A\mathbf{b}_i^{A\dagger}\mathbf{d}_i^A\mathbf{d}_i^{A\dagger}\right)_{\mathscr{I}}=1. \tag{4.30}$$

The steps in (4.30) require using the idempotentcy of $\mathbf{b}_i^A\mathbf{b}_i^{A\dagger}$ and $\mathbf{d}_i^A\mathbf{d}_i^{A\dagger}$ and decomposing $\mathbf{b}_i^A=(\mathbf{a}_i^A-i\gamma[\mathbf{a}_i^A])/2$ into self-conjugate parts as in (4.10).

So far we have seen how familiar quantities, like the vacuum and the Fock space of a spin-half quantum field theory, emerge naturally from various algebraic quantities, such as an idempotent and its minimal left ideal in second Cliffordization. Since the minimal left ideal $\mathscr{C}\mathscr{P}$ provides a spinor representation of the linear orthogonal transformations as defined in (4.16)–(4.19), we can also see how physical operators on the Fock space emerge as representations of the generators of inhomogeneous transformations. For example, for an infinitesimal translation by an amount ϵ^{μ}, we have from (4.19)

$$O_{(\epsilon,I)}\mathbf{b}_{\mathbf{q}}^A=e^{-iq_{\mu}\epsilon^{\mu}}\mathbf{b}_{\mathbf{q}}^A\approx(I-iq_{\mu}\epsilon^{\mu})\mathbf{b}_{\mathbf{q}}^A, \tag{4.31}$$

and from (4.21) the corresponding translation operator is

$$U\equiv e^{i\epsilon_{\mu}\mathbf{P}^{\mu}}\approx(\mathscr{I}+i\epsilon_{\mu}\mathbf{P}^{\mu}). \tag{4.32}$$

The momentum operators $\mathbf{P}^{\mu}$ are the representatives of translation generators and are only determined by (4.31) up to a generator of the kernel of the Pin covering as discussed in (4.21),

$$U\mathbf{b}_{\mathbf{q}}^{A}U^{\dagger}=O_{(\epsilon,I)}\mathbf{b}_{\mathbf{q}}^{A}$$
$$\Rightarrow[\mathbf{b}_{\mathbf{q}}^{A},\mathbf{P}^{\mu}]=q^{\mu}\mathbf{b}_{\mathbf{q}}^{A}$$
$$\Rightarrow\mathbf{P}^{\mu}=\sum_{A}\int dq\,q^{\mu}\mathbf{b}_{\mathbf{q}}^{A\dagger}\mathbf{b}_{\mathbf{q}}^{A}$$
$$+\text{commuting terms.}\qquad(4.33)$$

Similar consideration can be carried out for the d's. Then the total momentum operator becomes

$$\mathbf{P}^{\mu}=\sum_{A}\int dq\,q^{\mu}(\mathbf{b}_{\mathbf{q}}^{A\dagger}\mathbf{b}_{\mathbf{q}}^{A}+\mathbf{d}_{\mathbf{q}}^{A\dagger}\mathbf{d}_{\mathbf{q}}^{A})+(\text{const})\mathscr{I}.\qquad(4.34)$$

The arbitrary constant term comes from the generator of the U_1 kernel and can be eliminated by requiring that the vacuum be translationally invariant or equivalently

$$\mathbf{P}^{\mu}|0\rangle=0\quad\Rightarrow\text{const}=0.\qquad(4.35)$$

That is, we require the vacuum state to have zero energy and momentum. This is equivalent to the normal ordering procedure in quantum field theory. The resulting Hamiltonian for free electrons and positrons is

$$\mathbf{H}=c\mathbf{P}^{0}=c\sum_{A}\int dq\,q^{0}(\mathbf{b}_{\mathbf{q}}^{A\dagger}\mathbf{b}_{\mathbf{q}}^{A}+\mathbf{d}_{\mathbf{q}}^{A\dagger}\mathbf{d}_{q}^{A}),\qquad(4.36)$$

where

$$\mathbf{N}_{\mathbf{q}}^{A^{-}}\equiv\mathbf{b}_{\mathbf{q}}^{A\dagger}\mathbf{b}_{\mathbf{q}}^{A},\quad\mathbf{N}_{\mathbf{q}}^{A^{+}}\equiv\mathbf{d}_{\mathbf{q}}^{A\dagger}\mathbf{d}_{\mathbf{q}}^{A}\qquad(4.37)$$

are the number operators for the electrons and the positrons, respectively.

The final physical observable we obtain is the charge operator Q. It comes from the infinitesimal phase transformations (4.17),

$$O_{(0,u)}\mathbf{b}_{\mathbf{q}}^{A}=e^{+i\phi}\mathbf{b}_{\mathbf{q}}^{A}\approx(I+i\phi)\mathbf{b}_{\mathbf{q}}^{A},$$
$$O_{(0,u)}\mathbf{d}_{\mathbf{q}}^{A\dagger}=e^{+i\phi}\mathbf{d}_{\mathbf{q}}^{A\dagger}\approx(I+i\phi)\mathbf{d}_{\mathbf{q}}^{A\dagger}.\qquad(4.38)$$

If the corresponding U of (4.21) is written as

$$U=e^{i\phi\mathbf{Q}}\approx\mathscr{I}+i\phi\mathbf{Q},\qquad(4.39)$$

the reader finds by steps similar to (4.33)–(4.35) but even simpler,

$$\mathbf{Q}=\sum_{A}\int dq(-\mathbf{b}_{\mathbf{q}}^{A\dagger}\mathbf{b}_{\mathbf{q}}^{A}+\mathbf{d}_{\mathbf{q}}^{A\dagger}\mathbf{d}_{\mathbf{q}}^{A})+(\text{const})\mathscr{I}.\qquad(4.40)$$

Again, the constant term comes from the U_1 kernel and can be eliminated by requiring that the vacuum have no charge,

$$\mathbf{Q}|0\rangle=0\Rightarrow\text{const}=0.\qquad(4.41)$$

V. CONCLUSIONS

In this paper we have tried to "tie together" some of the "loose ends" in the free electron–positron field theory by showing how the appropriate construction of two successive Clifford algebras can result in the free quantum field theory. The first Clifford algebra was associated with the tangent space of any point in Minkowski space and its Lorentz invariant inner product. The second was associated with an infinite-dimensional Hilbert space and its Poincaré [enlarged (2.29)] invariant Hermitian inner product, which we constructed (via Wigner's procedure) using the spinor representation of the first Clifford algebra. All elements of the noninteracting theory seem to be accounted for by this "second Cliffordization." In particular, the operator algebra of the free field theory is just the second complex Clifford algebra. The familiar abstract Fock representation appears concretely as a spinor representation space in the infinite dimensional algebra analogous to four-component Dirac spinors in the finite Minkowski algebra. Two obvious extensions of this work are to higher dimensions and to the inclusion of interactions with external fields. Extensions to other spins as well as to massless fields seem straightforward.

ACKNOWLEDGMENTS

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

The authors wish to thank R. Ablamowicz and S.T. Ali for suggesting improvements in the first version of this manuscript.

[1] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).
[2] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).
[3] W. K. Clifford, Am. J. Math. **1**, 350 (1878).
[4] C. C. Chevalley, *The Algebraic Theory of Spinors* (Columbia Univ., New York, 1954).
[5] M. Riesz, *Clifford Numbers and Spinors*, Lecture Series No. 38 (Institute for Fluid Dynamics and Applied Mathematics, Univ. of Maryland, 1958).
[6] D. Hestenes, *Space-Time Algebra* (Gordon and Breach, New York, 1966).
[7] P. Budinich and A. Trautman, *The Spinoral Chessboard* (Springer, Berlin, 1988); P. Lounesto, Found. Phys. **11**, 721 (1981).
[8] I. M. Benn and R. W. Tucker, *An Introduction to Spinors and Geometry with Applications in Physics* (Hilger, Bristol, 1987).
[9] E. P. Wigner, Ann. Math. **40**, 149 (1939).
[10] E. P. Wigner, in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gursey (Gordon and Breach, New York, 1964).
[11] G. W. Mackey, Ann. Math. **55**, 101 (1952).
[12] F. R. Halpern, *Special Relativity and Quantum Mechanics* (Prentice-Hall, Englewood Cliffs, NJ, 1968).
[13] D. J. Simms, Lectures Notes in Mathematics Vol. 52 (Springer, Berlin, 1968).
[14] Y. S. Kim and M. E. Noz, *Theory and Applications of the Poincaré Group* (Reidal, Dordrecht, 1986).
[15] I. E. Segal, Ann. Math. **48**, 930 (1947).
[16] I. E. Segal, *Mathematical Problems in Relativistic Physics* (Am. Math. Soc., Providence, 1963).
[17] D. Shale and W. F. Stinespring, Ann. Math. **80**, 365 (1964).
[18] D. Shale and W. F. Stinespring, J. Math. Mech. **14**, 315 (1965).
[19] R. F. Streater, Rep. Prog. Phys. **38**, 847 (1975).
[20] J. M. Cook, Trans. Am. Math. Soc. **74**, 222 (1953).
[21] M. Weinless, J. Funct. Anal. **4**, 350 (1969).
[22] V. Fock, Z. Phys. **75**, 622 (1932).
[23] P. Lounesto and E. Latvamaa, Proc. AMS **79**, 533 (1980).
[24] P. Lounesto and G. P. Wene, Acta Appl. Math. **9**, 165 (1987).
[25] A. Crumeyrolle, *Algebres de Clifford et Spineurs* (Univ. of Toulouse, Toulouse, 1974).
[26] R. Ablamowicz and P. Lounesto, in *Clifford Algebras and Their Applications in Mathematical Physics*, edited by J. S. R. Chisholm and L. K. Common (Reidel, Dordrecht, 1985).
[27] K. Bugajska, J. Math. Phys. **27**, 143 (1986).
[28] P. J. M. Bongaarts, Ann. Phys. **56**, 108 (1970).
[29] P. J. M. Bongaarts, in *Mathematics of Contemporary Physics*, edited by R. F. Streater (Academic, New York, 1972).
[30] P. Broadbridge and C. A. Hurst, Ann. Phys. (N.Y.) **137**, 86 (1981).
[31] H. Araki, in *Quantum Theories and Geometry*, edited by M. Cahen and M. Flato (Kluwer, Netherlands, 1988), p. 1.
[32] G. P. Wene, J. Math. Phys. **30**, 249 (1989).
[33] A. Crumeyrolle, Rep. Math. Phys. **25**, 305 (1987).
[34] S. T. Ali has suggested to us a smoothing operation to legitimize the limit.

Dirac quantization of massive spin-one particles in an external symmetrical tensor field

Teymour Darkhosh
Division of Natural Science and Mathematics, St. Mary's College of Maryland, St. Mary's City, Maryland 20686

(Received 29 September 1989; accepted for publication 25 April 1990)

Dirac's method is used to quantize massive spin-one particles interacting with an external symmetric tensor field. It is shown that Dirac equations of motion are identical to Euler–Lagrange equations and that ϕ^0, the 0th component of the unknown field, is the only component that depends on the external field. Furthermore, Dirac commutators of the field and the Lorentz generators are calculated. It is shown that the field components except ϕ^0 transform like components of a free field, and that ϕ^0 transforms like a field component in an external potential.

I. INTRODUCTION

Wave equations for higher spin particles, i.e., particles with spin ≥ 1, in the presence of external interactions suffer from a variety of ill effects. The most fundamental flaw in the theory is the acausal behavior where certain components of the field propagate faster than the speed of light[1] and consequently violate a principle of the special theory of relativity. The source of the problem is that in the presence of external interactions, it is not possible to eliminate the unwanted field components using the constraint equations.

Although acausality surfaces in classical field equations, it is of interest to examine the presence of constraint equations in quantum field equations. Also, because acausality indirectly implies that the theory is not invariant, it is essential to see if the unknown field transforms properly under the Lorentz transformations.

In this paper, we will consider the Proca wave equation,[2] the simplest wave equation suffering this ill effect. We quantize the field using Dirac's method of quantization[3] and use the Hamiltonian method to derive the equations of motion. We will then show that the method of quantization is consistent with the Lagrangian equations of motion.

This consistency has been shown for the Rarita–Schwinger wave equation in the presence of an external electromagnetic field.[4] We will also show that ϕ^0 is the only component that depends on the external field and that this is consistent with the Lorentz transformation of ϕ^0.

Our notation is that of Bjorken and Drell.[5] The metric tensor, $g^{\alpha\beta}$, is defined as $g^{00}=1$, $g^{11}=g^{22}=g^{33}=-1$. Greek indices range from 0 to 3; Roman indices range from 1 to 3. The summation rule for repeated indices is used throughout the paper.

We assume that the external field is an explicit function of space-time, x^μ, and define the total derivatives, $d\mu$, as $d\mu = i[p\mu, \ ] + \partial\mu$, where $p\mu$ is the four momentum operator. All the derivatives that appear in the Lagrangian density and the Euler–Lagrange equation are total derivatives.

II. DIRAC METHOD OF QUANTIZATION

In this section, we begin by applying Dirac's method of quantization for constraint systems[6] to the Proca wave equation in the presence of an external field. The Lagrangian density that we are interested in is given by

$$\mathscr{L} = -\tfrac{1}{4}G^{\alpha\beta}G_{\alpha\beta} + m^2\phi\cdot\phi + (\lambda/2)\phi\cdot T\cdot\phi, \tag{2.1}$$

where $G^{\alpha\beta} = d^\alpha\phi^\beta - d^\beta\phi^\alpha$, and $T^{\alpha\beta}$ is a symmetric external field, $T^{\alpha\beta} = T^{\beta\alpha}$, which depends on space-time, x^μ, explicitly.

The conjugate momentum is defined

$$\pi^\mu \equiv \frac{\partial\mathscr{L}}{\partial\dot{\phi}_\mu} = G^{\mu^0}, \tag{2.2}$$

where $\dot{\phi}\mu$ is the total time derivative of $\phi\mu$. The primary constraint follows directly from (2.2):

$$\chi_1 \equiv G^{00} = \pi^0 \cong 0. \tag{2.3}$$

The expression (2.3) is weakly equal to zero, meaning that it should be set equal to zero after all the commutation relations have been calculated.

The Hamiltonian density is

$$\mathscr{H} \equiv \pi^\alpha\dot{\phi}_\alpha - \mathscr{L}, \tag{2.4}$$

and the Hamiltonian is given by

$$H = \int \mathscr{H}\, d^3x. \tag{2.5}$$

Using the explicit form of $\mathscr{L}$ and adding a three-divergent, the Hamiltonian density becomes

$$\mathscr{H} = -\tfrac{1}{2}\pi^i\pi_i + \tfrac{1}{4}G^{ij}G_{ij} - (m^2/2)\phi\cdot\phi - (\lambda/2)\phi\cdot T\cdot\phi - (d_i\pi^i)\phi^0. \tag{2.6}$$

Now we impose the following condition:

$$[\pi^\alpha(x),\phi_\beta(y)]_{x^0=y^0} = -ig^\alpha{}_\beta\delta^3(x-y). \tag{2.7}$$

Following Dirac, we define the total Hamiltonian density:

$$\mathscr{H}_t = \mathscr{H} + u_1\chi_1. \tag{2.8}$$

Note that u_1 is a function of the field components to be determined later.

To obtain the secondary constraints, we use

$$\chi_2 \equiv \dot{\chi}_1 = i[H_t,\chi_1] + \frac{\partial\chi_1}{\partial t} \cong 0. \tag{2.9}$$

After substituting for χ_1,π^0 and using (2.7), we get the secondary constraint:

 0022-2488/90/092201-03$03.00

$$\chi_2 = d_i\pi^i + m^2(1 + (\lambda/m^2)T^{00})\phi^0 + \lambda T^{0i}\phi_i = 0. \tag{2.10}$$

Following the same procedure, we do not get new constraints. However, we do obtain a relation for u_1:

$$u_1 = \frac{-1}{m^2(1 + (\lambda/m^2)T^{00})}\left[m^2 d_i\phi^i + \lambda d_i T^{i\alpha}\phi_2 + \lambda T^{0i} d_0\phi_i + \frac{\partial\chi_2}{\partial t}\right]. \tag{2.11}$$

Since the external field depends on x^μ explicitly, so does χ_2.

To calculate the generalized commutation relations, we need to calculate the matrix formed by the constraints, $[\chi_i,\chi_j]$. The matrix is

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} im^2\left(1 + \frac{\lambda}{m^2}T^{00}\right)\delta^3(x-y). \tag{2.12}$$

The inverse of (2.12), $\mathbf{C} = [\chi,\chi]^{-1}$, is

$$\mathbf{C} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\left[im^2\left(1 + \frac{\lambda}{m^2}T^{00}\right)\right]^{-1}\delta^3(x-y). \tag{2.13}$$

We define the generalized commutation relation between two field dependent quantities as

$$[\xi,\eta]^* \equiv [\xi,\eta] - [\xi,\chi_i]C_{ij}[\chi_j,\eta]. \tag{2.14}$$

Using the above commutation relation, the commutators of different field components are

$$\begin{aligned} &[\phi^i,\phi^j]^* = [\pi^i,\pi^j]^* = [\pi^0,\phi^0]^* = 0, \\ &[\pi^i,\phi^j]^* = -ig^{ij}\delta^3(x-y), \\ &[\phi^i,\phi^0]^* = \frac{-i}{m^2(1 + (\lambda/m^2)T^{00})}\partial^i\delta^3(x-y). \end{aligned} \tag{2.15}$$

Now we will compare the equations of motion derived using Dirac's generalized Hamiltonian method and those derived directly from the Lagrangian density. First consider

$$[H_t,\phi^i]^* = i\pi^i - id^i\phi^0 = -i\dot{\phi}i, \tag{2.16}$$

which implies that $\partial_0\phi^i = 0$, i.e., ϕ^i is not an explicit function of time and does not depend on the external field, $T^{\alpha\beta}$. Next we consider

$$[H_t,\pi^k]^* = -i[d_i G^{ik} + m^2\phi^k + \lambda T^k\cdot\phi]. \tag{2.17}$$

We set (2.17) equal to $-i\dot{\pi}^k$, and to be consistent with (2.7), we assume that π^k is not an explicit function of time. We get

$$d\mu\, G^{\mu k} + m^2\phi^k + \lambda T^k\cdot\phi = 0. \tag{2.18}$$

The expression (2.18), when combined with the secondary constraint, χ_2, gives the Euler–Lagrange equation

$$d\mu\, G^{\mu\alpha} + m^2\phi^\alpha + \lambda T^\alpha\cdot\phi = 0. \tag{2.19}$$

Following a similar procedure for ϕ^0, we get

$$[H_t,\phi^0]^* = -iu_1 - \frac{i}{(1 + (\lambda/m^2)T^{00})}\partial_0\chi_2, \tag{2.20}$$

and using the expression for u_1, (2.11), we can write

$$[H_t,\phi^0]^* = \frac{i}{m^2(1 + (\lambda/m^2)T^{00})}[m^2 d_i\phi^i + \lambda d_i T^{i\alpha}\phi_\alpha + \lambda T^{0i} d_0\phi_i]. \tag{2.21}$$

Setting (2.21) equal to $-i(\dot{\phi}^0 - \partial^0\phi^0)$, we have

$$m^2 d\cdot\phi + \lambda d\cdot T\cdot\phi - [m^2(1 + (\lambda/m^2)T^{00})\partial_0\phi^0 + \lambda(\partial\cdot T^{0\alpha}\phi_\alpha] = 0. \tag{2.22}$$

Now if we take the divergence of (2.19), we get

$$m^2 d\cdot\phi + \lambda d\cdot T\cdot\phi = 0, \tag{2.23}$$

which implies that

$$\partial_0\phi^0 = \frac{-\lambda}{(1 + (\lambda/m^2)T^{00})}(\partial_0 T^{0\alpha})\phi_\alpha. \tag{2.24}$$

The expression (2.24) shows that ϕ_0 depends on time explicitly and in turn on the external field. In our formulation, it is the only field component that depends on the external field. Therefore it is reasonable to assume that ϕ^0 is responsible for the acausal mode.

III. POINCARÉ GENERATORS

In this section, we use the Dirac commutation relation to examine the transformation of the field components under the Poincaré group.

The momentum operator is given by

$$P^i = \int \mathscr{T}^{0i}\, d^3x, \tag{3.1}$$

where $\mathscr{T}^{0i} = \pi^j d^i\phi_j$ is the momentum tensor.

Using the Dirac commutator, we find

$$[P^i,\phi^j]^* = -id^i\phi^j. \tag{3.2}$$

Combining (3.2) and (2.16), we can write

$$[P^\alpha,\phi^j]^* = -id^\alpha\phi^j. \tag{3.3}$$

In the above expression, we have taken $\mathscr{T}^{00}$ to be H_t. Comparing (3.3) with the general formula,

$$[P^\alpha,F]^* = -id^\alpha F + i\partial^\alpha F, \tag{3.4}$$

where F is a general field, we conclude that $\partial^\alpha\phi^j = 0$, i.e., ϕ^j is not an explicit function of space-time.

The commutation of P^i and ϕ^0 reveals

$$[P^i,\phi^0]^* = i[d_k d^i\pi^k + \lambda T^{0k} d^i\phi\mu]. \tag{3.5}$$

Substituting for $d_k\pi^k$ from (2.10), the above expression becomes

$$[P^i,\phi^0]^* = -id^i\phi^0 - \frac{i\lambda}{m^2(1 + (\lambda/m^2)T^{00})}(\partial^i T^{0\alpha})\phi_\alpha. \tag{3.6}$$

Relations (3.6) and (2.24) imply that the explicit dependence of ϕ^0 on x^μ is given by

$$\partial^\mu\phi^0 = -\frac{\lambda}{m^2(1 + (\lambda/m^2)T^{00})}(\partial^\mu T^{0\alpha})\phi_\alpha. \tag{3.7}$$

If we use the chain rule,

$$\partial^m\phi^0 = \frac{\partial\phi^0}{\partial T^{0\alpha}}\frac{\partial T_{0\alpha}}{\partial x\mu}, \tag{3.8}$$

we get the dependence of ϕ^0 on $T^{0\alpha}$, namely

$$\frac{\partial\phi^0}{\partial T^{0\alpha}} = -\frac{\lambda}{m^2(1 + (\lambda/m^2)T^{00})}\phi_\alpha. \tag{3.9}$$

Now to examine the transformation property of the field

components, we use the angular momentum operator, which is derived from the action integral and given by

$$\mathscr{M}^{\mu\alpha\beta} = (x^{\alpha}\mathscr{J}^{\mu\beta} - x^{\beta}\mathscr{J}^{\mu\alpha}) - G^{\mu\alpha}\phi^{\beta} + G^{\mu\beta}\phi^{\alpha}. \tag{3.10}$$

The calculation of the commutaton relations between $J^{\alpha\beta}$ and ϕ^{k}, where $J^{\alpha\beta}$ is defined as $J^{\alpha\beta} = \int d^{3}x\, \mathscr{M}^{0\alpha\beta}$, yields the following result:

$$[J^{\alpha\beta},\phi^{k}]^{*} = -i(x^{\alpha}d^{\beta} - x^{\beta}d^{\alpha})\phi^{k} - i(g^{\alpha k}g^{\beta\mu} - g^{\beta\mu}g^{\alpha\mu})\phi_{\mu}. \tag{3.11}$$

Expression (3.11) shows that ϕ^{k} transforms as a free spin-one field. A similar computation for ϕ^{0} yields

$$\begin{aligned}[J^{\alpha\beta},\phi^{0}]^{*} &= -i(x^{\alpha}d^{\beta} - x^{\beta}d^{\alpha})\phi^{0} - \frac{i\lambda}{m^{2}(1 + (\lambda/m^{2})T^{00})} \\ &\times \phi_{\sigma}(x^{\alpha}\partial^{\beta} - x^{\beta}\partial^{\alpha})T^{0\sigma} - i(g^{\alpha 0}g^{\beta 0} - g^{\beta 0}g^{\alpha 0})\phi_{\sigma} \\ &+ \frac{i\lambda}{m^{2}(1 + (\lambda/m^{2})T^{00})} \\ &\times (T^{0\alpha}\phi^{\beta} - T^{0\beta}\phi^{\alpha} - g^{\alpha 0}T^{\beta\cdot}\phi).\end{aligned} \tag{3.12}$$

When (3.9), the dependence of ϕ^{0} on the external field, is used, expression (3.12) takes the following form:

$$\begin{aligned}[J^{\alpha\beta},\phi^{0}]^{*} &= -i(x^{\alpha}d^{\beta} - x^{\beta}d^{\alpha})\phi^{0} \\ &+ i\frac{\partial\phi^{0}}{\partial T^{0\sigma}}(x^{\alpha}\partial^{\beta} - x^{\beta}\partial^{\alpha})T^{0\sigma} \\ &- i(g^{\alpha 0}g^{\beta\sigma} - g^{\beta 0}g^{\alpha\sigma})\phi_{\sigma} \\ &+ i\frac{\partial\phi^{0}}{\partial T^{0\sigma}}(g^{\alpha 0}T^{\sigma\beta} + g^{\beta 0}T^{\sigma\alpha} \\ &+ g^{\alpha\sigma}T^{0\beta} - g^{\beta\sigma}T^{0\alpha}).\end{aligned} \tag{3.13}$$

Expression (3.13) is consistent with the transformation of field components coupled to external interactions.[7] The first two parentheses are the transformation of ϕ^{0} and $T^{\alpha\beta}$ under rotation and the last two parentheses are the transformation of a vector field and a tensor of second rank.

Thus, we have shown that the ϕ^{k} components of the field are not affected by the external field and transform as a free spin-one field. However, ϕ^{0} transforms as a field component coupled to the external field, $T^{0\alpha}_{(x)}$, and is responsible for the acausal propagation.

[1] G. Velo and D. Zwanziger, Phys. Rev. **188**, 2218 (1969).
[2] T. Darkhosh, Phys. Rev. D **22**, 888 (1980).
[3] P. A. M. Dirac, *Lectures on Quantum Mechanics*, Belfer Graduate School of Science Monograph Series, No. 2 (Yeshiva University, New York, 1964).
[4] G. B. Mainland and E. C. G. Sudarshan, Phys. Rev. D **8**, 1088 (1973).
[5] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).
[6] For the Dirac method applied to the Rarita–Schwinger equation, see K. Inoue, M. Omote, and M. Kobayashi, Prog. Theor. Phys. **63**, 1413 (1980).
[7] G. B. Mainland and E. C. G. Sudarshan, Phys. Rev. D **10**, 3343 (1974).

Embedding of a Demianski cavity with small rotation parameter in a perturbation of a Friedmann universe with cosmological constant

J. M. Aguirregabiria, A. Chamorro, J. R. Etxebarria, and M. Rivas
Física Teórica, Facultad de Ciencias, Universidad del País Vasco, Apdo. 644, 48080 Bilbao, Spain

(Received 1 August 1989; accepted for publication 25 April 1990)

The problem of embedding a Demianski cavity with small rotation parameter in an appropriate rotational perturbation of a pressureless Friedmann universe with a Λ term is considered. The relation between the coordinate change introduced by Schücking [Z. Phys. **137**, 595 (1954)] for this kind of problems and that used for the simple model of Oppenheimer and Snyder [Phys. Rev. **56**, 455 (1939)] for gravitational collapse is also discussed.

I. INTRODUCTION

The problem of matching the two most important exact solutions of the Einstein equations, those of Friedmann–Robertson–Walker and Schwarzschild, was considered by Einstein and Straus,[1] who analyzed the influence of the universe expansion on the gravitational field surrounding an individual star. In the Einstein and Straus model, a spherical vacuum region containing at its center a Schwarzschild mass is cut out inside a pressureless cosmological fluid.The matching of metrics found by Einstein and Straus depends on the unknown solutions of some differential equations. A more explicit solution for the same problem was presented by Schücking[2] and his work has been recently extended to the case of the non-null cosmological constant by Balbinot *et al.*[3]

Related problems of embedding the Schwarzschild solution in cosmological backgrounds have been considered by McVittie,[4] Dirac,[5] and Gautreau.[6] Other spherical inhomogeneities in cosmology has been considered in the so-called "Swiss cheese" models.[7] In inflationary cosmology the dynamics of false-vacuum spherical bubbles with a domain wall have also been analyzed.[8]

On the other hand, the original approach of Einstein and Straus has been extended to the case of a small rotation by Chamorro,[9] keeping in mind that almost all large aggregations of matter in the universe have some form of rotation.In Chamorro's paper, the Kerr solution developed to first order in the rotation parameter is substituted for the Schwarzschild solution and a rotational perturbation of the Friedmann–Robertson–Walker solution is used as the exterior metric. This perturbation decays to zero as the spatial distance increases.

In this paper we simultaneously extend the works of Balbinot *et al.*[3] and Chamorro[9] by considering the matching of a Demianski solution[10] with a small rotation parameter in a spherical cavity cut out inside an external rotational perturbation of a Friedmann–Robertson–Walker universe with zero pressure and a non-null cosmological constant. Our results are valid to first order of perturbation theory. Instead of the original approach of Einstein and Straus[1] we shall start from the equivalent, but more explicit method of Schücking.[2]

In addition, the local equivalence of two problems which correspond to very different physical and topological conditions seems to have been largely overlooked (one exception would be the book by Stephani[11]). For example, the strictly local problems of matching the Schwarzschild and Robertson–Walker metrics in the Einstein and Straus vacuole and in the model for a gravitational collapse of Oppenheimer and Snyder,[12] where the exterior metric is Schwarzschild and the interior one is Friedmann, are exactly identical. In fact, the latter work has been extended to the case of the collapse of a slowly rotating dust cloud by Kegeles.[13]

We shall explicitly show the equivalence between the matching methods used in cosmological problems[2,3] and in the simplest models for gravitational collapse.[12–14]

II. THE PROBLEM

We shall consider a spherical cavity where the spacetime metric is the generalization of the Kerr solution to the case of the non-null cosmological constant given by Demianski.[10] Since we shall always keep only the first term in the expansions in the small rotation parameter ϵ, this metric reads, to this first approximation, as

$$ds_-^2 = -b\,dt^2 + (1/b)dr^2 + r^2\,d\omega^2 - 2\epsilon \sin^2\theta(1-b)dt\,d\varphi, \tag{1}$$

with

$$b = 1 - 2M/r - (\Lambda/3)r^2, \quad d\omega^2 = d\theta^2 + \sin^2\theta\,d\varphi^2. \tag{2}$$

This metric satisfies, to first order in ϵ, the vacuum field equations with a Λ term,

$$R_{\alpha\beta} + \Lambda g_{\alpha\beta} = 0, \tag{3}$$

and reduces to the Schwarzschild–de Sitter metric used in Ref. 3 when there is no rotation ($\epsilon = 0$) and to the expansion of the Kerr metric used in Ref. 9 when $\Lambda = 0$.

In the exterior of the cavity the metric will be a rotational perturbation of the Robertson–Walker metric in the form[13,9]

$$ds_+^2 = -d\tau^2 + R^2(C^{-2}\,d\rho^2 + \rho^2\,d\omega^2) - 2\epsilon\rho^2 R^2 \sin^2\theta(W\,d\tau + X\,d\rho)d\varphi, \tag{4}$$

 0022-2488/90/092204-04$03.00

where $C=\sqrt{1-k\rho^2}$ ($k=-1,0,1$), the scale factor R depends on τ, and the functions W and X depend on τ and ρ.

We suppose a cosmological fluid of pressureless dust moving with the four-velocity

$$u_\alpha=(-1,0,0,\epsilon L(\tau,\rho,\theta)),$$
$$u^\alpha=(1,0,0,\epsilon(W+L/\rho^2R^2\sin^2\theta)). \tag{5}$$

The stress-energy tensor is $T_{\alpha\beta}=\alpha u_\alpha u_\beta$ and its conservation gives the "total mass" $A\equiv\frac{1}{3}\alpha R^3$, which remains constant. Also, $\dot{L}\equiv\partial L/\partial\tau=0$: This condition also guarantees that the motion of the fluid is geodesic to first order in ϵ.

It can be seen that under these assumptions the field equations

$$R_{\alpha\beta}-\tfrac{1}{2}Rg_{\alpha\beta}-\Lambda g_{\alpha\beta}=-8\pi GT_{\alpha\beta} \tag{6}$$

give rise to the evolution equation for the scale factor,

$$\dot{R}=h(R)\equiv\sqrt{8\pi GAR^{-1}-k+\lambda R^2}\quad(\lambda=\Lambda/3), \tag{7}$$

and the following conditions on the functions L, W, and X:

$$L=Cl(\rho)\sin^2\theta/2\rho^2,\quad \dot{X}-W'=f(\rho)/\rho^4CR^3, \tag{8}$$

where $W'\equiv\partial W/\partial\rho$ and $l(\rho)$ and $f(\rho)$ are arbitrary except for the fact that they must satisfy

$$f'=24\pi GAl. \tag{9}$$

Finally, we shall also require the perturbation to vanish at infinite spatial distance, that is,

$$\lim_{\rho\to a}\frac{Cl(\rho)}{\rho^2}=\lim_{\rho\to a}\frac{f(\rho)}{\rho^4}$$
$$=\lim_{\rho\to a}W(\tau,\rho)$$
$$=\lim_{\rho\to a}X(\tau,\rho)=0, \tag{10}$$

where a stands for 1 when $k=1$ and for ∞ if $k=0,-1$.

The problem we face can be stated as follows: Given the values of the constants M, Λ, k, and A and the scale factor $R(\tau)$ satisfying Eq. (7), we seek a spherical surface Σ [with the equations $r=r_0(t)$ in internal coordinates and $\rho=\rho_0(\tau)$ in external coordinates] and the functions $L(\rho,\theta)$, $W(\tau,\rho)$, and $X(\tau,\rho)$ satisfying Eqs. (8)–(10) in such a way that the first and second fundamental forms are continuous across the surface.

We shall work to order ϵ throughout the paper, as indicated above, and will comment at the end on the approximate nature of our solution.

III. THE CONTINUITY OF THE METRIC

To analyze the continuity of the metric across the spherical surface, we shall closely follow the method of "curvature coordinates" of Refs. 2 and 3. Thus we shall change the coordinates for the exterior metric from (τ,ρ) to (t,r) by means of the implicit equations

$$R(\tau)=\phi(t,r),\quad \rho=r/\phi(t,r), \tag{11}$$

with $\phi(t,r)$ defined (implicitly, again) by

$$F_1(\phi(t,r))+F_2(r/\phi(t,r))=G(t), \tag{12}$$

with a function $G(t)$ to be determined later and with

$$F_1(x)=-2\int\frac{dx}{8\pi GA-kx+\lambda x^3},$$
$$F_2(x)=\frac{1}{k}\ln|1-kx^2|. \tag{13}$$

A more explicit expression for F_1 is discussed in Ref. 3.

Equations (12) and (13) are required in order to guarantee that in the new coordinates the coefficient of $dt\,dr$ vanishes. Indeed, one can easily see that in the coordinates (t,r) the exterior metric reads as

$$ds_+^2=-\frac{1}{4}\dot{G}^2\frac{H^2D^2}{B}dt^2$$
$$+\frac{1}{B}dr^2+r^2\,d\omega^2-2\epsilon r^2\sin^2\theta$$
$$\times\left\{\left(W-\frac{rH}{\phi^2}X\right)\frac{\dot{\phi}}{H}dt\right.$$
$$\left.+\left[\left(W-\frac{rH}{\phi^2}X\right)\frac{\phi'}{H}+\frac{X}{\phi}\right]dr\right\}d\varphi, \tag{14}$$

where

$$H(t,r)=\sqrt{8\pi GA\phi^{-1}-k+\lambda\phi^2},\quad D(t,r)=\sqrt{\phi^2-kr^2},$$
$$B(t,r)=1-8\pi GAr^2\phi^{-3}-\lambda r^2=\phi^{-2}(D^2-r^2H^2). \tag{15}$$

Next, we require continuity of the line element on the spherical surface Σ at $r=r_0(t)=R(\tau)\rho_0(\tau)=\phi_0(t,r_0(t))\rho_0(\tau)$. By comparing the coefficients of dr^2 in Eqs. (1) and (14), we see that[2,3]

$$\rho_0=(2M/8\pi GA)^{1/3}=\text{const.} \tag{16}$$

Using Eqs. (12) and (13) to analyze the continuity of the coefficients of dt^2, we find

$$\dot{r}_0=r_0^{-3/2}[r_0-(2M+\lambda r_0^3)](2M+\lambda r_0^3$$
$$-k\rho_0^2r_0)^{1/2}(1-k\rho_0^2)^{-1/2}. \tag{17}$$

Equations (16) and (17) give the radius of the matching surface in external and internal coordinates, respectively. By using Eq. (16) and a solution $r_0(t)$ to Eq. (17), the function $G(t)$ can be computed by means of the restriction of Eq. (12) to the surface of matching, which gives the relation

$$G(t)=F_1[r_0(t)/\rho_0]+F_2(\rho_0) \tag{18}$$

and then $\phi(t,r)$ can in principle be found from Eq. (12). This completely determines the change of coordinates in Eq. (11) and guarantees the continuity of the metric to zeroth order in ϵ.

In addition, continuity of the coefficients of $dt\,d\varphi$ and $dr\,d\varphi$, i.e., continuity of the metric to first order in ϵ, gives

$$W_0=C_0(1-B_0)/\rho_0^2R^2B_0,$$
$$X_0=\dot{R}(1-B_0)/\rho_0RC_0B_0, \tag{19}$$

where the subindex zero means that the expression is valid only on the matching surface. Thus, for example, we have

$$C_0=\sqrt{1-k\rho_0^2},$$
$$B_0=1-8\pi GAR^{-1}\rho_0^2-\lambda R^2\rho_0^2$$
$$=b_0=1-2M/r_0-\lambda r_0^2. \tag{20}$$

In the particular case in which $\Lambda = 0$, the results of Chamorro[9] are recovered.

The matching surface can be seen as made of points that slowly rotate along the geodesics, with equations given in external coordinates by (5) with $\rho = \rho_0$ and in internal coordinates by (17) and

$$\frac{d\varphi}{dt} = \epsilon \frac{1 + Nb_0}{r_0^2}, \quad N = \frac{l(\rho_0)}{2\rho_0^2} - 1 = \text{const.} \tag{21}$$

IV. THE CONTINUITY OF THE EXTRINSIC CURVATURE

Since we assume that there is no singular domain wall at the points on the matching surface—which are in free fall in both metrics—we must require not only the continuity of the metric, but also that of the extrinsic curvature.[15]

By using the results of Sec. III it is easy to see that the outward normal unit vector in internal and external coordinates is

$$n_\alpha^{(-)} = (-\rho_0 \dot{R}, C_0 B_0^{-1}, 0, 0),$$
$$n_\alpha^{(+)} = (0, RC_0^{-1}, 0, 0). \tag{22}$$

We choose the intrinsic coordinates for Σ as $(\xi_1, \xi_2, \xi_3) = (\tau, \theta, \varphi)$; the associated basis of tangent vectors in internal coordinates is

$$e^\alpha_{(\tau)} = (C_0 B_0^{-1}, \rho_0 \dot{R}, 0, 0), \quad e^\alpha_{(\theta)} = (0, 0, 1, 0),$$
$$e^\alpha_{(\varphi)} = (0, 0, 0, 1) \tag{23}$$

and in external coordinates the basis is

$$e^\alpha_{(\tau)} = (1, 0, 0, 0), \quad e^\alpha_{(\theta)} = (0, 0, 1, 0),$$
$$e^\alpha_{(\varphi)} = (0, 0, 0, 1). \tag{24}$$

It is possible to see that the components of the extrinsic curvature

$$K_{ij} = -n_\alpha \left(\frac{\partial e^\alpha_{(i)}}{\partial \xi^j} + \Gamma^\alpha_{\beta\gamma} e^\beta_{(i)} e^\gamma_{(j)} \right), \tag{25}$$

as computed in both types of coordinates, are exactly the same at zeroth order in ϵ,[3] but now we have the following first-order terms:

$$K^{(-)}_{\tau\varphi} = \frac{1}{2} \epsilon \sin^2 \theta \times \frac{(3B_0 - 2C_0^2)(1 - B_0) - \Lambda \rho_0^2 R^2 B_0}{\rho_0 R B_0},$$
$$K^{(+)}_{\tau\varphi} = \tfrac{1}{2} \epsilon \sin^2 \theta \rho_0 R C_0 [\rho_0 (\dot{X}_0 - W_0') - 2W_0]. \tag{26}$$

In consequence, using (19) we see that one must also require that W and X satisfy, at Σ,

$$\dot{X}_0 - W_0' = 6M / \rho_0^4 R^3 C_0, \tag{27}$$

which reduces to the condition found by Chamorro[9] when $\Lambda = 0$.

By a straightforward extension of the analysis in Ref. 9 it is possible to show that there exist the functions $L(\rho,\theta)$, $W(\tau,\rho)$, and $X(\tau,\rho)$ satisfying Eqs. (8)–(10), (19), and (27). This solves the proposed embedding problem.

V. FINAL COMMENTS

To match both metrics we have passed from the coordinates (τ,ρ) to (t,r) by means of (11). It is equally possible, of course, to express the coordinates (t,r) in terms of (τ,ρ). In fact, the necessary inverse change of coordinates is a slight extension of the change used by several authors[14,13] to deal with the model of gravitational collapse of Oppenheimer and Snyder[2] and its rotational perturbation.[13] Although the topology and physical meaning of both problems are completely different, the local problem of matching both metrics at a spherical surface is mathematically the same as the one discussed above. The only different minor details are the relative positions of the vacuum and dust solutions and the fact that in the problem of collapse one selects length units to have $k = 8\pi GA > 0$. Of course, there is no Λ term in the latter problem and instead of Eq. (10) other conditions[13] must be imposed.

In order to establish the relation between these two changes of coordinates used in different kinds of problems, we shall sketch the procedure to match both metrics in the coordinates (τ,ρ). The change from the coordinates (t,r) to (τ,ρ) is given by

$$t = -C_0 \int \left. \frac{dR}{B_0(R) h(R)} \right|_{R = S(\tau,\rho)}, \quad r = \rho R(\tau), \tag{28}$$

where the function $S(\tau,\rho)$ must be determined in the matching process. The latter can be accomplished in a way similar to that used in Secs. II–IV to obtain, obviously, the same final results. The relation between the changes (11) and (28) is given in terms of the functions defined in (13) by

$$S(\tau,\rho) = U(F_1(R(\tau)) + F_2(\rho)), \tag{29}$$

where the function U is implicitly defined by means of function $G(t)$:

$$G\left(-C_0 \int \left. \frac{dR}{B_0(R) h(R)} \right|_{R = U(x)} \right) = x. \tag{30}$$

Finally, we want to comment on the *approximate* nature of our solution. The problem has been solved to first order in the rotation parameter ϵ. If higher orders were considered, new features would appear in the situation. In fact, it is to be expected that the solution to second order in ϵ should require in general a nonspherical shape for the boundary of the Demianski cavity. This is so because the centrifugal force only becomes effective to second order in the angular velocity of the dust (second order in ϵ), therefore distorting to this order the originally spherical shape of the boundary. The rate of expansion of the boundary should also be in general latitude dependent in the second order, with that dependence determined by the initial conditions of the dust. This bears some resemblance to the results obtained by Brill and Cohen[16] and Pfister and Braun[17] in their studies of the Machian induction of the inertial forces by a rotating shell. Pfister and Braun were able to extend Brill and Cohen's induction of the Coriolis force (first order in the angular velocity of the shell) to the centrifugal force (second order in the angular velocity of the shell) by allowing for a prolate shell and a latitude-dependent mass density instead of the spherical and homogeneous shell considered by the latter authors.

It is perhaps worth stressing that our results do not guarantee the existence of an exact exterior cosmological solution matched to the Demianski cavity. However, the possibility of obtaining first-order results is a necessary con-

dition for such a solution to exist at all.

We do not see *a priori* reasons suggesting that an infinitesimally thin wall at the boundary of the cavity must become necessary at higher orders. Postulating singular domain walls in embedding problems relaxes the requirement of the matching of the extrinsic curvatures and therefore makes the embedding much easier. However, under the usual conditions of our present universe, most embeddings with an infinitesimal wall should be regarded as limiting cases of the more realistic smooth embeddings, where continuity of the extrinsic curvature holds in addition to that of the metric. An example in this direction is that of expanding voids in the universe: Their relativistic treatment has usually been undertaken within the context of the thin wall approximation.[18] However, the existence of expanding voids without thin walls smoothly embedded in asymptotically Friedmann–Tolman universes can be shown.[19]

ACKNOWLEDGMENTS

This work has been performed under contract Nos. 172.310-0030/88 and 172.310-0166/89 from the Universidad del País Vasco/Euskal Herriko Unibertsitatea.

[1] A. Einstein and E. G. Straus, Rev. Mod. Phys. **17**, 120 (1945).
[2] E. Schücking, Z. Phys. **137**, 595 (1954).
[3] R. Balbinot, R. Bergamini, and A. Comastri, Phys. Rev. D **38**, 2415 (1988).
[4] C. C. McVittie, Mon. Not. R. Astron. Soc. **93**, 325 (1937); Ann. Inst. H. Poincaré **40**, 235 (1984).
[5] P. A. M. Dirac, Proc. R. Soc. London Ser. A **345**, 19 (1978).
[6] R. Gautreau, Phys. Rev. D **27**, 764 (1983); **29**, 186 (1984); **29**, 198 (1984).
[7] C. C. Dyer, Mon. Not. R. Astron. Soc. **189**, 189 (1979); D. W. Olson and J. Silk, Astrophys. J. **233**, 395 (1979); C. Bona and J. Stela, Phys. Rev. D **36**, 2915 (1987).
[8] See, for instance, S. K. Blau, E. I. Guendelman, and A. H. Guth, Phys. Rev. D **35**, 1747 (1987) and references therein.
[9] A. Chamorro, Gen. Rel. Grav. **20**, 1309 (1988).
[10] M. Demianski, Acta Astron. **23**, 197 (1973); A. Krasinski, Ann. Phys. **112**, 22 (1978).
[11] H. Stephani, *General Relativity* (Cambridge U.P., Cambridge, 1982).
[12] J. R. Oppenheimer and H. Snyder, Phys. Rev. **56**, 455 (1939).
[13] L. S. Kegeles, Phys. Rev. D **18**, 1020 (1978).
[14] S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).
[15] W. Israel, Nuovo Cimento **44B**, 1 (1966); Nuovo Cimento **48B**, 463E (1967).
[16] D. Brill and J. M. Cohen, Phys. Rev. **143**, 1011 (1966).
[17] H. Pfister and K. H. Braun, Class. Quantum Grav. **2**, 909 (1985).
[18] See, for instance, K. Maeda and H. Sato, Prog. Theor. Phys. **70**, 772, 1276 (1983); K. Lake and R. Pim, Astrophys. J. **298**, 439 (1985).
[19] W. B. Bonnor and A. Chamorro (to be published).

Higher-dimensional Vaidya metric with an electromagnetic field

S. Chatterjee
New Alipore College, Calcutta 700053, India

B. Bhui and A. Banerjee
Department of Physics, Jadavpur University, Calcutta 700032, India

(Received 13 February 1990; accepted for publication 2 May 1990)

An exterior solution is obtained for a charged radiating sphere in higher dimensions. The solution reduces to an earlier one obtained by Krori and Barua [J. Phys. A 7, 2125 (1974)] when the space-time dimension is four, and to one obtained by Iyer and Vishveshwara [J. Phys. **32**, 749 (1989)] when the electromagnetic field is switched off.

I. INTRODUCTION

Finding a theory that unifies gravity with other forces in nature remains an elusive goal in quantum field theory. Most recent efforts in this search have been directed at studying theories in which the dimensions of space-time is greater than the $(3+1)$ of the world that we observe. The earlier suggestion of Kaluza and Klein that the topology of an extra dimension is a circle S' of very small radius, obtaining in this way a unified theory of gravitation and electromagnetism has now been replaced by what is called spontaneous compactification according to which solutions to $(4+k)$-dimensional Einstein's equations exist for which 4-D space-time expands while extra dimensions contract or remain constant at planckian length.[1] It has also been suggested[2] that the experimental detection of the time variation of fundamental constants could provide strong evidence for the existence of extra dimensions. Higher-dimensional cosmological models have been studied among others, by Chodos and Detweiler,[3] Demianski *et al.*[4] and Chatterjee.[5] In regarding to localized sources, mention may be made of Myers and Perry,[6] Dianyan,[7] Chatterjee[8] and Koikawa.[9]

But, to our knowledge, most of the works for localized bodies in higher dimensions are related to static sources. While, during many stages of stellar evolution, the variation of the physical parameters with time is so slow that a quasi-static approximation is justified, this method fails for stars evolving very rapidly from one stage to another. Following the detection of quasi-stellar objects (QSOs) and other extra galactic sources and their colossal energy requirements Hoyle and Fowler[10] suggested a theory of hot, convective supermassive stars where general relativistic effects can no longer be neglected. Further since a nonstatic starlike object, in general, would be radiating energy and it may contain electric charges as well,[11] models have been considered by allowing outgoing radiation to the collapsing body (Vaidya,[12,13] Lindquist,[14] Israel[15]). In the context of the above we have thought it worthwhile to investigate Vaidya's metric in higher dimensions with an electromagnetic field. Our solution may be described as a higher-dimensional generalization of earlier works in this field in the sense that when the space-time dimension is four our solution reduces to that of Krori and Barua[16] and when the electric field is absent our solution reduces to recent works of Iyer and Vishveshwara.[17]

II. BASIC EQUATIONS AND THEIR SOLUTIONS

An appropriate metric for a higher-dimensional spherically symmetric, nonstatic space-time may be taken as

$$ds^2 = e^{2\phi}\,dt^2 - e^{2\lambda}\,dr^2 - r^2\,d\Omega_n^2, \tag{2.1}$$

where ϕ and λ are functions of r and t and

$$d\Omega_n^2 = d\theta_n^2 + \sin^2\theta_n(d\theta_{n-1}^2 + \sin^2\theta_{n-1}(d\theta_{n-2}^2 + \cdots \sin^2\theta_2\,d\theta_1^2)) \tag{2.2}$$

is the metric on the n sphere in polar coordinates and $n = D-2$, where D is the total number of dimensions. The energy-momentum tensor corresponding to a charged radiating sphere is given by

$$T_{ab} = \rho V_a V_b + E_{ab}, \tag{2.3}$$

where ρ is the density of radiation and E_{ab} is the electromagnetic energy momentum tensor. Since the lines of flow are null geodesics,

$$V_\mu V^\mu = 0. \tag{2.4}$$

For radial outflow of radiation we have

$$V^3 = V^4 \cdots = V_{n+2}^{n+2} = 0, \tag{2.5}$$

and

$$\begin{aligned} T_1^1 &= \rho V_1 V^1 + \tfrac{1}{2}E, \\ T_2^2 &= \rho V_2 V^2 + \tfrac{1}{2}E, \\ T_2^1 &= \rho V_2 V^1, \\ T_3^3 &= T_4^4 = \cdots T_{n+2}^{n+2} = -E/2, \end{aligned} \tag{2.6}$$

where suffix 1 refers to time and 2 to radial coordinate.

Following *Xu* Dianyan one can write the Einstein–Maxwell equations in higher dimensions as follows:

$$E_{ab} = F_a^c F_{bc} - \tfrac{1}{4}g_{ab}F_{cd}F^{cd}, \tag{2.7}$$

$$F_{b;b}^a = 0, \tag{2.8}$$

$$F_{ab;c} + F_{bc;a} + F_{ca;b} = 0, \tag{2.9}$$

$$F_{ab} = A_{a,b} - A_{b,a}, \tag{2.10}$$

where A_a is the electromagnetic vector potential in D dimensions.

The components of the electromagnetic field tensor, not equal to zero, are

 0022-2488/90/092208-03$03.00

$$F_{12} = -F_{21} = q/r^n \tag{2.11}$$

and the corresponding vector potential is

$$A_1 = q/(n-1)r^{n-1} \tag{2.12}$$

where q is the total charge within the sphere and

$$E = q^2/r^{2n}. \tag{2.13}$$

From the field equations given by Iyer and Vishveshwara we get

$$e^{-2\lambda}\left(\frac{n\lambda'}{r} - \frac{n(n-1)}{2r^2}\right) + \frac{n(n-1)}{2r^2} = T_1^1 = \rho V_1 V^1 + \frac{E}{2}, \tag{2.14}$$

$$+ e^{-2\lambda}\left[-n\frac{\phi'}{r} - \frac{n(n-1)}{2r^2}\right] + \frac{n(n-1)}{2r^2} = T_2^2 = \rho V_2 V^2 + \frac{E}{2}, \tag{2.15}$$

$$- e^{-2\lambda}\left[+\phi'' + \phi'^2 - \phi'\lambda' - \frac{(n-1)(\lambda'-\phi')}{r} + \frac{(n-1)(n-2)}{2r^2}\right] + e^{-2\phi}(\ddot{\lambda} + \dot{\lambda}^2 - \dot{\lambda}\dot{\phi}) + \frac{(n-1)(n-2)}{2r^2} = T_3^3 = T_4^4 \cdots T_{n+2}^{n+2} = -E/2, \tag{2.16}$$

$$ne^{-2\phi}(\dot{\lambda}/r) = T_2^1 = \rho V_2 V^1, \tag{2.17}$$

where the overhead dot and prime refer to differentiation with respect to t and r, respectively.

From $T_2^1 \exp(\phi - \lambda) + T_1^1$ we get

$$e^{-2\lambda}\left[+n\frac{\lambda'}{r} - \frac{n(n-1)}{2r^2}\right] + \frac{n(n-1)}{2r^2} + ne^{-(\lambda+\phi)}\frac{\dot{\lambda}}{r} = \frac{q^2}{2r^{2n}}, \tag{2.18}$$

while Eqs. (2.14) and (2.15) give

$$ne^{-2\lambda}\left[+\frac{\lambda'-\phi'}{r} - \frac{n(n-1)}{r^2}\right] + \frac{n(n-1)}{r^2} = \frac{q^2}{r^{2n}}. \tag{2.19}$$

Assuming a particular form of the metric coefficient[7]

$$e^{-2\lambda} = 1 - \frac{2m(r,t)}{(n-1)r^{n-1}} + \frac{q^2}{n(n-1)r^{2n-2}}, \tag{2.20}$$

we get from Eq. (2.18)

$$e^{\phi} = -(\dot{m}/m')e^{\lambda}. \tag{2.21}$$

Using the operator

$$\frac{d}{d\tau} = V^2\frac{\partial}{\partial r} + V^1\frac{\partial}{\partial t}, \tag{2.22}$$

the last equation can be expressed as

$$\frac{dm}{d\tau} = 0. \tag{2.23}$$

From Eqs. (2.19), (2.20), and (2.21) we further get

$$\left(\frac{\dot{m}'}{\dot{m}} - \frac{m''}{m'}\right)\left[1 - \frac{2m}{(n-1)r^{n-1}} + \frac{q^2}{n(n-1)r^{2n-2}}\right] = \frac{2m}{r^n} - \frac{2q^2}{nr^{2n-1}}, \tag{2.24}$$

which yields a first integral

$$m'\left(1 - \frac{2m}{(n-1)r^{n-1}} + \frac{q^2}{n(n-1)r^{2n-2}}\right) = f(m), \tag{2.25}$$

where $f(m)$ is an arbitrary function of mass.

To check whether our solution satisfies the remaining field equation (2.16) also, we may use the rth component of the energy conservation equation

$$T^a_{b;a} = 0, \tag{2.26}$$

which gives

$$T^2_{2,2} + T^1_{2,1} + \phi'(T_2^2 - T_1^1) + \frac{n}{r}(T_2^2 - T_3^3) + T_2^1(\dot{\phi} + \dot{\lambda}) = 0. \tag{2.27}$$

We finally obtain via Eqs. (2.14), (2.18), and (2.19)

$$nT_3^3 = -r^{n+1}\exp(3\lambda)\frac{d}{d\tau}\left[m'\left(1 - \frac{2m}{(n-1)r^{n-1}} + \frac{q^2}{n(n-1)r^{2n-2}}\right)\right] - \frac{nE}{2}. \tag{2.28}$$

thus

$$T_3^3 = -E/2, \tag{2.29}$$

as is evident from Eqs. (2.25) and (2.23).

If we, at this stage, introduce a coordinate $u = u(m)$ defined by

$$du = -\frac{dm}{f(m)} = -\left(dr + \frac{\dot{m}}{m'}dt\right) \times \left(1 - \frac{2m}{(n-1)r^{n-1}} + \frac{q^2}{n(n-1)r^{2n-2}}\right) \tag{2.30}$$

the line element describing the radiation envelope of a charged sphere in D dimensions reduces to

$$ds^2 = \left(1 - \frac{2m(u)}{(n-1)r^{n-1}} + \frac{q^2}{n(n-1)r^{2n-2}}\right)du^2 + 2\,du\,dr - r^2\,d\Omega_n^2. \tag{2.31}$$

ACKNOWLEDGMENTS

S.C. wishes to thank the U.G.C. and B.B. and A.B. the C.S.I.R., New Delhi for financial support in this work.

[1] A. Salam and J. Strathdee, Ann. Phys. **141**, 316 (1982).
[2] W. J. Marciano, Phys. Rev. Lett. **52**, 489 (1984).
[3] A. Chodos and S. Detweiler, Phys. Rev. D **21**, 2167 (1980).

[4] M. Demianski, Z. Golda, L. M. Sokolowski, and P. Turkowski, J. Math. Phys. **28**, 171 (1987).
[5] S. Chatterjee, Astron. Astrophys. **179**, 1 (1987).
[6] R. Myer and M. Perry, Ann. Phys. (N.Y.) **172**, 304 (1986).
[7] Xu Dianyan, Class. Quantum Grav. **5**, 871 (1988).
[8] S. Chatterjee, Astron. Astrophys. (in press).
[9] T. Koikawa, Phys. Lett. A **17**, 273 (1986).
[10] F. Hoyle and W. A. Fowler, Nature **197**, 533 (1963).
[11] V. F. Shvartsman, Sov. Phys., JETP **33**, 475 (1971).
[12] P. C. Vaidya, Proc. Indian Acad. Sci. A **33**, 264 (1951).
[13] P. C. Vaidya, Astrophys. J. **144**, 943 (1966).
[14] R. W. Lindquist, R. A. Schwartz, and C. W. Misner, Phys. Rev. B **1364**, 137 (1965).
[15] W. Israel, Phys. Lett. A **24**, 184 (1967).
[16] K. D. Krori and J. Barua, J. Phys. A **7**, 2125 (1974).
[17] B. R. Iyer and C. V. Vishveshwara, Pramana J. Phys. **32**, 749 (1989).

Symmetries of space-time and geodesic symmetries

Diego Del-Castillo-Negrete and Sergio Hojman[a),b)]
Instituto de Ciencias Nuclears, Universidad Nacional Autónoma de México, Apartado Postal 70-543, 04510 México, D. F., Mexico

(Received 9 November 1988; accepted for publication 9 May 1990)

It is shown how to construct symmetries of the geodesic equation starting from space-time symmetries. The constants of motion associated with space-time symmetries are recovered and a few new ones are found using non-Noetherian conservation theorems. Explicit examples are presented.

I. INTRODUCTION

Symmetries of space-time have been extensively discussed by several authors.[1-3] Davis, Katzin, and Levine[2,3] have carefully studied their properties and constructed conserved quantities associated with geodesic motion on space-times that possess those symmetries. Nevertheless, certain space-time symmetries are not geodesic symmetries in the usual sense,[4-7] and, furthermore, there does not seem to be a systematic way to associate certain constants of motion to symmetries of space-time.

In this paper, we find geodesic symmetries associated with space-time symmetries and construct constants of motion using non-Noetherian conservation theorems developed earlier,[6,7] recovering known constants and exhibiting a few new ones.

The main idea is the following: The geodesic equation on any space-time is an autonomous equation and has therefore a symmetry vector $\dot{q}^i$ tangent to the geodesic curves. The space-time symmetries may be used to construct a tensor, which we call a special symmetry map, that maps the symmetry vector $\dot{q}^i$ into a new (sometimes non-Noetherian) symmetry vector of the geodesics. Special symmetry maps obey an equation that generalizes that of Killing tensors. Special symmetry maps may be interesting by themselves, even if not related to space-time symmetries.

In Sec. II we present a brief summary of previous results on symmetries of geodesic equations and non-Noetherian conservation laws. In Sec. III we define and explore the concept of symmetry maps. In Sec. IV we establish the relationship linking symmetries of space-time to geodesic symmetries and find the new constants of motion associated with them. In Sec. V we present explicit examples to apply the results obtained in Sec. IV. Section VI presents a summary and the conclusions.

II. SUMMARY OF RECENT RESULTS

In this section we summarize the relevant results recently obtained.[5,6] Consider the geodesic equations (space-time indices run from 1 to N);

$$\ddot{q}^i + \Gamma^i_{jk}\dot{q}^j\dot{q}^k = 0, \quad i,j,k = 1,...,N, \tag{2.1}$$

where

$$\dot{q}^j \equiv \frac{d}{ds} q^j, \tag{2.2}$$

with

$$ds^2 = g_{ij}(q^k)dq^i\,dq^j. \tag{2.3}$$

The Christoffel symbols Γ^i_{jk} are defined in terms of the Riemannian metric g_{ij} in the usual way,

$$\Gamma^i_{jk} = \tfrac{1}{2}g^{il}(-g_{jk,l} + g_{kl,j} + g_{lj,k}). \tag{2.4}$$

The geodesic equation (2.1) may be derived from the Lagrangian $\overline{L}$,

$$\overline{L} = \tfrac{1}{2}g_{ij}\dot{q}^i\dot{q}^j, \tag{2.5}$$

using the Euler–Lagrange operator E_i,

$$E_i \equiv \frac{d}{dt}\frac{\partial}{\partial \dot{q}_i} - \frac{\partial}{\partial q_i}. \tag{2.6}$$

There are several ways of associating conservation laws to the symmetries of the Lagrangian or the equations of motion. Noether's theorem is probably the best known of them all and we will not discuss it here. We will be mainly interested in non-Noetherian symmetries. It is easier to deal with them if the equations and Lagrangian are rewritten in first-order form. For that purpose consider the system

$$\dot{x}^a - f^a(x^b) = 0, \quad a,b = 1,...2N, \tag{2.7}$$

where

$$x^i = q^i, \quad x^{i+N} = \dot{q}^i, \tag{2.8}$$

$$f^i = x^{i+N}, \quad f^{i+N} = -\Gamma^i_{jk}x^{j+N}x^{k+N}. \tag{2.9}$$

It may be proved that the Lagrangian for Eq. (2.7) may always be written as[5-9]

$$L = \ell_a(x^b,s)(\dot{x}^a - f^a). \tag{2.10}$$

This fact may be easily understood considering Hamilton–Jacobi theory. Perform on *any* first-order Lagrangian $L_1 = p_i\dot{q}^i - H_1$ the canonical transformation (add to it a total time derivative), which leads from (p,q,H_1) to $(P,Q,K=0)$, where P and Q are constants of motion. The new Lagrangian L_2 can be written as

$$L_2 = P_i\dot{Q}^i.$$

The time derivative of *any* constant of motion is a linear combination of the left-hand side of the equations of motion of the system in question (i.e., it vanishes "on shell"), which leads to Eq. (2.10). The equations of motion for the Lagrangian (2.10) read

[a)] Current address: Center for Relativity, Department of Physics, University of Texas at Austin, Austin, Texas 78712.
[b)] Fellow of the John Simon Guggenheim Memorial Foundation.

 0022-2488/90/092211-06$03.00

$$\frac{d}{ds}\ell_b - \frac{\partial \ell_a}{\partial x^b}(\dot{x}^a - f^a) + \ell_a \frac{\partial f^a}{\partial x^b} = 0, \tag{2.11}$$

or

$$\frac{\partial}{\partial s}\ell_b + \frac{\partial \ell_b}{\partial x^a}\dot{x}^a - \frac{\partial \ell_a}{\partial x^b}(\dot{x}^a - f^a) + \ell_a \frac{\partial f^a}{\partial x^b} = 0. \tag{2.12}$$

Adding and subtracting $(\partial \ell_b/\partial x^a)f^a$, one gets

$$\left(\frac{\partial \ell_b}{\partial x^a} - \frac{\partial \ell_a}{\partial x^b}\right)(\dot{x}^a - f^a) + \frac{\partial}{\partial s}\ell_b + \frac{\partial \ell_b}{\partial x^a}f^a + \frac{\partial \ell_a}{\partial x^b}f^a = 0, \tag{2.13}$$

which leads to the condition

$$(\partial_s + \mathcal{L}_f)\ell_a = 0, \tag{2.14}$$

in order to reproduce Eqs. (2.7) (Here, ∂_s denotes the partial derivative with respect to s and $\mathcal{L}_f$ is the Lie derivative along f.) One may write ℓ_a [for Eqs. (2.7)–(2.9)] as

$$\ell_i = -\left(\frac{\bar{d}}{ds}\mu_i + \mu_j \frac{\partial f^{j+N}}{\partial x^{i+N}}\right), \quad \ell_{i+N} = \mu_i,$$
$$\mu_i = -sg_{ij}x^{j+N}, \tag{2.15}$$

and

$$\frac{\bar{d}}{ds} \equiv f^a \frac{\partial}{\partial x^a} + \partial_s. \tag{2.16}$$

The equations of motion for the Lagrangian (2.10) may be written as

$$-E_a L = \sigma_{ab}(\dot{x}^b - f^b) = 0, \tag{2.17}$$

when Eq. (2.14) is taken into account. The operator E_a is given by

$$E_a \equiv \frac{d}{ds}\frac{\partial}{\partial \dot{x}^a} - \frac{\partial}{\partial x^a}, \tag{2.18}$$

and the implectic (for inverse symplectic) two-form σ_{ab} is defined by

$$\sigma_{ab} = \frac{\partial \ell_a}{\partial x^b} - \frac{\partial \ell_b}{\partial x^a}. \tag{2.19}$$

One must then require

$$\det \sigma_{ab} \neq 0, \tag{2.20}$$

to guarantee that Eqs. (2.17) and (2.7) be equivalent. The Lagrangian (2.15) satisfies (2.20). It is straightforward to prove that σ_{ab} satisfies

$$(\partial_s + \mathcal{L}_f)\sigma_{ab} = 0, \tag{2.21}$$

on account of Eq. (2.14).

We define an infinitesimal transformation

$$x'^a = x^a + \epsilon \eta^a(x^b, s), \tag{2.22}$$

to be a symmetry transformation for Eq. (2.7) if it maps the space of solutions of Eq. (2.7) in itself, i.e., if η^a satisfies, to first order in ϵ,

$$(\partial_s + \mathcal{L}_f)\eta^a = 0. \tag{2.23}$$

Note that

$$\eta^a = f^a, \tag{2.24}$$

is always a solution to Eq. (2.23) for f^a given by Eq. (2.9), as it was mentioned in the Introduction.

Once a symmetry vector η^a is known, it is direct to prove that the one-form ℓ'_a

$$\ell'_a = \mathcal{L}_\eta \ell_a, \tag{2.25}$$

is Lagrangian, i.e.,

$$(\partial_s + \mathcal{L}_f)\ell'_a = 0. \tag{2.26}$$

To prove Eq. (2.26), it is helpful to use Eq. (2.23) and Eqs. (2.27)–(2.29) below:

$$\mathcal{L}_f \eta = [f, \eta], \tag{2.27}$$

$$[\mathcal{L}_f, \mathcal{L}_\eta] = \mathcal{L}_{[f,\eta]}, \tag{2.28}$$

and

$$\partial_s \mathcal{L}_\eta \ell = \mathcal{L}_\eta \partial_s \ell + \mathcal{L}_{\partial_s \eta}\ell. \tag{2.29}$$

The two form σ'_{ab}

$$\sigma'_{ab} = \mathcal{L}_\eta \sigma_{ab} = \frac{\partial \ell'_a}{\partial x^b} - \frac{\partial \ell'_b}{\partial x^a}, \tag{2.30}$$

is implectic, i.e.,

$$(\partial_s + \mathcal{L}_f)\sigma'_{ab} = 0. \tag{2.31}$$

Define Λ^b_a by

$$\Lambda^b_a = \sigma'_{ac}(\sigma^{-1})^{cb}, \tag{2.32}$$

then

$$(\partial_s + \mathcal{L}_f)\Lambda^b_a = 0, \tag{2.33}$$

because of Eqs. (2.21) and (2.31). It is then evident that

$$\ell''_a = \sigma_{ab}\eta^b, \tag{2.34}$$

and

$$\ell'''_a = \Lambda^b_a \ell_b \tag{2.35}$$

are Lagrangian one forms, while

$$\eta''^a = (\sigma^{-1})^{ab}\ell_b \tag{2.36}$$

$$\eta'''^a = \eta^b \Lambda^a_b \tag{2.37}$$

are symmetry vectors, because ℓ_a, η^b, σ_{ab}, and Λ^a_b satisfy Eqs. (2.14), (2.23), (2.21), and (2.33), respectively.

We end this section by writing down the following non-Noetherian conservation laws.[6] Let

$$\dot{J} = 0, \quad J = \ell_a \eta^a, \tag{2.38}$$

$$\dot{I}_k = 0, \quad I_k = \operatorname{tr}(\Lambda)^k, \tag{2.39}$$

$$\dot{M}_{12} = 0, \quad M_{12} = \eta^a_1 \sigma_{ab} \eta^b_2, \tag{2.40}$$

$$\dot{N}_{12} = 0, \quad N_{12} = \ell_{1a}(\sigma^{-1})^{ab}\ell_{2b}, \tag{2.41}$$

$$\dot{Q} = 0, \quad Q = \eta^a \Lambda^b_a \ell_b, \tag{2.42}$$

and

$$\dot{C}' = 0, \quad C' = \mathcal{L}_\eta C \text{ with } \dot{C} = 0. \tag{2.43}$$

Furthermore, if A is a constant of motion such that $A = dB/ds$ then

$$\dot{D} = 0, \quad D = As - B. \tag{2.44}$$

Equations (2.38) and (2.40)–(2.42) can also be found in Ref. 7, while Eq. (2.39) was first derived in Ref. 8. Katzin and Levine derived Eq. (2.43) in 1968 (see Ref. 2). Equations (2.40), (2.41), and (2.43) are, in some sense, generalizations of Poisson's theorem involving the Poisson bracket of two constants of motion.

III. SYMMETRY MAPS

In the preceding section we showed that Λ^b_a is a universal symmetry map, i.e., given any symmetry vector η^a, Λ^b_a produces a new symmetry vector according to Eq. (2.37).

This procedure is particularly useful in some instances, and, as a matter of fact, it provides one of the methods to completely solve some nonlinear problems such as the Korteweg–de Vries equation.[10] It may then be interesting to explore a bit more the concept of symmetry maps. Consider now a special symmetry map $\mathscr{K}^b_a$ such that for a *fixed* symmetry vector η^a, $\mathscr{K}^b_a$ produces $\tilde{\eta}^b$;

$$\tilde{\eta}^b = \eta^a \mathscr{K}^b_a, \tag{3.1}$$

which is also a symmetry vector.

We know that for geodesics, f_a is always a symmetry vector of the problem [see Eq. (2.24)]. We will then study the special symmetry maps $\mathscr{K}^b_a$ such that $\tilde{\eta}^a$ is a symmetry vector defined by

$$\tilde{\eta}^b = f^a \mathscr{K}^b_a. \tag{3.2}$$

It is now convenient to return to the second-order formalism and write down the symmetry equation for the first N components of the symmetry vector η^a. The first N equations contained in (2.23) state simply that (space-time indices run from 1 to N),

$$\eta^{i+N} = \frac{\bar{d}}{ds}\eta^i \quad (i = 1,...,N), \tag{3.3}$$

while the second N equations give rise to

$$\frac{\bar{D}}{DS}\frac{\bar{D}}{DS}\eta^i + \mathscr{R}^i_{jkl}\dot{q}^j\dot{q}^k\eta^l = 0 \tag{3.4}$$

where the Riemann tensor $\mathscr{R}^i_{jk,l}$ is given by

$$\mathscr{R}^i_{jkl} = \Gamma^i_{jk,l} - \Gamma^i_{jl,k} + \Gamma^m_{jk}\Gamma^i_{ml} - \Gamma^m_{jl}\Gamma^i_{mk}, \tag{3.5}$$

and

$$\frac{\bar{D}}{DS}\eta^i = \frac{\bar{d}}{ds}\eta^i + \Gamma^i_{jk}\dot{q}^j\eta^k, \tag{3.6}$$

once Eqs. (3.3) have been used. Here, $(\bar{D}/D_s)\,\eta^i$ is the "on shell" covariant derivative of η^i.

If we define $\tilde{\eta}^i = k^i_j\eta^j$ for $i,j = 1,...,N$, then it can be easily seen that the structure of $\mathscr{K}_{ab}$ is given by

$$\mathscr{K}_{ab} = \begin{pmatrix} k_{ij} & 0 \\ \dfrac{\bar{d}}{ds}k_{ij} & k_{ij} \end{pmatrix}, \tag{3.7}$$

when Eqs. (3.3) are taken into account.

We will consider the case in which k_{ij} is symmetric and depends on q^k only. After a bit of algebra, Eq. (3.4) reduces, for $\tilde{\eta}$ given by Eq. (3.2), to

$$k_{(ij;k);l} = 0. \tag{3.8}$$

Equation (3.8) is the basic equation to define the special symmetry maps with which we deal in this paper. Note that it provides a generalization of the concept of Killing tensors much in the same way affine collineations generalize the concept of Killing vectors. In the next section we will show how Eq. (3.8) has room enough to accommodate the symmetries of space-time (which are not symmetries of the geodesic equations) that have been extensively studied by Davis, Katzin, and Levine.[2,3] Katzin and Levine found equations similar to Eq. (3.8) in Ref. 3. We will also be able to link these space-time symmetries to geodesic symmetries.

IV. SYMMETRIES OF SPACE-TIME AND GEODESIC SYMMETRIES

Davis, Katzin, and Levine have very carefully classified and studied many symmetries ξ^i of Riemannian space-times.[2,3] They have also written down conserved quantities associated with geodesic motion on space-times having those symmetries. Nevertheless, most of the symmetries considered are *not* geodesic symmetries in the sense that they do *not* satisfy Eq. (3.4). Moreover, there does not seem to be a systematic way to assign conservation laws to space-time symmetries [simply because they do not seem directly related to the geodesic (or geodesic symmetry) equations].

In this section we prove that all space-time symmetries (which give rise to conservation laws) defined by Davis, Katzin, and Levine may be associated with special symmetry maps k_{ij} that satisfy Eq. (3.8). Therefore any space-time symmetry defines a geodesic (either Noetherian or non-Noetherian) symmetry η^i through

$$\eta^i = k^i_j\dot{q}^j, \tag{4.1}$$

which satisfies Eq. (3.4).

Furthermore, all the constants of motion found by Davis, Katzin, and Levine (plus a few others) are obtained by using the non-Noetherian conservation laws (2.38)–(2.44).

We summarize these results in three tables. In Table I we list the defining equations of the different kinds of space-time symmetries and show some ways of associating special symmetry maps k_{ij} [which satisfy Eq. (3.8)] with each of them.

In Table II, we show the constants of motion that can be associated with special symmetry maps, such as those constructed in Table I, according to the scheme developed in Sec. II. The numbers in parentheses refer to the formulas at the end of Sec. II, which define constants of motion associated with geodesic symmetries.

For those symmetries of space-time that are also geodesic symmetries (such as motions or affine collineations), the definition of the special symmetry map is, in some sense, irrelevant because one may simply define

$$\eta^i = \xi^i, \tag{4.2}$$

as a geodesic symmetry.

We will, nevertheless, include a symmetry map associated with affine collineations [which satisfies Eq. (3.8)] for completeness. In the case of motions the symmetry map is trivial and will be excluded.

In Table III we list constants of the motion associated with geodesic symmetries that we have not found in the literature. The equation numbers refer to Sec. II. All the known constants may also be obtained using the results of Sec. II.

Katzin and Levine[3] have constructed constants of motion that depend explicitly on s and are polynomial both in s and in the momentum p^μ. The coefficient in front of each

TABLE I. Definition of space-time symmetries and associated special symmetry maps.

Symmetry of space-time	Notation	Defining equation	Special symmetry map
Homothetic motion	HM	$\xi_{(i,j)} = \sigma_0 g_{ij}$ $\sigma_0 = \text{const}$	$k_{ij} = \xi_{(i,j)}$
Affine collineation	AC	$\xi_{(i,j);k} = 0$	$k_{ij} = \xi_{(i,j)}$
Projective collineation	PC	$2\xi_{(i,j);k} = 2g_{ij}\varphi_{,k} + g_{ik}\varphi_{,j} + g_{ik}\varphi_{,i}$	$k_{ij} = \xi_{(i,j)} - 2\varphi g_{ij}$
Special projective collineation	SPC	$2\xi_{(i,j);k} = 2g_{ij}\varphi_{,k} + g_{ik}\varphi_{,j} + g_{jk}\varphi_{,i}$ $\varphi_{;ij} = 0$	$k_{ij} = \xi_{(i,j)}$ $k_{ij} = \xi_{(i,j)} - 2\varphi g_{ij}$
Conformal motion	CM	$\xi_{(i,j)} = \sigma g_{ij}$	$k_{ij} = \xi_{(i,j)} - \sigma g_{ij}$
Special conformal motion	SCM	$\xi_{(i,j)} = \sigma g_{ij}$ $\sigma_{;ij} = 0$	$k_{ij} = \xi_{(i,j)}$ $k_{ij} = \xi_{(i,j)} - \sigma g_{ij}$
Conformal collineation	CONFC	$\xi_{(i,j);k} = \tau_{,k} g_{ij}$	$k_{ij} = \xi_{(i,j)} - \tau g_{ij}$
Special conformal collineation	SCONFC	$\xi_{(i,j);k} = \tau_{,k} g_{ij}$ $\tau_{;ij} = 0$	$k_{ij} = \xi_{(i,j)}$ $k_{ij} = \xi_{(i,j)} - \tau g_{ij}$
Special curvature collineation	SCC	$\xi_{(i,j);k,k} = 0$	$k_{ij} = \xi_{(i,j)}$

term with increasing powers of s and p^μ can be obtained in Katzin and Levine's case (up to a constant) by taking the covariant derivative of the coefficient in the preceding term of the polynomial. (See, for instance, Table IV in Ref. 3.)

Although we have followed a very different procedure, we have recovered their results, and the constants of motion that appear in Table II have exactly the structure described above.

Nevertheless, the constants listed in Table III have a different, more elaborate structure, and we have not found these new constants in the literature.

V. EXAMPLES

In this section we present some solutions to Eq. (3.8) for the special case of metrics representing plane-fronted gravitational waves with parallel rays or p-p waves.

TABLE II. Constant of motion associated with special symmetry maps k_{ij}.

	Constant of motion associated with SSM	Equation number
$C_{(1)} \equiv$	$g_{ij}k^{j}{}_{;m}p^i p^m p'$	(2.40)
$C_{(2)} \equiv$	$s g_{ij}k^{j}{}_{;m}p^i p^j p^m - g_{ij}k^{j}{}_{;l}p^i p^l$	(2.40)
$C_{(3)} \equiv$	$\xi_i p^i - s k_{ij}p^i p^j + (s^2/2)k_{ij,l}p^i p^j p^l$ [a]	(2.44)
$C_{(4)} \equiv$	$G_{ij}k^{j}{}_{;m}p^i p^j p^m$ [b]	(2.40) and (2.30)
$C_{(5)} \equiv$	$s G_{ij}k^{j}{}_{;m}p^i p^j p^m - G_{ij}k^{j}{}_{;l}p^i p^l$	(2.40) and (2.30)
$C_{(6)} \equiv$	$2\,\mathrm{tr}(G_{ij})^k,\ k = 1,2,\ldots$	(2.44)

$p^i \equiv \dfrac{dq^i}{ds}$

[a] $k_{ij} = \xi_{(i,j)}$.
[b] $G_{ij} = 3k_{(i,j,l)}p^l$.

We present some solutions to Eq. (3.8) for the special case of metrics representing plane-fronted gravitational waves with parallel rays or p-p waves.

The metric of a p-p wave can be written in terms of coordinates (ρ,σ,z,z^*):[11,12]

$$ds^2 = 2\,d\rho\,d\sigma - 2\,dz\,dz^* - 2H\,d\sigma^2, \tag{5.1}$$

where ρ,σ are real and z,z^* are a complex variable and its complex conjugate.

For the metric (5.1), the vacuum field equations require that

$$\frac{\partial H}{\partial \rho} = \frac{\partial^2 H}{\partial z\,\partial z^*} = 0. \tag{5.2}$$

The vectors

$$\begin{aligned} &\ell^i = (1,0,0,0), \qquad && n^i = (H,1,0,0), \\ &m^i = (1/\sqrt{2})(0,0,0,1), \qquad && m^{*i} = (1/\sqrt{2})(0,0,1,0), \end{aligned} \tag{5.3}$$

form a Newman–Penrose null tetrad.

In this formalism an arbitrary vector ξ is represented as (Greek indices run from 1 to 4);

$$\xi = \xi_\alpha e^\alpha = \xi^\alpha e_\alpha, \quad \alpha,\beta,\gamma,\delta,\ldots = 1,2,3,4, \tag{5.4}$$

where

$$\begin{aligned} &e_1 = e^2 = \ell, \qquad && e_2 = e^1 = n, \\ &e_3 = -e^4 = m, \qquad && e_4 = -e^3 = m^*, \end{aligned} \tag{5.5}$$

and a covariant tensor k of valence 2 as

$$k = k_{\alpha\beta} e^\alpha e^\beta. \tag{5.6}$$

The defining equation for special symmetry map (SSM), Eq. (3.8) is then:

$$k_{(\alpha\beta|\gamma)|\delta} = 0, \tag{5.7}$$

where the brackets denote total symmetrization, and the ver-

TABLE III. New constants of motion associated with symmetries of space-time.

Symmetry of space-time	New constants of motion	Equation number
SPC	$[3(\varphi_{,i}p^i)^2+(\varphi_{,i}\varphi^i{}_{,})]s-2\xi_{(i,j)}p^i\varphi^j,$	(2.40) and (2.30)
SCONFC	$\xi_{(i,j)}p^i[(\tau_{,k}p^k)p^j-\tau^j,]$	(2.40) and (2.30)
SCC	$h_{(ijk)}h^j{}_{l,m}p^ip^kp^lp^m$ [a]	(2.40) and (2.30)
SCC	$sh_{(ijk)}h^j{}_{l,m}p^ip^kp^lp^m-h_{(ijk)}h^i{}_{m}p^jp^kp^m$	(2.40) and (2.30)
SCC	$h_{(i}{}^{j_1}{}_{:k\,1)}h_{(j_1}{}^{j_2}{}_{:k\,2)}\cdots h_{(j_r}{}^{i}{}_{:k_r)}p^{k_1}\cdots p^{k_r}$ $r=1,2,\ldots$	(2.39)

$p^i\equiv\frac{dq^i}{ds}$

[a] $h_{ij}=\xi_{(i,j)}$.

tical bar represents the covariant derivated projected in the tetrad frame. As an example, for a vector this notation reads:

$$\xi_{\alpha|\beta}\equiv\xi_{i;j}e^i_\alpha e^j_\beta. \tag{5.8}$$

We now look for solutions to Eq. (5.7) using the metric defined by Eqs. (5.1) and (5.2).

Depending on the form of H,, several cases arise.

(*i*) *General p–p waves*: In this case, H is an arbitrary function of σ,z,z^* satisfying Eq. (5.2).

It is straightforward to prove that

$$k^{(1)}_{22}=A \tag{5.9}$$

is a (reducible) Killing tensor, where A is a constant. Upper indices in parenthesis are used to enumerate the different solutions, while lower indices denote as usual the covariant components of k on the tetrad.

Also, if B is a constant, the tensor,

$$k^{(2)}_{12}=B,\quad k^{(2)}_{34}=-B, \tag{5.10}$$

which is nothing but the metric multiplied by B, is trivially a Killing tensor.

In addition this space-time admits two SSM's, namely,

$$k^{(3)}_{22}=A\sigma, \tag{5.11}$$

and

$$k^{(4)}_{12}=B\sigma,\quad k^{(4)}_{34}=-B\sigma. \tag{5.12}$$

(*ii*) *Cylindrically symmetric p-p waves.* In this case, H takes the special form

$$H=\ell n(2zz^*)^{1/2}, \tag{5.13}$$

and

$$k^{(5)}_{22}=2AH,\quad k^{(5)}_{34}=-A, \tag{5.14}$$

is a Killing tensor.

We now establish the relation of SSM's with symmetries of space-time.

Some of these tensors can be written in terms of a vector ξ in the form;

$$k^{(i)}_{\alpha\beta}=\xi^{(i)}_{(\alpha|\beta)}. \tag{5.15}$$

For $k^{(1)}$, this vector is

$$\xi^{(1)}_2=A\sigma. \tag{5.16}$$

in the case of $k^{(5)}$, we have

$$\begin{aligned}&\xi^{(5)}_1=-A\sigma,\quad \xi^{(5)}_2=A(\rho+\sigma+\sigma H),\\ &\xi^{(5)}_3=-Az,\quad \xi^{(5)}_4=-z^*A.\end{aligned} \tag{5.17}$$

For $k^{(3)}$

$$\xi^{(3)}_2=(A/2)\sigma^2+B\sigma+C. \tag{5.18}$$

The tensor $k^{(2)}$ can be written in terms of a $\xi^{(2)}$ if we consider the restriction to the special case of general p-p waves known as plane linearly polarized waves defined by

$$H=\tfrac{1}{2}(z^2+z^{*2}). \tag{5.19}$$

In this case,the vector is

$$\begin{aligned}&\xi^{(2)}_1=0,\qquad \xi^{(2)}_2=2\beta\rho,\\ &\xi^{(2)}_3=-Bz,\quad \xi^{(2)}_4=-Bz^*.\end{aligned} \tag{5.20}$$

As far as we know the tensor $k^{(4)}$ cannot be written in the form (5.15). Due to the fact that $k^{(1)}$ is a Killing tensor, we conclude from (5.15) that $\xi^{(1)}$ defined in (5.16) is a *proper affine collineation* for general p-p waves.

Also, it is straightforward to verify that $\xi^{(2)}$ is a *homothetic motion* for plane linearly polarized waves.

On the other hand, the fact that $k^{(3)}$ is SSM implies that $\xi^{(3)}$ is a *proper special curvature collineation* for general p-p waves.

As far as we know, the vector $\xi^{(5)}$ used to write $k^{(5)}$ is not a space-time symmetry.

In what follows, we list the (independent) constants of motion associated with the SSM that we have found.

(*i*) *General waves.* Following the convention, an upper index in a constant indicates the number the tensor to which it corresponds and a lower index refers to the notation employed in Table II.

Let $\mathbf{p}\equiv d\mathbf{q}/ds$, then

$$C^{(1)}_{(2)}=A(\mathbf{p}\cdot\boldsymbol{\ell})^2 \tag{5.21}$$

$$C^{(1)}_{(3)}=A(\mathbf{p}\cdot\boldsymbol{\ell})[\sigma-s(\mathbf{p}\cdot\boldsymbol{\ell})] \tag{5.22}$$

$$C^{(3)}_{(3)}=A(\mathbf{p}\cdot\boldsymbol{\ell})[\sigma^2/2-s(\mathbf{p}\cdot\boldsymbol{\ell})+(s^2/2)(\mathbf{p}\cdot\boldsymbol{\ell})^2]. \tag{5.23}$$

(*ii*) *Cylindrically symmetric p-p waves*:

$$C^{(5)}_{(2)}=2A\,[H(\mathbf{p}\cdot\boldsymbol{\ell})^2-(\mathbf{p}\cdot\mathbf{m})(\mathbf{p}\cdot\mathbf{m}^*)]$$

$$C^{(5)}_{(3)} = C\,[-\sigma(\mathbf{p}\cdot\mathbf{n}) + \rho(\mathbf{p}\cdot\mathbf{l}) + \sigma(1+H)(\mathbf{p}\cdot\mathbf{l}) + z(\mathbf{p}\cdot\mathbf{m}^*) + z^*(\mathbf{p}\cdot\mathbf{m})] - 2As[H(\mathbf{p}\cdot\mathbf{l})^2 - (\mathbf{p}\cdot\mathbf{m})(\mathbf{p}\cdot\mathbf{m}^*)].$$

(*iii*) *Plane linearly polarized waves*:

$$C^{(2)}_{(3)} = B\,[2\rho(\mathbf{p}\cdot\boldsymbol{\ell}) + z(\mathbf{p}\cdot\mathbf{m}^*) + z^*(\mathbf{p}\cdot\mathbf{m}) - s(\mathbf{p}\cdot\mathbf{p})].$$

VI. SUMMARY AND CONCLUSIONS

We introduced the concept of symmetry maps, which generalizes the definition of Killing tensors and allows us to relate symmetries of space-time to geodesic symmetries. We recovered all the known constants associated with space-time symmetries and found some new ones.

Universal symmetry maps play an important role in the solutions of nonlinear problems, while special symmetry maps have proved to be useful in the study of geodesic motion. It may be interesting to consider generalizations of the concept of symmetry maps (using tensors with n indices and $n-1$ symmetry vectors, for instance) and to study in more detail and depth the possible applications of both universal and special symmetry maps.

[1] L. P. Eisenhart, *Riemannian Geometry* (Princeton U.P., Princeton, 1964).

[2] W. R. Davis and M. K. Moss, Nuovo Cimento **38**, 1558 (1965); G. H. Katzin and J. Levine, J. Math. Phys. **9**, 8 (1968); G. H. Katzin, J. Levine, and W. R. Davis, *ibid.* **10**, 617 (1969); W. R. Davis, M. K. Moss, and J. W. York, Jr., Nuovo Cimento B **65**, 19 (1970); G. H. Katzin and J. Levine, Colloq. Math. (Poland) **26**, 211 (1972).

[3] G. H. Katzin and J. Levine, J. Math. Phys. **18**, 1267 (1977); **22**, 1878 (1981).

[4] S. Lie, *Gesammelte Abhandlungen*, edited by F. Engel and P. Heegaard (Tuebner, Leipzig, 1922), Band III, and supplement; and *Vorlesungen uber Differentialgleichungen*, edited by G. Scheffers (Teubner, Leipzig, 1891); R. L. Anderson, S. Kumei, and C. E. Wulfman, Phys. Rev. Lwett. **28**, 988 (1972).

[5] G. Caviglia, C. Zordan, and F. Salmistraro, Int. J. Theo. Phys. **21**, 391 (1982); G. Caviglia, F. Salmistraro, and C. Zordan, J. Math. Phys. **23**, 2346 (1982); F. Salmistraro, Nuovo Cimento Lett. **36**, 35 (1983).

[6] S. Hojman, J. Phys. A **17**, 2399 (1984).

[7] S. Hojman, L. Núñez, A. Patiño, and H. Rago, J. Math. Phys. **27**, 281 (1986).

[8] S. Hojman and L. F. Urrutia, J. Math. Phys. **22**, 1896 (1981); S. Hojman and H. Harleston, *ibid.* **22**, 1414 (1981).

[9] R. Hojman, S. Hojman, and J. Sheinbaum, Phys. Rev. D **28**, 1333 (1983).

[10] B. Fuchssteiner and A. S. Fokas, Physica D **4**, 47 (1981).

[11] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge U.P., London, 1980).

[12] C. D. Collinson, J. Math. Phys. **11**, 818 (1970).

Regular double Riemann–Hilbert problems and double Kac–Moody algebraic structures for two-dimensional reduced gravity

Zai-Zhe Zhong
Department of Physics, Liaoning Normal University, Dalian 116022, Liaoning, People's Republic of China

(Received 28 September 1989; accepted for publication 18 April 1990)

By using the double-complex function method, two regular double Riemann–Hilbert problems are established; therefore two double Kac–Moody algebras are given. These structures show that in two-dimensional reduced gravity, in fact, there is more exquisite hidden symmetry than common nonlinear systems.

I. INTRODUCTION

By using a complex method, Wu and Ge[1] have given a Kac–Moody algebra structure for the gravitational plane waves. Recently, Chau and Ge[2] have further pointed out that a Kac–Moody algebraic structure can be obtained for many nonlinear systems from the infinitesimal regular Riemann–Hilbert transform. However, only the ordinary complex functions are used, hence the results are restricted. In this paper, it will be shown that the case is still deeper for the two-dimensional reduced gravity, i.e., there are four Kac–Moody algebras simultaneously, which are natural and isomorphic, and the result concerned in Ref. 2 is only one of these four algebraic structures. This multiple algebra reveals profoundly the hidden symmetry in two-dimensional reduced gravity. The thing that accounts for the occurrence is that in two-dimensional reduced gravity, there is an important dual symmetry; by the double-complex function method, this symmetry has been discussed by us in Refs. 3 and 4. The essence of this dual symmetry is the NK transformation[5] and analytic continuation.[3] In addiiton, the infinitesimal Riemann–Hilbert transform about the Ernst potential in Ref. 1, in fact, derives the Geroch group[6] in two-dimensional reduced gravity. However, recently we have proved[7] that by using the double-complex method, the double-complex realizations of the Geroch group can be obtained; therefore the appearance of some double algebraic structures, in fact, is to be expected.

In Sec. II, we discuss the regular double Riemann–Hilbert problem (RDRHP) of the axisymmetric stationary vacuum fields (ASVF). Two double Kac–Moody algebraic structures are derived in Sec. III. In Sec. IV, we calculate a concrete example. In the last section, we discuss the case of the cylindrically symmetric stational vacuum fields (CSVF).

II. RDRHP FOR ASVF

The general double-complex function method and the inverse scattering method have been discussed in Refs. 3 and 4; in the following, we directly use the results concerned. Let J denote the double-imaginary unit, i.e., $J=i(i^2=-1)$ or $J=\epsilon(\epsilon^2=+1,\epsilon\neq\pm 1)$. When the real series $\Sigma_{n=0}^{\infty}|a_n|$ is convergent, then

$$a(J)=\sum_{n=0}^{\infty}a_nJ^{2n} \tag{1}$$

is called a double-real number, and let $a_C=a(J=i)$, $a_H=a(J=\epsilon)$. If both $a(J)$ and $b(J)$ are double-real numbers, we call $Z(J)=a(J)+J\cdot b(J)$ a double-complex number, and write $Z_C=Z(J=i)$, $Z_H=Z(J=\epsilon)$.

Let the metric of the ASVF be the Papapetrous form

$$ds^2=f(dt-\omega\,d\theta)^2-f^{-1}[e^{\Gamma}(d\rho^2+dz^2)+\rho^2\,d\theta^2], \tag{2}$$

where f, ω, and Γ are real functions of ρ and z only. It is known that Γ is determined by f and ω. The double-complex Ernst equation is

$$\begin{aligned}&\mathrm{Re}(\mathscr{E}(J))\nabla^2\mathscr{E}(J)=\nabla\mathscr{E}(J)\cdot\nabla\mathscr{E}(J),\\&\nabla^2\equiv\partial_\rho^2+\rho^{-1}\partial_\rho+\partial_z^2,\\&\nabla\equiv(\partial_\rho,\partial_z).\end{aligned} \tag{3}$$

From a double solution $\mathscr{E}(J)=F(J)+J\cdot\Omega(J)$ of Eq. (3), a dual real gravitational solution pair $\{(f,\omega),(\hat{f},\hat{\omega})\}$ can be obtained simultaneously, where

$$\begin{aligned}&(f,\omega)=(F_C,V_{F_C}(\Omega_C)),\\&(\hat{f},\hat{\omega})=(T(F_H),\Omega_H),\end{aligned} \tag{4}$$

and (T,V) are the NK transformations[5]

$$T:\ f\to f'=T(f)=\rho/f,$$

$$V_f:\ \varphi\to\omega=V_f(\varphi)=\int\rho f^{-2}(\partial_\rho\varphi\,dz-\partial_z\varphi\,d\rho), \tag{5}$$

i.e.,

$$\partial_\rho\varphi=\rho^{-1}f^2\,\partial_z\omega,\quad \partial_z\varphi=-\rho^{-1}f^2\,\partial_\rho\omega.$$

Now let us consider how to relate the double-complex Ernst equation (3) with the RDRHP. For this purpose, let

$$\begin{aligned}&P(J)=\frac{1}{F(J)}\begin{bmatrix}1 & \Omega(J)\\ \Omega(J) & \Omega^2(J)-J^2F^2(J)\end{bmatrix},\\&\det(P(J))=-J^2;\end{aligned} \tag{6}$$

it is a 2×2 double-real symmetric matrix. It can be proved[3] that Eq. (3) is equivalent to the double-real Belinsky–Zakharov[8] equation

$$\begin{aligned}&\partial_\rho U(J)+\partial_z V(J)=0,\\&U(J)=\rho\,\partial_\rho P(J)\cdot P^{-1}(J),\\&V(J)=\rho\,\partial_z P(J)\cdot P^{-1}(J).\end{aligned} \tag{7}$$

The Lax pair for Eq. (7) has the form

$$[D_\kappa(\lambda)-\mathscr{A}_\kappa(\lambda;J)]\Psi(\lambda;J)=0\quad(\kappa=1,2), \tag{8}$$

where λ is an ordinary complex spectral parameter, and

 0022-2488/90/092217-04$03.00

$$D_1(\lambda)=\partial_z-\frac{2\lambda^2}{\lambda^2+\rho^2}\partial_\lambda,$$

$$D_2(\lambda)=\partial_\rho+\frac{2\lambda\rho}{\lambda^2+\rho^2}\partial_\lambda,$$

$$\mathscr{A}_1(\lambda;J)=\frac{\lambda U(J)-\rho V(J)}{\lambda^2+\rho^2},$$

$$\mathscr{A}_2(\lambda;J)=\frac{-\rho U(J)-\lambda V(J)}{\lambda^2+\rho^2}. \tag{9}$$

We must notice that where the "wave function" $\Psi(\lambda;J)$ is a double ordinary complex matrix, i.e., its matrix elements take the form as $a(\lambda;J)+ib(\lambda;J)$, in which both a and b are double-real functions depending on λ. If $\Psi(\lambda;J)$ is a double solution of Eq. (8), then

$$P(J)=\Psi(\lambda=0;J) \tag{10}$$

is a solution of Eq. (7), and

$$\mathscr{E}(J)=\frac{1}{[\Psi(\lambda=0;J)]_{11}}+J\cdot\frac{[\Psi(\lambda=0;J)]_{12}}{[\Psi(\lambda=0;J)]_{11}} \tag{11}$$

is a solution of Eq. (3), where $[M]_{ij}$ is a element of matrix M. Now we can propose a RDRHP for system (8) as follows. Let C be the circle surrounding the origin in the ordinary λ plane with radius ρ, and let $C_+(C_-)$ be the inside (outside) of C. Suppose that there has been a solution $\Psi(\rho,z;\lambda;J)$ satisfying the condition (10), and ψ is analytic in $C_+\cup C$. We seek ordinary complex matrices $X_\pm$ defined on $C_\pm\cup C$, which satisfy the following condition:

$$\begin{aligned}&X_-(\lambda;J)=X_+(\lambda;J)G(\lambda;J),\quad \lambda\in C,\\ &G(\lambda;J)=\Psi(\lambda;J)u(\lambda)\Psi^{-1}(\lambda;J),\\ &X_-(\lambda=\infty;J)=1,\end{aligned} \tag{12}$$

where $u(J)\in \mathrm{SL}(2,C)$, and

$$D_k(\lambda)u(\lambda)=0\quad (k=1,2). \tag{13}$$

Evidently, the discussion about this RDRHP is similar to a common regular Riemann–Hilbert problem, and we write it as RDRHP $(D,\mathscr{A},\Psi,C,u,X)$. We assume that there exists a pair of fundamental solutions $X_\pm(\lambda,J)$, nonsingular matrices on $C\cup C_\pm$. Let

$$\Psi'(\lambda;J)=\begin{cases}X_+(\lambda;J)\Psi(\lambda;J), & \lambda\in C_+,\\ X_-(\lambda;J)\Psi(\lambda;J)u^{-1}(\lambda), & \lambda\in C_-;\end{cases} \tag{14}$$

then it can be proved that $\Psi'(\lambda;J)$ is also a solution of Eq. (8), and

$$P'(J)=\Psi'(\lambda=0;J)=X(\lambda=0;J)\Psi(\lambda=0;J) \tag{15}$$

is a new solution of Eq. (7).

Since $P'(J)$ should correspond to a gravitational solution pair, the double-reality and symmetry of $P'(J)$ must be guaranteed. Therefore, similar to Belinsky and Zakharov,[8] Ψ, X, and u should satisfy the following additional requirements:

$$\begin{aligned}&\bar{\Psi}(\bar{\lambda};J)=\Psi(\lambda;J),\quad \bar{u}(\bar{\lambda})=u(\lambda),\\ &P'(J)=X(-\rho^2/\lambda;J)P(J)X^T(\lambda;J),\end{aligned} \tag{16}$$

where the bar denotes the complex conjugation and T denotes the transposition. In the following, we assume that these requirements have been satisfied (in Sec. IV an example is given).

Since U_C, V_C, U_H, and V_H are contained in $\mathscr{A}(J)$, the above RDRHP has related to the double-complex duality symmetry of the ASVF. However, it is interesting that by using this duality symmetry we can yet establish another RDRHP $(D,\hat{\mathscr{A}},\hat{\Psi},C,u,\hat{X})$, which is dual with the above RDRHP, as follows. Let the duality mapping $d(J)$ be defined as

$$\begin{aligned}&d(J):\mathscr{E}(J)=F(J)+J\cdot\Omega(J)\\ &\qquad\to\hat{\mathscr{E}}(J)=\hat{F}(J)+J\cdot\hat{\Omega}(J),\\ &\hat{F}(\mathring{J})=T(F(J)),\\ &\partial_\rho\hat{\Omega}(\mathring{J})=\frac{J^2\rho}{F^2(J)}\partial_z\Omega(J),\quad \partial_z\hat{\Omega}(\mathring{J})\\ &\qquad=-\frac{J^2\rho}{F^2(J)}\partial_\rho\Omega(J),\end{aligned} \tag{17}$$

where the overcircle denotes the commutation operation of an imaginary unit, i.e.,

$$\circ: J\to\mathring{J},\quad \mathring{i}=\epsilon,\quad \mathring{\epsilon}=i.$$

Therefore, from a solution $\mathscr{E}(J)$ of Eq. (3), we can obtain its other solution $\hat{\mathscr{E}}(J)$. According to Eqs. (6), (7), and (9), we write the corresponding results as $\hat{P}(J)$, $\hat{U}(J)$, $\hat{V}(J)$, and $\hat{\mathscr{A}}(\lambda;J)$. Of course, $\hat{P}(J)$ is also a solution of Eq. (7). Notice that, in view of gravitational fields, $\hat{\mathscr{E}}(J)$ [or $\hat{P}(J)$] is equivalent to $\mathscr{E}(J)$ [or $P(J)$], i.e., the gravitational field solutions obtained from $\mathscr{E}(J)$ and $\hat{\mathscr{E}}(J)$, in fact, are the same. However, it is important that $\hat{\mathscr{A}}(\lambda;J)\neq\mathscr{A}(\lambda;J)$; therefore by the formal substitution of $(\hat{\mathscr{A}},\hat{\Psi},\hat{X})$ for $(\mathscr{A},\Psi,X)$ in Eqs. (8) and (12)–(14), we can establish an RDRHP $(D,\hat{\mathscr{A}},\hat{\Psi},C,u,\hat{X})$, where $\hat{\Psi}(\lambda;J)$ must satisfy

$$\hat{\Psi}(\lambda=0;J)=\hat{P}(J)=d(\mathring{J})(P(\mathring{J})). \tag{18}$$

By a fundamental solution pair (X_+,X_-), we can obtain a new solution $P'(J)=X_+(\lambda=0;J)P(J)$ of Eq. (7) as well.

Summing up, where the doubleness of the Riemann–Hilbert problem reflects the duality symmetry of the ASVF, this doubleness makes just two dual RDRHP's appear. This is a notable difference from common nonlinear systems. Naturally, we expect that there can exist some method to determine a solution $\hat{\Psi}(\lambda;J)$ of Eq. (8) corresponding to $\hat{\mathscr{A}}(\lambda;J)$ from a solution $\Psi(\lambda;J)$ corresponding to $\mathscr{A}(\lambda;J)$. It is a pity that we have not yet found a concrete feasible way; only in a simple case can we obtain the result (see Sec. IV).

III. DOUBLE KAC–MOODY ALGEBRAS

Let us consider the infinitesimal transform corresponding to the above RDRHP. Let

$$f(\rho,z;\lambda)=\rho/\lambda-z-\lambda, \tag{19}$$

so

$$D_kf=0\quad (k=1,2). \tag{20}$$

Let I_A $(A=1,2,3)$ be the infinitesimal generator of group $\mathrm{SL}(2,R)$, and

$$\begin{aligned}&V_A(\lambda)=[f(\lambda)]^{-m}I_A,\\ &V_\alpha(\lambda)=\alpha^AV_A(\lambda),\end{aligned} \tag{21}$$

where α^A is an infinitesimal constant with the group index A, and m is an integer. It can be proved that the group element

$u(\lambda)$ generated from $V_\alpha(\lambda)$ satisfies Eq. (13). Therefore, according to the method of Ref. 2, an infinitesimal transform $\delta_\alpha \mathscr{A}$ can be derived, however, which is double, i.e.,

$$\mathscr{A}'_\kappa(\lambda;J) - \mathscr{A}_\kappa(\lambda;J) = \alpha^A \delta_A \mathscr{A}_\kappa(\lambda;J),$$
$$\delta_A \mathscr{A}_\kappa(\lambda;J) = \left[\mathscr{D}_\kappa(\lambda;J), \left(\frac{-1}{2\pi i}\right)\int_C \frac{dt}{t-\lambda} G_A(t;J)\right], \quad (22)$$

where

$$\mathscr{D}_\kappa(\lambda;J) = D_\kappa(\lambda) + \mathscr{A}_\kappa(\lambda;J) \quad (\kappa = 1,2),$$
$$G_A(\lambda;J) = \Psi(\lambda;J) V_A(\lambda) \Psi^{-1}(\lambda;J). \quad (23)$$

The algebraic structure is

$$[\delta_\alpha, \delta_\beta]\mathscr{A}_\kappa(\lambda;J) = \left[\mathscr{D}_\kappa(\lambda;J), \left(\frac{-1}{2\pi i}\right)\int_C \frac{dt}{t-\lambda} \Psi(\lambda;J) \times [V_\alpha(t), V_\beta(t)]\Psi^{-1}(\lambda;J)\right]. \quad (24)$$

If we write the infinitesimal transform corresponding to

$$G_A^{(m)}(\lambda;J) = \Psi(\lambda;J)[f(\lambda)]^{-m} V_A(\lambda)\Psi^{-1}(\lambda;J)$$

as $\delta_A^{(m)}$, then we have

$$[\delta_A^{(m)}, \delta_B^{(n)}]\mathscr{A}_\kappa = C^C_{AB}\delta_C{}^{(m+n)}\mathscr{A}_\kappa, \quad (25)$$

where m and n are integers and C^C_{AB} is the structure constant for SL(2,R).

Similarly, for the RDRHP $(D, \hat{\mathscr{A}}, \hat{\Psi}, C, u, \hat{X})$, we obtain a double Kac–Moody algebra,

$$\delta_A \hat{\mathscr{A}}_\kappa(\lambda;J) = \left[\hat{\mathscr{D}}_\kappa(\lambda;J), \left(\frac{-1}{2\pi i}\right)\int_C \frac{dt}{t-\lambda} \hat{G}_A(t;J)\right],$$
$$\hat{\mathscr{D}}_\kappa(\lambda;J) = D_\kappa(\lambda) + \hat{\mathscr{A}}(\lambda;J), \quad (26)$$
$$\hat{G}_A(\lambda;J) = \hat{\Psi}(\lambda;J) V_A(\lambda)\hat{\Psi}^{-1}(\lambda;J).$$

The above four Kac–Moody algebras obtained are isomorphic. In fact, let j denote the mapping derived from the substitution of ϵ for i in $\delta\mathscr{A}_C$ or $\delta\hat{\mathscr{A}}_C$, let $\mathscr{T}$ denote the mapping derived from the substitution for $(\hat{\mathscr{A}}, \hat{\Psi})$ of $(\mathscr{A}, \Psi)$, and let a corresponding Kac–Moody algebra be written simply as $(\delta\mathscr{A})$; therefore the relation among the above four algebras can be explained by the following diagram:

$$\begin{array}{ccc} (\delta\mathscr{A}_C) & \xrightarrow{\mathscr{T}} & (\delta\hat{\mathscr{A}}_H) \\ \downarrow j & & \downarrow j \\ (\delta\mathscr{A}_H) & \xrightarrow{\mathscr{T}} & (\delta\hat{\mathscr{A}}_C), \end{array} \quad (27)$$

where all arrows denote isomorphism mappings. The Ernst potentials corresponding to the above algebras, respectively, are $\mathscr{E}_C = F_c + i\Omega_c$, $\mathscr{E}_H = F_H + \epsilon\Omega_H$, $\hat{\mathscr{E}}_C = \rho F_H^{-1} + iV_{\rho F_H^{-1}}(\Omega_H)$, and $\hat{\mathscr{E}}_H = \rho F_C^{-1} + \epsilon V_{\rho F_C^{-1}}(\Omega_C)$. These four Kac–Moody algebras are equal in status; this fact itself is also a hidden symmetry in the ASVF.

IV. AN EXAMPLE

In correspondence to a Weyl-type solution, let $\mathscr{E} = e^\varphi$, where $\varphi(\rho,z)$ is a real function independent of J and

$$\nabla^2\varphi = 0. \quad (28)$$

The corresponding gravitational solution pair is $((f,\omega), (\hat{f},\hat{\omega})) = ((e^\varphi,0),(\rho e^{-\varphi},0))$. Now $\hat{\mathscr{E}} = \rho e^{-\varphi}$, and

$$P(J) = \begin{bmatrix} e^{-\varphi} & 0 \\ 0 & -J^2 e^{\varphi} \end{bmatrix}, \quad \hat{P}(J) = \begin{bmatrix} \rho^{-1}e^{\varphi} & 0 \\ 0 & -J^2\rho e^{-\varphi} \end{bmatrix},$$
$$U = \rho\,\partial_\rho\varphi \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{U} = (\rho\,\partial\varphi - 1)\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad (29)$$
$$V = \rho\,\partial_z\varphi \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{V} = \rho\,\partial_z\varphi \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Suppose that both $\Psi(\lambda;J)$ and $\hat{\Psi}(\lambda;J)$ are also diagonal, and

$$\Psi(\lambda;J) = \begin{bmatrix} e^{-E} & 0 \\ 0 & -J^2 e^{E} \end{bmatrix}, \quad \hat{\Psi}(\lambda;J) = \begin{bmatrix} e^{\hat{E}} & 0 \\ 0 & -J^2 e^{-\hat{E}} \end{bmatrix}, \quad (30)$$

where $E(\rho,z;\lambda)$ and $\hat{E}(\rho,z;\lambda)$ are ordinary real functions, and

$$\mathrm{Re}(E(\bar{\lambda})) = \mathrm{Re}(E(\lambda)), \quad \mathrm{Im}(E(\bar{\lambda})) = -\mathrm{Im}(E(\lambda)), \quad E(\lambda = 0) = \varphi,$$
$$\mathrm{Re}(\hat{E}(\bar{\lambda})) = \mathrm{Re}(\hat{E}(\lambda)), \quad \mathrm{Im}(\hat{E}(\bar{\lambda})) = -\mathrm{Im}(\hat{E}(\lambda)), \quad \hat{E}(\lambda = 0) = \varphi - \ln\rho. \quad (31)$$

Now, Eq. (8) is changed into

$$(\lambda\partial_\rho + \rho\partial_z)E = \rho\,\partial_z\varphi,$$
$$(\rho\partial_\rho - \lambda\partial_z + 2\lambda\partial_\lambda)E = \rho\,\partial_\rho\varphi, \quad (32a)$$
$$(\lambda\partial_\rho + \rho\partial_z)\hat{E} = \rho\,\partial_z\varphi,$$
$$(\rho\partial_\rho - \lambda\partial_z + 2\lambda\partial_\lambda)\hat{E} = \rho\,\partial_\rho\varphi - 1; \quad (32b)$$

thus we find that the relation between E and $\hat{E}$ is

$$\hat{E} = E - \tfrac{1}{2}\ln(\rho^2 - z^2 - 2\lambda^2). \quad (33)$$

As for how to solve Eq. (32a), see Ref. 9.

From Eq. (33), we have

$$\Psi\cdot\hat{\Psi} = 1/\sqrt{\rho^2 - z^2 - 2\lambda^2}. \quad (34)$$

If E is a solution of Eq. (32a), according to Eqs. (23), (24), and (26), every algebra can be easily calculated, in which

$$G_A(\lambda;J) = [f(\lambda)]^{-m}\Psi(\lambda;J) I_A \Psi^{-1}(\lambda;J),$$
$$\hat{G}_A(\lambda;J) = [f(\lambda)]^{-m}\hat{\Psi}(\lambda;J) I_A \hat{\Psi}^{-1}(\lambda;J) \quad (35)$$
$$= [f(\lambda)]^{-m}\Psi^{-1}(\lambda;J) I_A \Psi(\lambda;J).$$

Between G and $\hat{G}$, there is only a difference of factor order; the above four Kac–Moody algebras, obviously, are isomorphic. In fact, by the substitution of E for $-E$, one can be obtained from the other.

V. THE CASE OF CSVF

For the general discussion of CSVF, see Ref. 7, where we directly use some results to establish the RDRHP and the double Kac–Moody algebras for the CSVF. The steps are similar to the above; however, the results are more different from the ASVF.

For the CSVF, we use the metric

$$ds^2 = \Lambda(dt^2 - dz^2) - tg^{-1}[(dx - \mu\,dy)^2 + g^2\,dy^2], \quad (36)$$

where Λ, g, and μ are real functions of t and z only, and Λ is determined by g and μ (Ref. 9). Let us consider the hyperbolic type double-complex Ernst equation

$$\mathrm{Re}(\mathscr{E}(J))\tilde{\nabla}^2\mathscr{E}(J)=\tilde{\nabla}\mathscr{E}(J)\cdot\tilde{\nabla}\mathscr{E}(J), \tag{37}$$

where

$$\tilde{\nabla}^2\equiv\partial_t^2+t^{-1}\partial_t-\partial_z^2,\quad \tilde{\nabla}\equiv(\partial_t,i\partial_z). \tag{38}$$

Different from ASVF, the dual gravitational solution pair $((g,\mu),(\hat{g},\hat{\mu}))$, which is physical (real), is completely generated by the component $\mathscr{E}_c=\mathscr{E}(J=i)$ of $\mathscr{E}(J)$, when $\mathscr{E}(J)=G(J)+J\cdot M(J)$ is a double solution of Eq. (37), i.e.,

$$\begin{aligned}&(g,\mu)=(G_c,M_c),\\&(\hat{g},\hat{\mu})=(T(G_c),W_{G_c}(M_c)),\end{aligned} \tag{39}$$

where (T,W) is another kind of NK transformation, which differs from the ordinary NK transformation in a sign, i.e.,

$$\begin{aligned}&T:\quad g\to T(g)=t/g,\\&W_g:\quad \mu\to\psi=W_g(\mu),\\&\partial_t\psi=(t/g^2)\,\partial_z\mu,\quad \partial_z\psi=(t/g^2)\partial_t\mu.\end{aligned} \tag{40}$$

From $\mathscr{E}_H=G_H+\epsilon M_H$, we can only obtain a nonphysical (pure imaginary) solution pair as follows:

$$\begin{aligned}&(g',\mu')=(G_H,iM_H),\\&(\hat{g}',\hat{\mu}')=(T(G_H),iW_{T(G_H)}(M_H)).\end{aligned} \tag{41}$$

The duality mapping $\tilde{d}(J)$ is defined as [notice the difference from Eq. (17)]:

$$\begin{aligned}&\tilde{d}(J):\ \mathscr{E}(J)=G(J)+J\cdot M(J)\\&\qquad\to\hat{\mathscr{E}}(J)=\hat{G}(J)+J\cdot\hat{M}(J),\\&\hat{G}(J)=T(G(J)),\\&\partial_t\hat{M}(J)=\frac{t}{G^2(J)}\partial_z M(J),\quad \partial_z\hat{M}(J)=\frac{t}{G^2(J)}\partial_t M(J).\end{aligned} \tag{42}$$

Thus, if $\mathscr{E}(J)$ is a solution of Eq. (37), then $\hat{\mathscr{E}}(J)=\tilde{d}(J)(\mathscr{E}(J))$ is also a solution.

Let $Q(J)$ be a double-real symmetric matrix

$$Q(J)=\frac{1}{G(J)}\begin{bmatrix}1 & M(J)\\ M(J) & M^2(J)-J^2G^2(J)\end{bmatrix},$$

$$\det(M(J))=-J^2 \tag{43}$$

then the double-complex Ernst equation (37) is equivalent to the following double-real Belinsky–Zakharov equation

$$\begin{aligned}&\partial_t R(J)-\partial_z S(J)=0,\\&R(J)=t\,\partial_t Q(J)\cdot Q^{-1}(J),\\&S(J)=t\,\partial_z Q(J)\cdot Q^{-1}(J).\end{aligned} \tag{44}$$

From a solution $Q(J)$ of Eq. (44), we can obtain a solution of Eq. (37),

$$\mathscr{E}(J)=\frac{1}{Q_{11}(J)}+J\cdot\frac{Q_{12}(J)}{Q_{11}(J)}. \tag{45}$$

The Lax pair for Eq. (44) is

$$[\tilde{D}_\kappa(\lambda)-\mathscr{B}_\kappa(\lambda;J)]\Phi(\lambda;J)=0,$$

$$\begin{aligned}&\tilde{D}_1(\lambda)=\partial_t+\frac{2\lambda t}{t^2-\lambda^2}\partial_\lambda,\\&\tilde{D}_2(\lambda)=\partial_z+\frac{2\lambda^2}{t^2-\lambda^2}\partial_\lambda,\\&\mathscr{B}_1(\lambda;J)=\frac{tR(J)+\lambda S(J)}{t^2-\lambda^2},\\&\mathscr{B}_2(\lambda;J)=\frac{tS(J)+\lambda R(J)}{t^2-\lambda^2},\end{aligned} \tag{46}$$

and $\Phi(\lambda;J)$ is an ordinary complex 2×2 matrix such that

$$\Phi(\lambda=0;J)=Q(J). \tag{47}$$

Evidently, by the same explanation we can establish a RDRHP $(\tilde{D},\mathscr{B},\Phi,C,u,Y)$ as in Sec. II, and the addition condition can be taken as

$$\begin{aligned}&\bar{\Phi}(\bar{\lambda};J)=\Phi(\lambda;J),\\&Q'(J)=Y(t^2/\lambda;J)Q(J)Y^T(\lambda;J).\end{aligned} \tag{48}$$

For the infinitesimal transform, we take

$$\begin{aligned}&\tilde{f}(\lambda)=t^2/\lambda+2z+\lambda,\\&\tilde{D}_\kappa\tilde{f}=0,\\&V_\alpha^{(m)}(\lambda)=[\tilde{f}(\lambda)]^{-m}\alpha^A I_A,\ \text{etc.}\end{aligned} \tag{49}$$

Therefore we obtain a double Kac–Moody algebra as follows:

$$\begin{aligned}&\delta_\alpha^{(m)}\mathscr{B}_\kappa(\lambda;J)=\left[\mathscr{D}_\kappa(\lambda;J),\left(\frac{-1}{2\pi i}\right)\int_C\frac{dt}{t-\lambda}\tilde{G}_\alpha^{(m)}(t;J)\right],\\&\mathscr{D}_\kappa(\lambda;J)=\tilde{D}_\kappa(\lambda)+\mathscr{B}_\kappa(\lambda;J),\\&\tilde{G}_\alpha^{(m)}(\lambda;J)=\Phi(\lambda;J)V_\alpha^{(m)}(\lambda)\Phi^{-1}(\lambda;J),\\&[\delta_A^{(m)},\delta_B^{(n)}]\mathscr{B}_\kappa(\lambda;J)=C_{AB}^C\delta_C^{(m+n)}\mathscr{B}_\kappa(\lambda;J).\end{aligned} \tag{50}$$

Let $\hat{R}(J)$, $\hat{S}(J)$, $\hat{\mathscr{B}}(\lambda;J)$, and $\hat{\Phi}(\lambda;J)$ denote the results corresponding to $\hat{\mathscr{E}}(J)=\tilde{d}(J)(\mathscr{E}(J))$ in Eq. (46); then we can obtain another double Kac–Moody algebra as follows:

$$\begin{aligned}&\delta_\alpha^{(m)}\hat{\mathscr{B}}_\kappa(\lambda;J)=\left[\mathscr{D}_\kappa(\lambda;J),\left(\frac{-1}{2\pi i}\right)\int_C\frac{dt}{t-\lambda}\hat{\tilde{G}}_\alpha(t;J)\right],\\&\mathscr{D}_\kappa(\lambda;J)=\tilde{D}_\kappa(\lambda)+\hat{\mathscr{B}}_\kappa(\lambda;J),\\&\hat{\tilde{G}}_\alpha^{(m)}(\lambda;J)=\hat{\Phi}(\lambda;J)V_\alpha^{(m)}(\lambda)\hat{\Phi}^{-1}(\lambda;J),\\&\hat{\Phi}(\lambda=0;J)=\hat{\mathscr{E}}(J)=\tilde{d}(J)(\Phi(\lambda=0;J)).\end{aligned} \tag{51}$$

When $J=i$, the results in Eq. (50) correspond just to the results concerned in Ref. 2. This can be seen in that the above results more profoundly reveal the hidden symmetry in the CSVF. However, it should be pointed out that two gravitational field solution pairs related to Eqs. (50) and (51) are both mixtures consisting of physical and nonphysical solutions; this is not as ideal as in the case of ASVF discussed in Secs. II and III.

[1] Y. S. Wu and M. L. Ge, J. Math. Phys. **24**, 1187 (1983).
[2] L. L. Chau and M. L. Ge, J. Math. Phys. **30**, 166 (1989).
[3] Z. Z. Zhong, J. Math. Phys. **26**, 2589 (1985).
[4] Z. Z. Zhong, Sci. Sin. A **31**, 436 (1988).
[5] D. Kramer and G. Neugebauer, Commun. Math. Phys. **10**, 132 (1968).
[6] R. Geroch, J. Math. Phys. **12**, 918 (1971); **13**, 394 (1972).
[7] Z. Z. Zhong, Sci. Sin. A **33**(8), 75 (1990).
[8] V. A. Belinsky and V. E. Zakharov, Zh. Eksp. Teor. Fiz. **75**, 1953 (1978); **77**, 3 (1979).
[9] P. S. Letelier, J. Math. Phys. **25**, 2675 (1984); **26**, 326 and 467 (1985).

Second-order equations and quadratic Lagrangians

Richard T. Hammond
Department of Engineering Science, North Dakota State University, Fargo, North Dakota 58105

(Received 21 March 1989; accepted for publication 25 April 1990)

From a general Lagrangian that is quadratic in the Ricci tensor, independent variations of the metric and torsion tensors produce gravitational field equations that are of second differential order in the metric tensor. This reduction in order from four results from the use of the torsional field equations. Also, the equation of motion is derived and several special cases are considered.

I. INTRODUCTION

Over the years, a geometrical Lagrangian of the form

$$\delta \int \sqrt{-g}(R + AR^2 + BR_{\mu\nu}R^{\mu\nu}) = 0$$

has received considerable attention.[1] Weyl,[2] in his conformally invariant theory; Gregory;[3] and Eddington,[4] who used only the quadratic part, considered such forms soon after general relativity, and later this form was found to give renormalizable quantum field theories.[5] Soon thereafter, many papers with quadratic Langrangians (QLs) appeared.[6] However, in a V_4 metric theory of gravity the above equation yields fourth-order differential equations in the metric tensor. One problem introduced by the higher derivatives concerns the Cauchy problem, which may or may not be solved depending on the constants A and B.[7] For purely QLs, Havas[8] showed that in the linearized limit such equations do not produce the correct equation of motion for an extended source, and Folomeshkin[9] showed that sensible solutions only exist if T, the trace of the energy momentum tensor, vanishes.

However, recently it has been shown that a special case of the above Lagrangian (with $A = 0$) produces second-order differential equations in the metric tensor in U_4 spacetime.[10] This result was obtained in a metric theory of gravity by assuming independent variations of the metric and torsion tensors. The reduction in order occurs because the torsional field equations impose constraints that cause the higher order derivatives to drop out of the gravitational field equations.

It is the purpose of this paper to extend this result to the most general Lagrangian that is at most quadratic in the Ricci tensor, to include sources, and to derive the equation of motion. It is shown that a similar reduction occurs as in the special vacuum case and that the equations are second order in the metric tensor.

The inclusion of torsion into gravitation received an important boost after the work of Kibble[11] and Utiyama,[12] who showed that torsion can be viewed as the local gauge group of the Poincaré transformation. In this formalism, the independently varied quantities in the variational principle are taken to be the translational potential $e_\alpha{}^i$ (where Latin indices are nonholonomic) and the rotational potential $\Gamma_{\alpha i}{}^j$. With this, one obtains a first-order formalism[13] and general QLs yield second-order differential equations.[14] The use of QLs, besides producing propagating torsion, gives the torsion a conjugate momentum. These results do not occur in the linear case.

As a result of its elegance and success in particle physics, many favor the gauge approach to gravity; however, there are problems in this approach. The gauge theory does not restrict the Lagrangian very well, so that it may consist of a nine-parameter sum.[14,15] Moreover, there is no guarantee that the Poincaré invariance is the right one to gauge and one may consider generalizations such as the affine group or the conformal group, for example.[16] For that matter, there is no guarantee that the gauge approach correctly describes gravity.

Besides the gauge theory approach, there was also a first-order formalism by Israel and Trollope,[17] who considered QLs with independent variations of the (symmetric) metric tensor and the affine connection.

The view taken in this paper is that the fundamental quantities and unknowns in the theory are the metric and torsion tensors and that these are the objects to be varied. At first glance, this produces field equations for gravitation that are fourth-order differential equations in the metric tensor; this may stand as a strong argument for abandoning such a variational principle. However, as stated above, the gravitational equations will actually turn out to be of second differential order and the torsional equations will be of second differential order in the torsion tensor. However, the torsional field equations will contain third derivatives of the metric tensor. As a special case, we will show how to remove these third derivatives. We will also show how the Bianchi identity produces the equation of motion, which is done without resorting to the linearized version.

Following this, we show how the torsion tensor may be viewed as a gauge field of the conformal transformation. By assigning the correct transformation property to the torsion tensor, the Ricci tensor is made to be conformally invariant. Thus, by restricting the Langrangian to terms that are only quadratic in the Ricci tensor or curvature invariant, a vacuum conformally invariant theory results. However, matter will break conformal invariance.

Finally, special cases will be considered. We show how to remove the third derivative term by an appropriate choice of coupling constants. Another case produces nonpropagating torsion and finally, for a semisymmetric connection, we show how the torsion represents a massive vector field.

 0022-2488/90/092221-04$03.00

II. FOURTH- TO SECOND-ORDER EQUATIONS

It is assumed that the geometrical part of the Lagrangian consists, in addition to the usual scalar R, of the most general form that is quadratic in the curvature scalar and the Ricci tensor. Therefore, the variational principle takes the form

$$\delta\int\sqrt{-g}(R+AR^2+BR_{\mu\nu}R^{\mu\nu}+CR_{\mu\nu}R^{\nu\mu}+kL_m+2KL_s)=0, \tag{1}$$

where the definitions are those of Schouten[18] and the source tensors are defined by

$$\frac{\delta\tilde{L}_m}{\delta g_{\mu\nu}}=\tilde{T}^{\mu\nu} \tag{2}$$

and

$$\frac{\delta\tilde{L}_s}{\delta S_{\alpha\beta}{}^{\gamma}}=-\tilde{\tau}^{\alpha\beta}{}_{\gamma}, \tag{3}$$

where the overtilde denotes density. It is assumed that the covariant derivative of the metric tensor, which is assumed to be symmetric, vanishes, so that $\nabla_\sigma g_{\mu\nu}=0$. The gravitational field equations (GFEs) are obtained by varying the metric tensor, which gives

$$\begin{aligned}-G^{\mu\nu}&+\overset{*}{\nabla}_\sigma T^{\mu\nu\sigma}+A(\tfrac{1}{2}g^{\mu\nu}R^2-2RR^{(\mu\nu)})\\&+B(\tfrac{1}{2}g^{\mu\nu}R^{\alpha\beta}R_{\alpha\beta}-R^{\mu\sigma}R^{\nu}{}_{\sigma}-R^{\sigma\mu}R_{\sigma}{}^{\nu})\\&+C(\tfrac{1}{2}g^{\mu\nu}R^{\alpha\beta}R_{\beta\alpha}-R^{\sigma\mu}R^{\nu}{}_{\sigma}-R^{\mu\sigma}R_{\sigma}{}^{\nu})\\&+\mathrm{SYM}_{\mu\nu}\overset{*}{\nabla}_\sigma(-d^{\mu\nu\sigma}-d^{\mu\sigma\nu}+d^{\sigma\nu\mu})=-kT^{\mu\nu},\end{aligned} \tag{4}$$

where

$$\begin{aligned}d^{\mu\nu\sigma}=T^{\mu\nu\sigma}&+A(-g^{\sigma\nu}\nabla^\mu R+g^{\sigma\mu}\nabla^\nu R+2RT^{\mu\nu\sigma})\\&+B(-\nabla^\mu R^{\sigma\nu}+g^{\sigma\mu}\nabla_\phi R^{\phi\nu}+2R_\phi{}^{\nu}T^{\mu\phi\sigma})\\&+C(-\nabla^\mu R^{\nu\sigma}+g^{\sigma\mu}\nabla_\phi R^{\nu\phi}+2R^{\nu}{}_{\phi}T^{\mu\phi\sigma}),\end{aligned} \tag{5}$$

where the modified torsion tensor is defined by $T^{\alpha\beta\gamma}=S^{\alpha\beta\gamma}+S^\beta g^{\alpha\gamma}-S^\alpha g^{\beta\gamma}$, $\overset{*}{\nabla}_\gamma\equiv\nabla_\gamma+2S_\gamma$; the torsion vector is defined by $S_\gamma=S_{\gamma\beta}{}^{\beta}$; and brackets (parentheses) around indices imply antisymmetrization (symmetrization). Details concerning these variations may be found elsewhere.[19]

The torsional field equations (TFEs) are obtained by performing variations of $S_{\alpha\beta}{}^{\gamma}$. The resulting equations may be put in the form

$$d^{[\alpha\beta]\gamma}=K\tau^{\gamma[\beta\alpha]}. \tag{6}$$

As expected, the GFEs are of fourth differential order in the metric tensor. However, the TFEs impose additional constraints that can be used to eliminate the higher order derivatives from the GFEs. In fact, using the TFEs in (4) along with the Bianchi identity and the rule for commutation of covariant differentiation[20] (see Ref. 10 for details), one obtains

$$\begin{aligned}-G^{\mu\nu}&+A(\tfrac{1}{2}g^{\mu\nu}R^2-2RR^{\mu\nu})+B(\tfrac{1}{2}g^{\mu\nu}R^{\alpha\beta}R_{\alpha\beta}\\&-R^{\mu\sigma}R^{\nu}{}_{\sigma}-R^{\mu}{}_{\alpha\beta}{}^{\nu}R^{\alpha\beta})+C(\tfrac{1}{2}g^{\mu\nu}R^{\beta\alpha}R_{\alpha\beta}\\&-R^{\mu\sigma}R_{\sigma}{}^{\nu}-R^{\mu}{}_{\alpha\beta}{}^{\nu}R^{\beta\alpha})=-kT^{\mu\nu}-K\overset{*}{\nabla}_\sigma\tau^{\sigma\nu\mu}.\end{aligned} \tag{7}$$

Since $\tau^{\sigma\nu\mu}$ represents some prescribed distribution of "charge," which, of course, is zero in vaccum (torsion propagates here), (7) shows that the GFEs are of second differential order. It may also be shown that the antisymmetric part of (7), with the TFEs, vanishes. We point out that the torsion is nontrivial, i.e., it does not vanish identically, and these equations cannot reduce to Einsteinian equations unless the torsion vanishes. This will be seen explicitly in Sec. V, where special cases are considered.

Equation (7) has another interesting feature: The torsion source tensor contributes (interiorly) directly to the curvature of space and thus acts as a direct source for the gravitational field. This is over and above the usual coupling that arises from the mass–energy associated with the stress tensor contribution.

Thus the GFEs are of second differential order in $g_{\mu\nu}$ and the TFEs are of second differential order in the $S_{\alpha\beta}{}^{\gamma}$. However, in the TFEs there are terms that involve the third derivative of the metric tensor. Later, it is shown how these can be removed by an appropriate choice of constants, but these terms are more benign than the usual fourth derivatives that, without torsional variations, would appear in the GFEs.

III. EQUATION OF MOTION

We now show that the Bianchi identities of U_4 spacetime impose a differential constraint on the source tensor that produces the equation of motion.

The Christoffel derivative is useful here and is defined by

$$A^{\mu}{}_{;\sigma}=A^{\mu}{}_{,\sigma}+\{{}^{\mu}_{\sigma\beta}\}A^{\beta}, \tag{8}$$

where $\{{}^{\mu}_{\sigma\beta}\}$ is the Christoffel symbol. In order to find the equation of motion, operate with ∇_ν on both sides of (7) to obtain

$$\begin{aligned}kT^{\mu\nu}{}_{;\nu}=&-2k(S^{\mu}{}_{\eta\nu}T^{\eta\nu}-S_\eta T^{\mu\eta})+2S^{\mu}{}_{\eta\gamma}R^{\gamma\eta}-S_{\eta\gamma}{}^{\phi}R^{\mu}{}_{\phi}{}^{\gamma\eta}-\nabla_\sigma\{A(\tfrac{1}{2}g^{\mu\sigma}R^2-2RR^{\mu\sigma})+B(\tfrac{1}{2}g^{\mu\sigma}R^{\alpha\beta}R_{\alpha\beta}-R^{\mu\phi}R^{\sigma}{}_{\phi}\\&-R^{\mu}{}_{\alpha\beta}{}^{\sigma}R^{\alpha\beta})+C(\tfrac{1}{2}g^{\mu\sigma}R^{\alpha\beta}R_{\beta\alpha}-R^{\mu\phi}R_{\phi}{}^{\sigma}-R^{\mu}{}_{\alpha\beta}{}^{\sigma}R^{\beta\alpha})\}-K^{\mu},\end{aligned} \tag{9}$$

where $K^\mu\equiv K\nabla_\sigma\overset{*}{\nabla}_\phi\tau^{\phi\sigma\mu}$. In proceeding, the derivatives on the rhs of (9) can be simplified in the same way as that used to obtain (7) and the TFEs multiplied by the full curvature tensor can be used to eliminate $S_{\eta\gamma}{}^{\phi}R^{\mu}{}_{\phi}{}^{\gamma\eta}$ from (7). The result is

$$\begin{aligned}kT^{\mu\nu}{}_{;\nu}=&-2k(S^{\mu}{}_{\eta\nu}T^{\mu\nu}-S_\eta T^{\mu\eta})+2S^{\mu}{}_{\eta\gamma}R^{\gamma\eta}-2S_\gamma R^{\mu\gamma}+4AR(R^{\gamma\eta}S^{\mu}{}_{\eta\gamma}-R^{\mu\gamma}S_\gamma)+2B[R^{\mu\beta\alpha\gamma}R_{\beta\gamma}S_\alpha\\&-R^{\mu\gamma}R^{\sigma}{}_{\gamma}S_\sigma+R_{\sigma\rho}(S_\lambda{}^{\mu\phi}R^{\sigma}{}_{\phi}{}^{\rho\lambda}-S^{\sigma\mu\lambda}R_\lambda{}^{\rho})]+2C[R^{\mu\beta\alpha\gamma}R_{\gamma\beta}S_\alpha\\&-R^{\mu\gamma}R_\gamma{}^{\sigma}S_\sigma+R_{\rho\sigma}(S_\lambda{}^{\mu\phi}R^{\sigma}{}_{\phi}{}^{\rho\lambda}-S^{\sigma\mu\lambda}R_\lambda{}^{\rho})]-K^{\mu}+kR^{\mu\beta\alpha\gamma}\tau_{\beta\gamma\alpha}.\end{aligned} \tag{10}$$

Finally, the first four terms on the rhs can be eliminated by multiplying the torsion tensor and the torsion vector into the GFEs. After the ensuing cancellation, one obtains

$$kT^{\mu\nu}{}_{;\nu} = -K^{\mu} + KR^{\mu\beta\alpha\gamma}\tau_{\beta\gamma\alpha} + 2K(S^{\mu}{}_{\eta\gamma}\overset{*}{\nabla}_{\sigma}\tau^{\sigma\eta\gamma} - S_{\gamma}\overset{*}{\nabla}_{\sigma}\tau^{\sigma\gamma\mu}). \tag{11}$$

From Eq. (11), the actual equation of motion may be found by prescribing the source tensors of the material. Using standard techniques,[21] the equation of motion to any desired order may be found.

An important aspect of (11) is that for (torsionally) charge-free matter, the rhs of (11) vanishes and the equation of motion is that of standard general relativity. Equation (11) also shows that, at least as far as the equation of motion is concerned, the quadratic terms in the GFEs may be coupled to matter consistently even if the linear term in R from the Lagrangian is dropped. Also, in vacuum, Eq. (11) is an identity which shows that the Bianchi identities are consistent with the field equations.

As an example pertaining to the use of (11), consider the special case that the torsion source may be represented by a conserved four-current j^{α}, so that

$$\tau^{\alpha\beta}{}_{\gamma} \equiv j^{[\alpha}\delta^{\beta]}_{\gamma}. \tag{12}$$

Then (11) yields

$$kT_{\mu}{}^{\nu}{}_{;\nu} = Kj^{\nu}2S_{[\mu,\nu]}, \tag{13}$$

a useful result for later.

IV. GAUGING THE CONFORMAL TRANSFORMATION

The field equations of general relativity are not conformally invariant and to make them so, one has to introduce some new field or degree of freedom. Weyl introduced a nonmetricity four-vector, which he interpreted as the electromagnetic potential and which was assigned the correct transformation property concomitant with the conformal transformation of the metric tensor. Under the *combined* transformation, conformal invariance is obtained.

A similar scheme can be used here, although a metric theory of gravity is maintained. The new freedom is the torsion tensor, which can be assigned a transformation property to produce a conformally invariant Ricci tensor.

It may be noted that under a global or homothetic conformal transformation, the quadratic part of the Lagrangian density is invariant. In order to make this invariance local, one may assign a compensating transformation to the torsion tensor. In fact, under the combined local transformation

$$S_{\mu\nu}{}^{\sigma} \to S_{\mu\nu}{}^{\sigma} + 2\lambda_{,[\mu}\delta^{\sigma}_{\nu]} \tag{14a}$$

and

$$g_{\mu\nu} \to \omega g_{\mu\nu} \quad (g^{\mu\nu} \to g^{\mu\nu}/\omega), \tag{14b}$$

one may show that, calling $\Omega_{\beta} = 2\lambda_{,\beta} - \omega_{,\beta}/2\omega$,

$$R_{\alpha\beta} \to R_{\alpha\beta} + K_{\gamma\alpha}{}^{\gamma}\Omega_{\beta} - K_{\sigma\phi}{}^{\sigma}\Omega^{\phi}g_{\alpha\beta} + 2K_{\alpha\beta}{}^{\phi}\Omega_{\phi} + K_{\alpha\phi\beta}\Omega^{\phi} + K_{\phi\beta\alpha}\Omega^{\phi} + g_{\alpha\beta}(2\Box\lambda - \Box\omega/2\omega - 8\lambda_{,\phi}\lambda^{,\phi}) + 4\lambda_{,\beta;\alpha} + 8\lambda_{,\alpha}\lambda_{,\beta} + \omega^{-1}[-4(\omega_{,[\alpha}\lambda_{,\beta]}) + 4\lambda_{,\phi}\omega^{,\phi}g_{\alpha\beta} - \omega_{,\beta;\alpha} + 3\omega_{,\alpha}\omega_{,\beta}/2\omega]. \tag{15}$$

Equation (15) shows that if $\Omega_{\beta} = 0$, $R_{\alpha\beta}$ is invariant under the transformation (14) and, therefore, the quadratic Lagrangian densities are also invariant [contrary to a usual gauge approach, the covariant derivative is not invariant under (14)]. Like the Weyl case, the curvature scalar R has to be abandoned and only the quadratic terms in (1) may be retained in order to construct a conformally invariant Lagrangian density.

It has been shown above that the equation of motion can be obtained, and is sensible, without the necessity of keeping R in the Lagrangian. However, without R, the coupling constant k will not be the usual one and all interior solutions of gneral relativity are no longer valid. Moreover, $T^{\mu\nu}$ is not conformally invariant, so that matter breaks conformal invariance anyway. Thus, if one wants to construct a conformally invariant theory, there is still much work that needs to be done. The purpose of deriving (15) is to show that torsion may be viewed as allowing a global invariance inherent in QLs to be made local.

V. SPECIAL CASES

A few special cases will now be briefly examined. The strongest reason for considering these cases is that they simplify the above equations, reduce the number of arbitrary constants, and facilitate the interpretation of some of the above equations. It may also turn out that these represent approximate versions of the above, although this is speculative.

As stressed earlier, the GFEs are of second differential order in the metric tensor and the TFEs are of second differential order in the torsion tensor. However, the TFEs contain third derivatives of the metric tensor. It will now be shown how the third derivative terms can be removed by a suitable choice of constants.

In fact, the geometrical part of the Lagrangian contains three arbitrary constants. In a reasonable physical theory, unless each of these constants is related to a universal constant of nature, this number is too high. In order to reduce the number of arbitrary constants, one may seek some constraint or condition that can eliminate or determine some of these. A natural choice here is to impose the condition that the field equations nowhere contain derivatives of the metric tensor of order higher than 2. This can be achieved by letting $A = 0$ and $B = -C$. The TFEs then become

$$T^{\alpha\beta\gamma} + \underset{\alpha\beta}{\text{ANT}}2B\{-\nabla^{\gamma}\overset{*}{\nabla}_{\phi}T^{\alpha\beta\phi} + g^{\alpha\gamma}\nabla_{\sigma}\overset{*}{\nabla}_{\phi}T^{\sigma\beta\phi} + 2T^{\gamma}{}_{\sigma}{}^{\alpha}\overset{*}{\nabla}_{\phi}T^{\sigma\beta\phi}\} = KT^{\gamma[\beta\alpha]}. \tag{16}$$

The GFEs are given as before, with $A = 0$ and $B = -C$: These equations now represent second-order equations with one unknown constant in the geometrical part of the Lagrangian. As will be seen later under further special cases, this constant will acquire a physical interpretation.

If one considers the linearized (in the torsion) version, the vacuum equation becomes

$$\nabla_{\gamma}\nabla_{\phi}T^{\alpha\beta\phi} - S^{\alpha\beta}{}_{\gamma}/2B = 0. \tag{17}$$

Equation (17) represents 24 second-order differential equations in the modified torsion tensor.

Another special case is given by $B = C = 0$, which represents the simplest nonlinear Lagrangian. In this case, the torsion turns out to be nonpropagating, but is nonzero inside matter. The torsion is coupled to the gradient of the ordinary energy momentum tensor of matter and the equation of motion turns out to be that of conventional general relativity. Details of this case are presented elsewhere.[22]

As a final special case, reconsider $B = -C$, $A = 0$ for the semisymmetric case. In the semisymmetric case, the torsion tensor is written in terms of a four-vector (and therefore the source is represented by a vector), so that one can assume[23]

$$3S_{\alpha\beta\gamma} = 2S_{[\alpha}g_{\beta]\gamma}. \tag{18}$$

With (18), one has $R_{[\alpha\beta]} = 4S_{[\beta,\alpha]}/3$ and therefore that

$$P\{R_{[\alpha\beta],\sigma}\} = 0, \tag{19a}$$

where P stands for permutation of the indices. When (18) and (12) are used in (16), one obtains

$$R^{[\sigma\mu]}{}_{;\sigma} + 2S^{\mu}/B = -3Kj^{u}/2B. \tag{19b}$$

Thus the torsion vector may be interpreted as the potential for a massive vector (Proca) field and the antisymmetric part of the Ricci tensor may be interpreted as the field intensity. Alternatively, the field equations (19) for the torsion may be viewed as those of electromagnetism with massive photons, with S_{μ} taken as proportional to the electromagnetic field. The equation of motion is given by (13) and is seen to be the usual equation of the Einstein–Maxwell theory. This has been discussed elsewhere and details may be found there.[19]

In either case, there is left only one unknown constant of the geometrical part of the Lagrangian, which from the above special case may be viewed as being proportional to the range of the potential, or equivalently, the mass of the torsion quanta (or photon). Also, of course, the charge of torsion particles is not determined by these equations. Thus the only undetermined constants represent universal constants.

The conformal invariance discussed earlier is now seen to be related to a gauge transformation of the potential. However, with matter present ($R \neq 0$), the torsion quantum acquires a nonzero mass and breaks the gauge invariance, as well as the conformal invariance, of the theory.

VI. SUMMARY

This paper has examined the field equations resulting from the most general Lagrangian, which is at most quadratic in the Ricci tensor, under independent variations of the metric and torsion tensors. The main result shows that the gravitational field equations are of second differential order in the metric tensor. This reduction in order from four results from the use of the torsional field equations, which are second-order differential equations in the torsion. However, third derivatives of the metric tensor persist in the TFEs.

It was shown how the Bianchi identity produces the equation of motion in a simple way, without resorting to a linearized version. Also, it was shown how the torsion tensor can be viewed as the gauge field of a conformal transformation.

Several special cases were considered and, in particular, it was shown how to eliminate the third derivatives from the TFEs. As a further specialization, this case was considered for the semisymmetric situation, where the torsion vector may be viewed as the potential of a massive vector field. It was seen that the R contribution to the Lagrangian gave the torsion quanta a nonzero mass that broke the conformal invariance.

[1] L. Bel and S. Zia, Phys. Rev. D **32**, 3128 (1985) and the references therein; see, also, Ref. 5.
[2] H. Weyl, Ann. Phys. Leipzig **59**, 101 (1919).
[3] C. Gregory, Phys. Rev. **72**, 72 (1947).
[4] A. Eddington, *The Mathematical Theory of Relativity* (Cambridge U.P., Cambridge, 1923); see, also, R. Bach, Math. Z. **9**, 110 (1921).
[5] B. DeWitt, *Dynamical Theory of Groups and Fields* (Gordon and Breach, New York, 1965); see, also, K. S. Stelle, Phys. Rev. D **16**, 953 (1977); and for quantum fields in curved space-time, see R. Utiyama, Phys. Rev. **101**, 1597 (1956).
[6] V. Müller and H. -J. Schmidt, Gen. Rel. Grav. **17**, 769 (1985); H. Lenzen, Gen. Rel. Grav. **17**, 1137 (1985); A. Jakubiec and J. Kijowski, Phys. Rev. D **37**, 1406 (1988); V. Szczryba, Phys. Rev. D **36**, 351 (1987) and the many references therein.
[7] P. Teyssandier and P. Tourenc, J. Math. Phys. **24**, 2793 (1983).
[8] P. Havas, Gen. Rel. Grav. **8**, 631 (1977).
[9] V. Folomeshkin, Commun. Math. Phys. **22**, 115 (1971).
[10] R. Hammond, J. Math. Phys. **30**, 1115 (1988).
[11] T. Kibble, J. Math. Phys. **2**, 212 (1961).
[12] R. Utiyama, Phys. Rev. **101**, 1597 (1956).
[13] P. Von der Heyde, Phys. Lett. A **58**, 141 (1976); F. Heyl, J. Nitsch, and P. Von der Heyde, in *General Relativity and Gravitation. One Hundred Years after the Birth of Albert Einstein*, edited by A. Held (Plenum, New York, 1980), Vol. 1, 329; see, also, F. Hehl, P. Von der Heyde, G. Kerlick, and J. Nestor, Rev. Mod. Phys. **48**, 393 (1976).
[14] H. Lenzen, Gen. Rel. Grav. **17**, 1137 (1985).
[15] K. Hayashi and T. Shirafuji, Prog. Theor. Phys. **66**, 318 (1981) and their preceding work referenced therein.
[16] Y. Ne'eman, in *General Relativity and Gravitation. One Hundred years after the Birth of Albert Einstein*, edited by A. Held (Plenum, New York, 1980), Vol. 1, p. 309.
[17] W. Israel and R. Trollope, J. Math. Phys. **2**, 777 (1960).
[18] J. Schouten, *Ricci Calculus* (Springer-Verlag, Berlin, 1954).
[19] R. Hammond, Gen. Rel. Grav. **20**, 813 (1988).
[20] Reference 18, Chap. III.
[21] A. Papapetrou, *Lectures on General Relativity* (Reidel, Dordrecht, Holland, 1974).
[22] R. Hammond, submitted to Gen. Rel. Grav.
[23] This case is considered in Ref. 9, pp. 126 and 127, and is derived from a more specialized case given in Ref. 16.

Noether's theorem in nonlinear σ models with a Wess–Zumino term

Hishamuddin Zainuddin
Centre for Particle Theory and Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, England

(Received 9 January 1990; accepted for publication 18 April 1990)

The σ model with a Wess–Zumino term is treated as a system of a particle in a magnetic field on an infinite-dimensional configuration space. Noether's theorem has to be modified to take account of this background gauge field in the configuration space. The questions of possible "anomalous" constants of motion and an "Aharonov–Bohm effect" for nonsimply connected configuration spaces are also addressed.

I. INTRODUCTION

It is well known that continuous symmetry transformations that leave the action of a field theory invariant imply the existence of conserved currents and hence associated constants of motion (conserved charges). This is Noether's theorem. This idea still retains interest when the theory possesses gauge invariance. It has been demonstrated in Ref. 1 that Noether's theorem is modified for a system in a symmetric background gauge field. Under a transformation of space-time coordinates that leaves the gauge field invariant, there is a further contribution C_{gauge} to the usual constant of motion C_0, giving the constant of motion of the total system as

$$C = C_0 + C_{\text{gauge}}.$$

In this paper, we will discuss an example of such a modification for the case of nonlinear σ models with a Wess–Zumino term. It was shown by Wu and Zee[2] that the Wess–Zumino term provides an analog of a magnetic field in the configuration space of the σ model. We will use this idea to furnish these theories with a gauge symmetry and hence show the modification of Noether's theorem in the presence of a Wess–Zumino term. This will be illustrated in examples of σ models on simple target manifolds M in Secs. IV and V. Of particular interest to us in these discussions is the possible anomalous phenomenon of ill-defined constants of motion, as in the case of a particle in a constant magnetic field on a nonsimply connected space.[3–5] This will be discussed in conjunction with the examples.

II. WESS–ZUMINO TERM AND GAUGE SYMMETRY

To begin, we will give a construction of the Wess–Zumino term and hence the (total) Lagrangian density for the σ model in general. We will also demonstrate, from the Lagrangian density, the correspondence of these models with the case of a particle in a magnetic field. This leads us to the discussion of the gauge symmetry in the configuration space of the σ models.

Consider a σ model consisting of fields mapping a $(d+1)$-dimensional space-time into a manifold M. A Wess–Zumino action (term) is a topological action (term) added to the normal kinetic energy action (Lagrangian density) of the theory: It is topological in the sense that the physically relevant quantities derived from it are independent of transformations of the fields. A Wess–Zumino action consists of a $(d+2)$ form on M integrated over $(d+2)$ chains of M. [A d chain of a manifold M is an immersion of a d-dimensional surface into M, while a d cycle is a "boundaryless" d chain. Equipped with a boundary operator ∂, the equivalence classes of d cycles modulo boundaries of $(d+1)$ chains form the dth homology classes of M, $H^d(M)$.[6]] For simplicity, we will restrict our interest to $(1+1)$ dimensions of space-time, with the coordinates (x,t).

Let the fields mapping space (coordinate x) be denoted by Φ. We impose on Φ the boundary condition

$$\Phi \to \Phi_0 \in M \text{ as } |x| \to \infty. \tag{2.1}$$

Topologically, we are investigating loops in M with the basepoint Φ_0. The loops generate one-cycles of M and in effect we consider Φ to be the map

$$\Phi: S^1 \to M, \tag{2.2}$$

where S^1 is space with a distinguished point mapped to Φ_0. The image of Φ, which we shall also sometimes call Φ for simplicity, can be decomposed in terms of fundamental cycles of M as

$$\Phi = \sum_a n_a C_a + \partial\tilde{\phi}, \tag{2.3}$$

where C_a is a set of nontrivial loops generated by $\pi_1(M)$ with (nonzero) winding numbers n_a and $\tilde{\phi}$ is (the image of) an extension

$$\tilde{\phi}: D^2 \to M \tag{2.4}$$

of the map

$$\phi = \partial\tilde{\phi}: S^1 \to M, \tag{2.5}$$

where $\partial D^2 = S^1$ (with the distinguished point). Basically, ϕ is a map which is homotopic to the constant map. The $+$ symbol in (2.3) means the joining of oriented loops at the basepoint, as in the discussion of the group property of the fundamental group.[6] (See Fig. 1.) In general, this (group) operation may be non-Abelian. In such a case one considers only the Abelianized version of the group.[6] However, the examples of M considered here will all have the Abelian fundamental group. Thus we will not pursue the discussion any further.

The time dependence of the fields Φ is introduced by considering a family of maps $\Phi_t: S^1 \times \{t\} \to M$ parametrized

 0022-2488/90/092225-12$03.00

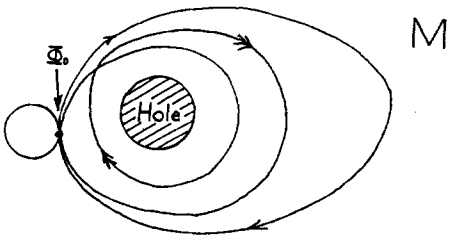


FIG. 1. Joining of loops in M.

by $t\in I=[t_0,t_1]\subset \mathbb{R}$. The whole family of maps will be denoted by

$$\Phi_I: S^1\times I\to M, \tag{2.6}$$

with

$$\Phi_I=\sum_a n_a C_{Ia}+(\partial\tilde{\phi})_I \tag{2.7}$$

(here, ϕ_I refers to the image of the map), where

$$\tilde{\phi}_I: D^2\times I\to M,$$
$$\phi_I=(\partial\tilde{\phi})_I: S^1\times I\to M. \tag{2.8}$$

[Note that one could take ϕ_I to be $(\partial\tilde{\phi}_I)$, where there will be an extra contribution from the endpoints of I. This amounts only to a total time derivative in the action and hence can be ignored. Further explanations can be found in note (i) and the discussions of gauge symmetry later in this section.] The relevant differential form on M for the construction of the Wess–Zumino term will be a three-form

$$\Omega+d\Lambda, \tag{2.9}$$

where Ω is a generator of the third cohomology class of M, $H^3(M)$ and Λ is some two-form on M. With Eq. (2.7) and form (2.9), the Wess–Zumino action may now be given by

$$\Gamma[\Phi_I]=\sum_a n_a\Gamma[C_{Ia}]+\Gamma[(\partial\tilde{\phi})_I]+\Gamma'[\Phi_I], \tag{2.10}$$

where

$$\Gamma[(\partial\tilde{\phi})_I]=\int_{D^2\times I}\tilde{\phi}_I^*\Omega \tag{2.11}$$

and

$$\Gamma'[\Phi_I]=\int_{S^1\times I}\Phi_I^*\Lambda. \tag{2.12}$$

(The asterisk denotes the pullback of the forms Ω and Λ by the maps $\tilde{\phi}_I$ and Φ_I, respectively.) The first term in Eq. (2.10) is an arbitrary fixed real number given by the following construction of Krichever *et al.*[7] This is done by realizing that C_a is a class of homologous one-cycles. Denote two cycles in this class by C_a and C_a'. Consider a pair (N^2,χ_a), where N^2 is a two-dimensional topological space whose boundary is $S^1\dot{\cup}(-S^{1'})$ ($S^{1'}$ being different from S^1) and is mapped by C_a' to an image of C_a' with the same basepoint as that of C_a, while χ_a is the mapping

$$\chi_a: N^2\to M \tag{2.13}$$

such that

$$\chi_a|_{S^1}=C_a, \tag{2.14}$$
$$\chi_a|_{S^{1'}}=C_a'. \tag{2.15}$$

We can now use the pari (N^2,χ_a) to construct $\Gamma[C_{Ia}]$ globally by defining

$$\Gamma[C_{Ia}]=\int_{N^2\times I}\chi_{Ia}^*\Omega, \tag{2.16}$$

where χ_{Ia} is simply the mapping

$$\chi_{Ia}: N^2\times I\to M. \tag{2.17}$$

It is now important to note the following.

(i) We have assumed that Ω is closed. This is necessary to make term (2.11) a topological term, i.e., to be independent of the way in which ϕ is extended to $\tilde{\phi}$. [The same can be done for term (2.16) simply by replacing ϕ and $\tilde{\phi}$ by C_a and χ_a, respectively.] This can be seen as follows. Let $\tilde{\phi}_I$ and $\tilde{\phi}_I'$ be two different extensions of ϕ_I. If $\pi_2(M)=0$, then a homotopy

$$\Psi: D^3\times I\to M \tag{2.18}$$

from $\tilde{\phi}_I$ to $\tilde{\phi}_I'$ always exists, where D^3 is a three-disk (see Fig. 2) and

$$\Psi|_{S\times I}=\tilde{\phi}_I, \tag{2.19}$$
$$\Psi|_{N\times I}=\tilde{\phi}_I', \tag{2.20}$$
$$\Psi|_{E\times I}=\phi_I. \tag{2.21}$$

If Ω is closed, then

$$\begin{aligned}0&=\int_{D^3\times I}\Psi^*\,d\Omega\\&=\int_{\partial(D^3\times I)}\Psi^*\Omega\\&=\int_{N\times I}\tilde{\phi}_I'^*\Omega-\int_{S\times I}\tilde{\phi}_I^*\Omega+\int_{D^2}\Psi^*\Omega|_{t=t_1}\\&\quad-\int_{D^2}\Psi^*\Omega|_{t=t_0}.\end{aligned} \tag{2.22}$$

Note that the last two terms in (2.22) combine to give a total time derivative under the integral $\int dt$ and may be ignored since they do not contribute to the dynamics. (Equivalently, one may use the gauge freedom discussed later in this section to gauge them away.) Hence we find that the Wess–Zumino action is independent of the extensions $\tilde{\phi}_I$ (or χ_{Ia}) modulo endpoint contributions and thus is well defined as a part of a physical action. For those manifolds M with $\pi_2(M)\neq 0$, there is an ambiguity in the possible extensions of ϕ_I (or C_{Ia}). Referring to the explanations above, the extensions $\tilde{\phi}_I$ and $\tilde{\phi}_I'$ are no longer homotopic to each other and hence are inequivalent. To resolve this problem one requires an extra

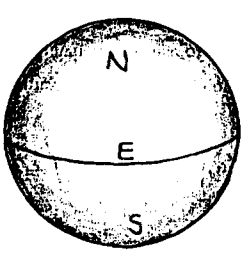


FIG. 2. Here we show D^3 (a solid ball) divided by the equator E and giving the two hemispheres N and S.

datum to make the Wess–Zumino action well defined. We will defer the discussion of such cases to Sec. VI after constructing the second example in Sec. V.

(ii) Most of the time we will restrict our attention to term (2.11) of the whole Wess–Zumino action. Term (2.11) in fact corresponds to the usual definition of the Wess–Zumino action, which is independent of any deformations of $\tilde{\phi}_I$ in M. Term (2.12) is uninteresting since it can be written consistently and globally as an integral over space-time without any difficulty. With this term in mind, one can always add further exact forms to Ω, with their integrals contributing only to integral (2.12). Hence to define (2.11) one uses only the generators of the third cohomology class of M, $H^3(M)$, as seen earlier. Thus we will have $b_3(M) = \dim H^3(M)$ independent Wess–Zumino terms. The term $\Gamma[C_{Ia}]$ will be treated as fixed numbers given by (2.16). We will in fact totally ignore the fields C_a later.

(iii) This construction of the Wess–Zumino action is different from its usual construction, e.g., that of Braaten *et al.*[8] Here, the normal construction involving Euclideanized space-time is avoided by using a time-parametrized family of maps Φ. An important consequence is that the fields Φ_I need no longer be cycles of M, but are general two-chains on M. (For explanations of the terms "cycles" and "chains" see the discussion regarding the d chain.)

To write terms (2.11) and (2.16) in the usual fashion of (the integral of) the Lagrangian density, we use the Poincaré lemma to write Ω as an exact form in some local patch of M:

$$\Omega = d\omega. \tag{2.23}$$

The integral (2.11) may now be written as

$$\begin{aligned}\Gamma[(\partial\tilde{\phi})_I] &= \int_{D\times I} \tilde{\phi}_I^* d\omega \\ &= \int_{S^1\times I} \phi_I^* \omega + \{\text{total time derivatives}\} \\ &= \int_{S^1\times I} dt\, dx \{\epsilon^{\mu\nu}\, \partial_\mu \phi_I^j\, \partial_\nu \phi_I^k\, \omega_{jk}(\phi)\},\end{aligned} \tag{2.24}$$

where $\omega(\phi)$ has singularities in ϕ. For the integral (2.16) there is no consistent way of writing the integral locally. This is due to the problem associated with the gauge transformations belonging only to trivial winding number sector of the fields Φ (this will be discussed at the end of this section). To avoid a cumbersome notation, the subscript I will be dropped from now on. Before discussing the total Lagrangian for the whole system, it is necessary to define the kinetic energy term: It consists of derivatives of the field Φ. Given a metric g on the target manifold, the kinetic energy term is written as

$$\int dx\, \frac{1}{2}\, \partial_\mu \Phi^j\, \partial^\mu \Phi^k\, g_{jk}(\Phi). \tag{2.25}$$

One can make simplifications for this term when one realizes that the fields C_a are linear functions of x (generating the nontrivial winding numbers). Furthermore, the term may be made independent of t owing to the topological nature of the fields C_a. Thus derivatives of the fields C_a in the kinetic term merely add constants to the kinetic term involving ϕ's [the cross terms in (2.25) may be integrated out] and, therefore, these fields will be ignored totally for the rest of this paper. Hence given Eq. (2.24), the total Lagrangian density for the nonlinear σ model with the Wess–Zumino term may now be written (modulo constants) as

$$\begin{aligned}\mathcal{L} &= \tfrac{1}{2}(\partial_\mu \phi^j)(\partial^\mu \phi^k) g_{jk}(\phi) \\ &\quad + \epsilon^{\mu\nu}(\partial_\mu \phi^j)(\partial_\nu \phi^k)\omega_{jk}(\phi).\end{aligned} \tag{2.26}$$

We will proceed to express the correspondence between the σ model and a particle in a magnetic field[2] by first writing the conjugate momenta to $\phi^j(x)$ from Eq. (2.26):

$$\begin{aligned}\pi_j(x) &= \frac{\delta \mathcal{L}}{\delta \dot{\phi}^j(x)} \\ &= \dot{\phi}^k(x) g_{jk}(\phi) + 2\,\partial_x \phi^k(x)\omega_{jk}(\phi).\end{aligned} \tag{2.27}$$

Note that the second term in (2.27) provides an analog of the gauge potential A_j [cf. $p_j = \dot{q}_j + A_j(q)$] in the space of field configurations ΩM (the space of loops in M with a given base point):

$$\mathcal{A}_j = 2\,\partial_x \phi^k(x)\omega_{jk}(\phi). \tag{2.28}$$

To relate the second term in (2.27) to topological properties of the configuration space, one must write it as a differential form on ΩM (cf. $A = A_j\, dx^j$). We will briefly digress to describe the relevant ideas of differential geometry on ΩM.

A vector field on ΩM is intuitively given by an infinitesimal deformation of based-point loops in M. Thus the set of basis vectors at each point of ΩM may be denoted by $\{\delta\phi^j\}$. (Similar definitions involving the other sectors C_a follow). To describe a one-form on ΩM, we think of it as an object which sends vectors on ΩM to real numbers. Noting that for every vector $\delta\phi^j$, $\delta\phi^j(x)$ is a number, we may denote the basis of one-forms on ΩM as $\{\delta\phi^j(x)\}$ (following the notation of Crnkovic and Witten[9]). These objects anticommute (resembling the wedge product of ordinary one-forms), i.e.,

$$\delta\phi^j(x)\delta\phi^k(x') = -\delta\phi^k(x')\delta\phi^j(x). \tag{2.29}$$

The exterior derivative is denoted by δ and obeys

$$\delta^2 = 0. \tag{2.30}$$

Interpreting the zero-form $\phi^j(x)$ as an object that sends the function ϕ^j to the number $\phi^j(x)$ for each x, we find that $\delta\phi^j(x)$ is a closed one-form, i.e., it obeys

$$\delta(\delta\phi^j(x)) = 0. \tag{2.31}$$

A general k form $\mathcal{K}$ may be written as

$$\begin{aligned}\mathcal{K} = \int_{S^1} dx^{(1)} \int_{S^1} dx^{(2)} \cdots \int_{S^1} dx^{(k)} \\ \times \mathcal{K}_{i_1 i_2 \cdots i_k}(\phi(x^{(1)}),\ldots,\phi(x^{(k)})) \\ \times \delta\phi^{i_1}(x^{(1)}) \cdots \delta\phi^{i_k}(x^{(k)}).\end{aligned} \tag{2.32}$$

where the parenthesized indices on the x's are just labels for different x's and $\mathcal{K}_{i_1 i_2 \cdots i_k}$ is a functional of ϕ at the various points. Note that apart from the usual summation convention over the indices $i_1, i_2, \ldots, i_k$, there is also a "summation" over the different x's, showing the infinite dimension of ΩM. Note that in our problem we will only be interested in a

subclass of such forms, in which the functional $\mathscr{K}_{i_1 i_2 \cdots i_k}$ will be a functional of ϕ at one point x. This is because such functionals will be appearing in the Lagrangian for the theory and hence required to be local. Last, given a vector $\delta\phi^j = v^j$ on ΩM, the contraction of v with $\mathscr{K}$ is given by

$$v \rfloor \mathscr{K} = \sum_{i_j} \int_{S^1} dx^{(1)} \int_{S^1} dx^{(2)} \cdots \int_{S^1} dx^{(k)}\, v^{i_j} \mathscr{K}_{i_1 i_2 \cdots i_k} \times \delta\phi^{i_1}(x^{(1)}) \cdots \widehat{\delta\phi^{i_j}(x^{(j)})} \cdots \delta\phi^{i_k}(x^{(k)}). \tag{2.33}$$

Returning to the discussion of $\mathscr{A}_j$, it can now be understood as the components of the gauge potential one-form

$$\mathscr{A}[\delta\phi^j] = \int_{S^1} dx\{2\,\partial_x \phi^k(x)\omega_{jk}(\phi)\delta\phi^j(x)\} = \int_{S^1} dx\{\mathscr{A}_j(\phi(x))\delta\phi^j(x)\}. \tag{2.34}$$

The field strength two-form on ΩM (cf. $F = dA$) can be obtained by applying the exterior derivative δ at point x, i.e.,

$$\begin{aligned}\mathscr{F}[\delta\phi^l,\delta\phi^j] &= \delta(\mathscr{A}[\delta\phi^j])[\delta\phi^l] \\ &= \int_{S_i} dx\{2(\partial_x \delta\phi^k(x))\omega_{jk}\,\delta\phi^j(x) \\ &\quad + 2(\partial_x\phi^k)\omega_{jk,l}\,\delta\phi^l(x)\delta\phi^j(x)\} \\ &= \int_{S^1} dx\{3(\partial_x\phi^k)\omega_{jk,l}\,\delta\phi^l(x)\delta\phi^j(x)\},\end{aligned} \tag{2.35}$$

where $\omega_{jk,l}$ denotes the derivative of ω_{jk} with respect to the field ϕ^l. In deriving Eq. (2.35), we have used the symmetries of ω,

$$\omega_{jk} = -\omega_{kj}, \quad \omega_{jk,l} = -\omega_{jl,k}, \tag{2.36}$$

and the fact that ϕ^j is periodic in x. The Wess–Zumino action can now be written in terms of forms on ΩM:

$$\begin{aligned}\Gamma[(\partial\tilde{\phi})_I] &= \int dt \int_{S^1} dx\{2\dot{\phi}^j(x)(\partial_x \phi^k(x))\omega_{jk}(\phi)\} \\ &= \int dt \int_{S^1} dx\{\dot{\phi}^j(x)\mathscr{A}_j(\phi)\} \\ &= \int_{S^1} dx \int dt \left\{\frac{\partial\phi^j}{\partial t}\mathscr{A}_j(\phi)\right\} \\ &= \int_{S^1} dx \int_\gamma \delta\phi^j(x)\mathscr{A}_j(\phi) = \int_\gamma \mathscr{A},\end{aligned} \tag{2.37}$$

where γ is the path traversed in ΩM. A more useful form will be

$$\Gamma[(\partial\tilde{\phi})_I] = \int dt\, \mathscr{A}[\dot{\phi}], \tag{2.38}$$

treating $\dot{\phi}$ as a vector at each point of ΩM and contracting it with $\mathscr{A}$ (cf. $\int dt\, A_i \dot{x}^i$).

Having mentioned the analogs of the gauge potential and field strength in ΩM, we must now check that they have the proper gauge symmetry properties. The analog of a gauge transformation of $\mathscr{A}$ is

$$\mathscr{A} \to \mathscr{A}' = \mathscr{A} + \delta\Lambda \tag{2.39}$$

for some zero-form Λ on ΩM. From the form of the Wess–Zumino Lagrangian, $\delta\Lambda$ is required to be

$$\delta\Lambda = \int_{S^1} dx\{2\,\partial_x\phi^k\,\alpha_{kj}\delta\phi^j(x)\} \tag{2.40}$$

for some functional one-form $\alpha(\phi)$ on M: This implies that

$$\Lambda = \int_{S^1} dx\{\partial_x\phi^k(x)\alpha_k(\phi)\}. \tag{2.41}$$

Note that in deriving Eq. (2.41), it is necessary that Λ is a functional of fields only from the trivial sector, namely $\phi^j(x)$. This means that the gauge transformation for the whole action comes from the trivial sector. This is precisely the reason why (2.16) cannot be put in a Lagrangian form, as it depends on the choice of function $\omega(C_a)$ [i.e., it is no longer gauge (quasi-) invariant]. We also find that as required, $\mathscr{F}$ is gauge invariant under the transformation (2.39):

$$\begin{aligned}\mathscr{F} \to \mathscr{F}' &= \int_{S^1} dx\{3(\omega_{jk,l} + \alpha_{k,jl})(\partial_x\phi^k)\delta\phi^l(x)\delta\phi^j(x)\} \\ &= \int_{S^1} dx\{3\omega_{jk,l}(\partial_x\phi^k)\delta\phi^l(x)\delta\phi^j(x)\} = \mathscr{F}.\end{aligned} \tag{2.42}$$

The Wess–Zumino action is also well defined under transformation (2.39) since it changes by a total time derivative given by

$$\begin{aligned}&\int dt \int_{S^1} dx\{\epsilon^{\mu\nu}\,\partial_\mu\phi^j\,\partial_\nu\phi^k\,\alpha_{j,k}\} \\ &= \int dt \int_{S^1} dx\{\partial_\nu(\epsilon^{\mu\nu}\,\partial_\mu\phi^j\,\alpha_j)\} \\ &= \int dt \frac{\partial}{\partial t}\left(-\int_{S^1} dx\{\partial_x\phi^j\,\alpha_j\}\right)\end{aligned} \tag{2.43}$$

[cf. $I_{\text{int}} = \int dt(A_i\dot{x}^i) \to I_{\text{int}} + \int dt(\partial_i\Lambda\dot{x}^i) = I_{\text{int}} + \int dt(d\Lambda/dt)$]: This can be ignored as it does not contribute to the dynamics of theory.

III. NOETHER'S THEOREM AND CONSTANTS OF MOTION

In Sec. II, we have seen how the Wess–Zumino action can be interpreted as the interacting part of a total action for a "particle" in a background "magnetic field" on an infinite-dimensional space exhibiting the appropriate gauge symmetries. Here, we will proceed by looking at space-time symmetries of the σ model and, by using the above interpretation, one can show that Noether's theorem gives constants of motion modified by a contribution from the "background field." This further elaborates the particle analogy.

Consider the Lagrangian density (2.26) as

$$\mathscr{L} = \mathscr{L}_{\text{KE}} + \mathscr{L}_{\text{WZ}},$$

where

$$\mathscr{L}_{\text{KE}} = \tfrac{1}{2}(\partial_\mu\phi^j)(\partial^\mu\phi^k)g_{jk}(\phi), \tag{3.1}$$

$$\mathscr{L}_{\text{WZ}} = \epsilon^{\mu\nu}(\partial_\mu\phi^j)(\partial_\nu\phi^k)\omega_{jk}(\phi). \tag{3.2}$$

Let ϕ^j transform as

$$\phi^j \to \phi^j + \delta\phi^j,$$

with

$$\delta\phi^j = v^j. \tag{3.3}$$

First, we let the Lagrangian density (3.1) be invariant under transformation (3.3) (for a more detailed discussion of the symmetries of $\mathscr{L}_{\mathrm{KE}}$ see Ref. 10:

$$\begin{aligned}\delta\mathscr{L}_{\mathrm{KE}} &= \tfrac{1}{2}(\partial_\mu v^j\,\partial^\mu\phi^k\,g_{jk} + \partial_\mu\phi^j\,\partial^\mu v^k\,g_{jk} \\ &\quad + \partial_\mu\phi^j\,\partial^\mu\phi^k\,g_{jk,l}v^l) \\ &= \tfrac{1}{2}(\partial_\mu\phi^l\,\partial^\mu\phi^k)(v^j{}_{,l}g_{jk} + v^j{}_{,k}g_{lj} + g_{lk,j}v^j) \\ &= 0;\end{aligned}$$

this implies

$$v^j{}_{,l}g_{jk} + v^j{}_{,k}g_{lj} + g_{lk,j}v^j = 0, \tag{3.4}$$

i.e., v must be a Killing vector on (M,g).

To see how $\mathscr{L}_{\mathrm{WZ}}$ responds to transformation (3.3) and, in particular, to see when (3.3) is a symmetry transformation, it is important to recall that the Wess–Zumino action may be written in terms of a gauge potential one-form [see (2.38)]. Interpreting $\delta\phi^j$ as a vector field on ΩM, the gauge potential one-form $\mathscr{A}$ transforms under (3.3) in a way given by its Lie derivative with respect to $\delta\phi = v$, i.e.,

$$\mathscr{A} \to \mathscr{A}' = \mathscr{A} + \pounds_v\mathscr{A}. \tag{3.5}$$

Thus to make $\mathscr{L}_{\mathrm{WZ}}$ invariant one can impose $\pounds_v\mathscr{A} = 0$, but note that we can use the gauge freedom to modify this into

$$\pounds_v\mathscr{A} = \delta\mathscr{W}_v(\phi) \tag{3.6}$$

for some scalar $\mathscr{W}_v(\phi)$ on ΩM: From Eq. (2.43) this is equivalent to the condition that $\mathscr{L}_{\mathrm{WZ}}$ may change by a total time derivative. The change in $\mathscr{L}_{\mathrm{WZ}}$ under transformation (3.3)is explicitly given by

$$\begin{aligned}\delta\mathscr{L}_{\mathrm{WZ}} &= 2\epsilon^{\mu\nu}(\partial_\mu v^j)(\partial_\nu\phi^k)\omega_{jk} \\ &\quad + \epsilon^{\mu\nu}(\partial_\mu\phi^j)(\partial_\nu\phi^k)\omega_{jk,l}v^l \\ &= \epsilon^{\mu\nu}(\partial_\mu\phi^j)(\partial_\nu\phi^k)(2\omega_{lk}v^l{}_{,j} + \omega_{jk,l}v^l).\end{aligned} \tag{3.7}$$

Equation (3.7) may be set to equal the total derivative.

$$\partial_\mu(\epsilon^{\mu\nu}\phi^j(\partial_\nu\phi^k)(2\omega_{lk}v^l{}_{,j} + \omega_{jk,l}v^l)),$$

provided that the following condition holds:

$$2\,\partial_{[m}\omega_{k]l}v^l{}_{,j} + \partial_{[m}\omega_{k]j,l}v^l = 0.$$

By using symmetries in (2.36) and from Eq. (2.43), a sufficient condition for Eq. (3.6) to hold is then

$$\partial_m(2\omega_{kl}v^l{}_{,j} + \omega_{kj,l}v^l) = 0. \tag{3.8}$$

Note that in general condition (3.8) is not true. For such cases it is necessary to treat Eq. (3.7) case by case for different M and ω (one such case is our second example in Sec. IV). For simplicity, we shall assume Eq. (3.8) in order to illustrate our point on the modified constants of motion.

Now the Lie derivative of $\mathscr{A}$ can also be expressed formally using the homotopy formula $\pounds_v(\cdot) = \delta(v\,\lrcorner\,\cdot) + v\,\lrcorner\,\delta(\cdot)$, in particular,

$$\pounds_v\mathscr{A} = \delta(v\,\lrcorner\,\mathscr{A}) + v\,\lrcorner\,(\delta\mathscr{A}) = \delta(\mathscr{A}[v]) + v\,\lrcorner\,\mathscr{F}. \tag{3.9}$$

Comparing Eq. (3.9) with Eq. (3.6) implies a new condition

$$v\,\lrcorner\,\mathscr{F} = -\delta\psi_v \tag{3.10}$$

for some scalar ψ_v on ΩM, with

$$\mathscr{W}_v = \mathscr{A}[v] - \psi_v \tag{3.11}$$

[cf. Eq. (1.8) in Ref. 1]. Of relevance to our discussion on the constants of motion is the gauge invariant object ψ_v. From Eq. (3.10) we note that ψ_v is globally well defined only if $v\,\lrcorner\,\mathscr{F}$ is exact. Note that $v\,\lrcorner\,\mathscr{F}$ is necessarily closed since $\mathscr{F}$, being gauge invariant, must be invariant under the symmetry transformation (3.3), i.e.,

$$\pounds_v\mathscr{F} = \delta(v\,\lrcorner\,\mathscr{F}) = 0. \tag{3.12}$$

Thus $v\,\lrcorner\,\mathscr{F}$ belongs to the first cohomology class of ΩM. A sufficient condition for a globally well defined ψ_v is then

$$H^1(\Omega M) = 0. \tag{3.13}$$

This is always the case for simply connected configuration spaces, i.e.,

$$\pi_1(\Omega M) \cong \pi_2(M) = 0. \tag{3.14}$$

However, for other spaces there is the possibility of $v\,\lrcorner\,\mathscr{F}$ being closed, but nonexact; thus Eq. (3.10) is only true locally.

In addition to Eq. (3.10), one can also obtain another equation for ψ_v involving further contraction of $\mathscr{F}$ with $v = [w,u]$ for some vector fields w, u on ΩM, namely,

$$\psi_{[w,u]} = \mathscr{F}[w,u] \tag{3.15}$$

[cf. Eq. (1.14b) in Ref. 1].

Proof: Consider the following identity:

$$\pounds_w\pounds_u\mathscr{A} - \pounds_u\pounds_w\mathscr{A} = \pounds_{[w,u]}\mathscr{A} = \delta\mathscr{W}_{[w,u]}, \tag{3.16}$$

The lhs of Eq. (3.16) gives

$$\begin{aligned}\pounds_w(\delta\mathscr{W}_u) - \pounds_u(\delta\mathscr{W}_w) &= \delta(w\,\lrcorner\,\delta\mathscr{W}_u) - \delta(u\,\lrcorner\,\mathscr{W}_w) \\ &= \delta(\pounds_w\mathscr{W}_u - \pounds_u\mathscr{W}_w).\end{aligned} \tag{3.17}$$

The rhs of Eq. (3.16) with Eq. (3.17) gives the identity

$$\mathscr{W}_{[w,u]} = \pounds_w\mathscr{W}_u - \pounds_u\mathscr{W}_w. \tag{3.18}$$

Using Eq. (3.11) in Eq. (3.18) we obtain

$$\begin{aligned}\mathscr{A}[[w,u]] - \psi_{[w,u]} &= \pounds_w\mathscr{A}[u] - \pounds_w\psi_u \\ &\quad - \pounds_u\mathscr{A}[w] + \pounds_u\psi_w \\ &= w(\mathscr{A}[u]) - (w\,\lrcorner\,\delta\psi_u) \\ &\quad - u(\mathscr{A}[w]) + (u\,\lrcorner\,\delta\psi_w).\end{aligned}$$

Thus

$$\begin{aligned}\psi_{[w,u]} &= u(\mathscr{A}[w]) - w(\mathscr{A}[u]) + \mathscr{A}[[w,u]] \\ &\quad - u\,\lrcorner\,(\delta\psi_w) + w\,\lrcorner\,(\delta\psi_u).\end{aligned} \tag{3.19}$$

Using Eq. (3.11) with the identity

$$\mathscr{F}[u,w] = u(\mathscr{A}[w]) - w(\mathscr{A}[u]) + \mathscr{A}[[w,u]] \tag{3.20}$$

in Eq. (3.19) will now give the desired identity

$$\psi_{[w,u]} = \mathscr{F}[u,w] + \mathscr{F}[w,u] - \mathscr{F}[u,w] = \mathscr{F}[w,u].$$

□

A useful computation is that of $\pounds_u\mathscr{A}$ using Eqs. (3.9) and (2.34):

$$\pounds_v \mathscr{A} = \int_{S^1} dx\{\delta(2(\partial_x\phi^k)\omega_{jk}v^j) + 3(\partial_x\phi^k)\omega_{jk,l}v^l\,\delta\phi^j(x) - 3(\partial_x\phi^k)\omega_{jk,l}v^j\,\delta\phi^l(x)\} = \int_{S^1} dx\{2\,\partial_x(\delta\phi^k(x)\omega_{jk}v^j)$$
$$- 2\,\delta\phi^k(x)\omega_{jk,l}v^j(\partial_x\phi^l) - 2\,\delta\phi^k(x)\omega_{jk}v^j{}_{,l}(\partial_x\phi^l) + 2(\partial_x\phi^k)\omega_{jk}v^j{}_{,l}\,\delta\phi^l(x) + 4(\partial_x\phi^k)\omega_{jk,l}v^l\,\delta\phi^j(x)\}$$
$$= \int_{S^1} dx\{2\omega_{jk}v^j{}_{,l}(\delta\phi^l(x)\partial_x\phi^k - \delta\phi^k(x)\partial_x\phi^l) + 2(\partial_x\phi^k)\omega_{jk,l}v^l\,\delta\phi^j(x)\}$$
$$= \int_{S^1} dx\{(\partial_x\phi^k\,\delta\phi^l(x) - \delta\phi^k(x)\partial_x\phi^l)(2\omega_{jk}v^j{}_{,l} + \omega_{lk,j}v^j)\}. \tag{3.21}$$

Equation (3.21) is in fact consistent with the change in $\mathscr{L}_{WZ}$ when the action is written in terms of $\mathscr{A}$:

$$\int_{S^1} dx\,\delta\mathscr{L}_{WZ}$$
$$= \pounds_v \mathscr{A}[\phi]$$
$$= \int_{S^1} dx\{(\dot\phi^l\partial_x\phi^k - \dot\phi^k\partial_x\phi^l)(2\omega_{jk}v^j{}_{,l} + \omega_{lk,j}v^j)\}$$
$$= \int_{S^1} dx\{\epsilon^{\mu\nu}\partial_\mu\phi^l\,\partial_\nu\phi^k(2\omega_{jk}v^j{}_{,l} + \omega_{lk,j}v^j)\}.$$

Given the above results one can now discuss conserved currents and hence the associated constants of motion with respect to the symmetry transformations (3.3). For completeness, we include the following standard discussion of Noether's theorem. A Lagrangian density $\mathscr{L}(\phi,\partial_\mu\phi)$, under transformation (3.3) changes (without using equations of motion) as

$$\mathscr{L} \to \mathscr{L}' = \mathscr{L} + \partial_\mu K^\mu \tag{3.22}$$

for some K^μ. With the equations of motion the Lagrangian density transforms as

$$\mathscr{L} \to \mathscr{L}' \approx \mathscr{L} + \partial_\mu\delta\phi^j\left(\frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi^j)}\right) + \delta\phi^j\left(\frac{\delta\mathscr{L}}{\delta\phi^j}\right)$$
$$= \mathscr{L} + \partial_\mu\left(\delta\phi^j\frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi^j)}\right), \tag{3.23}$$

Thus a conserved current can be constructed from the identity

$$\partial_\mu K^\mu = \partial_\mu\left(\delta\phi^j\frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi^j)}\right), \tag{3.24}$$

namely,

$$J^\mu = \delta\phi^j\frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi^j)} - K^\mu. \tag{3.25}$$

From the total Lagrangian density, (3.1) with (3.2), K^μ is given by

$$K^\mu = \epsilon^{\mu\nu}\phi^j(\partial_\nu\phi^k)(2\omega_{lk}v^l{}_j + \omega_{jk,l}v^l) \tag{3.26}$$

[see (3.7)]. Hence the current J^μ is

$$J^\mu = \partial^\mu\phi^j g_{jk}(\phi)v^k + 2\epsilon^{\mu\nu}\partial_\nu\phi^k\,\omega_{lk}v^l$$
$$- \epsilon^{\mu\nu}\phi^j(\partial_\nu\phi^k)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l). \tag{3.27}$$

One can verify using the equations of motion that

$$\partial_\mu J^\mu = 0.$$

Hence one can construct constants of motion C_v out of the time component of J^μ such that

$$\frac{\partial C_v}{\partial t} = \int_{S^1} dx\,\frac{\partial J^0}{\partial t} = -\int_{S^1} dx\,\frac{\partial J^1}{\partial x} = 0.$$

Computation of C_v from Eq. (3.27) gives

$$C_v = \int_{S^1} dx\{\dot\phi^k g_{jk}v^j + 2\,\partial_x\phi^k\omega_{jk}v^j$$
$$- \phi^j(\partial_x\phi^k)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l)\}. \tag{3.28}$$

The main point now is to understand what the terms in Eq. (3.28) mean. We begin by taking the last term in (3.28) and computing its exterior derivative:

$$\int_{S^1} dx\{\delta(\phi^j(\partial_x\phi^k)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l))\}$$
$$= \int_{S^1} dx\{\delta\phi^j(x)(\partial_x\phi^k)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l) + \phi^j(\partial_x\delta\phi^k(x))(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l)$$
$$+ \phi^j(\partial_x\phi^k)\partial_m(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l)\delta\phi^m(x)\}$$
$$= \int_{S^1} dx\{\delta\phi^j(x)(\partial_x\phi^k)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l) + \partial_x(\phi^j\,\delta\phi^k(x)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l))$$
$$- (\partial_x\phi^j)\delta\phi^k(x)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l) - \phi^j\,\delta\phi^k(x)\partial_m(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l)(\partial_x\phi^m)\}$$
$$= \int_{S^1} dx\{(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l)(\delta\phi^j(x)\partial_x\phi^k - \delta\phi^k(x)\partial_x\phi^j)\} = \pounds_v\mathscr{A}, \tag{3.29}$$

where we have used Eqs. (3.8) and (3.21) and the fact that the ϕ^i's are periodic functions of x. Hence Eqs. (3.29) and (3.8) imply that the third term in (3.28) is simply

$$\mathscr{W}_v = \int_{S^1} dx\{\phi^j(\partial_x\phi^k)(2v^l{}_j\omega_{lk} + \omega_{jk,l}v^l)\}. \tag{3.30}$$

The second term in (3.28) is straightforwardly given by $\mathscr{A}[v]$, while the first term is just the normal contribution from the kinetic term. Writing the first term in (3.28) C_{v_0}, we have

$$C_v = C_{v_0} + \mathscr{A}[v] - \mathscr{W}_v = C_{v_0} + \psi_v. \tag{3.31}$$

Thus we find that the normal constant of motion C_{v_0} is supplemented by ψ_v, the contraction of the field strength $\mathscr{F}$ with the Killing vector field v. This justifies the earlier claim that Noether's theorem gives a modified constant of motion that includes a contribution from the background field.

Having obtained these results, we will now illustrate them using specific examples of σ models with the Wess–Zumino term.

IV. THE σ MODEL ON $M = T^3$

The first example is the σ model on the target manifold $M = T^3$. Here, all the results derived in Secs. II and III hold. We will now make the results more explicit for this particular model.

From the construction of the Wess–Zumino action (in the usual sense), there is only one independent action given by the (integral of the) generator of $H^3(T^3)$, which is the volume form

$$\Omega = d\phi^1 \wedge d\phi^2 \wedge d\phi^3, \tag{4.1}$$

where ϕ^i $(i = 1, 2, 3)$ are the angular (field) variables of T^3. Here, Ω can be represented locally as the exterior derivative (d) of the two-form

$$\omega = \tfrac{1}{8}\epsilon_{ijk}\phi^i\, d\phi^j \wedge d\phi^k. \tag{4.2}$$

Given such an ω, the gauge potential one-form (2.34) is simply

$$\mathscr{A} = \int_{S^1} dx\left\{\frac{1}{3}\epsilon_{ijk}\phi^i\,\partial_x\phi^k\,\delta\phi^j(x)\right\}, \tag{4.3}$$

while the field strength two-form (2.35) is

$$\mathscr{F} = \int_{S^1} dx\left\{\frac{1}{2}\epsilon_{ijk}\,\partial_x\phi^k\,\delta\phi^i(x)\delta\phi^j(x)\right\}. \tag{4.4}$$

The Lagrangian density of this model may now be written as

$$\mathscr{L} = \tfrac{1}{2}\,\partial_\mu\phi^j\,\partial^\mu\phi^k\eta_{jk} + \tfrac{1}{8}\epsilon^{\mu\nu}\epsilon_{ijk}\phi^i\,\partial_\mu\phi^j\,\partial_\nu\phi^k, \tag{4.5}$$

where η is the flat metric on T^3.

The Killing vectors on (T^3,η) that will generate the symmetry transformations are just the vectors v generating translations, so that

$$v = v^1\frac{\partial}{\partial\phi^1} + v^2\frac{\partial}{\partial\phi^2} + v^3\frac{\partial}{\partial\phi^3}, \tag{4.6}$$

where the v^i's are constants. The induced vector field on the configuration space ΩT^3 is obtained by Lie dragging the coordinate functions by the vector field v, i.e.,

$$\delta\phi^j = \pounds_v\phi^j = v^j. \tag{4.7}$$

Under this symmetry transformation, the Lagrangian density (4.5) changes by a total derivative as in (3.7) [note that ω and v satisfy condition (3.8)]:

$$\delta\mathscr{L} = \partial_\mu K^\mu,$$

where

$$K^\mu = \tfrac{1}{8}\epsilon^{\mu\nu}\epsilon_{ijk}v^i\phi^j\,\partial_\nu\phi^k. \tag{4.8}$$

Thus the conserved current J^μ [(3.25)] is simply given by

$$J^\mu = v^j\,\partial^\mu\phi^k\eta_{jk} + \tfrac{1}{2}\epsilon^{\mu\nu}\epsilon_{ijk}v^j\phi^i\,\partial_\nu\phi^k. \tag{4.9}$$

The constant of motion associated with the current (4.9) is then

$$C_v = \int_{S^1} dx\, J^0 = \int_{S^1} dx\left\{v^j\dot{\phi}^k\eta_{jk} + \frac{1}{2}\epsilon_{ijk}v^j\phi^i\,\partial_x\phi^k\right\}. \tag{4.10}$$

Note that the second term in (4.10) may be written as the contraction of the field strength two-form (4.4) with ϕ and v, i.e., $\mathscr{F}[\phi,v]$ (with an abuse of notation; ϕ is not a vector on ΩT^3). This can be compared with the case of a particle in a magnetic field in which the analogous term is $F_{jk}x^jv^k$ (x^j is the coordinate function of the configuration space).

At this point, it is appropriate to address the aforementioned possibility of the constants of motion being ill defined. In Ref. 3 it is noted that for the case of a particle on T^n in a magnetic field, the term $F_{jk}x^jv^k$ is not globally defined owing to the multiple-valued coordinate function x^j on the nonsimply connected space T^n. However, in our example this problem does not occur. As the field variable $\phi^j(x)$ undergoes a translation of its period 2π,

$$\phi^j(x) \to \phi^j(x) + 2\pi, \tag{4.11}$$

the change in C_v is trivial:

$$\Delta C_v = \int_{S^1} dx\{\epsilon_{ijk}\pi v^j\,\partial_x\phi^k\} = 0 \tag{4.12}$$

since the function ϕ^k is periodic in x. This is consistent with the fact that the configuration space is now a loop space ΩT^3 of T^3 and is simply connected, i.e.,

$$\pi_1(\Omega T^3) = \pi_2(T^3) = 0. \tag{4.13}$$

In fact $\psi_v = \mathscr{F}[\phi,v]$ must be globally defined as a consequence of (4.13) [see the remarks after Eq. (3.13)].

Thus to find any possible phenomena of "anomalous" constants of motion, one must first require that $\pi_2(M)$ is nontrivial. Such an example will be discussed in Sec. V.

V. THE σ MODEL ON $M = S^2 \times S^1$

This model is a more interesting example than that in Sec. IV as the target manifold $M = S^2 \times S^1$ has a nontrivial second homotopy group, which means that the space of field configurations is no longer simply connected. However, this also means that one encounters an ambiguity in the construction of the Wess–Zumino action (see note (i) in Sec. II). We will nevertheless proceed as in Sec. III. A comment regarding the ambiguity will be made in Sec. VI.

The Wess–Zumino action is constructed from the one generator of $H^3(S^2 \times S^1)$ that is given by the volume form:

$$\Omega = \sin\phi^1\, d\phi^1 \wedge d\phi^2 \wedge d\phi^3, \tag{5.1}$$

where ϕ^1 and ϕ^2 are now spherical coordinates on S^2 and ϕ^3 is the angular coordinate on S^1. Locally, Eq. (5.1) is given by $\Omega = d\omega$, where

$$\omega = -(\cos\phi^1 \mp 1)d\phi^2 \wedge d\phi^3. \tag{5.2}$$

With the metric of $S^2 \times S^1$ given by

$$ds^2 = g_{jk}\, d\phi^j \otimes d\phi^k = (d\phi^1)^2 + \sin^2\phi^1 (d\phi^2)^2 + (d\phi^3)^2, \tag{5.3}$$

the total Lagrangian density of the σ model is

$$\mathscr{L} = \tfrac{1}{2}\partial_\mu\phi^j\,\partial^\mu\phi^k g_{jk}(\phi) - \epsilon^{\mu\nu}\partial_\mu\phi^2\,\partial_\nu\phi^3(\cos\phi^1 \mp 1). \tag{5.4}$$

The gauge potentials computed from the Lagrangian density (5.4) are given by

$$\mathscr{A}_N = \int_{S^1} dx\{-(\cos\phi^1 - 1)(\partial_x\phi^3\,\delta\phi^2(x) - \partial_x\phi^2\,\delta\phi^3(x))\}. \tag{5.5}$$

$$\mathscr{A}_S = \int_{S^1} dx\{-(\cos\phi^1 + 1)(\partial_x\phi^3\,\delta\phi^2(x) - \partial_x\phi^2\,\delta\phi^3(x))\}. \tag{5.6}$$

Note that (5.5) and (5.6) are only well defined locally in the regions

$$N_\epsilon = \{(\phi^1,\phi^2)\,|\,0 \leqslant \phi^1 < \pi/2 + \epsilon, 0 \leqslant \phi^2 < 2\pi\}, \tag{5.7}$$

$$S_\epsilon = \{(\phi^1,\phi^2)\,|\,\pi/2 - \epsilon < \phi^1 \leqslant \pi, 0 \leqslant \phi^2 < 2\pi\} \tag{5.8}$$

of S^2, respectively ($\epsilon > 0$). A gauge potential that is well defined on the whole of S^2 (and hence of ΩM) can then be given by

$$\mathscr{A} = \begin{cases} \mathscr{A}_N, & \text{on } N_\epsilon, \\ \mathscr{A}_S, & \text{on } S_\epsilon, \end{cases} \tag{5.9}$$

with the observation that $\mathscr{A}_N$ and $\mathscr{A}_S$ are gauge related on $(N_\epsilon \cap S_\epsilon)$ by

$$\xi = \mathscr{A}_S - \mathscr{A}_N = -2\int_{S^1} dx\{\partial_x\phi^3\,\delta\phi^2(x) - \partial_x\phi^2\,\delta\phi^3(x)\}. \tag{5.10}$$

Note that (5.10) is an exact one-form

$$\xi = \delta e, \tag{5.11}$$

where e is the zero-form

$$e = -2\int_{S^1} dx\{(\partial_x\phi^3)\phi^2(x)\} \tag{5.12}$$

(e is well defined under the translations $\phi^i \to \phi^i + 2\pi$ for $i = 2,3$). The field strength $\mathscr{F}$ is simply given by

$$\mathscr{F} = \int_{S^1} dx\left\{\frac{1}{2}\sin\phi^1 \epsilon_{ijk}\,\partial_x\phi^k\,\delta\phi^i(x)\delta\phi^j(x)\right\}. \tag{5.13}$$

For this model, the symmetry transformations are generated by the following Killing vectors on $S^2 \times S^1$:

$$v_{(1)} = \sin\phi^2 \frac{\partial}{\partial\phi^1} + \cot\phi^1\cos\phi^2\frac{\partial}{\partial\phi^2}, \tag{5.14}$$

$$v_{(2)} = \frac{\partial}{\partial\phi^2}, \tag{5.15}$$

$$v_{(3)} = \frac{\partial}{\partial\phi^3}, \tag{5.16}$$

$$v_{(4)} = -\left(\cos\phi^2\frac{\partial}{\partial\phi^1} - \cot\phi^1\sin\phi^2\frac{\partial}{\partial\phi^2}\right). \tag{5.17}$$

Here, the parenthesized indices are just labels denoting different Killing vectors. We can now construct the associated constants of motion for each of the symmetry transformations given by $\delta\phi^j = v^j_{(i)}$ ($i = 1,\ldots,4$): It is important to note that condition (3.8) does not hold for these cases and one has to repeat any necessary computations of Sec. III that assume this condition.

A. $i=1$

The vector field ΩM is given by

$$\delta\phi^j = \sin\phi^2\,\delta^{j1} + \cot\phi^1\cos\phi^2\delta^{j2}, \tag{5.18}$$

where δ^{ij} on the rhs is the Kronecker delta. The change in Lagrangian density by such a transformation is a total derivative, i.e., $\delta\mathscr{L} = \partial_\mu K^\mu_{(1)}$, where

$$K^\mu_{(1)} = -\epsilon^{\mu\nu}(\csc\phi^1 \mp \cot\phi^1)\cos\phi^2\,\partial_\nu\phi^3. \tag{5.19}$$

Hence the associated constant of motion constructed from transformation (5.18) will be

$$C_{v_{(1)}} = \int_{S^1} dx\{\dot\phi^1\sin\phi^2 + \dot\phi^2\sin\phi^1\cos\phi^1\cos\phi^2 + (\partial_x\phi^3)\sin\phi^1\cos\phi^2\}. \tag{5.20}$$

Here the contribution from the field strength $\mathscr{F}$,

$$\psi_{v_{(1)}} = \int_{S^1} dx\{\sin\phi^1\cos\phi^2\,\partial_x\phi^3\}, \tag{5.21}$$

is no longer as transparent as ψ_v in Sec. III. However, if we take the exterior derivative δ of $\psi_{v_{(1)}}$ we find

$$\delta\psi_{v_{(1)}} = \int_{S^1} dx\{\cos\phi^1\cos\phi^2(\partial_x\phi^3\,\delta\phi^1(x) - \partial_x\phi^1\,\delta\phi^3(x)) + \sin\phi^1\sin\phi^2 \times(\partial_x\phi^2\,\delta\phi^3(x) - \partial_x\phi^3\,\delta\phi^2(x))\} = -v_{(1)} \lrcorner \mathscr{F}, \tag{5.22}$$

thus confirming our previous results.

B. $i=2$

The vector field on ΩM induced by $v_{(2)}$ is simply given by

$$\delta\phi^j = \delta^{j2} \tag{5.23}$$

and hence the change in Lagrangian density $\delta\mathscr{L}$ is trivial. The constant of motion is then

$$C_{v_{(2)}} = \int_{S^1} dx\{\dot\phi^2\sin^2\phi^1 - (\partial_x\phi^3)\cos\phi^1\} = C_{v_{(2)0}} + \psi_{v_{(2)}} \tag{5.24}$$

($C_{v_{(2)0}}$ is the normal kinetic term contribution). As before, we find that

$$\delta\psi_{v_{(2)}} = \int_{S^1} dx\{\sin\phi^1(\partial_x\phi^3\,\delta\phi^1(x) - \partial_x\phi^1\,\delta\phi^3(x))\}$$
$$= -v_{(2)} \rfloor \mathscr{F}. \tag{5.25}$$

C. $i=3$

For this case similar calculations as above will produce the following results:

$$\delta\phi^j = \delta^3, \tag{5.26}$$

$$C_{v_{(3)}} = \int_{S^1} dx\{\dot{\phi}^3 + (\partial_x\phi^2)\cos\phi^1\} = C_{v_{(3)0}} + \psi_{v_{(3)}}, \tag{5.27}$$

$$\delta\psi_{v_{(3)}} = \int_{S^1} dx\{\sin\phi^1(\partial_x\phi^1\,\delta\phi^2(x) - \partial_x\phi^2\,\delta\phi^1(x))\}$$
$$= -v_{(3)} \rfloor \mathscr{F}. \tag{5.28}$$

Using Eq. (3.15), one finds the field strength contribution to the constant of motion

D. $i=4$

For this case we could proceed with calculations similar to the above; however, we will instead make use of the observation that

$$v_{(4)} = [v_{(1)}, v_{(2)}]. \tag{5.29}$$

Using Eq. (3.15), one finds the field strength contribution to the constant of motion associated with the symmetry transformation $\delta\phi^j = v^j_{(4)}$ to be

$$\psi_{v_{(4)}} = \mathscr{F}[v_{(1)}, v_{(2)}] = \int_{S^1} dx\{\sin\phi^1 \sin\phi^2\,\partial_x\phi^3\}. \tag{5.30}$$

One finds again that

$$\delta\psi_{v_{(4)}} = \int_{S^1} dx\{\cos\phi^1 \sin\phi^2(\partial_x\phi^3\,\delta\phi^1(x) - \delta\phi^3(x)\partial_x\phi^1)$$
$$+ \sin\phi^1\cos\phi^2(\partial_x\phi^3\,\delta\phi^2(x) - \partial_x\phi^2\,\delta\phi^3(x))\}$$
$$= -v_{(4)} \rfloor \mathscr{F} \tag{5.31}$$

Having constructed the constants of motion, we need to check whether or not the $\psi_{v_{(i)}}$'s are globally well defined. As discussed earlier, possible obstructions may occur when $\pi_1(\Omega M)$ is nontrivial. Thus it is natural to look at a noncontractible loop in the configuration space which is generated by this homotopy group. One such loop is shown in Fig. 3.

The reason why such a loop can give a possible obstruction to a well-defined constant of motion can be seen as follows. Consider the loop on S^2 in Fig. 3 as given by the map ϕ. In defining the Wess–Zumino action, the map ϕ has to be extended to $\tilde{\phi}$ (see Sec. II). Replacing the loop given by the map ϕ in Fig. 3 by the image of (a particular) extension $\tilde{\phi}$ gives the corresponding Fig. 4. Figure 4 shows that the Wess–Zumino action changes value as ϕ evolves under such a nontractible loop. Thus constants of motion derived from such an action can also change values under such an evolution of ϕ and hence are ill defined. One has to check this explicitly.

A construction of one such noncontractible loop in ΩM is given as follows. First, the submanifold S^2 of M is embedded in $\mathbb{R}^3$ with a triplet of coordinate functions x_i's:

$$\hat{n} = (x_1, x_2, x_3) = (\sin\phi^1\cos\phi^2, \sin\phi^1\sin\phi^2, \cos\phi^1), \tag{5.32}$$

which satisfies $\hat{n}\cdot\hat{n} = 1$ (the ϕ^i's are coordinates on M). The loop may now be constructed via such a triplet of functions, where they now map $[0,\pi]\times S^1$ to S^2, i.e.,

$$\hat{n} = (\sin\lambda\sin x, \sin^2\lambda\cos x$$
$$+ \cos^2\lambda, \sin\lambda\cos\lambda(\cos x - 1)), \tag{5.33}$$

where $\lambda\in[0,\pi]$ is some parameter and $x\in S^1$ is the coordinate of space. The vector $\hat{n}$ has all the properties required of a noncontractible loop in ΩM, as follows.

(i) $\hat{n}\cdot\hat{n} = 1$.

(ii) For fixed x,

$$\hat{n}|_{\lambda=0} = \hat{n}|_{\lambda=1} = (0,1,0)$$

is a fixed point through which the one-parameter family of loops (parametrized by λ) appears on the submanifold S^2. The map (5.33) actually describes the intersection of a plane with the two-sphere of unit radius, as shown in Fig. 5.

The map (5.33) possesses the required looplike property in ΩS^2 as λ goes from 0 to π.

(iii) The map (5.33) has a topological winding number 1. This can be seen by noting that with the coordinate functions (5.32), the volume form of S^2 is given by

$$\Omega = \epsilon^{ijk}x_i\,dx_j\wedge dx_k. \tag{5.34}$$

For the map (5.33), the volume form is

$$\Omega = \sin\lambda(1-\cos x)d\lambda\wedge dx. \tag{5.35}$$

Integrating (5.35) gives the winding number multiplied by the volume:

$$\int_0^{\pi} d\lambda \int_0^{2\pi} dx \sin\lambda(1-\cos x) = 4\pi.$$

Hence the winding number is 1.

Having obtained the map (5.33), it is now easily verified that the constants of motion are globally well defined [with respect to the function (5.33)]:

$$\Delta\psi_{v_{(i)}} = \psi_{v_{(i)}}|_{(\lambda=\pi)} - \psi_{v_{(i)}}|_{(\lambda=0)} = 0 \quad (i = 1,\ldots,4). \tag{5.36}$$

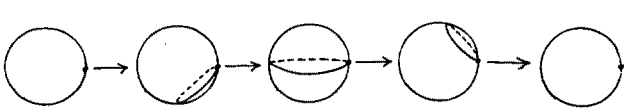


FIG. 3. Evolution of a loop around S^2 of M (S^1 not shown), giving a resultant noncontractible loop of ΩM.

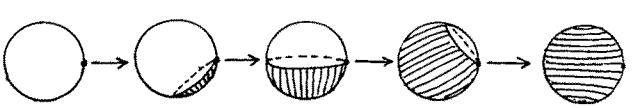


FIG. 4. The corresponding evolution of the image of the extension $\tilde{\phi}$ of σ.

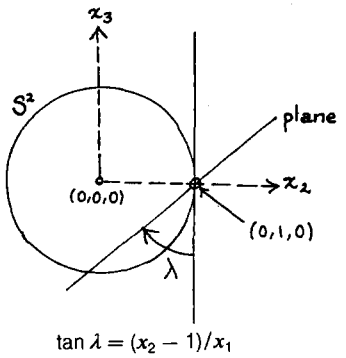


FIG. 5. The intersection of a plane with S^2 in the x_2-x_3 plane.

$\tan\lambda = (x_2 - 1)/x_1$

In fact, one finds that $v_{(i)} \rfloor \mathscr{F}$ is an exact form for each $i = 1,...,4$ (using the triplet of coordinate functions), i.e.,

$$v_{(1)} \rfloor \mathscr{F} = \int_{S^1} dx\, \delta\{-x_1\, \partial_x \phi^3\}, \tag{5.37}$$

$$v_{(2)} \rfloor \mathscr{F} = \int_{S^1} dx\, \delta\{x_3\, \partial_x \phi^3\}, \tag{5.38}$$

$$v_{(3)} \rfloor \mathscr{F} = \int_{S^1} dx\, \delta\{-x_1 \partial_x \phi^2\}, \tag{5.39}$$

$$v_{(4)} \rfloor \mathscr{F} = \int_{S^1} dx\, \delta\{-x_2\, \partial_x \phi^3\}. \tag{5.40}$$

This brings us to the conclusion that while the requirement of ΩM being nonsimply connected is necessary for the existence of "anomalous" constants of motion, it is not sufficient.

VI. SOME REMARKS AND CONCLUSIONS

We have seen from the above discussions how the correspondence between a σ model with a Wess–Zumino term and a particle in a magnetic field can be made closer through a discussion of Noether's theorem. In fact, these σ models may be treated as systems of a particle in a magnetic field on an infinite-dimensional configuration space ΩM. One can further elaborate this by investigating the idea of an "Aharonov–Bohm effect" in ΩM. This is a topological effect which always exists when the configuration space is no longer simply connected [i.e., $\pi_1(\Omega M) = \pi_2(M) \neq 0$ in our case]. Note that the example discussed in Sec. V B has this feature: We will use this example to demonstrate the "Aharonov–Bohm effect" in ΩM. However, prior to this, some comments on the ambiguity in the construction of the Wess–Zumino action associated with the nontrivial $\pi_2(M)$ (see note (i) of Sec. II) are necessary.

Consider the σ model on $M = S^2 \times S^1$ of Sec. V with ϕ mapping the space S^1 into the submanifold S^2 of M. This map has different extensions $\tilde{\phi}$ which are not deformable to each other as a result of the "obstruction" from S^2 of M. For example, the map ϕ that sends S^1 to the equator of S^2, e.g.,

$$\phi^1 = \pi/2, \quad \phi^2 = x, \quad \phi^3 = t \tag{6.1}$$

has the extensions

$$\tilde{\phi}^1 = r\pi/2, \quad \tilde{\phi}^2 = x, \quad \tilde{\phi}^3 = t, \tag{6.2}$$

$$\tilde{\phi}^1 = r\pi/2 + n(1-r)\pi, \quad \tilde{\phi}^2 = x, \quad \tilde{\phi}^3 = t, \tag{6.3}$$

where $r \in [0,1]$ is the radial coordinate of the two-dimensional disk D^2 ($\partial D^2 = S^1$) and n is a positive integer [n is actually the number of times the map (6.2), together with map (6.3), winds around S^2 of M]. (see Fig. 6.)

These extensions (6.2) and (6.3) in fact give different values to the Wess–Zumino action, i.e.,

$$\begin{aligned}\int_{D^2\times I} \tilde{\phi}^*\Omega &= \int_{D^2\times I} \epsilon^{\mu\nu\rho} \sin\tilde{\phi}^1\, \partial_\mu\tilde{\phi}^1\, \partial_\nu\tilde{\phi}^2\, \partial_\rho\tilde{\phi}^3\, r\, dr\, dx\, dt \\ &= \int dt \int_0^{2\pi} dx \int_0^1 dr \left\{ r \sin\left(\frac{r\pi}{2} + n(1-r)\pi \right) \right\} \\ &= \frac{(-1)^n 4I}{(2n-1)} \left\{ \cos\left(\frac{(2n-1)\pi}{2} \right) \right. \\ &\quad \left. - \frac{2}{(2n-1)\pi} \sin\left(\frac{(2n-1)\pi}{2} \right) \right\},\end{aligned} \tag{6.4}$$

where I is the length of the time interval and n takes values from 0, 1, 2,... [$n = 0$ corresponds to extensions (6.2)]. Thus to resolve the ambiguity of the extensions (6.2) and (6.3) one needs to specify this "winding number" n, which then gives a unique Wess–Zumino action.

Having done this, one can now discuss the "Aharonov–Bohm effect" in ΩM. An essential ingredient in this topological effect is that one can obtain a different gauge $\mathscr{A}'$ by performing a singular gauge transformation on $\mathscr{A}$.[11] Consider, then,

$$\begin{aligned}\mathscr{A}_N = \int_{S^1} dx \{ &-(\cos\phi^1 - 1)(\partial_x\phi^3\, \delta\phi^2(x) \\ &- \partial_x\phi^2\, \delta\phi^3(x))\}\end{aligned} \tag{6.5}$$

from Sec. V. We can perform a singular gauge transformation on $\mathscr{A}_N$ by the (nonexact) one-form

$$\xi' = \int_{S^1} dx\, \delta[-(\cos\phi^1 - 1)\phi^3(\partial_x\phi^2 - \partial_x\phi^1)] \tag{6.6}$$

in order to give the gauge potential

$$\begin{aligned}\mathscr{A}'_N = \int_{S^1} dx \{ &\phi^3 \sin\phi^1 (\partial_x\phi^2\, \delta\phi^1(x) - \partial_x\phi^1\, \delta\phi^2(x)) \\ &+ (\cos\phi^1 - 1)(\partial_x\phi^1\, \delta\phi^3(x) - \partial_x\phi^3\, \delta\phi^1(x))\}.\end{aligned} \tag{6.7}$$

Similarly, one can do the same for $\mathscr{A}_S$ to obtain

$$\begin{aligned}\mathscr{A}'_S = \int_{S^1} dx \{ &\phi^3 \sin\phi^1 (\partial_x\phi^2\, \delta\phi^1(x) - \partial_x\phi^1\, \delta\phi^2(x)) \\ &+ (\cos\phi^1 + 1)(\partial_x\phi^1\, \delta\phi^3(x) - \partial_x\phi^3\, \delta\phi^1(x))\}.\end{aligned} \tag{6.8}$$

Both $\mathscr{A}'_N$ and $\mathscr{A}'_S$ can now be "patched" up in the same way as $\mathscr{A}_N$ and $\mathscr{A}_S$ in Sec. V to obtain the desired new gauge

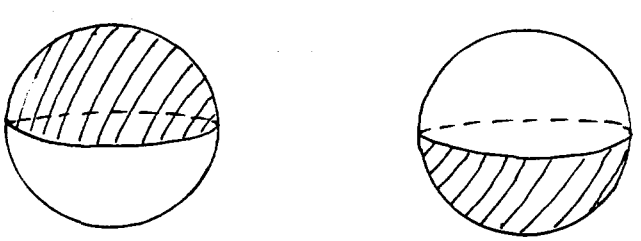


FIG. 6. The shaded regions are the image of different extensions of ϕ [(5.1)] given by (5.2) and (5.3) (with $n = 1$), respectively.

potential $\mathscr{A}'$ on ΩM, which gives the same $\mathscr{F}$ [(5.13)]. However, in the Aharonov–Bohm effect the relevant physical quantity to be determined is the phase factor[12]

$$\exp i\left(\oint \mathscr{A}\right), \tag{6.9}$$

where $\oint$ denotes an integral over a noncontractible closed path in ΩM. (See Fig. 7.)

Thus we only need to obtain different holonomies $\oint \mathscr{A}$, giving different phase factors, to demonstrate the existence of Aharonov–Bohm effect in ΩM. Here, the different holonomies are easily given by the two gauges $\mathscr{A}$ and $\mathscr{A}'$. To show that this is the case we shall use a particular mapping ϕ, namely that given by (5.33), i.e.,

$$\hat{n} = (\sin t \sin x,\ \sin^2 t \cos x + \cos^2 t,\ \sin t \cos t(\cos x - 1)) \tag{6.10}$$

and

$$\phi^3 = 1 \ \text{(a constant mapping)}. \tag{6.11}$$

The parameter t in $\hat{n}$ now denotes the time which parametrizes the noncontractible closed path traversed in ΩM. Using these set of functions, we find that the holonomy of $\mathscr{A}$ is just trivial:

$$\begin{aligned}\oint \mathscr{A} = &\int_0^{\pi/2} dt \int_0^{2\pi} dx\{-(\cos\phi^1 - 1)(\partial_x\phi^3\,\dot{\phi}^2 - \partial_x\phi^2\,\dot{\phi}^3)\} \\ &+ \int_{\pi/2}^{\pi} dt \int_0^{2\pi} dx\{-(\cos\phi^1 + 1) \\ &\times(\partial_x\phi^3\,\dot{\phi}^2 - \partial_x\phi^2\,\dot{\phi}^3)\} = 0\end{aligned} \tag{6.12}$$

as $\dot{\phi}^3 = \partial_x \phi^3 = 0$. for $\mathscr{A}'$, the computation of its holonomy,

$$\oint \mathscr{A}' = \int_0^{\pi} dt \int_0^{2\pi} dx\{\phi^3 \sin\phi^1(\partial_x\phi^2\,\dot{\phi}^1 - \partial_x\phi^1\,\dot{\phi}^2)\}, \tag{6.13}$$

is messy. The integral was done numerically, which gives

$$\oint \mathscr{A}' = -0.4159 \neq 0 \tag{6.14}$$

(to the fourth decimal place). These two results (6.12) and (6.14) then give different phase factors and hence imply the significance of the gauge potentials themselves (as in a normal Aharonov–Bohm effect). It is now important to note that the use of a different gauge potential $\mathscr{A}'$ implies the use of a different local expression of the Wess–Zumino Lagrangian density from that of (5.4), namely

$$\begin{aligned}\mathscr{L}_{WZ} = \epsilon^{\mu\nu}(&\phi^3 \sin\phi^1\,\partial_\mu\phi^1\,\partial_\nu\phi^2 \\ &+ (\cos\phi^1 \pm 1)\partial_\mu\phi^3\,\partial_\nu\phi^1).\end{aligned} \tag{6.15}$$

Thus the Aharonov–Bohm effect in ΩM would then imply that the local expression for the Wess–Zumino Lagrangian density has a physical significance. At this point, it is also tempting to relate the ambiguity of extensions (6.2) and (6.3) for the construction of the Wess–Zumino action [which comes from $\pi_2(M \neq 0]$ with this Aharonov–Bohm effect [which comes from $\pi_1(\Omega M) \neq 0$]. However, earlier we noted that the extensions are labeled by integral winding numbers, while the Aharonov–Bohm effect is labeled by a continuous parameter, say $\lambda \in \mathbb{R}$, given by the holonomy of $\lambda\mathscr{A} + (1-\lambda)\mathscr{A}'$. Therefore, there is no obvious relation.

Finally, we conclude this paper by summarizing the main results obtained above.

(i) A σ model with a Wess–Zumino term on a target manifold M may be treated as a system of a particle in a background magnetic field on a configuration space ΩM. The Wess–Zumino term provides an analog of the gauge potential on ΩM, which then gives the Lagrangian on the σ model a gauge symmetry.

(ii) Like other systems in a background gauge field, the constants of motion associated with the symmetry transformations of the total system are modified by a contribution from the gauge field $\mathscr{F}$ on ΩM.

(iii) There is no "anomalous" phenomenon of ill-defined constants of motion for σ models on $M = T^3$ and $M = S^2 \times S^1$. The second example shows that $\pi_1(\Omega M) \neq 0$ is not a sufficient condition for such phenomena.

(iv) For nonsimply connected configuration spaces ΩM, there is a functional analog of the Aharonov–Bohm effect in ΩM. The holonomy of the gauge potential $\mathscr{A}$ has to be specified to obtain a unique theory. This implies that the local expression of the Wess–Zumino Lagrangian density has a physical significance. In addition to this, one has to specify the "winding number" n to have a well-defined Wess–Zumino action.

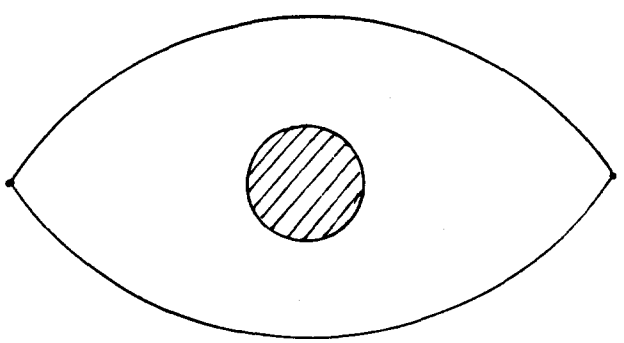


FIG. 7. A schematic diagram of a noncontractible loop in ΩM, where the shaded region is $\text{int}(S^2) \times S^1$.

ACKNOWLEDGMENTS

I am most grateful to Richard S. Ward for his valuable suggestions and discussions and Robert A. Leese for his help with numerical work and useful comments. Thanks are also given to P. E. Dorey and P. Fletcher for helpful comments.

I am also grateful to the Malaysian Public Services Department and the University of Agriculture, Malaysia for their support throughout this work.

[1] R. Jackiw and N. S. Manton, Ann. Phys. **127**, 257 (1980).
[2] Y. S. Wu and A. Zee, Nucl. Phys. B **272**, 322 (1986).
[3] R. S. Ward, Phys. Rev. D **36**, 640 (1987).
[4] N. S. Manton, Ann. Phys. **159**, 220 (1985).
[5] H. Zainuddin, Phys. Rev. D **40**, 636 (1989).
[6] M. J. Greenberg and J. R. Harper, *Algebraic Topology—A First Course* (Benjamin-Cummings, Reading, MA, 1981); J. S. Dowker, *Selected Topics in Topology and Quantum Field Theory* (University of Texas, Austin preprint, 1979).
[7] I. M. Krichever, M. A. Olshanetsky, and A. M. Perelomov, Nucl. Phys. B

264, 415 (1986).
[8] E. Braaten, T. L. Curtright, and C. K. Zachos, Nucl. Phys. B **260**, 630 (1985).
[9] C. Crnkovic and E. Witten, in *300 Years of Gravitation,* edited by S. W. Hawking and W. Israel (Cambridge U. P., Cambridge, 1988).
[10] C. M. Hull, *Lectures on Non-Linear Sigma Models and Strings* (DAMTP preprint, Cambridge, 1986).
[11] B. Felsager, *Geometry, Particles and Fields* (Odense U. P., Odense, 1981).
[12] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).

Matrix methods for the numerical solution of relativistic wave equations

Loyal Durand and Alan Gara[a)]
Department of Physics, University of Wisconsin—Madison, Madison, Wisconsin 53706

(Received 16 November 1988; accepted for publication 9 May 1990)

An efficient new method is presented for the solution of eigenvalue problems that involve nonlocal operators of the type that appear in the solution of relativistic wave equations. The method, which has wider utility, allows very accurate results to be obtained with small matrix approximations to the eigenvalue equation. The method is illustrated for the equation $[2(-\nabla^2+m^2)^{1/2}+V(r)-M]\psi=0$.

I. INTRODUCTION

In this paper, we describe a matrix method for the solution of eigenvalue problems that involve nonlocal differential operators of the type $E=[-\nabla^2+m^2]^{1/2}$ typically encountered in the solution of relativistic wave equations. The method is far superior to conventional finite-difference methods, and is applicable to a wide range of problems. In the case of local operators, it is a version of the collocation method.

We developed this method during an investigation of relativistic quark–antiquark bound state problems using a reduced Salpeter equation.[1] Because of our emphasis in that work on the form of the quark–antiquark interaction, the problem was most clearly formulated in configuration space where the interaction can be described in terms of local potentials and nonlocal operators that involve E, E^{-1},$(E+m)^{-1}$, and ordinary differential operators. For example, the reduced Salpeter equation with a Lorentz scalar interaction between quarks gives a spin-averaged radial wave equation,[1]

$$
\begin{aligned}
&[M-2E_\ell]R_\ell(r)\\
&\quad=\frac{1}{4E_\ell^2}\Bigg[(E_\ell+m)^2V(r)+2(E_\ell+m)\\
&\qquad\times\left(\frac{dV(r)}{dr}\frac{d}{dr}+V(r)\nabla_\ell^2\right)\frac{1}{E_\ell+m}\\
&\qquad+\left(\frac{d^2V(r)}{dr^2}\frac{d^2}{dr^2}+\frac{dV(r)}{dr}\left(\nabla_\ell^2\frac{d}{dr}+\frac{d}{dr}\nabla_\ell^2\right)\right.\\
&\qquad\left.+V(r)\nabla_\ell^2\nabla_\ell^2\right)\frac{1}{(E_\ell+m)^2}\Bigg]R_\ell(r),
\end{aligned}
\tag{1}
$$

where

$$
\nabla_\ell^2=\frac{1}{r}\frac{d^2}{dr^2}r-\frac{\ell(\ell+1)}{r^2},
\tag{2}
$$

and

$$
E_\ell=[-\nabla_\ell^2+m^2]^{1/2}.
\tag{3}
$$

The methods developed here allowed us to reduce the solution of this and more general equations to small matrix problems. Thus, in the work reported in Ref. 1, we were able to determine the five lowest eigenvalues M of Eq. (1) to an accuracy of $\sim 10^{-4}$ for singular $q\bar{q}$ potentials of the general form

$$
V(r)=-\alpha/r+Br,
\tag{4}
$$

using 25×25 matrices. We will illustrate the methods below using a simpler relativistic wave equation that has been used in a number of analyses of $q\bar{q}$ bound states,[2-10]

$$
[2E_\ell+V(r)-M]R_\ell(r)=0.
\tag{5}
$$

Matrix methods for solving this simplified equation efficiently were developed in earlier work.[8,11] The present methods are a considerable improvement, as we show later. However, their full advantage is only evident when one considers more singular wave equations such as that in Eq. (1).

The remainder of the paper is as follows. In Secs. II A and II B we describe two versions of our matrix method for treating the nonlocal operator E_ℓ. The methods are very efficient, as we illustrate in Sec. II C with tests of the accuracy and rate convergence with matrix size, but remain incomplete in the sense that we have not established rigorous bounds on the error. We discuss the relations of our approach to other methods briefly in Sec. II D. Finally, in an Appendix, we apply our method to an exact integral formulation of the operator E_ℓ. The results, while easier to analyze theoretically, are unfortunately cumbersome even for fairly small values of ℓ, and are not especially useful for the solution of Eq. (1).

II. MATRIX APPROXIMATIONS FOR E_ℓ

A. The square root matrix $\mathbf{E}_\ell$

Our approach to the development of matrix representations for the operator E_ℓ is motivated by the observation that the bound-state wave functions in a confining potential are spatially compact. It is therefore reasonable to suppose that they can be approximated by a finite expansion in functions orthogonal on the interval $0\leqslant r<\infty$. We choose, then, to construct a matrix $\mathbf{E}_\ell$ that represents the action of E_ℓ on the basis set as accurately as possible at a selected set of points x_i in the (finite) interval that is actually relevant for the problem. (The method developed in the Appendix is exact in this respect; the methods developed below are approximate, but much simpler.)

Our construction of $\mathbf{E}_\ell$ is closely related to the method of orthogonal collocation.[12] We select a finite basis of N orthogonal functions $\{\mathcal{L}_j(r)=p_j(r)\sqrt{\mu(r)},$

[a)] Present address: Fermilab—E790, P. O. Box 500, Batavia, IL 60510.

 0022-2488/90/092237-07$03.00

$j = 0,1...,N-1\}$, $p_j(r)$ a jth-order polynominal, and define a matrix $\mathbf{L}$ as the matrix of basis functions evaluated at a set of N points $\{r_i, i = 1,2,...,N\}$,

$$\mathbf{L}_{ij} = \mathscr{L}_j(r_i). \tag{6}$$

The most appropriate choice of basis functions depends on the problem to be solved. To optimize the convergence of the method, we will choose the r_i as the zeros of $\mathscr{L}_N(r)$, the choice appropriate for Gaussian integration with the weight $\mu(r)$. The reason for this choice will be evident later.

In our first method, we construct an $N \times N$ matrix $\mathbf{A}_\ell$ that gives the *exact* action of the operator $D_\ell = E_\ell^2 = -\nabla_\ell^2 + m^2$ on the basis functions at the points r_i,

$$(\mathbf{A}_\ell)_{ij} = [(-\nabla_\ell^2 + m^2)\mathscr{L}_j](r_i). \tag{7}$$

This action can be determined for the bases we considered using the recurrence relations for the functions $\mathscr{L}_j(r)$, but can also be determined using the Lagrange differentiation formulas.[13] The matrix representation for $\mathbf{D}$ is then given by

$$\mathbf{D}_\ell = \mathbf{A}_\ell \mathbf{L}^{-1}. \tag{8}$$

The eigenvalues of $\mathbf{D}$ are determined by the relation

$$\mathbf{D}_\ell \mathbf{U} = \mathbf{U}\boldsymbol{\lambda}_\ell. \tag{9}$$

Here, $\mathbf{U}$ is the matrix of eigenvectors, $\mathbf{U}_{ij} = U_i^{(j)}$ with $U^{(j)}$ the jth eigenvector, and $\boldsymbol{\lambda}_\ell$ is the diagonal matrix of eigenvalues, $(\boldsymbol{\lambda}_\ell)_{ij} = \delta_{ij}\lambda_j$. Thus,

$$\mathbf{D}_\ell = \mathbf{U}\boldsymbol{\lambda}_\ell \mathbf{U}^{-1}. \tag{10}$$

The operator D_ℓ is the formal square of E_ℓ. We therefore *define* the matrix $\mathbf{E}_\ell$ as the square root of $\mathbf{D}_\ell$, the finite matrix approximation for D_ℓ,

$$\mathbf{E}_\ell = \mathbf{D}_\ell^{1/2} = \mathbf{U}\boldsymbol{\lambda}^{1/2}\mathbf{U}^{-1}. \tag{11}$$

With this definition, the relativistic wave equation in Eq. (5) reduces to a standard matrix eigenvalue problem,

$$[2\mathbf{E}_\ell + \mathbf{V}]\mathbf{R}_\ell = M\mathbf{R}_\ell, \tag{12}$$

where $\mathbf{V}$ is the potential matrix, $(\mathbf{V})_{ij} = \delta_{ij}V(r_i)$, and $\mathbf{R}_\ell$ is the column vector with components $(\mathbf{R}_\ell)_i = R_\ell(r_i)$. We will use this equation in Sec. II C to test the accuracy and rate of convergence of our approximation.

The extra nonlocal operators that appear in the more general relativistic wave equation in Eq. (1) can be defined in terms of $\mathbf{E}_\ell$ as

$$E_\ell^{-1} \to \mathbf{E}_\ell^{-1} = \mathbf{U}\boldsymbol{\lambda}^{-1/2}\mathbf{U}^{-1}, \tag{13}$$

$$(E_\ell + m)^{-1} \to (\mathbf{E}_\ell + m)^{-1} = \mathbf{U}(\boldsymbol{\lambda}^{1/2} + \mathbf{m})^{-1}\mathbf{U}^{-1}. \tag{14}$$

Operators such as ∇_ℓ^2 and d/dr can also be defined in terms of their action on the basis functions, e.g.,

$$\frac{d}{dr} \to \mathbf{d}, \quad (\mathbf{d})_{ij} = \sum_n \left(\frac{d}{dr}\mathscr{L}_n(r_i)\right)(\mathbf{L}^{-1})_{nj}, \tag{15}$$

and the equation reduced to a matrix eigenvalue problem.

Several remarks are in order here. First, the definition of $\mathbf{E}_\ell$ given above is not exact even at the points r_i. (The definition given in the Appendix *is* exact at the r_i, but involves considerably more numerical computation to implement.) We do not have a theoretical analysis of the possible errors in the action of $\mathbf{E}_\ell$. However, $\mathbf{E}_\ell^2 = \mathbf{D}_\ell$ exactly, and the errors in the action of $\mathbf{D}_\ell$ can be estimated from standard results on orthogonal collocation,[12] and are small for appropriate choices of the basis function and points r_i. Second, the matrices $\mathbf{E}^{-1}$ and $(\mathbf{E} + \mathbf{m})^{-1}$ defined above are not the same as the matrices that would be constructed, for example, by determining the exact action of E_ℓ^{-2} on the basis set, and taking the square root of the resulting matrix. "Exact" matrix operators for E_ℓ^{-1} and $(E_\ell + m)^{-1}$ can again be constructed using the methods of the Appendix, but there does not appear to be any practical advantage to that construction. The present definition of the inverse operators preserves such relations as $E_\ell E_\ell^{-1} = 1$ without error.

B. The symmetrical square root matrix $\hat{\mathbf{E}}_\ell$

We have implicitly assumed above that the eigenvalues of $\mathbf{D}_\ell$ are real and positive. This is generally expected to be the case since, $D_\ell = -\nabla_\ell^2 + m^2$ is a positive Hermitian operator. However, the matrix $\mathbf{D}$, while real, is not symmetric, so reality of the λ's is not guaranteed, and positivity may also be lost in the approximations. We therefore present a second method of constructing $\mathbf{E}_\ell$ that is free of this potential problem.

The method that we will use to construct a positive, symmetric matrix $\mathbf{D}_\ell$ has been discussed elsewhere.[14] We begin with a finite-basis Rayleigh–Ritz variational problem, and seek to minimize the matrix elements of D_ℓ subject to the condition that R_ℓ be normalized,

$$\begin{aligned}
&\delta \int_0^\infty dr\, r^2 R_\ell(r)(-\nabla_\ell^2 + m^2 - \lambda)R_\ell(r) \\
&\quad = \delta \int_0^\infty dr\, u_\ell(r)\left[-\frac{d^2}{dr^2} + \frac{\ell(\ell+1)}{r^2} + m^2 - \lambda\right]u_\ell(r) \\
&\quad = \delta \int_0^\infty dr\left[\left(\frac{du_\ell(r)}{dr}\right)^2 + \left(\frac{\ell(\ell+1)}{r^2} + m^2 - \lambda\right)u_\ell^2(r)\right] \\
&\quad = 0,
\end{aligned} \tag{16}$$

where $u_\ell(r) = rR_\ell(r)$ and λ is a Lagrange multiplier used to enforce the normalization condition. Minimization of the expression in Eq. (16) without restriction on the u's gives the equation for the continuum eigenfunctions of D_ℓ. Minimization on the functions spanned by the basis $\{\mathscr{L}_j(r), j = 0,1...,N-1\}$ gives the best approximation of D_ℓ in the mean on the set of functions presumed to give a good description of the solutions to the complete wave equation, e.g., Eq. (5).

We will suppose that the basis functions $\mathscr{L}_j(r)$ are of the form $p_j(r)\sqrt{\mu(r)}$ with the polynomials p_j orthogonal on $[0,\infty]$ with respect to the weight $\mu(r)$. Then writing $u_\ell(r)$ as $v_\ell(r)\sqrt{\mu(r)}$ and extracting a factor of μ, we can rewrite Eq. (16) as

$$\begin{aligned}
\delta \int_0^\infty dr\,\mu(r)\Bigg[&\left(\frac{dv_\ell(r)}{dr} + \frac{1}{2\mu(r)}\frac{d\mu(r)}{dr}v_\ell(r)\right)^2 \\
&+ (\ell(\ell+1)/r^2 + m^2 - \lambda)v_\ell^2(r)\Bigg] = 0.
\end{aligned} \tag{17}$$

The integral can now be approximated using generalized Gaussian integration with respect to the weight $\mu(r)$,[15]

$$\int_0^\infty dr\,\mu(r)f(r) = \sum_{i=1}^{N} w_i f(r_i), \tag{18}$$

where the points r_i are the zeros of $p_N(r)$ or $\mathscr{L}_N(r)$, and the weights w_i can be determined by standard methods.[15] The result is exact for $f(r)$ a polynomial of order $2N-1$ or less. With this approximation, we obtain a discrete variational problem,

$$\delta \sum_{i=1}^{N} w_i \left[\left(\frac{dv_\ell(r_i)}{dr} + \frac{1}{2\mu(r_i)} \frac{d\mu(r_i)}{dr} v_\ell(r_i) \right)^2 + \left(\frac{\ell(\ell+1)}{r_i^2} + m^2 - \lambda \right) v_\ell^2(r_i) \right] = 0, \tag{19}$$

or, in matrix form,

$$\delta \left[\mathbf{v}_\ell^T \mathbf{A}_\ell^T \mathbf{w} \mathbf{A}_\ell \mathbf{v}_\ell + \mathbf{v}_\ell^T \left(\frac{\ell(\ell+1)}{\mathbf{r}^2} + \mathbf{m}^2 - \lambda \mathbf{1} \right) \mathbf{w} \mathbf{v} \right] = 0. \tag{20}$$

Here $\mathbf{v}$ is a column vector, $\mathbf{w}$ is the diagonal matrix of integration weights, and $\mathbf{A}$ is the matrix

$$\mathbf{A}_{ij} = \sum_{n=0}^{N-1} \left(\frac{dp_n(r_i)}{dr} + \frac{1}{2\mu(r_i)} \frac{d\mu(r_i)}{dr} p_n(r_i) \right) (\mathbf{p}^{-1})_{nj} = \left(\frac{1}{\sqrt{\mu}} \mathbf{d} \sqrt{\mu} \right)_{ij}, \tag{21}$$

where $\mathbf{d}$ is the matrix defined in Eq. (15).

It is convenient to define a new column vector $\hat{\mathbf{v}}$ and a matrix $\hat{\mathbf{d}}$ by

$$\hat{\mathbf{v}} = \mathbf{w}^{1/2}\mathbf{v}, \quad \hat{\mathbf{d}} = (\mathbf{w}/\mu)^{1/2}\mathbf{d}(\mu/\mathbf{w})^{1/2}. \tag{22}$$

With these definitions, the variational problem reduces to

$$\delta[\hat{\mathbf{v}}^T \hat{\mathbf{D}} \hat{\mathbf{v}} - \lambda \hat{\mathbf{v}}^T \hat{\mathbf{v}}] = 0, \tag{23}$$

where

$$\hat{\mathbf{D}}_\ell = \hat{\mathbf{d}}^T \hat{\mathbf{d}} + \ell(\ell+1)/\mathbf{r}^2 + \mathbf{m}^2. \tag{24}$$

Varying with respect to the components of $\hat{\mathbf{v}}^T$, we obtain the matrix eigenvalue problem

$$\hat{\mathbf{D}}_\ell \hat{\mathbf{v}} = \lambda \hat{\mathbf{v}}, \quad \hat{\mathbf{D}}_\ell \hat{\mathbf{V}} = \hat{\mathbf{V}} \boldsymbol{\lambda}, \tag{25}$$

where $\hat{\mathbf{V}}$ and $\boldsymbol{\lambda}$ are the matrices of eigenvectors and eigenvalues.

The matrix $\hat{\mathbf{d}}^T\hat{\mathbf{d}} = \hat{\mathbf{d}}^\dagger\hat{\mathbf{d}}$ is real, symmetric, and positive, so $\hat{\mathbf{D}}_\ell$ is also, and the eigenvalues λ are guaranteed to be real and positive. The eigenvectors $\hat{\mathbf{v}}$ can be chosen real. With this convention, $\hat{\mathbf{V}}$ is a real orthogonal matrix, and $\hat{\mathbf{D}}_\ell$ can be written in the symmetric form

$$\hat{\mathbf{D}}_\ell = \hat{\mathbf{V}} \boldsymbol{\lambda}_\ell \hat{\mathbf{V}}^T. \tag{26}$$

The $\hat{\mathbf{u}}$'s are related to the original eigenvectors $\mathbf{u}$ with components $\mathbf{u}_i = u(r_i)$ by $\mathbf{u} = (\mu/\mathbf{w})^{1/2}\hat{\mathbf{v}}$, hence

$$\mathbf{U} = (\mu/\mathbf{w})^{1/2}\hat{\mathbf{V}}, \quad \hat{\mathbf{V}} = (\mathbf{w}/\mu)^{1/2}\mathbf{U}. \tag{27}$$

The remainder of the construction of a matrix representation of E_ℓ follows that given in Sec. II A. We define the symmetrical square root operator $\mathbf{E}_\ell$ as the square root of $\hat{\mathbf{D}}_\ell$, transformed to the u basis $\{\mathscr{L}_j, j = 0,\ldots,N-1\}$,

$$\hat{\mathbf{E}}_\ell = \hat{\mathbf{U}} \boldsymbol{\lambda}^{1/2} \mathbf{U}^T, \tag{28}$$

where $\mathbf{U}$ and $\boldsymbol{\lambda}$ can be determined directly as the matrices of eigenvectors and eigenvalues associated with the equation

$$\left[\frac{\mu}{\mathbf{w}} \mathbf{d}^T \frac{\mathbf{w}}{\mu} \mathbf{d} + \frac{\ell(\ell+1)}{\mathbf{r}^2} + \mathbf{m}^2 \right] \mathbf{U} = \mathbf{U} \boldsymbol{\lambda}. \tag{29}$$

The relativistic wave equation in Eq. (5) can be formulated as the variational problem

$$\delta \int_0^\infty dr [2u_\ell(r)(E_\ell u_\ell)(r) + u_\ell(r)(V(r) - M)u_\ell(r)] = 0. \tag{30}$$

If we again use Gaussian integration with respect to the weight $\mu(r)$ to convert Eq. (30) to a discrete variational problem, it assumes the form

$$\delta \left[2\mathbf{u}_\ell^T \frac{\mathbf{w}}{\mu} \hat{\mathbf{E}}_\ell \mathbf{u}_\ell + \mathbf{u}_\ell^T (\mathbf{V} - M\mathbf{1}) \mathbf{u}_\ell \right] = 0 \tag{31}$$

or

$$[2\hat{\mathbf{E}}_\ell + \mathbf{V} - M\mathbf{1}]\mathbf{u}_\ell = 0, \quad [2\hat{\mathbf{E}}_\ell + \mathbf{V}]\mathbf{U} = \mathbf{U}\mathbf{M}. \tag{32}$$

Here, $\hat{\mathbf{E}}_\ell$ gives the best representation of E_ℓ in the mean on the basis states for this discrete problem, and $\mathbf{M}$ gives the variationally best set of eigenvalues. It was shown in Ref. 14 that the errors in the eigenvalues for a standard Sturm–Liouville problem, e.g., the problem above with $\hat{\mathbf{E}}_\ell$ replaced by $\hat{\mathbf{E}}_\ell^2$, decrease as

$$\delta M_n/M_n \sim 2(\pi N)^{1/2}(n\pi e/4N)^{2N} \times O(1), \quad n = 1,\ldots,N, \tag{33}$$

for an N-dimensional basis. We expect essentially the same error estimate to hold here. The extremely rapid convergence results from the optimization of the choice of points r_i in the Gaussian integration. The mean error in the nth eigenvector is expected to be of order $|\delta M_n/M_n|^{1/2}$.

C. Tests of square-root matrices

We have conducted a number of tests of the accuracy of the matrix operators $\mathbf{E}_\ell$, $\hat{\mathbf{E}}_\ell$, and the operator $\mathscr{E}_\ell$ constructed in the Appendix by applying the operators to simple functions. More realistic assessments of the accuracy and usefulness of these matrices can be obtained by applying the methods to the solution of a realistic problem, and checking the rate of convergence of the results with matrix size. We have used the equation

$$[2E_\ell - \alpha/r + Br - M]\phi_\ell(r) = 0, \tag{34}$$

for this purpose, using the parameters $m = 1.45$ GeV, $\alpha = 0.25$, and $B = 0.18$ GeV^2 used for the same purpose in Ref. 5. (These potential and mass parameters are characteristic of those encountered in the treatment of charmonium.[8])

In the radial form of the equation used in Sec. II A, $\phi_\ell = R_\ell$, the basis functions $\mathscr{L}_j(r)$ were chosen as associated Laguerre functions,[16]

$$\mathscr{L}_j(r) = r^\ell e^{-(1/2)cr} L_j^{(2\ell)}(cr), \tag{35}$$

with

$$\mu(r) = r^{2\ell} e^{-cr}, \tag{36}$$

a choice that builds in the correct behavior of the wave functions for $r \to 0$.

Here c is a scale parameter that was chosen so that the range of the points r_i [the zeros of $L_N^{(2\ell)}(cr)$] covered the region in which the low eigenfunctions are large. We could use c as an extra variational parameter; we did not. The integration weights w_i were determined using standard methods.[15]

The results of the convergence tests using the matrix $\mathbf{E}_\ell$ of Sec. II A with c chosen so that $r_N \approx 4$ fm are shown in Table I for

TABLE I. Convergence of the low eigenvalues M_n (in GeV) of Eq. (34) with increasing size N of the basis set using the matrix representations $\mathbf{E}_\ell$ and $\hat{\mathbf{E}}_\ell$ for $E_\ell = (-\nabla_\ell^2 + m)^{1/2}$.

			$\ell = 0$		
$N \backslash n$	1	2	3	4	5
			$\mathbf{E}_\ell$		
10	3.392 375	3.902 142	4.292 493	4.626 686	4.908 474
15	3.392 383	3.902 150	4.292 488	4.625 953	4.923 683
20	3.392 385	3.902 152	4.292 489	4.625 955	4.923 694
25	3.392 386	3.902 153	4.292 490	4.625 956	4.923 695
30	3.392 386	3.902 153	4.292 491	4.625 957	4.923 695
			$\hat{\mathbf{E}}_\ell$		
10	3.392 374	3.902 138	4.292 428	4.6093	4.685
25	3.392 386	3.902 153	4.292 490	4.625 956	4.923 695
			$\ell = 1$		
$N \backslash n$	1	2	3	4	5
			$\mathbf{E}_\ell$		
10	3.734 535	4.148 087	4.495 910	...	...
15	3.734 535	4.148 082	4.496 254	4.804 522	5.085 011
20	3.734 535	4.148 082	4.496 254	4.804 522	5.084 965
25	3.734 535	4.148 082	4.496 254	4.804 522	5.084 965
30	3.734 535	4.148 082	4.496 254	4.804 522	5.804 965
			$\hat{\mathbf{E}}_\ell$		
10	3.734 535	4.148 075	4.492 772	4.730	4.83
25	3.734 535	4.148 082	4.496 254	4.804 522	5.084 965

$\ell = 0,1$. The convergence is extremely rapid. A 15×15 matrix gives results which are sufficiently accurate for practical purposes for all the states shown. However, for $N = 10$, the method fails as low-lying complex-conjugate eigenvalues appear in the spectrum. This problem can be eliminated by using the symmetrical matrix $\hat{\mathbf{E}}_\ell$ developed in Sec. II B, as shown in the extra lines in Table I. In the latter calculations, we used the u representation, $\phi_\ell = u_\ell(r) = rR_\ell(r)$ in Eq. (34), and the corresponding basis functions

$$\mathscr{L}_j(r) = r^{\ell+1} e^{-(1/2)cr} L_j^{(2\ell+2)}(cr), \tag{37}$$

with

$$\mu(r) = r^{2\ell+2} e^{-cr}. \tag{38}$$

The symmetrical square root or $\hat{\mathbf{E}}_\ell$ method based on Gaussian integration is somewhat less accurate than the unsymmetrical $\mathbf{E}_\ell$ method for small matrix sizes, assuming that the latter method works. The difference for small N is apparently in the approximate nature of the Gaussian integration, since the differential operators are treated equivalently. The two methods are completely equivalent for practical purposes for $N \gtrsim 20$.

To further illustrate the advantages of the present methods, we show in Table II the results obtained in Refs. 8 and 11 using conventional finite-difference methods to construct the (much larger) square root matrices $\mathbf{E}_\ell$. The savings in computer time are substantial, especially for more complicated and more singular problems such as those in Eq. (1) and Refs. 1. In the latter work on solution of the complete, spin-dependent reduced Salpeter equation for $b\bar{b}$, $c\bar{c}$, and $s\bar{s}$ quark–antiquark bound states,

TABLE II. Convergence of the eigenvalues of Eq. (34) with increasing matrix size N using the square root method of Refs. 8 and 11. The numbers are from Ref. 8. M_n is given in GeV. The lines labeled $\mathbf{E}_\ell$ are from Table I with $N = 25$.

			$\ell = 0$		
$N \backslash n$	1	2	3	4	5
$\mathbf{E}_0$	3.392 386	3.902 153	4.292 490	4.625 956	4.923 695
25	3.3924	3.9022	4.2925	4.6252	4.9205
33	3.3924	3.9022	4.2925	4.6258	4.9228
49	3.3924	3.9022	4.2925	4.6260	4.9236
77	3.3924	3.9022	4.2925	4.6260	4.9237
			$\ell = 1$		
$N \backslash n$	1	2	3	4	5
E_1	3.734 535	4.148 082	4.496 254	4.804 533	5.084 965
25	3.7345	4.1481	4.4960	4.8033	5.0793
33	3.7345	4.1481	4.4962	4.8042	5.0838
49	3.7345	4.1481	4.4963	4.8045	5.0848
77	3.7345	4.1481	4.4963	4.8045	5.0850

we found the 25×25 matrix representation of $\mathbf{E}_\ell$ to be quite satisfactory: it is accurate, and the matrices can be manipulated quickly enough that it was possible to use conventional nonlinear regression methods to fit the spin-dependent quark–antiquark potential. Furthermore, there was no particular advantage at this matrix size to using the symmetrical matrix $\hat{\mathbf{E}}_\ell$, as the operators that appear are not easily symmetrized, see, e.g., Eq. (1).

The final method developed in the Appendix gives a matrix $\mathscr{E}_\ell$ that reproduces the *exact* action of E_ℓ on the basis functions at the points r_i. However, this method requires the accurate evaluation of N^2 integrals, and the advantage of computational speed characteristic of the square root methods is lost. Moreover, the integrands are singular and must be treated with great care, e.g., by extracting the singular pieces and treating them exactly, if one is to obtain results with accuracy comparable to that shown in Tables I and II. If this is done, the results obtained using $\mathscr{E}_\ell$ are essentially the same for the test problem above as the results obtained with $\mathbf{E}_\ell$ or $\hat{\mathbf{E}}_\ell$. The $\mathscr{E}_\ell$ method does provide a useful way of checking the accuracy of the square root matrices in reproducing the action of the operator E_ℓ on simple functions.

In Table III we compare the $\mathscr{E}_\ell$ method for $\ell=0$ and $N=15$ with a finite-difference approximation for the same integral with $N=100$ from Refs. 3 and 5. The singular integrals in $\mathscr{E}_\ell$ were evaluated in this calculation using a standard adaptive integration routine, with results of limited accuracy. The singularities were treated exactly in the calculations in Ref. 5. The improvement in accuracy with the present methods is obvious; the results from Ref. 5 are still inaccurate in the third decimal place even for the very large matrix used.

D. Discussion

The general advantages of the method of solution of relativistic wave equations presented here are the fact that it allows one to work in position space where interaction potentials are easily understood and easily varied, its very rapid (exponential or faster) convergence with increasing matrix size, and its simplicity. As shown in the Tables, the "spinless Salpeter equation"

$$[2E_\ell + V(r) - M]\phi_\ell(r) = 0\,, \tag{39}$$

can be solved to high accuracy for potentials $V(r)$ of the type which appear in analyses of relativistic quark–antiquark bound states using quite small matrix approximations for E_ℓ. Thus, 10×10 matrices already give results for the "Coulomb-plus-linear" potential in Eq. (34) that are accurate to better than 1/5000 for all the observed S and P states in the $c\bar{c}$ and $b\bar{b}$ systems, while the results for 15×15 matrices are accurate to about $1/10^6$. The very singular effective interactions encountered in the solution of the spin-dependent reduced Salpeter equation,[1] or its spin-independent part given in Eq. (1), require somewhat larger matrices (25×25 matrices are more than adequate,[1,10] to obtain accuracies $\sim1/10^4$), but no change in procedure: operators such as E_ℓ^{-1} or $(E_\ell+m)^{-1}$ are simply represented as the matrix inverses of $\mathbf{E}_\ell$ or $(\mathbf{E}_\ell+\mathbf{m})$, and differential operators are defined by their exact action on the basis functions.

TABLE III. Comparison of solutions of Eq. (34) for $\ell=0$ obtained using the finite-difference approximation of Ref. 11 to the integral operator E_ℓ in Eq. (A4), with the solutions obtained using the collocation operator $\mathscr{E}_\ell$ developed in the Appendix, and the square root operator $\mathbf{E}_\ell$ of Sec. II A.

n	Ref.5, $N=100$	$\mathscr{E}_\ell$, $N=15$	"Exact", $\mathbf{E}_\ell$, $N=30$
1	3.3925	3.392 35	3.392 386
2	3.9023	3.902 13	3.902 153
3	4.2928	4.2923	4.292 498
4	4.6263	4.6260	4.625 957
5	4.9240	4.9238	4.923 695

Other methods are of course available for the solution of relativistic wave equations, for example, treatment of the equation as an integral equation in momentum space with solution by finite difference methods. The most powerful (and popular) alternative is the Rayleigh–Ritz–Galerkin method[12] in which the wave functions $\phi_\ell(r)$ are expanded in terms of a finite basis, and the wave equation is reduced to a matrix equation for the expansion coefficients. A complication of the method is the need to determine the matrix elements of the operator $[2E_\ell + V - M]$ in the chosen basis. Various choices for the basis set have been used in the literature on quarkonium. For example, Gupta, Radford, and Repko[4] used a basis consisting of functions of the form $x^L e^{-x}$ that give simple integrals for the matrix elements of V, and then reduced the calculation of the matrix elements of E_ℓ to the numerical evaluation of single integrals involving trigonometric functions, one integral for each choice of the initial and final basis states. Jacobs, Olsson, and Suchyta[6] used a basis of Laguerre functions in position space, and their Fourier transforms—a set of Jacobi polynomials with a somewhat complicated argument—in momentum space, and evaluated the matrix elements of V and E_ℓ in position and momentum space, respectively, by using generalized Gauss–Laguerre and Gauss–Jacobi integration. Stanley and Robson[3] and Godfrey and Isgur[5] used radial harmonic oscillator bases that have simple properties under Fourier transform, but were probably a less appropriate choice otherwise. In all cases, there were extra numerical integrations relative to the $\mathbf{E}_\ell$ and $\hat{\mathbf{E}}_\ell$ methods presented here. Jacobs, Olsson, and Suchyta[6] also used the Rayleigh–Ritz–Galerkin method to solve the very singular spin-dependent equations for the $q\bar{q}$ bound states obtained from the reduced Salpeter equation, but found that the calculation of the extra matrix elements which appear in that case was not completely straightforward.[9] In contrast, we encountered no particular difficulties in the treatment of this problem, or its spin-averaged version given in Eq. (1), using either the $\mathbf{E}_\ell$ or the $\hat{\mathbf{E}}_\ell$ method. Other authors have modified the extra interactions relative to Eq. (39) to make them less singular, or have treated them as perturbations, thus avoiding the potential problems.

There are clearly several methods that can be used effectively to solve relativistic wave equations. We recommend the methods presented here for their simplicity, accuracy, and flexibility.

ACKNOWLEDGMENTS

This work was supported by the U. S. Department of Energy under Contract No. DE-AC02-76ER00881, and by the University of Wisconsin Graduate School with funds granted by the Wisconsin Alumni Research Foundation.

APPENDIX

The action of the nonlocal operator $E = [-\nabla^2 + m^2]^{1/2}$ is defined in terms of its Fourier transform. If $F(r)$ is a well behaved, but otherwise arbitrary test function,

$$[-\nabla^2+m^2]^{1/2}F(r)Y_\ell^m(\hat{r})$$
$$\equiv\frac{1}{(2\pi)^3}\int\int d^3r'\,d^3p[p^2+m^2]^{1/2}e^{i\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')}F(r')Y_\ell^m(\hat{r}'). \tag{A1}$$

This expression can be rewritten in the form[11]

$$[-\nabla^2+m^2]^{1/2}F(r)\,Y_{\ell m}(\hat{r})$$
$$=\frac{1}{(2\pi)^3}\int\int d^3r'\,d^3p\frac{1}{[p^2+m^2]^{1/2}}F(r')Y_\ell^m(\hat{r}')(-\nabla_{r'}^2+m^2)e^{i\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')}$$
$$=\frac{1}{(2\pi)^3}\int\int d^3r'\,d^3p\frac{1}{[p^2+m^2]^{1/2}}e^{i\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')}(-\nabla_{r'}^2+m^2)F(r')Y_\ell^m(\hat{r}')$$
$$=\frac{1}{(2\pi)^3}\int\int d^3r'\,d^3p\frac{1}{[p^2+m^2]^{1/2}}e^{i\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')}Y_\ell^m(\hat{r}')(-\nabla_\ell^2+m^2)F(r'), \tag{A2}$$

where ∇_ℓ^2 is defined in Eq. (12). The exponential in Eq. (A2) can be expanded in terms of spherical harmonics and spherical Bessel functions and the angular integrations performed by using the relation

$$e^{i\mathbf{p}\cdot\mathbf{r}}=4\pi\sum_{\ell,m}i^\ell j_\ell(pr)Y_\ell^m(\hat{p})Y_\ell^{m*}(\hat{r}), \tag{A3}$$

and the orthogonality of the spherical harmonics, with the result

$$[-\nabla^2+m^2]^{1/2}F(r)\,Y_{\ell m}(\hat{r})$$
$$=Y_{\ell m}(\hat{r})\,[-\nabla_\ell^2+m^2]^{1/2}F(r)$$
$$=Y_\ell(\hat{r})\frac{2}{\pi}\int_0^\infty dr'\,r'^2I_\ell(r,r')(-\nabla_\ell'^2+m^2)F(r'), \tag{A4}$$

where

$$I_\ell(r,r')=\int_0^\infty dp\,p^2\frac{j_\ell(pr)j_\ell(pr')}{[p^2+m^2]^{1/2}}. \tag{A5}$$

The integral $I_\ell(r,r')$ can be evaluated exactly in terms of hyperbolic Bessel functions $K_n(z)$,[11] with a result which increases rapidly in complexity with increasing ℓ,

$$I_\ell(x,x')=\frac{1}{2rr'}\cdot 2^\ell z'^{\ell+1}\left[\frac{1}{z}\frac{\partial}{\partial z}\right]^\ell$$
$$\times\frac{1}{z}\left[(y-z)^{\ell/2}K_\ell((y-z)^{1/2})\right.$$
$$\left.-(y+z)^{\ell/2}K_\ell((y+z)^{1/2})\right]. \tag{A6}$$

evaluated for

$$y=m^2(r^2+r'^2),\quad z=2m^2rr'. \tag{A7}$$

Using the ideas presented earlier, we can develop a matrix representation for $E_\ell=[-\nabla_\ell^2+m^2]^{1/2}$ by using each of our basis functions as the test function F, and evaluating the integral for each value r_i. Let A_{ij} be the matrix that gives the exact action of E_ℓ on $\mathcal{L}_j$ evaluated at r_i,

$$A_{ij}=\int_0^\infty dr'\,r'^2I_\ell(r_i,r')(-\nabla_\ell'^2+m^2)\mathcal{L}_j(r'). \tag{A8}$$

Then

$$\mathcal{E}_\ell=\mathbf{A}\mathbf{L}^{-1}. \tag{A9}$$

This is essentially a collocation method for approximating the action of $E_\ell=(-\nabla_\ell^2+m^2)^{1/2}$. The errors in the approximation are not known, but should be similar to the errors in ordinary collocation. The integrals can be evaluated to the accuracy necessary for consistency.

The result in Eq. (A9) is easily extended to the full relativistic equation

$$[2E_\ell+V-M]R_\ell(r)=0\,, \tag{A10}$$

considered in the text by introducing the diagonal matrix $\mathbf{V}$ with elements $\mathbf{V}_{ij}=V(r_i)\delta_{ij}$. The result is the matrix eigenvalue problem

$$[2\mathcal{E}_\ell+\mathbf{V}]\mathbf{R}_\ell=M\mathbf{R}_\ell, \tag{A11}$$

where $\mathbf{R}_\ell$ is the column vector with components $R_\ell(r_i)$. The Laguerre functions in Eq. (35) provide an appropriate basis set for potentials of the type of interest for quark–antiquark systems. With this basis and a reasonable choice of the scale factor c in Eq. (35), we have found this approach to be much more accurate for a given matrix size than the finite-difference methods used in Ref. 11 to treat Eq. (A10), and to converge much more rapidly with increasing matrix size, as shown in Table III. The results on accuracy and convergence of this method are similar to those discussed for the square root matrices $\mathbf{E}_\ell$ and $\hat{\mathbf{E}}_\ell$ in Sec. II C. The usefulness of the method is hampered by the necessity of evaluating N^2 integrals accurately for a basis of size N, and by the complexity of the kernels $I_\ell(r,r')$. However, the method provides a useful check on the simpler heuristically motivated methods considered in the text.

[1] A. Gara, B. Durand, L. Durand, and L. J. Nickisch, Phys. Rev. D **40**, 843 (1989); A. Gara, B. Durand, and L. Durand, University of Wisconsin-Madison report MAD/TH/90-4.

[2] A. B. Henriques, B. H. Kellett, and R. G. Moorhouse, Phys. Lett. **64** B, 85 (1976); P. Ditsas, N. A. McDougall, and R. G. Moorhouse, Nucl. Phys. B **146**, 191 (1978).

[3] D. P. Stanley and D. Robson, Phys. Rev. D **21**, 3180 (1980); C. Long and D. Robson, *ibid.* **27**, 644 (1983).

[4] S. N. Gupta, S. F. Radford, and W. W. Repko, Phys. Rev. D **31**, 160 (1985); **34**, 201 (1986); S. N. Gupta, W. W. Repko, and C. J. Suchyta, III, *ibid.* **39**, 974 (1989).

[5] S. Godfrey and N. Isgur, Phys. Rev. D **32**, 189 (1985).

[6] S. Jacobs, M. G. Olsson, and C. Suchyta, III, Phys. Rev. D **33**, 3338 (1986), **34**, 3536 (*E*) (1986); **35**, 2448 (1987).

[7] K. Igi and S. Ono, Phys. Rev. D **33**, 3349 (1986).
[8] L. J. Nickisch, Ph.D. thesis, University of Wisconsin—Madison, 1984 (unpublished).
[9] S. Jacobs, Ph.D. thesis, University of Wisconsin—Madison, 1986.
[10] A. Gara, Ph.D. thesis, University of Wisconsin—Madison, June 1987 (unpublished).
[11] L. J. Nickisch, L. Durand, and B. Durand, Phys. Rev. D **30**, 660 (1984).
[12] See, for example, P. M. Prenter, *Splines and Variational Methods* (Wiley, New York, 1975), Chap. 8.
[13] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965), Sec. 25.3.
[14] L. Durand, in *Polynômes Orthogonaux et Applications*, edited by C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux (Springer, Berlin, 1985), pp. 331–339.
[15] A. Ghizzetti and A. Ossicini, *Quadrature Formulae* (Birkhauser, Basel, 1970); P. J. Davis and P. Rabinowitz, *Numerical Integration* (Blaisdell, New York, 1967).
[16] Reference 13, Chap. 22.

Collision theory for massless particles in the framework of a Wightman field theory

Dieter Strube
Institut für Theoretische Physik, Universität Göttingen, Bunsenstrasse 9, 3400 Göttingen, Federal Republic of Germany

(Received 30 January 1990; accepted for publication 2 May 1990)

A collision theory of massless Fermions and Bosons is constructed within the framework of a Wightman field theory. To this end, using essentially the temperedness of the Wightman distributions, the main ideas of the collision theory of Buchholz formulated in a field theory of bounded operators are carried over.

I. INTRODUCTION

More than 10 years ago, Buchholz developed his famous collision theory for massless Fermions and Bosons starting from a field algebra of bounded operators.[1-3]

In this paper, we show that a collision theory for massless particles in four space-time dimensions also exists within a Wightman field theory.[4]

Using essentially the temperedness of the Wightman distributions, we succeed in adopting the main ideas from the collision theory of Buchholz in order to construct the collision states also in our case of a Wightman field theory.

In contrast to Buchholz,[1,3] we construct the asymptotic fields of massless Fermions and Bosons in an analogous way. The main effort in our investigation consists in carrying over Lemma 2 in Ref. 3 to the present case.

This lemma, which we also formulate for Fermions, plays a central part in the collision theory for massless Bosons and Fermions. It says that suitable spherical means of vacuum expectation values of local operators have clustering properties.

Since we are dealing with a field algebra of unbounded operators, the present construction is burdened with technicalities.

II. THE ASYMPTOTIC FIELDS

In a Wightman field theory, it turns out to be advantageous to use for Fermions the same construction of asymptotic fields as for Bosons.

As Buchholz,[3] we define the asymptotic fields as strong limits of certain sequences of operators.

Let us first list our assumptions and notations. We consider a Wightman field theory without a mass gap[4] given by a finite set of local fields $\{\phi_i(x)\}$. By $\mathscr{P}_{SL}$ we denote the polynomial algebra spanned by the $\phi_i(f)$ with $f\in\mathscr{D}(\mathbf{R}^4)$ (test functions on $\mathbf{R}^4$ with compact support), by Ω the vacuum state, and by $\mathscr{P}_{SL}(O)$ the subalgebra of $\mathscr{P}_{SL}$ with $\text{supp}\, f\subset O\subset\mathbf{R}^4$.

The elements in $\mathscr{P}_{SL}$ are called strictly local operators and those in $\mathscr{P}_{SL}\Omega$ strictly local states.

In addition to the usual assumptions of a Wightman framework we suppose that there exists a continuous unitary representation $L\to U(L)$ of the covering group of the Poincaré group $\mathscr{P}$ in a Hilbert space $\mathscr{H}$ in which there is a subspace $\mathscr{H}_1\subset\mathscr{H}$, the space of massless one-particle states, on which the $U(L)$, $L\in\mathscr{P}$ act like a representation of the Poincaré group $\mathscr{P}$ with mass $m=0$ (by a paper of Yngvason,[5] there are no unphysical representations with $m=0$ and continuous spin in a Wightman field theory). To begin with, the construction of asymptotic fields let A_+ and $A_-\in\mathscr{P}_{SL}$ be a strictly local Bose and Fermi operator and define

$$A_+(x)=U(x)A_+U(x)^{-1}$$

and

$$A_-(x)=U(x)A_-U(x)^{-1},$$

where $(x_0,\vec{x})=x\to U(x)$ is the unitary representation of the translations.

For each $t\in\mathbf{R}$, we define furthermore a spherical mean of A_+ and A_-,

$$A_{+t}=-2t\int d\omega\,\partial_0 A_+(t,t\vec{e}),$$

$$A_{-t}=-2t\int d\omega\,\partial_0 A_-(t,t\vec{e}).$$

Here, $d\omega=d\omega(\vec{e})$ is the normalized measure on the unit sphere S^2 in $\mathbf{R}^3$, $\vec{e}$ a unit vector that runs over the sphere, and ∂_0 denotes differentiation with respect to the time component of the translations.

Let us finally define sequences of functions:

$$h_T(t):=\frac{1}{\ln|T|}\,h\left(\frac{1}{\ln|T|}(t-T)\right),\quad |T|>1,$$

where $h\in\mathscr{D}(\mathbf{R})$ is real and normalized according to $\int dt\,h(t)=1$. Thus we have also $\int dt\,h_T(t)=1$ and h_T has support in an interval around T of a length proportional to $\ln|T|$ (instead of the logarithm, one could use any other slowly increasing function).

Consider now, for $|T|>1$,

$$A_{+T}:=\int dt\,h_T(t)A_{+t}\quad\text{and}\quad A_{-T}=\int dt\,h_T(t)A_{-t}.$$

By an explicit calculation in Ref. 1, one then establishes the existence of the strong limits

$$s-\lim_{T\to\pm\infty}A_{+T}\Omega=P_1A_+\Omega$$

and

 0022-2488/90/092244-09$03.00

$$s-\lim_{T\to\pm\infty} A_{-T}\Omega = P_1 A_-\Omega,$$

where P_1 is the projection onto the space $\mathscr{H}_1$ of massless one-particle states.

In the next step, we want to prove that A_{+T} and A_{-T} converge not only on the vacuum but also on a dense set of vectors. To this end, we need some notations introduced by Buchholz[3] and a lemma that also will play a central part in the further construction of asymptotic many particle states. It is suitable to distinguish a family of subsets $\mathscr{P}_N$, $N\in\mathbf{N}$ in $\mathscr{P}_{\mathrm{SL}}$. The elements of $\mathscr{P}_N$ are all finite sums of operators of the form $\int dt\,\varphi(t)A(tn)$, $A\in\mathscr{P}_{\mathrm{SL}}$. Here, A is a strictly local Bose or Fermi operator, n is a positive timelike four-vector, and $\varphi(t)\in\mathscr{D}(\mathbf{R})$ has a Fourier transform $\tilde{\varphi}(\omega)$ with an N-fold zero at $\omega=0$. We formulate now the above mentioned lemma that we need in contrast to Buchholz[1,3] also in the Fermi case.

Lemma 2.1: (i) Let $A_1,\ldots,A_n\in\mathscr{P}_N$ be n Bose operators and N sufficiently large (depending on n). Then,

$$\lim_{T\to\pm\infty}(\Omega,A_{1T}\cdots A_{nT}\Omega) = \sum(\Omega,A_{i_1}P_1A_{i_2}\Omega)\cdots(\Omega,A_{i_{n-1}}P_1A_{i_n}\Omega),$$

if n is even. The sum extends over all distinct partitions of $(1,\ldots,n)$ into ordered pairs. For odd n, the limit vanishes.

(ii) Let $\psi,\ldots,\psi_n\in\mathscr{P}_N$ be n Fermi operators and N sufficiently large (depending on n). Then,

$$\lim_{T\to\pm\infty}(\Omega,\psi_{1T}\cdots\psi_{nT}\Omega) = \sum\sigma_P(\Omega,\psi_{i_1}P_1\psi_{i_2}\Omega)\cdots(\Omega,\psi_{i_{n-1}}P_1\psi_{i_n}\Omega),$$

if n is even. Here, the sum is given as in (i) and $\sigma_P=\pm1$, if the permutation $P=(i_1,i_2,\ldots,i_n)$ of $(1,\ldots,n)$ is even or odd.

Before proving this lemma, we establish the strong convergence of A_{+T} and A_{-T} on a dense set of vectors. To specify this set, we need a geometrical notation. We call with Buchholz[1] the positive cone O_+ of all points that have a positive timelike separation from a bounded region $O\subset\mathbf{R}^4$ the future tangent of O (the past tangent that we use later is defined analogously).

Lemma 2.2: Let A_+ be an element of $\mathscr{P}_{\mathrm{SL}}(O)\cap\mathscr{P}_{N_0}$, N_0 sufficiently large where O is some bounded region $O\subset\mathbf{R}^4$. Then, the strong limit

$$\begin{aligned}A_+^{\mathrm{out}}F_+\Omega &:= s-\lim_{T\to\infty}A_{+T}F_+\Omega\\ &= s-\lim_{T\to\infty}F_+A_{+T}\Omega\\ &= F_+P_1A_+\Omega\end{aligned}$$

exists on the dense set of vectors $\{F_+\Omega : F_+\in\mathscr{P}_{\mathrm{SL}}(O_+),\ F_+ \text{ closed}\}$. This defines a linear operator A_+^{out}.

The operator A_-^{out} is defined analogously by

$$\begin{aligned}A_-^{\mathrm{out}}F_-\Omega &= s-\lim_{T\to\infty}A_{-T}F_-\Omega\\ &= -s-\lim_{T\to\infty}F_-A_{-T}\Omega\\ &= -F_-P_1A_-\Omega\end{aligned}$$

on the dense set of vectors $\{F_-\Omega : F_-\in\mathscr{P}_{\mathrm{SL}}(O_+),\ F_- \text{ closed}\}$. The operators A_+^{out} and A_-^{out} are closable and we also denote the least closed extension of these operators by A_+^{out} and A_-^{out}.

Proof: We prove the statement for the asymptotic Bose operator. The proof of the strong convergence of the sequence $A_TF\Omega$, where $A_+\equiv A\in\mathscr{P}_{\mathrm{SL}}(O)\cap\mathscr{P}_{N_0}$, N_0 sufficiently large, and $F\equiv F_+$ closed and localized in O_+, can be reduced to the proof of the following two conditions:

$$\text{(i) } w-\lim_{T\to\infty}A_TF\Omega = FP_1A\Omega,$$

$$\text{(ii) } \lim_{T\to\infty}\|A_TF\Omega\| = \|FP_1A\Omega\|.$$

To prove (i), we show that for $A\in\mathscr{P}_{N_0}$, $A_TF\Omega$ is uniformly bounded in T:

$$\begin{aligned}\lim_{T\to\infty}\|A_TF\Omega\|^2 &= \lim_{T\to\infty}(A_TF\Omega,A_TF\Omega)\\ &= \lim_{T\to\infty}(F^*F\Omega,A_T^*A_T\Omega),\end{aligned}$$

since F commutes with A_T for sufficiently large T due to the definition of A_T and locality.

The last term can be estimated by

$$\leqslant\|F^*F\Omega\|\lim_{T\to\infty}\|A_T^*A_T\Omega\|\leqslant c,$$

where the constant c does not depend on T.

In the last step, we used the first part of Lemma 2.1 for $n=4$ [i.e., $N_0=N(n=4)$]. Thus the sequence $A_TF\Omega$ is uniformly bounded:

$$\|A_TF\Omega\|\leqslant c\,\forall A\in\mathscr{P}_{N_0}.$$

Therefore, it suffices to prove the convergence of the sequence $A_TF\Omega$ on the dense set of vectors $C\Omega$, $C\in\mathscr{P}_{\mathrm{SL}}$:

$$\begin{aligned}\lim_{T\to\infty}(C\Omega,A_TF\Omega) &= \lim_{T\to\infty}(C\Omega,FA_T\Omega)\\ &= \lim_{T\to\infty}(F^*C\Omega,A_T\Omega)\\ &= (F^*C\Omega,P_1A\Omega)\\ &= (C\Omega,FP_1A\Omega),\end{aligned}$$

where we used the fact that $A_T\Omega$ converges strongly to $P_1A\Omega$ and that the vector $P_1A\Omega$ lies in the domain of $F^{**}=F$ due to the uniform boundedness of the sequence $A_TF\Omega$. The proof of (ii) follows from

$$\lim_{T\to\infty}\|A_T F\Omega\|^2 = \lim_{T\to\infty}(A_T F\Omega, A_T F\Omega)$$
$$= \lim_{T\to\infty}(A_T\Omega, A_T F^* F\Omega)$$
$$= (P_1 A\Omega, F^* F P_1 A\Omega)$$
$$= (FP_1 A\Omega, FP_1 A\Omega).$$

Here, we considered that $A_T\Omega$ converges strongly to $P_1 A\Omega$ and $A_T F^* F\Omega$ converges weakly to $F^* FP_1 A\Omega$ where the latter can be shown as in the proof of (i).

That the above defined operator A^{out} is closable follows from the relation

$$(F'\Omega, A^{\text{out}} F\Omega) = \lim_{T\to\infty}(F'\Omega, A_T F\Omega)$$
$$= \lim_{T\to\infty}(F' A_T^* \Omega, F\Omega)$$
$$= (F' P_1 A^*\Omega, F\Omega),$$

which holds for arbitrary closed F, $F'\in\mathscr{P}_{\mathrm{SL}}(O_+)$. The proof of the statement concerning the operator A_-^{out} can be given analogously using the fact that the vectors $A_T^* A_{-T}\Omega$ are uniformly bounded owing to the second part of Lemma 2.1.

Proof of Lemma 2.1: (i) For the first part we adopt the main steps given in the appendix of Ref. 3, which is split into four parts. We give here a short review of the first two parts of this appendix. The strategy of proof consists in converting the vacuum expectation value

$$(\Omega, A_{1T}\cdots A_{nT}\Omega)$$
$$= \int dt_1\cdots dt_n\, h_T(t_1)\cdots h_T(t_n) t_1\cdots t_n$$
$$\times \int d\omega_1\cdots d\omega_n (\Omega, B_1(t_1, t_1\vec{e}_1)\cdots B_n(t_n, t_n\vec{e}_n)\Omega),$$

where $B_i = -2\partial_0 A_i$, i = 1,...,n into a sum of vacuum expectation values containing only commutators to which the consequences of locality can be applied (examples up to $n = 4$ are given in the above-mentioned appendix).

This procedure is possible because by the spectrum condition we may replace each operator B_i acting on the vacuum by a creation operator B_i^+ such that $B_i\Omega = B_i^+\Omega$ and $(B_i^+)^*\Omega = 0$.

Here, for every $B\in\mathscr{P}_N$, B_+ is given by

$$B^+ = \int dt\, \phi^+(t) A(tn), \quad A\in\mathscr{P}_{\mathrm{SL}},$$

with

$$\varphi^+(t) = (2\pi)^{-1}\int_0^\infty d\omega\, \tilde{\varphi}(\omega) e^{-it\omega}$$

[$\tilde{\varphi}(\omega)$ has an N-fold zero at $\omega = 0$ and n is a positive timelike four-vector]. Here, B^+ is quasilocal of order N; i.e.,

$$\|(B^+ - B_R^+)\Omega\| \leqslant cR^{-N}, \quad \forall R\leqslant R_0.$$

Here, the local approximation B_R^+ of B^+, which is localized in the double cone $\mathscr{C}_R$ of radius R, is defined for $R\leqslant R_0$ by

$$B_R^+ = \int_{-a(R)}^{a(R)} dt\, \varphi^+(t) A(tn),$$
$$a(R) = 2^{-1/2}|n|^{-1}(R - R_0),$$

where $A\in\mathscr{P}_{\mathrm{SL}}(\mathscr{C}_{R_0})$ and $|n|$ is the Euclidean length of n. After this preparation, we turn now to the detailed proof that we split into the following three sections A–C.

A. Bounds on products of multiple commutators

We shall estimate now products of multiple commutators of the operators $A_t^{(+)} = t\int d\omega\, B^{(+)}(t, t\vec{e})$ and their time averages where $B^{(+)}$ stands for B or B^+ [here $d\omega = d\omega(\vec{e})$ is the normalized measure on the unit sphere S^2 in $\mathbf{R}^3$].

For notational convenience, we define for $m\geqslant 2$

$$[m] := [B_1(x_1), [B_2(x_2),...,B_m(x_m)]\cdots],$$

and

$$\left[\overset{+}{m}\right] := [B_{1R}(x_1), [B_2^+(x_2),...,B_m^+(x_m)]\cdots],$$

with

$$x_i = (t_i, t_i\vec{e}_i), B_{1R}\in\mathscr{P}_{\mathrm{SL}}(\mathscr{C}_R)$$

and

$$B_i\in\mathscr{P}_{\mathrm{SL}}(\mathscr{C}_{R_i}) \quad i = 1,...,m.$$

Lemma: Let $[m_i]$ and $\left[\overset{+}{m_i}\right]$ for $i = 1,...,k$ be defined as above. If all t_i, $i = 1,...,n$ are positive (or negative) and $\Sigma_{i=1}^k m_i = n$, then

(a) $$\int d\omega_1\cdots d\omega_n \|[m_1]\cdots[m_k]\Omega\|$$
$$\leqslant c_1 \prod_{s=1}^{k}\prod_{j=1}^{m_s-1}\sum_{l=j+1}^{m_s}(2t_j t_l)^{-1}(R_{jl} + 2R_{jl}|t_j - t_l|),$$

where $R_{kl} = R_k + R_l$ and the constant c_1 does not depend on $t_1,...,t_n$.

(b) $$\int d\omega_1\cdots d\omega_n \left\|\left[\overset{+}{m_1}\right]\cdots\left[\overset{+}{m_k}\right]\Omega\right\|$$
$$\leqslant c_2\left\{R^{-N} + \prod_{s=1}^{k}\prod_{j=1}^{m_s-1}\sum_{l=j+1}^{m_s}(2t_j t_l)^{-1} \times (R^2 + 2R|t_j - t_l|)\right\}$$

and the constant c_2 does not depend on R and $t_1,...,t_n$.

Proof: (a) Due to locality, we only have to integrate in $\int d\omega_1\cdots d\omega_n\|[m_1]\cdots[m_k]\Omega\|$ over a certain region $G\subset S^2\times\cdots\times S^2$. To determine this region, we consider first the simplest case $k = 1$ and $m_1 = 2$; i.e.,

$$\int d\omega_1\, d\omega_2 \|[B_1(t_1, t_1\vec{e}), B_2(t_2, t_2\vec{e}_2)]\Omega\|.$$

Owing to locality, the integrand of this expression vanishes for all $\vec{e}_1$, $\vec{e}_2$ for which the two inequalities $|t_1 - t_2 \pm R_{12}|^2 \leqslant |t_1\vec{e}_1 - t_2\vec{e}_2|^2$ hold. Hence, we have to integrate only over the region $G_{12}\subset S^2\times S^2$:

$$G_{12} = \{\vec{e}_1, \vec{e}_2 : 0\leqslant 1 - \vec{e}_1\vec{e}_2 \leqslant (2t_2 t_2)^{-1}(R_{12}^2 + 2R_{12}|t_1 - t_2|)\}.$$

Thus, if χ_{12} denotes the characteristic function of G_{12}, we can write

$$\int d\omega_1\, d\omega_2 \|[B_1(x_1),B_2(x_2)]\Omega\| = \int d\omega_1\, d\omega_2\, \chi_{12} \|[B_1(x_1),B_2(x_2)]\Omega\|.$$

With the help of this relation, one can prove by induction the following statement:

$$\int d\omega_1 \cdots d\omega_n \|[B_1(x_1),[B_2(x_2),...,B_m(x_m)]\cdots]\Omega\| = \int d\omega_1 \cdots d\omega_n \prod_{k=1}^{m-1} \sum_{l=k+1}^{m} \chi_{kl} \times \|[B_1(x_1),[B_2(x_2),...,B_m(x_m)]\cdots]\Omega\|.$$

Here, χ_{kl} is the characteristic function of the region G_{kl} which is defined in analogy to G_{12}.

Using this result, we get

$$I_1 := \int d\omega_1 \cdots d\omega_n \|[m_1]\cdots[m_k]\Omega\| = \int d\omega_1 \cdots d\omega_n \prod_{s=1}^{k} \prod_{j=1}^{m_s-1} \sum_{l=j+1}^{m_s} \chi_{jl} \times \|[m_1]\cdots[m_k]\Omega\|.$$

The norm appearing in the integrand can be estimated by terms in the form of $\|B_{i_1}(x_{i_1})\cdots B_{i_n}(x_{i_n})\Omega\|$. We consider one of these norms, say $N = \|B_1(x_1)\cdots B_n(x_n)\Omega\|$ and use the invariance of the vacuum under translations to get

$$N = \|B_1(\alpha x_1)B_2(x_2 + (\alpha-1)x_1)\cdots B_n(x_n + (\alpha-1)x_1)\Omega\|,$$

where $\alpha \in \mathbf{R}$ is arbitrary.

Now the $B_i \in \mathscr{P}_{SL}(\mathscr{C}_{R_i})$ are finite sums of operators

$$\phi_{l_1}(f_{l_1})...\phi_{l_m}(f_{l_m}),$$

where $f_{l_j} \in \mathscr{D}(\mathbf{R}^4)$ with $\operatorname{supp} f_{l_j} \subset \mathscr{C}_{R_i}$. Then $B_i(x) = U(x)B_iU(x)^{-1}$ is a finite sum of operators $\phi_{l_1}(f_{l_1,x})\cdots\phi_{l_m}(f_{l_m,x})$ where $f_{l_j,x}$ is the test function f_{l_j} shifted by x.

A typical term of the above norm N reads

$$N' = \|\phi_1(f^{(1)}_{1,\alpha x_1})\cdots\phi_{m_1}(f^{(1)}_{m_1,\alpha x_1}) \times \phi_1(f^{(2)}_{1,x_2+(\alpha-1)x_1})\cdots\phi_{m_2}(f^{(2)}_{m_2,x_2+(\alpha-1)x_1})\cdots \times \phi_1(f^{(n)}_{1,x_n+(\alpha-1)x_1})\cdots \times \phi_{m_n}(f^{(n)}_{m_n,x_n+(\alpha-1)x_1})\Omega\|.$$

According to the temperedness of the Wightman distributions, which are continuous in each argument, we can estimate further

$$N' \leqslant c\|f^{(1)}_{1,\alpha x_1}\|_{s_1 s'_1}\cdots\|f^{(1)}_{m_1,\alpha x_1}\|_{s_{m_1}s'_{m_1}} \times \|f^{(2)}_{1,x_2+(\alpha-1)x_1}\|_{t_1t'_1}\cdots\|f_{m_n,x_n+(\alpha-1)x_1}\|_{r_{m_n}r'_{m_n}}.$$

Here $\|\cdot\|_{ss'}$ denotes the s,s' norm on $\mathscr{S}^4$ which, for the translated function f_x, is bounded by a polynomial in the components of x.

Denoting such a polynomial by P^c, we have

$$N' \leqslant P_1^c(\alpha x_1)P_2^c(x_2 + (\alpha-1)x_1)\cdots P_n^c(x_n + (\alpha-1)x_1).$$

Now, the components of the vector $\alpha x_1 = (\alpha t_1, \alpha t_1 \vec{e}_1)$ can be estimated by $|\alpha t_1| = |\alpha t_1 \vec{e}_1| \leqslant |\alpha||t_1|$ and those of the vectors

$$x_i + (\alpha-1)x_1 = (t_i + (\alpha-1)t_1, t_i\vec{e}_i + (\alpha-1)t_1\vec{e}_1)$$

by $|t_i\vec{e}_i + (\alpha-1)t_1\vec{e}_1| \leqslant |t_i| + |\alpha-1||t_1|$, $i = 2,...,n$. We get, therefore,

$$N' \leqslant c(1 + |\alpha|^{r_1}|t_1|^{r_1})(1 + [|t_2| + |\alpha-1||t_1|]^{r_2})\cdots (1 + [|t_n| + |\alpha-1||t_1|]^{r_n}) \leqslant c'(1 + |\alpha|^{r_1}|t_1|^{r_1}[|t_2| + |\alpha-1||t_1|]^{r_2}\cdots [|t_n| + |\alpha-1||t_1|]^{r_n}),$$

where c, c' are constants and r_i, $i = 1,...,n$ are the degrees of polynomials appearing above.

Since $\alpha \in \mathbf{R}$ is arbitrary we choose especially $\alpha = 0$. Then N' and, therefore, also N is bounded by a constant so that we get for the original expression

$$I_1 \leqslant c_1 \int d\omega_1 \cdots d\omega_n \prod_{s=1}^{k} \prod_{j=1}^{m_s-1} \sum_{l=j+1}^{m_s} \chi_{jl},$$

where c_1 is a constant. Using

$$\int d\omega_i\, \chi_{ik} = \int_0^{(2t_it_k)^{-1}(R_{ik}^2 + 2R_{ik}|t_i - t_k|)} d\xi,$$

we can perform the spherical integrations to get

$$I_1 \leqslant c_1 \prod_{s=1}^{k} \prod_{j=1}^{m_s-1} \sum_{l=j+1}^{m_s} (2t_jt_l)^{-1}(R_{jl}^2 + 2R_{jl}|t_j - t_l|),$$

which proves the first part of the lemma.

(b) We split every operator B_i^+ in the expression $\left[\overset{+}{m_1}\right]\cdots\left[\overset{+}{m_k}\right]\Omega$ into two parts, $B_i^+ = B_{iR}^+ + \hat{B}_i$, $\hat{B}_i = B_i^+ - B_{iR}$ and get

$$\left[\overset{+}{m_1}\right]\cdots\left[\overset{+}{m_k}\right]\Omega = \left[\overset{R}{m_1}\right]\cdots\left[\overset{R}{m_k}\right]\Omega + \sum_{\Lambda_1} [m_1]\overset{\Lambda_1}{\cdots}[m_k]\Omega + \cdots + \sum_{\Lambda_r} [m_1]\overset{\Lambda_r}{\cdots}[m_k]\Omega + \cdots + [\hat{m}_1]\cdots[\hat{m}_k]\Omega \quad (*),$$

where we used the definitions (suppressing the coordinates)

$$\left[\overset{R}{m_i}\right] = [B_{1R},[B_{2R}^+,...,B_{m_iR}^+]\cdots],$$

$$[\hat{m}_i] = [B_{1R},[\hat{B}_2,...,\hat{B}_{m_i}]\cdots],$$

and $[m_1]\overset{\Lambda_r}{\cdots}[m_k]$ means that one has to substitute in $\left[\overset{R}{m_1}\right]\cdots\left[\overset{R}{m_k}\right]$ r operators by the corresponding operators $\hat{B}$ such that the first operator in each multiple commutator does not change.

Finally Σ_{Λ_r} denotes the sum over all r such distinct selections. Now we substitute relation (*) into the expression

$$I_2 = \int d\omega_1 \cdots d\omega_n \left|\left|\left[\overset{+}{m_1}\right]\cdots\left[\overset{+}{m_k}\right]\Omega\right|\right|$$

which we have to consider. To the first term in (*) we can apply part (a) of the lemma. A contribution of the second term in (*) yields

$$I_2' = \int d\omega_1 \cdots d\omega_n \| B_{1R}(x_1) \cdots (B_k^+ - B_{kR}^+)(x_k) \cdots \times B_{nR}^{(+)}(x_n)\Omega \|$$
$$= \int d\omega_1 \cdots d\omega_n \| B_{1R}(x_1) \cdots \left(\int_{-\infty}^{-a(R)} + \int_{a(R)}^{\infty} \right) dt' \times \varphi^+(t') A_k(t'n + x_k) \cdots B_{nR}^{(+)}(x_n)\Omega \|,$$

due to the definition of the operators B_k^+ and B_{kR}^+. Thus we get

$$I_2' \leqslant \int d\omega_1 \cdots d\omega_n \int_{\mathbf{R}\setminus(-a,a)} dt' |\varphi^+(t')| \times \| B_{1R}(x_1) B_{2R}^{(+)}(x_2) \cdots A_k(t'n + x_k) \cdots \times B_{nR}^{(+)}(x_n)\Omega \|.$$

Now, the norm N appearing in the integrand can further be converted and estimated as shown in the first part of the proof:

$$N = \| B_{1R}(\beta x_1) B_{2R}^{(+)}(x_2 + (\beta - 1)x_1) \cdots \times A_k(t'n + x_2 + (\beta - 1)x_1) \cdots \times B_{nR}^{(+)}(x_n + (\beta - 1)x_1)\Omega \|,$$

where $\beta \in \mathbf{R}$ is arbitrary. Then,

$$N \leqslant c'(1 + |\beta|^{r_1}|t_1|^{r_1}[|t_2| + |\beta - 1||t_1|]^{r_2} \cdots \times [|t'|n_0 + |t_k| + |\beta - 1||t_1|]^{r_k} \cdots \times [|t_n| + |\beta - 1||t_1|]^{r_n},$$

with a constant c' and positive integers r_i, $i = 1,...,n$. Putting especially $\beta = 0$ implies $N \leqslant c'$ and, therefore,

$$I_2' \leqslant c' \int_{\mathbf{R}\setminus(-a,a)} dt' |\varphi^+(t')|.$$

Hence, from $a(R) = 2^{-1/2}|n|^{-1}(R - R_0)$ and $|\varphi^+(t)| \leqslant c|t|^{-N-1}$, which one can verify taking into account the N-fold zero of $\tilde{\varphi}(\omega)$ at $\omega = 0$ we get the inequality $I_2' \leqslant cR^{-N}$ for sufficiently large $R > R_0$.

Those terms in I_2 that contain two or more operators $\hat{B}$ can be estimated analogously. Altogether, we get

$$I_2 \leqslant c_1 \prod_{s=1}^{k} \prod_{j=1}^{m_s - 1} \left(\sum_{l=j+1}^{m_s} (2t_j t_l)^{-1}(R^2 + 2R|t_j - t_l|) \right) + c_2(R^{-N} + R^{-2N} + \cdots + R^{-(n-k)N})$$
$$\leqslant c\left\{ R^{-N} + \prod_{s=1}^{k} \prod_{j=1}^{m_s - 1} \left(\sum_{l=j+1}^{m_s} (2t_j t_l)^{-1} \times (R^2 + 2R|t_j - t_l|) \right) \right\},$$

where c_1, c_2, and c are suitable constants and this completes the proof of the lemma. We apply now the above lemma for an estimate of products of multiple commutators of the time averages $A_T^{(+)} = \int dt\, h_T(t) A_t^{(+)}$.

To this end, we define for $m \geqslant 2$,

$$\left[m \right]_T^{(+)} = [A_{1T}, [A_{2T}^{(+)}, ..., A_{mT}^{(+)}] \cdots].$$

Proposition I: Let $B_1,...,B_n$ be operators in $\mathscr{P}_N$ and $\left[m_i \right]_T^{(+)}$ for $i = 1,...,k$ be defined as above. If $\Sigma_{i=1}^k m_i = n$, then,

$$\left\| \left[m_1 \right]_T^{(+)} \cdots \left[m_k \right]_T^{(+)} \Omega \right\| \leqslant c|T|^{-[N(n-2k) - 2n(n-k)]/[2(n-k)+N]},$$

for large $|T|$. The constant c does not depend on T.

Remarks: (i) If $k = 1$, we get the bound given by Buchholz for the norm of a multiple commutator[3] which, for $n \geqslant 3$, decreases with $|T|$ like $|T|^{-(n-2)}$.

(ii) If $k = n/2$, we have a product of simple commutators and the above bound increases with $|T|$ like $|T|^{n^2/(n+N)}$.

(iii) However, if there is at least one multiple commutator with $m_i \geqslant 3$ and if $k \geqslant 2$, the two relations $\Sigma_{i=1}^k m_i = n$ and $m_i \geqslant 2$ imply the inequality $n - 2k \geqslant 1$. Thus for sufficiently large N the bound decreases like $|T|^{-(n-2k)}$.

Proof: By definition we have

$$M := \left\| \left[m_1 \right]_T^{(+)} \cdots \left[m_k \right]_T^{(+)} \Omega \right\| \leqslant \int dt_1 \cdots dt_n |h_T(t_1) \cdots h_T(t_n)||t_1 \cdots t_n| \times \int d\omega_1 \cdots d\omega_n \left\| \left[m_1 \right]^{(+)} \cdots \left[m_k \right]^{(+)} \Omega \right\|.$$

To the spherical integrations, we apply now the second part of the above lemma. This is possible since for $R \geqslant R_i$, we can substitute every local operator

$$B_i = \int_{-a(R_i)}^{a(R_i)} dt\, \varphi_i(t) A_i(tn) \in \mathscr{P}_{\mathrm{sl}}(\mathscr{C}_{R_i}),$$
$$\varphi_i \in \mathscr{D}([-a(R_i), a(R_i)]),$$

appearing in the above expression by

$$B_{iR} = \int_{-a(R_i)}^{a(R_i)} dt\, \varphi_i(t) A_i(tn) \in \mathscr{P}_{\mathrm{SL}}(\mathscr{C}_R)$$
$$(B_{iR} = B_i \text{ for } R \geqslant R_i).$$

Further, we see from the proof of the second part of the above lemma that all terms containing less than $(n - k)$ quasilocal operators can be estimated by the bound given in this lemma which yields

$$M \leqslant c' \int dt_1 \cdots dt_n |h_T(t_1) \cdots h_T(t_n)||t_1 \cdots t_n| \times \left\{ R^{-N} + \prod_{s=1}^{k} \prod_{j=1}^{m_s - 1} \left(\sum_{l=j+1}^{m_s} (2t_j t_l)^{-1} \times (R^2 + 2R|t_j - t_l|) \right) \right\}.$$

Taking into account the support properties of h_T we get after integration

$$M \leqslant c|T|^n\left\{R^{-N}+\left(\frac{R^2+2R\ln|T|}{T^2}\right)^{\sum_{i=1}^{k}(m_i-1)}\right\}$$
$$=c|T|^n\left\{R^{-N}+\left[\frac{R^2}{T^2}\left(1+\frac{2\ln|T|}{R}\right)\right]^{n-k}\right\},$$

for sufficiently large $|T|$ with a constant c, which does not depend on R and $|T|$.

This inequalilty holds for arbitrary $R>0$ and if we put $R=|T|^{2(n-k)/2(n-k)+N}$ the statement follows.

B. The vacuum expectation value of two commutators

For later applications our bound on the norm of $[A_{1T},A_{2T}^{(+)}]\Omega$ is too weak. We shall estimate in the following the vacuum expectation value of two such commutators. To this end, we adopt, with an obvious change of notation, the strategy of proof given in part (c) of the appendix in Ref. 3 to get the following.

Proposition II: Let $B_1,...,B_4$ be operators in $\mathscr{P}_N$. Then, for large $|T|$,

$$|(\Omega,[A_{1T},A_{2T}^{+}](1-P_0)[A_{3T},A_{4T}^{+}]\Omega)|$$
$$\leqslant c\ln|T|\,|T|^{-2(N-14)/(N+7)},$$

where P_0 is the projection onto the vacuum and c is a constant which does not depend on T.

C. Proof of part (i) of Lemma 2.1

We are now prepared to complete our argument. By converting the vacuum expectation values $(\Omega,A_{1T}\cdots A_{nT}\Omega)$ into a sum of terms containing only commutators there are two types of contributions. In the first one,

$$\left(\Omega,\left[m_1^{(+)}\right]_T\cdots\left[m_k^{(+)}\right]_T\Omega\right),$$

there is at least one multiple commutator with $m_i\geqslant 3$. Such a term can be estimated by

$$\left|\left|\left[m_1^{(+)}\right]_T\cdots\left[m_k^{(+)}\right]_T\Omega\right|\right|$$

which, by the remark after Proposition I, coverges to zero in the limit of large $|T|$ provided all operators $B_1...B_n$ are in $\mathscr{P}_N$ with N sufficiently large.

If n is even the remaining contributions are of the form $M_n(T):=(\Omega[A_{1T},A_{2T}^{+}],...,[A_{n-1T},A_{nT}^{+}]\Omega)$. We shall prove by induction that for sufficiently large N, $M_n(T)$ converges in the limit of large $|T|$ to a product of one particle scalar products.

For $n=2$, the statement follows from the strong convergence of the vectors $A_T\Omega$:

$$\lim_{T\to\pm\infty}M_2(T)=\lim_{T\to\pm\infty}(\Omega,[A_{1T},A_{2T}^{+}]\Omega)$$
$$=\lim_{T\to\pm\infty}(A_{1T}^{*}\Omega,A_{2T}\Omega)$$
$$=(\Omega,[A_1,P_1A_2]\Omega).$$

Let us assume now that the statement holds for $(n-2)$.

We split $M_n(T)$ into two terms:

$$M_n(T)=(\Omega,[A_{1T},A_{2T}^{+}]\Omega)(\Omega,[A_{3T},A_{4T}^{+}]\cdots$$
$$\times[A_{n-1T},A_{nT}^{+}]\Omega)+(\Omega,[A_{1T},A_{2T}^{+}]$$
$$\times(1-P_0)[A_{3T},A_{4T}^{+}]\cdots$$
$$\times[A_{n-1T},A_{nT}^{+}]\Omega).$$

The first term converges by assumption to $(\Omega,A_1P_1A_2\Omega)\cdots(\Omega,A_{n-1}P_1A_n\Omega)$.

The second term can be estimated using Propositions I and II, by

$$|(\Omega,[A_{1T},A_{2T}^{+}](1-P_0)[A_{3T},A_{4T}^{+}]\cdots[A_{n-1T},A_{nT}^{+}]\Omega)|$$
$$\leqslant\|(1-P_0)[A_{1T},A_{2T}^{+}]^{*}\Omega\|\,\|[A_{3T},A_{4T}^{+}]\cdots$$
$$\times[A_{n-1T},A_{nT}^{+}]\Omega\|$$
$$\leqslant c(\ln|T|)^{1/2}|T|^{-(N-14)/(N+7)}|T|^{(n-2)^2/(n-2+N)}.$$

If N is sufficiently large (depending on n) this expression converges to zero in the limit of large $|T|$ and therefore $\lim_{T\to\pm\infty}M_n(T)=(\Omega,A_1P_1A_2\Omega)\cdots(\Omega,A_{n-1}P_1A_n\Omega)$. Summing up all contributions of the type $M_n(T)$ the first part of the lemma follows after some combinatorics.

(ii) For the proof of the second part of the lemma we proceed as in part (i) substituting the Bose operators A_i by the Fermi operators ψ_i and converting the vacuum expectation values $(\Omega,\psi_{1T}\cdots\psi_{nT}\Omega)\equiv(1\cdots n)$ into a sum of vacuum expectation values containing only commutators and anticommutators. We give the first two expressions using an obvious notation:

$$n=2:(12)=(\overset{+}{1\,2})=(\{1,\overset{+}{2}\}).$$

$$n=4:(1234)=(\{1,\overset{+}{4}\}\{2,\overset{+}{3}\})-(\{2,\overset{+}{4}\}\{1,\overset{+}{3}\})$$
$$+(\{3,\overset{+}{4}\}\{1,\overset{+}{2}\})-\left(\left[\left[\{1,\overset{+}{4}\},\overset{+}{3}\right],\overset{+}{2}\right]\right)$$
$$-\left(\left[1,\{2,\overset{+}{4}\}\right],\overset{+}{3}\right])-\left(\left[\left[\{3,\overset{+}{4}\},\overset{+}{2}\right],\overset{+}{1}\right]\right)$$
$$+\left(\left[\left[\{2,\overset{+}{4}\},\overset{+}{3}\right],\overset{+}{1}\right]\right)+\left(\left[\left[1,\{3,\overset{+}{4}\}\right],\overset{+}{2}\right]\right)$$
$$+\left(\left[\left[2,\{3,\overset{+}{4}\}\right],\overset{+}{1}\right]\right)+\left(\left[1,\left[2,\{3,\overset{+}{4}\}\right]\right]\right).$$

Analogously to the proof of part (i) one can show that those terms which contain multiple commutators vanish in the limit of large $|T|$ and that the remaining vacuum expectation values containing only products of simple anticommutators converge to the sum given in the second part of Lemma 2.1.

Remark: An elementary consequence of Lemma 2.1 is that the sequences $A_{1T}\cdots A_{nT}\Omega$ and $\psi_{1T}\cdots\psi_{nT}\Omega$ are uniformly bounded in T:

$$\|A_{1T}\cdots A_{nT}\Omega\|\leqslant c,\quad\|\psi_{1T}\cdot\psi_{nT}\Omega\|\leqslant c.$$

III. THE COLLISION STATES

A. Construction of asymptotic collision states for massless Fermions

We construct the asymptotic many particle states for massless Fermions with the help of the asymptotic fields $\psi^{\text{out}}(\psi\equiv A_{\_})$ using essentially part (ii) of Lemma 2.1 and methods already developed in Refs. 1 and 3.

The main step of this construction consists in showing that the anticommutator of two asymptotic operators is a c number. Consequently, ψ^{out} may be extended to a bounded operator on the whole Hilbert space. Further one can point out that ψ^{out} has all the properties of a smeared free field.

Then, as in Ref. 1, the asymptotic outgoing n-particle states are defined by

$$\Phi_1 \overset{\text{out}}{\times} \cdots \overset{\text{out}}{\times} \Phi_n := \psi_1^{\text{out}(+)} \cdots \psi_n^{\text{out}(+)} \Omega,$$

where $\psi_i^{\text{out}(+)}$, $i = 1,...,n$ is the creation part of ψ_i^{out}. To begin with, we show, adopting the proof of Lemma 3 in Ref. 3, that the sequence of vectors $\psi_{1T} \cdots \psi_{nT} \Omega$ converges weakly in the limit of large T due to their uniform boundedness.

Lemma 2.3: Let $\psi, \psi_1, ..., \psi_n$ be elements in $\mathscr{P}_N$ and N sufficiently large.

(a) Then the weak limit

$$w - \lim_{T \to \infty} \psi_{1T} \cdots \psi_{nT} \Omega = \Psi^{\text{out}}(\psi_1, ..., \psi_n)$$

exists. It is multilinear in $\psi_1, ..., \psi_n$ and depends only on the one particle states $P_1 \psi_1 \Omega, ..., P_1 \psi_n \Omega$.

(b) $\Psi^{\text{out}}(\psi_1, ..., \psi_n)$ is in the domain of $\psi^{\text{out}*}$ and $\psi^{\text{out}*} \Psi^{\text{out}}(\psi_1, ..., \psi_n) = \Psi^{\text{out}}(\psi^*, \psi_1, ..., \psi_n)$.

(c) If, in addition, ψ is closed and localized in $\mathscr{O}$ and $\psi_1, ..., \psi_n$ are localized in the future tangent $\mathscr{O}_+$ of $\mathscr{O}$, then $\Psi^{\text{out}}(\psi, ..., \psi_n)$ is also in the domain of ψ^{out} and $\psi^{\text{out}} \Psi^{\text{out}}(\psi_1, ..., \psi_n) = \Psi^{\text{out}}(\psi, \psi_1, ..., \psi_n)$.

Owing to the part (c) of this lemma $\psi_2^{\text{out}} \Omega$ is in the domain $D(\psi_1^{\text{out}})$ of ψ_1^{out} and $\psi_1^{\text{out}} \psi_2^{\text{out}} \Omega = \Psi^{\text{out}}(\psi_1, \psi_2)$ provided the ψ_i are localized in $\mathscr{O}_i$, $i = 1,2$, ψ_1 is closed and $\mathscr{O}_2 \subset (\mathscr{O}_1)_+$.

In the next step, we want to extend this relation to arbitrary regions of localization $\mathscr{O}_1, \mathscr{O}_2$. To this end it is convenient to define the following vectors for any closed $\psi_i, \psi_j \in \mathscr{P}_N$, N sufficiently large

$$\Phi_i = \Psi^{\text{out}}(\psi_i) = \psi_i^{*\text{out}*} \Omega = P_1 \psi_i \Omega,$$

$$\Phi_i \overset{\text{out}}{\times} \Phi_j = \Psi^{\text{out}}(\psi_i, \psi_j) - (\Omega, \psi_i P_1 \psi_j \Omega) \Omega$$

$$= \psi_i^{*\text{out}*} \psi_j^{*\text{out}*} \Omega - (\Omega, \psi_i P_1 \psi_j \Omega) \Omega.$$

Here, Φ_i is a massless one-particle state and $\Phi_i \overset{\text{out}}{\times} \Phi_j$ is the normal ordered product of the operators $\psi_i^{*\text{out}*} \psi_j^{*\text{out}*}$ applied to the vacuum.

Now we can proceed as after the proof of Lemma 4 in Ref. 3. Accordingly, for any vectors $\Phi_1, \Phi_2 \in \mathscr{H}_1$ we can specify sequences of operators $\psi_1^{(n)}, \psi_2^{(n)} \in \mathscr{P}_N$, localized in $\mathscr{O}_1^{(n)}, \mathscr{O}_2^{(n)}$ with $\mathscr{O}_2^{(n)} \subset (\mathscr{O}_1^{(n)})_+$ such that the strong limits $s - \lim_n P_1 \psi_1^{(n)} \Omega = \Phi_1$ and $s - \lim_n P_1 \psi_2^{(n)} \Omega = \Phi_2$ exist.

Further, we can show that

$$s = \lim_n \Phi_1^{(n)} \times \Phi_2^{(n)} = \Phi_1 \overset{\text{out}}{\times} \Phi_2,$$

where $\Phi_1 \overset{\text{out}}{\times} \Phi_2$ is the vector which we introduced above for arbitrary $\psi_2, \psi_2 \in \mathscr{P}_N$.

By definition, we get for $\mathscr{O}_2 \subset (\mathscr{O}_1)_+$

$$\psi_1^{\text{out}} \psi_2^{\text{out}} \Omega = \psi_1^{\text{out}} \Phi_2 = \Phi_1 \overset{\text{out}}{\times} \Phi_2 + (P_1 \psi_1^* \Omega, \Phi_2) \Omega.$$

This relation can be extended to arbitrary $\Phi_1, \Phi_2 \in \mathscr{H}_1$. To this end let $\Phi_i^{(n)}$ be defined as above, i.e,

$$s - \lim_n \Phi_i^{(n)} = \Phi_i, \quad i = 1,2$$

and

$$s - \lim_n \Phi_1^{(n)} \overset{\text{out}}{\times} \Phi_2^{(n)} = \Phi \overset{\text{out}}{\times} \Phi_2.$$

Now $\Phi_2^{(n)}$ is in the domain of the closed operator ψ_1^{out} and $\psi_1^{\text{out}} \Phi_2^{(n)}$ converges strongly:

$$s - \lim_n \psi_1^{\text{out}} \Phi_2^{(n)}$$

$$= s - \lim_n \left\{ \Phi_1^{(n)} \overset{\text{out}}{\times} \Phi_2^{(n)} + (P_1 \psi_1^* \Omega, \Phi_2^{(n)}) \Omega \right\}$$

$$= \Phi_1 \overset{\text{out}}{\times} \Phi_2 + (P_1 \psi_1^* \Omega, \Phi_2) \Omega.$$

Since ψ_1^{out} is closed, we have the desired extension: $\psi_2^{\text{out}} \Omega$ is an element from $D(\psi_1^{\text{out}})$ and

$$\psi_1^{\text{out}} \psi_2^{\text{out}} \Omega = \Phi_1 \overset{\text{out}}{\times} \Phi_2 + (\Omega, \psi_1 P_1 \psi_2 \Omega) \Omega,$$

for all ψ_1, ψ_2 localized in arbitrary regions $\mathscr{O}_1$, $\mathscr{O}_2$. This implies the following lemma.

Lemma 2.4: Let $\psi_i \in \mathscr{P}_{\text{SL}}(\mathscr{O}_i)$, $i = 1,2$ be in $\mathscr{P}_N$, N sufficiently large. If $F \in \mathscr{P}_{\text{SL}}((\mathscr{O}_1 \cup \mathscr{O}_2)_+)$ is closed then $\psi_2^{\text{out}} F \Omega$ is in $D(\psi_1^{\text{out}})$ and $\psi_1^{\text{out}} \psi_2^{\text{out}} F \Omega = F \psi_1^{\text{out}} \psi_2^{\text{out}} \Omega$.

This relation defines the operator $\psi_1^{\text{out}} \psi_2^{\text{out}}$ on the dense set of vectors $\{F\Omega : F \in \mathscr{P}_{\text{SL}}((\mathscr{O}_1 \cup \mathscr{O}_2)_+), F \text{ closed}\}$.

Proof: According to the proof of part (c) of Lemma 1 in Ref. 3, we can verify that the relations

$$F D(\phi^{\text{out}}) \subset D(\psi^{\text{out}}) \quad \text{and} \quad \{\psi^{\text{out}}, F\} \Phi = 0$$

hold for arbitrary $\Phi \in D(\phi^{\text{out}})$ provided $\psi \in \mathscr{P}_N$ is localized in $\mathscr{O}$, N is sufficiently large, and F is closed and localized in $\mathscr{O}_+$ (the existence of the strong limits appearing in the proof can be shown in analogy to Lemma 2.2). If we put $\Phi = \psi_2^{\text{out}} \Omega \in D(\psi_1^{\text{out}})$ in the above relation, the statement follows.

The next lemma will be used in order to prove that the anticommutator of two asymptotic operators is a c number.

Lemma 2.5: Let $\psi_i \in \mathscr{P}_{\text{SL}}(\mathscr{O}_i)$, $i = 1,2$ be in $\mathscr{P}_N$, N sufficiently large. Then

$$\{\psi_1^{\text{out}}, \psi_2^{\text{out}}(x)\} \Omega = 0,$$

provided $\mathscr{O}_2 + x$ lies in the future or past tangent of $\mathscr{O}_1$.

Proof: It suffices to show that $\{\psi_1^{\text{out}}, \psi_2^{\text{out}}\} \Omega = 0$ if $\mathscr{O}_2$ lies in the future or past tangent of $\mathscr{O}_1$.

For $F \in \mathscr{P}_{\text{SL}}((\mathscr{O}_1 \cup \mathscr{O}_2)_+)$ and $\mathscr{O}_2 \subset (\mathscr{O}_1)_+$, we consider the scalar product

$$(F\Omega,\Psi^{\text{out}}(\psi_1,\psi_2)) = \lim_{T\to\infty}(F\Omega,\psi_{1T}\psi_{2T}\Omega) = \lim_{T\to\infty}(\psi_{1T}^{*}F\Omega,\psi_{2T}\Omega) = \lim_{T\to\infty}(\psi_1^{*^{\text{out}}}F\Omega,\psi_{2T}\Omega) = \lim_{T\to\infty}(F\Omega,\psi_1^{*^{\text{out}*}}\psi_{2T}\Omega)$$

$$= -\lim_{T\to\infty}(F\Omega,\psi_{2T}\psi_1^{*^{\text{out}*}}\Omega) = -\lim_{T\to\infty}(\psi_{2T}^{*}F\Omega,\psi_1^{\text{out}}\Omega) = -(\psi_2^{\text{out}*}F\Omega,\psi_1^{\text{out}}\Omega) = -(F\Omega,\psi_2^{\text{out}}\psi_1^{\text{out}}\Omega).$$

Here, we used that $\psi_{2T}\Omega$ is for large T in $D(\psi_1^{*^{\text{out}*}})$ owing to locality and further to the fact that $\psi_1^{\text{out}}\Omega$ is in $D(\psi_2^{\text{out}})$ also for arbitrary regions $\mathcal{O}_1$, $\mathcal{O}_2$.

From the last equation, we get for $\mathcal{O}_2\subset(\mathcal{O}_1)_+$:

$(F\Omega,\{\psi_1^{\text{out}},\psi_2^{\text{out}}\}\Omega) = 0.$

Since the set $\{F\Omega: F\in\mathcal{P}_{\text{SL}}((\mathcal{O}_1\cup\mathcal{O}_2)_+)\}$ is dense the statement follows for $\mathcal{O}_2\subset(\mathcal{O}_1)_+$. If $\mathcal{O}_2$ is localized in the past tangent of $\mathcal{O}_1$ we have $\mathcal{O}_1\subset(\mathcal{O}_2)_+$. Therefore, the statement follows also in this case (change the indices 1 and 2 in the above proof).

In analogy to the proof of Lemma 4 in Ref. 1, we are now able to prove that the anticommutator of two asymptotic operators is a c number.

Lemma 2.6: Let ψ_1^{out} and ψ_2^{out} be two asymptotic operators. Then,

$$\{\psi_1^{\text{out}},\psi_2^{\text{out}}\} = (\Omega,\{\psi_1^{\text{out}},\psi_2^{\text{out}}\}\Omega)\mathbf{1}.$$

From this lemma, we can conclude that ψ^{out} may be extended to a bounded operator on the whole Hilbert space.

We denote this bounded operator also by ψ^{out}.

The following lemma which also can be proven as in Ref. 1 tells us that the operators ψ^{out} have all properties of a smeared free field.

Lemma 2.7: Let ψ^{out}, ψ_1^{out} and ψ_2^{out} be asymptotic operators. Then, (a) $\psi^{\text{out}}(x)$ is a solution of the wave equation $\partial_\mu\partial^\mu\psi^{\text{out}}(x) = 0$ and (b) $\{\psi_1^{\text{out}},\psi_2^{\text{out}}\} = 0$ if ψ_1 and ψ_2 are localized in two spacelike or timelike separated double cones.

We construct now the collision states for massless Fermions with the help of the operators ψ^{out}. To this end, we split ψ^{out} into a creation part $(\psi^{\text{out}})^{(+)}$ and a destruction part $(\psi^{\text{out}})^{(-)}$

$$\psi^{\text{out}} = (\psi^{\text{out}})^{(+)} + (\psi^{\text{out}})^{(-)}, (\psi^{\text{out}})^{(-)}\Omega = 0.$$

Now let $\psi_1^{\text{out}},\ldots,\psi_n^{\text{out}}$ be n asymptotic operators that create one-particle states $\Phi_1,\ldots,\Phi_n\in\mathcal{H}_1$ from the vacuum. Then, the outgoing collision states of these particles are defined by

$$\Phi_1 \overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_n = (\psi_1^{\text{out}})^{(+)}\cdots(\psi_n^{\text{out}})^{(+)}\Omega.$$

Following Buchholz in Ref. 1, one can prove finally that the Hilbert space $\mathcal{H}^{\text{out}}$, which is generated by $\Phi_1\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_n$, $n\in\mathbf{N}$ and Ω is a Fock space over the one-particle space $\mathcal{H}_1$ of massless Fermions.

Thus the vectors $\Phi_1\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_n$ can be interpreted as outgoing configurations of noninteracting particles $\Phi_1,\ldots,\Phi_n$. Of course the incoming collision states can be defined analogously. This allows the usual definition and interpretation of an S-matrix that completes our investigation for the massless Fermions.

B. Construction of asymptotic collision states for massless Bosons

In the case of massless Bosons, we can adopt the construction given by Buchholz in Ref. 3. Accordingly, one considers first for n Bose operators $A_1,\ldots,A_n\in\mathcal{P}_N$, N, sufficiently large, the vectors

$$\Psi^{\text{out}}(A_1,\cdots,A_n) = w-\lim_{T\to\infty}A_{1T}\cdots A_{nT}\Omega$$

$$= A_1^{*^{\text{out}*}}\cdots A_n^{*^{\text{out}*}}\Omega,$$

which are well defined.

Then the outgoing asymptotic n-particle states $\Phi_1\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_n$ are defined recursively with the help of the vectors $\Psi^{\text{out}}(A_1,\ldots,A_n)$ as follows:

$$\Phi_i = A_i^{*^{\text{out}*}}\Omega = P_1A_i\Omega,$$

$$\Phi_i\overset{\text{out}}{\times}\Phi_j = A_i^{*^{\text{out}*}}\Phi_j - (\Omega,A_iP_1A_j\Omega)\Omega$$

$$\vdots$$

$$\Phi_{i_1}\overset{\text{out}}{\times}\Phi_{i_2}\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_{i_n} = A_{i_1}^{*^{\text{out}*}}\Phi_{i_2}\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_{i_n}$$

$$-\sum_{k=2}^{n}(\Omega,A_{i_1}P_1A_{i_k}\Omega)\Phi_{i_2}\overset{\text{out}}{\times}\cdots\overset{i_k}{\vee}\cdots\overset{\text{out}}{\times}\Phi_{i_n},$$

where the symbol $\overset{j}{\vee}$ denotes omission of the one-particle state Φ_j.

One can show that $\Phi_1\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_n$ is the normal-ordered product of the operators $A_1^{*^{\text{out}*}},\ldots,A_n^{*^{\text{out}*}}$ applied to the vacuum. The most important properties of the vectors $\Phi_1\overset{\text{out}}{\times}\cdots\overset{\text{out}}{\times}\Phi_n$, $n\in\mathbf{N}$ are that they, together with the vacuum Ω, generate the Hilbert space $\mathcal{H}^{\text{out}}$, which is the usual Fock space over the one-particle space $\mathcal{H}_1$ and that they can be

interpreted as outgoing configurations of noninteracting particle $\Phi_1,\ldots,\Phi_n$.

This allows the usual definition and interpretation of an S matrix.

To conclude our investigation let us list some relevant facts which follow with the help of the collision states.

One can show as in Ref. 3 that the sequence of vectors $A_{1T}\cdots A_{nT}\Omega$ converges even strongly and that due to Theorem 8 and Theorem 9 in Ref. 3 the asymptotic operator A^{out} has all properties of a smeared free field.

ACKNOWLEDGMENT

I am indebted to Professor Dr. H. Reeh for helpful discussions and for reading the manuscript.

[1] D. Buchholz, Commun. Math. Phys. **42**, 269 (1975).
[2] D. Buchholz, Commun. Math. Phys. **45**, 1 (1975).
[3] D. Buchholz, Commun. Math. Phys. **52**, 147 (1977).
[4] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics and All That* (Benjamin, New York, 1964).
[5] J. Yngvason, Commun. Math. Phys. **18**, 195 (1970).

High symmetry solutions of the anti-self-dual Yang–Mills equations

Sarbarish Chakravarty
Program of Applied Mathematics, University of Colorado, Boulder, Colorado 80309

S. L. Kent
Department of Mathematical and Computer Sciences, Youngstown State University, Youngstown, Ohio 44555

E. T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 13 February 1990; accepted for publication 18 April 1990)

Beginning with the anti-self-dual Yang–Mills (ASDYM) equations for an arbitrary Lie algebra on Minkowski space, this paper specializes to the case in which the vector potentials are independent of all the space-time coordinates, i.e., are space-time constants. The resulting equations are three algebraic equations on the algebra. These equations are then simplified by using a null basis. Two of the equations can be immediately solved while the third remains, in general, quite difficult to deal with. Two general cases are considered: finite-dimensional Lie groups and the infinite-dimensional diffeomorphism groups on finite-dimensional manifolds. In a few of the special cases, e.g., SL(2,C) and the Virasoro algebra, the solutions can easily be found. The study of the the diffeomorphism groups leads unexpectedly to the Monge–Ampère equation. In particular, the four-dimensional volume preserving diffeomorphism group is identical with the vacuum anti-self-dual Einstein equations. In conclusion, the question of the associated Lax pair equations and its relation to the Riemann–Hilbert splitting problem on S^2 is examined.

I. INTRODUCTION

This paper is motivated by two main considerations: (a) In the past several years, a great deal of progress has been made in understanding the structure of the anti-self-dual Yang–Mills (ASDYM) equations and various solution generating methods. In one approach,[1,2] the ASDYM equations are thought of as the integrability condition on a pair of first-order linear differential equations, the so-called "Lax pair." The Lax pair leads, by several ingenious techniques, to the subject of Bäcklund transformations,[3–5] where the basic idea is to generate new solutions of the ASDYM equations from a seed solution. Another approach, due to Ward,[6] and with a slight variation by Sparling,[7] establishes the correspondence between solutions of the ASDYM equations and holomorphic vector bundles on twistor space. Solutions are then generated by a Riemann–Hilbert splitting of the patching functions defining the bundle.

(b) Recently, Mason and Newman[8] showed there is an unusual relationship between the Yang–Mills equations and the Einstein equations by first considering the Yang–Mills theory for an arbitrary Lie algebra with the symmetry condition that the connection one-form and curvature are *constant on Minkowski space*. This leads to a set of *algebraic* equations on the connection components. When the Lie algebra is specialized to be the (*infinite-dimensional*) Lie algebra of the group of diffeomorphisms of some auxiliary manifold, the algebraic equations become differential equations for vector fields on this auxiliary manifold. In the anti-self-dual case, if one chooses the connection components from the Lie algebra of the volume preserving four-dimensional diffeomorphisms, the resulting equations become the anti-self-dual Einstein vacuum equations. Other generalizations are also possible and will be discussed.

Motivated by the above considerations, this paper will be split into two parts: (i) First we will impose the symmetry restrictions on the connection one-form and show how the ASDYM equations can be simplified to three algebraic equations. We then study these three equations for two general classes of Lie algebras, the finite-dimensional ones and the diffeomorphism algebras on an n-dimensional manifold. Several special cases are worked out in detail. (ii) We then show how the above approach is related to the more general Riemann–Hilbert splitting problem on S^2 for the general ASDYM equations. Specifically, in Sec. II, we discuss the ASDYM equations on Minkowski space and the algebraic equations obtained by imposing the maximal symmetry conditions. In Sec. III, we study solution generating methods for the two classes of Lie algebras just mentioned, and finally in Sec. IV, we examine the associated Riemann–Hilbert problem.

II. THE REDUCED YANG–MILLS EQUATIONS

Consider a vector potential or connection γ_a on Minkowski space, M, where for each $a = 0,...,3$, $\gamma_a \in l$, for some Lie algebra l. The curvature is then given by

$$F_{ab} \equiv [D_b, D_a] = [\partial_b - \gamma_b,\ \partial_a - \gamma_a]$$
$$= 2\partial_{[a}\gamma_{b]} - [\gamma_a, \gamma_b]. \tag{2.1}$$

The full Yang–Mills equations are

$$D^a F_{ab} = 0, \tag{2.2}$$

and the anti-self-dual Yang–Mills (ASDYM) equations are

 0022-2488/90/092253-05$03.00

$$F^*_{ab} \equiv \tfrac{1}{2}\epsilon_{ab}{}^{cd}F_{cd} = -iF_{ab}, \tag{2.3}$$

where the asterisk is the Hodge duality operator. It is now possible to impose various symmetry reductions on the connection γ. For example, if we consider the case in which the γ_a only depend on $t = x^0$, we are led (in the anti-self-dual case) to Nahm's equation.[8] In our case though, we wish to impose the maximum symmetry by assuming that the connection components, in some choice of gauge, are independent of *all* the space-time coordinates, i.e., the γ_a are each constant elements of the Lie algebra. The full Yang–Mills equations then reduce to

$$[\gamma^a, [\gamma_a, \gamma_b]_l]_l = 0, \tag{2.4}$$

and the ASDYM equations reduce to

$$[\gamma_a, \gamma_b]_l = i\tfrac{1}{2}\epsilon_{ab}{}^{cd}[\gamma_c, \gamma_d]_l, \tag{2.5}$$

where the brackets, $[\ ,\ ]_l$ are the Lie algebra brackets for l. The Jacobi identity, $\epsilon^{abcd}[\gamma_b, [\gamma_c, \gamma_d]_l]_l = 0$, implies that solutions of (2.5) also satisfy (2.4). If we introduce null coordinates instead of the ordinary Minkowski coordinates, i.e.,

$$u = \tfrac{1}{2}(t - z), \quad v = \tfrac{1}{2}(t + z),$$
$$\omega = \tfrac{1}{2}(x + iy), \quad \bar{\omega} = \tfrac{1}{2}(x - iy), \tag{2.6}$$

so that the Minkowski metric is

$$ds^2 = 4(du\, dv - d\omega\, d\bar{\omega}), \tag{2.7}$$

the transformed field equations (2.5) become the triple

$$[\gamma_u, \gamma_\omega]_l = 0, \tag{2.8a}$$
$$[\gamma_v, \gamma_{\bar{\omega}}]_l = 0, \tag{2.8b}$$
$$[\gamma_u, \gamma_v]_l - [\gamma_\omega, \gamma_{\bar{\omega}}]_l = 0, \tag{2.8c}$$

where $\gamma_u, \gamma_v, \gamma_\omega, \gamma_{\bar{\omega}}$ are the connection components in the new coordinate system. The study of properties and solutions of Eqs. (2.8) will be our main concern in the next two sections.

III. SOLUTIONS

We will consider Eqs. (2.8) for two different classes of Lie algebras: finite-dimensional algebras and the diffeomorphism algebras on finite-dimensional manifolds.

A. Finite-dimensional algebras

If we consider the set of γ_u, γ_v, γ_ω, $\gamma_{\bar{\omega}}$ as finite-dimensional matrices and assume that they are all diagonalizable, then it is easy to see that the general solution to (2.8a) and (2.8b) is

$$\gamma_u = G_1\gamma_u^D G_1^{-1}, \quad \gamma_\omega = G_1\gamma_\omega^D G_1^{-1}, \tag{3.1a}$$
$$\gamma_v = G_2\gamma_v^D G_2^{-1}, \quad \gamma_{\bar{\omega}} = G_2\gamma_{\bar{\omega}}^D G_2^{-1}, \tag{3.1b}$$

where D indicates an arbitrary *diagonal* matrix, and G_1 and G_2 are arbitrary nonsingular matrices (needed to diagonalize the γ's). When (3.1) are substituted into (2.8c), one obtains the algebraic equation

$$\gamma_u^D J\gamma_v^D J^{-1} - J\gamma_v^D J^{-1}\gamma_u^D = \gamma_{\bar{\omega}}^D J\gamma_\omega^D J^{-1} - J\gamma_\omega^D J^{-1}\gamma_{\bar{\omega}}^D, \tag{3.2}$$

for the determination of

$$J = G_1^{-1}G_2. \tag{3.3}$$

(With no loss in generality one can always take, say, $G_1 = I$.)

Example: SL(2,C).

In this case, each component of γ is a 2×2 trace-free matrix. Assuming that the four γ's $\neq 0$, then a simple calculation shows that

$$\gamma_u = \lambda\gamma_\omega \quad \text{and} \quad \gamma_v = \mu\gamma_{\bar{\omega}}. \tag{3.4}$$

The third equation of (2.8) implies that

$$(\lambda\mu - 1)[\gamma_\omega, \gamma_{\bar{\omega}}]_l = 0. \tag{3.5}$$

Thus, either γ has at most one linearly independent component, or else $\lambda\mu = 1$. In the former case, the Yang–Mills field would be trivial. Thus the only nontrivial solution for this case is that γ_ω and $\gamma_{\bar{\omega}}$ are two arbitrary linearly independent components of γ, and

$$\gamma_u = \lambda\gamma_\omega, \quad \gamma_v = \lambda^{-1}\gamma_{\bar{\omega}},$$

where λ is an arbitrary nonzero complex constant. From this we can write down the components of the field F_{ab} as

$$F_{v\bar{\omega}} = F_{u\omega} = 0,$$
$$F_{u\bar{\omega}} = -\lambda F_{\omega\bar{\omega}}, \quad F_{v\omega} = \lambda^{-1}F_{\omega\bar{\omega}}, \tag{3.6}$$
$$F_{uv} = F_{\omega\bar{\omega}} = -[\gamma_\omega, \gamma_{\bar{\omega}}]_l.$$

An alternate promising method of studying the field equations (2.8), which has not yet been fully explored, is to change the order of solving them; namely first solve just (2.8a), i.e., as in (3.1a) with $\gamma^D{}_u$ and $\gamma^D{}_\omega$ in diagonal form, then substitute this into (2.8c), which can be solved for the γ_v in terms of the unknown $\gamma_{\bar{\omega}}$ and known $\gamma^D{}_u$ and $\gamma^D{}_\omega$; and finally use (2.8b) to determine the $\gamma_{\bar{\omega}}$.

More specifically, for SL(n,C), given $\gamma^D{}_u = \mathrm{diag}(u_i)$ and $\gamma^D{}_\omega = \mathrm{diag}(\omega_i)$, then from (2.8c), with $[\gamma_v] = v_{ij}$ and $[\gamma_{\bar{\omega}}] = \bar{\omega}_{ij}$, we have

$$v_{ij} = ((\omega_i - \omega_j)/(u_i - u_j))\bar{\omega}_{ij}, \tag{3.7}$$

for the *off-diagonal* terms. The diagonal terms of v_{ij} and $\bar{\omega}_{ij}$ are undetermined and are given freely. When (3.7) (with the diagonal terms) is substituted into (2.8b), i.e., into

$$[\gamma_v, \gamma_{\bar{\omega}}]_l = 0,$$

there are $n(n-1)$ quadratic equations to determine the $n(n-1)$ off-diagonal elements of $\bar{\omega}_{ij}$ (and hence v_{ij}) in terms of the *diagonal* elements of all the γ's. In the case of $n = 3$ when this procedure was carried out, four of the six equations were independent, thus determining only four of the six off-diagonal $\bar{\omega}_{ij}$; the other two being freely chosen. In this calculation, we assumed that the diagonal elements were chosen generically, avoiding special cases of vanishing coefficients. We thus can solve the generic SL(3,C) case.

Equivalent to the above method of solving (2.8) is the following: Again (2.8a) is solved by selecting $\gamma_u = \gamma^D{}_u$ and $\gamma_\omega = \gamma^D{}_\omega$. Eq. (2.8c) is solved by constructing the commutator of (2.8a) with an arbitrary matrix f (to be determined later). The Jacobi identity then leads to

$$[\gamma_u, [f, \gamma^D{}_\omega]_l]_l = [\gamma_\omega, [f, \gamma^D{}_u]_l]_l, \tag{3.8}$$

which yields, from (2.8c), that

$$\gamma_v = [f,\gamma^D{}_\omega]_l + \gamma^D{}_v \quad \text{and} \quad \gamma_{\bar\omega} = [f,\gamma^D{}_u]_l + \gamma^D{}_{\bar\omega}. \tag{3.9}$$

Note that we have added on the diagonal elements to γ_v and $\gamma_{\bar\omega}$, as they are not determined by (3.8). Finally, γ_ν and $\gamma_{\bar\omega}$ are substituted into $[\gamma_\nu,\gamma_{\bar\omega}]_l = 0$, i.e., (2.8b), yielding the equation for the determination of the *off-diagonal* terms of $[f] = f_{ij}$ (the diagonal terms can be taken as zero), namely

$$(u_i w_j - u_j w_i)\sum{}_k f_{ik}f_{kj} - (u_i - u_j)\sum{}_k w_k f_{ik}f_{kj}$$

$$+ (w_i - w_j)\sum{}_k u_k f_{ik}f_{kj} - f_{ij}[(u_i - u_j)(v_i - v_j)$$

$$- (w_i - w_j)(\bar w_i - \bar w_j)] = 0. \tag{3.10}$$

This again is easily solved in the generic SL(3,C) case.

B. Diffeomorphism algebras

We now consider a manifold $\mathfrak{M}$ of dimension n and the module $\mathfrak{m}$ of C^∞ vector fields, $\{\xi\}$, on it. In local coordinates, we have that

$$\xi = \xi^\mu \frac{\partial}{\partial x^\mu}.$$

An infinite-dimensional Lie algebra can be constructed on $\mathfrak{m}$ by defining the bracket as the Lie derivative, i.e.,

$$[\xi,\eta]_l \equiv \underset{\xi}{\mathcal{L}}\,\eta \equiv [\xi,\eta]. \tag{3.11}$$

We now ask for four vector fields, $\{\xi_1,\xi_2,\xi_3,\xi_4\} = \{\gamma_a\}$ that satisfy (2.8), yielding three *differential* equations,

$$[\gamma_u,\gamma_\omega] = 0, \tag{3.12a}$$

$$[\gamma_v,\gamma_{\bar\omega}] = 0, \tag{3.12b}$$

$$[\gamma_u,\gamma_\nu] - [\gamma_\omega,\gamma_{\bar\omega}] = 0. \tag{3.12c}$$

Equations (3.12a) and (3.12b) can be solved immediately in the following manner: For (3.12a) one can always introduce a coordinate system x^μ, so that γ_u and γ_ω are coordinate derivatives, i.e.,

$$\gamma_u = \frac{\partial}{\partial x^1} \quad \text{and} \quad \gamma_\omega = \frac{\partial}{\partial x^2}. \tag{3.13a}$$

Likewise, for (3.12b), we have in a different coordinate system x'^μ,

$$\gamma_v = \frac{\partial}{\partial x'^3} \quad \text{and} \quad \gamma_{\bar\omega} = \frac{\partial}{\partial x'^4}. \tag{3.13b}$$

The last equation, (3.12c), determines the coordinate transformation between the x'^μ and x^μ coordinates.

A second possible procedure is to again first solve (3.12a) by (3.13a) and note that (analogous to the finite-dimensional case) (3.13c) can be solved (using the Jacobi identity) by

$$\gamma_v = [f,\gamma_\omega] \quad \text{and} \quad \gamma_{\bar\omega} = [f,\gamma_u]. \tag{3.14}$$

Substituting (3.14) into (3.12b) yields a differential equation for the vector field f.

We illustrate these procedures in several special cases.

Case 1: The one (complex) dimensional diffeomorphism algebra, i.e., the Virasoro algebra, is trivial. We can take $\gamma_a = f_a\,\partial/\partial\zeta$ where $f_a = (f_u f_\omega f_v f_{\bar\omega})$ are four analytic functions of the complex variable ζ. The bracket $[\gamma_a,\gamma_b]$ is given by

$$[\gamma_a,\gamma_b] = f_a\frac{\partial f_b}{\partial\zeta} - f_b\frac{\partial f_a}{\partial\zeta} = f_a f_{b,\zeta} - f_b f_{a,\zeta},$$

and (3.12) becomes the system of differential equations for the unknown functions $f_a(\zeta)$,

$$\begin{aligned} &f_u f_{\omega,\zeta} - f_\omega f_{u,\zeta} = 0,\\ &f_v f_{\bar\omega,\zeta} - f_{\bar\omega} f_{v,\zeta} = 0,\\ &f_u f_{v,\zeta} - f_v f_{u,\zeta} = f_\omega f_{\bar\omega,\zeta} - f_{\bar\omega} f_{\omega,\zeta}. \end{aligned} \tag{3.15}$$

The first two of these yield

$$f_u = \lambda f_\omega, \quad f_v = \mu f_{\bar\omega}, \tag{3.16}$$

where λ and μ are two complex constants. The third equation of (3.7) yields

$$(\lambda\mu - 1)(f_\omega f_{\bar\omega,\zeta} - f_{\bar\omega} f_{\omega,\zeta}) = 0, \tag{3.17}$$

which again implies either γ_ω and $\gamma_{\bar\omega}$ are linearly independent and $\lambda\mu = 1$, or that there is at most one independent component of γ and a zero curvature. Thus the solution in the case of the Virasoro Lie algebra is similar to that of SL(2,C), and the relationships between the components of the Yang–Mills field are given by Eqs. (3.6).

Case 2: In this case, we will consider the special (volume preserving) diffeomorphisms in two dimensions. Equation (3.12a) can be solved, in general, by choosing two coordinate vectors, namely

$$\gamma_u = \frac{\partial}{\partial x^1} = \frac{\partial}{\partial u} \quad \text{and} \quad \gamma_\omega = \frac{\partial}{\partial x^2} = \frac{\partial}{\partial \omega}. \tag{3.13a'}$$

Rather than choosing (3.13b) as the solution to (3.12b), we will write

$$\gamma_v = v^i\partial/\partial x^i, \quad \gamma_{\bar\omega} = \bar\omega^i\partial/\partial x^i, \quad i = (1,2),$$

and substitute them into (3.12c), yielding

$$v^i_{,1} - \bar\omega^i_{,2} = 0,$$

which implies that

$$v^i = f^i_{,2}, \quad \bar\omega^i = f^i_{,1},$$

where f^i is an arbitrary vector function of x^i to be determined by (3.12b). If we now impose the (coordinate) divergence-free condition on v^i and $\bar\omega^i$, we can write $f^i = (\phi_{,2}, -\phi_{,1})$, which when substituted into (3.12b) yields

$$(\phi_{,11}\phi_{,22} - \phi_{,12}\phi_{,12})_{,i} = 0,$$

which in turn yields the "real" Monge–Ampère equation,

$$|\phi_{,ij}| = (\phi_{,11}\phi_{,22} - \phi_{,12}\phi_{,12}) = \text{constant}, \tag{3.18}$$

as the final equation for the "special" two-dimensional diffeomorphisms.

Case 3: In the case of the four-dimensional diffeomorphism group, it has been shown[8] that the field equations (2.5) or (2.8), when augmented by the condition that they be volume preserving (or equivalently, that the vector fields be divergence-free), become equivalent to the *anti-self-dual vacuum Einstein equations.*

More specifically, suppose that there is some nonvanishing four-form α, on a four-manifold, such that

$$\mathcal{L}_{\gamma_a}\alpha = 0, \tag{3.19}$$

for all four vector fields γ_a. Then satisfaction of (2.8) or (3.12) implies that the γ_a are proportional to a normalized null tetrad, i.e., $\sigma_a = f^{-1}\gamma_a$, with the scalar f defined by

$$f^2 = \alpha(\gamma_u, \gamma_\omega, \gamma_v, \gamma_{\bar\omega}), \tag{3.20}$$

so that the frame σ_a defines a metric $g = \sigma_u \otimes \sigma_v - \sigma_\omega \otimes \sigma_{\bar\omega}$, which satisfies the anti-self-dual Einstein equations.

Conversely, given an anti-self-dual space-time, there will always exist a null tetrad σ_a and a nonvanishing function f, such that $\gamma_a = f\sigma_a$ preserves some volume form α, and also satisfies (3.12).

Though the details will be given elsewhere, we mention that the field equations (3.12) and the volume preservation condition are equivalent to the following.

Defining T_a and V^a $(a = 1,2)$ by

$$\gamma_u = T_1, \quad \gamma_\omega = T_2, \quad \gamma_v = V^1, \quad \gamma_{\bar\omega} = V^2, \tag{3.21}$$

the field equations (3.12) become

$$[T_a, T_b] = 0, \quad [V^a, V^b] = 0, \quad [T_a, V^a] = 0. \tag{3.22}$$

Introducing the coordinate system (q^a and $Q^{a'}$, with a and $a' = 1,2$), the solutions can be written as

$$T_a = \frac{\partial}{\partial q^a}, \quad V^a = S^{aa'}\frac{\partial}{\partial Q^{a'}}, \tag{3.23}$$

with $S^{aa'}$ being the inverse matrix to

$$S_{aa'} = \frac{\partial^2 S}{\partial q^a \, \partial Q^{a'}}, \tag{3.24}$$

and with $S_{aa'}$ satisfying, this time, the "complex" Monge–Ampère equation, i.e.,

$$\det|S_{aa'}| = 1. \tag{3.25}$$

IV. THE RIEMANN–HILBERT SPLITTING PROBLEM

In the last section, we saw how the ASDYM equations could be studied and solved, at least for certain simple algebras, both finite and infinite dimensional, under the strong symmetry assumption. In this section, we will approach the problem from a different point of view, namely in terms of a Riemann–Hilbert splitting.

With this in mind, we review the Riemann–Hilbert approach to (2.3). A set of equations equivalent[5] to (2.3) is

$$F_{ab}L^{ab} = F_{ab}\overline{M}^{ab} = F_{ab}N^{ab} = 0, \tag{4.1}$$

where L^{ab}, $\overline{M}^{ab}$, and N^{ab} are any three independent self-dual antisymmetric tensors. Equation (4.1) follows from the orthogonality of self-dual and anti-self-dual tensors. A succinct version of (4.1) is

$$F_{ab}\overline{m}^{ab} = 0, \tag{4.2}$$

with $\overline{m}^{ab}$ a self-dual skew tensor written as

$$\overline{m}^{ab} = L^{ab} + \zeta\overline{M}^{ab} + \zeta^2 N^{ab}, \tag{4.3}$$

where ζ is an arbitrary point on the (complex) Riemann sphere, $\mathbf{C} + (\infty)$.

A normalized null tetrad, associated with the null coordinate system (2.6), is defined by

$$D = l^a\nabla_a = \partial_t + \partial_z = \partial_v,$$
$$\Delta = n^a\nabla_a = \partial_t - \partial_z = \partial_u,$$
$$\delta = m^a\nabla_a = -\partial_x + i\partial_y = -\partial_\omega, \tag{4.4}$$
$$\bar\delta = \overline{m}^a\nabla_a = -\partial_x - i\partial_y = -\partial_{\bar\omega},$$

with $l_a n^a = -m_a\overline{m}^a = 1$ and all other products vanishing. From this null tetrad, we define the following vectors:

$$L^a(\zeta) = l^a + \zeta m^a, \quad \overline{M}^a(\zeta) = \overline{m}^a + \zeta n^a. \tag{4.5}$$

We then take the skew tensor $L_{[a}\overline{M}_{b]}$ as $\overline{m}_{ab}$ in (4.3). This tensor, at any point P, defines a self-dual two surface through that point. As ζ ranges over the complex Riemann sphere, this tensor ranges over all self-dual totally null two-planes through that point. The vectors $L^a(\zeta)$ and $\overline{M}^a(\zeta)$ are two independent vectors in these planes. The set of all such two-surfaces in Minkowski space is (projective) twistor space **PT**, with coordinates $L = L_a x^a$, $\overline{M} = \overline{M}_a x^a$, and ζ.

For future reference, we note that

$$L = 2u + 2\zeta\bar\omega, \quad \overline{M} = 2\omega + 2\zeta v, \tag{4.6}$$

where u, v, ω, $\bar\omega$ are again defined by (2.6).

In a similar fashion, by contracting both sides of (4.5) with γ_a (instead of with x^a), we obtain

$$\gamma_a L^a \equiv \gamma_L = \gamma_v - \zeta\gamma_\omega, \quad \gamma_a\overline{M}^a \equiv \gamma_{\overline{M}} = -\gamma_{\bar\omega} + \zeta\gamma_u. \tag{4.7}$$

We now exhibit the linear differential equation (the Lax pair) for a function $G(x^a,\zeta)$ whose integrability conditions are the ASDYM equations:

$$L^a(\zeta)\nabla_a G = \gamma_L G, \quad \overline{M}^a(\zeta)\nabla_a G = \gamma_{\overline{M}} G. \tag{4.8}$$

The integrability conditions are precisely (4.2) with

$$\overline{m}^{ab}(\zeta) = L^{[a}\overline{M}^{b]}.$$

Knowledge of a solution $G(x^a,\zeta)$ allows one, directly from (4.8), to construct the γ_L and $\gamma_{\overline{M}}$ and hence the full set of γ_a's. If a solution, $G_0(x^a,\zeta)$, which is analytic in ζ in the neighborhood of $\zeta = 0$, is known, a second solution $G_\infty(x^a,\zeta)$ can be easily constructed in the following manner:

$$G_0(x^a,\zeta) = G_\infty(x^a,\zeta)P(L,\overline{M},\zeta), \tag{4.9}$$

where P is an arbitrary function of its three arguments. This result follows from the fact that the operators $L^a(\zeta)\nabla_a$ and $\overline{M}^a(\zeta)\nabla_a$ both annihilate $P(L,\overline{M},\zeta)$. With the proper choice of P, one can make $G_\infty(x^a,\zeta)$ analytic near $\zeta = \infty$.

This procedure of beginning with a given solution and finding a different solution via an arbitrary P can be reversed. One can begin with an arbitrary $P(L,\overline{M},\zeta)$ analytic in an annular region between $\zeta = 0$ and ∞ and try to find the two functions $G_\infty(x^a,\zeta)$ and $G_0(x^a,\zeta)$ that satisfy (4.9) or

$$P(L,\overline{M},\zeta) = G_\infty^{-1}(x^a,\zeta)G_0(x^a,\zeta), \tag{4.10}$$

so that $G_\infty(x^a,\zeta)$ and $G_0(x^a,\zeta)$ are analytic, respectively, around $\zeta = \infty$ and 0. Note that P is referred to as the patching matrix, and finding the two functions G is referred to as a matrix Riemann–Hilbert problem or Riemann–Hilbert splitting. A solution to this problem automatically solves the Lax pair and hence yields a solution to the ASDYM equations. Though there is no known method to accomplish this splitting for general P, there are large classes or choices of P where the splitting can be accomplished, e.g., P's that are either upper or lower triangular.

To see the relationship of the Riemann–Hilbert problem to the symmetry reduction of Sec. II, we note that when the connection components are independent of the space-time coordinates, the ASDYM equations (4.2) become

$$[\gamma_L,\gamma_{\bar{M}}]=0, \tag{4.11}$$

which are equivalent to (2.8). Solving the Lax pair with constant γ's satisfying (4.11), we obtain

$$G_0(x^a,\zeta)=\exp\{\gamma_L v-\gamma_{\bar{M}}\bar{\omega}\}g_0,$$
$$G_\infty(x^a,\zeta)=\exp\{-\gamma_L\omega\zeta^{-1}+\gamma_{\bar{M}}u\zeta^{-1}\}g_\infty, \tag{4.12}$$

with g_0 and g_∞ arbitrary matrix functions of L, $\bar{M}$, and ζ that are analytic, respectively, near $\zeta=0$ and ∞. From this, we see that the patching matrix must have the form

$$P(L,\bar{M},\zeta)=g_\infty^{-1}\exp\{\gamma_L\omega\zeta^{-1}-\gamma_{\bar{M}}u\zeta^{-1}\}\times\exp\{\gamma_L v-\gamma_{\bar{M}}\bar{\omega}\}g_0, \tag{4.13}$$

or since the γ_L and $\gamma_{\bar{M}}$ commute,

$$P(L,\bar{M},\zeta)=g_\infty^{-1}P_0(L,\bar{M},\zeta)g_0, \tag{4.14}$$

with [using (4.6)]

$$P_0(L,\bar{M},\zeta)=\exp\tfrac{1}{2}\{\gamma_L\bar{M}\zeta^{-1}-\gamma_{\bar{M}}L\zeta^{-1}\}.$$

Though a patching matrix of this form always splits into (4.12), it is of no use, since one must already know the commuting pair $\gamma_L,\gamma_{\bar{M}}$ in order to write (4.14). We would have to find a more general form for P, a form that is translational invariant up to gauge, i.e., has the following property.

If the $(u,v,\omega,\bar{\omega})$, in the defining equations for the L and $\bar{M}$, are replaced by $(u',v',\omega',\bar{\omega}')=(u,v,\omega,\bar{\omega})+(a,b,c,\bar{c})$, i.e., undergo a translation, then $L\to L'=L+2(a+\zeta\bar{c})$ and $\bar{M}\to\bar{M}'=\bar{M}+2(c+\zeta b)$. We then have that translational invariance, up to gauge, is defined by

$$P(L',\bar{M}',\zeta)=g_0^{-1}P(L,\bar{M},\zeta)g_\infty, \tag{4.15}$$

where g_0 and g_∞ are holomorphic functions of L, $\bar{M}$, and ζ, analytic, respectively, near ζ equals zero and infinity. Unfortunately, though from the general theory one knows that P's satisfying (4.15) exist, it is not known how to find or construct them.

(An alternate approach to this problem of finding a patching matrix for ASDYM with symmetries has been developed by Mason and Woodhouse[9] and applied successfully to stationary-axial symmetry. Again, unfortunately, it has not been applied to our case of maximal symmetry.)

ACKNOWLEDGMENTS

We would like to thank Lionel Mason for valuable conversations.

S.K. wishes to acknowledge support by a University Research Council Grant (No. 701) from Youngstown State University. E.N. thanks the NSF for support under Grant No. PHYS 8803073, and S.C. thanks the University of Colorado for support.

[1] P. Forgacs, in *Non-Linear Equations in Classical Quantum Field Theory*, edited by N. Sanchez (Springer, Berlin, 1985).
[2] A. A. Belavin and V. E. Zakharov, Phys. Lett. B **73**, 53 (1978).
[3] K. Pohlmeyer, Commun. Math. Phys. **72**, 37 (1980).
[4] M. K. Prasad, A. Sinha, and L. L. Wang, Phys. Lett. B **87**, 237 (1979).
[5] L. Mason, S. Chakravarty, and E. T. Newman, J. Math. Phys. **29**, 1005 (1988).
[6] R. Ward, Commun. Math. Phys. **80**, 563 (1981).
[7] G. Sparling, unpublished.
[8] L. Mason and E. T. Newman, Commun. Math. Phys. **121**, 659 (1989).
[9] N. M. J. Woodhouse and L. J. Mason, Non-Linearity **1**, 73 (1988).

Singularly perturbed Chern–Simons theory

Charles Nash
Department of Mathematical Physics, St. Patrick's College, Maynooth, Ireland

(Received 22 September 1989; accepted for publication 25 April 1990)

The Chern–Simons theory of an SU(2) gauge theory in three dimensions is looked at from a perturbative point of view. The pure Chern–Simons action is generalized by adding a conventional Yang–Mills action term. This acts as a singular perturbation. The resulting theory has a moduli space containing that of the pure Chern–Simons version; for certain discrete values of the perturbation parameter lying in the spectrum of an appropriate elliptic operator the enlargement of the moduli space can be made explicit. The extrema can be classified by a Hessian with a finite index and nullity without recourse to spectral flow. Corrections to the resultant quantum theory are also calculated. Also, the quantum theory of the present model should be better behaved than in the unperturbed case.

I. INTRODUCTION

We look at the three-dimensional Chern–Simons theory from a different point of view. This leads to advantages in the mathematics and physics. On th mathematical side we shall see that there is no need to employ the notion of spectral flow. On the physical side we shall have a quantum theory that we expect to have improved properties.

II. THEORY

In three dimensions Yang–Mills theories can be constructed in several important ways: One can take the point of view that the three-dimensional theory is a dimensionally reduced four-dimensional theory, or one can work directly in three dimensions. To do the former choose the four-dimensional theory to be a pure Yang-–Mills theory in $\mathbf{R}^4$, with the Lagrangian L given by

$$L = -\operatorname{tr}(F \wedge {}^*F), \tag{1}$$

where F is the curvature form of a connection A and tr denotes the trace defined by the Killing form on the Lie algebra, which we take in this paper to be su(2). Then demanding that the connection A be independent of time, one obtains a Yang–Mills–Higgs system with the three-dimensional interpretation that the finite-energy solutions to the equations of motion are monopoles in $\mathbf{R}^3$ (see Ref. 1); if one replaces invariance under time translations by invariance under rotations one obtains *hyperbolic monopoles*.[2–4]

However, one can start in three dimensions and use the Lagrangian

$$L = -\operatorname{tr}(\mathbf{F} \wedge {}^*\mathbf{F}), \tag{2}$$

where $\mathbf{F}$ denotes the curvature form of a connection $\mathbf{A}$ defined on an SU(2) bundle over a three-manifold M; this three-manifold has a Riemannian metric g_{ij} and the Hodge * is with respect to this metric. Now if the manifold M is taken to be $\mathbf{R}^3$ and the action S is given by the usual expression

$$S = -\operatorname{tr}\int_{\mathbf{R}^3} (F \wedge {}^*\mathbf{F}), \tag{3}$$

then the critical points of the action are given by the usual equation

$$d_{\mathbf{A}} {}^*\mathbf{F} = 0, \tag{4}$$

where $d_{\mathbf{A}}$ denotes the covariant exterior derivative with respect to the connection $\mathbf{A}$. However, there are no finite-action solutions to Eq. (4) except $\mathbf{F} = 0$, which for these circumstances is trivial. Throughout this paper we shall be interested in finite-action theories and so we can improve things by modifying either the action S or the manifold $\mathbf{R}^3$—in fact, in the end we shall do both.

If we just substitute a compact, closed three-manifold M for $\mathbf{R}^3$, then the zero-curvature solutions $\mathbf{F} = 0$ are no longer always trivial. Instead, they depend on the holonomy of the connection $\mathbf{A}$ around closed loops and this is classified by a group homomorphism from the fundamental group into the gauge group SU(2); thus such flat connections are classified generally by the space of all such homomorphisms, which we write as $\operatorname{Hom}(\pi_1(M), \mathrm{SU}(2))$. Rather than considering this situation we wish to modify the action by adding a Chern–Simons term to the action, so that it becomes

$$S = -\operatorname{tr}\int_M \left\{(\mathbf{F} \wedge {}^*\mathbf{F}) + \alpha\left(d\mathbf{A} \wedge \mathbf{A} + \frac{2}{3}\mathbf{A} \wedge \mathbf{A} \wedge \mathbf{A}\right)\right\}, \tag{5}$$

where α is a constant. This results in the topologically massive gauge theories of Refs. 5–7. In such theories the coefficient α has the dimensions of a mass whch must be quantized in order to render the amplitude $\exp[iS]$ single-valued; cf., also, Refs. 8–10.

Recently, a bold step has been taken by Floer (cf. Ref. 11 and references therein) which is to drop the conventional Yang–Mills term $\operatorname{tr}(\mathbf{F} \wedge {}^*\mathbf{F})$ and to regard the Chern–Simons term as an action in its own right and to study its critical points. To this end we set

$$S = S_{\mathrm{CS}} = \frac{1}{8\pi^2}\operatorname{tr}\int_M \left\{d\mathbf{A} \wedge \mathbf{A} + \frac{2}{3}\mathbf{A} \wedge \mathbf{A} \wedge \mathbf{A}\right\}, \tag{6}$$

whose critical points are given by

$$\mathbf{F} = 0. \tag{7}$$

Thus we know that the space of critical points consists just of the flat connections and we are led to consider the space $\operatorname{Hom}(\pi_1(M), \mathrm{SU}(2))$ of representations of $\pi_1(M)$, SU(2);

 0022-2488/90/092258-05$03.00

we shall restrict ourselves to (nontrivial) irreducible representations. We observe that if we restrict ourselves to M's, which are homology three-spheres, so that $H_1(M;\mathbf{Z})=0$, then the "Abelian part" of $\pi_1(M)$ is trivial; now, because an Abelian $\pi_1(M)$ can only be represented reducibly in SU(2) this has the consequence that the only nontrivial representations will be the irreducible ones.

A further point is that the group SU(2) acts on Hom $(\pi_1(M),\ \mathrm{SU}(2))$ by conjugation and representations that differ by this adjoint action are equivalent; thus we form the quotient

$$\mathrm{Hom}(\pi_1(M),\mathrm{SU}(2))/\mathrm{Ad}(\mathrm{SU}(2)), \tag{8}$$

from which we delete the trivial representation, leaving us with the irreducible ones.

Now the curvature $\mathbf{F}\equiv\mathbf{F}(\mathbf{A})$ can be regarded as a function on the space $\mathscr{A}$ of all connections whose *zeros* give the critical points of the Chern–Simons action S_{CS}. [Actually, for a fixed $\mathbf{A}$, $\mathbf{F}(\mathbf{A})$ is a one-form on the space $\mathscr{A}$, where we are thinking of a one-form as being a linear functional on the tangent space $T_{\mathbf{A}}\mathscr{A}\simeq\Omega^1(M)\times\mathrm{su}(2)$; we then denote the action of $\mathbf{F}(\mathbf{A})$ on an arbitrary $\mathbf{a}\in T_{\mathbf{A}}\mathscr{A}$ by $\mathbf{F}_a(\mathbf{A})$, where $\mathbf{F}_a(\mathbf{A})=\int_M \mathrm{tr}(\mathbf{a}\wedge\mathbf{F}(\mathbf{A}))$.] However, the zeros of $\mathbf{F}$ are gauge-invariant quantities and thus we should pass to the space of gauge orbits given by the quotient $\mathscr{A}/\mathscr{G}$, where $\mathscr{G}$ is the group of gauge transformations. An interesting technical matter here is that the Chern–Simons action is not gauge invariant and hence is not a single-valued function on the orbit space $\mathscr{A}/\mathscr{G}=\mathscr{C}$, say; however, under a gauge transformation g, which can be regarded as a map $g:M\to\mathrm{SU}(2)$, S_{CS} always changes by an integer k, where k is the degree of g. Thus S_{CS} actually takes values in $\mathbf{R}/\mathbf{Z}$; this is one way of regarding S_{CS}. Alternatively, and equivalently, one can pass to the covering space $\mathscr{C}_0$ of $\mathscr{C}$ on which S_{CS} is a single-valued function.

Given the zeros of $\mathbf{F}$ one can imitate a standard construction in ordinary differential topology and construct a "Euler characteristic" for $\mathbf{F}$ by taking the signed sum of the zeros provided that this sum converges in this infinite-dimensional context. This imitation is successful and the resulting integer is the "Euler characteristic" for the Floer homology of M.

This homology theory is constructed[11] using the properties of the critical points of the function S_{CS}; crucial in this construction is the role played by the Hessian of a critical point. In the present situation the Hessian is a differential operator with real eigenvalues and one would like to define the number of negative eigenvalues to be the index of a critical point. However, it is easy to calculate the Hessian and discover it to be the differential operator $2*d_{\mathbf{A}}$. One then encounters the problem that this operator has no lower bound, so that the index of a critical point would in general be infinite. Floer gets around this problem in his homology construction because he only needs to define the *difference* between the index of a pair of critical points; this is a quantity which can be "renormalized" by using the spectral flow of the Hessian along a gradient flow path in $\mathscr{A}$ connecting the pair of critical points.

In this paper we would like to replace the function S_{CS} by the function $\tilde{S}$, where

$$\tilde{S}=kS_{\mathrm{CS}}+\lambda\,\mathrm{tr}\int_M(\mathbf{F}\wedge{*}\mathbf{F}), \tag{9}$$

with $k\neq 0$ an integer and M any orientable closed three-manifold. Thus for nonzero λ we have essentially a perturbation of the Chern–Simons function; for reasons that will be made more precise below we will refer to $\tilde{S}$ as a singular perturbation of the Chern–Simons function. The critical points of $\tilde{S}$ are, of course, not the same as those of S_{CS} and indeed, the equation for the critical points is more complicated, being given by

$$(k/8\pi^2)\mathbf{F}(\mathbf{A})+\lambda d_{\mathbf{A}}{*}\mathbf{F}(\mathbf{A})=0. \tag{10}$$

It is still true that flat connections or irreducible representations of the fundamental group give critical points; however, the space of solutions requires some analytic investigation in the present case since there are nonflat critical points. Also, we do not assume M to be a homology sphere, so that there can be a moduli space of critical points.

With $\lambda\neq 0$ the equation for the critical points of $\tilde{S}$ is a partial differential equation of second order, whereas that for S_{CS} is only of *first* order. Thus as $\lambda\to 0$, $\tilde{S}$ can be viewed as a singular perturbation of S_{CS}. We wish to investigate the neighborhood of a general critical point, say $\mathbf{A}$, and also to classify it by assigning it some kind of index. To this end consider a path of connections through $\mathbf{A}$ given by $\mathbf{A}(t)$. For small t we can write $\mathbf{A}(t)=\mathbf{A}+t\mathbf{a}+\cdots$, with $\mathbf{a}\in\Omega^1(M)\times\mathrm{su}(2)$. Then we wish to calculate the tangent space to the moduli space: Using the path $\mathbf{A}(t)$ and Eq. (10) we find that this is given in part by

$$d_{\mathbf{A}}^{*}\,d_{\mathbf{A}}\mathbf{a}+{*}({*}\mathbf{F}\wedge\mathbf{a})+k/8\pi^2\lambda\,{*}d_{\mathbf{A}}\mathbf{a}=0. \tag{11}$$

However, we also have to project Eq. (11) onto a gauge orbit. To do this we make use of the fact that when we work in the space of connections $\mathscr{A}$ and construct the tangent space $T_{\mathbf{A}}\mathscr{A}$, then this space decomposes into a direct sum of the tangent space comprising directions within the orbit $\mathbf{A}_g$ through $\mathbf{A}$ plus its orthogonal complement. More precisely, we have

$$T_{\mathbf{A}}\mathscr{A}=T_{[\mathbf{A}_g]}\mathscr{A}\oplus\ker d_{\mathbf{A}}^{*} \tag{12}$$

and $d_{\mathbf{A}}^{*}$ is the adjoint of the operator $d_{\mathbf{A}}$: $\Omega^0(M)\times\mathrm{su}(2)\to\Omega^1(M)\times\mathrm{su}(2)$. Thus in order to be properly gauge invariant we must project from $T_{\mathbf{A}}\mathscr{A}$ onto $\ker d_{\mathbf{A}}^{*}$. The tangents to the moduli space must also therefore be realized as a quotient by the space $T_{[\mathbf{A}_g]}\mathscr{A}$; its full description is that it is given by those $\mathbf{a}\in\Omega^1(M)\times\mathrm{su}(2)$ that satisfy

$$d_{\mathbf{A}}^{*}\,d_{\mathbf{A}}\mathbf{a}+{*}({*}\mathbf{F}\wedge\mathbf{a})+k/8\pi^2\lambda\,{*}d_{\mathbf{A}}\mathbf{a}=0,\quad d_{\mathbf{A}}^{*}\mathbf{a}=0 \tag{13}$$

but because we are working effectively with $d_{\mathbf{A}}^{*}=0$ we can modify this to the *elliptic* form

$$d_{\mathbf{A}}^{*}d_{\mathbf{A}}\mathbf{a}+d_{\mathbf{A}}d_{\mathbf{A}}^{*}\mathbf{a}+{*}({*}\mathbf{F}\wedge\mathbf{a})+k/8\pi^2\lambda\,{*}d_{\mathbf{A}}\mathbf{a}=0,\quad d_{\mathbf{A}}^{*}\mathbf{a}=0. \tag{14}$$

Now, since $d_{\mathbf{A}}^{*}d_{\mathbf{A}}\mathbf{a}+d_{\mathbf{A}}d_{\mathbf{A}}^{*}\mathbf{a}$ is elliptic it has a finite-dimensional kernel on M and thus the moduli space is finite dimen-

sional, having a finite-dimensional tangent space. This dimension, say D, depends on λ, but will in general be positive and indeed, we have the inequality

$$D \geqslant \dim \mathrm{Hom}(\pi_1(M), \mathrm{SU}(2))/\mathrm{Ad}(\mathrm{SU}(2)). \tag{15}$$

Moreover, for appropriate λ we can be certain that inequality (15) is a strict one. To see this we assume that $\mathbf{A}$ is flat, so that the tangent directions are given by those $\mathbf{a} \in \ker d_{\mathbf{A}}^*$ satisfying

$$*d_{\mathbf{A}}\mathbf{b} + (k/8\pi^2\lambda)\mathbf{b} = 0, \tag{16}$$

where $\mathbf{b} = *d_{\mathbf{a}}\mathbf{a}$. Then since this is an eigenvalue problem for $*d_{\mathbf{A}}$, we know that there is a solution when $k/(8\pi^2\lambda)$ belongs to its spectrum: Denoting the spectrum of $*d_{\mathbf{A}}$ by $\{\ldots, \mu_{-1}, \mu_0, \mu_1, \ldots\}$, then if $\mathbf{e}_n \in \Omega^1(M) \times \mathrm{su}(2)$ has the eigenvalue $\mu_n \neq 0$, we require λ to satisfy $8\pi^2\lambda = -k/\mu_n$ and $\mathbf{a}$ to satisfy $*d_{\mathbf{A}}\mathbf{a} = c\mathbf{e}_n$, with c a constant. Thus $\mathbf{a}$ has the eigenvalue μ_n. Of course, the eigenvalue μ_n may be degenerated, but if $\Gamma_n \geqslant 1$ is its degeneracy, then the moduli space will have increased its dimension by at least this amount.

A further property of $\tilde{S}$ to examine is its Hessian at the critical points. The Hessian is obtained by expanding $\tilde{S}(\mathbf{A} + t\mathbf{a})$ in t:

$$\tilde{S}(\mathbf{A} + t\mathbf{a}) \equiv \tilde{S}(t) = \tilde{S}(0) + t\frac{d\tilde{S}(0)}{dt} + \frac{t^2}{2}\frac{d^2\tilde{S}(0)}{dt^2} + \cdots. \tag{17}$$

Having made expansion (17) we define the Hessian H by writing

$$\frac{d^2\tilde{S}(0)}{dt^2} = \langle \mathbf{a}, H\mathbf{a} \rangle, \tag{18}$$

so that

$$H = 2\{\lambda(d_{\mathbf{A}}^* d_{\mathbf{A}} + 2*(*\mathbf{F} \wedge \cdot)) + (k/8\pi^2)*d_{\mathbf{A}}\}. \tag{19}$$

Thus $H: \Omega^1(M) \times \mathrm{su}(2) \to \Omega^1(M) \times \mathrm{su}(2)$ is a partial differential operator which we must restrict to $\ker d_{\mathbf{A}}^*$ in order to obey correctly the requirements of gauge invariance. Now, on $\ker d_{\mathbf{A}}^*$, the combination $\lambda d_{\mathbf{A}}^* d_{\mathbf{A}}$ is a non-negative Hermitian elliptic operator and a standard coercivity argument[12] can then be applied to deduce that H is bounded below and positive if λ is large enough. Thus we can define the index of a critical point to be the dimension of the largest subspace of $\ker d_{\mathbf{A}}^*$ on which $H \equiv H(\mathbf{A})$ is negative definite. It follows that all the critical points of $\tilde{S}$ have *finite index*. On the other hand, if in H we allow $\lambda \to 0$, then we revert to the pure Chern–Simons case and the index of all critical points becomes infinite. We have seen already, though, that the Hessian in this case is $2*kd_{\mathbf{A}}$, so that H is indeed a singular perturbation of $2*kd_{\mathbf{A}}$; this accounts for the difference between the spectral behavior in the two cases.

We can view our model as being a regularization of the pure Chern–Simons model—a regularization that is λ dependent and which also can depend on the metric used. Notice, also, that $\tilde{S}$ is also the action for the topologically massive gauge theory. Thus our model provides us with a physically nontrivial interesting model in which to study the phenomena of the Chern–Simons theory, as well as providing us with a small finite-dimensional Hilbert space of moduli.

We proceed now to examine some properties of the quantum theory based on the action $\tilde{S}$. This leads us to consider the partition function

$$Z = \int \mathscr{D}\mathscr{A} \exp[2\pi i\tilde{S}], \tag{20}$$

where $\mathscr{D}\mathscr{A}$ denotes some appropriately chosen integration measure and the 2π is inserted for single-valuedness. Note that in the pure Chern–Simons case, where $\lambda = 0$, the action is *odd* in the gauge fields. This leads to problems in the functional integral since the action is not bounded below. This sort of problem also occurs in simpler theories such as the scalar ϕ^3 theory, which is similarly afflicted. Here, provided that $\lambda \neq 0$, we do not have this difficulty; this is part of the benefit of studying this model. We wish to work in a limit where the stationary phase approximation is applicable. To do this we can take the limit where $k \to \infty$ and λ is fixed; other limits in which λ increases are also of interest, but we shall look at these elsewhere. In any case, in the former limit, which corresponds to weak coupling, we write

$$Z = \int \mathscr{D}\mathscr{A} \exp\left\{\frac{ik}{4\pi}\left[\mathrm{tr}\int_M d\mathbf{A} \wedge A + \frac{2}{3}\mathbf{A} \wedge \mathbf{A} \wedge \mathbf{A} + \frac{8\pi^2\lambda}{k}\,\mathrm{tr}\int_M (\mathbf{F} \wedge *\mathbf{F})\right]\right\}. \tag{21}$$

Then as $k \to \infty$ the partition function should be dominated by those configurations that have stationary phase and obey the extremal equation

$$(1/8\pi^2)\mathbf{F}(\mathbf{A}) + (\lambda/k)d_{\mathbf{A}}*\mathbf{F}(\mathbf{A}) = 0. \tag{22}$$

Thus in this limit we expand about these extremal connections, say $\mathbf{A}_E$, and obtain a large k limiting form Z_L of Z, which we write as the expression

$$Z_L = \sum_{\mathbf{A}_E} \exp[2\pi i k\tilde{S}(\mathbf{A}_E)]F(\mathbf{A}_E), \tag{23}$$

where $\mathbf{F}(\mathbf{A}_E)$ is a factor that is about to be calculated: To do the integration over the space $\mathscr{A}$ we have to fix a gauge so as to integrate correctly over the orbits. In view of our above discussion of the orbits the natural gauge choice is

$$d_{\mathbf{A}}^*\mathbf{a} = 0, \tag{24}$$

where we have set $\mathbf{A} = \mathbf{A}_E + \mathbf{a}$. Let $L\mathscr{G}$ denote the Lie algebra of the group of gauge transformations, so that $L\mathscr{G} = T_e\mathscr{G}$, where e is the identity gauge transformation; and as a space, $L\mathscr{G} \simeq \Omega^0(M) \times \mathrm{su}(2)$. With this completed the gauge fixing is carried out by enlarging $T_{\mathbf{A}}\mathscr{A}$ to $T_{\mathbf{A}}\mathscr{A} \oplus T_e\mathscr{G}$ and then projecting onto $\ker d_{\mathbf{A}}^*$. We also have that $T_{\mathbf{A}}\mathscr{A} \simeq \mathrm{Im}\, d_{\mathbf{A}} \oplus \ker d_{\mathbf{A}}^*$, so that the ghosts are represented by the space $L\mathscr{G} \oplus \mathrm{Im}\, d_{\mathbf{A}}$. The gauge fixed action with its attendant ghosts then results in a new form for the limiting partition function Z_L:

$$Z_L = \sum_{A_E} \exp\left[2\pi i k \tilde{S}(A_E)\right] \int \mathcal{D}a \mathcal{D}b \mathcal{D}g \mathcal{D}\bar{g}$$
$$\times \exp\left[\frac{ik}{4\pi}\left\{\left\langle a, \left\{ *d_A + \frac{8\pi^2\lambda}{k}(d_A^* d_A + 2*(*F\wedge\cdot)\right\} a\right\rangle + 2\langle b, d_A^* a\rangle + \langle \bar{g}, (d_A^* d_A + d_A d_A^*) g\rangle\right\}\right]. \tag{25}$$

In Eq. (25) the inner product has the usual definition $\langle \omega, \nu\rangle = \int_M \mathrm{tr}(\omega \wedge *\nu)$, with $\omega, \nu \in \Omega^p(M)\times su(2)$. Also, g represents an anticommuting ghost field belonging to $\Omega^0(M)\times su(2)$ and b is necessarily an element of $\Omega^3(M)\times su(2)$, which implements the gauge condition. Thus the integration is over the space

$$(\Omega^0(M)\oplus\Omega^1(M)\oplus\Omega^3(M))\times su(2). \tag{26}$$

To carry out the integration we single out the subspace $(\Omega^1(M)\oplus\Omega^3(M))\times su(2)$ and note that we can write

$$\langle a, Qa\rangle + 2\langle b, d_A^* a\rangle = \langle v, Pv\rangle,$$

with

$$v = \begin{bmatrix} a \\ b \end{bmatrix}, \quad P = \begin{bmatrix} Q & d_A* \\ d_A* & 0 \end{bmatrix},$$
$$Q = (8\pi^2\lambda/k)(d_A^* d_A + 2*(*F\wedge\cdot)) + *d_A. \tag{27}$$

Assembling these facts gives us the result that the partition function Z_L is given by the expression

$$\sum_{A_E} \exp[2\pi i k \tilde{S}(A_E)] \frac{\det(d_A^* d_A + d_A d_A^*)}{\sqrt{\det(P)}},$$

so that

$$F(A_E) = \frac{\det(d_A^* d_A + d_A d_A^*)}{\sqrt{\det(P)}}. \tag{28}$$

For convenience in further elucidating the expression for $F(A_E)$ we introduce the notation $\Delta_p(A)$, which stands for the *covariant* Laplacian on $\Omega^p(M)\times su(2)$, i.e., $\Delta_p(A)$ is the operator $(d_A^* d_A + d_A d_A^*):\Omega^p(M)\times su(2)\to\Omega^p(M)\times su(2)$. We can then realize $\det(P)$ as $\sqrt{\det(P^2)}$ and if we let $h = 8\pi^2(d_A^* d_A + 2*(*F\wedge\cdot))$ we find that

$$P^2 = \begin{bmatrix} Q^2 + d_A *d_A * & Qd_A * \\ d_A *Q & \Delta_3(A) \end{bmatrix}$$
$$= \begin{bmatrix} (\lambda^2/k^2)h^2 + (\lambda/k)(h*d_A + *d_A h) + \Delta_1(A) & Qd_A* \\ d_A*Q & \Delta^3(A) \end{bmatrix}. \tag{29}$$

Actually, the off-diagonal terms in (29) can be shown to contribute zero; thus we obtain

$$F(A_E) = \frac{\det(\Delta_0(A))}{\det((\lambda^2/k^2)h^2 + (\lambda/k)(h*d_A + *d_A h) + \Delta_1(A))\det(\Delta_3(A))^{1/4}}. \tag{30}$$

Also, using Hodge duality to relate the various $\Delta_p(a)$ we find $\det(\Delta_3(A)) = \det(\Delta_0(A))$ and so we now have

$$Z_L = \sum_{A_E} \exp[2\pi i k \tilde{S}(A_E)] \frac{\det(\Delta_0(A))^{3/4}}{\det((\lambda^2/k^2)h^2 + (\lambda/k)(h*d_A + *d_A h) + \Delta_1(A))^{1/4}}. \tag{31}$$

Now we employ a formal rewriting of the determinant in the denominator of expression (31) for Z_L as

$$\det(\Delta_1(A))\det\left(I + \frac{(\lambda^2/k^2)h^2 + (\lambda/k)(h*d_A + *d_A h)}{\Delta_1(A)}\right). \tag{32}$$

This allows us to write

$$Z_L = \sum_{A_E} \exp[2\pi i k \tilde{S}(A_E)] \frac{\det(\Delta_0(A))^{3/4}}{\det(\Delta_1(A))^{1/4}} f(\lambda),$$

where

$$f(\lambda) = \exp\left[\frac{-1}{4}\,\mathrm{tr}\ln\left(I + \frac{(\lambda^2/k^2)h^2 + (\lambda/k)(h*d_A + *d_A h)}{\Delta_1(A)}\right)\right]. \tag{33}$$

Thus if we denote by Z_L^{CS} the pure Chern–Simons contribution to Z_L, which is, of course, a phase times the Ray–Singer analytic torsion $T(M)$,[13–15] then formula (33) allows us to calculate corrections:

$$Z_L = Z_L^{CS}(1 + c_0\lambda + \cdots),$$
$$c_0 = (-1/4k)\,\mathrm{tr}((h*d_A + *d_A h)/\Delta_1(A)). \tag{34}$$

In the analysis above the summation Σ_{AE} can be a sum over discrete configurations or an integral over the moduli space of extrema: If it is the latter, then $Q^2 + d_A *d_A *$ has zero modes which must be projected out before calculating its determinant; alternatively, one can perturb λ slightly so as to avoid them. Also, the determinants of each elliptic operator O are calculated using their associated zeta function $\zeta_O(s)$ and, because they are real and positive, there is no need to calculate a phase, as is necessary in the pure Chern–Simons theory cf. the breakthrough made by Witten[13] in establishing the connection between Chern–Simons theory and the Jones polynomial.

III. CONCLUDING REMARKS

In conclusion, we point out that to study the present model a certain price has to be paid; we have in mind the fact that in order to introduce our perturbation, a metric has to be introduced, while in the unperturbed case no metric is needed at the defining stage. However, in the pure Chern–Simons case a metric is also needed to introduce the spectral flow and to do the Fadeev–Popov gauge fixing. In the end one can show that some of the data, such as the Ray–Singer torsion, are nevertheless metric independent; in the perturbed case there may also be metric-independent features. In addition, we find that the quantum theory of the present model is both calculable and should be better behaved for large fluctuations in the gauge field. A further avenue of investigation is provided by considering M, which are, at least locally, of the form $\Sigma \times \mathbf{R}$, where Σ is a Riemann surface. This effectively allows us to employ results and techniques from two-dimensional conformal and Yang–Mills theories. This will be reported on elsewhere.

[1] A. Jaffe and C. H. Taubes, *Vortices and Monopoles* (Birkhäuser, Boston, 1980).
[2] M. F. Atiyah, Commun. Math. Phys. **93**, 437 (1984).
[3] C. Nash, J. Math. Phys. **27**, 2160 (1986).
[4] A. Chakrabarti, J. Math. Phys. **27**, 340 (1986).
[5] R. Jackiw and S. Templeton, Phys. Rev D **23**, 2291 (1981).
[6] J. Schonfeld, Nucl. Phys. B **185**, 157 (1981).
[7] S. Deser, R. Jackiw, and S. Templeton, Phys. Rev. Lett. **48**, 975 (1982).
[8] G. V. Dunne, R. Jackiw, and C. A. Trugenberger, MIT preprint, CTP #1711 (1989).
[9] R. Jackiw and S. Templeton, Phys. Rev. D **23**, 2291 (1981).
[10] P. Horváthy and C. Nash, Phys. Rev. D **33**, 1822 (1986).
[11] A. Floer, Commun. Math. Phys. **118**, 215 (1988).
[12] F. Treves, *Basic Linear Partial Differential Equations* (Academic, New York, 1975).
[13] E. Witten, Commun. Math. Phys. **121**, 351 (1989).
[14] A. S. Schwarz, Lett. Math. Phys. **2**, 247 (1978).
[15] D. B. Ray and I. M. Singer, Adv. Math. **7**, 145 (1971).

The Fock representation of osp($\infty|\infty$) and the orthogonal symplectic super KP hierarchy

Kaoru Ikeda
Department of Mathematics, Tokyo Metropolitan University, 2-1-1 Fukasawa, Setagaya-ku, Tokyo 158, Japan

(Received 12 October 1989; accepted for publication 25 April 1990)

The Fock representation of the Lie superalgebra osp($\infty|\infty$), from which is derived the super boson–fermion correspondence of osp($\infty|\infty$) is discussed. Solutions of the OSp–SKP hierarchy in terms of the neutral super free fermions are constructed.

I. INTRODUCTION

In this paper we investigate a relationship between a representation of the infinite-dimensional Lie superalgebra osp($\infty|\infty$) and the orthogonal symplectic super KP (OSp–SKP) hierarchy. In the investigation of the super Toda lattice hierarchy, we have found osp($\infty|\infty$) as a symmetry of the equation.[1,2] On the other hand, osp($\infty|\infty$) emerges in the theory of the SKP hierarchy.[3] We are now interested in the Lie superalgebra osp($\infty|\infty$) itself and its representations. Kac and van de Leur studied the representation theory of the Lie superalgebra $a_{\infty|\infty}$ (Ref. 4) and $b_{\infty|\infty}$ (Ref. 5). We apply their theory of $a_{\infty|\infty}$ to B-type Lie superalgebra osp($\infty|\infty$). Then we get the formula of boson–fermion correspondence of osp($\infty|\infty$). Through this paper we employ $\mathbb{Z}$ as the indices of osp($\infty|\infty$) and the theory of its representation. On the other hand, in Ref. 5, they use $\frac{1}{2}\mathbb{Z}$ for indices. By the transposition of indices, we see that the theory of osp($\infty|\infty$) and its representation in this paper corresponds to the theory of $b_{\infty|\infty}$.

For the representation of osp($\infty|\infty$), we deal with the superalgebra SBCL generated by the "neutral super free fermions" $\phi^{\mu}{}_{n}$ ($n\in\mathbb{Z}, \mu=0,1$) and the unit 1. And we construct the fock spaces F^{s} and F^{s^*} from SBCL. Let $\mathcal{S}$ be a space of superfields with (anti) commutative variables. We denote the tensor product of SBCL and $\mathcal{S}$ by $\tilde{\text{SBCL}}$. The supergroup $G(\tilde{V},\tilde{V})$ is defined by $G(\tilde{V},\tilde{V}) = \{g\in\tilde{\text{SBCL}} | g\tilde{V}g^{-1}=\tilde{V}\}$, where $\tilde{V}$ is the space of linear combination of $\phi^{\mu}{}_{n}$ over $\mathcal{S}$. Then $g\in G(\tilde{V},\tilde{V})$ defines the elements $g|0\rangle\in\tilde{F}^{s}$ and $\langle 0|g\in\tilde{F}^{s^*}$, where $\tilde{F}^{s}$ and $\tilde{F}^{s^*}$ are Fock spaces defined by $\tilde{\text{SBCL}}$. We call a basis vector of $\tilde{F}^{s}$ and of $\tilde{F}^{s^*}$ a state vector. By definition of $G(\tilde{V},\tilde{V})$, the state vector $\phi^{0}{}_{0}g|0\rangle$ (resp. $\langle 0|g\phi^{0}{}_{0}$) breaks into the sum of new state vectors, $g\phi^{\mu}{}_{n}|0\rangle$ (resp. $\langle 0|\phi^{\mu}{}_{n}g$). We call the equation

$$\phi^{0}{}_{0}g|0\rangle = \sum_{\mu=0}^{1}\sum_{n\in\mathbb{Z}} a^{0\mu}{}_{0n}g\phi^{\mu}{}_{n}|0\rangle$$

$$\left(\text{resp.}\langle 0|g\phi^{0}{}_{0} = \sum_{\mu=0}^{1}\sum_{n\in\mathbb{Z}} b^{0\mu}{}_{0n}\langle 0|g\phi^{\mu}{}_{n}\right),$$

"the scattering of $g|0\rangle$ (resp. $\langle 0|g$)," where $a^{0\mu}{}_{0n}$ and $b^{0\mu}{}_{0n}$ are superfields. For the Hamiltonian Φ, we consider the element $g_0\Phi g\in G(\tilde{V},\tilde{V})$, where g_0, $g\in G(\tilde{V},\tilde{V})$ and g_0 is a constant element. Considering the scattering of $\langle 0|g_0\Phi g$ under the condition, $\langle 0|g_0\Phi g\phi^{0}{}_{0} = \alpha\langle 0|\phi^{0}{}_{0}g_0\Phi g$, where α is a certain superfield, we get the Grassmann equation:[6]

$$'(\omega_j)\exp\left(\theta\Lambda + x\Lambda^2 + \sum_{n\equiv 2,3(\text{mod }4)} t_n\Gamma^n\right)\Xi = 0,$$

where ω_j ($j\in\mathbb{Z}$), is a superfield that depends on g, and Ξ is a superframe of the orthogonal universal super Grassmann manifold (USGM) determined by g_0. Thus we have a solution of the OSp–SKP hierarchy.

The knowledge of the representation theory of the infinite-dimensional Lie algebra of B type o(∞) (cf. Refs. 7–9) is indispensable to our study. We review the representations of o(∞). Let E_{ij} ($i,j\in\mathbb{Z}$) be the matrix unit. Put $Z_{ij} = E_{ij} - (-)^{i+j}E_{-j,-i}$.

We define the Lie algebra o(∞) over $\mathbb{C}$ by

$$\text{o}(\infty) = \left\{\sum_{i,j\in\mathbb{Z}} a_{ij}Z_{ij} | a_{ij}=0 \text{ if } |i-j|\gg 0\right\}.$$

The Lie bracket of o(∞) is defined by

$$[Z_{mn},Z_{kl}] = \delta_{nk}Z_{ml} - \delta_{lm}Z_{kn} - (-)^{k+l}\delta_{n,-l}Z_{m,-k} + (-)^{k+l}\delta_{k,-m}Z_{-l,n}.$$

The Chevalley generators of o(∞) are $h_i = Z_{ii}$ ($i>0$), $e_i = Z_{i,i+1}$ and $f_i = Z_{i+1,i}$ ($i\geqslant 0$). Let us consider the one-dimensional central extension $\text{o}(\infty)^{\sim} = \text{o}(\infty)\oplus\mathbb{C}c$. The Lie bracket of $\text{o}(\infty)^{\sim}$ is defined by

$$[Z_{mn}+\lambda c, Z_{kl}+\mu c]^{\sim} = [Z_{mn},Z_{kl}] + C(Z_{mn},Z_{kl})c,$$

where [ ] is the Lie bracket of o(∞). The two-cocycle $C(\cdot,\cdot)$ is defined by

$$C(Z_{mn},Z_{kl}) = (\delta_{nk}\delta_{ml} - (-)^{k+l}\delta_{m,-k}\delta_{l,-n}) \times (Y_B(-m) - Y_B(-n)),$$

where

$$Y_B(i) = \begin{cases} 1, & i>0, \\ \frac{1}{2}, & i=0, \\ 0, & i<0. \end{cases}$$

Let us consider a representation of $\text{o}(\infty)^{\sim}$. Consider the Clifford algebra BCL generated by ϕ_n ($n\in\mathbb{Z}$), and the unit 1 with the defining relation

$$\phi_n\phi_m + \phi_m\phi_n = (-)^{mn}\delta_{n,-m}, \quad n,m\in\mathbb{Z}.$$

Put

$$W_{ann} = \bigoplus_{i<0}\mathbb{C}\phi_i \text{ and } W_{cr} = \bigoplus_{i>0}\mathbb{C}\phi_i.$$

We define the Fock space $F = \text{BCL}/(\text{BCL}\cdot W_{ann})$. We de-

 0022-2488/90/092263-08$03.00

note the residue class of 1 in F by $|0\rangle$. Note that F is a left BCL module with vacuum vector $|0\rangle$ satisfying $\phi_n|0\rangle = 0$ for $n<0$. Similarly, we introduce a right BCL module $F^* = (W_{cr}\cdot \mathrm{BCL})\backslash \mathrm{BCL}$. The residue class of 1 in F^*, denoted by $\langle 0|$, satisfies $\langle 0|\phi_n = 0$ for $n>0$. Let us define the representation of $\mathrm{o}(\infty)\tilde{}$ on F as follows. For $v\in F$, put $\rho(z_{ij})v = (-)^j{:}\phi_i\phi_{-j}{:}v$ and $\rho(c)v = v$, where

$$:\phi_i\phi_j: \begin{cases} \phi_i\phi_j, & j<0, \\ 0, & i=j=0, \\ -\phi_j\phi_i, & j\geqslant 0 (i,j)\neq(0,0). \end{cases}$$

We see that ρ is a representation of $\mathrm{o}(\infty)\tilde{}$, that is, $\rho([A,B]\tilde{}) = [\rho(A),\rho(B)]$, where the bracket on the right-hand side is that of $\mathrm{End}_{\mathbb{C}} F$. We see that F breaks into two irreducible components as an $\mathrm{o}(\infty)$ module such as $F = F^0 \oplus F^1$, where F^0 and F^1 are the $\mathrm{o}(\infty)$ module with highest weight vectors $|0\rangle$ and $\phi_0|0\rangle$, respectively. We define the Hisenberg subalgebra $H = \oplus_{n\in 2\mathbb{Z}+1} H_n \oplus \mathbb{C}c$ of $\mathrm{o}(\infty)\tilde{}$, where $H_n = \Sigma_{i\in\mathbb{Z}} Z_{i,i+n}$. We see that H_n satisfies the relation $[H_n,H_m]\tilde{} = 2n\delta_{n,-m}c$. Hence, we have

$$\exp\left(\sum_{n>0}\frac{\lambda_n}{2n}\rho(H_n)\right) f(\rho(H_{-1}),\rho(H_{-3}),\ldots) = f(\rho(H_{-1}) + \lambda_1, \rho(H_{-3}) + \lambda_3, \ldots),$$

and

$$[(1/2n)\rho(H_n), f(\rho(H_{-1}),\rho(H_{-3}),\ldots)] = \frac{\partial f}{\partial x_n}(\rho(H_{-1}),\rho(H_{-3}),\ldots), \quad n>0,$$

for $f(x)\in\mathbb{C}[[x_1,x_3,\ldots]]$ in the universal enveloping algebra $U(\mathrm{End}_{\mathbb{C}} F)$. The linear functional on BCL, $\langle\ \rangle$: BCL$\in g \to \langle g\rangle\in\mathbb{C}$, is defined by $\langle 1\rangle = 1$, $\langle g\rangle = 0$ if $g|0\rangle = 0$ or $\langle 0|g = 0$. We define the map σ from F to $\mathbb{C}[[x_1,x_2,\ldots]]$ by $\sigma(g|0\rangle) = \langle \exp(H(x))g\rangle$, where

$$H(x) = \sum_{n=0}^{\infty} x_{2n+1}\rho(H_{2n+1}).$$

Let α be an element of $\mathrm{End}_{\mathbb{C}} F$ and suppose $\alpha(F^i)\subset F^i$ $(i=0,1)$. We define the differential operator $E(\alpha)$ by

$$E(\alpha)\langle \exp(H(x))g\rangle = \langle \exp(H(x))\alpha g\rangle.$$

The map E is an algebra homomorphism from $\mathrm{End}_{\mathbb{C}} F$ to the algebra of differential operators on $\mathbb{C}[[x_1,x_3,\ldots]]$. One can verify that

$$E(\rho(H_{2n+1})) = \frac{\partial}{\partial x_{2n+1}}$$

and

$$E(\rho(H_{-2n-1})) = (2n+1)x_{2n+1} \quad \text{for } n\geqslant 0.$$

Put $\phi(z) = \Sigma_{n\in\mathbb{Z}}\phi_n z^n$, $z\in\mathbb{C}^\times$. We have the following theorem.

Theorem 1.1 (cf. Refs. 8 and 10): Let q be a linear operator on F such that $q\phi_n = \phi_n q n\in\mathbb{Z}$ and $q|0\rangle = \phi_0|0\rangle$. Then $\phi(z)$ is represented as

$$\phi(z) = q\Gamma_-(z)\Gamma_+(z), \tag{1.1}$$

where

$$\Gamma_\pm(z) = \exp\left(\mp\sum_{n>0}\frac{\rho(H_{\pm(2n+1)})z^{\mp(2n+1)}}{2n+1}\right).$$

Proof: As an element of $\mathrm{End}_{\mathbb{C}} F$, $\phi(z)$ acts on F transposes one component for another, that is, $\phi(z)F^0\subset F^1$ and $\phi(Z)F^1\subset F^0$. Therefore, we can suppose that $\phi(z) = q\alpha(z)$, where $\alpha(z)F^i\subset F^i$. One can easily verify that

$$[\phi(z),\rho(H_n)] = -2z^n\phi(z), n\in 2\mathbb{Z}+1.$$

This means that $[\alpha(z),\rho(H_n)] = -2z^n\alpha(z)$. Put $X(z) = E(\alpha(z))$. Then we have

$$\left[X(z),\frac{\partial}{\partial x_{2n+1}}\right] = -2z^{2n+1}X(z), \tag{1.2}$$

$$[X(z),x_{2n+1}] = -2(2n+1)^{-1}z^{-(2n+1)}X(z). \tag{1.3}$$

From (1.2), (1.3), and Lemmas A and B below (shown in Ref. 8), we have

$$X(z) = m\cdot\exp\left(\sum_{n>0}x_{2n+1}z^{2n+1}\right)\exp\left(\sum_{n>0}z^{-(2n+1)}\times(2n+1)^{-1}\frac{\partial}{\partial x_{2n+1}}\right),$$

where m is a constant. Then $\alpha(z) = m\cdot\Gamma_-(z)\Gamma_+(z)$ and $\phi(z) = m\cdot q\Gamma_+(z)\Gamma_-(z)$. Comparing both sides of

$$\phi(z)|0\rangle = m\cdot q\Gamma_+(z)\Gamma_-(z)|0\rangle.$$

and

$$\phi(z)\phi_0|0\rangle = m\cdot q\Gamma_-(z)\Gamma_+(z)\phi_0|0\rangle,$$

we see that $m=1$.

Lemma A: Suppose that $X(z)$ satisfies (1.3). Then $X(z)$ is represented as

$$X(z) = M(x,z)\exp\left(-\sum_{n>0}z^{-(2n+1)}(2n+1)^{-1}\frac{\partial}{\partial x_{2n+1}}\right),$$

where $M(x,z)\in\mathbb{C}[[x_1,x_3,\ldots,z,z^{-1}]]$.

Lemma B: Suppose that $M(x,z)\in\mathbb{C}[[x_1,x_3,\ldots,z,z^{-1}]]$ satisfies

$$\left[M(x,z),\frac{\partial}{\partial x_{2n+1}}\right] = -2z^{2n+1}M(x,z).$$

Then $M(x,z) = m\cdot\exp(\Sigma_{n>0}x_{2n+1}z^{2n+1})$, where m is a constant number. Q.E.D.

The formula (1.1) is called a boson–fermion correspondence of $\mathrm{o}(\infty)$ and $\Gamma_\pm(z)$ are called the vertex operators.

In Sec. II we will introduce $\mathrm{osp}(\infty|\infty)$ and its central extension $\mathrm{osp}(\infty|\infty)\tilde{}$. We realize the $\mathrm{osp}(\infty|\infty)\tilde{}$ in terms of the super Clifford algebra generated by the neutral super free fermions and discuss thh Fock representation of $\mathrm{osp}(\infty|\infty)\tilde{}$. As an analogy of the theory of Kac and van de Leur,[4] we deduce the formula of boson–fermion correspondence of $\mathrm{osp}(\infty|\infty)\tilde{}$ and the super vertex operator. We also mention the correspondence the theory of $b_{\infty|\infty}$ (Ref. 5) to our $\mathrm{osp}(\infty|\infty)$. In Sec. III, the transformation group $G(\tilde{V},\tilde{V})$ will be defined. And we give the precise definition of the "state vector" $g|0\rangle$ and $\langle 0|g$ ($g\in\tilde{\mathrm{SBCL}}$), and the "scattering" of the state vector. We define the Hamiltonian in terms of the Heisenberg algebra of $\mathrm{osp}(\infty|\infty)\tilde{}$ and obtain solutions of the OSp–SKP hierarchy under some "scattering" conditions.

II. FOCK REPRESENTATION OF osp(∞|∞)

Let $E^{\mu\nu}{}_{ij}$ $(i,j\in\mathbb{Z},\ \mu,\nu\in\{0,1\})$ be the following blocks of $\mathbb{Z}\times\mathbb{Z}$ matrices:

$$E^{00}{}_{ij}=\begin{pmatrix}E_{ij} & 0\\ 0 & 0\end{pmatrix},\quad E^{01}{}_{ij}=\begin{pmatrix}0 & E_{ij}\\ 0 & 0\end{pmatrix},$$

$$E^{10}{}_{ij}=\begin{pmatrix}0 & 0\\ E_{ij} & 0\end{pmatrix},\quad E^{11}{}_{ij}=\begin{pmatrix}0 & 0\\ 0 & E_{ij}\end{pmatrix}.$$

Put $Z^{00}{}_{ij}=E^{00}{}_{ij}-(-)^{i+j}E^{00}{}_{-j,-i}$, $Z^{11}{}_{ij}=E^{11}{}_{ij}-(-)^{i+j}E^{11}{}_{-j-1,-i-1}$, and $Z^{01}{}_{ij}=E^{01}{}_{ij}+(-)^{i+j}E^{10}{}_{-j-1,-i}$. The infinite-dimensional Lie superalgebra osp(∞|∞) over $\mathbb{C}$ is defined by

$$\mathrm{osp}(\infty|\infty)=\Big\{\sum_{\rho,\mu=0,1}\sum_{i,j\in\mathbb{Z}}a^{\rho\mu}{}_{ij}Z^{\rho\mu}{}_{ij}\,|\,a^{\rho\mu}{}_{ij}=0 \text{ if } |i-j|\gg 0\Big\}.$$

The element $Z^{\rho\mu}{}_{ij}$ satisfies the following bracket relations:

$$[Z^{00}{}_{ij},Z^{00}{}_{kl}]=\delta_{jk}Z^{00}{}_{il}-\delta_{li}Z^{00}{}_{kj}-(-)^{k+l}\delta_{j,-l}Z^{00}{}_{i,-k}+(-)^{k+l}\delta_{k,-i}Z^{00}{}_{-lj},\tag{2.1}$$

$$[Z^{11}{}_{ij},Z^{11}{}_{kl}]=\delta_{jk}Z^{11}{}_{il}-\delta_{li}Z^{11}{}_{kj}+(-)^{l+k}\delta_{i,-k-1}Z^{11}{}_{-l-1,j}-(-)^{l+k}\delta_{-j-1,l}Z^{11}{}_{i,-k-1},\tag{2.2}$$

$$[Z^{00}{}_{ij},Z^{01}{}_{kl}]=\delta_{jk}Z^{01}{}_{il}-(-)^{i+j}\delta_{-i,k}Z^{01}{}_{-jl},\tag{2.3}$$

$$[Z^{11}{}_{ij},Z^{01}{}_{kl}]=-\delta_{il}Z^{01}{}_{kj}+(-)^{i+j}\delta_{j,-l-1}Z^{01}{}_{k,-l-1},\tag{2.4}$$

$$[Z^{01}{}_{ij},Z^{01}{}_{kl}]_{+}=(-)^{k+l}\delta_{j,-l-1}Z^{00}{}_{i,-k}+(-)^{k+l}\times\delta_{-i,k}Z^{11}{}_{-l-1,j}.\tag{2.5}$$

The bracket relation (2.1) is that of o(∞). There is a triangular decomposition

$$\mathrm{osp}(\infty|\infty)=n_{+}\oplus h\oplus n_{-},$$

where

$$h=\bigoplus_{i>0}\mathbb{C}Z^{00}{}_{ii}\oplus\bigoplus_{i\geqslant 0}\mathbb{C}Z^{11}{}_{ii},$$

is a Cartan subalgebra,

$$n_{+}=\bigoplus_{\substack{i<j\\ j>0}}\mathbb{C}Z^{00}{}_{ij}\oplus\bigoplus_{i<j}\mathbb{C}Z^{01}{}_{ij}\oplus\bigoplus_{\substack{i<j\\ j>0}}\mathbb{C}Z^{11}{}_{ij}\quad\Big(\text{resp.}\,n_{-}=\bigoplus_{\substack{i>j\\ j<0}}\mathbb{C}Z^{00}{}_{ij}\oplus\bigoplus_{i>j}\mathbb{C}Z^{01}{}_{ij}\oplus\bigoplus_{\substack{i>j\\ j<0}}\mathbb{C}Z^{11}{}_{ij}\Big)$$

is upper (resp. lower) triangular part.

Furthermore, we consider the one-dimensional central extension $\widetilde{\mathrm{osp}(\infty|\infty)}=\mathrm{osp}(\infty|\infty)\oplus\mathbb{C}c$. The bracket relation of $\widetilde{\mathrm{osp}(\infty|\infty)}$ is defined by $[A+\mu c,B+\lambda c]^{\sim}_{(+)}=[A,B]_{(+)}+C(A,B)c$, for $A,B\in\mathrm{osp}(\infty|\infty),\mu,\lambda\in\mathbb{C}$, where

$$C(Z^{00}{}_{mn},Z^{00}{}_{kl})=(\delta_{nk}\delta_{ml}-(-)^{l+k}\delta_{m,-k}\delta_{n,-l})\times(Y_B(-m)-Y_B(-n)),$$

$$C(Z^{11}{}_{mn},Z^{11}{}_{kl})=(\delta_{ml}\delta_{nk}-(-)^{k+l}\delta_{m,-k-1}\delta_{n,-l-1})\times(Y_C(-n-1)-Y_C(-m-1)),$$

$$C(Z^{01}{}_{mn},Z^{01}{}_{kl})=(-)^{k+l}\delta_{n,-l-1}\delta_{m,-k}(Y_B(k)-Y_C(-n-1)),$$

$$C(Z^{01}{}_{mn},Z^{\mu\mu}{}_{kl})=C(Z^{00}{}_{mn},Z^{11}{}_{kl})=0,\quad \mu=0,1,$$

with

$$Y_C(i)=\begin{cases}1, & i\geqslant 0,\\ 0, & i<0.\end{cases}$$

Let us consider the representation of $\widetilde{\mathrm{osp}(\infty|\infty)}$. We introduce the superalgebra SBCL whose generators are $\phi^{\mu}{}_{n}$ $(n\in\mathbb{Z}$ $\mu=\{0,1\})$ and the unit 1 satisfying the relation,

$$\phi^{\mu}{}_{n}\phi^{\nu}{}_{m}+(-)^{\mu\nu}\phi^{\nu}{}_{m}\phi^{\mu}{}_{n}=(-)^{m}\delta_{\mu\nu}\delta_{n,-m-\nu}.$$

We call $\phi^{\mu}{}_{n}$ the neutral super free fermion. Let g be a monomial of SBCL, such as

$$g=\phi^{\mu_1^{m_1}}_{i_1}\cdots\phi^{\mu_j^{m_j}}_{i_j},\ i_t\in\mathbb{Z},\mu_t=0,1,m_t\geqslant 1\ (1\leqslant t\leqslant j),$$

and $m_t=1$ if $\mu_t=0$. We define

$$\deg_{\bar{1}}(g)=\sum_{t=1}^{j}\mu_t m_t.$$

The $\mathbb{Z}_2$-grading of SBCL is defined by $\mathrm{SBCL}=\mathrm{SBCL}_{\bar{0}}\oplus\mathrm{SBCL}_{\bar{1}}$, where the $\deg_{\bar{1}}$ of monomial of $\mathrm{SBCL}_{\bar{i}}$ is congruent to i mod 2. Put

$$W^{s}{}_{cr}=\bigoplus_{\mu=0,1}\bigoplus_{n>0,\mu+n>0}\mathbb{C}\phi^{\mu}{}_{n}$$

and

$$W_{ann}=\bigoplus_{\mu=0,1}\bigoplus_{n<0}\mathbb{C}\phi^{\mu}{}_{n}.$$

We define the super Fock space F^{s} by $F^{s}=\mathrm{SBCL}/(\mathrm{SBCL}\cdot W_{ann})$. We denote $|0\rangle$ the image of 1 of the canonical map $\mathrm{SBCL}\to F^{s}$. Note that $\phi^{\mu}{}_{n}|0\rangle=0$, for $n<0,\mu=0,1$. We also define the right SBCL module F^{s^*} by $(W_{cr}\cdot\mathrm{SBCL})\backslash\mathrm{SBCL}$. We denote the image of 1 of the canonical map $\mathrm{SBCL}\to F^{s^*}$ by $\langle 0|$. The vector $\langle 0|$ satisfies

$$\langle 0|\phi^{\mu}{}_{n}=0 \text{ for } n\geqslant 0,\mu=0,1,(\mu,n)\neq(0,0).$$

The $\mathbb{Z}_2$-grading of F^{s} and F^{s^*} is introduced canonically from SBCL.

We consider the representation of $\widetilde{\mathrm{osp}(\infty|\infty)}$ on F^{s}. Put $X^{\mu\nu}{}_{ij}=(-)^{j+\mu+\nu}{:}\phi^{\mu}{}_{i}\phi^{\nu}{}_{-j-\nu}{:}$ $(i,j\in\mathbb{Z}$ and $\mu,\nu=0,1)$, where

$$:\phi^{\mu}{}_{i}\phi^{\nu}{}_{j}: = \begin{cases} \phi_i\phi_j, & j<0, \\ 0, & i=j=\mu=\nu=0, \\ -(-)^{\mu\nu}\phi^{\nu}{}_{j}\phi^{\mu}{}_{i}, & j\geqslant 0 (i,j,\mu,\nu)\neq(0,0,0,0). \end{cases}$$

For $Z^{\mu\nu}{}_{ij}$, we define $\gamma(Z^{\mu\nu}{}_{ij})\in \mathrm{End}_{\mathbb{C}}F^s$ by $\gamma(Z^{\mu\nu}{}_{ij})v = (\sqrt{-1})^{\nu-\mu}X^{\mu\nu}{}_{ij}v (v\in F^s)$. Furthermore, define $\gamma(c)v = v (v\in F^s)$. We obtain the representation of $\mathrm{osp}(\infty|\infty)^{\sim}$ on F^s, that is, the following equation holds

$$\gamma([A,B]^{\sim}_{(+)}) = [\gamma(A),\gamma(B)]_{(+)}, \quad A,B\in \mathrm{osp}(\infty|\infty)^{\sim}.$$

As an $\mathrm{osp}(\infty|\infty)^{\sim}$ module F_s breaks into two irreducible components F^{s0} and F^{s1}, where $F^{s0}(\text{resp.}F^{s1})$ is the $\mathrm{osp}(\infty|\infty)^{\sim}$ module with the highest weight vector $|0\rangle$ (resp. $\phi^0{}_0|0\rangle$).

The Heisenberg algebra H^s of $\mathrm{osp}(\infty|\infty)^{\sim}$ is defined by

$$H^s = \bigoplus_{n\in\mathbb{Z}} \mathbb{C}H^{00}{}_{2n+1} \oplus \bigoplus_{n\in\mathbb{Z}} \mathbb{C}H^{11}{}_{2n+1} \oplus \bigoplus_{n\in\mathbb{Z}} \mathbb{C}H^{01}{}_{n} \oplus \mathbb{C}c,$$

where

$$H^{\mu\mu}{}_{n} = \sum_{i\in\mathbb{Z}} Z^{\mu\mu}{}_{i,i+n}, \ \mu=0.1, \ n\in 2\mathbb{Z}+1 \text{ and } H^{01}{}_{n} = \sum_{i\in\mathbb{Z}} Z^{01}{}_{i,i+n}, \ n\in\mathbb{Z}.$$

We use the notation $H^{\mu\nu}{}_{n}$ $(\mu,\nu=0.1, \ n\in\mathbb{Z})$, regarding $H^{00}{}_{2n}$ and $H^{11}{}_{2n}=0$ for $n\in\mathbb{Z}$.

Proposition 2.1: The elements $H^{\mu\nu}{}_{n}$ satisfy the following relations:

$$[H^{00}{}_{n},H^{00}{}_{m}]^{\sim} = 2n\delta_{n,-m}c, \tag{2.6}$$

$$[H^{11}{}_{n},H^{11}{}_{m}]^{\sim} = -2n\delta_{n,-m}c, \tag{2.7}$$

$$[H^{00}{}_{n},H^{01}{}_{m}]^{\sim} = 2H^{01}{}_{n+m}, \tag{2.8}$$

$$[H^{11}{}_{n},H^{01}{}_{m}]^{\sim} = -2H^{01}{}_{n+m}, \tag{2.9}$$

$$[H^{01}{}_{n},H^{01}{}_{m}]^{\sim}_{+} = (-)^{m}(H^{00}{}_{m+n+1} + H^{11}{}_{m+n+1}) + (-)^{m}(n+\tfrac{1}{2})\delta_{m,-n-1}c. \tag{2.10}$$

Proof: We only show (2.10). The left-hand side of (2.10) is calculated as

$$[H^{01}{}_{n},H^{01}{}_{m}]^{\sim}_{+} = \sum_{i,j\in\mathbb{Z}} [Z^{01}{}_{i,i+n},Z^{01}{}_{j,j+m}]^{\sim}_{+}$$

$$= (-)^{m}\sum_{i\in\mathbb{Z}} Z^{00}{}_{i,i+m+n+1} + (-)^{m}\sum_{i\in\mathbb{Z}} Z^{11}{}_{i,i+m+n+1} + \sum_{i,j\in\mathbb{Z}} C(Z^{01}{}_{i,i+n},Z^{01}{}_{j,j+m})c$$

$$= (-)^{m}(H^{00}{}_{m+n+1} + H^{11}{}_{m+n+1}) + (-)^{m}\sum_{j\in\mathbb{Z}}\delta_{n,-m-1}(Y_B(j) - Y_C(j+m))c.$$

Suppose $m<0$. Then we have

$$\sum_{j\in\mathbb{Z}}\delta_{n,-m-1}(Y_B(j) - Y_C(j+m)) = \delta_{n,-m-1}\sum_{0<j<-m}(Y_B(j) - Y_C(j+m))$$

$$= \delta_{n,-m-1}(-m-1+\tfrac{1}{2}) = (n+\tfrac{1}{2})\delta_{n,-m-1}.$$

Hence we have (2.10). If we suppose $m\geqslant 0$, we get the same formula. One can show (2.6)–(2.9) similarly. Q.E.D.

As an H^s module, F^s also breaks into two irreducible components, F^{s0} and F^{s1}. To show this fact we refer to super boson–fermion correspondence of $\mathrm{osp}(\infty|\infty)^{\sim}$. Put $\phi_{\mu}(z) = \Sigma_{n\in\mathbb{Z}}\phi^{\mu}{}_{n}z^{n}$ $(\mu=0,1, \ z\in\mathbb{C}^{\times})$. We get the following result.

Theorem 2.2: Let Q be the operator on F^s such as $Q\phi^{\mu}{}_{n} = (-)^{\mu}\phi^{\mu}{}_{n}Q$ and $Q|0\rangle = \phi^0{}_0|0\rangle$. Then

$$\phi_0(z) = Q\Gamma_-(z)\Gamma_+(z), \tag{2.11}$$

$$\sqrt{-1}z\phi_1(z) = -2Q\Gamma_-(z)\gamma(H^{01}(z))\Gamma_+(z), \tag{2.12}$$

where

$$\Gamma_{\pm}(z) = \exp\left(\mp\sum_{m>0}\frac{\gamma(H^{00}{}_{\pm(2m+1)})z^{\mp(2m+1)}}{2m+1}\right)$$

and

$$H^{01}(z) = \sum_{j\in\mathbb{Z}}(-)^{j}H^{01}{}_{j}z^{-j}.$$

Proof: The formula (2.11) is nothing but Theorem 1.1. We show (1.12). By a simple calculation, one sees

$$[\gamma(H^{01}{}_{m}),\phi_0(z)] = (-)^{m}\sqrt{-1}z^{m+1}\phi_1(z),$$

especially, $[\gamma(H^{01}{}_{0}),\phi_0(z)] = \sqrt{-1}z\phi_1(z)$. From (2.11) we have

$$[\gamma(H^{01}{}_{0}),\phi_0(z)] = -Q(\gamma(H^{01}{}_{0})\Gamma_-(z)\Gamma_+(z) + \Gamma_-(z)\Gamma_+(z)\gamma(H^{01}{}_{0})). \tag{2.13}$$

The first term of the right-hand side of (2.13) is calculated as follows:

$$\gamma(H^{01}{}_{0})\Gamma_-(z)\Gamma_+(z) = \Gamma_-(z)\exp\left(-\sum_{m\geqslant 0}\frac{z^{2m+1}\ \mathrm{ad}\ (\gamma(H^{00}{}_{-2m-1}))}{2m+1}\right) \times\gamma(H^{01}{}_{0})\Gamma_+(z)$$

$$= \Gamma_-(z)\sum_{j=0}^{\infty}\tilde{p}_j(-\mathrm{ad}\ \gamma(H^{00}{}_{-2m-1}):m\geqslant 0) \times z^{j}\gamma(H^{01}{}_{0})\Gamma_+(z), \tag{2.14}$$

where the polynomial $\tilde{p}_j(x)$ is defined by

$$\exp\left(\sum_{n=0}^{\infty}x_{2n+1}(2n+1)^{-1}z^{2n+1}\right) = \sum_{j=0}^{\infty}\tilde{p}_j(x)z^{j}.$$

From (2.8) and (2.14), we have

$$\gamma(H^{01}{}_{0})\Gamma_-(z)\Gamma_+(z) = \Gamma_-(z)\sum_{j=0}^{\infty}\tilde{p}_j(-2)z^{j}\gamma(H^{01}{}_{-j})\Gamma_+(z). \tag{2.15}$$

Similarly, we have

$$\Gamma_-(z)\Gamma_+(z)\gamma(H^{01}{}_0)$$

$$= \Gamma_-(z)\sum_{j=0}^{\infty}\tilde{p}_j(-2)z^{-j}\gamma(H^{01}{}_j)\Gamma_+(z). \tag{2.16}$$

Note that

$$\exp\left(\sum_{n>0}(-2/2n+1)z^{2n+1}\right) = \sum_{j>0}\tilde{p}_j(-2)z^j$$

and

$$\sum_{n>0} -2(2n+1)^{-1}z^{2n+1} = \log\frac{1-z}{1+z}.$$

Then

$$p_j(-2) = \begin{cases} 1, & j=0, \\ 2(-)^j, & j>0. \end{cases}$$

From (2.15) and (2.16), we have

$$[\rho(H^{01}{}_0),\phi_0(z)] = -2Q\Gamma_-(z)\gamma(H^{01}(z))\Gamma_+(z).$$

Q.E.D.

From Theorem 2.2, we have the following corollary.

Corollary: As H^s modules, F^{s0} and F^{s1} are irreducible.

In conclusion of this section, we mention the relation between our result and that of Kac and van de Leur[5]. In Ref. 5, the indices of neutral super free fermion and matrix units belong to $\frac{1}{2}\mathbb{Z}$. Through the following correspondence, we see that the notations of the present paper are equivalent to that of 5:

Kac and van de Leur[5]		this paper
ϕ_i	→	$(\sqrt{-1})^{2(i-[i])}\phi^{2(i-[i])}{}_{[i]}$
E_{ij}	→	$E^{2(i-[i]),2(j-[j])}{}_{[i],[j]}, i,j\in\frac{1}{2}\mathbb{Z},$

where $[i] = \sup\{k\in\mathbb{Z} | k\leqslant i\}$.

III. THE ORTHOGONAL UNIVERSAL SUPER GRASSMANN MANIFOLD (USGM) AND THE OSp–SKP HIERARCHY

We denote by $\mathscr{A}$ the arbitrary superalgebra. Let us recall some notations of Lie supergroup. The Lie supergroup $\mathrm{SGL}(\mathscr{A})$ is defined by

$$\mathrm{SGL}(\mathscr{A}) = \left\{\begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11}\end{pmatrix} | A_{ij}\in\mathrm{Mat}(\mathbb{Z}\times\mathbb{Z}|\mathscr{A}_{i+j}),\right.$$
$$\left. A_{00} \text{ and } A_{11} \text{ are invertible}\right\}.$$

The supertranspose "st" is defined by

$${}^{\mathrm{st}}\begin{pmatrix} A_{00} & A_{01} \\ A_{10} & A_{11}\end{pmatrix} = \begin{pmatrix} {}^tA_{00} & {}^tA_{10} \\ -{}^tA_{01} & {}^tA_{11}\end{pmatrix}.$$

Put $J = ((-)^i\delta_{i,-j})_{i,j\in\mathbb{Z}}$ and $K = \Lambda J$, where $\Lambda = (\delta_{i+1,j})_{i,j\in\mathbb{Z}}$. The Lie supergroup $\mathrm{OSp}(\mathscr{A})$ is defined by

$$\mathrm{OSp}(\mathscr{A}) = \left\{A\in\mathrm{SGL}(\mathscr{A}) | \begin{pmatrix} J & 0 \\ 0 & -{}^tK\end{pmatrix}{}^{\mathrm{st}}A\begin{pmatrix} J & 0 \\ 0 & -K\end{pmatrix} = A^{-1}\right\}.$$

We define $\tilde{\mathrm{SBCL}}$ by $\mathrm{SBCL}\otimes\mathscr{A}$. The $\mathbb{Z}_2$-grading $\tilde{\mathrm{SBCL}} = \tilde{\mathrm{SBCL}}_{\bar{0}}\oplus\tilde{\mathrm{SBCL}}_{\bar{1}}$ is given $\tilde{\mathrm{SBCL}}_{\bar{l}} = \oplus_{\alpha+\beta\equiv l(\mathrm{mod}\,2)}\mathrm{SBCL}_{\bar{\alpha}}\otimes\mathscr{A}_{\bar{\beta}}$. We assume the following (anti) commutation relation of $\phi^\mu{}_n$ $n\in\mathbb{Z}$, $\mu = 0,1$, and $\mathscr{A}$.

$$\phi^\mu{}_n a - (-)^{\mu i}a\phi^\mu{}_n = 0, \text{ where } a\in\mathscr{A}_i.$$

Put

$$V = \bigoplus_{\mu=0,1}\bigoplus_{n\in\mathbb{Z}}\mathbb{C}\phi^\mu{}_n$$

and $\tilde{V} = V\otimes\mathscr{A}$. We introduce the transformation group of $\tilde{V}$ as $G(\tilde{V},\tilde{V}) = \{g\in\tilde{\mathrm{SBCL}}_{\bar{0}} | g \text{ is invertibe}, g\tilde{V}g^{-1} = \tilde{V}\}$. For $g\in G(\tilde{V},\tilde{V})$, we define the matrix $\eta(g)$ by $\eta(g) = (\eta(g)^{\mu\nu}{}_{ij})_{\mu,\nu=0,1}{}_{i,j\in\mathbb{Z}}$, where

$$g\phi^\mu{}_n g^{-1} = \sum_{\nu=0}^{1}\sum_{m\in\mathbb{Z}}\eta(g)^{\mu\nu}{}_{nm}\phi^\nu{}_m.$$

If $g\in\tilde{\mathrm{SBCL}}_{\bar{0}}$, then $g\phi^\mu{}_n g^{-1}\in\tilde{\mathrm{SBCL}}_{\bar{\mu}}$ and $\deg(\eta(g)^{\nu\mu}{}_{nm}) + \nu \equiv\mu(\mathrm{mod}\,2)$. Then one sees that $\deg(\eta(g)^{\mu\nu}{}_{nm}) \equiv\mu+\nu(\mathrm{mod}\,2)$. We show the following fact.

Fact: If $g,h\in G(\tilde{V},\tilde{V})$. Then $\eta(gh) = \eta(h)\eta(g)$.

Proof: By definition, $gh\phi^\mu{}_n(gh)^{-1} = \sum_{\nu=0}^{1}\sum_{m\in\mathbb{Z}}\eta(gh)^{\mu\nu}{}_{nm}\phi^\nu{}_m$. Notice that $g\in\tilde{\mathrm{SBCL}}_{\bar{0}}$, we have

$$gh\phi^\mu{}_n(gh)^{-1} = g(h\phi^\mu{}_n h^{-1})g^{-1}$$
$$= g\left(\sum_{\rho=0}^{1}\sum_{l\in\mathbb{Z}}\eta(h)^{\mu\rho}{}_{nl}\phi^\rho{}_l\right)g^{-1}$$
$$= \sum_{\rho=0}^{1}\sum_{l\in\mathbb{Z}}\eta(h)^{\mu\rho}{}_{nl}g\phi^\rho{}_l g^{-1}$$
$$= \sum_{\nu=0}^{1}\sum_{m\in\mathbb{Z}}\left(\sum_{\rho=0}^{1}\sum_{l\in\mathbb{Z}}\rho(h)^{\mu\rho}{}_{nl}\eta(g)^{\rho\nu}{}_{lm}\right)\phi^\nu{}_m.$$

Then we have the conclusion. Q.E.D.

From the above fact, we see $\eta(g^{-1}) = \eta(g)^{-1}$ for $g\in G(\tilde{V},\tilde{V})$. One can see $\eta(g)\in\mathrm{SGL}(\mathscr{A})$ for $g\in G(\tilde{V},\tilde{V})$.

Proposition 3.1: $\eta(g)\in\mathrm{OSp}(\mathscr{A})$ for $g\in G(\tilde{V},\tilde{V})$.

Proof: From the relation of $\phi^\mu{}_n$, we have

$$g\phi^\mu{}_m g^{-1}g\phi^\nu{}_n g^{-1} + (-)^{\mu\nu}g\phi^\nu{}_n g^{-1}g\phi^\mu{}_m g^{-1} = (-)^n\delta_{\mu\nu}\delta_{m,-n-\nu}. \tag{3.1}$$

The left-hand side of (3.1) is calculated as

$$\sum_{\rho_1,\rho_2=0}^{1}\sum_{i,j\in\mathbb{Z}}\{(-)^{(\nu+\rho_2)\rho_1}\eta(g)^{\mu\rho_1}{}_{mi}\eta(g)^{\nu\rho_2}{}_{nj}\phi^{\rho_1}{}_i\phi^{\rho_2}{}_j + (-)^{\mu\nu+(\mu+\rho_1)\rho_2+(\nu+\rho_2)(\nu+\rho_1)}\eta(g)^{\mu\rho_1}{}_{mi}\eta(g)^{\nu\rho_2}{}_{nj}\phi^{\rho_2}{}_j\phi^{\rho_1}{}_i\}$$
$$= \sum_{\rho_1,\rho_2=0}^{1}\sum_{i,j\in\mathbb{Z}}(-)^{(\nu+\rho_2)\rho_1}\eta(g)^{\mu\rho}{}_{mi}\eta(g)^{\nu\rho_2}{}_{nj}\{\phi^{\rho_1}{}_i\phi^{\rho_2}{}_j + (-)^{\mu\nu+(\mu+\rho_1)\rho_2+(\nu+\rho_1)(\mu+\rho_2)+(\nu+\rho_1)\rho_2}\phi^{\rho_2}{}_j\phi^{\rho_1}{}_i\}. \tag{3.2}$$

Note that

$$\mu\nu+(\mu+\rho_1)\rho_2+(\nu+\rho_1)(\mu+\rho_2)+(\nu+\rho_1)\rho_2 \equiv\rho_1\rho_2(\mathrm{mod}\ 2).$$

Then we have

$$(3.2)=\sum_{\rho_1,\rho_2=0}\sum_{i,j\in\mathbf{Z}}(-)^{(\nu+\rho_2)\rho_1}\eta(g)^{\mu\rho_1}{}_{mi}\eta(g)^{\nu\rho_2}{}_{nj}(\phi^{\rho_1}{}_i\phi^{\rho_2}{}_j + (-)^{\rho_1\rho_2}\phi^{\rho_2}{}_j\phi^{\rho_1}{}_i).$$

From (3.1), we have

$$\sum_{\rho=0}^{1}\sum_{i\in\mathbf{Z}}(-)^{\rho\nu+i+n}\eta(g)^{\mu\rho}{}_{mi}\eta(g)^{\nu\rho}{}_{n,-i-\rho}(g) =\delta_{\mu\nu}\delta_{m,-n-\nu}.$$

This implies that

$$\eta(g)\begin{pmatrix}J & 0\\ 0 & -{}^tK\end{pmatrix}^{st}\eta(g)\begin{pmatrix}J & 0\\ 0 & -K\end{pmatrix}=1.$$

Then we see that $\eta(g)\in\mathrm{OSp}(\mathscr{A})$. Q.E.D.

In the rest of this section we construct the OSp–SKP hierarchy in terms of the Fock representation. Let us review the theory of the OSp–SKP hierarchy. Let θ and t_{4n+3} $(n\geqslant 0)$ be odd variables and x and t_{4n+2} $(n\geqslant 0)$ be even variables. The superalgebra $\mathscr{S}$ of superfields defined by $\mathscr{S}=\mathbf{C}[[x,\theta,t_{4n+2},t_{4n+3},n\geqslant 0]]\otimes\mathscr{V}$, where $\mathscr{V}$ is a Grassmann algebra. Super derivation on $\mathscr{S}$ is defined by

$$D=\frac{\partial}{\partial\theta}+\theta\frac{\partial}{\partial x}.$$

The super vector fields on $\mathscr{S}$ is defined by

$$D_{4n+2}=\frac{\partial}{\partial t_{4n+2}},$$

$$D_{4n+3}=\frac{\partial}{\partial t_{4n+3}}+\sum_{k>0}t_{4k+3}\frac{\partial}{\partial t_{4n+4k+6}},\quad n\geqslant 0.$$

We can verify that $[D_{4n+3},D_{4m+3}]_+=2D_{4m+4n+6}$. The operator

$$W=\sum_{i=0}^{\infty}w_jD^{-j}(w_j\in\mathscr{S}_j,\ w_0=1)$$

is called the wave operator of the OSp–SKP hierarchy if W satisfies

$$D_{4n+2}W=-(B_{4n+2}W-WD^{4n+2}),$$
$$D_{4n+3}W=B_{4n+3}W-WD^{4n+3},$$
$$D^{-1}W^*D=W^{-1},$$

where W^* is the formal adjoint of W defined by

$$W^*=\sum_{j=0}^{\infty}(-)^{\epsilon_j}D^{-j}w_j,$$

where $\epsilon_j=j(j+1)/2$, and B_n is a differential operator part of WD^nW^{-1}. To solve the OSp–SKP hierarchy we consider the following linear algebraic equation (Grassmann equation):

$${}^t\mathbf{w}(\theta,x,t)\exp\Big(\theta\Lambda+x\Lambda^2+\sum_{n>0}t_{4n+2}\Gamma^{4n+2} +\sum_{n>0}t_{4n+3}\Gamma^{4n+3}\Big)\Xi=0. \tag{3.3}$$

Here we have put $\mathbf{w}(\theta,x,t)=(\ldots w_1,w_0,w_{-1}\ldots)$ $(w_j\in\mathscr{S}_j$ and $w_0=1$, $w_j=0$ for $j<0)$, $\Gamma=((-)^i\delta_{i+1,j})_{i,j\in\mathbf{Z}}$ and $\Xi=(\xi_{ij})_{i,j\in\mathbf{Z}}$ with $\xi_{ij}\in\mathscr{V}_{i+j}$ and $\xi_{ij}=\delta_{ij}$ for $i\leqslant j$; Ξ is the superframe of USGM$^\phi$ the largest cell in USGM. In general, if ${}^t\mathbf{w}={}^t\mathbf{w}(\theta,x,t)$ satisfies (3.3), then the operator

$$W=\sum_{j\geqslant 0}w_jD^{-j}$$

solves the SKP hierarchy for $t_n[n\equiv 2,3(\mathrm{mod}\ 4)]$. For a matrix $A=(a_{ij})_{\substack{i\in\mathbf{Z}\\ j\in\mathbf{Z}(\mathbf{N}^c)}}$, put $\check{A}=(a^{\mu\nu}{}_{ij})_{\substack{i\in\mathbf{Z}\\ j\in\mathbf{Z}(\mathbf{N}^c)\\ \mu,\nu=0,1}}$ where $a^{\mu\nu}{}_{ij}=a_{2i+\mu,2j+\nu}$. Put $\check{\Xi}=(\xi_j{}^{\mu\nu})_{\substack{j<0\\ \mu,\nu=0,1}}$, where $\xi^{\mu\nu}{}_j=(\xi^{\mu\nu}{}_{ij})_{i\in\mathbf{Z}}$. Add the constraint (3.4) to Ξ such as

$$\begin{aligned}&\langle\xi^{00}{}_i,\xi^{00}{}_j\rangle_B-\langle\xi^{10}{}_i,\xi^{10}{}_j\rangle_C=0,\\ &\langle\xi^{00}{}_i,\xi^{01}{}_j\rangle_B-\langle\xi^{10}{}_i,\xi^{11}{}_j\rangle_C=0,\\ &\langle\xi^{00}{}_i,\xi^{00}{}_j\rangle_B+\langle\xi^{11}{}_i,\xi^{10}{}_j\rangle_C=0,\\ &\langle\xi^{01}{}_i,\xi^{01}{}_j\rangle_B+\langle\xi^{11}{}_i,\xi^{11}{}_j\rangle_C=0,\end{aligned} \tag{3.4}$$

where

$$\langle(a_i)_{i\in\mathbf{Z}},(b_j)_{j\in\mathbf{Z}}\rangle_B=\sum_{k\in\mathbf{Z}}(-)^ka_kb_{-k},$$

and

$$\langle(a_i)_{i\in\mathbf{Z}},(b_j)_{j\in\mathbf{Z}}\rangle_C=\sum_{k\in\mathbf{Z}}(-)^ka_kb_{-k-1}.$$

The set of superframes satisfying the condition (3.4) is called orthogonal USGM$^\phi$. We quote the following theorem from Ref. 3.

Theorem 3.2: Suppose that Ξ is the superframe of the orthogonal USGM$^\phi$ and ${}^t\mathbf{w}={}^t(w_j)_{j\in\mathbf{Z}}$ satisfies (3.3). Then the operator

$$W=\sum_{j\geqslant 0}w_jD^{-j}$$

solves the OSp–SKP hierarchy.

For $g\in G(\tilde{V},\tilde{V})$, the state vectors $g|0\rangle\in\tilde{F}^s$ and $\langle 0|g\in\tilde{F}^{s*}$ are defined, where $\tilde{F}^s(\mathrm{resp.}\tilde{F}^{s*})=F^s\otimes\mathscr{S}(\mathrm{resp.}\ F^{s*}\otimes\mathscr{S})$. Put

$$\Phi(\theta,x,t)=\exp\Big(\theta j^o{}_{-1}-xj^e{}_0+\sum_{n>0}t_{4n+2}j^e{}_n +\sum_{n>0}t_{4n+3}j^o{}_n\Big),$$

where

$$j^e{}_n=-\tfrac{1}{2}\{\gamma(H^{00}{}_{-2n-+})+\gamma(H^{11}{}_{-2n-1})\},$$
$$j^o{}_{-1}=\sqrt{-1}\gamma(H^{01}{}_{-1})$$

and

$$j^o{}_n=\sqrt{-1}\gamma(H^{01}{}_{-2n-2})(n\geqslant 0).$$

Theorem 3.3: Suppose $g\in G(\tilde{V},\tilde{V})$ satisfies the following scattering conditions:

(i) $\phi^0{}_0g^{-1}|0\rangle=g^{-1}\phi^0{}_0|0\rangle$,

(ii) $\langle 0|g_0\Phi g\phi^0{}_0=\alpha\langle 0|\phi^0{}_0g_0\Phi g$, for a constant element $g_0\in G(\tilde{V},\tilde{V})$, where $\alpha\in\mathscr{S}$. If $\eta(g_0)^{\mu\mu}{}_{ij}=\delta_{ij}$ for $j\geqslant i$ and $\phi(g_0)^{\mu\nu}{}_{ij}=0$ for $\mu\neq\nu$ and $j+\mu\geqslant i$, then the operator

$$W=\sum_{\mu=0}^{1}\sum_{n>0}\eta(g)^{0\mu}{}_{0,-n}D^{-2n-\mu}$$

solves the OSp–SKP hierarchy.

Proof: We first show the lemma

Lemma C:

$$\eta(g)^{0\mu}{}_{0j}=\delta_{0\mu}\delta_{0j} \quad \text{for } j\geqslant 0.$$

Proof: By definition we have

$$\phi^0{}_0 g^{-1}|0\rangle=\sum_{\mu=0}^{1}\sum_{j>0}\eta(g)^{0\mu}{}_{0j}g^{-1}\phi^{\mu}{}_j|0\rangle.$$

From (i), one sees that

$$\sum_{\mu=0}^{1}\sum_{j>0}\eta(g)^{0\mu}{}_{0j}g^{-1}\phi^{\mu}{}_j|0\rangle=g^{-1}\phi^0{}_0|0\rangle. \tag{3.5}$$

Multiplying by g the both sides of (3.5) from the left, we have

$$\sum_{\mu=0}^{1}\sum_{j>0}\eta(g)^{0\mu}{}_{0j}\phi^{\mu}{}_j|0\rangle=\phi^0{}_0|0\rangle.$$

Since the elements $\phi^{\mu}{}_j|0\rangle$ $(\mu=0,1, j\geqslant 0)$ are linearly independent over $\mathscr{S}$, we have $\eta(g)^{0\mu}{}_{0j}=\delta_{\mu 0}\delta_{j0}$ for $\mu=0,1$ and $j\geqslant 0$. Q.E.D.

From (ii), we get the equations

$$\eta(g_0\Phi g)^{0\mu}{}_{0j}=0 \quad \text{for } j<0 \text{ and } \mu=0,1. \tag{3.6}$$

The equations (3.6) are equivalent to

$${}^t(\eta(g)^{0\mu}{}_{0j})_{\substack{j\in\mathbf{Z}\\ \mu=0,1}}\ \eta(\Phi)\eta(g_0)\breve{\Xi}_{\phi}=0,$$

where

$$\Xi_{\phi}=(\delta_{ij})_{\substack{i\in\mathbf{Z}\\ j<0}}. \tag{3.7}$$

Lemma D:

$$\eta(\Phi)=\exp\Big(\theta\breve{\Lambda}+x\breve{\Lambda}^2+\sum_{n>0}t_{4n+2}\breve{\Gamma}^{4n+2}+\sum_{n>0}t_{4n+3}\breve{\Gamma}^{4n+3}\Big).$$

Proof: From the commutation relations of $H^{\mu\nu}{}_n$, we have $[j^e{}_n j^0{}_m]=0$, $n\geqslant 0$ and $m\geqslant -1$. $[j^0{}_{-1}j^0{}_m]_+=0$, $m\geqslant 0$. and $[j^e{}_n j^e{}_m]=0$, $n,m\geqslant 0$.

One sees that

$$[t_{4n+2}j^e{}_n, t_{4n+3}j^0{}_n]=0, \quad n,m\geqslant 0, \tag{3.8}$$

$$[xj^e{}_0, t_{4n+3}j^0{}_n]=0, \quad n\geqslant 0, \tag{3.9}$$

$$[\theta j^0{}_{-1}, t_{4n+3}j^0{}_n]=0, \quad n\geqslant 0. \tag{3.10}$$

From (3.8), (3.9), and (3.10), we have

$$\eta(\Phi)=\eta\Big(\exp\Big(\sum_{n>0}t_{4n+3}j^0{}_n\Big)\Big)\eta\Big(\exp\Big(\sum_{n>0}t_{4n+2}j^e{}_n\Big)\Big)\times\eta(\exp(\theta j^0{}_{-1}-xj^e{}_0)).$$

First, we calculate

$$\eta\Big(\exp\Big(\sum_{n>0}t_{4n+2}j^e{}_n\Big)\Big).$$

We have

$$\exp\Big(\sum_{n>0}t_{4n+2}j^e{}_n\Big)\phi^{\mu}{}_m\exp\Big(-\sum_{n>0}t_{4n+2}j^e{}_n\Big)=\exp\Big(\sum_{n>0}t_{4n+2}\operatorname{ad}j^e{}_n\Big)\phi^{\mu}{}_m.$$

One can easily verify that $\operatorname{ad}j^e{}_n\phi^{\mu}{}_m=-\phi^{\mu}{}_{m+2n+1}$. Therefore,

$$\exp\Big(\sum_{n>0}t_{4n+2}\operatorname{ad}j^e{}_n\Big)\phi^{\mu}{}_m=\sum_{l>0}p_l(t_{4n+2}\operatorname{ad}j^e{}_n;n\geqslant 0)\phi^{\mu}{}_m=\sum_{l>0}p_l(-t_{4n+2};n\geqslant 0)\phi^{\mu}{}_{m+l},$$

where we have defined

$$\exp\Big(\sum_{n>0}t_{4n+2}z^{2n+1}\Big)=\sum_{l>0}p_l(t_{4n+2};n\geqslant 0)z^l.$$

Thus we have

$$\eta\Big(\exp\Big(\sum_{n>0}t_{4n+2}j^e{}_n\Big)\Big)=\exp\Big(\sum_{n>0}t_{4n+2}\breve{\Gamma}^{4n+2}\Big).$$

Next let us calculate

$$\eta\Big(\exp\Big(\sum_{n>0}t_{4n+3}j^0{}_n\Big)\Big).$$

In general, we have

$$\exp\Big(\sum_{n>0}t_{4n+3}j^0{}_n\Big)\phi^{\mu}{}_m\exp\Big(-\sum_{n>0}t_{4n+3}j^0{}_n\Big)=\exp\Big(\sum_{n>0}\operatorname{ad}t_{4n+3}j^0{}_n\Big)\phi^{\mu}{}_m.$$

Note that

$$\operatorname{ad}t_{4n+3}j^0{}_n\operatorname{ad}t_{4k+3}j^0{}_k\phi^{\mu}{}_m=(-)^{\mu}t_{4n+3}t_{4k+3}\phi^{\mu}{}_{m+2n+2k-1}.$$

Then we see that $[\operatorname{ad}t_{4n+3}j^0{}_n, \operatorname{ad}t_{4k+3}j^0{}_k]_+=0$. Thus we have

$$\exp\Big(\sum_{n>0}\operatorname{ad}t_{4n+3}j^0{}_n\Big)=1+\sum_{n>0}\operatorname{ad}t_{4n+3}j^0{}_n.$$

Proceeding the calculation we see that

$$\Big(1+\sum_{n>0}\operatorname{ad}t_{4n+3}j^0{}_n\Big)\phi^{\mu}{}_m=\phi^{\mu}{}_m-(-)^{\mu}\sum_{l>0}t_{4l+3}\phi^{\mu+1}{}_{l+2n+2+\mu},$$

where we regard $\phi^{1+1}{}_n=\phi^0{}_n$, and we have

$$\eta\Big(\exp\Big(\sum_{n>0}t_{4n+3}j^0{}_n\Big)\Big)=\exp\Big(\sum_{n>0}t_{4n+3}\breve{\Gamma}^{4n+3}\Big).$$

One can similarly verify that

$$\eta(\exp(\theta j^0{}_{-1}-xj^e{}_0)=\exp(\theta\breve{\Lambda}+x\breve{\Lambda}^2).$$

The three matrices

$$\exp\Big(\sum_{n>0}t_{4n+3}\breve{\Gamma}^{4n+3}\Big),\ \exp\Big(\sum_{n>0}t_{4n+2}\breve{\Gamma}^{4n+2}\Big),$$

and

$$\exp(\theta\breve{\Lambda}+x\breve{\Lambda}^2)$$

commute with each other. Hence we can write

$$\eta(\Phi) = \exp\left(\theta\check{\Lambda} + x\check{\Lambda}^2 + \sum_{n\equiv 2,3(\mathrm{mod}\,4)} t_n \check{\Gamma}^n\right). \quad \text{(Q.E.D.)}$$

From Lemma D, Eq. (3.7) is

$$'(\eta(g)^{0\mu}{}_{0j})_{j\in\mathbf{Z}}{}_{\mu=0,1} \exp\left(\theta\check{\Lambda} + x\check{\Lambda} + \sum_{n\equiv 2,3(\mathrm{mod}\,4)} t_n \check{\Gamma}^n\right)$$
$$(\eta(g_0)^{\mu\nu}{}_{ij})_{i\in\mathbf{Z}}{}_{j<0\mu,\nu=0,1} = 0. \tag{3.11}$$

From the hypothesis for g_0, (3.11) coincides with the Grassmann equation (3.3). Since $g_0 \in G(\tilde{V},\tilde{V})$, $\eta(g_0)\check{\Xi}_\phi$ $\in$orthogonal USGM$^\phi$. Hence, from Theorem 3.2, the operator

$$W = \sum_{\mu=0}^{1} \sum_{n>0} \eta(g)^{0\mu}{}_{0-n} D^{-2n-\mu}$$

solves the OSp–SKP hierarchy. This completes the proof of Theorem 3.3. Q.E.D.

In Theorem 3.3, we assume that $\eta(g_0)\check{\Xi}_\phi$ belong to USGM$^\phi$. For $g'_0 = 1 - ((-)^i\phi^0{}_i + (-)^j\phi^0{}_{-j}) \times (\phi^0{}_{-i} + \phi^0{}_j), j > i > 0$, one can verify that $\eta(g'_0)\check{\Xi}_\phi$ does not belong to USGM$^\phi$ but to another cell of USGM. Unfortunately it is not unclear how to deduce the OSp–SKP hierarchy in case of $\eta(g_0)\check{\Xi}_\phi \not\subset$USGM$^\phi$. In Ref. 11, Bergvelt constructs the representation of $(\mathrm{gl}(\infty|\infty)\otimes\nu)_0$ on the holomorphic section of dual Berezinian bundle of the USGM. We expect that the OSp–SKP hierarchy can be analyzed over the whole USGM applying his idea to $\mathrm{osp}(\infty|\infty)$.

ACKNOWLEDGMENTS

The author would like to express his sincere thanks to Professor Hirofumi Yamada and the referee for their helpful comments.

[1]K. Ikeda, Lett. Math. Phys. **14**, 321 (1987).
[2]K. Ikeda, "The super Toda lattice hierarchy," to be published in Publ. RIMS.
[3]K. Ueno, H. Yamada, and K. Ikeda, Comm. Math. Phys. **124**, 57 (1989).
[4]V. G. Kac and J. W. van de Leur, Ann. Inst. Fourier **37**, 99 (1987).
[5]V. G. Kac and J. W. van de Leur, *Infinite Dimensional Lie Algebras and Groups*, edited by V. G. Kac, Advanced Series in Mathematical Physics Vol. 7 (World Scientific, Singapore, 1989), pp. 369–406.
[6]K. Ueno and H. Yamada, Adv. Stud. Pure Math. **16**, 373 (1987).
[7]M. Date, M. Jimbo, M. Kashiwara, and T. Miwa, Proc. RIMS *Symposium on Non-Linear Integrable Systems*, 1981, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983), pp. 39–120.
[8]V. G. Kac, *Infinite Dimensional Lie Algebras* (Cambridge U. P.), Cambridge, 1985, 2nd ed.
[9]K. Ueno and K. Takasaki, Adv. Stud. Pure Math. **4**, 1 (1984).
[10]V. G. Kac and D. Peterson, Seminaire de Math. Superiures, Les Presses de Montreal. **102**, 141 (1986).
[11]M. J. Bergvelt, *Infinite Dimensional Lie Algebras and Groups*, edited by V. G. Kac, Advanced Series in Mathematical Physics Vol. 7 (World Scientific, Singapore, 1989), pp. 343–351.
[12]M. Awada and A. H. Chamseddine, Phys. Lett. B **206**, 437 (1988).
[13]K. Takasaki, Lett. Math. Phys. **17**, 351 (1989).

Local supersymmetry in nonrelativistic systems

L. F. Urrutia[a] and J. Zanelli[b]
International Centre for Theoretical Physics, Trieste, Italy

(Received 3 November 1989; accepted for publication 25 April 1990)

Classical and quantum nonrelativistic interacting systems invariant under local supersymmetry are constructed by the method of taking square roots of the bosonic constraints that generate timelike reparametrization, leaving the action unchanged. In particular, the square root of the Schrödinger constraint is shown to be the nonrelativistic limit of the Dirac constraint. Contact is made with standard models of supersymmetric quantum mechanics through the reformulation of locally invariant systems in terms of their true degrees of freedom. Contrary to the field theory case, it is shown that locally invariant systems are completely equivalent to corresponding globally invariant systems, where the latter are the Heisenberg picture description of the former with respect to some fermionic time.

I. INTRODUCTION

Supersymmetric quantum mechanical (SSQM) models viewed as one-dimensional field theories provide explicit realization of the basic supersymmetry algebra.[1] The interest in such SSQM models has been mainly twofold: On one hand we find the search for realistic applications of them,[2] while on the other hand, they are considered merely as a simplified arena where new ideas in supersymmetry are generated, tested, and subsequently generalized. A distinguished example of the latter point of view constitutes the introduction of the Witten index as a characterization for spontaneously broken supersymmetric theories.[3]

Generally speaking, the construction of SSQM models has been characterized by the use of supercharges that are linear in the Grassmann variables and generate a global symmetry of the system. That is to say, the parameters of the induced transformation do not depend on time, being just constant Grassmann numbers. Having in mind the richer structure and interesting possibilities exhibited by systems possessing gauge supersymmetry, i.e., those coupled to supergravity,[4] the question of how to extend the above-mentioned quantum mechanical models to include time-dependent supersymmetry transformations naturally arises. Most of the work along these lines has been done in the realm of the relativistic point particle, which then naturally includes the concept of the relativistic spinning particle because the additional Grassmann variables can be interpreted as intrinsic spin degrees of freedom.[5] There are some recent works that deal with the problem of constructing nonrelativistic systems having local supersymmetry.[6–8] In this paper we present an alternative systematic method for such construction, which is based on the idea that the generators of time-dependent supersymmetry, when considered as constraints on the system, are square roots of the constraints that generate the reparametrization invariance of the same system.[9] Within the context of quantum mechanics, this method has been applied mostly to the relativistic point particle, where the reparametrization invariance of the original action is manifest.[10] The corresponding constraint is the Klein–Gordon operator, while its square root, leading to the fermionic constraint, is the Dirac operator.

One can proceed along similar lines in the nonrelativistic case by considering the free particle as a constrained system invariant under time reparametrization. As is well known, this is achieved, for example, by introducing a new coordinate $t(\tau)$ and rewriting the free particle action as[11]

$$S=\int_0^1 d\tau \frac{1}{2m}\frac{\dot{\mathbf{x}}^2}{\dot{t}}. \tag{1}$$

The condition $\delta S=0$ reproduces the usual dynamical equations if the coordinates $x(\tau)$, $t(\tau)$ are kept fixed at the endpoints $\tau=0$, $\tau=1$. The canonical action is given by

$$S=\int_0^1 d\tau[p_0\dot{t}+\mathbf{p}\cdot\dot{\mathbf{x}}-N\mathscr{H}], \tag{2}$$

where p_0, $\mathbf{p}$ are the momenta canonically conjugated to t, $\mathbf{x}$, respectively; N is a Lagrange multiplier; and

$$\mathscr{H}=p_0+(1/2m)\mathbf{p}^2 \tag{3}$$

is a first-class constraint. The two extra degrees of freedom t and p_0 that we have added to the phase space are removed by constraint (3) upon gauge fixing. The action (2) is invariant under the local transformations generated by $\mathscr{H}$:

$$\delta V=(V,\epsilon(\tau)\mathscr{H}), \tag{4}$$

where V is any function of the canonical variables and $(,)$ denotes the usual Poisson bracket. The Lagrange multiplier transformation is given by $\delta N=\dot{\epsilon}(\tau)$.[12] The parameter $\epsilon(\tau)$ is restricted only at the endpoints by $\epsilon(0)=\epsilon(1)=0$, as required by the action principle. Using the standard language, we refer to the above symmetry of the action as a gauge symmetry, even though one is dealing with a noninternal symmetry.

Upon Dirac quantization, $\mathscr{H}$ becomes the Schrödinger operator and the constraint condition becomes the Schrödinger equation

$$\left(\frac{1}{i}\frac{\partial}{\partial t}+\frac{\mathbf{p}^2}{2m}\right)\psi=0. \tag{5}$$

[a] On sabbatical leave from Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de Mexico, Circuito Exterior, C. U. 04510 Mexico, D. F. Also at Centro de Estudios Científicos de Santiago, Casilla 16443, Santiago, Chile.
[b] Centro de Estudios Científicos de Santiago, Casilla 16443, Santiago, Chile. Also at Departmento de Física, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago.

 0022-2488/90/092271-07$03.00

This can be naturally extended to the interacting case by redefining the constraint as

$$\mathscr{H} = p_0 + H, \tag{6}$$

where H is now the full physical (gauge invariant) Hamiltonian

$$H = (1/2m)\mathbf{p}^2 + V(\mathbf{x}). \tag{7}$$

The paper is organized as follows. In Sec. II we construct the square root of the Schrödinger operator in the free case, which is obtained by taking the nonrelativistic limit of the Dirac operator, as suggested by the fact that the Schrödinger operator is the nonrelativistic limit of the Klein–Gordon operator. In Sec. III the interacting case is disucssed. It is shown that the gauge field associated to the invariance under local supersymmetry can be completely eliminated. This is in contrast with the case for a field theory, where local supersymmetry requires the introduction of a physical spin-3/2 field: the gravitino. In fact, one can see that the local supersymmetric Witten model reduces to the global one in the appropriate coordinates through a finite supersymmetry rotation. This transformation can be interpreted as the passage from a Schrödinger to a Heisenberg picture with respect to some fermionic time. Finally, Sec. IV contains a short summary and the conclusions.

II. THE SQUARE ROOT OF THE FREE SCHRÖDINGER EQUATION

In Sec. I we briefly reviewed how the Schrödinger equation can be understood as a bosonic constraint restricting the allowed states in a reparametrization invariant description of a nonrelativistic system. This is a useful remark that allows us to construct a locally supersymmetric action for a nonrelativistic system. There is a standard procedure for constructing a locally supersymmetric extenstion of a bosonic system invariant under general coordinate transformations: The generators of the local supersymmetry transformations are the square roots of the bosonic constraints responsible for the invariance under general coordinate transformations.[9]

In the case of the relativistic point particle such a procedure starts from the Klein–Gordon equation as the original bosonic constraint and leads to the Dirac equation as the resulting fermionic constraint. Both constraints obey a closed graded supersymmetry algebra. Having in mind that the Schrödinger equation is the nonrelativistic limit of the Klein–Gordon equation, we look for a fermionic constraint that can be obtained as the corresponding nonrelativistic limit of the Dirac operator. We first study the noninteracting case in order to determine the basic structure of the constraints.

Let us consider the following form of the Dirac equation:

$$(i\gamma_5\gamma^\mu p_\mu + m\gamma_5)\psi = 0, \tag{8}$$

where $p_\mu = (1/i)\partial_\mu$, $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$, $\eta^{\mu\nu} = \text{diag}(-,+,+,+)$, $\gamma_5 = \gamma^0\gamma^1\gamma^2\gamma^3$, $\gamma_5^2 = -1$, and $\hbar = 1$. Using a representation for the Dirac matrices given by

$$\gamma^0 = i\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \quad \gamma^i = i\begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \quad \gamma_5 = -i\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix},$$

we obtain the nonrelativistic limit for Eq. (8):

$$\begin{pmatrix} \boldsymbol{\sigma}\cdot\mathbf{p} & p_0 \\ 2m\cdot & -\boldsymbol{\sigma}\cdot\mathbf{p} \end{pmatrix}\begin{pmatrix} \psi \\ \chi \end{pmatrix} = 0, \tag{9}$$

where $p_0 = (1/i)(\partial/\partial t) \ll m$.

Defining

$$\hat{\mathscr{S}} = -\frac{i}{2\sqrt{m}}\begin{pmatrix} \boldsymbol{\sigma}\cdot\mathbf{p} & p_0 \\ 2m & -\boldsymbol{\sigma}\cdot\mathbf{p} \end{pmatrix}, \tag{10}$$

we easily verify that

$$\{\hat{\mathscr{S}}, \hat{\mathscr{S}}\} = -(\mathbf{p}^2/2m + p_0)I \equiv -\hat{\mathscr{H}}I, \tag{11a}$$

$$[\hat{\mathscr{H}}, \hat{\mathscr{S}}] = 0, \quad [\hat{\mathscr{H}}, \hat{\mathscr{H}}] = 0. \tag{11b}$$

Here, I is the 4×4 identity matrix. The normalization of $\hat{\mathscr{S}}$ is such that when we consider the classical theory according to the prescription

$$(1/i)[\hat{A}, \hat{B}]_\pm \to (A, B), \tag{12}$$

with $\hat{A}\to A$, the first relation in (11) reads as $(\mathscr{S}, \mathscr{S}) = i\mathscr{H}$. In order to better understand the nature of the symmetries generated by $\hat{\mathscr{S}}$, we look at the classical limit. Following Ref. 13 we introduce the new variables

$$\hat{\theta}^\mu = (1/\sqrt{2})i\gamma_5\gamma^\mu,$$

$$\hat{\theta}_5 = (1/\sqrt{2})\gamma_5, \tag{13}$$

which allow us to rewrite the fermionic constraint as

$$\hat{\mathscr{S}} = (1/\sqrt{2m})(\mathbf{p}\cdot\hat{\boldsymbol{\theta}} + p_0(\hat{\theta}_5 + \hat{\theta}^0)/2 + 2m(\hat{\theta}_5 - \hat{\theta}^0)/2). \tag{14}$$

The classical limit is now obtained by considering the bosonic operators x^μ, p_μ as real numbers with the usual Poisson bracket relations, while the fermionic operators are replaced by the real Grassmann variables θ^μ, θ_5. According to prescription (12), the only nonvanishing Poisson brackets of θ^μ, θ_5 are[13]

$$(\theta^\mu, \theta^\nu) = i\eta^{\mu\nu}, \quad (\theta_5, \theta_5) = i. \tag{15}$$

The physical system that we are considering at this stage is the free, nonrelativistic spinning point particle of mass m. The fact that a constrained description is employed means that the phase space contains extra coordinates, bosonic as well as fermionic, besides the dynamical ones $(\mathbf{x}, \mathbf{p}, \boldsymbol{\theta})$ and that the action is to be varied with respect to all of them. The classical version of the algebra (11) ensures that both constraints are first class and consequently, we need two extra variables of each type: t, p_0 and θ^0, θ_5, respectively. The fermionic contribution to the kinetic part of the action is

$$\int_0^1 d\tau \frac{i}{2}(\dot{\boldsymbol{\theta}}\cdot\boldsymbol{\theta} - \dot{\theta}^0\theta^0 + \dot{\theta}_5\theta_5). \tag{16}$$

We introduce the following combinations of the extra fermionic variables:

$$\theta_+ = \tfrac{1}{2}(\theta_5 + \theta^0), \quad \theta_- = \theta_5 - \theta^0; \tag{17}$$

the corresponding kinetic term in the Lagrangian is

$$-\dot\theta^0\theta^0+\dot\theta_5\theta_5=2\dot\theta_+\theta_- -\frac{d}{d\tau}(\theta_+\theta_-). \tag{18}$$

The complete action for our system, which generalizes (2), is then

$$S=\int_0^1 d\tau\Big[\dot{\mathbf{x}}\cdot\mathbf{p}+\dot t p_0+\frac{i}{2}\dot{\boldsymbol\theta}\cdot\boldsymbol\theta+i\dot\theta_+\theta_- - N(\tau)\Big(p_0+\frac{\mathbf{p}^2}{2m}\Big)-iM(\tau)\frac{1}{\sqrt{2m}} \times(\boldsymbol\theta\cdot\mathbf{p}+\theta_+p_0+\theta_-m)\Big]+\frac{i}{2}\,[\boldsymbol\theta(0)\cdot\boldsymbol\theta(1) +(\theta_+(0)-\theta_+(1))(\theta_-(0)+\theta_-(1))], \tag{19}$$

where $M(\tau)$ is a fermionic Lagrange multiplier. The variations at the endpoints are restricted so that the action is stationary under arbitrary changes of the coordinates around a classical trajectory for $0<\tau<1$; they are

$$\delta b(0)=\delta b(1)=0,\quad \delta f(0)+\delta f(1)=0, \tag{20}$$

where b and f denote the bosonic coordinates $\mathbf{x}$, t and the fermionic coordinates $\boldsymbol\theta$, θ_+, θ_-, respectively. Conditions (20) are chosen in such a way that they provide a unique solution for the equations of motion.[10]

The action (19) possesses two kinds of local symmetries generated by

$$\mathscr{H}=p_0+\mathbf{p}^2/2m, \tag{21a}$$

$$\mathscr{S}=(1/\sqrt{2m})(\boldsymbol\theta\cdot\mathbf{p}+\theta_+p_0+\theta_-m), \tag{21b}$$

respectively. The local supersymmetry transformations of the dynamical variables are induced by $\mathscr{S}$ and given by

$$\delta x^i=(1/\sqrt{2m})\eta\theta^i,\quad \delta p_i=0,$$
$$\delta t=(1/\sqrt{2m})\eta\theta_+,\quad \delta p_0=0,$$
$$\delta\theta^i=-(i/\sqrt{2m})\eta p_i,\quad \delta\theta_+=-i(\sqrt{m/2})\eta,$$
$$\delta\theta_-=-(i/\sqrt{2m})p_0\eta,\quad \delta N=iM\eta,\quad \delta M=\dot\eta, \tag{22}$$

where $\eta=\eta(\tau)$ is an arbitrary local Grassmann parameter restricted by $\eta(0)=\eta(1)=0$ because of the fixed endpoint conditions.

The action (19) can be rewritten in second-order form by eliminating $\mathbf{p}$, p_0, N in favor of the velocities from the following equations of motion:

$$N=\dot t-iM\theta_+/\sqrt{2m}\,, \tag{23a}$$

$$p_0=-\mathbf{p}^2/2m, \tag{23b}$$

$$\frac{p_i}{m}=\frac{\dot x^i}{\dot t}+\frac{iM}{\sqrt{2m}}\Big(\frac{\dot x^i}{\dot t^{\,2}}\theta_+-\frac{\theta^i}{\dot t}\Big). \tag{23c}$$

There is no conflict between the already prescribed endpoint conditions and the defining relations (23). Substituting (23) into (19) we obtain

$$S=\int_0^1 d\tau\Big(\frac{m}{2}\frac{\dot{\mathbf{x}}^2}{\dot t}+\frac{i}{2}\dot{\boldsymbol\theta}\cdot\boldsymbol\theta+i\dot\theta_+\theta_- +iM\sqrt{\frac{m}{2}}\Big(-\frac{\dot{\mathbf{x}}\cdot\boldsymbol\theta}{\dot t}+\frac{\dot{\mathbf{x}}^2}{2\dot t^{\,2}}\theta_+-\theta_-\Big)\Big)+\mathrm{BT}, \tag{24}$$

where BT stands for the boundary terms.

The result (24) coincides with the starting point taken in Ref. 8 in discussing the free spinning nonrelativistic particle.

It is sometimes found in the literature that $e=\dot t$ is considered as an independent coordinate[6,7] However, this is incorrect unless e is also the Lagrange multiplier that goes together with the generator of time reparametrizations in the canonical action.

In closing, let us identify the true degrees of freedom of the theory described by the action (19). This can be done by rewriting the action of the system in terms of supersymmetric invariant quantities.[14] In order to achieve this, the variables p_0, N, M are substituted in the action (19) using Eqs. (23a), (23b), and the equation obtained by varying θ_-. The result is

$$S=\int_0^1 d\tau\Big[\dot{\mathbf{x}}\cdot\mathbf{p}-\frac{\mathbf{p}^2}{2m}\dot t+\frac{i}{2}\dot{\boldsymbol\theta}\cdot\boldsymbol\theta-\frac{i}{m}\dot\theta_+ \times\Big(\boldsymbol\theta\cdot\mathbf{p}-\frac{\mathbf{p}^2}{2m}\theta_+\Big)\Big]+\mathrm{BT}. \tag{25}$$

We observe that θ_- has dropped out from the action: The only reference to this variable remains in the boundary term through the combination $\alpha_-=\theta_-(0)+\theta_-(1)$. When the action (25) is varied with respect to θ_+, we find the boundary contribution

$$-i(\delta\theta_+/m)(\boldsymbol\theta\cdot\mathbf{p}-(\mathbf{p}^2/2m)\theta_+)\big|_0^1 +(i/2)(\delta\theta_+(0)-\delta\theta_+(1))\alpha_-, \tag{26}$$

which must be set equal to zero. Recalling the endpoint restriction $\delta\theta_+(0)+\delta\theta_+(1)=0$ we obtain

$$\delta\theta_+(1)[\theta_-(1)+\theta_-(0)-\alpha_-]=0. \tag{27}$$

Here, θ_- is a shorthand notation for

$$-(1/m)(\boldsymbol\theta\cdot\mathbf{p}-(\mathbf{p}^2/2m)\theta_+). \tag{28}$$

It would be incorrect to consider (28) as a definition of θ_- valid for the whole history of the system because the condition $\delta(\theta_-(1)+\theta_-(0))=0$ is not recovered from (28). Nevertheless, imposing the correct boundary condition (27) effectively means that we are using such a definition, but only at the endpoints.

Returning to the action (25) we notice that it is invariant under reparametrization and also under the following local supersymmetry transformations:

$$\delta\mathbf{x}=(i/m)\eta(\tau)(\boldsymbol\theta-\mathbf{p}(\theta_+/m)),\quad \delta\mathbf{p}=0,\quad \delta t=0,$$
$$\delta\boldsymbol\theta=\eta(\tau)(\mathbf{p}/m),\quad \delta\theta_+=\eta(\tau), \tag{29}$$

as can be verified by direct substitution. The transformations (29) are generated by $s=\Pi_+ +i(\boldsymbol\theta\cdot\mathbf{p}-(\mathbf{p}^2/2m)\theta_+)$, which arises as a constraint $s\simeq 0$ from the action (25). This results when Π_+, the momentum canonically conjugated to θ_+, is defined to be the left derivative of the Lagrangian in (25) with respect to $\dot\theta_+$, as is usually done. The constraint s must be proportional to $\mathscr{S}$, with $p_0=-\mathbf{p}^2/2m$ in the latter. This allows us to identify $\Pi_+=i\theta_-$, together with the Poisson bracket $(\theta_+,\Pi_+)=-1$. Let us remark that Π_+ is purely imaginary with the conventions adopted.

Let us now introduce the combinations

$$\mathbf{X}=\mathbf{x}-(i/m)\theta_+\boldsymbol\theta,\quad \boldsymbol\Theta=\boldsymbol\theta-(\mathbf{p}/m)\theta_+, \tag{30}$$

which are invariant under the local transformations (29). In terms of these new variables, the action (25) is rewritten as

$$S \equiv \int_0^1 d\tau \left[\dot{\mathbf{X}} \cdot \mathbf{p} - \frac{\mathbf{p}^2}{2m} \dot{t} + \frac{i}{2} \dot{\boldsymbol{\Theta}} \cdot \boldsymbol{\Theta} \right] + \mathrm{BT}, \tag{31}$$

which corresponds to the first-order formulation of noninteracting three-dimensional bosonic and fermionic degrees of freedom and where all information concerning local supersymmetry invariance is lost.

III. LOCAL SUPERSYMMETRIC SYSTEMS WITH INTERACTIONS

In this section we extend the previous results in order to include interactions which will produce after quantization the locally invariant generalization of the standard supersymmetric quantum mechanics.[1] To do this we redefine the fermionic constraint (21b) in the form

$$\mathscr{S} = (1/\sqrt{2m})(Q + p_0\theta_+ + m\theta_-) \simeq 0, \tag{32}$$

where the supercharge Q is required to be a function of the variables $\mathbf{x}$, $\mathbf{p}$, $\boldsymbol{\theta}$ only. In the general case we would like to maintain the basic square root relation between $\mathscr{S}$ and $\mathscr{H}$, where the constraint

$$\mathscr{H} = p_0 + H(\mathbf{x}, \mathbf{p}, \boldsymbol{\theta}) \simeq 0 \tag{33}$$

now includes interactions. The already assumed dependence of Q upon the dynamical variables implies that the supercharge Q turns out to be the square root of the physical Hamiltonian H. That is,

$$H = (1/2im)(Q, Q). \tag{34}$$

As a consequence of the Jacobi identity and Eq. (34) we obtain

$$(Q, H) = 0. \tag{35}$$

Equation (35) implies that the constraints $\mathscr{S}$ and $\mathscr{H}$ remain first class, exactly as in the noninteracting case.

We recognize the basic structure of Witten supersymmetric quantum mechanics[15] in Eqs. (34) and (35). It is now possible to promote this supersymmetry to a local one generated by $\mathscr{S}$.

The above construction can also be generalized to include n fermionic constraints $\mathscr{S}_a$, $a = 1, \ldots, n$, which are required to satisfy the following algebra:

$$(\mathscr{S}_a, \mathscr{S}_b) = i\delta_{ab}\mathscr{H}, \quad (\mathscr{S}_a, \mathscr{H}) = 0. \tag{36}$$

A realization of an algebra such as (36) is given by extending definition (32) in an obvious manner:

$$\mathscr{S}_a = (1/\sqrt{2m})(Q_a + p_0\theta_{+a} + m\theta_{-a}), \tag{37}$$

where Q_a is still a function of $\mathbf{x}$, $\mathbf{p}$, $\boldsymbol{\theta}$ only. The n supercharges Q_a are such that

$$(Q_a, Q_b) = 2im\delta_{ab}H, \quad (Q_a, H) = 0, \tag{38}$$

which again guarantee that $\mathscr{S}_a$ and $\mathscr{H}$ are first-class constraints.

We will further comment on Eqs. (38) after we discuss the one-dimensional locally invariant Witten model as a particular realization of the previous ideas. The Witten model[15] corresponds to $n = 2$, with

$$Q_1 = \theta_1 p + \theta_2 V(x), \quad Q_2 = \theta_2 p - \theta_1 V(x) \tag{39}$$

leading to the physical Hamiltonian

$$H = \tfrac{1}{2}(p^2 + V^2 + 2i\theta_1\theta_2 V'), \tag{40}$$

with $V' = dV/dx$ and $m = 1$.

The action for the system is

$$S = \int_0^1 d\tau \left[\dot{q}p + i\dot{p}_0 + \frac{i}{2}\dot{\theta}\theta + \dot{\theta}_{+a}\theta_{-a} - N\mathscr{H} - M_a\mathscr{S}_a \right] + \mathrm{BT}. \tag{41}$$

The action (41) possesses the same two local gauge invariances as the free case: (i) under the transformations generated by $\mathscr{S}_a$ and (ii) under time reparametrizations generated by $\mathscr{H}$. This fact, together with the explicit form (40) of the physical Hamiltonian, allows us to interpet the action (41) as describing the original Witten model having its global supersymmetry promoted to a gauge supersymmetry.

In order to distinguish the true dynamical degrees of freedom from the gauge variables associated to local supersymmetry, we proceed in complete analogy to the free case and eliminate the variables N, p_0, and M_a from (41). The result is

$$S = \int_0^1 d\tau \left[\dot{q}p - H\dot{t} + \frac{i}{2}\dot{\theta}\theta - i\dot{\theta}_{+a}(Q_a - H\theta_{+a}) \right] + \mathrm{BT}. \tag{42}$$

Again, the variables θ_{-a} automatically drop out from the action and the corresponding boundary conditions α_{-a} are correctly recovered in the same manner as described following Eq. (25).

The action (42) is invariant under the local supersymmetry transformations

$$\delta x = i\epsilon_a(\theta_a - p\theta_{+a}), \quad \delta p = i\epsilon_a\left(-\tilde{\theta}_a V' + \frac{\partial H}{\partial x}\theta_{+a} \right),$$
$$\delta t = 0, \quad \delta\theta_a = \epsilon_a p - \tilde{\epsilon}_a V + iV'\tilde{\theta}_a\epsilon_b\theta_{+b}, \tag{43}$$

generated by the constraints $s_a = \Pi_{+a} + i(Q_a - H\theta_{+a})$ arising from the action (42) in a manner similar to the noninteracting case discussed after Eq. (29). Here, $\epsilon_a = \epsilon_a(\tau)$ and $\tilde{\theta}_a = \epsilon_{ab}\theta_b$, with $\epsilon_{12} = -\epsilon_{21} = 1$. Now we can introduce new variables (denoted by the corresponding capital letter) that are invariant under the transformations (43) and generalize the analogous expressions of the free case given in (30): They are

$$X = x - i\theta_{+a}\theta_a - iV\theta_{+1}\theta_{+2},$$
$$P = p(1 - iV'\theta_{+1}\theta_{+2}) + iV'(\theta_1\theta_{+2} - \theta_2\theta_{+1}),$$
$$\Theta_a = \theta_a - p\theta_{+a} + V\tilde{\theta}_{+a} + iV'\theta_a\theta_{+1}\theta_{+2}, \tag{44}$$

and can be shown to satisfy the Poisson brackets corresponding to the bosonic and fermionic canonical degrees of freedom. In terms of these variables, the action (42) reduces to

$$S = \int_0^1 d\tau \left[\dot{X}P - \overline{H}\dot{t} + \frac{i}{2}\dot{\Theta}_a\Theta_a \right] + \frac{i}{2}\Theta_a(1)\Theta_a(0), \tag{45}$$

where $\overline{H} \equiv H(x(X, P, \Theta), p(X, P, \Theta), \theta(X, P, \Theta))$. The result is

$$\overline{H} = \tfrac{1}{2}(P^2 + V^2(X) + 2i\Theta_1\Theta_2 V'(X)). \tag{46}$$

Let us observe that the functional form of $\overline{H}(X, P, \Theta)$ is the same as that of $H(x, p, \theta)$.

If the action (45) is deparametrized, setting $\dot{t} = 1$ in the Lagrangian, the result is the canonical first-order action of an unconstrained system with the phase space coordinates $X(t)$, $P(t)$, $\Theta_a(t)$ and Hamiltonian given by (46). At this level of the description, all reference to the previous invariance under local supersymmetry is lost and only a global supersymmetry remains. This global invariance cannot be considered as a restriction of the local supersymmetry because, for example, it would be inconsistent with the fixed coordinate conditions at the endpoints. However, the algebra of the global symmetry is isomorphic to (36):

$$(\overline{Q}_a, \overline{Q}_b) = i\delta_{ab}\,\overline{H}, \tag{47a}$$

$$(\overline{Q}_a, \overline{H}) = 0, \tag{47b}$$

where $\overline{Q}_a = Q_a - 2iH\theta_{+a}$ is obtained by replacing $x^i \to X^i$, $p_i \to P_i$, $\theta^i \to \Theta^i$ in Q_a. It would not surprise us to find that under the change of variables (44) H is form invariant, whereas Q_a is not. The reason is that (44) is a canonical transformation (generated by Q_a) and H commutes with its generator, while Q_a does not. Now it is easy to see that the supersymmetry transformation generated by $\overline{Q}_a$ on each variable produce a net shift at the endpoints of the action (45), as appropriate to a global symmetry. These transformations are precisely those of global supersymmetry in the Witten model.

Nevertheless, our calculation has shown that the local supersymmetric Witten model is just a parametrized form of the globally invariant model and hence they are physically equivalent. In order to understand this point further let us discuss the quantum version of the theory described by the action (41). For simplicity we will assume only one supercharge Q ($n = 1$), so that the classical properties previously discussed will have a more transparent origin. We proceed in the most direct way by changing the Poisson brackets into commutators or anticommutators and by imposing the first-class constraints as null conditions upon the wavefunction that depends on the real coordinates t, x, θ, θ_+. We recall that θ, θ_+ are Grassmann numbers with the derivatives taken from the left and that they anticommute (commute) with every fermionic (bosonic) operator. The wavefunction then satisfies

$$\left(\hat{Q} + \hat{p}_0\theta_+ + m\frac{\partial}{\partial\theta_+}\right)\psi = 0, \tag{48a}$$

$$(\hat{p}_0 + \hat{H})\psi = 0, \tag{48b}$$

where the Schrödinger equation (48b) is a consequence of Eq. (48a).

The fact that in the classical case we were able to effectively eliminate the coordinate θ_+ has as a counterpart here the fact that Eq. (48a) has the general solution

$$\psi = e^{-\theta_+\hat{Q}/m}\phi(x, \theta, t). \tag{49}$$

Substituting (49) into (48a) we obtain

$$\theta_+(\hat{p}_0 + \hat{H})\phi = 0, \tag{50}$$

which is satisfied in terms of the Schrödinger equation for ϕ obtained from (48b).

The solution (49) suggests that the coordinate θ_+ can be interpreted as a kind of fermionic time whose dynamics is governed by the evolution operator $\hat{Q}$.[16] In this sense, the description of the system in terms of ψ or ϕ is like going from the Schrödinger to the Heisenberg picture with respect to this fermionic time. To carry the analogy further we can shift the θ_+ independent operators $\hat{a}$ to their Heisenberg picture expression by means of the unitary transformation

$$\hat{A}(\theta_+) = e^{+\theta_+\hat{Q}/m}\,\hat{a}\,e^{-\theta_+\hat{Q}/m}. \tag{51}$$

This expression is the quantum mechanical analog of the classical variables, invariant under local supersymmetry, introduced in (44). It should be stressed that this supersymmetry is the one generated by $\hat{s} = \Pi_+ + i(\hat{Q} - i\hat{H}\hat{\theta}_+)$. Using the explicit expression (51) it is possible to verify that $\delta\hat{A} = [\hat{A}, \eta(\tau)\hat{s}]$ is indeed equal to zero for any bosonic for fermionic operator.

The quantum analog of the description of the system in terms of the classical action given by Eqs. (45) and (46) for the $n = 1$ case corresponds to the use of a full Heisenberg representation for both the bosonic and fermionic times t and θ_+, respectively. The transition from the Schrödinger picture ($\psi(x, \theta, \theta_+, t)$, $\hat{a}$) to the Heisenberg picture ($\phi(x, \theta)$, $\overline{A}(t, \theta_+)$) is achieved by means of the following unitary transformation:

$$\bar{\phi}(x, \theta) = V^\dagger\psi(x, \theta, t, \theta_+), \quad \hat{\overline{A}}(t, \theta_+) = V\hat{a}V^\dagger, \tag{52}$$

where

$$V = e^{i\hat{H}t}e^{-\hat{\theta}_+\hat{Q}/m}. \tag{53}$$

The wavefunction $\bar{\phi}(x, \theta)$ describes the initial condition of the system at $t = 0$ and $\theta_+ = 0$ and the arguments x, θ are the eigenvalues of the operators $\hat{\overline{X}}$, $\hat{\overline{\Theta}}$ at this particular instant. The dynamical evolution is now fully contained in the operators $\hat{\overline{A}}$, which are the quantum analogs of the classical canonical variables $A(t, \theta_+)$ introduced in Eq. (44). In particular, the physical evolution with respect to the real time t is given by the Hamiltonian

$$\hat{\overline{H}} = e^{-i\hat{H}t}e^{\theta_+\hat{Q}/m}\hat{H}(x, p, \theta)e^{-\theta_+\hat{Q}/m}e^{i\hat{H}t}. \tag{54}$$

By explicit application of Eq. (52), we conclude that $\hat{\overline{H}} = \hat{H}(\hat{\overline{X}}, \hat{\overline{P}}, \hat{\overline{\Theta}})$. On the other hand, using $[\hat{Q}, \hat{H}] = 0$ we obtain $\hat{\overline{H}} = \hat{H}(x, p, \theta)$. In this way we have recovered in quantum mechanics the form invariance of $\hat{\overline{H}}$ and $\hat{H}$ which we found previously at the classical level.

The above exercise using a simple model with one supercharge shows that given a globally invariant supersymmetric theory [which means knowledge of the corresponding algebra (38)] it is always possible to construct a locally invariant extension of it by taking the square root of the Schrödinger operator in the form given by Eq. (37). Nevertheless, both theories are in fact the same because the locally invariant extension turns out to be just a Heisenberg picture description of the globally invariant original theory. This result can be easily extended to the general case of a system possessing n-global supersymmetry and any number of physical degrees of freedom. Again, the locally invariant extension is con-

structed by adding $2n$ extra variables θ_{+a}, θ_{-a} and defining the generators of local supersymmetry as in Eq. (37). When quantizing the system, there will be one equation of the form (48a) for each of the n constraints and the Schrödinger equation (48b) will be the square of any one of them. The general solution of the n fermionic equations will again be of the form (49), with $\theta_+\hat{Q}$ replaced by $\theta_{+a}\hat{Q}_a$ and ϕ depending on the original physical degrees of freedom only. The reason why this method works is that the graded algebra (38) implies $[\theta_{+a}\hat{Q}_a, \theta_{+b}\hat{Q}_b] = 0$ for fixed a and b with $a \neq b$. Now, the local extension of the original theory will correspond to a Heisenberg picture defined by n fermionic times θ_{+a}.

IV. CONCLUSIONS

We have presented a systematic way to construct nonrelativistic systems having bosonic and fermionic degrees of freedom $(\mathbf{x}, \mathbf{p}, \theta)$ that are invariant under local supersymmetry transformations. This is achieved by reformulating the original bosonic problem in terms of a gauge-type action that is invariant under time reparametrizations, in analogy with the relativistic point particle. The associated constraint is the Schrödinger operator. Subsequently, and according to Refs. 9 and 10, local supersymmetry is introducted by constructing the corresponding generators $\mathscr{S}_a$ as square roots of the Schrödinger constraint in the sense of satisfying the graded algebra of first-class constraints given in Eq. (36). These generators are obtained as the nonrelativistic limit of the Dirac operator and hence local supersymmetry is understood as the noninternal gauge symmetry associated with these fermionic contraints. Consequently, the fermionic phase space also had to be enlarged by adding the variables θ_{+a} and θ_{-a} for each supersymmetry generator $\mathscr{S}_a$ considered, in complete analogy with the bosonic situation.

The basic structure underlying local supersymmetry is understood from the analysis of the free nonrelativistic spinning particle. An important point that arises from the noninteracting situation is that a reparametrization-invariant description of nonrelativistic systems is correctly achieved only in terms of a velocity $\dot{t}$, according to Eq. (1), for example, instead of a coordinate e, as sometimes done in the literature.

The construction is extended to include interactions by redefining the constraints $\mathscr{S}_a$ [cf. Eq. (37)] via the introduction of the supercharges Q_a, which are assumed to depend upon the physical degrees of freedom only. The closure of the graded algebra of first-class constraints $\mathscr{H}$, $\mathscr{S}_a$ then requires that the physical Hamiltonian H and the supercharges Q_a satisfy the graded algebra of standard SSQM. This algebra is the usual starting point for discussing global supersymmetry and in our approach is recovered from the locally invariant formulation.

In particular, the local supersymmetric version of the Witten model is analyzed [cf. Eq. (41)] in order to better elucidate the relation between local and global supersymmetry. The basic idea is to identify the true degrees of freedom of the model by rewriting the action in terms of canonical variables invariant under local supersymmetry.[14] After this is done, a first-order action corresponding precisely to the global Witten model described by the physical Hamiltonian $\bar{H} = (1/2i)(\bar{Q}_1, \bar{Q}_1) = (1/2i)(\bar{Q}_2, \bar{Q}_2)$ is found. Besides, in terms of these new variables, the Hamiltonian $\bar{H}$ has the same functional form as the one appearing in the original constraint $\mathscr{H}$. All reference to local supersymmetry is then lost and the extra variables introduced previously are completely eliminated from the action.

The above result is an indication that in the nonrelativistic case, any system invariant under local supersymmetry is equivalent to the corresponding globally invariant one constructed by using gauge-invariant degrees of freedom. In order to test this idea at the quantum level we finally considered the quantization of a simple model with the constraints $\mathscr{H}$ and $\mathscr{S}$. In this case, the relation between the locally and globally invariant descriptions is achieved by means of a unitary transformation that can be interpreted as describing the passage from a Schrödinger to a Heisenberg picture defined with respect to the real time t, together with a fictitious fermionic time θ_+. The new operators in this Heisenberg picture correspond to the canonical variables invariant under local supersymmetry introduced at the classical level. Naturally, the form invariance of the respective Hamiltonians is a direct consequence of the unitary transformation involved. This argument was generalized to arbitrary global systems and no inconsistencies, as previously reported,[6] were found.

The general point of view taken in this article starts with the well-known assertion that any bosonic Lagrangian theory can be written in a parametrized form, for instance, introducing an extra time parameter τ, as is done here, or the extra space-time coordinates τ, σ^i. The resulting theory naturally possesses a gauge invariance corresponding to the group of reparametrizations and, therefore, has a set of first-class constraints satisfying a closed algebra. The results obtained in the case of SSQM discussed in this article suggest that the square root of such constraints will give rise to a local supersymmetric theory that is merely equivalent to a globally invariant theory. The passage from the local to the global theory could then be obtained by deparametrizing the former (e.g., setting $t = \tau, \theta_+ = 0$). This deparametrization cannot be achieved by a gauge transformation (it is not a gauge choice) and corresponds to a canonical transformation to gauge invariant (physical) coordinates in phase space. In quantum mechanics this transformation is seen to correspond to a unitary transformation, which can be interpreted as the change from a Schrödinger to a Heisenberg picture in some fermionic time.

Clearly, local and global supersymmetry give rise to different dynamical systems if the former theory cannot be deparametrized. This happens, for instance, when the parameters are coordinates on a manifold with dynamics of its own, i.e., strings, membranes, gravity, etc. This explains why in the case of a point particle the two theories are equivalent: The world line of the particle is flat and can sustain no geometrodynamics.

ACKNOWLEDGMENTS

The authors would like to thank C. Teitelboim for many enlightening discussions, suggestions, and correspondence regarding this work. They would also like to thank Professor

Abdus Salam, the International Atomic Energy Agency, and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste.

Author LFU was partially supported by a grant from the International Centre for Theoretical Physics, Trieste to the Centro de Estudios Científicos de Santiago (CECS) in the form of a Visiting Scholar fellowship. He also wants to thank C. Teitelboim for the warm hospitality extended to him at CECS.

This work was supported in part by FONDECYT-Chile under Contract Nos. 297/88 and 927/89 and also by CONACYT-Mexico under Contract No. PCEXCEU-022621.

[1]M. de Crombrugge and V. Rittenberg, Ann. Phys. (NY) **151**, 99 (1983).
[2]See, for example, L. E. Gendenshtein and I. V. Krive, Sov. Phys. Usp. **28**, 645 (1985).
[3]E. Witten, Nucl. Phys. B **202**, 253 (1982).
[4]D. Z. Freedman, P. van Nieuwenhuizen, and S. Ferrera, Phys. Rev. D **13**, 3214 (1976); S. Deser and B. Zumino, Phys. Lett. B **62**, 335 (1976).
[5]See, for example, R. Casalbuoni, Nuovo Cimento A **33**, 115 (1976); Nuovo Cimento A **33**, 286 (1976); F. A. Berezin and M. S. Marinov, Ann. Phys. (NY) **104**, 336 (1977).
[6]E. Alvarez, Phys. Rev. D **29**, 320 (1984).
[7]P. Mertens, *Mechanische and Quantenmechanische Systeme mit lokaler Supersymmetrie* (Diplomarbeit, Universitat Bonn, Physikalisches Institute, June 1986).
[8]J. Gomis and M. Novell, Phys. Rev. D **33**, 2212 (1986).
[9]C. Teitelboim, Phys. Rev. Lett. **38**, 1106 (1977); R. Tabensky and C. Teitelboim, Phys. Lett. B **69**, 453 (1977); J. Gamboa and J. Zanelli, Ann. Phys. (NY) **188**, 239 (1988).
[10]C. A. P. Galvao and C. Teitelboim, J. Math. Phys. **21**, 1863 (1980).
[11]See, for example, C. Lanczos, *The Variational Principles of Mechanics* (Univ. of Toronto, Toronto, 1962), 2nd ed., p. 133.
[12]C. Teitelboim, Phys. Rev. D **25**, 3159 (1982).
[13]A. Barducci, R. Casalbuoni, and L. Lusanna, Nuovo Cimento A **35**, 377 (1976); L. Brink, S. Deser, B. Zumino, P. DiVecchia, and P. Howe, Phys. Lett. B **64**, 435 (1976); F. A. Berezin and M. S. Marinov, Ann. Phys. (NY) **104**, 336 (1977).
[14]C. Bachas, Phys. Lett. B **29**, 62 (1978).
[15]E. Witten, Nucl. Phys. B **188**, 513 (1981).
[16]C. Teitelboim, Phys. Lett. B **96**, 77 (1980).

Character formulas for irreducible modules of the Lie superalgebras sl(m/n)

J. Van der Jeugt[a)]
Faculty of Mathematical Studies, University of Southampton, Southampton SO9 5NH, United Kingdom and Laboratorium voor Numerieke Wiskunde en Informatica, Rijksuniversiteit Gent, Krijgslaan 281-S9, 9000 Gent, Belgium

J. W. B. Hughes
School of Mathematical Sciences, Queen Mary College and Westfield, Mile End Road, London E1 4NS, United Kingdom

R. C. King
Faculty of Mathematical Studies, University of Southampton, Southampton SO9 5NH, United Kingdom

J. Thierry-Mieg
Groupe d'Astrophysique Relativiste, CNRS, Observatoire de Paris-Meudon, F-92195 Meudon, France

(Received 13 November 1989; accepted for publication 14 February 1990)

Kac distinguished between typical and atypical finite-dimensional irreducible representations of the Lie superalgebras sl(m/n) and provided an explicit character formula appropriate to all the typical representations. Here, the range of validity of some character formulas for atypical representations that have been proposed are discussed. Several of them are of the Kac–Weyl type, but then it is proved that all formulas of this type fail to correctly give the character of one particular atypical representation of sl(3/4). Having ruled out, therefore, all such formulas, a completely new extension of the Kac–Weyl character formula is proposed. The validity of this formula in the case of all covariant tensor irreducible representations is proved, and some evidence in support of the conjecture that it covers all irreducible representations of sl(m/n) is presented.

I. INTRODUCTION

Lie superalgebras and their representations play an important role in the understanding and exploitation of supersymmetry in physical systems. An early review of their use was given by Corwin *et al.*[1] Since then, the Lie superalgebras of the type under consideration here, namely sl(m/n), have found applications, for example, in quantum mechanics,[2] nuclear physics,[3] particle physics,[4] and string theory.[5]

A complete classification of the finite-dimensional simple Lie superalgebras over $\mathbb{C}$ has been given by Kac[6–8] and by Scheunert *et al.*[9,10] Kac showed that the subclass of these with a nondegenerate bilinear form, called basic classical Lie superalgebras, are closely analogous to the finite-dimensional complex simple Lie algebras. In particular, they can be constructed from a (super)Cartan matrix or, equivalently, from a Kac–Dynkin diagram.[7,11] However, unlike the Lie algebra case the Kac–Dynkin diagram of a given Lie superalgebra is not unique, but depends on the choice of a particular Borel subalgebra.

In his seminal paper[12] on the representations of the basic classical Lie superalgebras Kac proved that all inequivalent finite-dimensional complex irreducible representations may be labelled by means of Kac–Dynkin labels that serve to specify the highest weight Λ, of the corresponding irreducible module, $V(\Lambda)$. In the case of sl(m/n), choosing what Kac called the distinguished Borel subalgebra, these labels

$$[a_1,a_2,\ldots,a_{m-1};a_m;a_{m+1},\ldots,a_{m+n-1}]$$

are such that $a_i \in \mathbb{N} = \{0,1,2,\ldots\}$ for $i \neq m$ and $a_m \in \mathbb{C}$.

For any complex reductive Lie algebra, $G_{\bar{0}}$, each finite-dimensional irreducible module over $\mathbb{C}$ has a weight structure that is completely determined by the Weyl character formula[13]

$$\text{ch } V^0(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) e^{w(\Lambda+\rho_0)}, \tag{1.1}$$

where $V^0(\Lambda)$ is the irreducible module of highest weight Λ, W is the Weyl group of $G_{\bar{0}}$, ρ_0 is half the sum of the positive roots Δ_0^+ of $G_{\bar{0}}$, and

$$L_0 = \prod_{\alpha \in \Delta_0^+} (e^{\alpha/2} - e^{-\alpha/2}). \tag{1.2}$$

Kac[12,14] showed that the irreducible finite-dimensional modules of a basic classical Lie superalgebra $G = G_{\bar{0}} \oplus G_{\bar{1}}$ fall into two classes, referred to as typical and atypical. For $G = \text{sl}(m/n)$, the irreducible module $V(\Lambda)$ of highest weight Λ is said to be typical if and only if

$$\langle \Lambda + \rho | \beta \rangle \neq 0, \quad \text{for all } \beta \in \Delta_1^+, \tag{1.3}$$

where $\langle . | . \rangle$ is a nondegenerate bilinear form, and $\rho = \rho_0 - \rho_1$, with ρ_0 and ρ_1 half the sums of the even and odd positive roots Δ_0^+ and Δ_1^+, respectively, of sl(m/n). Conversely, if

$$\langle \Lambda + \rho | \beta \rangle = 0, \quad \text{for any } \beta \in \Delta_1^+, \tag{1.4}$$

the module is said to be atypical.

The weight structure of a typical module $V(\Lambda)$ is completely determined by the Kac character formula:[12,14]

$$\text{ch } V(\Lambda) = \frac{L_1}{L_0} \sum_{w \in W} \epsilon(w) e^{w(\Lambda+\rho)}, \tag{1.5}$$

[a)] Aangesteld Navorser N.F.W.O. (National Fund for Scientific Research of Belgium).

 0022-2488/90/092278-27$03.00

where L_0 and W are defined as above for $G_{\bar{0}} = \mathrm{sl}(m) \oplus \mathbb{C} \oplus \mathrm{sl}(n)$, and

$$L_1 = \prod_{\beta \in \Delta_1^+} (e^{\beta/2} + e^{-\beta/2}). \tag{1.6}$$

For atypical modules, for which necessarily $a_m \in \mathbb{Z}$, the situation is far from clear and no such classification of the set of weights in terms of character formulas has been given so far, although many partial results have been obtained. In the case of $\mathrm{sl}(m/n)$ the application of formula (1.5) to the case of an atypical module with highest weight Λ does not give the character of that irreducible module, but instead gives the character of a reducible indecomposable module, $\overline{V}(\Lambda)$, with highest weight Λ, which we shall refer to as the Kac module. This Kac module possesses a unique maximal proper submodule M such that the irreducible module $V(\Lambda)$ is the quotient $\overline{V}(\Lambda)/M$. In this paper we are not concerned with the explicit construction of the modules $V(\Lambda)$ but only with their character formulas that determine and are determined by their weight structure. Nonetheless we exploit certain general properties of the underlying modules that have been established through a variety of approaches.

Progress has been made in understanding the properties of atypical modules of $\mathrm{sl}(m/n)$ through the use of a rich variety of methods: the decomposition of tensor powers using a graded version of the symmetric group action;[15-17] the use of power sum supersymmetric functions;[18] the exploitation of Young diagrams and supertableaux;[19-25] the explicit determination of the action of the superalgebra on weight vectors;[26-29] the consideration of induced modules and the identification of submodules by various means.[30-38] To date, however, these methods have not provided character formulas for all atypical modules of $\mathrm{sl}(m/n)$, except in such special cases as $\mathrm{sl}(2/1)$,[26] $\mathrm{sl}(3/1)$,[28] $\mathrm{sl}(3/2)$,[23] or $\mathrm{sl}(m/1)$.[21,29-31,37]

It is possible to express the characters of $\mathrm{sl}(m/n)$ modules in terms of Schur functions, also known in the literature as S functions.[39,40] This approach was followed by Berele and Regev[16] and Serge'ev[17] who derived a character formula, first given by Dondi and Jarvis,[15] which is appropriate to all irreducible covariant tensor representations of $\mathrm{sl}(m/n)$.

The extension of this approach to the case of irreducible mixed tensor representations was thwarted by the fact that mixed tensor products give rise to reducible, but indecomposable, representations.[15,20-22] Interesting results were certainly obtained[15,19-25] but the separation of the resulting characters into their irreducible constituents was not accomplished in general.

A number of quite different investigations[30,33,41,42] have led independently to the consideration of character formulas of the Kac–Weyl type, by which we mean any formula of the type:

$$\mathrm{ch}\, V(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) e^{w(\Lambda + \rho_0)} \prod_{\beta \in \Delta(\Lambda)} (1 + e^{-w\beta}), \tag{1.7}$$

where $\Delta(\Lambda)$ is some subset of Δ_1^+. This is a generalization of the formula (1.5) due to Kac in the sense that (1.5) may be recovered from (1.7) merely by setting $\Delta(\Lambda) = \Delta_1^+$ and exploiting the Weyl invariance of Δ_1^+. This covers the case of all typical modules.

In trying to accommodate atypical modules an important formula of the Kac–Weyl type (1.7) was reported by Bernstein and Leites.[30] This formula, which we refer to as the Leites formula, is given by (1.7) with

$$\Delta(\Lambda) = \{\beta \in \Delta_1^+ \,|\, \langle \Lambda + \rho | \beta \rangle \neq 0\}. \tag{1.8}$$

Whereas this formula is certainly correct for all typical modules and a large number of atypical modules, it is not correct[35,41-43] for some "low-lying" modules, i.e., modules for which Λ is close to the Weyl reflection planes. In particular, if $n, m \geqslant 2$, it does not give correctly the trivial character of the identity module of $\mathrm{sl}(m/n)$ with highest weight $\Lambda = 0$.

On the other hand, if $\#\{\beta \in \Delta_1^+ \,|\, \langle \Lambda + \rho | \beta \rangle = 0\} = 1$, in which case we say that Λ is singly atypical, the Leites formula, (1.7) with (1.8), does give the correct character for $V(\Lambda)$, as we have recently shown elsewhere[44] by exploiting a technique outside the domain of the present article.

Realizing the failure of the Leites formula in certain multiply atypical cases, for which $\#\{\beta \in \Delta_1^+ \,|\, \langle \Lambda + \rho | \beta \rangle = 0\} > 1$, new attempts were made[35,41,42] at obtaining a character formula with a wider range of validity. In particular, both Hughes and King[41] and Serganova and Serge'ev[42] found formulas of the Kac–Weyl type that yield the correct character in many cases where the Leites formula failed, including the identity module. One great merit of the Serganova–Serge'ev formula is that Pragacz[45] has recently shown, by exploiting a characterization of supersymmetric polynomials,[46] that this formula is equivalent to the S function formula of Berele and Regev[16] in the case of all covariant tensor representations. It is therefore valid in all these cases, whether or not they are typical, singly atypical, or multiply atypical. Unfortunately, for the simplest of all mixed tensor representations, namely the adjoint representation, the Serganova–Serge'ev formula[42] fails if $m,n \geqslant 2$. The Hughes–King formula[41] has a similar defect for other cases.

Bearing these successes and failures in mind we have explored more systematically all formulas of the Kac–Weyl type, arriving at yet another Kac–Weyl type formula that is identical to the Serganova–Serge'ev formula for these covariant tensor modules, but which differs from it for certain multiply atypical modules. It yields the correct character formula in very many cases, including the adjoint modules for which the Serganova–Serge'ev formula fails.

Despite our new formula surviving many checks in cases where the characters are unambiguously known and a variety of self-consistency checks in others, the computer programs developed to carry out these checks eventually revealed an example of an irreducible module of $\mathrm{sl}(3/4)$ for which closer scrutiny seemed to be required. By means of further tests described here we were finally able to show that for this module no Kac–Weyl type formula, including our own generalization of the Serganova–Serge'ev formula, can yield the correct character.

Faced with the failure of Kac–Weyl type formulas (1.7) to give the correct weight structure of all $\mathrm{sl}(m/n)$ modules,

but having acquired a great deal of information concerning the cases for which breakdowns occur, we are forced to consider other types of formula.

It is not difficult to see that the characters of all irreducible modules may be expressed in terms of Kac characters, (1.5), even though the Kac modules are themselves reducible in general. Moreover the Leites character formula, (1.7) with (1.8), can be re-expressed as an infinite alternating sum of Kac characters. This has led us to conjecture the validity of what we call an extended Kac–Weyl character formula. This also involves an infinite sum of Kac characters, but in certain critical cases all terms corresponding to weights beyond certain truncation planes in the weight space, and which are included in the Leites formula, are excluded. These truncation planes are uniquely determined, for each highest weight module $V(\Lambda)$, as symmetry planes under the so-called "dot" action of particular elements of the Weyl group that connect the various atypical roots. For singly atypical modules, no truncation planes exist and so the extended formula is identical with the Leites formula, which as we have already mentioned, is correct for all these cases.

In the case of irreducible covariant tensor modules, whether they be typical, singly or multiply atypical, we prove in this paper that the extended formula gives the same character as the Berele–Regev formula, and so it is correct for these modules. For the other multiply atypical modules, we have no proof that the extended formula is correct. However, using Pascal programs, we have tested it out on a large number of modules and have found no counterexamples, even amongst those modules for which all the Kac–Weyl type formulas discussed above fail. We therefore conjecture that the extended Kac–Weyl formula is indeed the correct character formula for all irreducible $\mathrm{sl}(m/n)$ modules.

The structure of the paper is as follows: In Sec. II the Lie superalgebra $\mathrm{sl}(m/n)$ is introduced, and in Sec. III we give some general definitions concerning modules and characters. In Sec. IV the important construct known as the atypicality matrix is defined for $\mathrm{sl}(m/n)$ and used to determine two generating matrices whose specification allow us to write down the two Kac–Weyl type character formulas: χ_S, due to Serganova and Serge'ev, and χ_J, our own. Their relationships to other formulas are discussed and their ranges of validity are described, and we discuss in detail our counterexample to all possible formulas of the Kac–Weyl type. In Sec. V we restrict attention to covariant tensor irreducible representations and discuss the equivalence of the χ_S formula and the Berele–Regev formula.

Section VI is concerned with singly atypical modules: In particular we show that for such modules the Leites formula follows from a simple proposition, Proposition 6.8, concerning primitive weight vectors of the Kac module, $\overline{V}(\Lambda)$, whose validity is proved elsewhere.[44] The connection is made with one particular supertableaux method[34] and it is shown that each character defined by the Leites formula can be written as a sum of Weyl conjugates of the character of an induced module.

The above-mentioned extended Kac–Weyl formula is then introduced for doubly atypical representations in Sec. VII, Definition 7.19, and for multiply atypical representations in Sec. VIII, Definition 8.11. We show how the application of this extended formula to the identity module gives rise to Cauchy's identity. We also prove the validity of the extended formula, χ_T, in the case of all covariant tensor representations of $\mathrm{sl}(m/n)$, and in Sec. IX summarize the evidence in favor of the validity of our Conjecture 8.14 for the characters of all irreducible modules of $\mathrm{sl}(m/n)$.

II. NOTATION AND CONVENTIONS

A complex Lie superalgebra G is a $\mathbb{Z}_2$-graded linear vector space, $G = G_{\bar{0}} \oplus G_{\bar{1}}$ over $\mathbb{C}$ with a bracket $[\ ,\ ]$ such that $\forall a \in G_\alpha$, $\forall b \in G_\beta$, $\forall c \in G$, and $\forall \alpha, \beta \in \mathbb{Z}_2$:

$$[a,b] \in G_{\alpha+\beta},$$
$$[a,b] = -(-1)^{\alpha\beta}[b,a],$$
$$[a,[b,c]] = [[a,b],c] + (-1)^{\alpha\beta}[b,[a,c]]. \tag{2.1}$$

The simplest example is given by $\mathrm{gl}(m/n)$[7,10] with $m,n \in \mathbb{N}$. Its natural matrix realization takes the form:

$$\mathrm{gl}(m/n) = \left\{ x = \begin{pmatrix} A & B \\ C & D \end{pmatrix} : A \in M_{m\times m}, B \in M_{m\times n}, \right.$$
$$\left. C \in M_{n\times m}, D \in M_{n\times n} \right\}, \tag{2.2}$$

where $M_{p\times q}$ is the set of all $p\times q$ complex matrices. The even subspace $\mathrm{gl}(m/n)_{\bar{0}}$ has $B=0$ and $C=0$; the odd subspace $\mathrm{gl}(m/n)_{\bar{1}}$ has $A=0$ and $D=0$. Note that $\mathrm{gl}(m/n)_{\bar{0}} = \mathrm{gl}(m) \oplus \mathrm{gl}(n)$. In the case of $G = \mathrm{gl}(m/n)$, the bracket is determined in the natural matrix representation by

$$[a,b] = ab - (-1)^{\alpha\beta}ba, \quad \forall a \in G_\alpha \text{ and } \forall b \in G_\beta, \tag{2.3}$$

where on the right-hand side, juxtaposition denotes matrix multiplication.

If we denote by $\mathrm{gl}(m/n)_{+1}$ the space spanned by matrices $\left(\begin{smallmatrix} 0 & B \\ 0 & 0 \end{smallmatrix}\right)$ and by $\mathrm{gl}(m/n)_{-1}$ the space of matrices $\left(\begin{smallmatrix} 0 & 0 \\ C & 0 \end{smallmatrix}\right)$, then $G = \mathrm{gl}(m/n)$ has a $\mathbb{Z}$ grading that is consistent with the $\mathbb{Z}_2$ grading:

$$G = G_{-1} \oplus G_0 \oplus G_{+1}, \quad G_{\bar{0}} = G_0,$$

and

$$G_{\bar{1}} = G_{-1} \oplus G_{+1}. \tag{2.4}$$

With the definition of *supertrace*[1,7,10] as $\mathrm{str}(x) = \mathrm{tr}(A) - \mathrm{tr}(D)$ one can define the subalgebra $\mathrm{sl}(m/n)$:

$$\mathrm{sl}(m/n) = \{x \in \mathrm{gl}(m/n) : \mathrm{str}(x) = 0\}. \tag{2.5}$$

If $m \neq n$ then $\mathrm{sl}(m/n)$ is the *simple* Lie superalgebra $A(m-1,n-1)$, otherwise it contains a one-dimensional ideal $\mathbb{C}I_{2m}$ and then $A(m-1,m-1) = \mathrm{sl}(m/m)/\mathbb{C}$.

A Cartan subalgebra of $\mathrm{gl}(m/n)$ is given by the vector space $\mathfrak{d}$ of diagonal matrices and has dimension $m+n$. The restriction to $\mathrm{sl}(m/n)$ requires the supertrace condition to be satisfied. Hence the Cartan subalgebra $\mathfrak{h}$ of $\mathrm{sl}(m/n)$ has dimension $m+n-1$ and is spanned by

$$H_i = E_{ii} - E_{i+1,i+1}, \quad 1 \leqslant i < m,$$
$$H_m = E_{mm} + E_{m+1,m+1}, \tag{2.6}$$
$$H_{m+a} = E_{m+a,m+a} - E_{m+a+1,m+a+1}, \quad 1 \leqslant a < n,$$

where E_{ij} is the matrix with entry 1 at position (i,j) and 0 elsewhere. For $m=n$ the H_i in $A(m-1,m-1)$ satisfy the extra relation

$$\sum_{i=1}^{m} iH_i - \sum_{a=1}^{m} (m-a)H_{m+a} = 0. \tag{2.7}$$

The Lie superalgebra $\mathrm{sl}(m/n)$ $(m\neq n)$ contains the special element

$$c = \frac{n}{n-m}\sum_{i=1}^{m} iH_i - \frac{m}{n-m}\sum_{a=1}^{n} (n-a)H_{m+a}, \tag{2.8}$$

which is a central element in $\mathrm{sl}(m/n)_{\bar{0}} = \mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$. In particular, with the $\mathbb{Z}$ grading (2.4) we have the property[10]

$$[c,x] = kx, \quad \text{for } x\in G_k \quad (k\in\mathbb{Z}). \tag{2.9}$$

The dual space $\mathfrak{h}^*$ is described in the basis of forms ϵ_i $(i=1,\ldots,m)$ and δ_a $(a=1,\ldots,n)$, where $\epsilon_i\colon x\to a_{ii}$ and $\delta_a\colon x\to d_{aa}$. Then the roots and corresponding root vectors of $\mathrm{sl}(m/n)$ are given by[12]

$$\begin{aligned}
&\epsilon_i-\epsilon_j \leftrightarrow E_{ij} \quad (1\leqslant i,j\leqslant m) \quad \text{(even)},\\
&\delta_a-\delta_b \leftrightarrow E_{m+a,m+b} \quad (1\leqslant a,b\leqslant n) \quad \text{(even)},\\
&\epsilon_i-\delta_a \leftrightarrow E_{i,m+a} \quad (1\leqslant i\leqslant m, 1\leqslant a\leqslant n) \quad \text{(odd)},\\
&-\epsilon_i+\delta_a \leftrightarrow E_{m+a,i} \quad (1\leqslant a\leqslant n, 1\leqslant i\leqslant m) \quad \text{(odd)}.
\end{aligned} \tag{2.10}$$

We denote by Δ the set of all roots, by Δ_0 the set of even roots, by Δ_1 the set of odd roots, and by $e(\alpha)$ the root vector corresponding to the root $\alpha\in\Delta$. Then G has the root space decomposition

$$G = \mathfrak{h}\oplus\Big(\bigoplus_{\alpha\in\Delta}\mathbb{C}e(\alpha)\Big). \tag{2.11}$$

There are several choices for a basis of simple roots for Δ. Here we shall take the following distinguished basis[12] as the simple root system $\Pi=\{\alpha_1,\alpha_2,\ldots,\alpha_{m+n-1}\}$:

$$\begin{array}{ccccccccccc}
\epsilon_1-\epsilon_2 & \epsilon_2-\epsilon_3 & \cdots & \epsilon_m-\delta_1 & \delta_1-\delta_2 & \cdots & \delta_{n-2}-\delta_{n-1} & \delta_{n-1}-\delta_n\\
\mathrm{O} & \mathrm{O} & \cdots & \otimes & \mathrm{O} & \cdots & \mathrm{O} & \mathrm{O};
\end{array} \tag{2.12}$$

then the even and odd positive roots are given by

$$\Delta_0^+ = \{\epsilon_i-\epsilon_j : 1\leqslant i<j\leqslant m, \delta_a-\delta_b : 1\leqslant a<b\leqslant n\},$$

and

$$\Delta_1^+ = \{\epsilon_i-\delta_a : 1\leqslant i\leqslant m, 1\leqslant a\leqslant n\}, \tag{2.13}$$

respectively. It will be convenient to denote the odd positive roots by

$$\beta_{ia} = \epsilon_i-\delta_a \quad (1\leqslant i\leqslant m; 1\leqslant a\leqslant n). \tag{2.14}$$

If we put

$$\mathfrak{n}^{\pm} = \mathrm{span}\{e(\alpha)\,|\,\alpha\in\Delta^{\pm}\}, \tag{2.15}$$

then the root space decomposition (2.11) implies that

$$G = \mathfrak{n}^-\oplus\mathfrak{h}\oplus\mathfrak{n}^+, \tag{2.16}$$

where $\mathfrak{n}^-$ and $\mathfrak{n}^+$ are nilpotent subalgebras. As usual,

$$B = \mathfrak{h}\oplus\mathfrak{n}^+ \tag{2.17}$$

is called a Borel subalgebra.

The invariant nondegenerate inner product on G is given by $\langle x|y\rangle = \mathrm{str}(xy)$. The restriction of this to $\mathfrak{h}$ is also nondegenerate and therefore for $\alpha\in\mathfrak{h}^*$ there is a corresponding element $h_\alpha\in\mathfrak{h}$ defined by

$$\langle h\,|h_\alpha\rangle = \alpha(h)h_\alpha, \quad \forall h\in\mathfrak{h}. \tag{2.18}$$

This induces a nondegenerate inner product on $\mathfrak{h}^*$:

$$\langle\alpha|\beta\rangle = \langle h_\alpha|h_\beta\rangle. \tag{2.19}$$

The inner product for the weight space of $\mathrm{gl}(m/n)$ is explicitly determined by

$$\langle\epsilon_i|\epsilon_j\rangle = \delta_{ij}, \quad \langle\epsilon_i|\delta_a\rangle = 0, \quad \langle\delta_a|\delta_b\rangle = -\delta_{ab}. \tag{2.20}$$

An element $\Lambda\in\mathfrak{h}^*$ with $\Lambda=\Sigma_i\mu_i\epsilon_i+\Sigma_a\nu_a\delta_a$ can be written in terms of its components in this $\epsilon\delta$ basis as $\Lambda=(\mu_1\mu_2\cdots\mu_m|\nu_1\nu_2\cdots\nu_n)$. If $G=\mathrm{sl}(m/n)$ with $m\neq n$ then these components satisfy the extra condition $\Sigma_i\mu_i+\Sigma_a\nu_a=0$, while if $G=A(m-1,m-1)$ then both $\Sigma_i\mu_i=0$ and $\Sigma_a\nu_a=0$.

With the choice of the distinguished Borel subalgebra of $\mathrm{sl}(m/n)$ implicit in (2.12), the element $\Lambda\in\mathfrak{h}^*$ may be written in terms of Kac–Dynkin labels as $\Lambda=[a_1a_2\cdots a_{m-1};a_m;a_{m+1}\cdots a_{m+n-1}]$, where $a_i=\Lambda(H_i)$ for $1\leqslant i\leqslant m+n-1$.[7,12] To be explicit:

$$\begin{aligned}
&a_i = \mu_i-\mu_{i+1} \quad (1\leqslant i<m),\\
&a_m = \mu_m+\nu_1,\\
&a_{m+b} = \nu_b-\nu_{b+1} \quad (1\leqslant b<n).
\end{aligned} \tag{2.21}$$

If $m=n$ and $G=A(m-1,m-1)$, it follows from (2.7) that

$$a_1+2a_2+3a_3+\cdots+ma_m-(m-1)a_{m+1}-(m-2)a_{m+2}-\cdots-a_{2m-1}=0. \tag{2.22}$$

The *level* of an element $\Lambda\in\mathfrak{h}^*$ is defined for $\mathrm{sl}(m/n)$ $(m\neq n)$ by

$$\mathrm{level}(\Lambda) = \Lambda(c), \tag{2.23}$$

where c is the element (2.8). Hence

$$\mathrm{level}(\Lambda) = \frac{n}{n-m}\sum_{i=1}^{m}\mu_i + \frac{m}{n-m}\sum_{b=1}^{n}\nu_b. \tag{2.24}$$

For $\mathrm{sl}(m/m)$ (2.24) is not appropriate as it stands, but since $\Sigma_i\mu_i+\Sigma_b\nu_b=0$ we can define the level by

$$\mathrm{level}(\Lambda) = \sum_{i=1}^{m}\mu_i. \tag{2.25}$$

There is a partial ordering for the elements of $\mathfrak{h}^*$: for $\Lambda,\Lambda'\in\mathfrak{h}^*$ we say

$$\Lambda\geqslant\Lambda' \Leftrightarrow \Lambda-\Lambda' = \sum_{\alpha\in\Delta^+} t_\alpha\alpha, \quad \text{with every } t_\alpha\geqslant 0. \tag{2.26}$$

The *Weyl group* of G ($=\mathrm{gl}(m/n)$ or $\mathrm{sl}(m/n)$ or $A(m-1,m-1)$) is by definition the Weyl group W of $G_{\bar 0}$. Hence W is the direct product of the Weyl group of $\mathrm{sl}(m)$ with the Weyl group of $\mathrm{sl}(n)$, or to be explicit, $W = S_m \times S_n$, where S_p denotes the symmetric group of order $p!$. There are two distinctly defined actions of W on $\Lambda \in \mathfrak{h}^*$. First, the direct action indicated by juxtaposition. For $\Lambda \in \mathfrak{h}^*$ with components $(\mu_1,\mu_2,\ldots,\mu_m|\nu_1,\nu_2,\ldots,\nu_n)$, an element $w = \sigma\times\tau \in W = S_m \times S_n$ acts on Λ to give $w\Lambda$ by permuting the components of Λ so that $\mu_i \to \mu_{\sigma(i)}$ and $\nu_b \to \nu_{\tau(b)}$. The signature $\epsilon(w)$ of w is the product of the signatures of σ and τ, and takes the values ± 1. Second, the "dot" action whereby

$$w\cdot\Lambda = w(\Lambda+\rho) - \rho, \quad \text{where } \rho = \rho_0 - \rho_1, \tag{2.27}$$

with

$$\rho_0 = \frac{1}{2}\sum_{\alpha\in\Delta_0^+}\alpha, \quad \text{and} \quad \rho_1 = \frac{1}{2}\sum_{\beta\in\Delta_1^+}\beta. \tag{2.28}$$

For simple Lie algebras, all simple root systems are W equivalent [i.e., if Π_1 and Π_2 are two simple root systems, then there exists a $w \in W$ such that $\Pi_2 = w(\Pi_1)$]. Note that for Lie superalgebras in general, and for $\mathrm{sl}(m/n)$ in particular, this is not the case. A different choice of simple roots is in general not W equivalent to the distinguished choice made here. As a consequence, another Borel subalgebra is in general not W equivalent to (2.17). For the distinguished choice[12] corresponding to (2.12) and (2.13) we have

$$\rho_0 = \frac{1}{2}\sum_{i=1}^{m}(m-2i+1)\epsilon_i + \frac{1}{2}\sum_{b=1}^{n}(n-2a+1)\delta_b$$

and

$$\rho_1 = \frac{n}{2}\sum_{i=1}^{m}\epsilon_i - \frac{m}{2}\sum_{b=1}^{n}\delta_b. \tag{2.29}$$

It is to be noted that Δ_1^+ is invariant under the action of W for the distinguished choice of Borel subalgebra, so that for all $w \in W$

$$w\rho_1 = \rho_1, \tag{2.30}$$

and consequently

$$w\cdot\Lambda = w(\Lambda+\rho) - \rho = w(\Lambda+\rho_0) - \rho_0. \tag{2.31}$$

III. REPRESENTATIONS AND CHARACTERS

Let $V = V_{\bar 0} \oplus V_{\bar 1}$ be a $\mathbb{Z}_2$-graded linear vector space over $\mathbb{C}$, and denote by $\mathrm{gl}(V)$ the space of endomorphisms of V. Then $\mathrm{gl}(V)$ is naturally graded: $\mathrm{gl}(V) = \mathrm{gl}(V)_{\bar 0} \oplus \mathrm{gl}(V)_{\bar 1}$. A representation ϕ is a linear mapping from G to $\mathrm{gl}(V)$ such that:

$$\phi:\ x \to \phi(x), \quad \text{with } \phi(x) \in \mathrm{gl}(V)_\alpha \text{ for } x \in G_\alpha,$$
$$\phi([x,y]) = \phi(x)\phi(y) - (-1)^{\alpha\beta}\phi(y)\phi(x), \tag{3.1}$$
$$\forall x \in G_\alpha \text{ and } \forall y \in G_\beta.$$

In this case V is a G module with the action of G defined by $gv = \phi(g)v$ for $g \in G$ and $v \in V$. If V is a G module, then V is naturally a $\mathrm{U}(G)$ module, where $\mathrm{U}(G)$ denotes the universal enveloping algebra of G.

A module V is called a *highest-weight module* if it contains a vector v_Λ such that

$$\mathfrak{n}^+ v_\Lambda = 0, \quad hv_\Lambda = \Lambda(h)v_\Lambda \quad \forall h \in \mathfrak{h}, \quad V = \mathrm{U}(G)v_\Lambda. \tag{3.2}$$

Then v_Λ is called a highest weight vector and Λ is the highest weight. Such modules have a weight space decomposition:

$$V = \bigoplus_{\mu\le\Lambda} V_\mu, \tag{3.3}$$

where

$$V_\mu = \{v \in V: hv = \mu(h)v, \quad \forall h \in \mathfrak{h}\} \tag{3.4}$$

is called the weight space of weight μ.

Restricting ourselves to $G = \mathrm{sl}(m/n)$ with $G_{\bar 0} = \mathrm{sl}(m) \oplus \mathbb{C} \oplus \mathrm{sl}(n)$ it is convenient to distinguish between various types of weight $\Lambda \in \mathfrak{h}^*$.

Definition 3.5: The weight $\Lambda = [a_1 a_2 \cdots a_{m-1}; a_m; a_{m+1}\cdots a_{m+n-1}]$ with $a_i = \Lambda(H_i) \in \mathbb{C}$ is said to be *admissible* if $a_i \in \mathbb{Z}$ for $i \neq m$ and $a_m \in \mathbb{C}$; *dominant* if $a_i \in \mathbb{N}$ for $i \neq m$ and $a_m \in \mathbb{C}$; *integral* if $a_i \in \mathbb{Z}$ for all i; and *integral dominant* if $a_i \in \mathbb{N}$ for $i \neq m$ and $a_m \in \mathbb{Z}$.

Let Λ be dominant. From the theory of reductive Lie algebras it follows that there exists a unique finite-dimensional irreducible $G_{\bar 0}$ module $V^0(\Lambda)$ with highest weight Λ. Putting $G_{+1}V^0(\Lambda) = 0$, this becomes a $G_0 \oplus G_{+1}$ module. The *Kac module* $\overline{V}(\Lambda)$ is defined as the induced module:[12]

$$\overline{V}(\Lambda) = \mathrm{Ind}_{G_0\oplus G_{+1}}^{G} V^0(\Lambda) \cong \mathrm{U}(G_{-1}) \otimes V^0(\Lambda). \tag{3.6}$$

But $[x,y] = 0$ for $x,y \in G_{-1}$, so $\mathrm{U}(G_{-1}) \cong \Lambda(G_{-1})$, the exterior algebra over G_{-1}. Since $\dim(G_{-1}) = mn$, the dimension of $\mathrm{U}(G_{-1})$ is 2^{mn}, and thus $\overline{V}(\Lambda)$ is a finite-dimensional G module of dimension $2^{mn}\times\dim(V^0(\Lambda))$. Unfortunately, $\overline{V}(\Lambda)$ is not always an irreducible module. In general, $\overline{V}(\Lambda)$ contains proper submodules. If M is the unique maximal submodule of $\overline{V}(\Lambda)$ such that $M \neq \overline{V}(\Lambda)$, then the quotient module

$$V(\Lambda) = \overline{V}(\Lambda)/M \tag{3.7}$$

is a finite-dimensional irreducible G module. Kac proved further the following result:[12]

Theorem 3.8: Every finite-dimensional irreducible G module is isomorphic to a module of the type $V(\Lambda) = \overline{V}(\Lambda)/M$, where Λ is dominant, $\overline{V}(\Lambda)$ is the corresponding Kac module and $M \neq \overline{V}(\Lambda)$ is the maximal submodule of $\overline{V}(\Lambda)$. Moreover, every finite-dimensional irreducible G module is uniquely characterized by its dominant highest weight Λ.

In what follows it is crucial to subdivide the class of dominant weights by means of

Definition 3.9: If Λ is a dominant weight of G then Λ is said to be *typical* if $\langle\Lambda+\rho|\beta\rangle \neq 0$ for all $\beta \in \Delta_1^+$; *atypical* if there exists $\beta \in \Delta_1^+$ such that $\langle\Lambda+\rho|\beta\rangle = 0$; *atypical of type* β if $\langle\Lambda+\rho|\beta\rangle = 0$ with $\beta \in \Delta_1^+$; and *atypical of degree* d if $\#\{\beta \in \Delta_1^+ : \langle\Lambda+\rho|\beta\rangle = 0\} = d$ with $d > 0$.

This definition is involved in Kac's theorem[12] regarding the conditions for $\overline{V}(\Lambda)$ to be irreducible.

Theorem 3.10: Let Λ be a dominant weight of G. The Kac module $\overline{V}(\Lambda)$ is an irreducible G module if and only if its highest weight Λ is typical, that is,

$$\langle\Lambda+\rho|\beta\rangle \neq 0, \quad \forall\beta \in \Delta_1^+. \tag{3.10}$$

In this case, we call $V(\Lambda) = \overline{V}(\Lambda)$ a *typical* module, otherwise $V(\Lambda) \neq \overline{V}(\Lambda)$ is called an *atypical* module.

The *character* ch V of a G module V with weight space decomposition (3.3) is defined as

$$\text{ch } V = \sum_{\mu} \dim(V_\mu) e^{\mu}, \tag{3.11}$$

where the summation is over all $\mu \in \mathfrak{h}^*$ for which $V_\mu \neq 0$ and e^{μ} is the formal exponential.

The action of the Weyl group on such formal exponentials is defined by $w(e^{\mu}) = e^{w\mu}$. Let

$$L_0 = \prod_{\alpha \in \Delta_0^+} (e^{\alpha/2} - e^{-\alpha/2}),$$

and

$$L_1 = \prod_{\beta \in \Delta_1^+} (e^{\beta/2} + e^{-\beta/2}). \tag{3.12}$$

From (3.5) it follows that the Kac module has character

$$\text{ch } \overline{V}(\Lambda) = \prod_{\beta \in \Delta_1^+} (1 + e^{-\beta}) \text{ch } V^0(\Lambda), \tag{3.13}$$

where ch $V^0(\Lambda)$ is given by Weyl's character formula (1.1):

$$\text{ch } V^0(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) e^{w(\Lambda + \rho_0)}. \tag{3.14}$$

From the definitions (2.28) and (3.12) and the Weyl invariance of ρ_1 in our distinguished basis, we have

$$\prod_{\beta \in \Delta_1^+} (1 + e^{-\beta}) = L_1 e^{-\rho_1} = L_1 e^{-w\rho_1}, \quad \forall w \in W, \tag{3.15}$$

from which we obtain Kac's character formula:

$$\text{ch } \overline{V}(\Lambda) = \frac{L_1}{L_0} \sum_{w \in W} \epsilon(w) e^{w(\Lambda + \rho)}. \tag{3.16}$$

Thanks to Kac's Theorem 3.10 this formula gives the character of each typical irreducible module $V(\Lambda) = \overline{V}(\Lambda)$.

Again, in our distinguished basis, $w(\Delta_1^+) = \Delta_1^+$ for all $w \in W$, so that $w(L_1) = L_1$, and hence (3.16) can be rewritten as[41]

$$\text{ch } \overline{V}(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\Lambda + \rho_0} \prod_{\beta \in \Delta_1^+} (1 + e^{-\beta}) \right\}. \tag{3.17}$$

This formula indicates that since Λ is dominant, all weights of $\overline{V}(\Lambda)$ are admissible in the sense of Definition 3.5. By virtue of Kac's Theorem 3.8 the same is true of all weights of any finite-dimensional irreducible module. For any admissible $\Lambda \in \mathfrak{h}^*$ we define the formal expression

$$\chi_K(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\Lambda + \rho_0} \prod_{\beta \in \Delta_1^+} (1 + e^{-\beta}) \right\}, \tag{3.18}$$

which we refer to as the Kac-character formula.

Kac having evaluated the characters of typical modules,[12,14] one of the most interesting problems remaining in Lie superalgebra theory is the determination of characters of atypical modules. Although the characters of atypical modules of sl(m/n) have been the subject of several studies, the problem of determining such characters has not yet been completely solved. The general *ansatz* used in various attempts to do so has been a formula of the Kac–Weyl type

$$\chi_X(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\Lambda + \rho_0} \prod_{\beta \in \Delta_X(\Lambda)} (1 + e^{-\beta}) \right\}, \tag{3.19}$$

where $\Delta_X(\Lambda) \subseteq \Delta_1^+$ and the subscript X is used to distinguish between a number of character formulas of this type. They are distinct by virtue of the fact that the prescription necessary to determine $\Delta_X(\Lambda)$ for a given weight Λ varies from one formula to another. Clearly every such prescription should yield $\Delta_X(\Lambda) = \Delta_1^+$ when applied to a typical highest weight Λ. We would justify calling (3.19) a formula of the Kac–Weyl type by noting that if we set $\Delta_K(\Lambda) = \Delta_1^+$ then $\chi_K(\Lambda)$ is the Kac character (3.18), whilst if we set $\Delta_W(\Lambda) = \varnothing$ then

$$\chi_W(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) e^{w(\Lambda + \rho_0)} \tag{3.20}$$

is the formal expression, for any admissible $\Lambda \in \mathfrak{h}^*$, which coincides with the Weyl character (3.14) if Λ is dominant.

Bernstein and Leites published[30] another formula, $\chi_L(\Lambda)$, of the Kac–Weyl type (3.19), with

$$\Delta_L(\Lambda) = \{\beta \in \Delta_1^+ : \langle \Lambda + \rho | \beta \rangle \neq 0\}, \tag{3.21}$$

so that

$$\chi_L(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) \times w \left\{ e^{\Lambda + \rho_0} \prod_{\beta \in \Delta_1^+, \langle \Lambda + \rho | \beta \rangle \neq 0} (1 + e^{-\beta}) \right\}. \tag{3.22}$$

This is one natural generalization of the character formula appropriate to typical modules. Unfortunately, it does not always yield the character of $V(\Lambda)$ when Λ is atypical. In the next section we shall discuss the range of validity of $\chi_L(\Lambda)$ and other character formulas of the Kac–Weyl type (3.19).

One great merit of (3.19) is that the expression of this character of sl(m/n) as a linear combination of Weyl characters (3.14) of irreducible $G_{\bar{0}}$ modules is a simple two-stage operation since the expansion of (3.19) immediately gives

$$\chi_X(\Lambda) = \sum_{\mu} p_X(\Lambda - \mu) \chi_W(\mu), \tag{3.23}$$

where p_X is the partition function defined in such a way that $p_X(\Lambda - \mu)$ is the number of ways of writing $\Lambda - \mu$ in the form

$$\Lambda - \mu = \sum_{\beta \in \Delta_X(\Lambda)} k_\beta \beta, \quad \text{with } k_\beta \in \{0,1\}, \tag{3.24}$$

and, by virtue of (2.31), the relationship between formal Weyl characters (3.20) and characters (3.14) of irreducible $G_{\bar{0}}$ modules is such that

$$\chi_W(\mu) = \begin{cases} \epsilon(w) \text{ch } V^0(w \cdot \mu), & \text{if } w \cdot \mu \text{ is dominant for some } w \in W, \\ 0, & \text{if } w \cdot \mu = \mu \text{ for some } w \in W \text{ with } \epsilon(w) = -1. \end{cases} \tag{3.25}$$

IV. CHARACTER FORMULAS OF THE KAC–WEYL TYPE

In this section we shall introduce a number of character formulas of the Kac–Weyl type (3.19) that each serve to give correctly the characters of certain finite-dimensional irreducible modules $V(\Lambda)$ of $\mathrm{sl}(m/n)$ for particular highest weights Λ. In doing so we shall give some indication of their range of validity. Unfortunately, we shall also show quite explicitly that no single formula of this type can correctly give the character of all finite-dimensional irreducible modules.

The key combinatorial tool in the study of character formulas of the Kac–Weyl type is the atypicality matrix:

Definition 4.1: Let Λ be an element of $\mathfrak{h}^* \subset \mathrm{sl}(m/n)$. Then the *atypicality matrix* A_Λ is the $m \times n$ matrix $A_\Lambda = (A(\Lambda)_{ij})$ with

$$A(\Lambda)_{ij} = \langle \Lambda + \rho | \beta_{ij} \rangle, \quad 1 \leqslant i \leqslant m \text{ and } 1 \leqslant j \leqslant n, \tag{4.1a}$$

so that in terms of the Kac–Dynkin labels:

$$A(\Lambda)_{ij} = \sum_{r=i}^{m-1} a_r + a_m - \sum_{s=1}^{j-1} a_{m+s} + m - i - j + 1, \tag{4.1b}$$

or equivalently in terms of the $\epsilon\delta$ components of Λ:

$$A(\Lambda)_{ij} = \mu_i + \nu_j + m - i - j + 1, \quad 1 \leqslant i \leqslant m \text{ and } 1 \leqslant j \leqslant n. \tag{4.1c}$$

The following properties are an immediate consequence of this definition: from (4.1a),

$$A(w \cdot \Lambda)_{ij} = A(\Lambda)_{\sigma(i)\tau(j)}, \tag{4.2a}$$

where $w \cdot \Lambda = w(\Lambda + \rho) - \rho$ and $w \in W = S_m \times S_n$ with $w^{-1} = \sigma \times \tau$; from (4.1b),

$$\begin{aligned} &A(\Lambda)_{ij} - A(\Lambda)_{i+1,j} = a_i + 1, \quad 1 \leqslant i < m \\ &A(\Lambda)_{m1} = a_m, \\ &A(\Lambda)_{ij} - A(\Lambda)_{ij+1} = a_{m+j} + 1, \quad 1 \leqslant j < n; \end{aligned} \tag{4.2b}$$

from (4.1c),

$$A(\Lambda)_{ij} + A(\Lambda)_{kl} = A(\Lambda)_{il} + A(\Lambda)_{jk}. \tag{4.2c}$$

It should be noted that (4.2b) implies $A_\Lambda = A_\Sigma$ if and only if $\Lambda = \Sigma$. Moreover, by virtue of our previous definitions (3.5) and (3.9):

$$\begin{aligned} \Lambda \text{ is dominant } &\Leftrightarrow A(\Lambda)_{ij} - A(\Lambda)_{i+1,j} - 1 \in \mathbb{N}, \\ &\quad 1 \leqslant i < m, \quad 1 \leqslant j \leqslant n, \\ &\quad A(\Lambda)_{m1} \in \mathbb{C}, \\ &\quad A(\Lambda)_{ij} - A(\Lambda)_{ij+1} - 1 \in \mathbb{N}, \\ &\quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j < n, \end{aligned} \tag{4.3b}$$

Λ is integral dominant $\Leftrightarrow \Lambda$ is integral and dominant, (4.3c)

Λ is typical $\Leftrightarrow \Lambda$ is dominant and $A(\Lambda)_{ij} \neq 0$ for

$$1 \leqslant i \leqslant m \text{ and } 1 \leqslant j \leqslant n, \tag{4.3d}$$

Λ is atypical of type $\beta = \beta_{ij} = \epsilon_i - \delta_j$

$$\Leftrightarrow \Lambda \text{ is dominant and } A(\Lambda)_{ij} = 0, \tag{4.3e}$$

Λ is atypical of degree $d \Leftrightarrow \Lambda$ is dominant and $d > 0$,

where $d = \#\{A(\Lambda)_{ij} : A(\Lambda)_{ij} = 0, 1 \leqslant i \leqslant m \text{ and } 1 \leqslant j \leqslant n\}$. (4.3f)

In what follows we are concerned with the study only of finite-dimensional irreducible modules for which the highest weight Λ is integral dominant, so that all atypicality matrices have integer elements. By virtue of the above conditions this includes all atypical modules. From the atypicality matrices we first construct certain sequences of elements.

Definition 4.4: If $A(\Lambda)_{ij} = 0$ for given integral dominant Λ, we define the *upper* and *lower sequences* through (i,j) as sequences, U_{ij} and L_{ij}, of matrix positions given schematically by

$$(i_0 j_0) \begin{matrix} \nearrow (i_1 j_1) \to (i_2 j_2) \to \cdots \to (i_t j_t) \to \cdots \to (i_p j_p), \\ \searrow (i_{-1} j_{-1}) \to (i_{-2} j_{-2}) \to \cdots \to (i_{-t} j_{-t}) \to \cdots \to (i_{-q} j_{-q}), \end{matrix} \tag{4.4a}$$

with $(i_0 j_0) = (i,j)$. The upper sequence proceeds upwards a row at a time with column jumps to the right determined by the Kac–Dynkin labels $a_{i-1}, a_{i-2}, \ldots, a_{i-p}$. To be explicit, it is defined by

$$\begin{aligned} U_{ij} = (i,j) &\to (i-1, j + a_{i-1}) \to (i-2, j + a_{i-1} + a_{i-2}) \to \\ &\cdots \to \left(i - t, j + \sum_{r=1}^{t} a_{i-r}\right) \to \\ &\cdots \to \left(i - p, j + \sum_{r=1}^{p} a_{i-r}\right). \end{aligned} \tag{4.4b}$$

Similarly, the lower sequence proceeds to the right a column at a time with upward row jumps determined by the Kac–Dynkin labels $a_{m+j}, a_{m+j+1}, \ldots, a_{m+j+q-1}$. It is defined by

$$\begin{aligned} L_{ij} = (i,j) &\to (i - a_{m+j}, j+1) \\ &\to (i - a_{m+j} - a_{m+j+1}, j+2) \\ &\to \cdots \to \left(i - \sum_{r=1}^{t} a_{m+j+r-1}, j+t\right) \to \\ &\cdots \to \left(i - \sum_{r=1}^{q} a_{m+j+r-1}, j+q\right). \end{aligned} \tag{4.4c}$$

The sequences are constrained by the following conditions:

$$1 \leqslant i_t \leqslant m \text{ and } 1 \leqslant j_t \leqslant n, \quad \text{for } -q \leqslant t \leqslant p, \tag{4.5a}$$

$$\sum_{r=1}^{s} a_{m+j+r-1} < t \text{ with } s = \sum_{r=1}^{t} a_{i-r}, \quad \text{for } t = 1,2,\ldots,q \tag{4.5b}$$

and

$$\sum_{r=1}^{s} a_{i-r} < t \text{ with } s = \sum_{r=1}^{t} a_{m+j+r-1},$$

$$\text{for } t = 1,\ldots,p. \tag{4.5c}$$

The values of p and q are the maximum values consistent with these constraints (4.5).

The set of all matrix positions on the upper and lower sequences through $(i_0 j_0)$ is denoted by $S_\Lambda(i_0 j_0)$. For given Λ the number of upper and lower sequences is equal to the degree of atypicality of Λ. It is convenient to denote the set of positions of the zeros of the atypicality matrix by Z_Λ and the set of positions through which either an upper or a lower sequence passes by S_Λ. Thus

$$Z_\Lambda = \{(i,j) : A(\Lambda)_{ij} = 0\}, \tag{4.6a}$$

and

$$S_\Lambda = \{(i,j) : (i,j) \in S_\Lambda(i_0 j_0), \text{ with } (i_0 j_0) \in Z_\Lambda\}. \tag{4.6b}$$

The first of the above conditions, (4.5a), ensures that the sequences remain wholly within the $m \times n$ atypicality matrix and the last two, (4.5b) and (4.5c) ensure that the upper and lower sequences do not "cross or meet" in a graphical presentation, except of course at the position of atypicality $(i_0 j_0)$ from which they emanate. Such a graphical presentation of the sequences is given, for example, in the case $\Lambda = [11;0;010]$ of sl(3/4) by

$$A_\Lambda = \begin{array}{c} \\ 1 \\ 1 \\ \\ \end{array} \begin{pmatrix} 4 & & 3 & & 1 & & 0 \\ & \nearrow & & & & & \\ 2 & & 1 & & -1 & \to & -2 \\ & \nearrow & & & \nearrow & & \\ 0 & \to & -1 & & -3 & & -4 \end{pmatrix}. \quad \begin{matrix} \\ \\ \\ \\ 0 \quad 1 \quad 0 \end{matrix} \tag{4.7}$$

The Dynkin labels a_i alongside the matrix A_Λ, as in the above example, determine the column and row jumps of the upper and lower sequences. Those at the left describe the rightward jumps of the upper sequences, and those at the bottom describe the upward jumps of the lower sequences.

The significance of these sequences is that they allow us to specify in a unified way a number of rival character formulas of the Kac–Weyl type (3.19). This is most usefully done in terms of appropriate generating matrices.

Definition 4.8: The *generating matrix* $G_X(\Lambda)$ associated with the Kac–Weyl character formula (3.19) is the $m \times n$ matrix with entries 0 or 1 defined by

$$G_X(\Lambda)_{ij} = \begin{cases} 1, & \beta_{ij} \in \Delta_X(\Lambda), \\ 0, & \beta_{ij} \in \Delta_1^+(\Lambda) \setminus \Delta_X(\Lambda). \end{cases} \tag{4.8}$$

With this definition it follows that the formal character (3.19) can be re-expressed in the form

$$\chi_X(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) \times w\left\{ e^{\Lambda+\rho_0} \prod_{i=1}^{m} \prod_{j=1}^{n} (1 + G_X(\Lambda)_{ij} e^{-\beta_{ij}}) \right\}. \tag{4.9}$$

Formulas of this type have already appeared in the literature. We have already cited the extreme cases: the Kac formula[12] (3.18), $\chi_K(\Lambda)$, and the Weyl formula[13] (3.20), $\chi_W(\Lambda)$. In addition there are three intermediate formulas: the Leites formula[30] (3.22), $\chi_L(\Lambda)$, the Hughes–King formula,[41] $\chi_H(\Lambda)$, and the Serganova–Serge'ev formula,[42] $\chi_S(\Lambda)$. This latter formula was also discovered independently by the present authors as part of this investigation, together with one new formula $\chi_J(\Lambda)$ that we propose here for the first time. These formulas are given in Table I by specifying for each formula the set of positions (i,j) in the matrix $G_X(\Lambda)$ such that $G_X(\Lambda)_{ij} = 0$.

As in (4.6), Z_Λ is the set of positions of the zeros of the atypicality matrix and S_Λ is the set of positions through which there passes either an upper or a lower sequence. The distinction between the various formulas lies merely in the extent to which zeros of the generating matrix are associated with positions on the upper and lower sequences emanating from the positions in the set Z_Λ.

The characters have been arranged in such a way that the successive subsets of Δ_1^+ denoted by $\Delta_1^+ \setminus \Delta_X(\Lambda)$ are included in one another reading from top to bottom. It should be further pointed out that just as the set Z_Λ is the set of positions of the zeros of the atypicality matrix A_Λ, so the set S_Λ is precisely the set of positions of zeros of $G_S(\Lambda)$. It follows that the character $\chi_S(\Lambda)$ is generated by precisely those odd roots β_{ij} such that (i,j) does not lie on any upper or lower sequence.

In the case $\Lambda = [11;0;010]$ of sl(3/4) for which the atypicality matrix and the associated sequences have been given in (4.7), we have the following generating matrices:

$$G_K(\Lambda) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad G_L(\Lambda) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix},$$

$$G_H(\Lambda) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad G_J(\Lambda) = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

$$G_S(\Lambda) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad G_W(\Lambda) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.10}$$

Of course in many other cases there are some coincidences between these six generating matrices. Moreover, even if the generating matrices differ, it is by no means the

TABLE I. Intermediate formulas.

$\chi_X(\Lambda)$	$\{(i,j) : \beta_{ij} \in \Delta_1^+(\Lambda) \setminus \Delta_X(\Lambda)\}$
$\chi_K(\Lambda)$	ϕ
$\chi_L(\Lambda)$	$\{(i_0 j_0) \in Z_\Lambda\}$
$\chi_H(\Lambda)$	$\{(i_t,j_t) \in S_\Lambda : (i_t,j_t) = (i_0, -t j_0)\ 0 \leqslant t \leqslant p$ or $(i_0 j_0 + t),\ 0 \leqslant t \leqslant q\}$
$\chi_J(\Lambda)$	$\{(i_t,j_t) \in S_\Lambda : (i_t,j_t) = t, -q \leqslant t \leqslant p\}$
$\chi_S(\Lambda)$	$\{(i_t,j_t) \in S_\Lambda : -q \leqslant t \leqslant p\}$
$\chi_W(\Lambda)$	$\{\{i,j\} : 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n\}$

case that the corresponding formal characters are themselves different.

Let Λ be an integral dominant weight of $\mathrm{sl}(m/n)$ and let $V(\Lambda)$ be the irreducible finite-dimensional module of $\mathrm{sl}(m/n)$ of highest weight Λ. Then we can state one theorem and two conjectures as follows:

If Λ is typical then no zeros appear in the atypicality matrix and $G_K(\Lambda) = G_L(\Lambda) = G_H(\Lambda) = G_J(\Lambda) = G_S(\Lambda)$. Hence by virtue of Kac's result, Theorem 3.10, we have the following theorem.

Theorem 4.11: If Λ is typical then

$$\chi_K(\Lambda) = \chi_L(\Lambda) = \chi_H(\Lambda) = \chi_J(\Lambda) = \chi_S(\Lambda) = \mathrm{ch}\ V(\Lambda). \tag{4.11}$$

Let Λ be atypical with $A(\Lambda)_{ij} = 0$ for some particular (i,j). If the upper and lower sequences through (i,j) defined by (4.4) consist simply of the single matrix position $(i_0,j_0) = (i,j)$, and the same is true of all such upper and lower sequences, then we say that Λ is *generic*. It follows from the conditions (4.5) that Λ is generic provided that for each $(i,j) \in Z_\Lambda$ we have both $a_{i-1} > 0$ if $i > 1$ and $a_{m+j} > 0$ if $j < n$. In such a case $S_\Lambda = Z_\Lambda$ and consequently $G_L(\Lambda) = G_H(\Lambda) = G_J(\Lambda) = G_S(\Lambda)$. Our analysis of many specific cases of this type leads us to the following conjecture.

Conjecture 4.12: If Λ is atypical but generic then

$$\chi_L(\Lambda) = \chi_H(\Lambda) = \chi_J(\Lambda) = \chi_S(\Lambda) = \mathrm{ch}\ V(\Lambda). \tag{4.12}$$

Similarly if $G_J(\Lambda) = G_S(\Lambda)$ we say that Λ is *normal*. Once again the analysis of many specific cases of this type leads us to a new conjecture.

Conjecture 4.13: If Λ is atypical but normal then

$$\chi_J(\Lambda) = \chi_S(\Lambda) = \mathrm{ch}\ V(\Lambda). \tag{4.13}$$

Unfortunately, if Λ is atypical and *abnormal* in that $G_J(\Lambda) \neq G_S(\Lambda)$ then our hoped-for third conjecture: $\chi_J(\Lambda) = \mathrm{ch}\ V(\Lambda)$ for all Λ, turns out to be false in general, although true for very many particular cases. Moreover, our counterexample to this formula suffices to rule out the general validity of any formula of the Kac–Weyl type as we shall now explain.

If Λ is atypical then, in the absence of any general proofs regarding the range of validity of the various formulas, it is necessary to carry out a case by case study of particular irreducible modules $V(\Lambda)$ of $\mathrm{sl}(m/n)$. This exercise is itself inhibited by the absence of many cases for which the character is actually known.

However a systematic study has revealed integral dominant weights Λ for which $\chi_X(\Lambda)$ for some X does not give correctly the required character of the irreducible module $V(\Lambda)$. First of all there are cases for which $\chi_X(\Lambda)$ is manifestly not the character of any module by virtue of the fact that its expansion in terms of irreducible characters of the even subalgebra $\mathrm{sl}(m) \oplus \mathbb{C} \oplus \mathrm{sl}(n)$ gives rise to negative multiplicities. For $\chi_L(\Lambda)$ one such example is provided by the sl(3/5) highest weight $\Lambda = [01;0;0002]$, while for $\chi_H(\Lambda)$ the same is true in the case of the sl(4/5) highest weight $\Lambda = [200;0;0001]$. No such example has been found for either $\chi_J(\Lambda)$ or $\chi_S(\Lambda)$.

TABLE II. Examples of a breakdown of validity of our formulas.

$\chi_X(\Lambda)$	Λ for which $\chi_X(\Lambda) \neq \mathrm{ch}\ V(\Lambda)$ for $\mathrm{sl}(m/n)$
$\chi_K(\Lambda)$	$\Lambda = [0;0;00]$ for sl(2/3)
$\chi_L(\Lambda)$	$\Lambda = [0;0;00]$ for sl(2/3)
$\chi_H(\Lambda)$	$\Lambda = [10;0;01]$ for sl(3/3)
$\chi_J(\Lambda)$	$\Lambda = [11;0;010]$ for sl(3/4)
$\chi_S(\Lambda)$	$\Lambda = [10;0;001]$ for sl(3/4)
$\chi_W(\Lambda)$	$\Lambda = [1;0;00]$ for sl(2/3)

Second, there are other examples, as shown in Table II, where there is a clear breakdown of the validity of each of our formulas $\chi_X(\Lambda)$ for some integral dominant Λ.

Of these results, only that appropriate to $\chi_J(\Lambda)$ requires any comment since all the others depend on the very well-known characters of the identity representation, the defining representation or the adjoint representation with highest weights $\Lambda = [0\cdots0;0;0\cdots0]$, $[10\cdots0;0;0\cdots0]$, or $[10\cdots0;0;0\cdots01]$, respectively.

Throughout the remainder of this section we set $G = \mathrm{sl}(3/4)$ and $\Lambda = [11;0;010]$. The corresponding atypicality matrix is displayed in (4.7), from which it can be seen that Λ is doubly atypical of type β_{31} and β_{14}. The various generating matrices discussed so far are shown in (4.10). The first of these generates the Kac character $\chi_K(\Lambda)$. By virtue of (3.17) this is in turn equal to the character of the Kac module $\overline{V}(\Lambda)$ of the superalgebra $G = \mathrm{sl}(3/4)$. This character may be readily expanded in terms of characters of the even subalgebra $G_{\bar{0}} = \mathrm{sl}(3) \oplus \mathbb{C} \oplus \mathrm{sl}(4)$ by making use of (3.22). Rather than give the complete expansion we content ourselves with noting that the level structure takes the form

$$1\ 6\ 18\ 34\ 56\ 70\ 79\ 70\ 56\ 34\ 18\ 6\ 1. \tag{4.14}$$

These numbers indicate at each level the number, together with their multiplicity, of $G_{\bar{0}}$ highest weights Σ, each associated with an irreducible module $V^0(\Sigma)$ of $G_{\bar{0}}$ that appears in the decomposition of the Kac module $\overline{V}(\Lambda)$ on restriction from G to $G_{\bar{0}}$. The level of Σ relative to that of Λ is determined by the number of positive odd roots which it is necessary to subtract from Λ to obtain Σ.

At the highest level there exists just one $G_{\bar{0}}$ highest weight, namely Λ itself, which may be written in the form $\Lambda = [11] \oplus [0] \oplus [010]$ appropriate to $G_{\bar{0}} = \mathrm{sl}(3) \oplus \mathbb{C} \oplus \mathrm{sl}(4)$. At the next highest level there are precisely six $G_{\bar{0}}$ highest weights:

$$\begin{aligned}
\Sigma_{11} &= \Lambda - \beta_{11} = [01] \oplus [1] \oplus [110], \\
\Sigma_{13} &= \Lambda - \beta_{13} = [01] \oplus [0] \oplus [001], \\
\Sigma_{21} &= \Lambda - \beta_{21} = [20] \oplus [1] \oplus [110], \\
\Sigma_{23} &= \Lambda - \beta_{23} = [20] \oplus [0] \oplus [001], \\
\Sigma_{31} &= \Lambda - \beta_{31} = [12] \oplus [0] \oplus [110], \\
\Sigma_{33} &= \Lambda - \beta_{33} = [12] \oplus [-1] \oplus [001].
\end{aligned} \tag{4.15}$$

Of these only Σ_{31} involves an atypical odd root, namely β_{31}.

There exists a second-order Casimir operator C in the center $Z(G)$ of the enveloping algebra $U(G)$ of G such that

for any vector $v \in \overline{V}(\Lambda)$ we have[23]

$$Cv = C(\Lambda)v,$$

where

$$C(\Lambda) = \langle \Lambda + \rho | \Lambda + \rho \rangle - \langle \rho | \rho \rangle. \tag{4.16}$$

Moreover, if v_Σ is an eigenvector of the Borel subalgebra $\mathfrak{b}$ such that $\mathfrak{n}^+ v_\Sigma = 0$ and $hv_\Sigma = \Sigma(h)v_\Sigma$, then v_Σ is of weight Σ and

$$Cv_\Sigma = (\langle \Sigma + \rho | \Sigma + \rho \rangle - \langle \rho | \rho \rangle) v_\Sigma. \tag{4.17}$$

It follows that, if any vector $v_\Sigma \in \overline{V}(\Lambda)$ of weight Σ is to be a highest weight vector not only of $G_{\bar{0}}$ but also of G then, we must necessarily have

$$\langle \Sigma + \rho | \Sigma + \rho \rangle = \langle \Lambda + \rho | \Lambda + \rho \rangle. \tag{4.18}$$

For $\Sigma = \Lambda - \beta$ with $\langle \beta | \beta \rangle = 0$ this implies the atypicality condition $\langle \Lambda + \rho | \beta \rangle = 0$.

All this means that, of the 6 weights Σ_{ij} listed in (4.15), only Σ_{31} may correspond to a G highest weight vector of a submodule of $\overline{V}(\Lambda)$. It follows that corresponding to each of the other five weights we have $G_{\bar{0}}$ highest weight vectors that necessarily belong to the irreducible G module $V(\Lambda)$. Turning to the case $\Sigma = \Sigma_{31}$ and $\beta = \beta_{31} = \beta_{m1}$, the vector $v_\Sigma = e(-\beta)v_\Lambda$ is a $G_{\bar{0}}$ highest weight vector. This may be seen by noting that for each $\alpha \in \Delta_0^+$

$$\begin{aligned} e(\alpha)v_\Sigma &= e(\alpha)e(-\beta)v_\Lambda \\ &= [e(\alpha), e(-\beta)]v_\Lambda + e(-\beta)e(\alpha)v_\Lambda = 0, \end{aligned} \tag{4.19}$$

since $[e(\alpha), e(-\beta_{m1})] = 0$ and $e(\alpha)v_\Lambda = 0$. Moreover

$$\begin{aligned} e(\beta)v_\Sigma &= e(\beta)e(-\beta)v_\Lambda \\ &= [e(\beta), e(-\beta)]v_\Lambda + e(-\beta)e(\beta)v_\Lambda \\ &= h_\beta v_\Lambda = \Lambda(h_\beta)v_\Lambda = \langle \Lambda | \beta \rangle v_\Lambda \\ &= \langle \Lambda + \rho | \beta \rangle v_\Lambda = 0, \end{aligned} \tag{4.20}$$

since $[e(\beta), e(-\beta)]v_\Lambda = h_\beta v_\Lambda$ and $e(\beta)v_\Lambda = 0$. Together, (4.19) and (4.20) imply

$$e(\alpha_i)v_\Sigma = 0, \quad \text{for } \alpha_i \in \Pi, \tag{4.21}$$

where Π is the set of simple roots defined, in the distinguished basis, by (2.12). Since $\mathfrak{n}^+$ is generated by $\{e(\alpha): \alpha \in \Pi\}$ it follows that $\mathfrak{n}^+ v_\Sigma = 0$. Hence v_Σ is a G highest weight vector.

Correspondingly there exists a submodule of the Kac module $\overline{V}(\Lambda)$ with highest weight vector v_Σ that is to be factored out of $\overline{V}(\Lambda)$, along perhaps with other submodules, in forming the irreducible module $V(\Lambda)$.

Putting this information together we see that any formula of the Kac–Weyl type (3.19) must, when expressed in the form (4.9), be associated with a generating matrix $G_X(\Lambda)$ of the form:

$$G_X(\Lambda) = \begin{pmatrix} 1 & x & 1 & u \\ 1 & y & 1 & v \\ 0 & z & 1 & w \end{pmatrix}, \tag{4.22}$$

where each letter indicates a matrix element that could be 0 or 1. The six matrix elements that are fixed are those determined by the above arguments regarding the weight vectors corresponding to (4.17). Notice that we have already eliminated not only $\chi_K(\Lambda)$ and $\chi_W(\Lambda)$ but also $\chi_S(\Lambda)$ by this argument since the corresponding generating matrices are not of the form (4.22).

Now comes the \$64,000 question: Do any of the 64 possible generating matrices, with each letter in (4.22) equal to 0 or 1, give rise to a character $\chi_X(\Lambda)$ that could possibly be the required irreducible character $\mathrm{ch}(V(\Lambda))$?

The answer is no! But first we have to discuss tests involving more than just the top two levels. Exactly 32 cases could be ruled out by the fact that the expansion of $\chi_X(\Lambda)$ in terms of $\mathrm{sl}(3) \oplus \mathbb{C} \oplus \mathrm{sl}(4)$ characters gives rise to negative multiplicities, but we can do better than this. Our previous analysis has already demonstrated that the Kac module $\overline{V}(\Lambda)$ with level structure 1 6 18 $\cdots$ contains a composition factor with highest weight vector v_Σ. Applying exactly the same analysis to the Kac module $\overline{V}(\Sigma)$, which has level structure 1 9 32 $\cdots$, indicates the existence of a composition factor in $\overline{V}(\Sigma)$ with highest weight vector v_Ω where $\Omega = \Sigma - \beta = \Lambda - 2\beta$. Moreover, the remaining 8 $G_{\bar{0}}$ highest weights at this level are of the form $\Sigma - \beta'$ with β' not atypical. The Casimir argument then implies that the corresponding 8 $G_{\bar{0}}$ highest weight vectors belong to the irreducible module $V(\Sigma)$. This must therefore have level structure 1 8 $\cdots$. Subtracting this from the level structure (4.14) of $\overline{V}(\Lambda)$ leads to a level structure

$$1\ 5\ 10\ \cdots. \tag{4.23}$$

Once again, recourse to the Casimir argument implies that just as the 5 $G_{\bar{0}}$ highest weights at level 1 do not satisfy the condition (4.18), nor do any of the 10 at level 2. We conclude that (4.23) gives the first portion of the level structure of the irreducible module $V(\Lambda)$.

The 64 possible generating matrices given by (4.22) lead to a level structure consistent with (4.23) in precisely 12 cases: any combination of $(xyz) = (111)$, (110), (100), (010), with $(uvw) = (111)$, (101), (011). However of these, 6 cases lead to negative multiplicities. The remaining 6 cases lead to one or other of the following level structures

$$1\ 5\ 10\ 10\ 5\ 2\ 3\ 1, \tag{4.24a}$$

$$1\ 5\ 10\ 10\ 7\ 3\ 3\ 1, \tag{4.24b}$$

$$1\ 5\ 10\ 12\ 16\ 14\ 9\ 4\ 1, \tag{4.24c}$$

$$1\ 5\ 10\ 11\ 8\ 5\ 4\ 3\ 2\ 1. \tag{4.24d}$$

Now for the final nail in the coffin of any formula of the Kac–Weyl type! Consider the completely new module of highest weight Λ formed from the tensor product of two irreducible $\mathrm{sl}(3/4)$ modules $V(\Lambda_1)$ and $V(\Lambda_2)$ with

$$\Lambda_1 = [11;0;000] \quad \text{and} \quad \Lambda_2 = [00;0;010]. \tag{4.25}$$

If v_{Λ_1} and v_{Λ_2} are the highest weight vectors of $V(\Lambda_1)$ and $V(\Lambda_2)$, respectively, then $v_{\Lambda_1} \otimes v_{\Lambda_2}$ is necessarily the highest weight vector of the tensor product module $V(\Lambda_1) \otimes V(\Lambda_2)$. Moreover, Λ_1 and Λ_2 have been chosen so that this weight vector $v_{\Lambda_1} \otimes v_{\Lambda_2}$ is of weight $\Lambda = \Lambda_1 + \Lambda_2$. It follows that $V(\Lambda)$ is a quotient module of $V(\Lambda_1) \otimes V(\Lambda_2)$ so that the character $\mathrm{ch}(V(\Lambda_1) \otimes V(\Lambda_2))$ of the tensor product module must contain the character of the irreducible module $V(\Lambda)$ as a summand. The weights Λ_1 and Λ_2 have also been chosen so that Λ_1 is the highest weight of a covar-

iant tensor irreducible module and Λ_2 is the highest weight of a contravariant tensor irreducible module. Although the proof of this will only be described in the next section, these two weights are normal and their characters are well known, being determined by means of the character formula of Berele and Regev[16] or, equivalently, by $\chi_S(\Lambda_1)$ and $\chi_S(\Lambda_2)$. Correspondingly the $\mathrm{sl}(3/4)\to\mathrm{sl}(3)\oplus\mathbb{C}\oplus\mathrm{sl}(4)$ decompositions are given by

$$[11;0;000]\to[11]\times[0]\times[000]+[20]\times[1]\times[100]+[01]\times[1]\times[100]+[10]\times[1]\times[010]+[10]\times[2]\times[200]+[00]\times[2]\times[110], \tag{4.26}$$

$$[00;0;010]\to[00]\times[0]\times[010]+[01]\times[-1]\times[001]+[02]\times[-2]\times[000].$$

Decomposing the tensor product of the corresponding $G_{\bar{0}}$ characters in terms of $G_{\bar{0}}$ characters then gives an expression that must contain ch $V(\Lambda)$ as a proper summand. This simple criterion immediately eliminates 60 of our possible 64 generating matrices, including those corresponding to $\chi_L(\Lambda)$, $\chi_H(\Lambda)$, and $\chi_J(\Lambda)$. This serves to justify the entry for $\chi_J(\Lambda)$ in our earlier tabulation.

The remaining 4 of our 64 cases give rise to one or other of the level structures

$$1\ 5\ 8\ 7\ 2, \tag{4.27a}$$

$$1\ 5\ 6\ 5\ 1, \tag{4.27b}$$

in direct contradiction with (4.24). Indeed from (4.25) it can be seen that the total number of levels of $V(\Lambda_1)$ and $V(\Lambda_2)$ below Λ_1 and Λ_2 are 3 and 2, respectively, so that the irreducible quotient module $V(\Lambda)$ of $V(\Lambda_1)\otimes V(\Lambda_2)$ has a maximum of 5 levels below $\Lambda=\Lambda_1+\Lambda_2$. This immediately rules out all possibilities (4.24).

We are therefore led to the following conclusion.

Proposition 4.28: For $G=\mathrm{sl}(3/4)$ and $\Lambda=[11;0;010]$, there exists no generating matrix $G_X(\Lambda)$ such that

$$\chi_X(\Lambda)=\mathrm{ch}\ V(\Lambda). \tag{4.28}$$

Therefore we have the following corollary.

Corollary 4.29: No formula of the Kac–Weyl type (3.19) can give correctly the characters of all the finite-dimensional irreducible modules $V(\Lambda)$ of $\mathrm{sl}(m/n)$ for all m and n.

V. CHARACTERS OF IRREDUCIBLE COVARIANT TENSOR MODULES

Quite apart from the distinction between typical and atypical irreducible finite-dimensional modules of $\mathrm{sl}(m/n)$, it is possible to distinguish between such modules on the basis of their relationship to tensor modules of various kinds. Each irreducible module $V(\Lambda)$ for which the highest weight Λ is not only dominant but also integral in the sense of Definition 3.5, is equivalent either to some tensor module or to some quotient of a tensor module. The requirement that Λ be integral dominant excludes some modules $V(\Lambda)$ for which Λ is admissible but not integral, but these modules are all typical.

The tensor modules we have in mind are submodules or quotient modules of the tensor algebra of V and V^*, where V is the $(m+n)$-dimensional vector space carrying the natural matrix representation (2.2) of $\mathrm{gl}(m/n)$, and V^* is the conjugate of V carrying the matrix representation contragredient to (2.2). Such tensor modules have been considered by a number of authors,[15–25] with some successes and some failures in the effort to obtain character formulas for those that were irreducible.

Berele and Regev,[16] and later Serge'ev,[17] showed that the tensor product $V^{\otimes N}$ of N copies of the natural $(m+n)$-dimensional representation V of $\mathrm{gl}(m/n)$ is completely reducible, and that the irreducible components, V_λ, can be labeled by means of a partition λ of N of length $l(\lambda)$ and weight $|\lambda|$, where $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_p)$, with $l(\lambda)=p$, $|\lambda|=\lambda_1+\lambda_2+\cdots+\lambda_p=N$, and $\lambda_i\geqslant\lambda_{i+1}>0$ for $i=1,2,\ldots,p-1$, satisfying the condition $\lambda_{m+1}\leqslant n$. These representations V_λ are known as irreducible *covariant tensor representations*. Furthermore, by exploiting the properties of the symmetric group S_N, Berele and Regev[16] and Serge'ev[17] established an explicit character formula appropriate to all irreducible covariant tensor representations of $\mathrm{gl}(m/n)$.

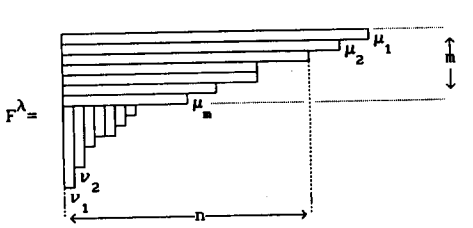
Similar results are valid for tensor products $(V^*)^{\otimes N}$ of the contragredient V^* of V. Each irreducible component, $V_{\bar{\lambda}}$, may be labelled by means of a partition λ of N with $\lambda_{m+1}\leqslant n$. These representations $V_{\bar{\lambda}}$ are known as irreducible *contravariant tensor representations* and their character formula is obtained trivially from that appropriate to the covariant case. Unfortunately, tensor products involving both V and V^* are, in general, not completely reducible and to date there exists no character formula for all the irreducible *mixed tensor representations*. However, they may conveniently be denoted by $V_{\bar{\tau},\sigma}$ since they occur as the leading components in the reduction of tensor products of the form $V_{\bar{\tau}}\otimes V_\sigma$.

Returning to the covariant tensor case, the irreducible submodule V_λ of $V^{\otimes N}$ specified by the partition λ, with $\lambda_{m+1}\leqslant n$, is necessarily finite dimensional. Hence, by virtue of Theorem 3.8, there must exist a dominant weight Λ_λ such that V_λ is isomorphic to $V(\Lambda_\lambda)$. The relation between $\Lambda_\lambda=(\mu_1\mu_2\cdots\mu_m|\nu_1\nu_2\cdots\nu_n)$ and $\lambda=(\lambda_1\lambda_2\cdots)$ is such that:[17,25]

$$\mu_i=\lambda_i\quad(1\leqslant i\leqslant m), \tag{5.1a}$$

$$\nu_a=\langle\lambda'_a-m\rangle=\max\{0,\lambda'_a-m\}\quad(1\leqslant a\leqslant n), \tag{5.1b}$$

where λ' is the partition conjugate to λ. In terms of the Young diagram F^λ specified by λ, the parts λ_i of λ and λ'_a of λ' are the row and column lengths, respectively, while the components μ_i and ν_a of Λ_λ can be identified as shown below:



(5.2)

The Dynkin labels corresponding to the representation V_λ then follow from (2.21). It is not difficult to see from (4.1) that such a representation is typical or atypical according as $\lambda_m \geqslant n$ or $\lambda_m < n$, respectively.

In precisely the same way the irreducible contravariant module $V_{\bar\lambda}$ specified by $\lambda = (\lambda_1 \lambda_2 \cdots)$ is isomorphic to $V(\Lambda_{\bar\lambda})$ with $\Lambda_{\bar\lambda} = (\mu_1 \mu_2 \cdots \mu_m | \nu_1 \nu_2 \cdots \nu_n)$, where

$$\mu_i = -\langle \lambda_{m-i+1} - n \rangle \quad (1 \leqslant i \leqslant m) \tag{5.3a}$$

$$\nu_a = -\lambda'_{n-a+1} \quad (1 \leqslant a \leqslant n). \tag{5.3b}$$

As before, this module is typical or atypical according as $\lambda_m \geqslant n$ or $\lambda_m < n$, respectively.

Conversely if $\Lambda = (\mu_1 \mu_2 \cdots \mu_m | \nu_1 \nu_2 \cdots \nu_n)$ is integral dominant with $\mu_i \geqslant 0$ for $1 \leqslant i \leqslant m$, $\nu_a \geqslant 0$ for $1 \leqslant a \leqslant n$ and $\mu_m \geqslant \#\{a : \nu_a > 0, 1 \leqslant a \leqslant n\}$, then there exists λ such that $V(\Lambda)$ is isomorphic to the irreducible covariant tensor module V_λ. Similarly if $\Lambda = (\mu_1 \mu_2 \cdots \mu_m | \nu_1 \nu_2 \cdots \nu_n)$ is integral dominant with $\mu_i \leqslant 0$ for $1 \leqslant i \leqslant m$, $\nu_a \leqslant 0$ for $1 \leqslant a \leqslant n$ and $-\nu_1 \geqslant \#\{i : \mu_i < 0, 1 \leqslant i \leqslant m\}$ then there exists λ such that $V(\Lambda)$ is isomorphic to the irreducible contravariant module $V_{\bar\lambda}$. For $\mathrm{sl}(m/n)$ these conditions can be expressed in terms of the Kac–Dynkin labels: let $\Lambda = [a_1 a_2 \cdots a_{m-1}; a_m; a_{m+1} \cdots a_{m+n-1}]$ be integral dominant. Then $V(\Lambda)$ is isomorphic to an irreducible covariant tensor module provided that $a_m \geqslant 0$ and either $a_{m+b} = 0$ for $b = 1,2,\ldots,n-1$ or

$$a_m \geqslant c + \sum_{b=1}^{c} a_{m+b}, \text{ where } c = \max(b) \text{ such that } a_{m+b} > 0, \tag{5.4a}$$

and to an irreducible contravariant tensor module if $a_m \leqslant 0$ and either $a_i = 0$ for $i = 1,2,\ldots,m-1$ or

$$-a_m \geqslant k + \sum_{j=1}^{k} a_{m-j}, \text{ where } k = \max(j) \text{ such that } a_{m-j} > 0. \tag{5.4b}$$

Of course, not all irreducible modules $V(\Lambda)$ are either covariant or contravariant. For example the adjoint module has highest weight $\Lambda = [10\cdots0;0;0\cdots01]$, thereby violating both (5.4a) and (5.4b).

The importance of the irreducible covariant tensor modules of $\mathrm{gl}(m/n)$ lies in the fact that their characters are known.[16,17] Just as the characters of irreducible covariant tensor modules of $\mathrm{gl}(m)$ may be expressed in terms of S functions,[39] so those of $\mathrm{gl}(m/n)$ may be expressed in terms of supersymmetric S functions.

We adopt the notation and terminology of Macdonald[40] whereby the S function of $(\mathbf{x}) = (x_1 x_2 \cdots x_m)$ specified by the partition σ is denoted by $s_\sigma(\mathbf{x})$. Such S functions satisfy the product and quotient rules:

$$s_\sigma(\mathbf{x}) s_\tau(\mathbf{x}) = \sum_\lambda c^\lambda_{\sigma\tau} s_\lambda(\mathbf{x}) \tag{5.5}$$

$$s_{\lambda/\tau}(\mathbf{x}) = \sum_\sigma c^\lambda_{\sigma\tau} s_\sigma(\mathbf{x}), \tag{5.6}$$

where the coefficients $c^\lambda_{\sigma\tau}$ are the famous Littlewood–Richardson coefficients, and the summations are over partitions λ and σ, as indicated, with $|\lambda| = |\sigma| + |\tau|$. With the identification $x_i = e^{\epsilon_i}$ for $1 \leqslant i \leqslant m$ the S function $s_\lambda(\mathbf{x})$, with $l(\lambda) \leqslant m$ is just the Weyl character (1.1), ch $V^0(\lambda)$, of the irreducible module of $\mathrm{gl}(m)$ with highest weight λ. By generalizing the notion of an S function, Berele and Regev[16] proved the following.

Theorem 5.7: Let $V(\Lambda_\lambda)$ be the irreducible $\mathrm{gl}(m/n)$ module isomorphic to the covariant tensor module V_λ specified by the partition λ, with Λ_λ related to λ by (5.1), and let

$$x_i = e^{\epsilon_i}, \quad \text{for } 1 \leqslant i \leqslant m,$$

and

$$y_a = e^{\delta_a}, \quad \text{for } 1 \leqslant a \leqslant n.$$

Then the character of $V(\Lambda_\lambda)$ is given by

$$\text{ch } V(\Lambda_\lambda) = s_\lambda(\mathbf{x}/\mathbf{y}), \tag{5.7a}$$

where $s_\lambda(\mathbf{x}/\mathbf{y})$ is the supersymmetric S function of $\mathbf{x} = (x_1,\ldots,x_m)$ and $\mathbf{y} = (y_1,\ldots,y_n)$ defined by

$$s_\lambda(\mathbf{x}/\mathbf{y}) = \sum_\tau s_{\lambda/\tau}(\mathbf{x}) s_{\tau'}(\mathbf{y}) = \sum_{\sigma,\tau} c^\lambda_{\sigma\tau} s_\sigma(\mathbf{x}) s_{\tau'}(\mathbf{y}), \tag{5.7b}$$

with $l(\sigma) \leqslant m$ and $l(\tau') \leqslant n$.

In the notation of (3.20), this immediately gives the following.

Corollary 5.8:

$$\text{ch } V(\Lambda_\lambda) = \sum_{\sigma,\tau} c^\lambda_{\sigma\tau} \chi_W(\Sigma_{\sigma\tau'}), \tag{5.8a}$$

with

$$\Lambda_\lambda = (\lambda_1,\lambda_2,\ldots,\lambda_m | \langle \lambda'_1 - m \rangle, \langle \lambda'_2 - m \rangle, \ldots, \langle \lambda'_n - m \rangle), \tag{5.8b}$$

$$\Sigma_{\sigma\tau'} = (\sigma_1,\sigma_2,\ldots,\sigma_m | \tau'_1,\tau'_2,\ldots,\tau'_n). \tag{5.8c}$$

Similarly, we have in the case of irreducible contravariant tensor representations:

$$\text{ch } V(\Lambda_{\bar\lambda}) = s_\lambda(\overline{\mathbf{x}/\mathbf{y}}) = \sum_{\sigma,\tau} c^\lambda_{\sigma\tau} s_\sigma(\bar{\mathbf{x}}) s_{\tau'}(\bar{\mathbf{y}}). \tag{5.9a}$$

where $\bar{\mathbf{x}} = (x_1^{-1},\ldots,x_m^{-1})$ and $\bar{\mathbf{y}} = (y_1^{-1},\ldots,y_n^{-1})$, so that

$$\text{ch } V(\Lambda_{\bar\lambda}) = \sum_{\sigma,\tau} c^\lambda_{\sigma\tau} \chi_W(\Sigma_{\overline{\sigma\tau'}}), \tag{5.9b}$$

with

$$\Lambda_{\bar\lambda} = (\langle \lambda_1 - n \rangle, \langle \lambda_2 - n \rangle, \ldots, \langle \lambda_m - n \rangle | -\lambda'_n, \ldots, -\lambda'_2, -\lambda'_1), \tag{5.9c}$$

$$\Sigma_{\overline{\sigma\tau'}} = (-\sigma_m,\ldots,-\sigma_2,-\sigma_1 | -\tau'_n,\ldots,-\tau'_2,-\tau'_1). \tag{5.9d}$$

Unfortunately, a possible extension of (5.8) and (5.9) to mixed tensor representations, namely,[25]

$$s_{\bar\tau,\sigma}(\mathbf{x}/\mathbf{y}) = \sum_\zeta (-1)^{|\zeta|} s_{\sigma/\zeta}(\mathbf{x}/\mathbf{y}) s_{\tau/\zeta'}(\overline{\mathbf{x}/\mathbf{y}}), \tag{5.10a}$$

where, for all $\mathbf{z}$,

$$s_{\lambda/\mu}(\mathbf{z}) = \sum_\nu c^\lambda_{\mu\nu} s_\nu(\mathbf{z}), \tag{5.10b}$$

does not, in general, yield characters of irreducible modules, although it does coincide with certain mixed tensor determinantal characters defined elsewhere.[19,24]

In the case of irreducible covariant tensor representations it is possible to make contact with the work of Sec. IV. We have the following lemma.

Lemma 5.11: If Λ is the highest weight of a covariant tensor representation specified by a partition λ, then Λ is normal, in the sense that $G_J(\Lambda) = G_S(\Lambda)$, and

$$G_S(\Lambda)_{ib} = \begin{cases} 1, & \text{if } (i,b) \in F^\lambda, \\ 0, & \text{if } (i,b) \in F^\lambda, \end{cases} \quad 1 \leqslant i \leqslant m \text{ and } 1 \leqslant b \leqslant n, \tag{5.11}$$

where $(i,b) \in F^\lambda$ if and only if there is a box in the ith row and b th column of the Young diagram F^λ.

Proof: Since the hook length, h_{ib}, of the box in the ith row and b th column of F^λ is always positive and given by

$$h_{ib} = \lambda_i + \lambda'_b - i - b + 1, \tag{5.12}$$

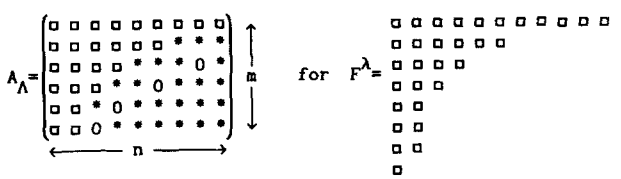
it follows from the atypicality condition (4.1c)

$$A(\Lambda)_{ib} = \mu_i + \nu_b + m - i - b + 1 = 0,$$
$$1 \leqslant i \leqslant m \text{ and } 1 \leqslant b \leqslant n,$$

and the relation (5.1) that any zero of the atypicality matrix lies at a position $(i,b) \notin F^\lambda$. To be more precise, the zeros of the atypicality matrix are given by

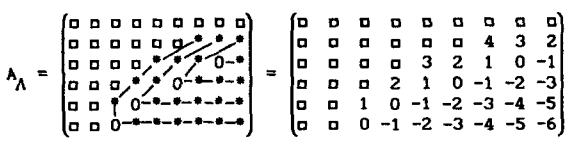
$$A(\Lambda)_{ib} = 0 \Leftrightarrow b = \lambda_i + m - i + 1,$$
$$1 \leqslant i \leqslant m \text{ and } 1 \leqslant b \leqslant n. \tag{5.13}$$

Schematically, the relation between the atypicality matrix A_Λ and the Young diagram F^λ is as follows for $\lambda = (11,6432221)$, $m = 6$ and $n = 9$:



(5.14)

In A_Λ, □ or * stand for a nonzero integer, and 0 stands for zero. Since the Dynkin labels a_{m+j} vanish for $j \geqslant \lambda_m + 1$ each lower sequence is horizontal. On the other hand, since the horizontal steps in each upper sequence are determined by the same Dynkin labels $a_i = \lambda_i - \lambda_{i+1}$ $(1 \leqslant i \leqslant m-1)$, it follows that upper sequences are parallel to one another. Hence the pairs of upper and lower sequences are nested as shown below on the left:



(5.15)

Moreover, another consequence of the fact that the Dynkin labels a_{m+j} with $j \geqslant \lambda_m + 1$ are all zero, is that the elements of the atypicality matrix to the right of the □'s in each row are consecutive integers. This is illustrated on the right in the above example. Starting from the known positions of the zeros and the nested nature of the pairs of upper and lower sequences, it can immediately be seen that the normality condition, $(i_t, b_t) = t$ for $-q \leqslant t \leqslant p$, is satisfied for each sequence. Therefore the highest weight Λ corresponding to λ is normal. Moreover the union of the sequences is the set of points $(i,b) \in F^\lambda$ with $1 \leqslant i \leqslant m$ and $1 \leqslant b \leqslant n$. The result (5.11) then follows from the definition of $G_S(\Lambda)$ given in Sec. IV.

Example: Let $\lambda = (5,2,1,1,1)$ for sl(3/4). Then $\Lambda = (5,2,1|2,0,0,0) = [3,1;3;2,0,0]$ in terms of $\epsilon\delta$ components and Kac–Dynkin labels, respectively. The Young diagram F^λ, A_Λ and G_S are now given by

$$F^\lambda = \begin{array}{l} \blacksquare\blacksquare\blacksquare\blacksquare\square \\ \blacksquare\blacksquare \\ \blacksquare \\ \square \\ \square \end{array},$$

$$A_\Lambda = \begin{pmatrix} 9 & 6 & 5 & 4 \\ 5 & 2 & 1 & 0 \\ 3 & 0 & -1 & -2 \end{pmatrix},$$

$$G_S = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{5.16}$$

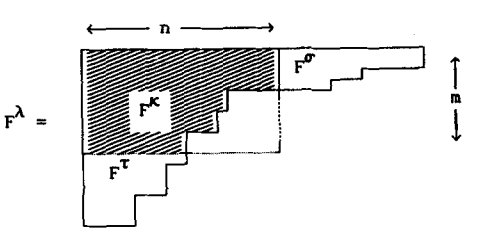
In F^λ we have shaded the boxes that fall within the $m \times n$ rectangle and which determine the positions of the entries 1 in G_S.

An immediate consequence of Lemma 5.11 is the following corollary.

Corollary 5.17: If Λ is the highest weight of the irreducible covariant tensor representation specified by the partition λ with $\lambda_{m+1} \leqslant n$, then

$$\chi_S(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\Lambda + \rho_0} \prod_{(i,b) \in F^\kappa} (1 + e^{-\beta_{ib}}) \right\}, \tag{5.17}$$

where F^κ is indicated in the following diagram:



(5.18)

In other words, F^κ is that part of F^λ inside the $m \times n$ rectangle, and F^σ and F^τ are the parts of F^λ outside this rectangle.

We are now in a position to formulate the main result of this section.

Theorem 5.19: If Λ is the highest weight of the irreducible module $V(\Lambda)$ isomorphic to the irreducible covariant tensor module specified by the partition λ, then

$$\text{ch } V(\Lambda) = \chi_S(\Lambda), \tag{5.19}$$

with $\chi_S(\Lambda)$ given by (5.17).

In order to prove this theorem it is only necessary to establish that $\chi_S(\Lambda) = s_\lambda(\mathbf{x/y})$, with $\mathbf{x}$ and $\mathbf{y}$ defined as in Theorem 5.7. To illustrate how this is done we first establish some further properties of $s_\lambda(\mathbf{x/y})$.

Lemma 5.20: The generating function for the functions $s_\lambda(\mathbf{x/y})$ is given by

$$G(\mathbf{x},\mathbf{y},\mathbf{z}) = \prod (1 + y_a z_t) \Big/ \prod (1 - x_i z_s)$$

$$= \sum_\lambda s_\lambda(\mathbf{x/y}) s_\lambda(\mathbf{z}), \tag{5.20}$$

where $\mathbf{x} = (x_1,...,x_m)$, $\mathbf{y} = (y_1,...,y_n)$, $\mathbf{z} = (z_1,...,z_p)$, $1 \leqslant i \leqslant m$, $1 \leqslant a \leqslant n$, $1 \leqslant s$, $t \leqslant p$, and the summation is over all partitions λ of length $\leqslant p$.

Proof: Making use of the S function expansions of the numerator and denominator products appearing in (5.20)[40] one obtains

$$\frac{\prod (1 + y_a z_t)}{\prod (1 - x_i z_s)} = \sum_\nu s_\nu(\mathbf{y}) s_\nu(\mathbf{z}) \sum_\mu s_\mu(\mathbf{x}) s_\mu(\mathbf{z}) \tag{5.21a}$$

$$= \sum_{\mu,\nu,\lambda} s_\mu(\mathbf{x}) s_\nu(\mathbf{y}) c^\lambda_{\mu\nu} s_\lambda(\mathbf{z}), \tag{5.21b}$$

where the summation is over all partitions μ, ν, and λ, and where $c^\lambda_{\mu\nu}$ are the Littlewood–Richardson coefficients.[39,40] Comparison with our definition, (5.7), of $s_\lambda(\mathbf{x/y})$ proves the validity of (5.20).

Now let $\mathscr{P}(\mathbf{x})$ denote the ring consisting of all symmetric polynomials in $x_1, x_2, ..., x_m$ with coefficients in $\mathbb{Z}$, and define the ring of *doubly-symmetric* polynomials as $\mathscr{P}(\mathbf{x},\mathbf{y}) = \mathscr{P}(\mathbf{x}) \otimes_{\mathbb{Z}} \mathscr{P}(\mathbf{y})$. An element $p \in \mathscr{P}(\mathbf{x},\mathbf{y})$ is said to have the *cancellation property* if p is such that when the substitution $x_1 = t$, $y_1 = -t$ is made in p, the resulting polynomial is independent of t. The elements of $\mathscr{P}(\mathbf{x},\mathbf{y})$ satisfying the cancellation property are known as supersymmetric polynomials. They form a subring of $\mathscr{P}(\mathbf{x},\mathbf{y})$ which we denote by $\mathscr{P}(\mathbf{x/y})$. One can see from Lemma 5.20 that every $s_\sigma(\mathbf{x/y})$ satisfies the cancellation property, and we have the following theorem due to Stembridge.[46]

Theorem 5.22: Every element of $\mathscr{P}(\mathbf{x/y})$ is a $\mathbb{Z}$-linear combination of the $s_\sigma(\mathbf{x/y})$; the set of $s_\sigma(\mathbf{x/y})$ form a $\mathbb{Z}$ basis of $\mathscr{P}(\mathbf{x/y})$.

There exist at least two proofs of this theorem in the literature.[45,46]

Let us now return to the formula (5.17) for $\chi_S(\Lambda)$ from which it follows, by making use of the parameters $x_i = e^{\epsilon_i}$ and $y_a = e^{\delta_a}$ in the known expression for L_0 and the fact that $W = S_m \times S_n$, that

$$\chi_S(\Lambda) = s^{(w)}_\lambda(\mathbf{x/y}), \tag{5.23}$$

where

$$s^{(w)}_\lambda(\mathbf{x/y}) = \left[\prod_{i<j}(x_i - x_j) \prod_{a<b}(y_a - y_b)\right]^{-1} \times \sum_{w \in S_m \times S_n} \epsilon(w) w\left\{x_1^{\sigma_1 + m - 1} \cdots x_m^{\sigma_m} \times y_1^{\tau_1 + n - 1} \cdots y_n^{\tau_n} \prod_{(i,a) \in F^\kappa} (x_i + y_a)\right\}. \tag{5.24}$$

Now the final link needed to prove Theorem 5.19 is provided by the following theorem.

Theorem 5.25:

$$s^{(w)}_\sigma(\mathbf{x/y}) = s_\sigma(\mathbf{x/y}). \tag{5.25}$$

This identity appears to have been discovered independently by ourselves and by Serge'ev. The latter communicated it to Pragacz who provided a proof of its validity.[45] The first step in his very elegant proof is to show that $s^{(w)}_\sigma(\mathbf{x/y})$ satisfies the cancellation property. Then Theorem 5.22 enables him to write $s^{(w)}_\sigma(\mathbf{x/y})$ as a $\mathbb{Z}$-linear combination of $s_\lambda(\mathbf{x/y})$ functions. Finally, he uses a specialization argument to show that there is only one term in the linear combination. We shall not reproduce Pragacz's proof here, but simply refer the interested reader to his paper.[45]

Despite the negative conclusion reached in the last section, we would contend that character formulas of the Kac–Weyl type are still of some significance. The main result in this section, namely Theorem 5.19 whose proof is now completed, supports this contention. Furthermore, Theorem 5.19 also affords support for the Conjecture 4.13 of the last section, in that all the atypical highest weights Λ involved in Theorem 5.19 are normal, as proved in Lemma 5.11. In addition, it is easy to see that the theorem may be extended so that it covers not only all irreducible covariant tensor modules but also all irreducible contravariant tensor modules.

Of course our counterexample, Proposition 4.28, to the general validity of Kac–Weyl character formulas corresponds to an irreducible mixed tensor module whose highest weight is not only atypical but also abnormal. However, extensive computer calculations involving the characters of more than 100 000 irreducible representations of $\mathrm{sl}(m/n)$, with $1 \leqslant m \leqslant n \leqslant 6$, all having normal atypical highest weights, have revealed no counterexample to Conjecture 4.13.

Returning for the moment to our atypical, abnormal example of the last section, we should point out that the argument leading to Proposition 4.13 leaned heavily on the claimed decompositions (4.26) of the irreducible modules $V(\Lambda_1)$ and $V(\Lambda_2)$ on restriction from sl(3/4) to $\mathrm{sl}(3) \oplus \mathbb{C} \oplus \mathrm{sl}(4)$. From (4.23) it is easy to see that $V(\Lambda_1)$ is isomorphic to the irreducible covariant tensor module V_λ with $\lambda = (2,1)$, while $V(\Lambda_2)$ is ismorphic to the irreducible contravariant tensor module $V_{\bar{\lambda}}$ with $\lambda = (2)$. These observations, together with (5.7) and (5.8), do indeed serve to confirm the validity of (4.26).

Although we have not been able to prove the fact, we believe that the character ch $V(\Lambda)$ of the irreducible module $V(\Lambda)$ of sl(3/4) of Proposition 4.3, with $\Lambda = [11;0;010]$, is actually defined by the $\mathrm{sl}(3/4) \to \mathrm{sl}(3) \oplus \mathbb{C} \oplus \mathrm{sl}(4)$ decomposition:

$$[11;0;010]\to[11]\times[0]\times[010]+[01]\times[1]\times[110]+[01]\times[0]\times[001]$$
$$+[20]\times[1]\times[110]+[20]\times[0]\times[001]+[12]\times[-1]\times[001]$$
$$+[10]\times[2]\times[210]+[10]\times[1]\times[020]+2[10]\times[1]\times[101]$$
$$+[10]\times[0]\times[000]+[02]\times[0]\times[101]+[21]\times[0]\times[101]$$
$$+[21]\times[-1]\times[000]+[02]\times[-1]\times[000]+[13]\times[-2]\times[000]$$
$$+[00]\times[1]\times[011]+[00]\times[2]\times[120]+[00]\times[2]\times[201]$$
$$+[00]\times[1]\times[100]+[11]\times[1]\times[201]+[11]\times[0]\times[011]$$
$$+2[11]\times[0]\times[100]+[03]\times[-1]\times[100]+[22]\times[-1]\times[100]$$
$$+[01]\times[1]\times[111]+[01]\times[0]\times[010]+[01]\times[1]\times[200]$$
$$+[12]\times[0]\times[200]+[12]\times[-1]\times[010]+[02]\times[0]\times[110]. \tag{5.26}$$

The level structure is

$$1\ 5\ 10\ 10\ 5\ 1, \tag{5.27}$$

and the lowest $\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$-highest weight is $\Sigma_L=[02;0;110]$.

It is, we believe, no accident that although this decomposition does not coincide with that specified by $\chi_S(\Lambda)$, which has level structure

$$1\ 3\ 5\ 5\ 3\ 1, \tag{5.28}$$

nonetheless the lowest $\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$-highest weight specified by $\chi_S(\Lambda)$ does coincide with Σ_L. Indeed, on the basis of our computer calculations, we are led to the following.

Conjecture 5.29: Let $V(\Lambda)$ be a finite-dimensional irreducible module of $\mathrm{sl}(m/n)$ whose highest weight Λ is integral dominant. Then the lowest $\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$-highest weight, Σ_L, appearing in the decomposition of $V(\Lambda)$ on restriction from $\mathrm{sl}(m/n)$ to $\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$ is given by the lowest $\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$-highest weight of the Weyl characters appearing in the expansion of $\chi_S(\Lambda)$.

The importance of this observation is that, if true, it would enable the highest weight of the module contragredient to $V(\Lambda)$ to be found from $\chi_S(\Lambda)$. This is because the highest weight of the contragredient module $V(\bar{\Lambda})$ is given by $\bar{\Lambda}=-w_0\Sigma_L$, where w_0 is the Coxeter element of the Weyl group $W=S_m\times S_n$. In our example the lowest $\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$-highest weight of $V(\Lambda)$ is $\Sigma_L=[02;0;110]$ and the highest weight of the contragredient irreducible module, $V(\bar{\Lambda})$, is given by $\bar{\Lambda}=[20;0;011]$.

We have also observed that in all the cases we have examined it is possible to write

$$\mathrm{ch}\,V(\Lambda)=\sum_{\lambda}q_\lambda\chi_S(\lambda), \tag{5.30}$$

where the summation is over integral dominant weights λ, and $q_\lambda\in\mathbb{N}$ for all such λ, with $q_\lambda\geqslant 0$ and $q_\Lambda=1$. In our example we believe that

$$\mathrm{ch}\,V([11;0;010])$$
$$=\chi_S([11;0;010])+\chi_S([01;0;001])$$
$$+\chi_S([20;0;001])+\chi_S([10;0;000]), \tag{5.31}$$

which agrees with (5.26).

VI. SINGLY ATYPICAL REPRESENTATIONS

In the rest of this paper we shall be dealing with a different kind of character formula. The new character formula will not be of the Kac–Weyl type (3.19), i.e., it can usually not be generated by a generating set $\Delta_\chi(\Lambda)$. On the other hand, one can think of the new character formula as being an expansion of ch $V(\Lambda)$ in terms of the Kac characters $\chi_K(\lambda)$ of (3.18). We shall introduce the new formula step by step: singly atypical representations in this section, doubly atypical in Sec. VII, and atypical of arbitrary degree in Sec. VIII, where the formula is shown to cover all irreducible covariant tensor representations. In this section we state a crucial proposition (Proposition 6.8), which we have proved elsewhere,[44] upon which we now base a proof of the validity of the Leites character formula for all singly atypical irreducible modules of $\mathrm{sl}(m/n)$. This proof involves an expansion in terms of Kac characters, showing the way to a generalization made in the next section.

For two integral weights λ and μ of $\mathrm{sl}(m/n)$ we say that λ is *W equivalent* to μ and write $\lambda\overset{w}{\cong}\mu$ if there exists an element $w\in W$ such that $w\cdot\lambda=\mu$; that is $w(\lambda+\rho)=\mu+\rho$ or, equivalently, by virtue of (2.31), $w(\lambda+\rho_0)=\mu+\rho_0$. In such a case $\chi_K(\lambda)=\epsilon(w)\chi_K(\mu)$ and $\chi_W(\lambda)=\epsilon(w)\chi_W(\mu)$, as can be seen from the definitions (3.18) and (3.20) of the formal Kac and Weyl characters, respectively. Moreover, λ is W equivalent to μ if and only if the atypicality matrix A_μ may be obtained from A_λ by suitable permutations of rows and columns. This follows from the definition (4.1) of A_λ and properties (4.2a) and (4.2d), which imply that $\mu=w\cdot\lambda$ if and only if $A_\mu=A_{w\cdot\lambda}=w^{-1}(A_\lambda)$.

If there exists $w\in W$ such that $w\cdot\lambda=\lambda$ with $\epsilon(w)=-1$, then both $\chi_K(\lambda)=0$ and $\chi_W(\lambda)=0$, and we say that λ is *vanishing*. Otherwise, λ is said to be *nonvanishing*, and both $\chi_K(\lambda)$ and $\chi_W(\lambda)$ are nonzero, as can be seen from the relation $\chi_K(\lambda)=e^{\rho_1}L_1\chi_W(\lambda)$, the fact that $L_1\neq 0$, and the well-known properties of the Weyl characters. Furthermore, these properties also ensure that λ is nonvanishing if and only if there exists $w\in W$ such that $w\cdot\lambda$ is dominant. It then follows from (4.1) and (4.2) that λ is

vanishing if and only if A_λ has two identical rows or two identical columns. To see this it should be noted first that, if A_λ has either a pair of identical rows or a pair of identical columns, then there exists $w \in W$ with $\epsilon(w) = -1$ such that $A_\lambda = A_{w\cdot\lambda}$. Hence $\lambda = w\cdot\lambda$ and λ is vanishing. Conversely, if A_λ has no such pair of identical rows or identical columns then there exist $w \in W$ such that the elements of $A_{w\cdot\lambda}$ are strictly decreasing from left to right across rows and from top to bottom down columns. It then follows from the definition (4.3c) that $w\cdot\lambda$ is dominant and therefore λ is nonvanishing.

It should further be recalled that the odd roots $\beta \in \Delta_1^+$ may be partially ordered according to (2.26):

$$\beta_{ia} \leqslant \beta_{jb} \Leftrightarrow i \geqslant j \quad \text{and} \quad a \leqslant b, \tag{6.1a}$$

$$\beta_{ia} < \beta_{jb} \Leftrightarrow \beta_{ia} \leqslant \beta_{jb} \quad \text{and} \quad \beta_{ia} \neq \beta_{jb}. \tag{6.1b}$$

In particular, $\beta_{m1} < \beta$ for every $\beta \in \{\Delta_1^+ \setminus \beta_{m1}\}$.

In this section, from now on, we let Λ be an integral dominant weight that is singly atypical of type β:

$$\langle \Lambda + \rho | \beta \rangle = 0 \text{ and } \langle \Lambda + \rho | \gamma \rangle \neq 0 \text{ for } \gamma \neq \beta (\gamma, \beta \in \Delta_1^+). \tag{6.2}$$

Then we can prove the following lemmas.

Lemma 6.3: Let A be the atypicality matrix of Λ, where Λ is singly atypical of type $\beta = \beta_{ia}$. Then

$$\{A_{ib} | 1 \leqslant b \leqslant a\} \cap \{-A_{ja} | i \leqslant j \leqslant m\} = \{0\}. \tag{6.3}$$

Proof: From (4.2c) and the fact that $A_{ia} = 0$, one deduces that $A_{ib} + A_{ja} = A_{jb}$. But Λ is singly atypical of type β_{ia}, so that $A_{jb} = 0$ if and only if $j = i$ and $b = a$. The result (6.3) then follows. □

By way of example, consider the weight $\Lambda = [1023;1;13020]$ of sl(5/6). The corresponding atypicality matrix is

$$A_\Lambda = \begin{pmatrix} 11 & 9 & 5 & 4 & 1 & 0 \\ 9 & 7 & 3 & 2 & -1 & -2 \\ 8 & 6 & 2 & 1 & -2 & -3 \\ 5 & 3 & -1 & -2 & -5 & -6 \\ 1 & -1 & -5 & -6 & -9 & -10 \end{pmatrix} \tag{6.4}$$

and Λ is singly atypical of type $\beta = \beta_{16}$. Lemma 6.3 is equivalent to the statement that the set of numbers in the same row but to the left of the single zero in the atypicality matrix has no element in common with the set formed by taking the negative of the entries in the same column but below the zero. In our example we have $\{11,9,5,4,1,0\} \cap \{0,2,3,6,10\} = \{0\}$, in conformity with Lemma 6.3.

Let Λ be given as in Lemma 6.3. Let k be the length of the longest sequence of consecutive integers $(0,1,2,\ldots,k-1)$ all contained in the union $\{A_{ib} | 1 \leqslant b \leqslant a\} \cup \{-A_{ja} | i \leqslant j \leqslant m\}$. In the above example the union is $\{0,1,2,3,4,5,6,9,10,11\}$, hence the value of k is 7.

Lemma 6.5: There exists a unique sequence of distinct elements $\beta_1 = \beta, \beta_2, \beta_3, \ldots, \beta_k$ from Δ_1^+ such that the chain $\nu_0 = \Lambda$, $\nu_1 = \Lambda - \beta_1$, $\nu_2 = \nu_1 - \beta_2$, $\ldots, \nu_k = \nu_{k-1} - \beta_k \equiv \Phi$ satisfies $\langle \nu_j + \rho | \beta_j \rangle = 0$ for $i = 1,2,\ldots,k$, with ν_j vanishing for $1 \leqslant j \leqslant k$ and $\nu_k = \Phi$ dominant. Moreover, $\nu_j \overset{w}{\cong} \Lambda - j\beta$ and $\epsilon(w) = (-1)^{j-1}$ for $1 \leqslant j \leqslant k$.

Proof: First a special case: if $\beta_1 = \beta = \beta_{m1}$ then $\Lambda - \beta_1$ is always dominant, hence $k = 1$ with $\Phi = \Lambda - \beta_{m1}$. More generally, suppose now that Λ is singly atypical of type $\beta_1 = \beta_{ia} > \beta_{m1}$. Since $\langle \beta_{ia} | \beta_{jb} \rangle = \delta_{ij} - \delta_{ab}$, the definition (4.1) implies that $A_{\Lambda - \beta_{ia}}$ is obtained from A_Λ by decreasing the elements of row(i) by one and increasing the elements of the column(a) by one, leaving the entry 0 at their intersection. If $\nu_1 = \Lambda - \beta_1 = \Lambda - \beta_{ia}$ is dominant, then it is still singly atypical of type β_{ia}, and $k = 1$ with $\Phi = \Lambda - \beta_{ia}$. Otherwise, if Λ is dominant but $\nu_1 = \Lambda - \beta_{ia}$ is nondominant, $A_{\Lambda - \beta_{ia}}$ must have two equal rows or columns: either row$(i+1)$ = row(i) or else column $(a-1)$ = column(a). Both cannot occur since this would imply that $A(\Lambda - \beta_{ia})_{i+1,a-1} = A(\Lambda)_{i+1,a-1} = 0$, which is not possible because Λ is singly atypical of type β_{ia}. Hence $\nu_1 = \Lambda - \beta_{ia}$ is vanishing and doubly atypical of type β_1 and β_2, with $\beta_1 = \beta_{ia}$ and either $\beta_2 = \beta_{i+1,a}$ or $\beta_2 = \beta_{i,a-1}$. Moreover, because of the identical pair of rows or columns in $A_{\Lambda - \beta_{ia}}$, $A_{\Lambda - \beta_{ia} - \beta_2}$ may be obtained from $A_{\Lambda - 2\beta_{ia}}$ by one transposition either of row(i) and row$(i+1)$, or of column(a) and column$(a-1)$. Thus $\nu_2 = \nu_1 - \beta_2 = \Lambda - \beta_1 - \beta_2$ satisfies $\nu_2 \overset{w}{\cong} \Lambda - 2\beta_1$ with $\epsilon(w) = -1$.

It is clear that the same reasoning now applies to $\Lambda - \beta_1 - \beta_2$, etc. Indeed, the atypicality matrix A_{ν_j} is obtained from $A_{\nu_{j-1}}$ by subtracting 1 in the row of β_j and adding 1 in the column of β_j; this produces a new zero in the position of β_{j+1}. This continues as long as the differences between the relevant rows or columns is 1, i.e., this continues for k steps. The atypicality matrix A_{ν_j} is obtained by reordering the rows and columns of $A_{\Lambda - j\beta_{ia}}$, and one can check that there are $j-1$ transpositions involved in going from $A_{\Lambda - j\beta_{ia}}$ to A_{ν_j}. Hence, $\nu_j \overset{w}{\cong} \Lambda - j\beta_{ia}$ and $\epsilon(w) = (-1)^{j-1}$ for $1 \leqslant j \leqslant k$. Because of the definition of k and the method of construction, every $A_{\Lambda - j\beta_{ia}}$ has either two equal columns or two equal rows for $1 \leqslant j < k$. The same must then be true of A_{ν_j}, and ν_j is therefore vanishing for $1 \leqslant j < k$. Moreover, the construction and definition of k implies that $\Lambda - k\beta_{ia}$ is the first nonvanishing element in the sequence $\Lambda - j\beta_{ia}$ with $j = 1,2,\ldots$. It then follows that ν_k, which is W equivalent to $\Lambda - k\beta_{ia}$, is also nonvanishing. This guarantees that no two rows or two columns of $A_{\Lambda - k\beta_{ia}}$ are identical. However, the construction procedure ensures that the matrix elements of A_{ν_j} are nonincreasing across rows from left to right and down columns from top to bottom for $1 \leqslant j \leqslant k$. Hence the matrix elements of A_{ν_k} must be strictly decreasing across rows from left to right and down columns from top to bottom, so that $\nu_k = \Phi$ is dominant. □

By way of illustration, in the case $\Lambda = [1023;1;13020] = (87752|-1-2-5-5-7-7)$ of sl(5/6) we have:

$$A_\Lambda = \begin{pmatrix} 11 & 9 & 5 & 4 & 1 & 0 \\ 9 & 7 & 3 & 2 & -1 & -2 \\ 8 & 6 & 2 & 1 & -2 & -3 \\ 5 & 3 & -1 & -2 & -5 & -6 \\ 1 & -1 & -5 & -6 & -9 & -10 \end{pmatrix}, \quad A_{\nu_1} = \begin{pmatrix} 10 & 8 & 4 & 3 & 0 & 0 \\ 9 & 7 & 3 & 2 & -1 & -1 \\ 8 & 6 & 2 & 1 & -2 & -2 \\ 5 & 3 & -1 & -2 & -5 & -5 \\ 1 & -1 & -5 & -6 & -9 & -9 \end{pmatrix},$$

$$A_{\nu_2} = \begin{pmatrix} 9 & 7 & 3 & 2 & 0 & -1 \\ 9 & 7 & 3 & 2 & 0 & -1 \\ 8 & 6 & 2 & 1 & -1 & -2 \\ 5 & 3 & -1 & -2 & -4 & -5 \\ 1 & -1 & -5 & -6 & -8 & -9 \end{pmatrix}, \quad A_{\nu_3} = \begin{pmatrix} 9 & 7 & 3 & 2 & 1 & -1 \\ 8 & 6 & 2 & 1 & 0 & -2 \\ 8 & 6 & 2 & 1 & 0 & -2 \\ 5 & 3 & -1 & -2 & -3 & -5 \\ 1 & -1 & -5 & -6 & -7 & -9 \end{pmatrix},$$

$$A_{\nu_4} = \begin{pmatrix} 9 & 7 & 3 & 2 & 2 & -1 \\ 8 & 6 & 2 & 1 & 1 & -2 \\ 7 & 5 & 1 & 0 & 0 & -3 \\ 5 & 3 & -1 & -2 & -2 & -5 \\ 1 & -1 & -5 & -6 & -6 & -9 \end{pmatrix}, \quad A_{\nu_5} = \begin{pmatrix} 9 & 7 & 3 & 3 & 2 & -1 \\ 8 & 6 & 2 & 2 & 1 & -2 \\ 6 & 4 & 0 & 0 & -1 & -4 \\ 5 & 3 & -1 & -1 & -2 & -5 \\ 1 & -1 & -5 & -5 & -6 & -9 \end{pmatrix},$$

$$A_{\nu_6} = \begin{pmatrix} 9 & 7 & 4 & 3 & 2 & -1 \\ 8 & 6 & 3 & 2 & 1 & -2 \\ 5 & 3 & 0 & -1 & -2 & -5 \\ 5 & 3 & 0 & -1 & -2 & -5 \\ 1 & -1 & -4 & -5 & -6 & -9 \end{pmatrix}, \quad A_{\nu_7} = \begin{pmatrix} 9 & 7 & 5 & 3 & 2 & -1 \\ 8 & 6 & 4 & 2 & 1 & -2 \\ 5 & 3 & 1 & -1 & -2 & -5 \\ 4 & 2 & 0 & -2 & -3 & -6 \\ 1 & -1 & -3 & -5 & -6 & -9 \end{pmatrix}.$$

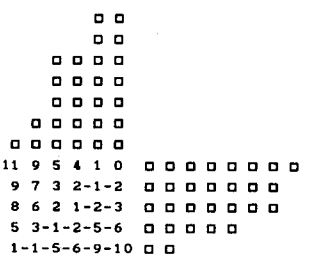
Note that every A_{ν_j} is obtained from $A_{\Lambda - j\beta_{16}}$ by transpositions of rows and columns. Also note that the sequence of β's described in Lemma 6.5 can be found from A_Λ by considering the integers on the row through 0, and the negatives of the integers on the column through 0, and then following a path in A_Λ of horizontal and vertical steps starting at 0 and following the directions dictated by consecutive integers. In the example (6.4) this sequence is (0,1,2,3,4,5,6), and we can augment the atypicality matrix, as shown below, in order to identify the corresponding sequence of β's, signified on the right by x's:

$$A_\Lambda = \begin{pmatrix} & & 5 & 4 & 1 & \\ 11 & 9 & 5 & 4 & 1 & 0 \\ 9 & 7 & 3 & 2 & -1 & -2 \\ 8 & 6 & 2 & 1 & -2 & -3 \\ 5 & 3 & -1 & -2 & -5 & -6 \\ 1 & -1 & -5 & -6 & -9 & -10 \end{pmatrix} \begin{matrix} \\ 2 \\ 3 \\ 6 \\ \\ \end{matrix} \qquad \begin{pmatrix} 0 & 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & 0 & x & 0 \\ 0 & 0 & x & x & x & 0 \\ 0 & 0 & x & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{6.5}$$

In this example the length of the sequence is given by $k = 7$, and we have $A_{\nu_7} = A_\Phi$, with

$$\begin{aligned} \Phi &= \Lambda - (21310|00 - 2 - 1 - 3 - 1) \\ &= (66442| - 1 - 2 - 3 - 4 - 4 - 6) \\ &= [0202;1;11102]. \end{aligned}$$

It should be pointed out that this procedure for obtaining Φ from Λ may be described, as proposed by Hurni,[34] in terms of Young diagrams. To each $\Lambda = [a_1 a_2 \cdots a_{m-1}; a_m; a_{m+1} \cdots a_{m+n-1}]$ we can associate a composite Young diagram by writing Λ in the $\epsilon\delta$ basis as $\Lambda = (\sigma_1 \sigma_2 \cdots \sigma_m | - \tau'_n \cdots - \tau'_2 - \tau'_1)$ with $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_m$ and $\tau'_1 \geqslant \tau'_2 \geqslant \cdots \geqslant \tau'_n$. Without loss of generality we may choose both σ_m and τ'_n to be positive, with $a_m = \sigma_m - \tau'_n$ in accordance with (2.21). This allows us to associate with Λ the composite Young diagram[25] $F^{\bar{\tau},\sigma}$ formed from F^σ and $F^{\bar{\tau}}$, where $\sigma_1, \sigma_2, \ldots, \sigma_m$ are the row lengths of F^σ, and $\tau'_1, \tau'_2, \ldots, \tau'_n$ are the column lengths of $F^{\bar{\tau}}$. In the case of our sl(5/6) example, we have $\Lambda = [1023;1;13020] = (87752| - 1 - 2 - 5 - 5 - 7 - 7)$, so that the associated composite Young diagram takes the form:



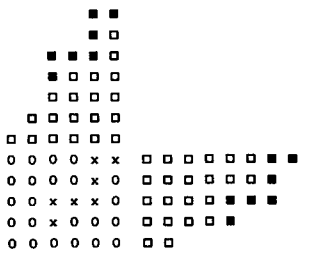
(6.6a)

where the atypicality matrix has also been displayed.

The heuristic method of Hurni[34] may then be interpreted more precisely in the singly atypical case of type $\beta = \beta_{ia}$ as follows. It consists of identifying the ith row and $(n - a + 1)$th column of F^σ and $F^{\bar{\tau}}$, respectively, which contain the single zero of the atypicality matrix, and then deleting from F^σ and $F^{\bar{\tau}}$ two continuous boundary strips, starting at the ends of the ith row and $(n - a + 1)$th column, and continuing until the resulting composite Young diagram is once again regular. In such a case the row and column lengths of the undeleted boxes define a new weight Φ in the $\epsilon\delta$ basis, which may then be rewritten in terms of Kac–Dynkin labels. That this weight Φ coincides with the weight

Φ of Lemma 6.5 can be seen by noting that the deleted continuous boundary strips are necessarily of the same shape as each other and as the path in the atypicality matrix corresponding to the chain defined in Lemma 6.5.

This is illustrated as follows in our example (6.6a) by



(6.6b)

where ■ signifies a deleted box and, as in (6.6a), x specifies an element of the sequence of β's defined in Lemma 6.5. The undeleted boxes signified by □ imply that $\Phi = (66442|-1-2-3-4-4-6)$ as before.

We have seen that the Kac module $\overline{V}(\Lambda)$ is irreducible when Λ is typical. In the present case of a singly atypical dominant weight Λ, $\overline{V}(\Lambda)$ is reducible and contains a unique maximal proper submodule M, and $V(\Lambda) = \overline{V}(\Lambda)/M$ is irreducible. Let us recall the following definition.

Definition 6.7: A vector v_λ, of weight λ, of a G module V is called a *primitive vector* if v_λ is annihilated by $\mathfrak{n}^+ : \mathfrak{n}^+ v_\lambda = 0$.

When Λ is typical, then $\overline{V}(\Lambda) = V(\Lambda)$ contains no primitive vectors other than scalar multiples of v_Λ.

The following proposition[44] gives an explicit description of the primitive vectors of the Kac module $\overline{V}(\Lambda)$ if Λ is singly atypical of type β.

Proposition 6.8: Let Λ be an integral dominant weight of $\mathrm{sl}(m/n)$ which is singly atypical of type β, and let Φ be the weight defined in Lemma 6.5. Then

(i) $\overline{V}(\Lambda)$ contains a primitive weight vector v_Φ ($\neq c v_\Lambda$) for which $\mathrm{U}(G)v_\Phi$ is irreducible and isomorphic to $V(\Phi)$;

(ii) $\mathrm{U}(G)v_\Phi$ is the maximal submodule M of $\overline{V}(\Lambda)$ and hence $\overline{V}(\Lambda)/\mathrm{U}(G)v_\Phi$ is irreducible.

Part (i) may be proved by the explicit construction of a primitive vector $v_\Phi \in U(G)v_\Sigma$, where $v_\Sigma = \Pi_{\beta\in\Delta_1^+} e(-\beta)v_\Lambda$ is the lowest $G_{\bar{0}}$ highest weight of $\overline{V}(\Lambda)$. The irreducibility of $\mathrm{U}(G)v_\Sigma$ has recently been proved by Gould.[36] It is the identification of Φ as the highest weight of this module which is new. The proof of part (ii) then follows from a rather different argument involving the fact that v_Λ is an eigenvector of the operator $\Pi_{\beta\in\Delta_1^+} e(+\beta)\Pi_{\beta\in\Delta_1^+} e(-\beta)$, whose eigenvalue $Q(\Lambda)$[12] possesses a zero of multiplicity one, when viewed as a polynomial in the odd Kac–Dynkin label a_m, if Λ is singly atypical. The details of this argument are presented elsewhere.[44]

An immediate consequence of the proposition is the following corollary.

Corollary 6.9: With the same notation as in Proposition 6.8

$$\mathrm{ch}\,\overline{V}(\Lambda) = \mathrm{ch}\,V(\Lambda) + \mathrm{ch}\,V(\Phi). \tag{6.9}$$

This gives a precise statement of the irreducible content of the Kac module in the singly atypical case: the second summand is the character of the maximal submodule of the Kac module and the first that of the corresponding quotient module. Both these modules are irreducible in this singly atypical case, and the result serves to justify Hurni's procedure[34] in such cases.

We now consider the implications of Proposition 6.8 and its Corollary 6.9, firstly for the case of singly atypical representations of type β_{m1}, where β_{m1} is the unique *simple* odd root. Let $\Lambda = (\mu_1,\mu_2,\ldots,\mu_m|\nu_1,\nu_2,\ldots,\nu_n)$ be integral dominant (i.e., $\mu_i - \mu_{i+1} \in \mathbb{N}$ for $1 \leqslant i < m$, $\nu_a - \nu_{a+1} \in \mathbb{N}$ for $1 \leqslant a < n$ and $\mu_m + \nu_1 \in \mathbb{Z}$). Then $\Lambda - \beta_{m1} = (\mu_1,\mu_2,\ldots,\mu_m - 1|\nu_1 + 1,\nu_2,\ldots,\nu_n)$ is also integral dominant, with $\langle \Lambda - \beta_{m1} + \rho | \beta_{m1} \rangle = \langle \Lambda + \rho | \beta_{m1} \rangle = 0$. Hence $\Phi = \Lambda - \beta_{m1}$, from Lemma 6.5, and (6.9) then implies that

$$\mathrm{ch}\,V(\Lambda) = \chi_K(\Lambda) - \mathrm{ch}\,V(\Lambda - \beta_{m1}). \tag{6.10}$$

But $\Lambda - \beta_{m1}$ is again integral dominant and singly atypical of type β_{m1}, hence we can use (6.10) as a recursion relation to find:

$$\begin{aligned}
&\mathrm{ch}\,V(\Lambda)\\
&\quad = \chi_K(\Lambda) - (\chi_K(\Lambda - \beta_{m1}) - \mathrm{ch}\,V(\Lambda - 2\beta_{m1}))\\
&\quad = \chi_K(\Lambda) - \chi_K(\Lambda - \beta_{m1}) + (\chi_K(\Lambda - 2\beta_{m1})\\
&\qquad - \mathrm{ch}\,V(\Lambda - 3\beta_{m1})) = \cdots = \chi_K(\Lambda)\\
&\qquad - \chi_K(\Lambda - \beta_{m1}) + \chi_K(\Lambda - 2\beta_{m1})\\
&\qquad - \chi_K(\Lambda - 3\beta_{m1}) + \cdots,
\end{aligned} \tag{6.11}$$

which becomes a formal infinite series since (6.10) can be applied for every $V(\Lambda - t\beta_{m1})$ with $t \in \mathbb{N}$. Therefore, from (3.15) and (3.18):

$$\begin{aligned}
&\mathrm{ch}\,V(\Lambda)\\
&\quad = \frac{L_1}{L_0} \sum_{w\in W} \epsilon(w) w\left(e^{\Lambda+\rho} - e^{\Lambda-\beta_{m1}+\rho} + e^{\Lambda-2\beta_{m1}+\rho}\right.\\
&\qquad \left. - e^{\Lambda-3\beta_{m1}+\rho}\cdots\right)\\
&\quad = L_0^{-1} \sum_{w\in W} \epsilon(w) w\left(e^{\Lambda+\rho} L_1 (1 + e^{-\beta_{m1}})^{-1}\right)\\
&\quad = \chi_L(\Lambda),
\end{aligned}$$

where $\chi_L(\Lambda)$ has been defined by (3.22). Hence, we are led to the following corollary.

Corollary 6.12: If Λ is integral dominant and singly atypical of type β_{m1} then

$$\mathrm{ch}\,V(\Lambda) = \chi_L(\Lambda). \tag{6.12}$$

Let us now go back to establishing from Proposition 6.8 a character formula for $V(\Lambda)$ when Λ is singly atypical of any type β. Remembering that $\Delta_L(\Lambda) = \{\beta \in \Delta_1^+ \,|\, \langle \Lambda + \rho | \beta \rangle \neq 0\}$ as in (3.21), it is convenient to use the notation $\chi_\beta(\lambda)$ for $\chi_L(\lambda)$. This serves to emphasize the fact that the factor $(1 + e^{-\beta})$ is missing in the character formula for $\chi_L(\lambda)$ as compared with $\chi_K(\lambda)$.

From the definition of $\chi_K(\Lambda)$ one finds:

$$\mathrm{ch}\,\overline{V}(\Lambda) = \chi_K(\Lambda) = \chi_\beta(\Lambda) + \chi_\beta(\Lambda - \beta), \tag{6.13}$$

but from the expansion of the term $(1 + e^{-\beta})^{-1}$ one finds

$$\begin{aligned}
&\chi_\beta(\Lambda - \beta)\\
&\quad = \chi_K(\Lambda - \beta) - \chi_K(\Lambda - 2\beta) + \cdots + (-1)^{k-1}\\
&\qquad \times \chi_K(\Lambda - k\beta) + \cdots.
\end{aligned} \tag{6.14}$$

Due to the vanishing terms for $\Lambda-\beta,\ldots,\Lambda-(k-1)\beta$ implied by Lemma 6.5, this reduces to

$$\begin{aligned}\chi_\beta(\Lambda-\beta)&=(-1)^{k-1}[\chi_K(\Lambda-k\beta)-\chi_K(\Lambda-(k+1)\beta)\\&\quad+\chi_K(\Lambda-(k+2)\beta)\cdots]\\&=(-1)^{k-1}\chi_\beta(\Lambda-k\beta).\end{aligned}\tag{6.15}$$

But $\Lambda-k\beta\overset{w}{\cong}\Phi$ where Φ is singly atypical of type $w(\beta)$ with $\epsilon(w)=(-1)^{k-1}$. Thus

$$\chi_L(\Phi)=\chi_{w(\beta)}(\Phi)=(-1)^{k-1}\chi_\beta(\Lambda-k\beta),\tag{6.16}$$

and therefore

$$\operatorname{ch}\overline{V}(\Lambda)=\chi_L(\Lambda)+\chi_L(\Phi).\tag{6.17}$$

Now we shall describe a construction that finally links Λ, which is singly atypical of some arbitrary type β, with some dominant weight singly atypical of type β_{m1}. For the given dominant weight Λ we apply Lemma 6.5 to obtain the dominant weight Φ; let $\Phi_0=\Lambda$ and $\Phi_1=\Phi$, and let $\beta_0=\beta$ and $\beta_1=w(\beta)$ be the odd roots with respect to which Φ_0 and Φ_1 are atypical. From the construction procedure $\beta_1\leqslant\beta_0$, with equality only if $\Phi_0-\beta_0$ is dominant.

Then we apply Lemma 6.5 to Φ_1, leading to Φ_2 and β_2 etc. So we build a sequence of dominant weights $\Phi_0,\Phi_1,\Phi_2,\ldots$ and associated roots $\beta_0,\beta_1,\beta_2,\ldots$ such that Φ_i is singly atypical of type β_i and Φ_{i+1} is the weight of the primitive vector of $\overline{V}(\Phi_i)$ that generates the maximal submodule.

Then, according to (6.17):

$$\operatorname{ch}\overline{V}(\Phi_i)=\chi_L(\Phi_i)+\chi_L(\Phi_{i+1}).\tag{6.18}$$

Also $\beta_0\geqslant\beta_1\geqslant\beta_2\geqslant\cdots$ But if $\beta_i\neq\beta_{m1}$ then the set $\Phi_i-t\beta_i$ for which $\Phi_i-t\beta_i$ is dominant is finite (this follows easily by considering the components of the weights in the $\epsilon\delta$ basis). Therefore, the subsequences of equal elements in the sequence $\beta_0\geqslant\beta_1\geqslant\beta_2\geqslant\cdots$ are finite. As a consequence, after a number of steps we necessarily end up with the lowest element of Δ_1^+, namely β_{m1}. At this point we stop our sequence: we now have the dominant weights $\Phi_0,\Phi_1,\Phi_2,\ldots,\Phi_s$ and associated roots $\beta_0\geqslant\beta_1\geqslant\beta_2\geqslant\cdots>\beta_s=\beta_{m1}$ (and $\beta_{s-1}\neq\beta_{m1}$), and (6.18) is valid at every step. Also (6.9) is true at every step. Hence,

$$\begin{aligned}&\operatorname{ch}\overline{V}(\Phi_i)=\operatorname{ch}V(\Phi_i)+\operatorname{ch}V(\Phi_{i+1}),\\&\operatorname{ch}\overline{V}(\Phi_i)=\chi_L(\Phi_i)+\chi_L(\Phi_{i+1})\\&\qquad(i=0,1,\ldots,s-1).\end{aligned}\tag{6.19}$$

But for Φ_s, which is atypical of type $\beta_s=\beta_{m1}$, we can apply Corollary 6.12, giving $\operatorname{ch}V(\Phi_s)=\chi_L(\Phi_s)$. Then the set of Eqs. (6.19) imply $\operatorname{ch}V(\Phi_{s-1})=\chi_L(\Phi_{s-1})$, and one can systematically proceed backwards in the sequence. Hence,

$$\operatorname{ch}V(\Phi_i)=\chi_L(\Phi_i),\quad\text{for } i=s-1,\ s-2,\ldots,1,0.\tag{6.20}$$

This establishes our final corollary of Proposition 6.8.

Corollary 6.21: If Λ is singly atypical then

$$\operatorname{ch}V(\Lambda)=\chi_L(\Lambda).\tag{6.21}$$

This asserts that the Bernstein–Leites character formula,[30] (3.22), is valid for all singly atypical representations of $\mathrm{sl}(m/n)$.

Finally, we shall show how the formula (6.21) can be rewritten in terms of characters of induced modules. Let Λ be singly atypical of type β, as in (6.3), and let w_β be an element of W such that $\beta=w_\beta(\beta_{m1})$. Then we find

$$\begin{aligned}\langle w_\beta^{-1}\cdot\Lambda|\beta_{m1}\rangle&=\langle w_\beta^{-1}(\Lambda+\rho)-\rho|\beta_{m1}\rangle\\&=\langle w_\beta^{-1}(\Lambda+\rho)|\beta_{m1}\rangle\\&=\langle\Lambda+\rho|w_\beta(\beta_{m1})\rangle=\langle\Lambda+\rho|\beta\rangle=0.\end{aligned}\tag{6.22}$$

Now we define the parabolic subalgebra

$$P=B\oplus\mathbb{C}e(-\beta_{m1}),\tag{6.23}$$

where B is the Borel subalgebra defined in (2.17). Note that P is a solvable subalgebra of G, since $[e(-\beta_{m1}),B]\subseteq B$. Let $\lambda=w_\beta^{-1}\cdot\Lambda$, with $\langle\lambda|\beta_{m1}\rangle=0$ as shown in (6.22), and define a one-dimensional B module, $\mathbb{C}v_\lambda$, by

$$\mathfrak{n}^+v_\lambda=0,\quad hv_\lambda=\lambda(h)v_\lambda\quad\forall h\in\mathfrak{h}.\tag{6.24}$$

Since $[e(+\beta_{m1}),e(-\beta_{m1})]=h_{\beta_{m1}}$, we obtain, using (2.18) and (6.24):

$$\begin{aligned}e(+\beta_{m1})e(-\beta_{m1})v_\lambda&=h_{\beta_{m1}}v_\lambda=\lambda(h_{\beta_{m1}})v_\lambda\\&=\langle\lambda|\beta_{m1}\rangle v_\lambda=0.\end{aligned}\tag{6.25}$$

Therefore, $\mathbb{C}v_\lambda$ can be naturally extended to become a one-dimensional P module by putting $e(-\beta_{m1})v_\lambda=0$. Consequently we can define the induced module

$$X=\operatorname{Ind}_P^G\mathbb{C}v_\lambda,\tag{6.26}$$

whose character is given by

$$\begin{aligned}\operatorname{ch}X&=e^\lambda\frac{\prod_{\gamma\in\Delta_1^+}(1+e^{-\gamma})}{\prod_{\alpha\in\Delta_0^+}(1-e^{-\alpha})}(1+e^{-\beta_{m1}})^{-1}\\&=\frac{L_1}{L_0}e^{\lambda+\rho}\ (1+e^{-\omega\beta_{m1}})^{-1}.\end{aligned}\tag{6.27}$$

Since $w(L_1)=L_1$, $w(L_0)=\epsilon(w)L_0$ and $w(\lambda+\rho)=ww_\beta^{-1}(\Lambda+\rho)$, we then find

$$\sum_{w\in W}w(\operatorname{ch}X)=\frac{L_1}{L_0}\sum_{w\in W}\epsilon(w)e^{ww_\beta^{-1}(\Lambda+\rho)}/(1+e^{-w\beta_{m1}}).\tag{6.28}$$

Setting $w'=ww_\beta^{-1}$, we have $w\beta_{m1}=w'w_\beta\beta_{m1}=w'\beta$, and (6.28) becomes

$$\begin{aligned}&\sum_{w\in W}w(\operatorname{ch}X)\\&\quad=\frac{L_1}{L_0}\sum_{w'\in W}\epsilon(w'w_\beta)e^{w'(\Lambda+\rho)}/(1+e^{-w'\beta})\\&\quad=\epsilon(w_\beta)\chi_L(\Lambda).\end{aligned}\tag{6.29}$$

Thus we have shown that $\chi_L(\Lambda)$ can be formally written as a Weyl average of the character $\operatorname{ch}X$ of an induced module.

Unfortunately, this approach has not enabled us to give a proof of (6.21). Note that trying to extend B with $e(-\beta)$

($\beta \neq \beta_{m1}$) in (6.23) would not give rise to a subalgebra. In particular, when trying to extend this approach to multiply atypical Λ, the most obvious parabolic subalgebra to consider might not exist.

VII. DOUBLY ATYPICAL REPRESENTATIONS

In considering singly atypical representations in the last section we eventually arrived at a familiar enough Kac–Weyl type character formula; in fact the Leites character formula. That we did so via an infinite expansion (6.11) in terms of Kac characters was apparently incidental. Rather than proceeding directly to the general case we prefer first to confine our attention to doubly atypical representations in this section. This will serve to illustrate the structure of our new formula and the crucial notion of *truncation* in a relatively simple context.

Throughout this section, therefore, it is to be assumed that the integral dominant weight Λ of $\mathrm{sl}(m/n)$ is doubly atypical of type (β_1,β_2) that is,

$$\langle \Lambda+\rho|\beta_1\rangle = \langle \Lambda+\rho|\beta_2\rangle = 0$$

and

$$\langle \Lambda+\rho|\beta\rangle \neq 0, \quad \text{if } \beta \neq \beta_1,\beta_2. \tag{7.1}$$

Without loss of generality we let $\beta_1=\beta_{jb}$ and $\beta_2=\beta_{ia}$ with $1\leqslant j<i\leqslant m$ and $1\leqslant a<b\leqslant n$, so that $\beta_1>\beta_2$. By virtue of (4.2c) and (4.3c) it follows that the atypicality matrix takes the form:

$$\begin{array}{c} \\ j \\ \\ i \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} & a & & b & \\ \end{array} \\ \begin{pmatrix} \cdots & x & \cdots & 0 & \cdots \\ & \vdots & & \vdots & \\ \cdots & 0 & \cdots & -x & \cdots \end{pmatrix} \end{array}, \tag{7.2}$$

where x is a strictly positive integer subject to the following constraint.

Lemma 7.3: With the notation of (7.2),

$$x \geqslant i-j+b-a-1. \tag{7.3}$$

Proof: Consider the part (A_{kt}) of the atypicality matrix A_Λ with $j\leqslant k\leqslant i$ and $a\leqslant t\leqslant b$, and $A_{kt}=(A_\Lambda)_{kt}$:

$$A_\Lambda = \begin{pmatrix} & \vdots & & & \vdots & \\ \cdots & x=A_{ja} & A_{j,a+1} & \cdots A_{j,b-1} & 0 & \cdots \\ & & & & A_{j+1,b} & \\ & \vdots & & & \vdots & \\ & & & & A_{i-1,b} & \\ \cdots & 0 & & \cdots & -x=A_{i,b} & \cdots \\ & \vdots & & & \vdots & \end{pmatrix}. \tag{7.4}$$

From the proof of Lemma 6.3 we deduce that

$$\{A_{jc}|a+1\leqslant c\leqslant b\}\cap\{-A_{kb}|j\leqslant k\leqslant i-1\}=\{0\}, \tag{7.5a}$$

and also

$$\{A_{jc}|a+1\leqslant c\leqslant b\}\cup\{-A_{kb}|j\leqslant k\leqslant i-1\} \subseteq \{0,1,2,\ldots,x-1\}. \tag{7.5b}$$

The cardinalities of the sets in (7.5) then lead to

$$(b-a)+(i-j)-1\leqslant x. \qquad \square$$

In the light of this lemma we make the following definition.

Definition 7.6: Λ is critical if $x=i-j+b-a-1$, otherwise it is *noncritical.*

It is to be noted that the critical value for x is equal to the *hook length* of the path connecting the two zeros in the atypicality matrix (where the zeros themselves are to be disregarded in the path):

$$\begin{array}{c} \\ j \\ \\ i \end{array}\!\!\begin{array}{c} \begin{array}{ccc} a & & b \end{array} \\ \begin{pmatrix} \vdots & & \vdots \\ \cdots\ x & \text{———} & 0\ \cdots \\ \quad | & & \\ \cdots\ 0 & & -x\ \cdots \\ \vdots & & \vdots \end{pmatrix} \end{array}. \tag{7.7}$$

Also note that it follows from the proof of Lemma 7.3 that in the case of a critical Λ, the elements of the atypicality matrix satisfy:

$$\{A_{jc}|a\leqslant c\leqslant b\}\cup\{-A_{kb}|j\leqslant k\leqslant i\} = \{0,1,2,\ldots,x\}, \tag{7.8a}$$

with

$$\{A_{jc}|a\leqslant c\leqslant b\}\cap\{-A_{kb}|j\leqslant k\leqslant i\}=\{0,x\}. \tag{7.8b}$$

Given $\beta_1=\beta_{jb}$ and $\beta_2=\beta_{ia}$ with $\beta_1>\beta_2$, they determine a unique element w_{12} of the Weyl group W with the following action on any $\lambda\in\mathfrak{h}^*$ where $\lambda=(\mu_1,\mu_2,\ldots,\mu_m|\nu_1,\nu_2,\ldots,\nu_n)$:

$$w_{12}(\ldots,\mu_j,\ldots,\mu_i,\ldots|\ldots,\nu_a,\ldots,\nu_b,\ldots) = (\ldots,\mu_i,\ldots,\mu_j,\ldots|\ldots,\nu_b,\ldots,\nu_a,\ldots), \tag{7.9}$$

i.e., μ_i and μ_j, and ν_a and ν_b are transposed and all the other components of λ are left invariant. Note that $\epsilon(w_{12})=+1$, but nevertheless if $w_{12}\cdot\lambda=\lambda$ then λ is vanishing (this is because w_{12} is the product of two commuting Weyl elements σ and τ in S_m and S_n, respectively, with $\epsilon(\sigma)=\epsilon(\tau)=-1$, such that $\sigma\cdot\lambda=\lambda$ and $\tau\cdot\lambda=\lambda$). We denote the hyperplane in $\mathfrak{h}^*$ which is invariant under the dot action of w_{12} by h_{12}:

$$h_{12}=\{\alpha\in\mathfrak{h}^*|w_{12}\cdot\alpha=\alpha\}. \tag{7.10}$$

Let Λ be doubly atypical of type $(\beta_1,\beta_2)=(\beta_{jb},\beta_{ia})$, as in (7.1) and (7.2). Then (4.1c) implies that

$$\mu_i+\nu_a+m-i-a+1=\mu_j+\nu_b+m-j-b+1=0, \tag{7.11a}$$

and

$$\mu_j + \nu_a + m - j - a + 1$$
$$= -(\mu_i + \nu_b + m - i - b + 1) = x. \tag{7.11b}$$

Hence

$$x = \mu_j - \mu_i - j + i = \nu_a - \nu_b - a + b, \tag{7.11c}$$

so that

$$\Lambda - x\beta_1$$
$$= (\cdots, \underset{(j)}{\mu_i + j - i}, \ldots, \underset{(i)}{\mu_i}, \ldots | \ldots, \underset{(a)}{\nu_a}, \ldots, \underset{(b)}{\nu_a + b - a}, \ldots). \tag{7.12}$$

It then follows from (2.29)–(2.31) that $w_{12} \cdot (\Lambda - x\beta_1) = w_{12}(\Lambda - x\beta_1 + \rho) - \rho = \Lambda - x\beta_1$. In other words, the string $\Lambda - t\beta_1$ ($t \in \mathbb{N}$) intersects the hyperplane h_{12} in $\Lambda - x\beta_1$. Similarly, one can show that the string $\Lambda + t\beta_2$ ($t \in \mathbb{N}$) intersects the hyperplane h_{12} in $\Lambda + x\beta_2$.

It follows that if Λ is doubly atypical of type (β_1, β_2) then the weights $\Lambda - x\beta_1$ and $\Lambda + x\beta_2$ are both vanishing. In the critical case one can further deduce the following lemma.

Lemma 7.13: Let Λ be doubly atypical and critical. Then every $\Lambda + t\beta_2 - s\beta_1$ with either $s \in \{1,2,\ldots,x-1\}$ and $t \in \mathbb{Z}$, or $t \in \{1,2,\ldots,x-1\}$ and $s \in \mathbb{Z}$ is vanishing.

Proof: We first consider elements of the form $\Lambda - s\beta_1$ with $s \in \{1,2,\ldots,x-1\}$. Remember that the atypicality matrix $A_{\Lambda - s\beta_1}$ is obtained from A_Λ by subtracting s from the numbers in the jth row and adding s to the numbers in the b th column. But from (7.8) it follows that $A_{\Lambda - s\beta_1}$ has either two equal rows or else two equal columns provided $s \in \{1,2,\ldots,x-1\}$ (for $s = x$, one would have two equal rows and two equal columns). And, as shown in Sec. VI, λ is vanishing if and only if A_λ has two equal rows or columns. This proves the statement for $s \in \{1,2,\ldots,x-1\}$ and $t = 0$. But if $A_{\Lambda - s\beta_1}$ has two equal rows, say row(j) = row(k) with $j < k < i$ then $A_{\Lambda - s\beta_1 + t\beta_2}$ also has these two rows equal, since $A_{\Lambda - s\beta_1 + t\beta_2}$ is obtained from $A_{\Lambda - s\beta_1}$ by adding t in row i and subtracting t in column a. A similar result applies if $A_{\Lambda - s\beta_1}$ has two equal columns. This proves the $s \in \{1,2,\ldots,x-1\}$ and $t \in \mathbb{Z}$ case, and the $t \in \{1,2,\ldots,x-1\}$ and $s \in \mathbb{Z}$ case has a similar proof. □

Further consideration of (7.4)–(7.8) along the same lines leads to an alternative prescription for criticality.

Lemma 7.14: Let Λ be doubly atypical of type (β_1, β_2) then Λ is critical if and only if $\Lambda - s\beta_1$ is vanishing for all $s \in \{1,2,\ldots,x-1\}$ or, equivalently, if and only if $\Lambda + t\beta_2$ is vanishing for all $t \in \{1,2,\ldots,x-1\}$, where x is such that $\Lambda - x\beta_1$ and $\Lambda + x\beta_2$ lie on the hyperplane h_{12} defined by (7.10).

If we consider the plane containing the lattice of weights $\lambda = \Lambda - k_1\beta_1 - k_2\beta_2$, with $k_1, k_2 \in \mathbb{Z}$, the previous lemmas may be illustrated by identifying lines of vanishing weights. This is done in the case of sl(3/4) for $\Lambda = [03;0;002]$, which is noncritical, and $\Lambda = [11;0;010]$, which is critical, in Figs. 1 and 2, respectively. The second example corresponds to the case discussed at length in Sec. IV, where it was used to rule out all formulae of the Kac–Weyl type.

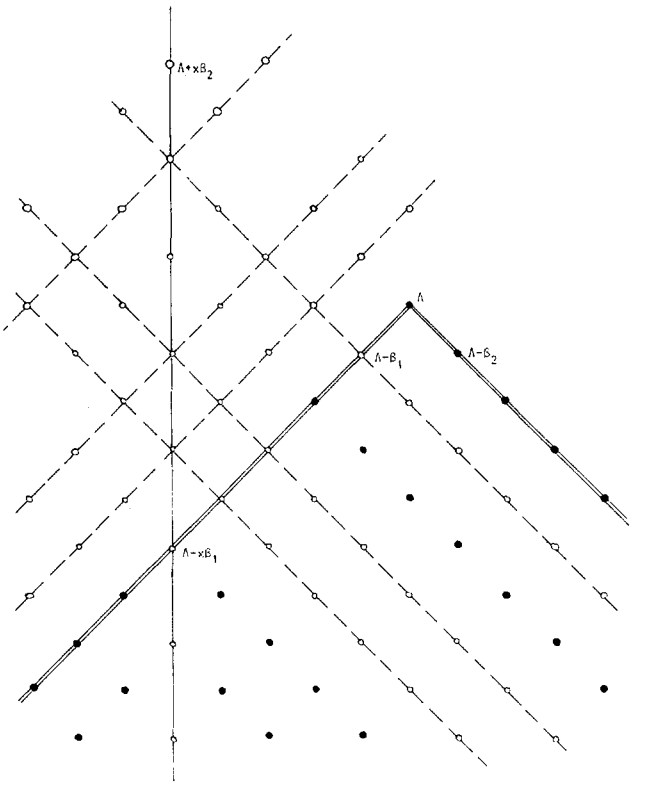


FIG. 1. The lattice of weights $\lambda \in \mathscr{C}_\Lambda$ contributing to the character $\chi_T(\Lambda) = \Sigma_{\lambda \in \mathscr{C}_\Lambda} (-1)^{|\Lambda - \lambda|} \chi_K(\lambda)$ in the noncritical, doubly atypical case $\Lambda = [03;0;002]$ of sl(3/4), for which

$$A_\Lambda = \begin{bmatrix} 5 & 4 & 3 & 0 \\ 4 & 3 & 2 & -1 \\ 0 & -1 & -2 & -5 \end{bmatrix},$$

so that $\beta_1 = \beta_{14}, \beta_2 = \beta_{31}$ and $x = h + 1 = 5$. The nonvanishing, contributing weights are indicated by ●, the nonvanishing but noncontributing weights by ·, and the vanishing weights by O. The boundary of the cone $\mathscr{C}_\Lambda$ is indicated by =, the hyperplane h_{12} by —O——O—, and the lines of vanishing weight by —O— — —O—.

We denote by $\mathscr{L}_\Lambda$ the following set of lattice points:

$$\mathscr{L}_\Lambda = \{\Lambda + k_1\beta_1 + k_2\beta_2 \,|\, k_1, k_2 \in \mathbb{Z}\}, \tag{7.15}$$

and denote by $\mathscr{C}_\Lambda$ the cone in $\mathscr{L}_\Lambda$ with vertex at Λ:

$$\mathscr{C}_\Lambda = \{\Lambda - k_1\beta_1 - k_2\beta_2 \,|\, k_1, k_2 \in \mathbb{N}\}. \tag{7.16}$$

From (3.18) and (3.22) it follows, in this doubly atypical case, that the Leites formula can be written as follows:

$$\chi_L(\Lambda)$$
$$= \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} (-1)^{k_1+k_2} \chi_K(\Lambda - k_1\beta_1 - k_2\beta_2)$$
$$= \sum_{\lambda \in \mathscr{C}_\Lambda} (-1)^{|\Lambda - \lambda|} \chi_K(\lambda), \tag{7.17}$$

where $(-1)^{|\Lambda - \lambda|}$ is defined for points λ of the lattice $\mathscr{L}_\Lambda$ by $(-1)^{|\Lambda - \lambda|} = (-1)^{k_1+k_2}$ for $\lambda = \Lambda - k_1\beta_1 - k_2\beta_2$. This can be interpreted as a formal, infinite, expansion of the Leites formula in terms of Kac characters, χ_K.

The cone $\mathscr{C}_\Lambda$ defined in (7.16) has an intersection with the hyperplane h_{12}, and we further define the *truncated cone*

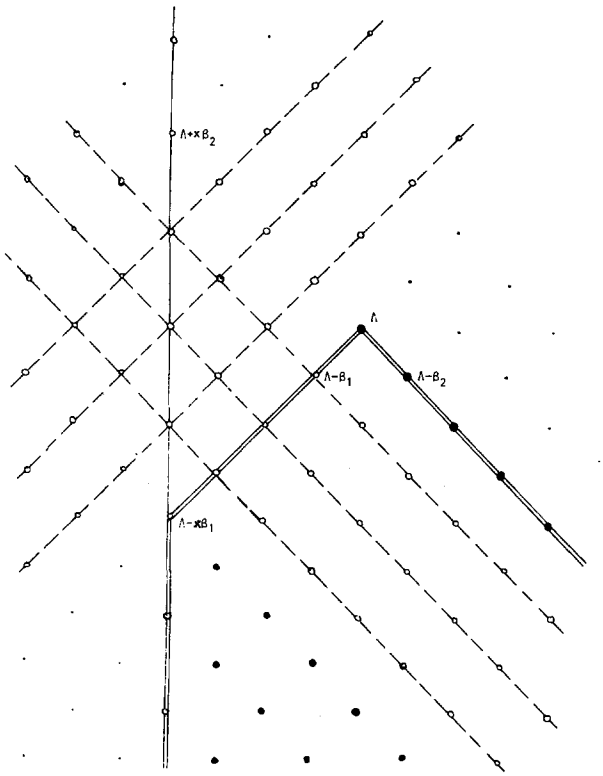


FIG. 2. The lattice of weights $\lambda \in \mathscr{C}_\Lambda^+$ contributing to the character $\chi_T(\Lambda) = \Sigma_{\lambda \in \mathscr{C}_\Lambda^+}(-1)^{|\Lambda-\lambda|}\chi_K(\lambda)$ in the critical, doubly atypical case $\Lambda = [11;0;010]$ of sl(3/4), for which

$$A_\Lambda = \begin{bmatrix} 4 & 3 & 1 & 0 \\ 2 & 1 & -1 & -2 \\ 0 & -1 & -3 & -4 \end{bmatrix},$$

so that $\beta_1 = \beta_{14}, \beta_2 = \beta_{31}$ and $x = h = 4$. The nonvanishing, contributing weights are indicated by ●, the nonvanishing but noncontributing weights by ·, and the vanishing weights by ○. The boundary of the truncated cone $\mathscr{C}_\Lambda^+$ is indicated by =, the hyperplane h_{12} by —○——○—, and the lines of vanishing weight by —○— — —○—.

$\mathscr{C}_\Lambda^+$ to be the subset of the set of weights of $\mathscr{C}_\Lambda$ that are on the same side of the hyperplane h_{12} as Λ (the weights of the intersection itself are to be excluded). If we denote by $\mathfrak{h}_{12}^+$ the half-space of $\mathfrak{h}^*$ that is on the same side of the hyperplane h_{12} as Λ, then

$$\mathscr{C}_\Lambda^+ = \mathscr{C}_\Lambda \cap \mathfrak{h}_{12}^+. \tag{7.18}$$

These notions are also illustrated in Figs. 1 and 2.

Finally, we introduce our new character formula.

Definition 7.19: If Λ is doubly atypical of type (β_1, β_2), then in the notation of (3.18), (3.22), (7.16), and (7.18):

$$\chi_T(\Lambda) = \begin{cases} \sum_{\lambda \in \mathscr{C}_\Lambda} (-1)^{|\Lambda-\lambda|}\chi_K(\lambda) = \chi_L(\Lambda), \\ \qquad \text{if } \Lambda \text{ is noncritical}, \\ \sum_{\lambda \in \mathscr{C}_\Lambda^+} (-1)^{|\Lambda-\lambda|}\chi_K(\lambda), \\ \qquad \text{if } \Lambda \text{ is critical}. \end{cases} \tag{7.19}$$

In other words, the expansion of $\chi_T(\Lambda)$ in terms of $\chi_K(\lambda)$ runs over the complete cone $\mathscr{C}_\Lambda$ if Λ is noncritical, and over the truncated cone $\mathscr{C}_\Lambda^+$ if Λ is critical.

For our two examples the nonvanishing weights contributing to the summations in (7.19) have been signified by full circles in Figs. 1 and 2. In the case of the second example the restriction of contributions to the truncated cone leads to the result (5.26) quoted in Sec. V as the $\mathrm{sl}(3) \oplus \mathbb{C} \oplus \mathrm{sl}(4)$ decomposition of the irreducible module $V(\Lambda)$ of sl(3/4) with $\Lambda = [11;0;010]$.

It should be stressed that our new character formula $\chi_T(\Lambda)$ is nothing other than the Leites formula $\chi_L(\Lambda)$ if Λ is doubly atypical and noncritical, as in the case of our first example.

We shall now establish some equivalent expressions for $\chi_T(\Lambda)$ when Λ is doubly atypical and critical. In the notation of (3.21), $\Delta_L(\Lambda) = \Delta_1^+ \setminus \{\beta_1, \beta_2\}$ and correspondingly $\chi_L(\Lambda)$ is defined by (3.19) with $X = L$. It is convenient to generalize the notation slightly so that

$$\chi_{L(\Lambda)}(\lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\lambda+\rho_0} \prod_{\beta \in \Delta_L(\Lambda)} (1 + e^{-\beta}) \right\}, \tag{7.20}$$

where the use of λ and Λ is to be noted, as well as the special case $\chi_{L(\Lambda)}(\Lambda) = \chi_L(\Lambda)$.

Lemma 7.21: Let Λ be doubly atypical and critical, then

$$\text{(i)} \quad \chi_T(\Lambda) = \chi_{L(\Lambda)}(\Lambda) + \tfrac{1}{2}\chi_{L(\Lambda)}(\Lambda - \beta_1) \tag{7.21a}$$

and

$$\text{(ii)} \quad \chi_T(\Lambda) = \tfrac{1}{2}(-1)^x \chi_{L(\Lambda)}(\Lambda + x\beta_2). \tag{7.21b}$$

Proof: Because of the lines with vanishing weights (see Lemma 7.13) it follows that

$$\begin{aligned} \chi_{L(\Lambda)}(\Lambda) = &\sum_{k=0}^{\infty} (-1)^k \chi_K(\Lambda - k\beta_2) \\ &+ (-1)^x \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} (-1)^{k_1+k_2} \\ &\times \chi_K(\Lambda - x\beta_2 - k_1\beta_1 - k_2\beta_2). \end{aligned} \tag{7.22}$$

The second summand in (7.22) can be rewritten as

$$\begin{aligned} &(-1)^x \sum_{\lambda \in \mathscr{C}_{\Lambda - x\beta_1}} (-1)^{|\Lambda - x\beta_1 - \lambda|} \chi_K(\lambda) \\ &\quad = 2(-1)^x \sum_{\lambda \in \mathscr{C}_{\Lambda - x\beta_1}^+} (-1)^{|\Lambda - x\beta_1 - \lambda|} \chi_K(\lambda), \end{aligned} \tag{7.23}$$

since h_{12} splits the cone $\mathscr{C}_{\Lambda - x\beta_1}$ into two parts $\mathscr{C}_{\Lambda - x\beta_1}^+$ and $\mathscr{C}_{\Lambda - x\beta_1}^-$ such that $\mathscr{C}_{\Lambda - x\beta_1}^- = w_{12} \cdot \mathscr{C}_{\Lambda - x\beta_1}^+$, and χ_K vanishes when evaluated at points of $h_{12} \cap \mathscr{C}_{\Lambda - x\beta_1}$. But again because of the lines of vanishing weights given by Lemma 7.13, we have,

$$\begin{aligned} \chi_T(\Lambda) = &\sum_{k=0}^{\infty} (-1)^k \chi_K(\Lambda - k\beta_2) \\ &+ \sum_{\lambda \in \mathscr{C}_{\Lambda - x\beta_1}^+} (-1)^{|\Lambda - \lambda|} \chi_K(\lambda), \end{aligned} \tag{7.24}$$

$$\begin{aligned} &\chi_{L(\Lambda)}(\lambda - \beta_1) \\ &\quad = (-1)^{x-1} \sum_{\lambda \in \mathscr{C}_{\Lambda - x\beta_2}} (-1)^{|\Lambda - x\beta_1 - \lambda|} \chi_K(\lambda). \end{aligned} \tag{7.25}$$

Combining (7.22)–(7.25) leads to (7.21a).

To prove (7.21b), note that there is the same symmetry for $\Lambda + x\beta_2$ as for $\Lambda - x\beta_1$, i.e.,

$$\chi_{L(\Lambda)}(\lambda + x\beta_2) = 2 \sum_{\lambda \in \mathcal{C}^+_{\Lambda + x\beta_1}} (-1)^{|\lambda + x\beta_2 - \lambda|} \chi_K(\lambda). \tag{7.26}$$

But, due to the vanishing weights described by Lemma 7.13, it is not difficult to see that

$$(-1)^x \sum_{\lambda \in \mathcal{C}^+_{\lambda + x\beta_2}} (-1)^{|\Lambda + x\beta_2 - \lambda|} \chi_K(\lambda) = \chi_T(\lambda), \tag{7.27}$$

which implies (7.21b). □

The character $\chi_X(\Lambda)$ was defined in (4.9) for a matrix $G_X(\Lambda)$ with entries 0 or 1. This restriction on the entries may be relaxed, and $G_X(\Lambda)$ replaced by an arbitrary matrix. In particular, for the doubly atypical and critical case under consideration, we may replace $G_X(\Lambda)$ by the matrix $M = (M_{kc})$ $(1 \leqslant k \leqslant m;\ 1 \leqslant c \leqslant n)$, where

$$M_{kc} = \begin{cases} 0, & \text{if } (k,c) = (i,a) \\ \frac{1}{2}, & \text{if } (k,c) = (j,b) \\ 1, & \text{otherwise.} \end{cases} \tag{7.28}$$

Denoting the corresponding characters by $\chi_M(\Lambda)$ we have the following lemma.

Lemma 7.29: Let Λ be doubly atypical and critical. Then

$$\chi_T(\Lambda) = \chi_M(\Lambda). \tag{7.29}$$

Proof: Using the explicit expression (7.20) for $\chi_{L(\Lambda)}(\lambda)$, (7.21b) gives

$$\chi_T(\Lambda) = L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\Lambda + \rho_0} \prod_{\beta \in \Delta_L(\Lambda)} (1 + e^{-\beta}) \right\}$$
$$+ \tfrac{1}{2} L_0^{-1} \sum_{w \in W} \epsilon(w)$$
$$\times w \left\{ e^{\Lambda - \beta_1 + \rho_0} \prod_{\beta \in \Delta_L(\Lambda)} (1 + e^{-\beta}) \right\}. \tag{7.30}$$

But the right-hand side of (7.30) is equal to

$$L_0^{-1} \sum_{w \in W} \epsilon(w) w \left\{ e^{\Lambda + \rho_0} \left(1 + \frac{1}{2} e^{-\beta_1}\right) \prod_{\beta \in \Delta_L(\Lambda)} (1 + e^{-\beta}) \right\},$$

and this is exactly $\chi_M(\Lambda)$ with $G_M(\Lambda)$ given by (7.28). □

Note that, although the new character $\chi_T(\Lambda)$ is defined as a formal infinite expansion of Kac characters, $\chi_K(\lambda)$, both Lemma 7.21 and Lemma 7.29 show that $\chi_T(\Lambda)$ is a finite sum of Weyl characters $\chi_W(\lambda)$ if Λ is critical. The same is true if Λ is noncritical by virtue of the definition (7.19).

In the next section we shall define $\chi_T(\Lambda)$ more generally for multiply atypical representations, and conjecture that ch $V(\Lambda) = \chi_T(\Lambda)$ for all irreducible modules of sl(m/n). Suffice to say at this stage that for all the doubly atypical cases we have examined, including that of the key counterexample of Sec. IV, the character of the irreducible module appears to be given correctly by

Conjecture 7.31: The character of the irreducible module $V(\Lambda)$ of sl(m/n), with Λ doubly atypical, is given by

$$\text{ch } V(\Lambda) = \chi_T(\Lambda), \tag{7.31}$$

with $\chi_T(\Lambda)$ defined in (7.19).

VIII. MULTIPLY ATYPICAL REPRESENTATIONS

Let Λ be an integral dominant weight of sl(m/n) that is multiply atypical of type $(\beta_1,\beta_2,\ldots,\beta_N)$, with $\beta_1 > \beta_2 > \cdots > \beta_N$, i.e.,

$$\langle \Lambda + \rho | \beta_i \rangle = 0, \quad \text{for } i \in \{1,2,\ldots,N\},$$
$$\{\beta_1,\beta_2,\ldots,\beta_N\} \subset \Delta_1^+,$$
$$\langle \Lambda + \rho | \beta \rangle \neq 0, \quad \text{for } \beta \in \Delta_1^+ \setminus \{\beta_1,\beta_2,\ldots,\beta_N\}, \tag{8.1}$$

We shall also assume that $N \geqslant 2$, since $N = 0$ corresponds to typical representations and the case $N = 1$ is covered by Corollary 6.21. The atypicality matrix A_Λ has the following form:

$$A_\Lambda = \begin{pmatrix} \cdots & x_{1N} & \cdots & x_{13} & \cdots & x_{12} & \cdots & 0 & \cdots \\ & \vdots & & \vdots & & \vdots & & & \\ \cdots & x_{2N} & \cdots & x_{23} & \cdots & 0 & \cdots & & \\ & \vdots & & \vdots & & & & & \\ \cdots & x_{3N} & \cdots & 0 & \cdots & & & & \\ & \vdots & & & & & & & \\ \cdots & 0 & \cdots & & & & & & \end{pmatrix}, \tag{8.2}$$

where the N zeros corresponding to $\beta_1,\beta_2,\ldots,\beta_N$ are ordered from the top right-hand corner of A_Λ to the bottom left-hand corner. We shall refer to these zeros as the first, second, $\cdots$, Nth zero, and call x_{ij} $(i<j)$ the integer in the corner of the hook connecting the ith and jth zero. Obviously, every x_{ij} is a positive integer, and from (4.2c)

$$x_{ij} = x_{ik} + x_{kj}, \quad i < k < j. \tag{8.3}$$

To every pair (β_i,β_j) with $i<j$ there corresponds a unique $w_{ij} \in W$, defined as before in (7.10), and a corresponding hyperplane h_{ij} defined by

$$h_{ij} = \{\alpha \in \mathfrak{h}^* : w_{ij} \cdot \alpha = \alpha\}. \tag{8.4}$$

With the given Λ, h_{ij} divides $\mathfrak{h}^*$ into two half-spaces and the one containing Λ is called $\mathfrak{h}^+_{ij}$. Just as in (7.15) and (7.16) we define the lattice $\mathscr{L}_\Lambda$ and the cone $\mathscr{C}_\Lambda$:

$$\mathscr{L}_\Lambda = \{\Lambda + \sum_{i=1}^N k_i \beta_i | k_i \in \mathbb{Z}\}, \tag{8.5}$$

$$\mathscr{C}_\Lambda = \{\Lambda - \sum_{i=1}^N k_i \beta_i | k_i \in \mathbb{N}\}. \tag{8.6}$$

For $\lambda \in \mathscr{L}_\Lambda$ we define

$$|\Lambda - \lambda| = \sum_{i=1}^N k_i, \quad \text{if } \lambda = \Lambda - \sum_{i=1}^N k_i \beta_i. \tag{8.7}$$

Note that the Leites character $\chi_L(\Lambda)$ has the following expansion:

$$\chi_L(\Lambda) = \sum_{\lambda \in \mathscr{C}_\Lambda} (-1)^{|\Lambda - \lambda|} \chi_K(\lambda). \tag{8.8}$$

Just as in Sec. VII, we shall now show how to truncate the cone $\mathscr{C}_\Lambda$ in order to define the new character formula.

For any two neighboring zeros, i and $i+1$, in the atypicality matrix, we can apply Lemma 7.3 as we did in the doubly atypical case, and establish that the element $x_{i,i+1}$ is always greater than or equal to the hook length connecting zero i and zero $i+1$. In the multiply atypical case, if $j>i+1$, x_{ij} can actually be less than the hook length of the path connecting zero i with zero j.

Definition 8.9: We say that the element x_{ij} is *critical* if

(i) $x_{ij} \leqslant$ hook length of the path connecting zero i and j; and

(ii) either $i=j-1$, or $i<j-1$ and $x_{i+1,j}$ is also critical.

Note that this definition is recursive in i for each value of j: the criticality of $x_{j-1,j}$ is well defined by (i), and then we have to use (i) and (ii) in order to determine successively the criticality of $x_{j-2,j}, x_{j-3,j},\ldots, x_{ij}$.

We are now in a position to define the truncated cone $\mathscr{C}_\Lambda^+$.

Definition 8.10: If Λ is multiply atypical as in (8.1), then with the notation of (8.2) and Definition 8.9, $\mathscr{C}_\Lambda^+ = \mathscr{C}_\Lambda \cap (\cap_{ij} \mathfrak{h}_{ij}^+)$, with (i,j) such that x_{ij} is critical.

Notice that this definition replaces that appropriate to the doubly atypical case, (7.18), thereby enabling us to introduce our new character formula by means of the following definition.

Definition 8.11: In the notation of Definition 8.10 and (3.18)

$$\chi_T(\Lambda) = \sum_{\lambda \in \mathscr{C}_\Lambda^+} (-1)^{|\Lambda-\lambda|} \chi_K(\lambda). \tag{8.11}$$

As a consequence, if none of the elements x_{ij} are critical, the summation in (8.11) is over the complete cone $\mathscr{C}_\Lambda$ and $\chi_T(\Lambda)$ is simply equal to the Leites character $\chi_L(\Lambda)$.

To turn to the critical case: in the extreme situation for which every x_{ij} in (8.2) is critical there exists an analog of (7.21b). Remember that the proof of (7.21b) made essential use of the facts that there were lines of vanishing weights, and that there was a complete symmetry when cutting the cone, with vertex at $\Lambda + x\beta_2$, into two pieces. In the more general case this complete symmetry is still present provided that every $x_{i,i+1}$ in (8.2) is critical. Indeed, one can verify that in such a case the N-dimensional cone, with vertex at

$$\Lambda_t = \Lambda + x_{12}\beta_2 + x_{13}\beta_3 + \cdots + x_{1N}\beta_N, \tag{8.12}$$

splits into $N!$ pieces, all of which give rise to the same contribution when the sum over $(-1)^{|\Lambda-\lambda|}\chi_K(\lambda)$ is taken, and one of which is precisely $\mathscr{C}_\Lambda^+$. The argument again involves an analysis of the vanishing weights, and a "geometrical" argument describing the position of the truncated cone as one of the $N!$ pieces of a larger cone with vertex at Λ_t. Hence, if Λ is atypical of degree N and totally critical, in the sense that every $x_{i,i+1}$ for $1 \leqslant i < N$ is critical, and hence also, every x_{ij} in (8.2) is critical, the analog of (7.21b) takes the form:

$$\chi_T(\Lambda) = (1/N!)(-1)^{|\Lambda_t-\Lambda|}\chi_{L(\Lambda)}(\Lambda_t). \tag{8.13}$$

An examination of both critical and noncritical cases has led us to the following generalization of Conjecture 7.31.

Conjecture 8.14: The character of the irreducible module, $V(\Lambda)$, of $\mathrm{sl}(m/n)$ having highest weight Λ is given by

$$\mathrm{ch}\, V(\Lambda) = \chi_T(\Lambda) = \sum_{\lambda \in \mathscr{C}_\Lambda^+} (-1)^{|\Lambda-\lambda|}\chi_K(\lambda), \tag{8.14}$$

where the summation over λ is carried out over the truncated cone $\mathscr{C}_\Lambda^+$ defined in (8.10), and $\chi_K(\lambda)$ is the Kac character (3.18).

So far we have not been able to prove (8.14) in general, but we now present some evidence in favor of its validity.

One surprisingly interesting case is that of the trivial one-dimensional representation with highest weight $\Lambda = 0$. For $\mathrm{sl}(m/n)$, with $m \leqslant n$, the corresponding atypicality matrix is given by

$$A_0 = \begin{pmatrix} m-1 & m-2 & m-3 & \cdots & m-n \\ \vdots & \vdots & \vdots & & \vdots \\ 2 & 1 & 0 & \cdots & 3-n \\ 1 & 0 & -1 & \cdots & 2-n \\ 0 & -1 & -2 & \cdots & 1-n \end{pmatrix}, \tag{8.15}$$

and $\beta_1 = \beta_{1m}$, $\beta_2 = \beta_{2,m-1}$, $\beta_3 = \beta_{3,m-3},\ldots, \beta_m = \beta_{m1}$. Also, every x_{ij} $(1 \leqslant i < j \leqslant m)$ is critical. Hence,

$$\begin{aligned}\mathscr{C}_0 &= \{-\sum_{i=1}^m k_i\beta_i : k_i \in \mathbf{N}\} \\ &= \{(-k_1, -k_2,\ldots, -k_m | k_m, k_{m-1}, \\ &\qquad \ldots, k_1, 0,\ldots,0) : k_i \in \mathbf{N}\}.\end{aligned} \tag{8.16}$$

But then the truncation (8.10) implies

$$\begin{aligned}\mathscr{C}_0^+ = \{&(-k_1, -k_2,\ldots, -k_m | k_m, k_{m-1},\ldots, k_1, 0,\ldots,0): \\ &k_i \in \mathbf{N} \text{ and } k_m \geqslant k_{m-1} \geqslant \cdots \geqslant k_1 \geqslant 0\}.\end{aligned} \tag{8.17}$$

Therefore, with the notation $\mathbf{k} = (k_m, k_{m-1},\ldots, k_1, 0,\ldots,0)$, $\mathbf{k}' = (-k_1, -k_2,\ldots, -k_m)$ and $|\mathbf{k}| = |k_1 + k_2 + \cdots + k_m|$, we obtain

$$\begin{aligned}\chi_T(0) &= \sum_{k_m=0}^{\infty} \sum_{k_{m-1}=0}^{k_m} \cdots \sum_{k_1=0}^{k_2} (-1)^{|\mathbf{k}|}\chi_k(\mathbf{k}'|\mathbf{k}) \\ &= \sum_{\mathbf{k}} (-1)^{|\mathbf{k}|}\chi_K(\mathbf{k}'|\mathbf{k}),\end{aligned} \tag{8.18}$$

where the summation is over all partitions $\mathbf{k}$ with $l(\mathbf{k}) \leqslant m$. Making use of the explicit form for $\chi_K(\lambda)$ in (3.12), this becomes

$$\begin{aligned}\chi_T(0) &= L_1 \sum_{\mathbf{k}} (-1)^{|\mathbf{k}|} L_0^{-1} \sum_{w \in W} \epsilon(w) e^{w(\mathbf{k}'|\mathbf{k})+\rho)} \\ &= \prod_{\beta \in \Delta_1^+} (1 + e^{-\beta}) \sum_{\mathbf{k}} (-1)^{|\mathbf{k}|}\chi_W(\mathbf{k}'|\mathbf{k}).\end{aligned} \tag{8.19}$$

where $\chi_W(\mathbf{k}'|\mathbf{k})$ is the Weyl character (3.20) of the irreducible $G_{\bar{0}}$ module $V^0(\mathbf{k}'|\mathbf{k})$. But this can be expressed in terms of S functions if we make the usual substitutions $x_i = e^{\epsilon_i}$ for $1 \leqslant i \leqslant m$ and $y_a = e^{\delta_a}$ for $1 \leqslant a \leqslant n$:

$$\begin{aligned}\chi_W(\mathbf{k}'|\mathbf{k}) &= \mathrm{ch}\, V^0(\mathbf{k}'|\mathbf{k}) = s_{\mathbf{k}'}(\mathbf{x}) s_{\mathbf{k}}(\mathbf{y}) \\ &= s_{\mathbf{k}}(\mathbf{x}^{-1}) s_{\mathbf{k}}(\mathbf{y}).\end{aligned} \tag{8.20}$$

Then, denoting the partition $\mathbf{k}=(k_m,...,k_1,0,...,0)$ by τ, (8.18) can be rewritten as

$$\chi_T(0)=\prod_{i=1}^{m}\prod_{a=1}^{n}(1+x_i^{-1}y_a)\sum_{\tau}(-1)^{|\tau|}s_\tau(\mathbf{x}^{-1})s_\tau(\mathbf{y}), \tag{8.21}$$

where in principle the summation is restricted to those partitions τ with $l(\tau)\leqslant m$, but may be extended to include *all* partitions since $s_\tau(x)=0$ if $l(\tau)>m$. Moreover,

$$\prod_{i=1}^{m}\prod_{a=1}^{n}(1+x_i^{-1}y_a)^{-1}=\sum_{\tau}(-1)^{|\tau|}s_\tau(\mathbf{x}^{-1})s_\tau(\mathbf{y}), \tag{8.22}$$

by virtue of Cauchy's classical identity,[40]

$$\prod_{i=1}^{m}\prod_{a=1}^{n}(1-u_iv_a)^{-1}=\sum_{\tau}s_\tau(\mathbf{u})s_\tau(\mathbf{v}). \tag{8.23}$$

Hence

$$\chi_T(0)=1=\mathrm{ch}\ V(0), \tag{8.24}$$

as required.

Turning to the more general case, described in Sec. V, of an arbitrary irreducible covariant tensor module $V(\Lambda_\sigma)$ labeled by the partition σ, the zeros of the atypicality matrix are at positions (i,b) specified by (5.13) and illustrated in (5.15). It follows that in the notation of (8.1) we have

$$(\beta_1,\beta_2,...,\beta_N)=(\cdots,\beta_{m-1,\sigma_{m-1}+2},\beta_{m,\sigma_m+1}). \tag{8.25}$$

Moreover, for $k=1,2,...,N-1$ we have

$$\begin{aligned}x_{N-k,N-k+1}&=A(\Lambda)_{m-k,\sigma_{m-k+1}+k}\\&=\sigma_{m-k}-\sigma_{m-k+1}+1,\end{aligned} \tag{8.26}$$

which is the hook length between neighboring zeros of $A(\Lambda)$. Hence $x_{i,i+1}$ is critical for all i. This implies that x_{ij} is critical for all i and j, so that Λ is totally critical.

From the definitions (8.11), (3.18), and (3.20) of $\chi_T(\Lambda)$, $\chi_K(\Lambda)$ and $\chi_W(\Lambda)$, respectively, it follows that

$$\chi_T(\Lambda)\prod_{\beta\in\Delta_1^+}(1+e^{-\beta})^{-1}=\sum_{\lambda\in\mathscr{C}_\Lambda^+}(-1)^{|\Lambda-\lambda|}\chi_W(\lambda). \tag{8.27}$$

Here $\lambda=\Lambda-\sum_{i=1}^{N}k_i\beta_i$, or more explicitly,

$$\begin{aligned}\lambda=(&...,\sigma_{m-1}-k_{m-1},\sigma_m-k_m|\nu_1,...,\nu_{\sigma_m},k_m,\\&0\cdots0,k_{m-1},0\cdots0,k_{m-2},...),\end{aligned} \tag{8.28}$$

where there are $\sigma_{i-1}-\sigma_i$ zeros between k_i and k_{i-1} in the second set of components of λ, for $i=m,\ m-1,...,\ m-N+1$. The restriction of λ to the truncated cone $\mathscr{C}_\Lambda^+$ in (8.27) may be expressed in terms of the parameters k_i appearing in (8.28) by making use of (7.18) and its generalizations. The fact that Λ is totally critical leads to a summation defined by

$$\sum_{\mathbf{k}}=\sum_{k_m=0}^{\infty}\sum_{k_{m-1}=0}^{k_m+(\sigma_{m-1}-\sigma_m)}\sum_{k_{m-2}=0}^{k_{m-1}+(\sigma_{m-1}-\sigma_m)}\cdots. \tag{8.29}$$

Since $G_{\bar{0}}=\mathrm{sl}(m)\oplus\mathbb{C}\oplus\mathrm{sl}(n)$, we can write

$$\chi_W(\lambda)=s_{(\sigma)_m-\mathbf{k}}(\mathbf{x})s_\eta(\mathbf{y}), \tag{8.30}$$

where, from (8.28),

$$(\sigma)_m-\mathbf{k}=(...,\sigma_{m-1}-k_{m-1},\sigma_m-k_m), \tag{8.31a}$$

$$\eta=(\nu_1,...,\nu_{\sigma_m},k_m,0,...,0,k_{m-1},0\cdots0,k_{m-2},...). \tag{8.31b}$$

With this notation we have the following lemma.

Lemma 8.32: Let

$$\begin{aligned}\chi_T(\Lambda)&\prod_{\beta\in\Delta_1^+}(1+e^{-\beta})^{-1}\\&=\sum_{\mathbf{k}}(-1)^{|\Lambda-\lambda|}s_{(\sigma)_m-\mathbf{k}}(\mathbf{x})s_\eta(\mathbf{y}).\end{aligned} \tag{8.32}$$

In principle the summation over $\mathbf{k}$ in (8.32) is a nested N-fold summation, where N is the degree of atypicality, as in (8.1). However, without loss of generality, the summation may be extended to an m-fold nested summation over $\mathbf{k}=(k_m,k_{m-1},...,k_1)$, since any additional terms included in this extension lead to S functions $s_\eta(\mathbf{y})$ that vanish identically by virtue of the fact that the corresponding sequence η contains positive elements beyond the nth position. It should perhaps be stressed at this point that in (8.30) and (8.32) we have again made use of the parametrizations:

$$x_i=e^{\epsilon_i}\quad(1\leqslant i\leqslant m)\ \text{ and }\ y_a=e^{\delta_a}\quad(1\leqslant a\leqslant n). \tag{8.33}$$

We shall now consider the function $s_\sigma(\mathbf{x}/\mathbf{y})$, given by (5.7b). Multiplication by the same factor as in (8.27) gives

$$\begin{aligned}s_\sigma(\mathbf{x}/\mathbf{y})&\prod_{\beta\in\Delta_1^+}(1+e^{-\beta})^{-1}\\&=\prod_{i,a}(1+x_i^{-1}y_a)^{-1}\sum_{\kappa}s_{\sigma/\kappa}(\mathbf{x})s_\kappa(\mathbf{y})\\&=\sum_{\tau}(-1)^{|\tau|}s_\tau(\mathbf{x}^{-1})s_\tau(\mathbf{y})\sum_{\kappa}s_{\sigma/\kappa}(\mathbf{x})s_\kappa(\mathbf{y}),\end{aligned} \tag{8.34}$$

by virtue of (8.22). Then the product rule (5.6a) for S functions gives:

$$\begin{aligned}&s_\tau(\mathbf{y})s_{\kappa'}(\mathbf{y})\\&\quad=\sum_{\nu}c_{\tau\kappa'}^{\nu}s_\nu(\mathbf{y}),\quad\text{with }|\nu|=|\tau|+|\kappa'|,\end{aligned} \tag{8.35}$$

whilst the quotient rule (5.6b) implies that

$$\sum_{\tau}c_{\tau\kappa'}^{\nu}s_\tau(\mathbf{x}^{-1})=s_{\nu/\kappa'}(\mathbf{x}^{-1}). \tag{8.36}$$

Using these identities in (8.34) gives

$$\begin{aligned}s_\sigma(\mathbf{x}/\mathbf{y})&\prod_{\beta\in\Delta_1^+}(1+e^{-\beta})^{-1}\\&=\sum_{\kappa,\nu}(-1)^{|\nu|+|\kappa'|}s_{\nu/\kappa'}(\mathbf{x}^{-1})s_{\sigma/\kappa}(\mathbf{x})s_\nu(\mathbf{y}),\end{aligned} \tag{8.37}$$

and finally by means of the definition[25]

$$s_{\bar{\nu},\sigma}(\mathbf{x})=\sum_{\kappa}(-1)^{|\kappa|}s_{\nu/\kappa'}(\mathbf{x}^{-1})s_{\sigma/\kappa}(\mathbf{x}), \tag{8.38}$$

we obtain the following lemma.

Lemma 8.39:

$$s_\sigma(\mathbf{x}/\mathbf{y})\prod_{\beta\in\Delta_1^+}(1+e^{-\beta})^{-1}=\sum_{\nu}(-1)^{|\nu|}s_{\bar{\nu},\sigma}(\mathbf{x})s_\nu(\mathbf{y}), \tag{8.39}$$

where the summation is over all partitions ν.

Our task is now to compare (8.32) and (8.39). In the former, the summation over $\mathbf{k}$ is restricted by (8.29) in just such a way that the parts of $(\sigma)_m - \mathbf{k}$ are weakly decreasing from left to right. It follows that

$$(\sigma)_m - \mathbf{k} = (\sigma_1 - k_1, \sigma_2 - k_2, \ldots, \sigma_m - k_m) = (\zeta_1, \zeta_2, \ldots, -\tau_2, -\tau_1), \tag{8.40}$$

where ζ and τ are both partitions. Clearly $l(\zeta) + l(\tau) \leqslant m$, so that $s_{(\sigma)_m - \mathbf{k}}(\mathbf{x}) = s_{\bar{\tau},\zeta}(\mathbf{x}) \neq 0$ is a standard sl(m) character defined as in (8.38). However, in general, η, as defined in (8.31b), is not a partition and it is necessary to apply the usual modification rules[39] for S functions to identify those η for which $s_\eta(\mathbf{y}) = \pm s_\nu(\mathbf{y}) \neq 0$ for some partition ν. In fact $s_\eta(\mathbf{x}) = 0$ unless the m tuple $\mathbf{k}$ is such that

$$k_i = (\sigma_i - i) - (\sigma_{t_i} - t_i), \quad 1 \leqslant i \leqslant m, \tag{8.41}$$

for some m tuple $\mathbf{t}$ such that $i \leqslant t_i < t_{i+1}$. Moreover, if (8.41) is satisfied, then the set

$$\{-\nu'_j + j | 1 \leqslant j < \infty\} = \{-\sigma_i - m + i | 1 \leqslant i < \infty\} \times \{-\sigma_{t_i} - m + t_i | 1 \leqslant i \leqslant m\}^{-1} \tag{8.42}$$

defines the partition ν' conjugate to ν such that $s_\eta(\mathbf{y}) = \pm s_\nu(\mathbf{y}) \neq 0$. That this is true follows from certain determinantal expansions of S functions and a combinatorial result given by Macdonald[40] linking a partition and its conjugate.

Conversely in (8.39), even if ν is a standard partition, its length may be such that $(\bar{\nu};\sigma)$ will not define, in general, a standard S function $s_{\bar{\nu},\sigma}(\mathbf{x})$ of sl(m). Once again recourse must be made to modification rules[25] to identify those terms for which $s_{\bar{\nu},\sigma}(\mathbf{x}) = \pm s_{\bar{\tau},\zeta}(\mathbf{x}) \neq 0$, with $s_{\bar{\tau},\zeta}(\mathbf{x})$ standard in the sense that both ζ and τ are partitions and $l(\zeta) + l(\tau) \leqslant m$. These modification rules, again based on determinantal expansions, imply that if ν is a partition with $s_\nu(\mathbf{y}) \neq 0$ then $s_{\bar{\nu},\sigma}(\mathbf{x})$ is nonvanishing if and only if

$$\{\sigma_j - j + 1 | 1 \leqslant j < \infty\} \setminus \{\nu'_i - i - m + 1 | 1 \leqslant i < \infty\} = \{\sigma_i - k_i - i + 1 | 1 \leqslant i \leqslant m\}, \tag{8.43}$$

for some m tuple $\mathbf{k}$. This forces $\mathbf{k}$ to be such that there exists $\mathbf{t}$ for which once again (8.41) is valid. Indeed, each nonvanishing term $s_{\bar{\nu},\sigma}(\mathbf{x})$ gives rise to a term $s_{\bar{\tau},\zeta}(\mathbf{x})$ where the connection is made through the deletion of the sequence $(\nu'_1 - m, \nu'_2 - m - 1)$ from the sequence $(\sigma_1, \sigma_2 - 1, \ldots)$ giving a new ordered list

$$\begin{aligned}(\zeta_1, \zeta_2 - 1, \ldots, -\tau_2 - m + 2, -\tau_1 - m + 1) &= (\sigma_{t_1} - t_1 + 1, \sigma_{t_2} - t_2 + 1, \ldots, \sigma_{t_m} - t_m + 1) \\ &= (\sigma_1 - k_1, \sigma_2 - k_2 - 1, \ldots, \sigma_m - k_m - m + 1),\end{aligned} \tag{8.44}$$

in agreement with (8.40) and (8.41).

Hence we have established

$$\sum_{\mathbf{k}} (-1)^{|\Lambda - \lambda|} s_{(\sigma)_m - \mathbf{k}}(\mathbf{x}) s_\eta(\mathbf{y}) = \sum_{\mathbf{t}} (\pm) s_{\bar{\tau},\zeta}(\mathbf{x}) s_\nu(\mathbf{y}), \tag{8.45a}$$

and

$$\sum_{\nu} (-1)^{|\nu|} s_{\bar{\nu},\sigma}(\mathbf{x}) s_\nu(\mathbf{y}) = \sum_{\mathbf{t}} (\pm) s_{\bar{\tau},\zeta}(\mathbf{x}) s_\nu(\mathbf{y}). \tag{8.45b}$$

where $\mathbf{t}$ runs over all m tuples in $\mathbf{N}^m$ with $t_1 < t_2 < \cdots < t_m$. All that remains in order to identify these two expressions is to verify that the sign factors are in agreement. This can be done. Hence, on comparing Lemmas (8.32) and (8.39), and using Theorem (5.7) we have the following theorem.

Theorem 8.46: If Λ_σ is the highest weight of an irreducible covariant tensor representation of sl(m/n) specified by the partition σ, then in the notation of (8.11), this representation has character given by

$$\text{ch } V(\Lambda_\sigma) = \chi_T(\Lambda_\sigma). \tag{8.46}$$

While this result (8.46) is only a special case of our Conjecture 8.14, the conjecture has thereby been proved in the case of all irreducible covariant tensor representations of sl(m/n), regardless of their degree of atypicality. In addition the Conjecture 8.14 is certainly correct in the case of all typical and all singly atypical irreducible representations, when it reduces to the Kac character formula and the Bernstein–Leites character formula, respectively. Moreover, it has stood up to the test of extensive computer calculations which, as explained in Sec. IV, led to the downfall of all formulas of the Kac-Weyl type. In the light of all these tests we remain optimistic concerning the validity of Conjecture 8.14, whose form we feel may well be amenable to rigorous derivation.

IX. CONCLUSION

In our analysis of character formulas for irreducible modules of sl(m/n) we have essentially introduced three new character formulas: denoted by $\chi_S(\Lambda)$, $\chi_J(\Lambda)$, and $\chi_T(\Lambda)$. The first two can be expressed in terms of a generating matrix, and hence they are of what we call the Kac–Weyl type (3.19). We have shown that $\chi_S(\Lambda)$ is equivalent to the Serganova–Serge'ev formula, and pointed out that, although $\chi_S(\Lambda)$ coincides with the Schur function formula of Berele and Regev for all irreducible covariant tensor modules, and is thus correct in these cases, it does not give the correct irreducible character in all cases. The formula $\chi_J(\Lambda)$ seems to cover many more cases than $\chi_S(\Lambda)$, but for $\chi_J(\Lambda)$ we have also found counterexamples to its validity. Moreover, we have demonstrated that there exist irreducible modules for which the character cannot be written in terms of any formula of the type (3.19).

We then introduced a different type of character formula, $\chi_T(\Lambda)$, which is a formal infinite expansion in terms of Kac characters $\chi_K(\lambda)$ (i.e., characters of Kac modules). This new character formula coincides with the Bernstein–Leites formula, $\chi_L(\Lambda)$, in the singly atypical case, and can be regarded as a truncation of the Bernstein–Leites formula for multiply atypical cases. Having proved that $\chi_L(\Lambda)$, and hence also $\chi_T(\Lambda)$, is correct in the case of all singly atypical cases, we have also proved that the character $\chi_T(\Lambda)$ gives the correct irreducible character when Λ is the highest weight of any irreducible covariant tensor representation.

Further extensive computer calculations lead us to conjecture that all irreducible characters of sl(m/n) are given

correctly by the extended Kac–Weyl formula $\chi_T(\Lambda)$ as in (8.14).

ACKNOWLEDGMENTS

We would like to thank the following people for stimulating discussions and/or sending us relevant preprints: A. J. Bracken and M. D. Gould (University of Queensland), C. J. Cummins (Concordia University, Montreal), S. Donkin (Queen Mary College, London), P. Pragacz (Polish Academy of Sciences), V. Serganova (Moscow).

This work was supported by the SERC (U.K.) Grant GR/D49909 (RCK, JVdJ), a Research Fellowship from NATO (Belgium) (JVdJ), by the sponsorship of a research visit to Paris (JWBH, RCK, and JVdJ) from CNRS (France), and a Royal Society European Exchange Programme Fellowship (JT-M). All of these contributions are gratefully acknowledged.

[1] L. Corwin, Y. Ne'eman, and S. Sternberg, Rev. Mod. Phys. **47**, 573 (1975).
[2] M. de Crombrugghe and V. Rittenberg, Ann. Phys. **151**, 99 (1983).
[3] F. Iachello, Phys. Rev. Lett. **44**, 772 (1980); A. B. Balantekin, I. Bars, and F. Iachello, Phys. Rev. Lett. **47**, 19 (1981); F. Iachello, Physica D **15**, 85 (1985).
[4] Y. Ne'eman, Phys. Lett. B **81**, 190 (1979); D. B. Fairlie, Phys. Lett. B **82**, 97 (1979); P. H. Dondi and P. D. Jarvis, Phys. Lett. B **84**, 75 (1979); I. Bars, Nucl. Phys. B **208**, 77 (1982).
[5] T. L. Curtright, G. I. Ghandour, and C. B. Thorn, Phys. Lett. B **182**, 45 (1986); R. J. Farmer, R. C. King, and B. G. Wybourne, J. Phys. A **21**, 3979 (1988).
[6] V. G. Kac, Funct. Anal. Appl. **9**, 263 (1975).
[7] V. G. Kac, Adv. Math. **26**, 8 (1977).
[8] V. G. Kac, Commun. Math. Phys. **53**, 31 (1977).
[9] M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **17**, 1626 (1976); **17**, 1640 (1976).
[10] M. Scheunert, *The Theory of Lie Superalgebras*, Lecture Notes in Mathematics, Vol. 716 (Springer, Berlin, 1979).
[11] J. Van de Leur, Ph.D. thesis, University of Utrecht, 1986.
[12] V. G. Kac, in Lecture Notes in Mathematics, Vol. 676, edited by K. Bleuler, H. Petry, and A. Reetz (Springer, Berlin, 1977) pp. 579–626.
[13] H. Weyl, Math. Z. **24**, 377 (1926).
[14] V. G. Kac, Commun. Algebra **5**, 889 (1977).
[15] P. H. Dondi and P. D. Jarvis, J. Phys. A **14**, 547 (1981).
[16] A. Berele and A. Regev, Bull. Am. Math. Soc. **8**, 337 (1983); A. Berele and A. Regev, Adv. Math. **64**, 118 (1987).
[17] A. Serge'ev, Math. USSR Sbornik **51**, 419 (1985).
[18] A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1149 (1981); R. C. King, Ars. Comb. A **16**, 269 (1983); B. G. Wybourne, J. Phys. A **17**, 1573 (1984).
[19] A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1810 (1981); **23**, 1239 (1982); I. Bars, Physica D **15**, 42 (1985); I. Bars, Lec. Appl. Math. **21**, 17 (1985).
[20] I. Bars, B. Morel, and H. Ruegg, J. Math. Phys. **24**, 2253 (1983).
[21] F. Delduc and M. Gourdin, J. Math. Phys. **25**, 1651 (1984).
[22] F. Delduc and M. Gourdin, J. Math. Phys. **25**, 1865 (1984).
[23] M. Gourdin, preprint: Université Pierre et Marie Curie, Paris, 1984, PAR LPTHE 84/26; M. Gourdin, preprint: Université Pierre et Marie Curie, Paris, 1984, PAR LPTHE 84/31.
[24] R. C. King, in Lecture Notes in Physics, Vol. 180, edited by M. Serdaroglu and E. Ínönü (Springer, Berlin, 1983) pp. 41–47; C. J. Cummins and R. C. King, J. Phys. A **20**, 3121 (1987).
[25] R. C. King, in *Invariant Theory and Tableaux*, edited by D. Stanton, IMA Volumes in Mathematics and its Applications (Springer, New York, 1990), Vol. 19, pp. 226–261.
[26] M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977).
[27] J. P. Hurni and B. Morel, J. Math. Phys. **24**, 157 (1983).
[28] T. D. Palev, J. Math. Phys. **26**, 1640 (1985); **27**, 1994–2001 (1986); **28**, 272 (1987).
[29] T. D. Palev, J. Math. Phys. **28**, 2280 (1987); **29**, 2589 (1988); **30**, 1433 (1989).
[30] I. N. Bernstein and D. A. Leites, C. R. Acad. Bulg. Sci. **33**, 1049 (1980); (in Russian); D. A. Leites, Funct. Anal. Appl. **14**, 106 (1980); D. A. Leites, Theor. Math. Phys. **52**, 764 (1982); D. A. Leites, J. Sov. Math. **25**, 2481 (1984).
[31] J. Thierry-Mieg, Phys. Lett. B **138**, 393 (1984).
[32] J. Thierry-Mieg, "Table des représentations irréductibles des superalgèbras de Lie su(m/n), su(n/n)/u(1), Osp(m/n), $D(2/1,\alpha)$, $G(3)$ et $F(4)$," unpublished (CNRS, Meudon, 1983); J. Thierry-Mieg, in Lecture Notes in Physics, Vol. 201, edited by G. Denardo, G. Ghirardi, and T. Weber (Springer, Berlin, 1984), pp. 94–98.
[33] J. Van der Jeugt, J. Phys. A **20**, 809 (1987).
[34] J. P Hurni, J. Phys. A **20**, 5755 (1987).
[35] I. Penkov and V. Serganova, Lett. Math. Phys. **16**, 251 (1988).
[36] M. D. Gould, J. Phys. A **22**, 1209 (1989).
[37] M. D. Gould, A. J. Bracken, and J. W. B. Hughes, J. Phys. A **22**, 2879 (1989).
[38] R. Le Blanc and D. J. Rowe, J. Math. Phys. **30**, 1415 (1989).
[39] D. E. Littlewood, *The Theory of Group Characters* (Oxford U.P., Oxford, 1940)
[40] I. G. Macdonald, *Symmetric Functions and Hall Polynomials* (Oxford U.P., Oxford, 1979).
[41] J. W. B. Hughes and R. C. King, J. Phys. A **20**, L1047 (1987).
[42] V. Serganova and A. Serge'ev, "Super Weyl groups and dominant weights" (unpublished).
[43] D. A. Leites, "Seminar on supermanifolds," University of Stockholm, (1987).
[44] J. Van der Jeugt, J. W. B. Hughes, R. C. King, and J. Thierry-Mieg, to be published in Commun. Algebra (1990).
[45] P. Pragacz, to be published in *Séminaire d'Algèbre Dubreil–Malliavin 1989/90,* Lecture Notes in Mathematics (Springer, Berlin).
[46] J. Stembridge, J. Algebra **95**, 439 (1985).

Classical scattering of a charged particle on an extended magnetic monopole

James B. Bowlin and Alfred S. Goldhaber
Institute for Theoretical Physics, State University of New York, Stony Brook, New York 11790-3840

(Received 6 December 1989; accepted for publication 14 February 1990)

The classical equations of motion for a charged particle scattering on an extended magnetic monopole are found in a model that includes the possibility of charge exchange between the particle and the pole. The special case of a spherical shell of monopole density is examined in detail. Deviations from point monopole scattering are analyzed.

I. INTRODUCTION

A necessary requirement for the consistency of quantum electrodynamics in the presence of magnetic monopoles is the well-known Dirac quantization condition,[1]

$$qg/\hbar c = n/2. \tag{1}$$

This formula together with the small size of the (experimentally measured) fine-structure constant implies that in order for the energy stored in the magnetic field surrounding a monopole to be less than or equal to the mass of the monopole (also clearly a necessary condition), some further modification to the laws of electrodynamics is required: The energy density must reach a "plateau" at a radius much larger than the Compton wavelength of the monopole, instead of continuing to grow as the radius decreases towards the Compton wavelength. This means that the monopole must have an internal structure described by a length scale so great that in a first approximation the pole may be considered as a classical object.[2,3] There are various ways this could happen.

One possibility is a generalization of the equation

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{2}$$

which governs the relationship between the vector potential A and the electromagnetic field $F \equiv (\mathbf{E},\mathbf{B})$. The field F in turn determines the energy density ($\sim \mathbf{E}^2 + \mathbf{B}^2$). The most elegant known modification to Eq. (2) is by generalization to non-Abelian gauge fields, permitting the type of stable classical monopole configurations first found by 't Hooft and by Polyakov in spontaneously broken SO(3).[4,5]

A second possibility is a modification to the metric that is used to integrate the energy density, giving decreasing weight to the region nearer the center of the monopole. This occurs automatically for a monopole that is also a black hole. A third, closely related case occurs if a new scalar field is present, which modifies the energy density as a multiplicative factor. The vanishing of this factor at the center of the monopole (this occurs for a monopole in Kaluza–Klein gravitoelectrodynamics) again permits a consistent semiclassical description of monopole structure, and again requires a generalization of standard electrodynamics.

What are the implications of monopole internal structure for scattering of charged particles? For an 't Hooft–Polyakov monopole the magnetic field strength is finite everywhere, so there exists a class of orbits with sufficiently high energy and small impact parameter to pass straight through the monopole and emerge on the other side. When this happens, some internal degree of freedom of the monopole (either its charge or internal angular momentum or both) must change in order to reconcile the conservation of total angular momentum with the behavior of the angular momentum carried by the crossed electric and magnetic fields (that would reverse sign if both charges stayed constant).

In this work we implement conservation of total angular momentum by considering a spherically symmetric classical Hamiltonian for charged particle motion in a monopole field like that of 't Hooft and Polyakov. This Hamiltonian reproduces point charge/point monopole dynamics at large distances from the monopole center, and reduces to free particle dynamics at very small distances. With this formalism we find that the possibility of charge exchange between the monopole and the charged particle emerges quite naturally, and that this exchange only occurs when the charged particle actually penetrates the monopole. In order to keep the problem tractable we work in the small coupling limit ($e^2/\hbar c \to 0$), so that we may ignore the additional interaction due to the electric charge that may be deposited on the magnetic monopole during the collision.

Our formalism is established and the general features of the solutions are explored in the second section. The third section is devoted to the special case of a spherical shell of monopole density. The conclusion contains a summary of the results of the previous two sections and mentions some further questions that have been raised in the course of this work. Several interesting mathematical problems are solved in the two appendices.

Our model of the interior of a magnetic monopole is not unique, since a variety of possibilities exist, including those mentioned above. In the absence of experimental observations, the only criteria for selecting among models consistent with known physics are elegance and simplicity. Our choice is unrealistically simple, but nonetheless has the merits that it respects gauge invariance (necessary for the existence of a correspondence between classical and quantum mechanics including monopoles) and is sufficiently tractable to analysis so that it becomes possible to study a rich and intricate dynamics in considerable detail. While we take that as sufficient justification, we should still note the simplifications imposed. First and most obvious is the fact that we are using classical mechanics, which implies that our charged particle is heavy enough so that its Compton wavelength is small compared to the radius of the monopole. By the same token, we are neglecting any effects on the particle trajectory of

 0022-2488/90/092305-09$03.00

recoil by the monopole, which must be assumed to be much heavier still. Also, we are using the simplest non-Abelian gauge theory with monopoles, spontaneously broken Yang–Mills theory. While it is not known if there is in nature a grand unified gauge field dynamics accounting for all observed phenomena, it is known that any such theory would have to involve a much bigger gauge group than the SO(3) of Yang and Mills. Finally, purely for calculational ease, we shall assume that the radius of our monopole marks a sharp transition between an exterior which influences particles exactly as would a point Dirac monopole, and an interior in which the particle moves with no acceleration at all. Of course the classical, static 't Hooft–Polyakov field configuration is completely smooth, so that the transition between exterior and interior occurs gradually, only becoming complete at the exact center of the monopole.

II. EQUATIONS OF MOTION

To find the classical motion of a particle of charge q in the field of an extended monopole with total magnetic charge g we start with the Hamiltonian[6]

$$H=(1/2m)(\mathbf{P}+(1-f(r))[(\mathbf{S}\times\mathbf{r})/r^2])^2, \tag{3}$$

where S has dimensions of angular momentum and is assumed to have Poisson bracket relations

$$\{S_i,S_j\}=\epsilon_{ijk}S_k, \tag{4}$$

with

$$(\mathbf{S}\cdot\hat{r})=-qg/c. \tag{5}$$

The function $f(r)$ may be loosely interpreted as a measure of the deviation from point monopole dynamics as a function of radial distance. For $f\equiv 0$ we have point monopole dynamics, while for $f\equiv 1$ we have free-particle dynamics. Thus f is usually chosen as a monotonically decreasing function of r obeying the conditions $\lim_{r\to\infty} f(r)=0$ and $f(0)=1$.

Using standard procedures we find the following equations of motion:

$$\dot{\mathbf{S}}=(1-f)\mathbf{S}\times[(\dot{\mathbf{r}}\times\mathbf{r})/r^2], \tag{6}$$

$$m\ddot{\mathbf{r}}=\frac{-\dot{\mathbf{r}}}{r^2}\times\left[(1-f)\mathbf{S}_r+\left(r\frac{\partial f}{\partial r}\right)\mathbf{S}_\perp\right], \tag{7}$$

where

$$\mathbf{S}_r\equiv(\mathbf{S}\cdot\hat{r})\hat{r}, \tag{8}$$

$$\mathbf{S}_\perp\equiv\mathbf{S}-\mathbf{S}_r. \tag{9}$$

If we define $\mathbf{L}\equiv m(\mathbf{r}\times\dot{\mathbf{r}})$ to be the "orbital" angular momentum, then

$$\mathbf{J}=(\mathbf{r}\times\mathbf{P})+\mathbf{S} \tag{10}$$

$$=\mathbf{L}+\mathbf{S}_r+f\mathbf{S}_\perp \tag{11}$$

is a constant of the motion. So are $|\mathbf{S}|$ and $|\dot{\mathbf{r}}|$. Finally we note

$$d(\mathbf{S}\cdot\hat{r})/dt=-f[(\mathbf{S}_\perp\cdot\dot{\mathbf{r}})/r]. \tag{12}$$

This implies that far away from the monopole, where $f\approx 0$, we find $\mathbf{S}\cdot\hat{r}$ is approximately constant, and the equations of motion for the charged particle become (in the limit $f\to 0$)

$$m\ddot{\mathbf{r}}=-(\mathbf{S}\cdot\hat{r})[(\dot{\mathbf{r}}\times\mathbf{r})/r^2], \tag{13}$$

which is the equation of motion for a particle with charge defined by Eq. (5) in the field of a point monopole.

For motion nearer the pole with f approaching unity we choose to identify the change in qg/c as due to a change in q, the charge of the particle. We see that $(\mathbf{S}\cdot\hat{r})$ is analogous to the third component of isospin,[7] but since in this model $(\mathbf{S}\cdot\hat{r})$ can change continuously, we must consider the charged particle as a member of an infinitely large isospin multiplet. Since $(\mathbf{S}\cdot\hat{r})$ can change while the particle is inside the monopole, the charge of the particle can be changed in the process of monopole scattering. So although J is conserved in this model, the charge of the particle by itself is no longer necessarily conserved.

In order to understand the scattering orbits, note that far away from the monopole we have

$$\mathbf{J}=\mathbf{L}+\mathbf{S}_r \tag{14}$$

and

$$\mathbf{L}\perp\mathbf{S}_r, \tag{15}$$

implying

$$J^2=L^2+S_r^2. \tag{16}$$

We see that a change in $(\mathbf{S}\cdot\hat{r})$ is accompanied by a change in $|\mathbf{L}|$;

$$\Delta L^2=-\Delta S_r^2. \tag{17}$$

If we start out with $\mathbf{S}\|\hat{r}$ then L^2 can only increase. This information suffices to give us a rough description of the orbits when the particle is far away from the pole. Well before the scattering event the particle's path is restricted to lie on a cone of half-angle $\theta_i=\sin^{-1}(L_i/J)$ centered about J with its apex on the center of the monopole.[6] Well after the event the path is restricted to a new cone also centered about J with its apex on the center of the monopole but with half-angle $\vartheta_f=\sin^{-1}(L_f/J)$, where

$$L_f=(L_i^2+\Delta(\mathbf{S}\cdot\hat{r})^2)^{1/2}. \tag{18}$$

If $(\mathbf{S}\cdot\hat{r})$ is positive, the new cone opens in the same direction along J as the original. If it is negative then the new cone opens in the opposite direction. To complete the description of the scattering event, the angle ω that the particle has revolved about the J axis must be specified. The polar scattering angle, ϑ is given by

$$\cos(\vartheta)=-(L_iL_f/J^2)\cos(\omega)-\mathrm{Sgn}(\mathbf{S}\cdot\hat{r})\times[(1-(L_i/J)^2)(1-(L_f/J))^2]^{1/2}. \tag{19}$$

So for a given $f(r)$ and S the differential scattering cross section can be determined by finding $\Delta(\mathbf{S}\cdot\hat{r})$ and ω as functions of b and $\dot{r}$. Unfortunately, we know of no general way to determine these functions except by direct numerical integration.

III. SPECIAL CASE OF A SPHERICAL SHELL

Many of the features mentioned above may be seen in the special case of a particle with $(\mathbf{S}\cdot\hat{r})$ initially equal to $|\mathbf{S}|$ scattering on an infinitely thin spherical shell of monopole density. Define f such that

$$f=\begin{cases}1, & r<r_0,\\ 1-(r-r_0)/\delta, & r_0<r<r_0+\delta,\\ 0, & r>r_0+\delta.\end{cases} \tag{20}$$

We then take the limit $\delta\to 0$. Inside the shell, the particle moves along straight line trajectories and $\mathbf{S}$ remains constant. Outside the shell, $(\mathbf{S}\cdot\hat{r})$ is conserved and the particle behaves as if it were in the field of a point monopole. The behavior of the particle in the region of nonzero monopole density, where $r_0<r<r_0+\delta$, is quite interesting and can be deduced directly from the conservation of

$$\mathbf{J}=\mathbf{L}+\mathbf{S}_r+f\mathbf{S}_\perp. \tag{21}$$

Since $\mathbf{S}$ cannot change discontinuously as the particle traverses the shell, $\mathbf{L}$ must change by $\mathbf{S}_\perp$ as f goes from one to zero. If the magnitude of the resulting L is greater than $L_c\equiv mr_0|\dot{r}|$ (L_c is the maximum possible magnitude of L for a given $|\dot{r}|$ at the radial distance r_0), then the particle cannot pass through the shell and it must undergo specular reflection, i.e., it must bounce. Of course if $\mathbf{S}_\perp$ is initially zero then L will not change in traversing the shell and the particle will be allowed to enter. The only aspect we are not assured of by this reasoning is whether the particle always passes through the shell when it is allowed to by the conservation law. We have calculated the orbits in the region $r_0<r<r_0+\delta$ in the limit $\delta\to 0$ and in addition to verifying the aforementioned deduction, we have found that the particle always passes through the shell when allowed.

A particle impinging on the shell from the outside will bounce off if $(\mathbf{L}-\mathbf{S}_\perp)^2>L_c^2$, and if it does pass through, the change in $\mathbf{L}$ will be equal to $-\mathbf{S}_\perp$. These same relations apply for a particle impinging on the shell from the inside except there is a relative sign change between $\mathbf{L}$ and $\mathbf{S}_\perp$.

In order to describe the classical orbits we define two dimensionless parameters

$$\beta\equiv b/r_0, \tag{22}$$

$$\alpha\equiv L_c/|\mathbf{S}|. \tag{23}$$

If $\beta>1$ then the particle never enters the shell and scatters as if from a point monopole. For $\beta<1$ the particle rotates about the $\mathbf{J}$ axis an angle

$$\omega_1=(J/L)\sin^{-1}(\beta) \tag{24}$$

before entering the shell. If $\mathbf{S}_\perp$ is initially zero (as we shall assume here), it always passes through the shell into the interior. Once inside, the motion lies in a plane perpendicular to $\mathbf{L}$ at the point of impact. This plane always contains the origin. Thus the motion inside the shell is restricted to a disk of radius r_0. After entering, the particle subtends an angle $\psi=2\cos^{-1}(\beta)$ around the center of the disk before striking the shell again. It then either bounces or passes through. In general it will bounce $(m-1)$ times (with m still to be determined) and then exit at an angle $\phi=m\psi$ relative to the point of impact. In order for the particle to pass through the shell on its way out we must have

$$S_\perp^2<L_c^2-L^2 \tag{25}$$

or

$$\sin^2(\phi)<\alpha^2(1-\beta^2). \tag{26}$$

Thus the particle can escape only if $m\psi$ is within ϕ_c of zero or π where

$$\phi_c=\sin^{-1}(\alpha(1-\beta^2)^{1/2}). \tag{27}$$

We define $m(\psi)$ as the smallest integer greater than zero such that

$$\sin^2(2m\cos^{-1}\beta)<\alpha^2(1-\beta^2) \tag{28}$$

or

$$\sin^2(m\psi)/\sin^2(\psi/2)<\alpha^2. \tag{29}$$

Regions where $m(\psi)=n$ are plotted in Fig. 1 as a function of α (vertical) and ψ (horizontal). The outlines of these regions were obtained by setting both sides of the previous equation equal and solving for α in terms of ψ. Where two different regions would overlap the region of lower n always prevails (since the particle will always escape at the first opportunity) and we say that the region of higher n has been occluded.

Since α scales with r_0, the "size" of the monopole, and ψ is explicitly a function of β only, we are particularly interested in the behavior of the function $m(\psi)$ for various fixed α's. In general this behavior is simple for large α and becomes more complicated as $\alpha\to 0$.

For $\alpha>2$, the particle always passes straight through without bouncing, so $m(\psi)=1$ and the outgoing charge is $q\cos(\psi)$. For $\alpha<2$ the function $m(\psi)$ takes on an infinite number of values and $\lim_{\psi\to 0}m(\psi)=\infty$. For $1<\alpha<2$, the angle ψ (the size of the angular steps the particle takes about the disk) is always less than twice the critical angle (ϕ_c), so the particle either passes straight through or bounces until it reaches the allowed region near π where it passes through. Thus $m(\psi)$ is a decreasing function, always changing in unit steps except for possibly the last step [to the region where $m(\psi)=1$] that may be greater than one. When α is less than one, it is possible for the particle to miss the allowed region near π on its first pass so $m(\psi)$ is no longer monotonic and

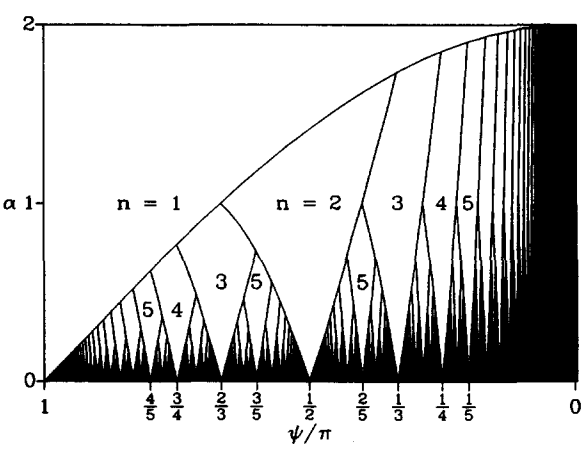


FIG. 1. Regions where $m(\psi)=n$ are plotted for given ψ/π (horizontal) and α (vertical). The regions with $n\leq 5$ have been labeled with their corresponding n-values. Extrapolation to higher values of n is straightforward. Note that since $b/r_0=\cos(\psi/2)$, $\psi/\pi=1$ corresponds to $b=0$ while $\psi/\pi=0$ corresponds to $b=r_0$.

develops a very complicated structure. A derivation of $P(n)$, the probability that $m(\psi)=n$ given that the particle has struck the shell in the limit $\alpha\to 0$, is contained in the appendices. The result is

$$P(n)=[4\alpha\phi(n)/3\pi]g(n), \tag{30}$$

where

$$g(n)=\begin{cases}1, & n<n_0,\\ -1+2\left(\frac{n_0}{n}\right)+2\left(\frac{n_0}{n}\right)^2-2\left(\frac{n_0}{n}\right)^3\left(1+2\ln\left(\frac{n_0}{n}\right)\right) & \\ +\left(2\left(\frac{n_0}{n}\right)-1\right)^2\left(\left(\frac{n_0}{n}\right)+1\right)\ln\left(2\left(\frac{n_0}{n}\right)-1\right), & n_0<n<2n_0,\\ \left(\frac{n_0}{n}\right)^3(4\ln 2+2), & n>2n_0.\end{cases} \tag{31}$$

Here $\phi(n)$ is *Euler's totient function* and is equal to the number of integers less than n that share no common divisors with n; we define $n_0\equiv\pi/2\alpha$.

IV. PHYSICAL OBSERVABLES

The results above suggest that one might be able to distinguish between pointlike and extended monopoles by studying the scattering of charged particles from the monopoles. Even when this is done, all physical observables will be identical in the two cases unless b, the impact parameter, is less than r_0, the size of the monopole. Below we shall discuss the difference in the charge distribution and the differential cross section, $d\sigma/d\vartheta$ assuming the charged particle has actually struck the monopole. In order to compare these predictions with experiment the contribution to the cross sections from orbits with $b>r_0$ must be added to the distributions discussed below. Such orbits are identical for the extended shell and the point monopole. In the limit $r_0\to 0$ the contribution from orbits in which the particle strikes the shell goes to zero, so that in our classical (nonquantum) model there is no observable difference between an infinitesimally small monopole shell and a point monopole.

For a point monopole there are an infinite number of cusps in the differential cross section, $d\sigma/d\vartheta$, even when we exclude the contribution due to orbits with $b>r_0$. The same is true for a monopole shell with α less than two. However, the infinite number of cusps in the point monopole distribution are due to orbits with b arbitrarily close to 0, while for the shell, they are due to orbits with b arbitrarily close to r_0. As $\alpha\to 0$ the number of cusps for any finite range of the impact parameter ($a<b<c$:$a,c<r_0$) tends to infinity. For a fixed $\alpha>0$ it is possible to calculate the contribution to $d\sigma/d\vartheta$ for $b<b_1<r_0$. In this case there will only be contributions from $n<n_{\max}$ where $n_{\max}=(\pi/\alpha)\sqrt{1-(b_1/r_0)^2}$. The resulting spectrum appears to be random, presumably because effects of motion before entering the shell, while inside the shell, and after leaving the shell combine in a haphazard manner.

There is a striking difference between the two differential cross sections. The cusps in the point monopole differential cross section are concentrated near $\vartheta=\pi$, while the cusps in the spherical shell cross section are more randomly distributed. For $\alpha<2$, the function $\vartheta(b)$ has an infinite number of discontinuities. Finally, we note that, in the case of a shell, $b<r_0$ implies

$$|\sin(\vartheta)|\leqslant 2\alpha. \tag{32}$$

So in the limit $\alpha\to 0$, we find that ϑ is always within 2α of either π or 0, and there are an infinite number of cusps in both of these regions.

The charge distribution, $(1/\sigma_0)d\sigma/dq$, for a monopole shell of course is different from that of a point monopole. The point monopole has $(1/\sigma_0)(d\sigma/dq)=\delta(q-q_0)$, where q_0 is the initial charge, while an extended shell always has a distribution of finite width. For $\alpha>2$ the distribution is continuous and varies from $+q_0$ to $-q_0$. For $\alpha<2$ there are an infinite number of discontinuities, with

$$1\geqslant|q/q_0|^2\geqslant 1-\alpha^2. \tag{33}$$

As α goes to zero the charge distribution peaks in two very small regions around q_0 and $-q_0$.

When the particle is on an orbit which corresponds to a region of even n (i.e., it bounces an odd number of times) then it must exit through the allowed region near $\psi=\pi$ which results in $\hat{S}\cdot\hat{r}<0$. Orbits which correspond to a region of odd n exit through both allowed regions with equal probability in the small α limit. Also, the average value of $\phi(n)/n$ differs for odd and even n by a factor of 2;

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{\phi(2k-1)}{2k-1}=2\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{\phi(2k)}{2k}. \tag{34}$$

The net result of these two effects is that the magnitude of the charge distribution near $q\approx -q_0$ is approximately twice that of the distribution near $q\approx q_0$ while the shapes of the distribution in these two regions are similar. Although the number of discontinuities in $(q_0/\sigma_0)(d\sigma/dq)$ increases without bound as $\alpha\to 0$, the distribution also approaches a continuous function in this limit. We found an approximation for the limit function using techniques similar to those

used in Appendix A:

$$\lim_{\alpha\to 0}\frac{q_0}{\sigma_0}\left(\frac{d\sigma}{dq}+\frac{d\sigma}{d(-q)}\right)\approx\frac{12}{\pi^2\alpha^2 t}\left(1+(1-t)^{1/2}-t\ln\left(\frac{1+(1-t)^{1/2}}{t^{1/2}}\right)-t\right), \tag{35}$$

where

$$t=(1-q^2)^{1/2}/\alpha. \tag{36}$$

This approximation, when integrated over q, overpredicts the total cross section by about 15%. The discrepancy is attributable to use of the approximate solution to

$$y=\frac{1-F}{1+F}\left(1-\ln\left(\frac{1-F}{1+F}\right)\right), \tag{37}$$

namely

$$F\approx(1-y)^{1/2}, \tag{38}$$

which was used to sum the contributions due to partially occluded regions of n. A numerical solution is straightforward but unilluminating.

V. CONCLUSIONS

Just as Rutherford was able to probe the structure of the nucleus by scattering charged particles on nuclei, the internal structure of a monopole could be probed with similar scattering experiments. In this work we have used a very simple model to explore the question, what happens to a charged particle in the interior of a magnetic monopole?

We have solved the special case of scattering on an infinitesimally thin monopole shell. For this case the deviations from point monopole scattering of the charge and angular distributions have been studied, with particular attention paid to the limit in which the size of the spherical shell approaches zero. Because of the almost random variation in the number of bounces made while the charge is inside the monopole, the detailed structure of these distributions becomes extremely complicated. Nevertheless we were able to determine many of the characteristics of the distributions. In addition, our explicit form of $P(n)$ makes feasible the calculation of average values of many physical properties (such as time spent inside the monopole).

Even though none have yet been found in nature, there are still many benefits to be gleaned from the theoretical study of magnetic monopoles. The concept provides a ground for exploring the structure, beauty, and consistency of theories, as well as seeking specific predictions of phenomena involving monopoles that differ from one theory to another. Thus, we learn from studying such questions today, and would only be rewarded further if monopoles were found in the future.

ACKNOWLEDGMENTS

We wish to thank the Department of Applied Mathematics and Theoretical Physics at Cambridge University whose warm and gracious hospitality we enjoyed while most of this work was being done. We are also most grateful to Professor Chih-Han Sah and Professor Patrick X. Gallagher for their indispensible help with Appendix B. Their speed, thoroughness, and courtesy were a delight.

This work was supported in part by the (U.S.) National Science Foundation (grant #PHY89-08495) and the (U.K.) Science and Engineering Research Council.

APPENDIX A

Here we find $P(n)$, the probability that a particle bounces exactly $(n-1)$ times given that it has struck a monopole shell in the limit $\alpha(\equiv mr_0\dot{r}/|S|)\to 0$ and for large n. These limits will be assumed implicitly throughout this appendix.

Define $x\equiv\sin(\psi/2)$ and $P(n,x)dx$ as the probability that the particle bounces exactly $(n-1)$ times for x within the small range dx of x. Then

$$P(n,x)=\widehat{\Delta x}_n(x)\rho_n(x)P(x)dx, \tag{A1}$$

where $\widehat{\Delta x}_n(x)$ is the average width (as a function of x) of all of the regions within dx such that $m(\psi)=n$, $\rho_n(x)dx$ is the number of regions such that $m(\psi)=n$ within dx, and $P(x)dx$ is the probability that $\sin(\psi/2)$ is within dx. Now we know

$$P(x)=\frac{1}{\sigma_0}\frac{d\sigma}{dx} \tag{A2}$$

$$=2x \tag{A3}$$

and also

$$\widehat{\Delta x}_n(x)\rho_n(x)=\widehat{\Delta\psi}_n(x)\rho_n(\psi). \tag{A4}$$

In Appendix B we prove that $\rho_n(\psi)=\phi(n)/\pi$ and we are left with finding $\widehat{\Delta\psi}_n(x)$.

For x and n such that

$$xn<\pi/2\alpha, \tag{A5}$$

none of the regions of n within dx are occluded by their neighbors so

$$\widehat{\Delta\psi}_n(x)=2/\left|\frac{\partial\alpha_n(\psi)}{\partial\psi}\right|_{\alpha\approx 0} \tag{A6}$$

$$=2x/n. \tag{A7}$$

Here we have used the small angle approximation with $\alpha_n(\psi)=\sin(n\psi)/\sin(\psi/2)$.

For x and n such that

$$xn>\pi/\alpha, \tag{A8}$$

then all of the regions of n within dx are totally occluded by their neighbors and we have simply

$$\widehat{\Delta\psi}_n(x)=0. \tag{A9}$$

In order to express $\widehat{\Delta\psi}_n(x)$ precisely for the intermediate case of

$$\pi/2\alpha<xn<\pi/\alpha, \tag{A10}$$

we define $m_{n,k}$ as the k th-largest integer that shares no common divisors with n, and define $\delta_{n,k}$ as $d_{n,k}/n$ where

$$d_{n,k} < n \tag{A11}$$

and

$$d_{n,k} m_{n,k} = 1 \pmod{n}. \tag{A12}$$

It is then easy to show that $m(\psi) = d_{n,k}$ in the region that occludes the kth region of n from the left and $m(\psi) = n - d_{n,k}$ in the region that occludes the k th region of n from the right. Furthermore, the k th region of n intersects the ψ axis at $\psi = \pi m_{n,k}/n$.

Now define the functions $\Delta\psi_{n,k}(x)$ as

$$\Delta\psi_{n,k}(x) \equiv \begin{cases} \dfrac{\alpha x}{n}, & \dfrac{\pi}{2\alpha} < xn < \dfrac{\pi}{\alpha(1+\delta_{n,k})}, \\ \dfrac{\alpha}{n}\left(\dfrac{\pi}{\alpha n} - x\right)\left(\dfrac{1}{\delta_{n,k}}\right), & \dfrac{\pi}{\alpha(1+\delta_{n,k})} < xn < \dfrac{\pi}{\alpha}. \end{cases} \tag{A13}$$

Then

$$\widehat{\Delta\psi}_n(x) = \langle \Delta\psi_{n,k}(x) + \Delta\psi_{n,(n-k)} \rangle, \tag{A14}$$

where the average is taken over all k such that $\sin(\pi m_{n,k}/n)$ is within dx. In Appendix B we prove that

$$\langle \Delta\psi_{n,k}(x) \rangle|_{\sin(\pi m_{n,k}/2n) \in dx} = \langle \Delta\psi_{n,k}(x) \rangle|_{1 < m_{n,k} < n}. \tag{A15}$$

The right-hand side is easy to evaluate. Let $f_n(x)$ be equal to the fraction of δ's such that

$$\pi/2\alpha < xn < \pi/\alpha(1+\delta_{n,k}). \tag{A16}$$

Then

$$f_n(x) = (\pi/\alpha xn - 1), \tag{A17}$$

since the δ's are uniformly distributed between zero and one. This also implies that if we define $f(\delta)$ as the fraction of δ's less than δ then $f(\delta) = \delta$. We have

$$\widehat{\Delta\psi}_n(x) = \frac{2\alpha}{n}\left(x f_n(x) + \left(\frac{\pi}{\alpha n} - x\right)\int_{fn(x)}^{1} \frac{df(\delta)}{\delta}\right) \tag{A18}$$

$$= \frac{2\alpha x}{n}\left(\frac{\pi}{\alpha xn} - 1\right)\left(1 - \ln\left(\frac{\pi}{\alpha nx} - 1\right)\right). \tag{A19}$$

Thus

$$P(n,x) = \frac{4\alpha\phi(n)x^2}{\pi n} \tag{A20}$$

$$\times \begin{cases} 1, & xn < \pi/2\alpha, \\ (\pi/\alpha xn - 1)(1 - \ln(\pi/\alpha nx - 1)), & \pi/2\alpha < xn < \pi/\alpha, \\ 0, & xn > \pi/\alpha \end{cases} \tag{A21}$$

and

$$P(n) = \int_0^1 P(n,x)dx \tag{A22}$$

$$= [4\alpha\phi(n)/3\pi]\, g(n), \tag{A23}$$

where

$$g(n) = \begin{cases} 1, & n < n_0, \\ -1 + 2\left(\dfrac{n_0}{n}\right) + 2\left(\dfrac{n_0}{n}\right)^2 - 2\left(\dfrac{n_0}{n}\right)^3\left(1 + 2\ln\left(\dfrac{n_0}{n}\right)\right) & \\ \quad + \left(2\left(\dfrac{n_0}{n}\right) - 1\right)^2\left(\left(\dfrac{n_0}{n}\right) + 1\right)\ln\left(2\left(\dfrac{n_0}{n}\right) - 1\right), & n_0 < n < 2n_0, \\ \left(\dfrac{n_0}{n}\right)^3 (4\ln 2 + 2), & n > 2n_0, \end{cases} \tag{A24}$$

and

$$n_0 = \pi/2\alpha. \tag{A25}$$

APPENDIX B

The results in this appendix are extracted from conversations between Professor Sah and Professor Gallagher of the Mathematics Departments at SUNY Stony Brook and Columbia University, respectively. According to them, these are well-known "folklores" in analytic number theory.

We want to show that for any finite segment (y_1, y_2), $0 \leqslant y_1 < y_2 \leqslant 1$, for all $n > n_0$ $(n_0 \equiv \pi/2\alpha)$

$$\lim_{\alpha \to 0} \langle W_{n,k}(x) \rangle|_{y_1 < m_{n,k} < y_2} = \langle W_{n,k}(x) \rangle|_{0 < m_{n,k} < 1}, \tag{B1}$$

the W functions are defined as

$$W_{n,k}(x)\equiv\begin{cases}1, & \dfrac{\pi}{2\alpha}<xn<\dfrac{\pi}{\alpha(1+\delta_{n,k})},\\ \dfrac{1}{\delta_{n,k}}\left(\dfrac{\pi}{\alpha nx}-1\right), & \dfrac{\pi}{\alpha(1+\delta_{n,k})}<xn<\dfrac{\pi}{\alpha},\end{cases} \tag{B2}$$

where

$$\delta_{n,k}\equiv d_{n,k}/n \tag{B3}$$

and

$$d_{n,k}m_{n,k}=1 \bmod n. \tag{B4}$$

Note that the functions $W_{n,k}(x)$ are related to the functions $\Delta\psi_{n,k}(x)$ of Appendix A by

$$W_{n,k}(x)=(n/\alpha x)\Delta\psi_{n,k}. \tag{B5}$$

Since the shapes of the functions $W_{n,k}(x)$ depend solely on the multiplicative inverses $d_{n,k}$, it suffices to show that in the limit $n\to\infty$, we can make the following statements.

(A) $m_{n,k}$ are uniformly distributed.

(B) The inverses $d_{n,k}$ from the interval (y_1,y_2) (i.e., $\{d_{n,k}:y_1<m_{n,k}/n<y_2\}$) are uniformly distributed.

Define $f(n;x,y)$ as the number of pairs (a,b) such that

(i) $ab=1 \bmod n$,

(ii) $0\leqslant a\leqslant nx\leqslant n$,

(iii) $0\leqslant b\leqslant ny\leqslant n$.

Then statements A and B above are equivalent to the claim

$$\lim_{n\to\infty}[1/\phi(n)]f(n;x,y)=xy. \tag{B6}$$

The remainder of this appendix is devoted to proving this claim.

Proof: We first prove statement A. Define $f(n;x)$ such that

$$f(n;x)\equiv\sum_{\substack{a<xn\\ \gcd(a,n)=1}}1. \tag{B7}$$

Thus $f(n;x)$ counts the fraction of $m_{n,k}$'s that are less than xn. Statement A is equivalent to the claim that

$$\lim_{n\to\infty}[f(n;x)/\phi(n)]=x. \tag{B8}$$

Express n as a product of unique primes

$$n=\prod_{i=1}^{m}P_i^{\alpha_i}, \tag{B9}$$

then

$$\phi(n)=n\prod_{i=1}^{m}\left(1-\frac{1}{P_i}\right), \tag{B10}$$

which may be reexpressed as

$$\phi(n)=n\left(1-\sum_i\frac{1}{P_i}+\sum_{i\neq j}\frac{1}{P_iP_j}-\cdots\pm\frac{1}{P_1P_2\cdots P_m}\right). \tag{B11}$$

We can express $f(n;x)$ as

$$f(n;x)=[xn]-\sum_i\left[\frac{xn}{P_i}\right]+\sum_{i\neq j}\left[\frac{xn}{P_iP_j}\right]-\cdots\pm\left[\frac{xn}{P_1P_2\cdots P_m}\right]. \tag{B12}$$

So

$$|x\phi(n)-f(n;x)|\leqslant 1+\sum_i 1+\sum_{i\neq j}1+\cdots+1 \tag{B13}$$

$$=\sum_{k=1}^{m}\binom{m}{k} \tag{B14}$$

$$=2^m. \tag{B15}$$

We have

$$\lim_{n\to\infty}|x-f(n;x)/\phi(n)|\leqslant 2^m/\phi(n) \tag{B16}$$

$$=\prod_{i=1}^{m}\frac{2}{(P_i-1)P_i^{(\alpha-1)}} \tag{B17}$$

$$\to 0. \tag{B18}$$

We now proceed to prove statement B.

We can express the function $f(n;x,y)$ as

$$f(n;x,y)=\sum_{\substack{a,b\\ ab=1 \bmod n}}\sum_{\alpha=1}^{nx-1}\delta_{a,\alpha}\sum_{\beta=1}^{ny-1}\delta_{b,\beta}. \tag{B19}$$

Now expand the delta functions as

$$\delta_{a,\alpha}=\frac{1}{n}\sum_{c=0}^{n-1}e_n((a-\alpha)c), \tag{B20}$$

where $e_n(x)\equiv\exp(2\pi ix/n)$. We then have

$$f(n;x,y)=\frac{1}{n^2}\sum_{\substack{a,b\\ ab=1 \bmod n}}\sum_{c=0}^{n-1}\sum_{d=0}^{n-1}\sum_{\alpha=1}^{nx-1}\sum_{\beta=1}^{ny-1}e_n((a-\alpha)c)\times e_n((b-\beta)d) \tag{B21}$$

$$=f_{c=0}+f_{d=0}-f_{c=d=0}+f_{\text{trash}}, \tag{B22}$$

where

$$f_{c=0}\equiv\frac{1}{n^2}\sum_{\substack{a,b\\ ab=1 \bmod n}}\sum_{c=0}^{0}\sum_{d=0}^{n-1}\sum_{\alpha=1}^{nx-1}\sum_{\beta=1}^{ny-1}e_n((a-\alpha)c)\times e_n((b-\beta)d), \tag{B23}$$

$$f_{d=0}\equiv\frac{1}{n^2}\sum_{\substack{a,b\\ ab=1 \bmod n}}\sum_{c=0}^{n-1}\sum_{d=0}^{0}\sum_{\alpha=1}^{nx-1}\sum_{\beta=1}^{ny-1}e_n((a-\alpha)c)$$

$$\times e_n((b-\beta)d), \tag{B24}$$

$$f_{c=d=0} \equiv \frac{1}{n^2} \sum_{\substack{a,b \\ ab=1 \bmod n}} \sum_{c=0}^{0} \sum_{d=0}^{0} \sum_{\alpha=1}^{nx-1} \sum_{\beta=1}^{ny-1} e_n((a-\alpha)c)$$

$$\times e_n((b-\beta)d), \tag{B25}$$

$$f_{\text{trash}} \equiv \frac{1}{n^2} \sum_{\substack{a,b \\ ab=1 \bmod n}} \sum_{c=1}^{n-1} \sum_{d=1}^{n-1} \sum_{\alpha=1}^{nx-1} \sum_{\beta=1}^{ny-1} e_n((a-\alpha)c)$$

$$\times e_n((b-\beta)d). \tag{B26}$$

The evaluation of the first three of these functions is trivial,

$$f_{c=d=0} = \frac{1}{n^2} \sum_{\substack{a,b \\ ab=1 \bmod n}} \sum_{\alpha=1}^{nx-1} \sum_{\beta=1}^{ny-1} 1 \times 1 \tag{B27}$$

$$= \sum_{\substack{a,b \\ ab=1 \bmod n}} \frac{nx\,ny}{n^2} \tag{B28}$$

$$= \phi(n)xy. \tag{B29}$$

Also,

$$f_{d=0} = \frac{1}{n^2} \sum_{\substack{a,b \\ ab=1 \bmod n}} \sum_{c=0}^{n-1} \sum_{\alpha=1}^{nx-1} \sum_{\beta=1}^{ny-1} e_n((a-\alpha)c) \tag{B30}$$

$$= y \sum_{\substack{a,b \\ ab=1 \bmod n}} \sum_{\alpha=1}^{nx-1} \delta_{a,\alpha} \tag{B31}$$

$$= \phi(n)xy. \tag{B32}$$

Likewise, $f_{c=0} = \phi(n)xy$, so

$$\lim_{n\to\infty} f(n;x,y) = \phi(n)xy + \lim_{n\to\infty} f_{\text{trash}}. \tag{B33}$$

We now proceed to show that

$$\lim_{n\to\infty} f_{\text{trash}}/\phi(n) = 0. \tag{B34}$$

We rearrange f_{trash} to obtain

$$f_{\text{trash}} = \frac{1}{n^2} \sum_{c=1}^{n-1} \sum_{d=1}^{n-1} \sum_{\substack{a,b \\ ab=1 \bmod n}} e_n(ac+bd)$$

$$\times \sum_{\alpha=1}^{nx-1} e_n(-\alpha c) \sum_{\beta=1}^{ny-1} e_n(-\beta d). \tag{B35}$$

This new version of f_{trash} contains the *Kloosterman sum* defined as

$$K(c,d;n) \equiv \sum_{\substack{a,b \\ ab=1 \bmod n}} e_n(ac+bd). \tag{B36}$$

Thus,

$$f_{\text{trash}} = \frac{1}{n^2} \sum_{c=1}^{n-1} \sum_{d=1}^{n-1} K(c,d;n) \sum_{\alpha=1}^{nx-1} e_n(-\alpha c)$$

$$\times \sum_{\beta=1}^{ny-1} e_n(-\beta d), \tag{B37}$$

and

$$|f_{\text{trash}}| \leqslant \frac{1}{n^2} \sum_{c=1}^{n-1} \sum_{d=1}^{n-1} |K(c,d;n)| \left| \sum_{\alpha=1}^{nx-1} e_n(-\alpha c) \right|$$

$$\times \left| \sum_{\beta=1}^{ny-1} e_n(-\beta d) \right|. \tag{B38}$$

The last two factors are geometric sums that are easily evaluated.

$$\sum_{\alpha=1}^{nx-1} e_n(-\alpha c) = \frac{1 - e_n([nx]c)}{1 - e_n(c)}, \tag{B39}$$

$$\left| \sum_{\alpha=1}^{nx-1} e_n(-\alpha c) \right| \leqslant \frac{2}{|1 - e_n(c)|} \tag{B40}$$

$$= \frac{2}{2\sin(\pi c/n)} \tag{B41}$$

$$= \frac{1}{\sin(\pi c/n)}. \tag{B42}$$

Our limit on the magnitude of f_{trash} becomes

$$|f_{\text{trash}}| \leqslant \frac{1}{n^2} \sum_{c=1}^{n-1} \sum_{d=1}^{n-1} |K(c,d;n)|$$

$$\times \left| \frac{1}{\sin(\pi c/n)} \right| \left| \frac{1}{\sin(\pi d/n)} \right| \tag{B43}$$

We now use a very nontrivial estimate of Kloosterman sums

$$|K(c,d;n)| \leqslant n^{1/2} \gcd(c,n)^{1/4} \gcd(d,n)^{1/4} \operatorname{div}(n), \tag{B44}$$

where $\gcd(c,n)$ is the greatest common divisor of c and n and $\operatorname{div}(n)$ is the number of divisors of n. For a historical account of this powerful estimate see Ref. 8.

Our bound on f_{trash} becomes

$$|f_{\text{trash}}| \leqslant \frac{n^{1/2}}{n^2} \operatorname{div}(n) \sum_{c=1}^{n-1} \frac{\gcd(c,n)^{1/4}}{\sin(\pi c/n)}$$

$$\times \sum_{d=1}^{n-1} \frac{\gcd(d,n)^{1/4}}{\sin(\pi d/n)}, \tag{B45}$$

$$= n^{1/2} \operatorname{div}(n) (g(n))^2 \tag{B46}$$

where

$$g(n) \equiv \frac{1}{n} \sum_{c=1}^{n-1} \frac{\gcd(c,n)^{1/4}}{\sin(\pi c/n)}. \tag{B47}$$

Now note that $\sin(\pi c/n) = \sin(\pi(n-c)/n)$ and $\gcd(c,n) = \gcd(n-c,n)$ so that

$$g(n) \leqslant \frac{2}{n} \sum_{c=1}^{[n/2]} \frac{\gcd(c,n)^{1/4}}{\sin(\pi c/n)}. \tag{B48}$$

Now for $c \leqslant n/2$ we have $\sin(\pi c/n) \geqslant 2c/n$ so

$$g(n) \leqslant \sum_{c=1}^{[n/2]} \frac{\gcd(c,n)^{1/4}}{c}. \tag{B49}$$

Let $d = \gcd(c,n)$; then d divides c and d divides n. We can then write

$$g(n) \leqslant \sum_{d \mid n} d^{1/4} \sum_{\substack{c<n/2 \\ d \mid c}} (1/c). \tag{B50}$$

Since $d \mid c$ there must be an integer m such that $c = md$ so that

$$\sum_{\substack{c<n/2 \\ d \mid c}} (1/c) = \sum_{m=1}^{[n/2d]} (1/c) \tag{B51}$$

$$\approx (1/d)\ln(n/2d) \tag{B52}$$

$$\leqslant (1/d)\ln(n) - (\ln(2d)) \tag{B53}$$

$$\leqslant \ln(n)/d. \tag{B54}$$

We obtain

$$g(n) \leqslant \sum_{\substack{d \\ d|n}} d^{-3/4} \ln(n) \tag{B55}$$

$$\leqslant \sum_{\substack{d \\ d|n}} \ln(n) \tag{B56}$$

$$= \mathrm{div}(n)\ln(n). \tag{B57}$$

So,

$$|f_{\mathrm{trash}}| \leqslant n^{1/2}(\mathrm{div}(n))^3(\ln(n))^2. \tag{B58}$$

Now $\mathrm{div}(n)$ is of order $\ln(n)$, which for arbitrarily large n is of order n^{ρ} for arbitrarily small ρ. Thus

$$|f_{\mathrm{trash}}|/\phi(n) \leqslant \mathcal{O}(n^{1/2+\rho}/\phi(n)) \tag{B59}$$

$$= \mathcal{O}\left(\prod_i \frac{P_i^{(1/2+\rho)\alpha_i}}{P_i^{\alpha_i - 1}(P_i - 1)}\right) \tag{B60}$$

$$= \mathcal{O}\left(\prod_i \frac{P^{(\rho - 1/2)\alpha_i + 1}}{P_i - 1}\right) \tag{B61}$$

$$\to 0. \tag{B62}$$

Note that with one possible exception all the terms in the product are less than one. So if any one of the prime factors is greater than $P_{\max} = (4\epsilon)^3$ then the entire product is less than ϵ since for large P_i

$$P_i^{(\rho - 1/2)\alpha_i + 1}/P_i - 1 \leqslant 2P_i^{-1/3}. \tag{B63}$$

Likewise if the sum $\alpha_{\mathrm{tot}} = \sum_i \alpha_i$ is greater than $\alpha_{\max} = 3\log_2(2/\epsilon)$ then

$$\prod_i \left(\frac{P_i^{(\rho - 1/2)\alpha_i + 1}}{P_i - 1}\right) \leqslant \prod_i 2^{(\rho - 1/2)\alpha_i + 1} \tag{B64}$$

$$= 2^{(-1/3)\alpha_{\mathrm{tot}} + 1} \tag{B65}$$

$$\leqslant \epsilon. \tag{B66}$$

So if

$$n > P_{\max}^{\alpha_{\max}} \tag{B67}$$

$$= (4/\epsilon)^{9(\log_2(2/\epsilon))} \tag{B68}$$

then

$$|f_{\mathrm{trash}}|/\phi(n) < \epsilon. \tag{B69}$$

[1] P. A. M. Dirac, Proc. R. Soc., London Ser. A **133**, 60 (1931).
[2] C. J. Goebel, in *Quanta*, edited by C. J. Goebel, P. G. O. Freund, and Y. Nambu (University of Chicago, Chicago, 1970), p. 338.
[3] A. S. Goldhaber, in *Magnetic Monopoles*, edited by Richard A. Carrigan, Jr. and W. Peter Trower (Plenum, New York, 1983), p. 1.
[4] G. 't Hooft, Nucl. Phys. B **79**, 276 (1974).
[5] A. M. Polyakov, JETP Lett. **20**, 194 (1974).
[6] A. S. Goldhaber, Phys. Rev. B **140**, 1407 (1965).
[7] D. G. Boulware *et al.*, Phys. Rev. D **20**, 2708 (1976).
[8] R. A. Smith, J. Number Theory **11**, 324 (1979).

Simple calculation of Löwdin's alpha function. III. Procedure for calculating $h_{n,2n-i}(LM|\ell)/(2n-i)!$ successively

Noboru Suzuki

Department of Applied Physics and Chemistry, The University of Electro-Communications, Chofu-Shi, Tokyo 182, Japan

(Received 30 October 1989; accepted for publication 9 May 1990)

When $r/a \approx 0$, Löwdin's α function $(1/r)\alpha_\ell(fLM|a,r)$ is expressed as $(1/r)\alpha_\ell(fLM|a,r) = 2\gamma(LM|\ell)a^{L-1}\Sigma_{n=\ell}^{\infty}\left[\Sigma_{i=\ell+M}^{i_{\max}}[h_{n,2n-i}(LM|\ell)/(2n-i)!]a^{2n-i}g^{(2n-i)}(a)\right](r/a)^{2n-\ell}$, where a is a separation between a new center placed on the origin and the old center located at a given point on the z axis that is the origin of the coordinate system defining a function $f(R)Y_L^M(\Theta,\Phi)$ to be expanded around the new center [$f(R)$ is an arbitrary radial function, and $Y_L^M(\Theta,\Phi)$, a complex spherical harmonics], and r is a distance from the new center. Here, $i_{\max} = \min\{2n, 2(L+\ell)\}$, and the function $g^{(j)}(a)$ is expressed as $g^{(j)}(a) = [(d/dR)^j(f(R)/R^{L-1})]_{R=a}$. The closed form of the coefficients $h_{n,2n-i}(LM|\ell)$ is given by Eq. (27) in J. Math. Phys. **26**, 3193 (1985) (referred to as Part II). Since they are expressed in terms of the different coefficients $b_{Kk}(LM|\ell)$ independent of the form of $f(R)$, $h_{n,2n-i}(LM|\ell)$ constitute a set of universal constants as do $b_{Kk}(LM|\ell)$. The explicit expression for the factor $\gamma(LM|\ell)$ in front of the summation symbol is given by Eq. (2.9) in J. Math. Phys. **25**, 1133 (1984). Hereafter the coefficients expressed by $h_{n,2n-i}(LM|\ell)/(2n-i)!$ are symbolized by $I_{n,2n-i}(LM|\ell)$. Although they are expressed in a complex form with a triple sum, that expression can be reduced to two considerably simple forms by changing the summation indices to others by using the addition theorem for binomial coefficients and the condition for the sum in this theorem to vanish. Moreover, in the two special cases that $i = \ell + M$ and $i = 2(L+\ell)$ or $2(L+\ell)-1$, those two expressions are reduced to two single-term forms, respectively. Introducing into the expression for $I_{n,2n-i}(LM|\ell)$ in terms of $b_{Kk}(LM|\ell)$ the recursion formula for $b_{Kk}(LM|\ell)$ in only M given by Eq. (24) in Part II, leads to the recursion formula for $I_{n,2n-i}(LM|\ell)$ with respect to M. On the other side, the recursion formulas for $I_{n,2n-i}(LM|\ell)$ with $M = L$ and ℓ as to i are obtained through a skillful manipulation. By connecting those recursion formulas for $I_{n,2n-i}(LM|\ell)$ to each other, we can obtain a procedure for calculating $I_{n,2n-i}(LM|\ell)$ successively. As the result of actual performance of the procedure, all the formulas expressing $I_{n,2n-i}(LM|\ell)$ with restriction $0 \leqslant M \leqslant \min\{L,\ell\} \leqslant \max\{L,\ell\} \leqslant 2$ are presented as functions of the parameter n in a table.

I. INTRODUCTION

In the preceding papers[1,2] (hereafter Refs. 1 and 2 will be referred to as Parts I and II, respectively) it was manifested that, when $r/a \approx 0$, Löwdin's α function $(1/r)\alpha_\ell(fLM|a,r)$ is computable by using the formula:

$$(1/r)\alpha_\ell(fLM|a,r) = 2\gamma(LM|\ell)a^{L-1} \times \sum_{n=\ell}^{\infty}\left[\sum_{i=\ell+M}^{i_{\max}} \frac{h_{n,2n-i}(LM|\ell)}{(2n-i)!}a^{2n-i}g^{(2n-i)}(a)\right] \times \left(\frac{r}{a}\right)^{2n-\ell}, \quad (1)$$

where a is a separation between a new center placed on the origin and an old center located at a given point on the z axis that is the origin of the coordinate system defining a function $f(R)Y_L^M(\Theta,\Phi)$ to be expanded around the new center [$f(R)$ is an arbitrary radial function and $Y_L^M(\Theta,\Phi)$ is a complex spherical harmonics], and r is a distance from the new center. Here $i_{\max} = \min\{2n, 2(L+\ell)\}$, and the function $g^{(j)}(a)$ is expressed as

$$g^{(j)}(a) = \left[\left(\frac{d}{dR}\right)^j \frac{f(R)}{R^{L-1}}\right]_{R=a}. \quad (2)$$

The closed form of the coefficients $h_{n,2n-i}(LM|\ell)/(2n-i)!$ is given by dividing Eq. (27) in Part II with $(2n-i)!$. These coefficients are also expressed in terms of the different coefficients $b_{Kk}(LM|\ell)$[1,2] that are independent of the form of $f(R)$. Hence they constitute a set of universal constants as do $b_{Kk}(LM|\ell)$. This is very important and interesting. The explicit expression for the factor $\gamma(LM|\ell)$ in front of the summation symbol is given by Eq. (2.9) in Part I.

The aim of this paper is to think out some easy procedure for evaluating $h_{n,2n-i}(LM|\ell)/(2n-i)!$ directly without computation of $b_{Kk}(LM|\ell)$. In this paper, hereafter, the notation:

 0022-2488/90/092314-13$03.00

$$I_{n,2n-i}(LM|\ell) \equiv h_{n,2n-i}(LM|\ell)/(2n-i)! \tag{3}$$

will be used. Throughout this paper, the restriction $M \geqslant 0$ will be assumed in calculation of $I_{n,2n-i}(LM|\ell)$ because $h_{n,2n-i}(L-M|\ell) = h_{n,2n-i}(LM|\ell)$. In addition, only the cases of $n \geqslant L+\ell$ will be treated in all the following sections, because in the other cases the permissible values of the parameter i are only limited to less than or equal to $2n$.

In Sec. II it will be demonstrated that two considerably simple expressions for $I_{n,2n-i}(LM|\ell)$ can be derived by changing the summation indices to different ones and then utilizing the condition that the sum in Eq. (B1) in Appendix B of part II expressing the addition theorem for binomial coefficients vanishes. Section III will prove that, in the two special cases that i takes the minimum value $\ell+m$ and the maximum value $2(L+\ell)$ or the next maximum value $2(L+\ell)-1$, those expressions are reduced to single-term forms. Section IV will show that the recursion formula for $I_{n,2n-i}(LM|\ell)$ with respect to M as well as i can be derived by introducing the recursion formula for $b_{Kk}(LM|\ell)$ in M alone given by Eq. (24) in Part II, into the expression for $I_{n,2n-i}(LM|\ell)$ in terms of $b_{Kk}(LM|\ell)$. There, it will become necessary to derive some recursion formulas for $I_{n,2n-i}(LM|\ell)$ with the specific values of M as to i, in order to evaluate $I_{n,2n-i}(LM|\ell)$ successively. To this end, Sec. V will manifest that two recursion formulas for $I_{n,2n-i}(LM|\ell)$ with $M=L$ and ℓ concerning i as well as L can be derived by the use of some manipulation. Finally, Sec. VI will present a procedure for computing $I_{n,2n-i}(LM|\ell)$ successively, and as the result of the actual use of this procedure, will exhibit all the formulas expressing $I_{n,2n-i}(LM|\ell)$ with $0 \leqslant M \leqslant \min\{L,\ell\} \leqslant \max\{L,\ell\} \leqslant 2$ as functions of the parameter n in a table.

II. DERIVATION OF TWO SIMPLE EXPRESSIONS FOR $I_{n,2n-i}(LM|\ell)$

We start by writing down the explicit expression[2] for $I_{n,2n-i}(LM|\ell)$ again [see Eq. (27) in Part II]:

$$\begin{aligned} I_{n,2n-i}(LM|\ell) = {} & \frac{(L-M)!(L+M)!(l-M)(\ell+M)!}{2(2n-i)!(L-\frac{1}{2})!(\ell-\frac{1}{2})!} \sum_{k=0}^{[i/2]} \sum_{K=[(i+1)/2]-k}^{L+\ell-k} \sum_{s=\max\{0,L-\ell\}}^{L-M} \frac{(n-k-\frac{1}{2})!(K-\frac{1}{2})!(L-K-\frac{1}{2})!}{(n-k+\frac{1}{2})![i/2]-k)![[(i+1)/2]-k-\frac{1}{2})!} \\ & \times \frac{(\ell-k-\frac{1}{2})!(-L+K+k+s-\frac{1}{2})!}{(K+k-[(i+1)/2])!(K+k-[i/2]-\frac{1}{2})!(L+\ell-K-k)!k!} \\ & \times \frac{1}{s!(L-M-s)!(-L+\ell+s)!(-L+s-\frac{1}{2})!(L+M-s)!}. \end{aligned} \tag{4}$$

The above expression can be reduced to considerably compact forms in the following manner: In the beginning, an equation obtained from the addition theorem for binomial coefficients [see Eq. (B1) in Appendix B of Part II]:

$$\begin{aligned} \frac{(L-K-\frac{1}{2})!}{(L+\ell-K-k)!} = {} & (-\ell+k-\tfrac{1}{2})!(-K-\tfrac{1}{2})!L! \\ & \times \sum_{t=0}^{\min\{L,L+\ell-K-k\}} [t!(L-t)!(L+\ell-K-k-t)!(-L-\ell+k-\tfrac{1}{2}+t)!]^{-1} \end{aligned} \tag{5}$$

is introduced into expression (4), and then the relation between factorials for half-integers, which is given by Eq. (2) in Part II, is used. Following this, extracting the sum over K only, which is hereafter denoted by S_1, we may write it as

$$S_1 = \sum_{K=[(i+1)/2]-k}^{L+\ell-t-k} (-1)^K \frac{(-L+K+k+s-\frac{1}{2})!}{(K+k-[(i+1)/2])!(K+k-[i/2]-\frac{1}{2})!(L+\ell-t-K-k)!}. \tag{6}$$

Replacing $(-L+K+k+s-\frac{1}{2})!$ by $[(-1)^{L-K-k-s}(L-K-k-s-\frac{1}{2})!]^{-1}$ according to the relation expressed by Eq. (2) in Part II and then changing the summation index K to a new index p defined as $p \equiv K+k-[(i+1)/2]$, we arrive at

$$\begin{aligned} S_1 = {} & (-1)^{-L+k+s} \sum_{p=0}^{L+\ell-[(i+1)/2]-t} \left[p!\left(L-\left[\frac{i+1}{2}\right]-s-\tfrac{1}{2}-p\right)!\left(L+\ell-\left[\frac{i+1}{2}\right]-t-p\right)!\right. \\ & \left.\times\left(\left[\frac{i+1}{2}\right]-\left[\frac{i}{2}\right]-\frac{1}{2}+p\right)!\right]^{-1}. \end{aligned} \tag{6'}$$

The result of summation over p in Eq. (6') is obtained directly by setting

$$\mu = L-\left[\frac{i+1}{2}\right]-s-\frac{1}{2}, \quad \rho = L+\ell-\left[\frac{i+1}{2}\right]-t, \quad \nu-\rho = \left[\frac{i+1}{2}\right]-\left[\frac{i}{2}\right]-\frac{1}{2},$$

and $\kappa = p$ in Eq. (B1) of Part II. Then S_1 becomes

$$S_1 = \begin{cases} (-1)^{-L+k+s} \dfrac{(2L+\ell-i-s-t-1)!}{(L-[(i+1)/2]-s-\frac{1}{2})!(L+\ell-[i/2]-t-\frac{1}{2})!(L+\ell-[(i+1)/2]-t)!(L-[i/2]-s-1)!}, \\ \qquad\qquad \text{for } L-[i/2]-s-1 \geqslant 0 \\ 0, \qquad\qquad \text{for } 2L+\ell-i-s-t-1 \geqslant 0 \text{ and } L-[i/2]-s-1<0 \\ (-1)^{\ell-[(i+1)/2]+k+s-t} \dfrac{(-L+[i/2]+s)!}{(L-[(i+1)/2]-s-\frac{1}{2})!(L+\ell-[i/2]-t-\frac{1}{2})!(L+\ell-[(i+1)/2]-t)!(-2L-\ell+i+s+t)!}, \\ \qquad\qquad \text{for } 2L+\ell-i-s-t-1<0. \end{cases} \tag{6''}$$

Here it should be noted that the relation $L-[i/2]-s-1 \leqslant 2L+\ell-i-s-t-1$ holds, independent of the value of t, because $L+\ell-[(i+1)/2]-t \geqslant 0$. The middle result in the above summation is derived from the fact that $2L+\ell-i-s-t-1 \geqslant 0$ and $L-[i/2]-s-1<0$ just agrees with the condition for the sums in Eq. (B1) of Part II to vanish.

Now we take the case of $L-[i/2]-s-1 \geqslant 0$ into consideration. Introducing the upper result of Eq. (6'') into expression (4) and then drawing out the sum over t only, which is hereafter denoted by T, we may write it as

$$T = \sum_{t=0}^{t_{\max}} \frac{(2L+\ell-i-s-t-1)!}{t!(L-t)!(L+\ell-[i/2]-\frac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell-\frac{1}{2}+k+t)!}, \tag{7}$$

with $t_{\max} = \min\{L, L+\ell-[(i+1)/2]\}$. Here note that the movable range of s is from $\max\{0, L-\ell\}$ to $\min\{L-M, L-[i/2]-1\}$. Therefore, $i \leqslant 2\times\min\{L,\ell\}-1$, and thus $L+\ell-[(i+1)/2] \geqslant \max\{L,\ell\}$. As the result, $t_{\max}$ must be L. Inserting an equation derived from Eq. (B1) of Part II:

$$\begin{aligned} &\frac{(2L+\ell-i-1-s-t)!}{(L+\ell-[i/2]-\frac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!} \\ &\quad = \left(L-\left[\frac{i+1}{2}\right]-\frac{1}{2}-s\right)!\left(L-\left[\frac{i}{2}\right]-1-s\right)! \\ &\qquad \times \sum_{q=0}^{L-[i/2]-s-1} \left[q!\left(L-\left[\frac{i+1}{2}\right]-\frac{1}{2}-s-q\right)!\left(L-\left[\frac{i}{2}\right]-1-s-q\right)!\left(\ell+\frac{1}{2}+s-t+q\right)!\right]^{-1} \end{aligned} \tag{8}$$

into sum (7) yields

$$\begin{aligned} T = &\left(L-\left[\frac{i+1}{2}\right]-\frac{1}{2}-s\right)!\left(L-\left[\frac{i}{2}\right]-1-s\right)! \\ &\times \sum_{q=0}^{L-[i/2]-1-s}\left[q!\left(L-\left[\frac{i+1}{2}\right]-\frac{1}{2}-s-q\right)!\left(L-\left[\frac{i}{2}\right]-1-s-q\right)!\right]^{-1} \\ &\times \sum_{t=0}^{L}\left[t!\left(\ell+\frac{1}{2}+s+q-t\right)!(L-t)!\left(-L-\ell-\frac{1}{2}+k+t\right)!\right]^{-1}. \end{aligned} \tag{7'}$$

Here, if $\mu, \rho, \nu-\rho$, and κ in Eq. (B1) of Part II are set equal to $\ell+\frac{1}{2}+s+q, L, -L-\ell-\frac{1}{2}+k$ and t, respectively, then $\mu+\nu = k+s+q \geqslant 0$ because $k,s,q \geqslant 0$, while $\mu+\nu-\rho = -L+k+s+q \leqslant k-[i/2]-1<0$ because $k \leqslant [i/2]$. Therefore, the inner sum over t vanishes, and so does T.

From the above consideration as well as the middle result of Eq. (6''), it can be understood that, if $2L+\ell-i-s-t-1 \geqslant 0$ holds for all sets of permissible values of s and t, $I_{n,2n-i}(LM\,|\,\ell)$ vanishes. This condition is just $i<\ell+M$ since $s+t \leqslant 2L-M$. This was already manifested in a different manner in Part II.

Next we proceed to the case of $2L+\ell-i-s-t-1<0$. Introducing the lower result of Eq. (6'') into Eq. (4) and then extracting the sum over k only, which is hereafter denoted by S_2, we may write it down as

$$S_2 = \sum_{k=0}^{[i/2]} \frac{(n-k-\frac{1}{2})!}{k!(n-k+\frac{1}{2})!([i/2]-k)!([(i+1)/2]-k-\frac{1}{2})!(-L-\ell-\frac{1}{2}+k+t)!}. \tag{9}$$

Inserting an equation derived from Eq. (B1) of Part II:

$$\frac{(n-k-\frac{1}{2})!}{(n-k+\frac{1}{2})!([i/2]-k)!} = \left(n-\left[\frac{i}{2}\right]-\frac{1}{2}\right)!\sum_{p=0}^{[i/2]-k}\frac{(-1)^p}{([i/2]-k-p)!(n-[i/2]+\frac{1}{2}+p)!} \tag{10}$$

into sum (9) leads to

$$\begin{aligned} S_2 = &\left(n-\left[\frac{i}{2}\right]-\frac{1}{2}\right)!\sum_{p=0}^{[i/2]}\frac{(-1)^p}{(n-[i/2]+\frac{1}{2}+p)!} \\ &\times \sum_{k=0}^{[i/2]-p}\left[k!\left(\left[\frac{i+1}{2}\right]-\frac{1}{2}-k\right)!\left(\left[\frac{i}{2}\right]-p-k\right)!\left(-L-\ell-\frac{1}{2}+t+k\right)!\right]^{-1}. \end{aligned} \tag{9'}$$

The sum over k is obtained easily by replacing μ, ρ, $\nu-\rho$, and κ in Eq. (B1) of Part II with $[(i+1)/2]-\frac{1}{2}$, $[i/2]-p$, $-L-\ell-\frac{1}{2}+t$ and k, respectively. Then, $\mu+\nu-\rho=-L-\ell+[(i+1)/2]-1+t<0$ for any allowable value of t because $t\leqslant L+\ell-[(i+1)/2]$. Therefore, if $\mu+\nu=-L-\ell+i-1+t-p\geqslant 0$, the sum over k vanishes. Otherwise, i.e., in the case of $p\geqslant -L-\ell+i+t$, it takes a nonzero value, which is given by

$$(-1)^{[i/2]-p}\frac{(L+\ell-[(i+1)/2]-t)!}{([(i+1)/2]-\frac{1}{2})!(-L-\ell+[i/2]+t-p-\frac{1}{2})!([i/2]-p)!(L+\ell-i-t+p)!}. \tag{11}$$

Here note that $p\geqslant -L-\ell+i+t\geqslant L-s\geqslant M\geqslant 0$ because $t\geqslant 2L+\ell-i-s$, and therefore the minimum value of p must be $-L-\ell+i+t$. Thus S_2 may be rewritten as

$$S_2=(-1)^{[i/2]}\frac{(n-[i/2]-\frac{1}{2})!(L+\ell-[(i+1)/2]-t)!}{([(i+1)/2]-\frac{1}{2})!}\sum_{p=-L-\ell+i+t}^{[i/2]}\left[\left(\left[\frac{i}{2}-p!\right)\right]\right.$$

$$\left.\times(L+\ell-i-t+p)!\left(n-\left[\frac{i}{2}\right]+\frac{1}{2}+p\right)!\left(-L-\ell+\left[\frac{i}{2}\right]-\frac{1}{2}+t-p\right)!\right]^{-1}. \tag{9''}$$

Replacing $[i/2]-p$ by q in the above sum transforms the sum into

$$\sum_{q=0}^{L+\ell-[(i+1)/2]-t}\left[q!\left(L+\ell-\left[\frac{i+1}{2}\right]-t-q\right)!\left(n+\frac{1}{2}-q\right)!\left(-L-\ell-\frac{1}{2}+t+q\right)!\right]^{-1}. \tag{12}$$

Here, if μ, ρ, $\nu-\rho$, and κ are set equal to $n+\frac{1}{2}$, $L+\ell-[(i+1)/2]-t$, $-L-\ell-\frac{1}{2}+t$ and q, respectively, then $\mu+\nu=n-[(i+1)/2]\geqslant 0$ since $i_{\max}=\min\{2n,2(L+\ell)\}$. Accordingly, if $\mu+\nu-\rho=-L-\ell+n+t<0$, sum (12) vanishes. Otherwise, i.e., when $-L-\ell+n+t\geqslant 0$, the sum amounts to

$$\frac{(n-[(i+1)/2])!}{(n+\frac{1}{2})!(-[(i+1)/2]-\frac{1}{2})!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!}. \tag{12'}$$

Thus, in this case S_2 is reduced to

$$S_2=(-1)^i\frac{2^{i+1}(2n-i)!n!}{(2n+1)!(-L-\ell+n+t)!}. \tag{9'''}$$

Here the relation between factorials for half-integers, and the two identities,

$$(n-[i/2]-\tfrac{1}{2})!(n-[(i+1)/2])!=(2n-i)!/2^{2n-i}\quad\text{and}\quad(n+\tfrac{1}{2})!=(2n+1)!/(2^{2n+1}n!),$$

have been used.

If the condition $n-L-\ell+t<0$ is fulfilled for all permissible values of t, S_2 vanishes, and so does $I_{n,2n-i}(LM|\ell)$. The condition is just $n-\ell<0$ because $t\leqslant L$. This was already proved in a different manner in Part II.

From the above discussion, the minimum and maximum values of t, $t_{\min}$ and $t_{\max}$ are given, respectively, by $\max\{0, 2L+\ell-i-s, L+\ell-n\}$ and $\min\{L, L+\ell-[(i+1)/2]\}$. Therefore, if $i\leqslant 2\ell$, $t_{\max}=L$ and thus $2L+\ell-i-s\leqslant L$, i.e., $s\geqslant L+\ell-i$. Otherwise, $t_{\max}=L+\ell-[(i+1)/2]$, and hence $2L+\ell-i-s\leqslant L+\ell-[(i+1)/2]$, i.e., $s\geqslant L-[i/2]$. On the other side, $L-[i/2]\leqslant L-\ell$ because $i>2\ell$. From this consideration, it can be seen that the minimum value of s, $s_{\min}$, is given by $\max\{0, L-\ell, L+\ell-i\}$.

Finally, introducing sum (9''') into expression (4), one obtains a simple expression for $I_{n,2n-i}(LM|\ell)$ that includes only one double sum:

$$I_{n,2n-i}(LM|\ell)=(-1)^{[i/2]}\frac{2^i(L-M)!(L+M)!(\ell-M)!(\ell+M)!L!n!}{(L-\frac{1}{2})!(\ell-\frac{1}{2})!(2n+1)!}$$

$$\times\sum_{s=s_{\min}}^{L-M}\sum_{t=t_{\min}}^{t_{\max}}(-1)^{s+t}\left(-L+\left[\frac{i}{2}\right]+s\right)!\Big/\left[\left(L-\left[\frac{i+1}{2}\right]-\frac{1}{2}-s\right)!s!(L-M-s)!\right.$$

$$\times(-L+\ell+s)!\left(-L-\frac{1}{2}+s\right)!(L+M-s)!t!(L-t)!\left(L+\ell-\left[\frac{i}{2}\right]-\frac{1}{2}-t\right)!$$

$$\left.\times\left(L+\ell-\left[\frac{i+1}{2}\right]-t\right)!(-L-\ell+n+t)!(-2L-\ell+i+s+t)!\right]. \tag{13}$$

Furthermore, substituting Eq. (6') for the lower result in Eq. (6'') on Eq. (13) and then performing the summation over t yields a slightly simpler expression than expression (13):

$$I_{n,2n-i}(LM|\ell)=(-1)^{L+\ell-i}\frac{2^i(L-M)!(L+M)!(\ell-M)!(\ell+M)!n!}{(L-\frac{1}{2})!(\ell-\frac{1}{2})!(2n+1)!(n-\ell)!}$$

$$\times\sum_{s=s_{\min}}^{L-M}\sum_{p=0}^{\min\{L+\ell-[(i+1)/2],n-[(i+1)/2]\}}(-1)^s$$

$$\times\left(n+L-\left[\frac{i+1}{2}\right]-p\right)!\Big/\,[s!(L-M-s)!(-L+\ell+s)!$$
$$\times\left(-L-\frac{1}{2}+s\right)!(L+M-s)!p!\left(L-\left[\frac{i+1}{2}\right]-\frac{1}{2}-s-p\right)!\left(\left[\frac{i+1}{2}\right]-\left[\frac{i}{2}\right]-\frac{1}{2}+p\right)!$$
$$\times\left(n-\left[\frac{i+1}{2}\right]-p\right)!\left(L+\ell-\left[\frac{i+1}{2}\right]-p\right)!\Big]. \tag{14}$$

It seems impossible to obtain any more compact expression for $I_{n,2n-i}(LM|\ell)$ than expression (14).

III. SINGLE-TERM EXPRESSIONS FOR $I_{n,2n-i}(LM|\ell)$ IN TWO SPECIAL CASES OF $i=\ell+M$ AND $i=2(L+\ell)$ OR $2(L+\ell)-1$

First, we take into account the special case that i takes the minimum value $\ell+M$. In this case the permissible values of s and t in expression (13) are, respectively, $L-M$ and L only, and thus expression (13) is reduced to a single-term form:

$$I_{n,2n-(\ell+M)}(LM|\ell)=(-1)^{\ell}\frac{2^{\ell-M}(L+M)!(\ell+M)!n!}{(L-\frac{1}{2})!(\ell-\frac{1}{2})!(2n+1)!M!(n-\ell)!}. \tag{15}$$

Here, the following two identities have been utilized:

$$\frac{(-M+[(\ell+M)/2])!}{(M-[(\ell+M+1)/2]-\frac{1}{2})!(\ell-[(\ell+M)/2]-\frac{1}{2})!(\ell-[(\ell+M+1)/2])!}=(-1)^{-M+[(\ell+M+1)/2]}, \tag{16}$$

and

$$(2M)!(-M-\tfrac{1}{2})!=(-1)^{M}2^{2M}M!. \tag{17}$$

Thus, the respective recursion formulas with respect to L, ℓ, and M are obtained readily from Eq. (15):

$$I_{n,2n-(\ell+M)}(L-1M|\ell)=[(L-\tfrac{1}{2})/(L+M)]\cdot I_{n,2n-(\ell+M)}(LM|\ell), \tag{18}$$

$$I_{n,2n-(\ell-1+M)}(LM|\ell-1)=-[(\ell-\tfrac{1}{2})/2(\ell+M)(n-\ell+1)]\cdot I_{n,2n-(\ell+M)}(LM|\ell), \tag{19}$$

and

$$I_{n,2n-(\ell+M-1)}(LM-1|\ell)=[2M/(L+M)(\ell+M)]\cdot I_{n,2n-(\ell+M)}(LM|\ell). \tag{20}$$

When M takes some special values, the above three relations are simplified: For example, when $M=L-1$, Eq. (18) is reduced to

$$I_{n,2n-(L+\ell-1)}(L-1L-1|\ell)=\tfrac{1}{2}\cdot I_{n,2n-(L+\ell-1)}(LL-1|\ell); \tag{21}$$

and for $M=L$, Eq. (20) is simplified into

$$I_{n,2n-(L+\ell-1)}(LL-1|\ell)=[1/(L+\ell)]\cdot I_{n,2n-(L+\ell)}(LL|\ell). \tag{22}$$

Combining Eq. (21) with Eq. (22) yields the relation:

$$I_{n,2n-(L+\ell-1)}(L-1L-1|\ell)=[1/2(L+\ell)]\cdot I_{n,2n-(L+\ell)}(LL|\ell). \tag{23}$$

On the other side, when $M=\ell-1$, Eq. (19) is reduced to

$$I_{n,2n-(2\ell-2)}(L\ell-1|\ell-1)=-[1/2^2(n-\ell+1)]\cdot I_{n,2n-(2\ell-1)}(L\ell-1|\ell). \tag{24}$$

Also, setting $M=\ell$ in Eq. (20) leads to the same relation as Eq. (22):

$$I_{n,2n-(2\ell-1)}(L\ell-1|\ell)=[1/(L+\ell)]\cdot I_{n,2n-2\ell}(L\ell|\ell). \tag{25}$$

Second, assuming that $L+\ell>0$, we deal with the case that i takes the first or second maximum value, i.e., $2(L+\ell)$ or $2(L+\ell)-1$. Then the allowable value of p in expression (14) is zero only. Thus, expression (14) is simplified into

$$I_{n,2n-i}(LM|\ell)=(-1)^{L+\ell-i}\frac{2^{i}(L-M)!(L+M)!(\ell-M)!(\ell+M)!n!}{(L-\frac{1}{2})!(\ell-\frac{1}{2})!(2n+1)!(L+\ell-[i/2]-\frac{1}{2})!(n-L-\ell)!}$$
$$\times\sum_{s=s_{\min}}^{L-M}\frac{(-1)^{s}}{s!(L-M-s)!(-L+\ell+s)!(-L-\frac{1}{2}+s)!(L+M-s)!(-\ell-\frac{1}{2}-s)!}, \tag{26}$$

where $s_{\min}=\max\{0,L-\ell\}$. The sum over s in Eq. (26) is obtained directly by replacing $(-\ell-\frac{1}{2}-s)!$ with $(-1)^{-\ell-s}/(\ell-\frac{1}{2}+s)!$ and then setting a, b, α, β, and γ in Eq. (A1) of Appendix A equal to $L-M, L+M, -L+\ell, -L$ and ℓ, respectively, because in this case, the equation $a+b+\alpha=\gamma-\beta$ holds, and the condition that $\gamma-\beta$ takes a non-negative integer is satisfied. This sum may be written as

$$(-1)^{\ell}\frac{(L+\ell)!(L-\frac{1}{2})!(\ell-\frac{1}{2})!}{(L-M)!(L+M)!(\ell-M)!(\ell+M)!(-M-\frac{1}{2})!(M-\frac{1}{2})!}. \tag{27}$$

Thus we reach the single-term expression for $I_{n,2n-i}(LM|\ell)$ with $i=2(L+\ell)$ or $2(L+\ell)-1$:

$$I_{n,2n-i}(\mathrm{LM}|\ell)=(-1)^{L+M-i}2^{2(L+\ell)}\frac{(L+\ell)!n!}{(2n+1)!(n-L-\ell)!}. \tag{28}$$

Here the relation: $(-M-\frac{1}{2})!(M-\frac{1}{2})!=(-1)^{-M}$ and the fact that the equality $2^i/(L+\ell-[i/2]-\frac{1}{2})!=2^{2(L+\ell)}$ holds for $i=2(L+\ell)$ or $2(L+\ell)-1$ have been used.

From Eq. (28) the relation between $I_{n,2n-i}(LM|\ell)$ with $i=2(L+\ell)$ and with $i=2(L+\ell)-1$ is derived immediately:

$$I_{n,2n-2(L+\ell)+1}(LM|\ell)=-I_{n,2n-2(L+\ell)}(LM|\ell). \tag{29}$$

On the other hand, the respective recursion formulas for $I_{n,2n-i}(LM|\ell)$ with $i=2(L+\ell)$ or $2(L+\ell)-1$ as to ℓ or L and M are obtained as

$$\begin{aligned}I_{n,2n-(i+2)}(LM|\ell+1)&=2^2(L+\ell+1)(n-L-\ell)I_{n,2n-i}(LM|\ell)\\&=-I_{n,2n-(i+2)}(L+1\,M|\ell),\end{aligned} \tag{30}$$

and

$$I_{n,2n-i}(L\,M+1|\ell)=-I_{n,2n-i}(L\,M|\ell)=I_{n,2n-i}(LM-1|\ell). \tag{31}$$

Combining Eq. (29) with Eq. (31) for $i=2(L+\ell)$ leads to the relation:

$$\begin{aligned}I_{n,2n-2(L+\ell)}(L\,M+1|\ell)&=I_{n,2n-2(L+\ell)+1}(L\,M|\ell)\\&=I_{n,2n-2(L+\ell)}(L\,M-1|\ell).\end{aligned} \tag{32}$$

One of our purposes in this paper is to express $I_{n,2n-i}(LM|\ell)$ with the specific values of L, M, ℓ, and i as a function of the parameter n. Therefore, in both the cases we have not written down the recursion formulas with respect to n only, although those formulas can be obtained without difficulty.

For these two special cases the five recursion formulas, Eqs. (18), (19), (20), (30), and (31) and the seven relations, Eqs. (21)–(25), (29), and (32), are very useful for calculating $I_{n,2n-i}(LM|\ell)$.

IV. RECURSION FORMULA FOR $I_{n,2n-i}(L\,M|\ell)$ WITH RESPECT TO M

We begin to write down the expression for $I_{n,2n-i}(LM|\ell)$ in terms of $b_{Kk}(LM|\ell)$ [see Eq. (5.6) in Part I]:

$$I_{n,2n-i}(LM|\ell)=\frac{1}{(2n-i)!}\sum_{k=0}^{[i/2]}\sum_{K=[(i+1)/2]-k}^{L+\ell-k}\frac{(2K)!}{(2n-2k+1)(i-2k)!(2K+2k-i)!}b_{Kk}(LM|\ell). \tag{33}$$

Introducing into expression (33) the recursion formula for $b_{Kk}(L\,M|\ell)$ in M only which is given by Eq. (24) in Part II yields

$$\begin{aligned}I_{n,2n-i}(LM-1|\ell)=\frac{1}{(L+M)(\ell+M)}\Bigg[&\frac{2M}{(2n-i)!}\sum_{k=0}^{[i/2]}\sum_{K=[(i+1)/2]-k}^{L+\ell-k}\frac{(L+\ell-2K-2k)(2K)!}{(2n-2k+1)(i-2k)!(2K+2k-i)!}\\&\times b_{Kk}(L\,M|\ell)+(L-M)(\ell-M)I_{n,2n-i}(L\,M+1|\ell)\Bigg].\end{aligned} \tag{34}$$

Inserting the identity:

$$L+\ell-2K-2k=-(2K+2k-i)\frac{(2n-2k+1)-(2n-i)}{i+1-2k}+(L+\ell-i) \tag{35}$$

into each term in the sum of Eq. (34) transforms the first term in the brackets of Eq. (34) into

$$\begin{aligned}2M\Bigg[&-\frac{1}{(2n-i)!}\sum_{k=0}^{[i/2]}\sum_{K=[i/2]+1-k}^{L+\ell-k}\frac{(2K)!}{(i+1-2k)!(2K+2k-i-1)!}b_{Kk}(LM|\ell)+\frac{1}{(2n-i-1)!}\\&\times\sum_{k=0}^{[i/2]}\sum_{K=[i/2]+1-k}^{L+\ell-k}\frac{(2K)!}{(2n-2k+1)(i+1-2k)!(2K+2k-i-1)!}b_{Kk}(LM|\ell)+(L+\ell-i)I_{n,2n-i}(LM|\ell)\Bigg].\end{aligned} \tag{36}$$

Here note that the minimum value of K may increase to $[i/2]+1-k$ for an even i, because the term with $K=[i/2]+1-k$ in the first sum compensates for the corresponding term in the second sum. The use of the notation:

$$J_{n,2n-i}(LM|\ell) \equiv \frac{1}{(2n-i+1)!}\sum_{k=0}^{[i/2]}\sum_{K=[(i+1)/2]-k}^{L+\ell-k}\frac{(2K)!}{(i-2k)!(2K+2k-i)!}b_{Kk}(LM|\ell) \tag{37}$$

leads to the following expression for the first term in the brackets of formula (36):

$$\begin{cases} -J_{n,2n-(i+1)}(LM|\ell)+\dfrac{1}{(2n-i)!}\displaystyle\sum_{K=0}^{L+\ell-(i+1)/2} b_{K(i+1)/2}(LM|\ell), & \text{if } i \text{ is odd}, \\ -J_{n,2n-(i+1)}(LM|\ell), & \text{otherwise}. \end{cases} \tag{38}$$

On the other side, the second term is expressed as

$$\begin{cases} I_{n,2n-(i+1)}(LM|\ell)-\dfrac{1}{(2n-i)!}\displaystyle\sum_{K=0}^{L+\ell-(i+1)/2} b_{K(i+1)/2}(LM|\ell), & \text{if } i \text{ is odd}, \\ I_{n,2n-(i+1)}(LM|\ell), & \text{otherwise}. \end{cases} \tag{39}$$

Hence formula (36) may be rewritten, regardless of whether i is odd or not, as

$$2M[-J_{n,2n-(i+1)}+I_{n,2n-(i+1)}(LM|\ell)+(L+\ell-i)I_{n,2n-i}(LM|\ell)]. \tag{36'}$$

As proved in Appendix B, $J_{n,2n-i}(LM|\ell)$ vanishes as far as M takes any nonzero value. Therefore, the product $2M[-J_{n,2n-(i+1)}(LM|\ell)]$ disappears for any value of M.

Ultimately, Eq. (34) is reduced to

$$\begin{aligned} I_{n,2n-i}(LM-1|\ell) = [(L+M)(\ell+M]^{-1}\{&2M[I_{n,2n-(i+1)}(LM|\ell)]+(L+\ell-i)I_{n,2n-i}(LM|\ell) \\ &+(L-M)(\ell-M)I_{n,2n-i}(LM+1|\ell)\}. \end{aligned} \tag{40}$$

Equation (40) is no other than a recursion formula for $I_{n,2n-i}(LM|\ell)$ with respect to M as well as i. Especially, when M is equal to L or ℓ, Eq. (40) is slightly simplified into

$$I_{n,2n-i}(LM-1|\ell) = (L+\ell)^{-1}[I_{n,2n-(i+1)}(LM|\ell)+(L+\ell-i)I_{n,2n-i}(LM|\ell)]. \tag{41}$$

Here it should be noted that the relation for the special case of $i=2(L+\ell)$ given by Eq. (31) can be derived independently by first using Eq. (41) for $M=L$ or ℓ and then utilizing Eq. (40) successively in descending order of the value of M. On the other side, setting i equal to the minimum value $\ell+M-1$ in Eq. (40) reproduces Eq. (20) immediately, because $I_{n,2n-(\ell+M-1)}(LM|\ell)$ and $I_{n,2n-(\ell+M-1)}(LM+1|\ell)$ both disappear.

The two recursion formulas Eqs. (40) and (41) are very useful for calculating $I_{n,2n-i}(LM|\ell)$ with $M<L$ or ℓ from the values of $I_{n,2n-i}(LL|\ell)$ and $I_{n,2n-i}(L\ell|\ell)$. Therefore, it becomes necessary to think out recursion formulas for $I_{n,2n-i}(LL|\ell)$ and $I_{n,2n-i}(L\ell|\ell)$ with respect to i.

V. RECURSION FORMULAS FOR $I_{n,2n-i}(LL|\ell)$ AND $I_{n,2n-i}(L\ell|\ell)$ WITH RESPECT TO i

In this section we derive the recursion formulas for $I_{n,2n-i}(LL|\ell)$ and $I_{n,2n-i}(L\ell|\ell)$ with respect to i in the following manner. At the outset, we take into consideration how the two sums appearing in expression (13) for $I_{n,2n-i}(LM-1|\ell)$ with $M=L$ or ℓ can be expressed in terms of some $I_{n,2n-i}(LM|\ell)$ with $M=L$ or ℓ that are different from each other in the value of i only. Then, if such expressions are obtained, the whole of them can be set equal to the right-hand side of Eq. (41) for $M=L$ or ℓ. Thus the recursion formula for $I_{n,2n-i}(LM|\ell)$ with $M=L$ or ℓ as to i can be derived.

First, we treat the case of $M=L$. The closed expression for $I_{n,2n-i}(LL|\ell)$ is derived from expression (13) by putting $M=L$ in it:

$$\begin{aligned} &I_{n,2n-i}(LL|\ell) \\ &\quad = (-1)^i\cdot 2^{2L}\cdot\frac{(L+\ell)!L!(i-2L)!n!}{(\ell-\tfrac{1}{2})!(2n+1)!} \\ &\qquad \times\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-[i/2]-\tfrac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!(-2L-\ell+i+t)!}, \end{aligned} \tag{42}$$

and also from expression (13) is done that for $I_{n,2n-i}(LL-1|\ell)$:

$$\begin{aligned} &I_{n,2n-i}(LL-1|\ell) \\ &\quad = (1-\delta_{i,L+\ell-1})(-1)^i\cdot 2^{2L}\cdot\frac{(-L+\ell+1)(L+\ell-1)!L!(i-2L)!n!}{(\ell-\tfrac{1}{2})!(2n+1)!} \end{aligned}$$

$$\times\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-[i/2]-\frac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!(-2L-\ell+i+t)!}$$
$$+(-1)^{i+1}\cdot 2^{2L-1}\cdot\frac{(L+\ell-1)!L!(i-2L+2)!n!}{(\ell-\frac{1}{2})!(2n+1)!}$$
$$\times\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-[i/2]-\frac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!(-2L-\ell+i+1+t)!}. \tag{43}$$

By comparing the first term of Eq. (43) with Eq. (42), it can be seen immediately that the term is expressed as

$$[(-L+\ell+1)/(L+\ell)]\cdot I_{n,2n-i}(LL|\ell). \tag{44}$$

Needless to say, $I_{n,2n-i}(LL|\ell)$ disappears when $i=L+\ell-1$, because i must be larger than or equal to $L+\ell$. If each term in the second sum of Eq. (43) is multiplied by the identity: $1\equiv(L+\ell-i/2-t)/(L+\ell-i/2)+t/(L+\ell-i/2)$, this sum is divided into two terms. Then, replacing the summation index t by $t+1$ in the second resulting term yields an alternative expression for the second sum of Eq. (43):

$$\frac{1}{L+\ell-i/2}\left\{\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-[(i+1)/2]-\frac{1}{2}-t)!(L+\ell-[i/2]-1-t)!(-L-\ell+n+t)!(-2L-\ell+i+1+t)!}\right.$$
$$\left.-\sum_t \frac{(-1)^t}{t!(L-1-t)!(L-1+\ell-[i/2]-\frac{1}{2}-t)!(L-1+\ell-[(i+1)/2]-t)!(-L+1-\ell+n+t)!(-2L+2-\ell+i+t)!}\right\}. \tag{45}$$

It can be seen readily that the first sum in the above formula agrees with the sum in the expression for $I_{n,2n-(i+1)}(LL|\ell)$, while the second sum agrees with the sum in the expression for $I_{n,2n-i}(L-1L-1|\ell)$. Finally, comparing the factor in front of the summation symbol of the second term of Eq. (43) with that in the expression for $I_{n,2n-(i+1)}(LL|\ell)$ and with that in the expression for $I_{n,2n-i}(L-1L-1|\ell)$, we obtain an expression for the whole of the second term of Eq. (43) in terms of some $I_{n,2n-i}(LL|\ell)$ with different values of i and L:

$$\frac{1}{2(L+\ell)-i}\left[\frac{i-2L+2}{L+\ell}\cdot I_{n,2n-(i+1)}(LL|\ell)+2^2\cdot L\cdot I_{n,2n-i}(L-1L-1|\ell)\right]. \tag{46}$$

Eventually, Eq. (43) is reduced to

$$I_{n,2n-i}(LL-1|\ell)=\frac{1}{2(L+\ell)-i}\left[\frac{i-2L+2}{L+\ell}\cdot I_{n,2n-(i+1)}(LL|\ell)+2^2\cdot L\cdot I_{n,2n-i}(L-1L-1|\ell)\right]$$
$$+\frac{-L+\ell+1}{L+\ell}\cdot I_{n,2n-i}(LL|\ell). \tag{47}$$

Equating the right-hand side of Eq. (47) with that of Eq. (41) for $M=L$ yields

$$I_{n,2n-i}(LL|\ell)=\frac{1}{[2(L+\ell)-i](i-2L+1)}$$
$$\times[2(2L+\ell-i-1)\cdot I_{n,2n-(i+1)}(LL|\ell)-2^2\cdot L\cdot(L+\ell)\cdot I_{n,2n-i}(L-1L-1|\ell)]. \tag{48}$$

Equation (48) is nothing but a recursion formula for $I_{n,2n-i}(LL|\ell)$ with respect to i as well as L. Here it should be noted that setting $i=L+\ell-1$ in Eq. (48) produces Eq. (23), and then introducing this result into Eq. (47) yields Eq. (22), while putting $i=2(L+\ell)-1$ in Eq. (48) leads to Eq. (29) with $M=L$, and then inserting this result into Eq. (47) yields Eq. (32) with $M=L-1$.

Second, we deal with the case of $M=\ell$. The closed expression for $I_{n,2n-i}(L\ell|\ell)$ is obtained from expression (13) by putting $M=\ell$ in it:

$$I_{n,2n-i}(L\ell|\ell)$$
$$=(-1)^{L-\ell+i}\cdot\frac{2^{2\ell}\cdot(L+\ell)!L!(i-2\ell)!n!}{(L-\frac{1}{2})!(2n+1)!}$$
$$\times\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-[i/2]-\frac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!(-L-2\ell+i+t)!}, \tag{49}$$

and also from expression (13) is done that for $I_{n,2n-i}(L\ell-1|\ell)$:

$$
\begin{aligned}
&I_{n,2n-i}(L\,\ell-1|\ell)\\
&\quad=(-1)^{L-\ell+i}\cdot\frac{2^{2\ell}(L-\ell+1)(L+\ell-1)!L!(i-2\ell)!n!}{(L-\tfrac{1}{2})!(2n+1)!}\\
&\qquad\times\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-[i/2]-\tfrac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!(-L-2\ell+i+t)!}\\
&\qquad+(-1)^{L-\ell+i+1}\cdot\frac{2^{2\ell-1}(L+\ell-1)!L!(i-2\ell+2)!n!}{(L-\tfrac{1}{2})!(2n+1)!}\\
&\qquad\times\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-[i/2]-\tfrac{1}{2}-t)!(L+\ell-[(i+1)/2]-t)!(-L-\ell+n+t)!(-L-2\ell+i+1+t)!}.
\end{aligned}
\tag{50}
$$

From comparison of the first term of Eq. (50) with Eq. (49), it can be seen easily that the term is expressed by

$$
[(L-\ell+1)/(L+\ell)]\cdot I_{n,2n-i}(L\,\ell|\ell). \tag{51}
$$

In contrast to the first term, it is not so easy to derive an expression for the second term of Eq. (50) in terms of some $I_{n,2n-i}(L\ell|\ell)$ with different values of i and L. We start with considering the case that i takes an even non-negative integer. In this case i may be set equal to $2m$, where m is an arbitrary non-negative integer. Then the sum over t in the second term of Eq. (50) may be written as

$$
\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-m-\tfrac{1}{2}-t)!(L+\ell-m-t)!(-L-\ell+n+t)!(-L-2\ell+2m+1+t)!}, \tag{52}
$$

while the sum over t in the expression for $I_{n,2n-i}(L\,\ell|\ell)$ with $i=2m+1$, as

$$
\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-m-\tfrac{1}{2}-t)!(L+\ell-m-1-t)!(-L-\ell+n+t)!(-L-2\ell+2m+1+t)!}. \tag{53}
$$

Multiplying each term in sum (53) with the identity: $1\equiv(L+\ell-m)/(L+\ell-m-t)-t/(L+\ell-m-t)$ to separate it into two terms and then replacing the summation index t by $t+1$ in the second resulting sum, we reach

$$
\begin{aligned}
&(L+\ell-m)\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-m-\tfrac{1}{2}-t)!(L+\ell-m-t)!(-L-\ell+n+t)!(-L-2\ell+2m+1+t)!}\\
&\quad+\sum_t\frac{(-1)^t}{t!(L-1-t)!(L-1+\ell-m-\tfrac{1}{2}-t)!(L-1+\ell-m-t)!(-L+1-\ell+n+t)!(-l+1-2\ell+2m+1+t)!}.
\end{aligned}
\tag{53'}
$$

The first sum over t in sum (53') is no other than sum (52). Therefore, by equating sum (53) with sum (53'), one obtains an expression for sum (52) in terms of two sums:

$$
\begin{aligned}
&\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-m-\tfrac{1}{2}-t)!(L+\ell-m-t)!(-L-\ell+n+t)!(-L-2\ell+2m+1+t)!}\\
&\quad=\frac{1}{L+\ell-m}\left\{\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-m-\tfrac{1}{2}-t)!(L+\ell-m-1+t)!(-L-\ell+n+t)!(-L-2\ell+2m+1+t)!}\right.\\
&\qquad\left.-\sum_t\frac{(-1)^t}{t!(L-1-t)!(L-1+\ell-m-\tfrac{1}{2}-t)!(L-1+\ell-m-t)!(-L+1-\ell+n+t)!(-L+1-2\ell+2m+1+t)!}\right\}.
\end{aligned}
\tag{54}
$$

The second sum on the right-hand side of Eq. (54) agrees with the second sum in the closed expression for $I_{n,2n-i}(L-1\,\ell-1|\ell)$ with $i=2m$ [see Eq. (50)]. Therefore, this sum is expressed in the formula obtained by replacing L with $L-1$ on the right-hand side of Eq. (54), and hence we substitute this formula for the second sum in Eq. (54). Iteration of the above substitution leads finally to

$$
\begin{aligned}
&\sum_t\frac{(-1)^t}{t!(L-t)!(L+\ell-m-\tfrac{1}{2}-t)!(L+\ell-m-t)!(-L-\ell+n+t)!(-L-2\ell+2m+1+t)!}\\
&\quad=\sum_{\lambda=0}^{L+\ell-m-1}(-1)^\lambda\cdot\frac{(L-\lambda+\ell-m-1)!}{(L+\ell-m)!}\\
&\qquad\times\sum_t\frac{(-1)^t}{t!(L-\lambda-t)!(L-\lambda+\ell-m-\tfrac{1}{2}-t)!(L-\lambda+\ell-m-1-t)!(-L+\lambda-\ell+n+t)!(-L+\lambda-2\ell+2m+1+t)!}\\
&\qquad+\frac{(-1)^{L+\ell-m}}{(L+\ell-m)!(m-\ell)!(n-m)!(m-\ell+1)!}.
\end{aligned}
\tag{55}
$$

Next, we take into consideration the case that i is an odd positive integer. In this case i may be set equal to $2m'-1$, where m' is an arbitrary positive integer. Then the sum over t in the second term of Eq. (50) may be written as

$$\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-m'+1-\frac{1}{2}-t)!(L+\ell-m'-t)!(-L-\ell+n+t)!(-L-2\ell+2m'+t)!}, \tag{56}$$

while the sum over t in the expression for $I_{n,2n-i}(L\ell|\ell)$ with $i=2m'$, as

$$\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-m'-\frac{1}{2}-t)!(L+\ell-m'-t)!(-L-\ell+n+t)!(-L-2\ell+2m'+t)!}. \tag{57}$$

In this case, if each term in sum (57) is multiplied by the identity:

$$1\equiv(L+\ell-m'+\tfrac{1}{2})/(L+\ell-m'+\tfrac{1}{2}-t)-t/(L+\ell-m'+\tfrac{1}{2}-t),$$

sum (57) is separated into two sums:

$$\begin{aligned}&(L+\ell-m+\tfrac{1}{2})\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-m'+1-\frac{1}{2}-t)!(L+\ell-'m-t)!(-L-\ell+n+t)!(-L-2\ell+2m'+t)!}\\&\times\sum_t \frac{(-1)^t}{t!(L-1-t)!(L-1+\ell-m'+1-\frac{1}{2}-t)!(L-1+\ell-m'-t)!(-L+1-\ell+n+t)!(-L+1-2\ell+2m'+t)!}.\end{aligned} \tag{58}$$

Here carrying out the same iterative substitution for the above second sum as led to Eq. (55) from Eq. (54) in the previous paragraph, one obtains

$$\begin{aligned}&\sum_t \frac{(-1)^t}{t!(L-t)!(L+\ell-m'+1-\frac{1}{2}-t)!(L+\ell-m'-t)!(-L-\ell+n+t)!(-L-2\ell+2m'+t)!}\\&=\sum_{\lambda=0}^{L+\ell-m'-1}(-1)^\lambda\cdot\frac{(L-\lambda+\ell-m'-\frac{1}{2})!}{(L+\ell-m'+1-\frac{1}{2})!}\\&\times\sum_t \frac{(-1)^t}{t!(L-\lambda-t)!(L-\lambda+\ell-m'-\frac{1}{2}-t)!(L-\lambda+\ell-m'-t)!(-L+\lambda-\ell+n+t)!(-L+\lambda-2\ell+2m'+t)!}\\&+\frac{(-1)^{L+\ell-m'}}{(L+\ell-m'+1-\frac{1}{2})!(m'-\ell)!(n-m')!(m'-\ell)!}.\end{aligned} \tag{59}$$

Ultimately, from observation of the two expressions given by Eqs. (55) and (59), we can obtain an expression for the sum in the second term of Eq. (50) common to the two cases that i takes an even and an odd integers. This expression is written as

$$\begin{aligned}&\sum_{\lambda=0}^{L+\ell-[(i+1)/2]-1}(-1)^\lambda\cdot\frac{(L-\lambda+\ell-i/2-1)!}{(L+\ell-i/2)!}\\&\times\sum_t \frac{(-1)^t}{t!(L-\lambda-t)!(L-\lambda+\ell-[(i+1)/2]-\frac{1}{2}-t)!(L-\lambda+\ell-[i/2]-1-t)!(-L+\lambda-\ell+n+t)!(-L+\lambda-2\ell+i+1+t)!}\\&+\frac{(-1)^{L+\ell-[(i+1)/2]}}{(L+\ell-i/2)!([(i+1)/2]-\ell)!(n-[(i+1)/2])![i/2]-\ell+1)!}.\end{aligned} \tag{60}$$

Expressing in formula (60) the sums over t in terms of $I_{n,2n-i}(L-\lambda\,\ell|\ell)$ and the last term by $I_{n,2n-i}([(i+1)/2]-\ell\ell|\ell)$ and then adding formula (51) to this expression, one obtains an alternative expression for $I_{n,2n-i}(L\,\ell-1|\ell)$:

$$\begin{aligned}I_{n,2n-i}(L\,\ell-1|\ell)=&\frac{L-\ell+1}{L+\ell}\cdot I_{n,2n-i}(L\,\ell|\ell)+\left(\frac{i}{2}-\ell+1\right)\frac{(L+\ell-1)!L!}{(L-\tfrac{1}{2})!(L+\ell-i/2)!}\\&\times\left[\sum_{\lambda=0}^{L+\ell-[(i+1)/2]-1}\frac{(L-\lambda+\ell-i/2-1)!(L-\lambda-\frac{1}{2})!}{(L-\lambda+\ell)!(L-\lambda)!}I_{n,2n-(i+1)}(L-\lambda\,\ell|\ell)\right.\\&\left.-\frac{2([(i+1)/2]-[i/2]-\frac{1}{2})!([i/2]-\ell+1-\frac{1}{2})!}{([i/2]-\ell+1)![(i+1)/2]!}I_{n,2n-i}\left(\left[\frac{i+1}{2}\right]-\ell\ell|\ell\right)\right].\end{aligned} \tag{61}$$

Here the identity

$$\frac{(i-2\ell+1)([(i+1)/2]-\ell-\frac{1}{2})!}{([i/2]-\ell+1)([(i+1)/2]-\ell)!}=\frac{2([i/2]-\ell+1-\frac{1}{2})!}{([i/2]-\ell+1)!} \tag{62}$$

has been used. Here it should be noted that when $i=2(L+\ell)$ or $2(L+\ell)-1$, Eq. (61) is reduced to Eq. (31) with $M=\ell$ or $\ell-1$. Combining Eq. (61) with Eq. (41) for $M=\ell$ yields the recursion formula for $I_{n,2n-i}(L\ell|\ell)$ with respect to i as well as L:

TABLE I. $(2n+1)!\cdot I_{n,2n-i}(LM|\ell)$ with $0\leqslant M\leqslant\min\{L,\ell\}\leqslant\max\{L,\ell\}\leqslant 2$.

$LM\ell$	i = 0	1	2	3	4	5	6	7	8
000	1	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
100	2	2^2n	-2^2n	⋯	⋯	⋯	⋯	⋯	⋯
001	⋯	-2^2n	2^2n	⋯	⋯	⋯	⋯	⋯	⋯
101	⋯	-2^3n	$-2^4n(n-1)$	$2^5n(n-1)$	$-2^5n(n-1)$	⋯	⋯	⋯	⋯
111	⋯	⋯	-2^4n	$-2^5n(n-1)$	$2^5n(n-1)$	⋯	⋯	⋯	⋯
200	$2^3/3$	$2^5n/3$	$2^5n(n-2)/3$	$-2^5n(n-1)$	$2^5n(n-1)$	⋯	⋯	⋯	⋯
002	⋯	⋯	$2^5n(n-1)/3$	$-2^5n(n-1)$	$2^5n(n-1)$	⋯	⋯	⋯	⋯
201	⋯	$-2^5n/3$	$-2^5n(4n-3)/3$	$-2^7n(n-1)\times(n-3)/3$	$2^7n(n-1)\times(4n-9)/3$	$-2^73n(n-1)\times(n-2)$	$2^73n(n-1)\times(n-2)$	⋯	⋯
211	⋯	⋯	-2^5n	$-2^7n(n-1)$	$-2^7n(n-1)\times(n-3)$	$2^73n(n-1)\times(n-2)$	$-2^73n(n-1)\times(n-2)$	⋯	⋯
102	⋯	⋯	$2^6n(n-1)/3$	$2^6n(n-1)\times(2n-5)/3$	$-2^6n(n-1)\times(8n-17)/3$	$2^73n(n-1)\times(n-2)$	$-2^73n(n-1)\times(n-2)$	⋯	⋯
112	⋯	⋯	⋯	$2^6n(n-1)$	$2^6n(n-1)\times(2n-5)$	$-2^73n(n-1)\times(n-2)$	$2^73n(n-1)\times(n-2)$	⋯	⋯
202	⋯	⋯	$2^8n(n-1)/3^2$	$2^8n(n-1)\times(4n-7)/3^2$	$2^9n(n-1)\times(2n^2-14n+21)/3^2$	$-2^{11}n(n-1)\times(n-2)\times(n-4)/3$	$2^{11}n(n-1)\times(n-2)\times(4n-13)/3$	$-2^{11}3n(n-1)\times(n-2)\times(n-3)$	$2^{11}3n(n-1)\times(n-2)\times(n-3)$
212	⋯	⋯	⋯	$2^7n(n-1)$	$2^9n(n-1)\times(n-2)$	$2^9n(n-1)\times(n-2)\times(n-5)$	$-2^9n(n-1)\times(n-2)\times(5n-17)$	$2^{11}3n(n-1)\times(n-2)\times(n-3)$	$-2^{11}3n(n-1)\times(n-2)\times(n-3)$
222	⋯	⋯	⋯	⋯	$2^9n(n-1)$	$2^{11}n(n-1)\times(n-2)$	$2^{11}n(n-1)\times(n-2)\times(n-4)$	$-2^{11}3n(n-1)\times(n-2)\times(n-3)$	$2^{11}3n(n-1)\times(n-2)\times(n-3)$

$$
\begin{aligned}
I_{n,2n-i}(L\,\ell|\ell) &= \frac{2(L+2\ell-i-1)}{(i-2\ell+1)(2L+2\ell-i)}\,I_{n,2n-(i+1)}(L\,\ell|\ell)-(i-2\ell+2)\,\frac{(L+\ell)!L!}{2(i-2\ell+1)(L-\tfrac{1}{2})!(L+\ell-i/2)!}\\
&\quad\times\sum_{\lambda=1}^{L+\ell-[(i+1)/2]-1}\frac{(L-\lambda+\ell-i/2-1)!(L-\lambda-\tfrac{1}{2})!}{(L-\lambda+\ell)!(L-\lambda)!}\,I_{n,2n-(i+1)}(L-\lambda\,\ell|\ell)\\
&\quad-\frac{2([(i+1)/2]-[i/2]-\tfrac{1}{2})!([i/2]-\ell+1-\tfrac{1}{2})!}{([i/2]-\ell+1)![(i+1)/2]!}\,I_{n,2n-i}\left(\left[\frac{i+1}{2}\right]-\ell\,\ell|\ell\right)\Bigg].
\end{aligned}
\tag{63}
$$

Here note that putting $i=2(L+\ell)-1$ in Eq. (63) reduces it to Eq. (29) with $M=\ell$.

By using the two recursion formulas Eqs. (48) and (63) as well as the expression Eq. (28) for $I_{n,2n-i}(LM|\ell)$ with the maximum value of i, we can calculate $I_{n,2n-i}(LL|\ell)$ and $I_{n,2n-i}(L\,\ell|\ell)$ for all permissible values of i without any difficulty.

VI. PROCEDURE FOR CALCULATING $I_{n,2n-i}(LM|\ell)$ SUCCESSIVELY

In this section we present a procedure for calculating $I_{n,2n-i}(LM|\ell)$ successively. First, a series of $I_{n,2n-i}(LL|\ell)$ ($i=L+\ell$ to $2(L+\ell)$) with $L=0,1,\ldots,\ell$ is taken up. We start from the case of $L=\ell=0$. Then the value of ℓ is increased by step 1, which is followed by increasing the value of L by step 1. In calculation of $I_{n,2n-i}(LL|\ell)$ with different values of i, Eq. (48) is utilized in descending order of the value of i. Second, we deal with another seires of $I_{n,2n-i}(L\,\ell|\ell)$ ($i=2\ell$ to $2(L+\ell)$) with $\ell=0,1,\ldots,L-1$. In this case, $I_{n,2n-i}(10|0)$ is computed at the start. Following this computation, the value of L is increased by step 1, and then the value of ℓ, by step 1. In computation of $I_{n,2n-i}(L\,\ell|\ell)$ having different values of i, Eq. (63) is used in descending order of the value of i. $I_{n,2n-i}(LL|\ell)$ and $I_{n,2n-i}(L\,\ell|\ell)$ with the maximum value $2(L+\ell)$ of i are computed by Eq. (28). Finally, $I_{n,2n-i}(LM|\ell)$ with $M<\min\{L,\ell\}$ ($i=\ell+M$ to $2(L+\ell)$) are calculated by using Eqs. (41) and (40) in descending order of the value of M. Also, Eq. (15) is very useful to check whether or not the formulas expressing $I_{n,2n-i}(LM|\ell)$ with the minimum value $\ell+M$ of i are correct.

According to the procedure described above, we obtained all the formulas expressing $I_{n,2n-i}(LM|\ell)$ with the restriction $0\leqslant M\leqslant\min\{L,\ell\}\leqslant\max\{L,\ell\}\leqslant 2$ as the functions of the parameter n. They are exhibited in Table I.

ACKNOWLEDGMENT

I would like to thank Masatoshi Ohtake for checking whether or not each of the formulas in the table is correct.

APPENDIX A: CALCULATION OF A SUM INCLUDING SIX FACTORIALS

The sum over s in Eq. (26) is transformed into the form of a sum including six factorials as given below:

$$\sum_{s=\max\{0,-\alpha\}}^{\min\{a,b\}} \frac{(\gamma-\frac{1}{2}+s)!}{s!(a-s)!(b-s)!(\alpha+s)!(\beta-\frac{1}{2}+s)!}. \tag{A1}$$

This sum can be calculated, provided that a specific condition is satisfied. This is proved in the following: In sum (A1) s must be a non-negative integer, and a and b both are positive integers, while α, β, and γ are all integers, irrespective of being non-negative or not. Hereafter, sum (A1) is symbolized by T. If in T the condition $\gamma-\beta\geqslant 0$ is satisfied, the equality:

$$\frac{(\gamma-\frac{1}{2}+s)!}{(\alpha+s)-(\beta-\frac{1}{2}+s)!}=(\gamma-\beta)!(\gamma-\alpha-\tfrac{1}{2})!$$
$$\times\sum_{t=0}^{\min\{\gamma-\beta,\,\alpha+s\}}[t!(\gamma-\beta-t)!(\alpha+s-t)!(-\alpha+\beta-\tfrac{1}{2}+t)!]^{-1} \tag{A2}$$

holds valid. Introducing Eq. (A2) into sum (A1) yields another expression for T:

$$T=(\gamma-\beta)!(\gamma-\alpha-\tfrac{1}{2})!\sum_{t=0}^{\min\{\gamma-\beta,\,a+\alpha,\,b+\alpha\}}[t!(\gamma-\beta-t)!(\beta-\alpha-\tfrac{1}{2}+t)!]^{-1}$$
$$\times\sum_{s=\max\{0,-\alpha+t\}}^{\min\{a,\,b\}}[s!(a-s)!(b-s)!(\alpha-t+s)!]^{-1}. \tag{A1'}$$

The inner sum over s in Eq. (A1') is obtained immediately from the addition theorem for binomial coefficients, Eq. (B1) of Ref. 2. The result may be written as

$$\frac{(a+b+\alpha-t)!}{a!b!(a+\alpha-t)!(b+\alpha-t)!}. \tag{A3}$$

Thus Eq. (A1') is reduced to

$$T=\frac{(\gamma-\beta)!(\gamma-\alpha-\frac{1}{2})!}{a!b!}\sum_{t=0}^{\min\{\gamma-\beta,\,a+\alpha,\,b+\alpha\}}\frac{(a+b+\alpha-t)!}{t!(\gamma-\beta-t)!(\beta-\alpha-\frac{1}{2}+t)!(a+\alpha-t)!(b+\alpha-t)!}. \tag{A1''}$$

Therefore, if $a+b+\alpha=\gamma-\beta$, the sum over t in Eq. (A1'') can be calculated as

$$T=\frac{(a+b+\alpha)!(a+b+\beta-\frac{1}{2})!(a+b+\alpha+\beta-\frac{1}{2})!}{a!b!(a+\alpha)!(b+\alpha)!(a+\beta-\frac{1}{2})!(b+\beta-\frac{1}{2})!}. \tag{A4}$$

Otherwise, it seems impossible to obtain the sum over t. Also, in the other cases that $\gamma-\beta<0$, it seems very difficult to evaluate sum (A1).

APPENDIX B: CALCULATION OF $J_{n,2n-l}(LM|\ell)$ DEFINED BY EQ. (37)

Performance of the summation in Eq. (37) defining $J_{n,2n-l}(LM|\ell)$ starts with the use of Eq. (6''). From the upper and middle results of Eq. (6''), it is apparent that, in the case of $2L+\ell-i-s-t-1\geqslant 0$, the sum over either K or t vanishes. Therefore, here we take only the other case $t\geqslant 2L+\ell-i-s$ into consideration. So we introduce the lower result of Eq. (6'') into Eq. (37), and then extract the sum over k only, which is hereafter denoted by U_1. It may be written as

$$U_1=\sum_{k=0}^{[i/2]}\left[k!\left(\left[\frac{i}{2}\right]-k\right)!\left(\left[\frac{i+1}{2}\right]-\frac{1}{2}-k\right)!\left(-L-\ell-\frac{1}{2}+t+k\right)!\right]^{-1}. \tag{B1}$$

Replacing μ, ρ, $\nu-\rho$, and κ respectively with $[(i+1)/2]-\frac{1}{2}$, $[i/2]$, $-L-\ell-\frac{1}{2}+t$ and k in Eq. (B1) of Ref. 2 yields the result of summation over k in Eq. (B1). Then $\mu+\nu-\rho=-L-\ell+[(i+1)/2]-1+t<0$ because $t\leqslant L+\ell-[(i+1)/2]$. Therefore, if $\mu+\nu=-L-\ell+i+t-1\geqslant 0$, i.e., $t\geqslant L+\ell-i+1$, U_1 vanishes. Otherwise, i.e., when $t\leqslant L+\ell-i$, U_1 amounts to

$$(-1)^{[i/2]}\frac{(L+\ell-[(i+1)/2]-t)!}{([(i+1)/2]-\frac{1}{2})!(-L-\ell+[i/2]-\frac{1}{2}+t)![i/2]!(L+\ell-i-t)!}. \tag{B2}$$

Next, drawing out the sum over t only, denoted by U_2, we may write it as

$$U_2=(-1)^{L+\ell-[i/2]}\sum_{t=\max\{0,\,2L+\ell-i-s\}}^{t_{\max}}[t!(L-t)!(L+\ell-i-t)!(-2L-\ell+i+s+t)!]^{-1}. \tag{B3}$$

Here the relation: $(L+\ell-[i/2]-t-\frac{1}{2})!(-L-\ell+[i/2]+t-\frac{1}{2})!=(-1)^{L+\ell-[i/2]-t}$ has been used. Note that $t_{\max}=L+\ell-i$ because $L+\ell-i\leqslant L-M\leqslant L$ and $L+\ell-i=L+\ell-[(i+1)/2]-[i/2]\leqslant L+\ell-[i/2]$, and there-

fore U_2 does not appear, provided that the condition $2L+\ell-i-s\leqslant L+\ell-i$ is not fulfilled. Thus, the value which s can take is limited to L only. Hence M must be zero. In this case, only one term remains in the sum of Eq. (B3). Then U_2 is written as

$$U_2=\frac{(-1)^{L+\ell-[i/2]}}{(L+\ell-i)!(i-\ell)!}. \tag{B4}$$

Eventually, in only the special case of $M=0$, $J_{n,2n-i}(LM|\ell)$ may remain nonzero, which is written as

$$J_{n,2n-\ell}(L0|\ell)=(-1)^{\ell}\frac{(L!)^2\ell!}{(2n-i+1)!(L+\ell-i)!(i-\ell)!(L-\frac{1}{2})!(\ell-\frac{1}{2})!}. \tag{B5}$$

In all the other cases, $J_{n,2n-i}(LM|\ell)$ disappears.

[1]N. Suzuki, J. Math. Phys. **25**, 1133 (1984); **25**, 3135 (E) (1984).
[2]N. Suzuki, J. Math. Phys. **26**, 3193 (1985).

Squeezing of free Bose fields

Wojtek Ślowikowski
Institute of Mathematics, University of Aarhus, DK-8000 Aarhus C, Denmark

Klaus Mølmer
Institute of Physics, University of Aarhus, DK-8000 Aarhus C, Denmark

(Received 8 September 1988; accepted for publication 18 April 1990)

The relevant structure of a Bose Fock space using the notion of free commutative algebra with unit element and with a scalar product is incorporated. The squeezed states are then the exponentials of quadratic forms and can be written in the normal form of a general squeeze group. The connection of this group to polarization of the initial Hilbert space is established and its infinitesimal generator is computed. In this way, a common denominator for a series of papers is provided in the topic of light squeezing and last but not least a mathematical formalism is provided that may prove convenient in the general treatment of quantum optics.

I. INTRODUCTION

In his article "The mathematical foundations of quantum theory," Dirac[1] expresses the view that mathematics should be the guide while exploring new physical ideas. This also applies partly in our case.

The concept of Bose Fock space is usually referred to as "Fock representation" where one builds a linear space on physically motivated premises and provides it with operators that again are expressing some physical concepts. In this paper, we substitute the traditional Fock space representation accompanied by suitably defined creation and annihilation operators by a commutative algebra freely generated by a unit element and a Hilbert space $\mathscr{H}$ corresponding, respectively, to the vacuum state and to the one particle space. This procedure enriches the initial mathematical structure of the Hilbert space if only we do not loose the scalar product. Hence, we extend the scalar product in such a manner that the extension is "free," i.e., no new relations appear. Mathematically it corresponds to the requirement that the adjoint to the operator of multiplication by a generator should be a derivation and the word "free" is manifested by a theorem of Nelson,[2] where he shows how any contraction in the generating Hilbert space can be lifted to a homomorphism of the algebra. The obtained object is called a Bose algebra. The advantage of considering an algebra instead of just a linear space lies in the fact that the operation of multiplication by elements of $\mathscr{H}$ and their adjoints correspond, respectively, to the operators of creation and annihilations by the same elements. This way the operations significant for physical interpretation are expressed by use of an abstract mathematical structure well known in many other connections. The exponentials of elements of $\mathscr{H}$ belong to the Fock space $\Gamma\mathscr{H}$ and are the well-known coherent vectors. Exponentials of forms of degree higher than 2 always diverge in $\Gamma\mathscr{H}$.

Not all exponentials of the quadratic forms converge in $\Gamma\mathscr{H}$ but those that do have been known for a long time as pure quasifree states,[3,4] and also as ultracoherent vectors.[5] They play an important role in constructing projective representations of the metaplectic group.[6,5]

Our interest is, however, directed toward the interpretation of ultracoherent vectors as the so-called squeezed states of light. Here, we get one more example of a concept developed on purely mathematical premises that directly interprets in physics contributing to understanding and organizing a physical theory.

In the first section we present the notation and more general concepts to be used throughout the paper. The physical system we have in mind is that of a light field, which is usually discussed in other terms, therefore the rather thorough presentation. Constructing the field theory by forming products of one-particle states may, however, impart strong means for interpretation and calculus in quantum optics. This is one of our general conclusions we elaborate in Ref. 7 and intend to discuss further in the future.

In this paper, mathematical rigor is essential, and physical interpretation will mostly consist of references to publications in physics, where similar results as ours appear. Although the paper concentrates upon a specific class of states, the mathematical framework and techniques are general and should be equally well suited for the discussion of other states of an electromagnetic field.

Squeezed states, introduced in Sec. III, are relevant in quantum optics, where they have recently been detected experimentally (for extensive discussions on squeezed light see Refs. 8–10). The term "squeezing" refers to the Heisenberg uncertainty relation for any pair of conjugate observables, which allows for a "transfer" of uncertainty from one observable to the other as long as the uncertainty product is maintained. These aspects are well understood and we shall not discuss them here.

Our aim is to analyze the construction of squeezed states of infinitely many modes of light.

Operators, to be called squeeze gauges, on the one-particle space, define the squeezed states and the pertaining unitary operators in Fock space. The Bogoliubov transformations associated with squeezing are briefly discussed in the end of Sec. III.

In Sec. IV we return to the unitary squeeze operators to derive the Hamiltonian of squeezing and the normal form of the squeeze operator.

 0022-2488/90/092327-07$03.00

II. PRELIMINARIES

A. Bose algebras

We shall use here an axiomatized version of Bose Fock space, where the operations of creations and anihilations are incorporated in an algebraic structure. In this way our description differs significantly from the one usually adopted in quantum optics where quantization of the classical field is provided by the introduction of creation and annihilation operators.[11] Our description is more particle oriented.

Given a one-boson space, $\mathscr{H},\langle,\rangle$, we consider the free commutative algebra generated by the space $\mathscr{H}$ and the unit element ϕ called the *vacuum*. We denote it by $\Gamma_0\mathscr{H}$ and call it the *Bose algebra* of $\mathscr{H}$. We provide $\Gamma_0\mathscr{H}$ with a scalar product determined by the requirement that the vacuum is a unit vector and that the adjoint $\mathbf{a}(x)$ to the operator $\mathbf{a}^+(x)$ of multiplication by $x\in\mathscr{H}$ is defined on the whole $\Gamma_0\mathscr{H}$ and is a derivation, i.e., fulfils the Leibnitz rule. Hence, we require that for all $x\in\mathscr{H}$ and $f,g\in\Gamma_0\mathscr{H}$ the following conditions are satisfied:

$$\begin{aligned}&\langle\mathbf{a}^+(x)f,g\rangle=\langle f,\mathbf{a}(x)g\rangle,\\&\langle\phi,\phi\rangle=1,\\&\mathbf{a}(x)(fg)=(\mathbf{a}(x)f)g+f(\mathbf{a}(x)g).\end{aligned}\tag{2.1}$$

The operators $\mathbf{a}^+(x),\mathbf{a}(x)$ shall be called the *creation* and the *annihilation* by x, respectively.

We write $\Gamma\mathscr{H}$ for the completion of $\Gamma_0\mathscr{H},\langle,\rangle$ and $\mathscr{H}^n$ for the closure in $\Gamma\mathscr{H}$ of the linear span $\mathscr{H}_0^n$ of all the n-fold products of elements of $\mathscr{H}$. The linear span of all $\mathscr{H}^n$ shall be denoted by $\Gamma_w\mathscr{H}$.

Since for $f\in\mathscr{H}_0^m$ and $g\in\mathscr{H}_0^n$ we have [Ref. 5, Eq. (2A1)]

$$|fg|\leqslant\binom{m+n}{n}^{1/2}|f||g|,\tag{2.2}$$

we can extend the multiplication over $\Gamma_w\mathscr{H}$ making out of it an algebra.

It is easy to see that the vector

$$\exp x=\sum_{n=0}^{\infty}n!^{-1}x^n,\tag{2.3}$$

exists in $\Gamma\mathscr{H}$ for every $x\in\mathscr{H}$ and that for $x,y\in\mathscr{H}$ we have

$$\langle\exp x,\exp y\rangle=\exp\langle x,y\rangle.\tag{2.4}$$

The corresponding unit vectors $(\exp-\frac{1}{2}|x|^2)\exp x$ constitute the well known *coherent states*.[11] The operator

$$\exp\mathbf{a}(x)=\sum_{n=0}^{\infty}n!^{-1}(\mathbf{a}(x))^n\tag{2.5}$$

is well defined on $\Gamma_0\mathscr{H}$ since every element of $\Gamma_0\mathscr{H}$ is annihilated by almost all elements of the series (2.5). We have the well-known Campbell–Hausdorff–Baker relation on $\Gamma_0\mathscr{H}$:

$$\begin{aligned}\exp(\mathbf{a}^+(x)+\mathbf{a}(y))&=\exp\tfrac{1}{2}\langle x,y\rangle\exp\mathbf{a}^+(x)\exp\mathbf{a}(y)\\&=\exp-\tfrac{1}{2}\langle x,y\rangle\exp\mathbf{a}(y)\exp\mathbf{a}^+(x).\end{aligned}\tag{2.6}$$

The *Weyl displacement operators* assign to each $x\in\mathscr{H}$ the unitary transformation:

$$\begin{aligned}D_x&=\exp(\mathbf{a}^+(x)-\mathbf{a}(x))\\&=\exp-\tfrac{1}{2}|x|^2\exp\mathbf{a}^+(x)\exp-\mathbf{a}(x).\end{aligned}\tag{2.7}$$

Given $f\in\Gamma\mathscr{H}$ such that the operation of multiplication by f can be naturally defined on a dense domain in $\Gamma\mathscr{H}$ and admits a densely defined adjoint, we shall write $\mathbf{a}^+(f)$ for the closure in $\Gamma\mathscr{H},\langle,\rangle$ of the operator of multiplication by f and $\mathbf{a}(f)$ for its adjoint. In particular, $\mathbf{a}^+(f)$ is well defined for f belonging to $\Gamma_w\mathscr{H}$.

Let us denote by $\Gamma_1\mathscr{H}$ the subspace of $\Gamma\mathscr{H}$ spanned by all elements of the form $f\exp x$, where $f\in\Gamma_0\mathscr{H}$ and $x\in\mathscr{H}$.

Since for $z\in\Gamma_1\mathscr{H}$, and $x,y\in\mathscr{H}$ we have

$$\lim_{m,n}\left\langle z,\left(\sum_{k=0}^{n}k!^{-1}x^k\right)\left(\sum_{k=0}^{m}k!^{-1}y^k\right)\right\rangle=\langle z,\exp(x+y)\rangle,\tag{2.8}$$

we can define the multiplication in $\Gamma_1\mathscr{H}$ setting

$$(f\exp x)(g\exp y)=fg\exp(x+y).\tag{2.9}$$

Provided with this multiplication $\Gamma_1\mathscr{H}$ shall be called the *extended Bose algebra*.

From the second identity of (2.6) we easily conclude that $\Gamma_1\mathscr{H}$ is contained in the domain of $\mathbf{a}(\exp x),x\in\mathscr{H}$, and that the identity itself holds on $\Gamma_1\mathscr{H}$ (Ref. 5).

To every $f\in\Gamma\mathscr{H}$ there is assigned a complex-valued function $f[\cdot]$ defined on $\mathscr{H}$:

$$f[x]=\langle\exp x,f\rangle.\tag{2.10}$$

The so-called Bargmann representation[12–14] was introduced to quantum optics by Glauber[11] and will be a very useful tool in what follows.

It is easy to check that for $f,g\in\Gamma_1\mathscr{H}$ and $x\in\mathscr{H}$ we have

$$(fg)[x]=(f[x])(g[x]).\tag{2.11}$$

Let $\gamma_{\mathscr{H}}^{1/2}$ denote the Gaussian measure sitting on Hilbert–Schmidt enlargements of $\mathscr{H}$ and acting on each finite dimensional $\mathscr{K}\subset\mathscr{H}$ as the measure $\pi^{-\dim\mathscr{K}}\exp-|x|^2\,dx$. Given $f\in\Gamma_1\mathscr{H}$ we shall use the same symbol $f[\cdot]$ for the continuous extension of $f[\cdot]$ over a suitable Hilbert–Schmidt enlargement of $\mathscr{H}$. Then for $f,g\in\Gamma_1\mathscr{H}$ we have

$$\langle f,g\rangle=\int\gamma_{\mathscr{H}}^{1/2}(dx)\,\overline{f[x]}g[x].\tag{2.12}$$

The formula extends over the whole $\Gamma\mathscr{H}$ by use of standard techniques.[5]

Treating the expression $\int\gamma_{\mathscr{H}}^{1/2}(dx)(\exp x)g[x],g\in\Gamma\mathscr{H}$, as a linear functional,

$$\begin{aligned}&\left\langle f,\int\gamma_{\mathscr{H}}^{1/2}(dx)(\exp x)g[x]\right\rangle\\&\quad=\int\gamma_{\mathscr{H}}^{1/2}(dx)\,\overline{f[x]}g[x],\quad f\in\Gamma\mathscr{H},\end{aligned}\tag{2.13}$$

in view of (2.12) we can shortly write a Pettis integral

$$g=\int\gamma_{\mathscr{H}}^{1/2}(dx)(\exp x)g[x],\tag{2.14}$$

getting a coherent vector integral representation for elements of $\Gamma\mathscr{H}$.

We shall need the following construction essentially due to Nelson.[2]

Theorem 2.1: Let for $j = 1,2$ $\mathscr{H}_j, \langle , \rangle_j$ be Hilbert spaces and $\Gamma_0\mathscr{H}_j, \langle , \rangle_j$ the corresponding Bose algebras. Given a linear bounded transformation A of $\mathscr{H}_1$ into $\mathscr{H}_2$, A can be extended to a homomorphism ΓA of $\Gamma_0\mathscr{H}_1$ into $\Gamma_0\mathscr{H}_2$ and if A is a contraction then ΓA is a contraction as well.

Proof: Once the result is established for contractions, it is easy to extend it over arbitrary bounded transformations. In the case of contractions following,[2] we easily construct ΓA for orthogonal projections and isometries and then use the well-known Halmos result that represent an arbitrary contraction as the composition of isometries and an orthogonal projection. (■)

Theorem 2.1 will be used here only for $\mathscr{H}_1 = \mathscr{H}_2 = \mathscr{H}$. For example, given a bounded operator A in $\mathscr{H}$, we have for each $x \in \mathscr{H}$:

$$\Gamma A(\exp x) = \exp(Ax). \tag{2.15}$$

III. SQUEEZED STATES

A. Squeeze gauges

Given $x,y \in \mathscr{H}$ we write

$$\rho(x,y) = \tfrac{1}{2}(\langle x,y\rangle + \langle y,x\rangle), \tag{3.1}$$

$$\sigma(x,y) = -\tfrac{1}{2}i(\langle x,y\rangle - \langle y,x\rangle). \tag{3.2}$$

Phase sensitivity introduced by (3.1) and (3.2) is essential to squeezing.[7]

We shall denote by $\mathscr{L}(\mathscr{H})$ the set of Hilbert–Schmidt conjugate-linear operators L that are strict contractions and are real self-adjoint, i.e.,

$$\|L\| < 1$$

and

$$\rho(x,Ly) = \rho(Lx,y), \tag{3.3}$$

for $x,y \in \mathscr{H}$. Let us denote by κ the Cayley-like transformation:

$$\kappa A = (I + A)^{-1}(I - A), \tag{3.4}$$

where A is an operator on $\mathscr{H}$ such that the operator $(I + A)^{-1}$ exists. It is easy to see that $\kappa^2 = I$, i.e., κ is a symmetry. We shall denote by $\kappa\mathscr{L}(\mathscr{H})$ the image of $\mathscr{L}(\mathscr{H})$ by κ.

A real-linear operator M in $\mathscr{H}$ shall be called symplectic iff it holds σ invariant, i.e., if for any pair $x,y \in \mathscr{H}$ we have

$$\sigma(Mx,My) = \sigma(x,y). \tag{3.5}$$

We have the following proposition.

Proposition 3.1:[5] The elements of $\kappa\mathscr{L}(\mathscr{H})$ are exactly the symplectic operators M that are real self-adjoint non-negative such that I–M is a Hilbert–Schmidt operator. (■)

The operators from $\mathscr{L}(\mathscr{H})$ shall be called squeeze gauges.

B. Squeezed states of Bose Fock space

Take $L \in \mathscr{L}(\mathscr{H})$. It is easy to check that the series

$$h_L = \sum_{n=1}^{\infty} (Le_n)e_n, \tag{3.6}$$

where $\{e_n\}$ is an orthonormal basis in $\mathscr{H}$, converges for Hilbert–Schmidt operators L and that

$$\tfrac{1}{2}|h_L|^2 = \sum_{n=1}^{\infty} |Le_n|^2 = \|L\|_{\mathrm{HS}}{}^2, \tag{3.7}$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm.

We shall introduce the element

$$\delta_L = \sum_{n=0}^{\infty} n!^{-1}\left(-\frac{1}{2}h_L\right)^n \in \Gamma\mathscr{H}. \tag{3.8}$$

Both in Refs. 15 and 5 it is shown that the series (3.8) converges and that we have

$$|\delta_L|^2 = \det(I - L^2)^{-1/2}. \tag{3.9}$$

One can easily check that

$$\delta_L[z] = \exp -\tfrac{1}{2}\langle z,Lz\rangle. \tag{3.10}$$

When L approximates a conjugation, δ_L considered in the real wave representation (Ref. 5, Sec. III B) approximates the Dirac δ function that accounts for the adopted notation. Normalizations $|\delta_L|^{-1}\delta_L$ of the elements $\delta_L \in \Gamma\mathscr{H}$ shall be called the squeezed states of Bose Fock space.

Since $h_L/|h_L|$ is a two-boson state, squeezed states in quantum optics are also termed two-photon coherent states.[9] In Ref. 5, δ_L are called ultracoherent vectors.

For $L \in \mathscr{L}(\mathscr{H})$ and $f \in \Gamma_1\mathscr{H}$ the products $\delta_L f$ are well defined so that one can talk about the multiplication operator $\mathbf{a}^+(\delta_L)$.

For $f \in \Gamma_0\mathscr{H}$ the terms $\mathbf{a}((-\tfrac{1}{2}h_L)^n)f$ are identically zero for almost all n so that the series

$$\mathbf{a}(\delta_L) = \sum_{n=0}^{\infty} n!^{-1}\mathbf{a}((-\tfrac{1}{2}(h_L))^n) \tag{3.11}$$

converges trivially on $\Gamma_0\mathscr{H}$ providing the adjoint $\mathbf{a}(\delta_L)$ of $\mathbf{a}^+(\delta_L)$ on $\Gamma_0\mathscr{H}$. Given $L \in \mathscr{L}(\mathscr{H})$ and $x \in \mathscr{H}$ we have the following easy to check commutation on $\Gamma_0\mathscr{H}$,

$$[\mathbf{a}((\tfrac{1}{2}h_L)^m), \mathbf{a}^+(x)] = m\mathbf{a}((\tfrac{1}{2}h_L)^{m-1})\mathbf{a}(Lx), \tag{3.12}$$

which at once yields also on $\Gamma_0\mathscr{H}$ the intertwining

$$\mathbf{a}(\delta_L)\mathbf{a}^+(x) = (\mathbf{a}^+(x) - \mathbf{a}(Lx))\mathbf{a}(\delta_L). \tag{3.13}$$

We use (3.13) to show that $\Gamma_1\mathscr{H}$ is contained in the domain of the closure of $\mathbf{a}(\delta_L)$ and that we have the following intertwining on $\Gamma_1\mathscr{H}$:

$$\mathbf{a}(\delta_L)\mathbf{a}^+(\exp x) = \exp -\tfrac{1}{2}\langle Lx,x\rangle \mathbf{a}^+(\exp x)\mathbf{a}(\exp -Lx)\mathbf{a}(\delta_L). \tag{3.14}$$

C. Relation between $\mathbf{a}(\delta_N)\mathbf{a}^+(\delta_M)$, $\mathbf{a}^+(\delta_M)\mathbf{a}(\delta_N)$, and ΓA

We start this section with exposing an elucidating identity which is easy to check and will be useful later on. Given an $L \in \mathscr{L}(\mathscr{H})$ and a bounded linear operator A such that A and L have a common basis of eigenvectors, we have on $\Gamma_1\mathscr{H}$

$$\mathbf{a}(\delta_L)\Gamma A^* = \Gamma A^*\mathbf{a}(\delta_{A^2L}). \tag{3.15}$$

Proposition 3.2: Consider $M,N \in \mathscr{L}(\mathscr{H})$ with common basis of eigenvectors. Then for $f,g \in \Gamma_1\mathscr{H}$ we have

$$\langle \mathbf{a}^+(\delta_M)f, \mathbf{a}^+(\delta_N)g\rangle = \det(I-MN)^{-1/2}\langle \mathbf{a}(\delta_N)\Gamma(I-MN)^{-1/2}f, \mathbf{a}(\delta_M)\Gamma(I-MN)^{-1/2}g\rangle, \tag{3.16}$$

where

$$\det(I-MN)^{-1/2} = \langle \delta_N, \delta_M\rangle. \tag{3.17}$$

Proof: Take $f=\exp x$ and $g=\exp y$, where $x,y\in\mathscr{H}$. For this choice of f,g the second term of (3.16) can be easily computed with help of (3.14). Using (2.12) one can write the first term of (3.16) as an integral that can be explicitly computed. This verifies (3.16) for the specially chosen f,g. To extend it over the whole $\Gamma_1\mathscr{H}$ we substitute $x+tu$ for x, $y+sv$ for y and differentiate with respect to s,t in 0 m,n times, respectively, getting the relation checked for $f=u^m\exp x, g=v^n\exp y$. Since both sides of (3.16) are sesquilinear, the proposition follows. (■)

Transporting in (3.16) all the operators from the right to the left sides of the scalar products and observing that $\Gamma_1\mathscr{H}$ is dense in $\Gamma\mathscr{H}$ we obtain the following corollary.

Corollary 3.3: Given $M,N\in\mathscr{L}(\mathscr{H})$ with a common basis of eigenvectors, the identity

$$\mathbf{a}(\delta_M)\mathbf{a}^+(\delta_N) = \det(I-MN)^{-1/2}\Gamma(I-MN)^{-1/2}\mathbf{a}^+(\delta_N) \times \mathbf{a}(\delta_M)\Gamma(I-MN)^{-1/2} \tag{3.18}$$

holds on $\Gamma_1\mathscr{H}$. (■)

Given an $L\in\mathscr{L}(\mathscr{H})$, we write briefly

$$\check{L} = (I-L^2)^{1/2} \tag{3.19}$$

and define the operator S_L transforming $\Gamma_1\mathscr{H}$ into $\Gamma\mathscr{H}$ setting

$$S_L = \det\check{L}^{1/2}\mathbf{a}^+(\delta_L)\Gamma\check{L}\mathbf{a}(\delta_{-L}). \tag{3.20}$$

Substituting in (3.16) $M=N=L$ and setting $\Gamma\check{L}\mathbf{a}(\delta_{-L})f$ for f and $\Gamma\check{L}\mathbf{a}(\delta_{-L})g$ for g we conclude that S_L is an isometry. It is easy to see that the adjoint S_L^* of S_L fulfils on $\Gamma_1\mathscr{H}$ the identity

$$S_L^* = S_{-L}, \tag{3.21}$$

which yields the following theorem.

Theorem 3.4: Given $L\in\mathscr{L}(\mathscr{H})$, the transformation S_L extends to a unitary mapping of $\Gamma\mathscr{H}$ and

$$S_L^{-1} = S_{-L}, \tag{3.22}$$

where we denote the extension by the same symbol. (■)

We shall call S_L the *L-squeeze operator.*

We conclude this section with a formula for composition of squeeze operators.

Proposition 3.5: Given commuting $M,N\in\mathscr{L}(\mathscr{H})$, we have

$$S_M S_N = S_{(M+N)(I+MN)^{-1}}. \tag{3.23}$$

Proof: It easily follows if we apply (3.18) together with (3.15) and its dual on the left side of (3.23). (■)

D. Bogoliubov transformations associated with the *L*-squeeze operator

In this section, we present a mathematical frame for considering the squeezed state as a new vacuum.[9] Let us fix an $L\in\mathscr{L}(\mathscr{H})$. To each $x\in\mathscr{H}$ we assign a pair of transformations $\mathbf{a}_L^+(x)$ and $\mathbf{a}_L(x)$ of $\Gamma_1\mathscr{H}$ into itself setting

$$\mathbf{a}_L^+(x) = \mathbf{a}^+(\check{L}^{-1}x) + \mathbf{a}(L\check{L}^{-1}x),$$
$$\mathbf{a}_L(x) = \mathbf{a}^+(L\check{L}^{-1}x) + \mathbf{a}(\check{L}^{-1}x). \tag{3.24}$$

It is easy to see that (3.24) are adjoint to each other. We shall denote their closures by the same symbols. It is easy to observe that $S_L(\Gamma_1\mathscr{H})$ is contained in the domains of the closure of the operators (3.24) and that the following intertwinings on $\Gamma_1\mathscr{H}$ are valid,

$$S_L\mathbf{a}^+(x) = \mathbf{a}_L^+(x)S_L,$$
$$S_L\mathbf{a}(x) = \mathbf{a}_L(x)S_L. \tag{3.25}$$

The transformations (3.24) are usually called the Bogoliubov transformations.

Consider the space $\mathscr{H}_L$ of all the elements of the form $f=S_Lx$, $x\in\mathscr{H}$, the L-squeeze space. Now, the Bose algebra $\Gamma_0\mathscr{H}_L$ of the L-squeeze space $\mathscr{H}_L$ is

$$\Gamma_0\mathscr{H}_L = S_L(\Gamma_0\mathscr{H}), \tag{3.26}$$

with the product $(fg)_L$ of $f,g\in\Gamma_0\mathscr{H}_L$ given by the formula

$$(fg)_L = S_L((S_L^{-1}f)(S_L^{-1}g)) \tag{3.27}$$

and with the new unit element, which is the squeezed vacuum,

$$\phi_L = S_L\phi = \det\check{L}^{1/2}\delta_L. \tag{3.28}$$

From (3.25) and (3.27) we can see that the creation by $a\in\mathscr{H}_L$ in $\Gamma_0\mathscr{H}_L$ consist of application of the operator $\mathbf{a}_L^+(S_L^{-1}a)$.

From (3.25) it easily follows that on $\Gamma_1\mathscr{H}$ we have

$$S_L(\mathbf{a}^+(x)-\mathbf{a}(x))S_L^{-1} = \mathbf{a}^+((\kappa L)^{1/2}x) - \mathbf{a}((\kappa L)^{1/2}x), \tag{3.29}$$

where κL is given by (3.4). Taking the exponential on both sides we arrive to the intertwining

$$S_L D_x = D_{(\kappa L)^{1/2}x}S_L. \tag{3.30}$$

It should be mentioned that this and some other algebraic identities considered here, have been derived in Ref. 16 in the single-mode case and by use of other techniques.

The displacement operator does not alter the quantum uncertainty properties, therefore the term squeezed states also embraces displaced squeezed states in the physical literature. From (3.30) we see that these states may also be obtained by squeezing a coherent state.

IV. GROUPS OF SQUEEZE OPERATORS AND THEIR GENERATORS

A. Unitary groups of squeeze operators

For every $L\in\mathscr{L}(\mathscr{H})$ there can be found an orthonormal system $\{e_n\}$ such that

$$L = \sum_{n=1}^{\infty} t_n\langle\cdot, e_n\rangle e_n, \tag{4.1}$$

where $|t_n|<1$ and $\Sigma|t_n|^2<\infty$. Clearly, there exists an orthonormal system such that in (4.1) all t_n are positive real numbers. We have then

$$A=\kappa L=\sum_{n=1}^{\infty}(r_n\rho(e_n,\cdot)e_n+r_n^{-1}\rho(ie_n,\cdot)ie_n),$$

$$r_n=\frac{1-t_n}{1+t_n}>0,\quad \sum_{n=1}^{\infty}(1-r_n)^2<\infty. \tag{4.2}$$

Since for any pair of commuting operators $A,B\in\kappa\mathscr{L}(\mathscr{H})$,

$$(\kappa A+\kappa B)(I+(\kappa A)(\kappa B))^{-1}=\kappa(AB), \tag{4.3}$$

we have from (3.23)

$$S_{\kappa A}S_{\kappa B}=S_{\kappa(AB)}, \tag{4.4}$$

which shows that the function

$$]-\infty,+\infty[\ni t\to S_{\kappa(A^t)}\in\{\text{unitary operators}\}, \tag{4.5}$$

where

$$A^t=\sum_{n=1}^{\infty}(r_n{}^t\rho(e_n,\cdot)e_n+r_n{}^{-t}\rho(ie_n,\cdot)ie_n), \tag{4.6}$$

constitutes a unitary group. Due to Lemma 3, p. 616 of Ref. 17 the group (4.5) is also continuous and we can search for its infinitesimal generator iH_L, i.e., we search for a self-adjoint H_L with the domain $D(H_L)$ such that

$$S_{B(t)}=\exp itH_L, \tag{4.7}$$

where we write briefly

$$B(t)=\kappa(A^t)=\sum_{n=1}^{\infty}(1-r_n{}^t)(1+r_n{}^t)^{-1}\langle\cdot,e_n\rangle e_n, \tag{4.8}$$

so that, using (3.15), its dual and (3.18), we have

$$\begin{aligned}S_{B(t)}&=\det(I-B(t)^2)^{1/2}\mathbf{a}^+(\delta_{B(t)})\Gamma(I-B(t)^2)^{1/2}\mathbf{a}(\delta_{-B(t)})\\&=\det(I-B(t)^2)^{1/2}\Gamma(I-B(t)^2)^{1/2}\mathbf{a}^+(\delta_{B(t)(I-B(t)^2)^{-1}})\mathbf{a}(\delta_{-B(t)})\\&=\det(I-B(t)^2)^{-1/2}\mathbf{a}(\delta_{B(t)})\mathbf{a}^+(\delta_{B(t)(I-B(t)^2)^{-1}})\Gamma(I-B(t)^2)^{-1/2}.\end{aligned} \tag{4.9}$$

Let $L\in\mathscr{L}(\mathscr{H})$ be given by (4.1) and consider $A=\kappa L$:

$$\begin{aligned}\log A&=\sum_{n=1}^{\infty}(\log r_n)(\rho(e_n,\cdot)e_n-\rho(ie_n,\cdot)ie_n)\\&=\sum_{n=1}^{\infty}(\log r_n)\langle\cdot,e_n\rangle e_n,\end{aligned} \tag{4.10}$$

where t_n in (4.1) are real and r_n are as in (4.2). It is easy to see that the definition of $\log A$ does not depend on the choice of orthonormal system $\{e_n\}$. Due to (4.2) the sequence $\{\log r_n\}$ is square summable so that $\log A$ is a Hilbert–Schmidt operator. Since it is also conjugate linear,

$$h_{\log A}=\sum_{n=1}^{\infty}(\log r_n)e_n{}^2 \tag{4.11}$$

is a well-defined element of $\mathscr{H}^2$. Now we are ready to examine the unitary group of squeeze operators.

B. Generators of the squeeze groups

Theorem 4.1: The domain $D(H_L)$ contains $\Gamma_w\mathscr{H}$ and on $\Gamma_w\mathscr{H}$ we have

$$\begin{aligned}iH_L&=\tfrac{1}{4}\sum_{n=1}^{\infty}(\log r_n)(\mathbf{a}^+(e_n{}^2)-\mathbf{a}(e_n{}^2))\\&=\tfrac{1}{4}(\mathbf{a}^+(h_{\log A})-\mathbf{a}(h_{\log A}))\\&=i\tfrac{1}{4}(\mathbf{a}^+(-ih_{\log A})+\mathbf{a}(-ih_{\log A}))\end{aligned} \tag{4.12}$$

where $h_{\log A}$ is defined by (4.10).

Proof: Take $f,g\in\Gamma_0\mathscr{H}$. Using the last part of (4.9) we obtain

$$\begin{aligned}\langle f,S_{B(t)}g\rangle&=\det(I-B(t)^2)^{-1/2}\langle\delta_{-B(t)}f,\delta_{B(t)(I-B(t)^2)^{-1}}\\&\quad\times\Gamma(I-B(t)^2)^{-1/2}g\rangle.\end{aligned} \tag{4.13}$$

For $f\in\Gamma_0\mathscr{H}$ we have

$$(\Gamma(I-B(t)^2)^{-1/2}f)[z]=f[(I-B(t)^2)^{-1/2}z] \tag{4.14}$$

so that applying (2.12) to the right side of (4.14) we obtain

$$\begin{aligned}\langle f,S_{B(t)}g\rangle&=\det(I-B(t)^2)^{-1/2}\\&\quad\cdot\int\overline{f[z]}\exp\tfrac{1}{2}(\langle B(t)z,z\rangle-\langle z,B(t)\\&\quad\times(I-B(t)^2)^{-1}z\rangle)g[(I-B(t)^2)z]\gamma_H^{1/2}(dz),\end{aligned} \tag{4.15}$$

the exponent term that originates from (3.10) is here understood continuously extended to a suitable Hilbert–Schmidt enlargement of $\mathscr{H}$.

Using the Lebesgue dominated convergence theorem we differentiate in $t=0$ the integral on the right side of (4.15). Since $d/dt\det(I-B(t)^2)^{1/2}|_{t=0}=0$ this provides the derivative of the whole (4.15) in $t=0$. It is relatively easy to compute the derivative explicitly and conclude that iH_L on $\Gamma_0\mathscr{H}$ is as given by (4.12).

To show that (4.12) holds on $\Gamma_w\mathscr{H}$ as well, we observe that due to (2.2) the operator $\mathbf{a}^+(h_{\log A})$ is continuous from $\mathscr{H}^n$ into $\mathscr{H}^{n+2}$ and its adjoint is continuous from $\mathscr{H}^{n+2}$ into $\mathscr{H}^n$ for all n and the theorem follows. (■)

Write

$$\delta_M\Gamma_w\mathscr{H}=\{\mathbf{a}^+(f)\delta_M:\ f\in\Gamma_w\mathscr{H}\}. \tag{4.16}$$

Lemma 4.2: The closure of the restriction of iH_L to $\Gamma_w\mathscr{H}$ contains in its domain $S_{B(t)}\Gamma_w\mathscr{H}$ for all $t\geqslant 0$ and we have

$$iH_L(\delta_M\Gamma_w\mathscr{H})\subset\delta_M\Gamma_w\mathscr{H} \tag{4.17}$$

Proof: For $M\in\mathscr{L}(\mathscr{H})$ we define

$$\delta_{M/n}=\sum_{j=1}^{n}j!^{-1}(-\tfrac{1}{2}h_M)^j. \tag{4.18}$$

Using the dual of (3.14) together with (3.7) of Ref. 5 we easily show that for a square summable $\{s_m\}\subset\mathbb{C}$ and any $f\in\Gamma_w\mathscr{H}$ we have

$$\lim_n \mathbf{a}\Big(\sum_{m=1}^{\infty} s_m e_m{}^2\Big)\delta_{M/n} f$$

$$= \delta_M \sum_{m=1}^{\infty} \bar{s}_m (\mathbf{a}(e_m) - \mathbf{a}^+(Me_m))^2 f \tag{4.19}$$

and the lemma follows. (■)

Theorem 4.3: The subspace $\Gamma_0\mathscr{H}$ constitutes a core for iH_L.

Proof: Due to the remark at the end of the proof of Theorem 4.1 the generator iH_L extends from $\Gamma_0\mathscr{H}$ to $\Gamma_w\mathscr{H}$ and it follows from Lemma 4.2 that it can be further extended over to the linear hull of all $S_{B(t)}\Gamma_w\mathscr{H}$. But this hull is $S_{B(t)}$ invariant so by Theorem X.49 of Ref. 18 it is a core for iH_L. (■)

Hence, for $t = 1$ we have

$$S_L = \exp \tfrac{1}{4}(\mathbf{a}^+(h_{\log \kappa L}) - \mathbf{a}(h_{\log \kappa L}))$$
$$= \det \check{L}^{1/2} \exp\, -\tfrac{1}{2}\mathbf{a}^+(h_L)\Gamma\check{L} \exp\, -\tfrac{1}{2}\mathbf{a}(h_L). \tag{4.20}$$

Observe that definition (3.6) does not require that $\log \kappa L$ is a strict contraction. It is sufficient that it is a Hilbert–Schmidt operator.

To obtain the normal form of S_L we only have to derive the following.

C. The normal form of $\Gamma\check{L}$

Let us write $\mathbf{N}$ for the set of all finite tuples $\mathscr{k} = (k_1,...,k_n)$. Given an orthonormal basis $\{e_n\}$ in $\mathscr{H}$, we construct an orthonormal system $\{e_\mathscr{k}\}$, setting

$$e_\mathscr{k} = \mathscr{k}!^{-1/2} e^\mathscr{k}, \tag{4.21}$$

where

$$e^\mathscr{k} = e_1{}^{k_1} \cdots e_n{}^{k_n} \tag{4.22}$$

and

$$\mathscr{k}! = k_1! \ldots k_n!. \tag{4.23}$$

For any fixed d the elements $e_\mathscr{k}$ with $|\mathscr{k}| = k_1 + \cdots + k_n$, length of the tuples, equal to d form an orthonormal basis in $\mathscr{H}^d$.

Theorem 4.4: Given a bounded operator R in $\mathscr{H}, \langle,\rangle$, we have on $\Gamma_1\mathscr{H}$

$$\Gamma(I + R) = \sum_{\mathscr{k}\in\mathbf{N}} \mathbf{a}^+(\Gamma R e_\mathscr{k})\mathbf{a}(e_\mathscr{k}). \tag{4.24}$$

Proof: Observe that

$$\mathbf{a}(a_1{}^{k_1}) \cdots \mathbf{a}(a_n{}^{k_n}) x^m$$
$$= \begin{cases} m(m - |\mathscr{k}|)!^{-1} \langle a_1, x\rangle^{k_1} \cdots \langle a_n, x\rangle^{k_n} x^{m-|\mathscr{k}|} \\ \quad \text{for } |\mathscr{k}| \leqslant m, \\ 0, \text{ otherwise.} \end{cases} \tag{4.25}$$

Furthermore, due to the continuity of $\mathscr{H} \ni z \to z^m \in \mathscr{H}^m$ we have

$$z^m = \lim_n \Big(\sum_{j=1}^{n} \langle e_j, z\rangle e_j\Big)^m = m! \sum_{|\mathscr{k}|=m} \mathscr{k}!^{-1} \langle e, z\rangle^\mathscr{k} e^\mathscr{k}, \tag{4.26}$$

where we write briefly

$$\langle e, z\rangle^\mathscr{k} = \langle e_1, z\rangle^{k_1} \cdots \langle e_n, z\rangle^{k_n}. \tag{4.27}$$

Consequently, for $k \leqslant m$ we have

$$\sum_{|\mathscr{k}|=k} \mathscr{k}!^{-1} (\Gamma R e^\mathscr{k}) \mathbf{a}(e^\mathscr{k}) z^m$$
$$= m!(m - |\mathscr{k}|)!^{-1} \Big(\Gamma R \sum_{|\mathscr{k}|=k} \mathscr{k}!^{-1} \langle e, z\rangle^\mathscr{k} e^\mathscr{k}\Big) z^{m-k}$$
$$= m!(m-k)!^{-1} (Rz)^k z^{m-k}. \tag{4.28}$$

Summing over k we get

$$\sum_{|\mathscr{k}|\leqslant m} \mathbf{a}^+(\Gamma R e_\mathscr{k})\mathbf{a}(e_\mathscr{k}) z^m = \Gamma(I + R) z^m. \tag{4.29}$$

Since ΓR is a homomorphism in $\Gamma_1\mathscr{H}$, we have

$$\sum_\mathscr{k} \mathbf{a}^+(\Gamma R e_\mathscr{k})\mathbf{a}(e_\mathscr{k}) z^m \sum_{j=0}^{n} j!^{-1} x^j$$
$$= \sum_{j=0}^{n} \sum_{|\mathscr{k}|<j+m} \mathbf{a}^+(\Gamma R e_\mathscr{k})\mathbf{a}(e_\mathscr{k}) z^m j!^{-1} x^j$$
$$= \sum_{j=0}^{n} \Gamma(I + R)(z^m j!^{-1} x^j)$$
$$= \Big(\sum_{j=0}^{n} j!^{-1} [(I + R)x]^j\Big) \Gamma(I + R) z^m. \tag{4.30}$$

Since due to (4.29) and (2.2) we have

$$\Big|\sum_{|\mathscr{k}|<j+m} \mathbf{a}^+(\Gamma R e_\mathscr{k})\mathbf{a}(e_\mathscr{k}) z^m j!^{-1} x^j\Big|$$
$$= |\Gamma(I + R) z^m j!^{-1} x^j|$$
$$\leqslant \binom{m+j}{m}^{1/2} |\Gamma(I + R) z^m| j!^{-1} |x|^j, \tag{4.31}$$

the series of the last term of (4.30) converges so that passing to the limit we obtain

$$\sum_\mathscr{k} \mathbf{a}^+(\Gamma R e_\mathscr{k})\mathbf{a}(e_\mathscr{k}) z^m \exp x = \Gamma(I + R)(z^m \exp x), \tag{4.32}$$

where the left side of (4.32) is defined as the limit of the first term of (4.30). Since the powers of elements of $\mathscr{H}$ span the whole $\Gamma_0\mathscr{H}$, the theorem follows. (■)

In the case of L squeeze we are interested in writing the middle operator $\Gamma\check{L} = \Gamma(I - L^2)^{1/2}$ of S_L in the normal form. In this case the operator R of Theorem 4.4 is chosen as the Hilbert–Schmidt contraction,

$$R = \check{L} - I = -\sum_{n=1}^{\infty} t_n{}^2 [1 + (1 - t_n{}^2)^{1/2}]^{-1} \langle e_n, \cdot\rangle e_n. \tag{4.33}$$

From Theorem 4.4 we obtain at once the following.

Corollary 4.5: The identity

$$S_L = \det \check{L}^{1/2} \sum_\mathscr{k} (t^2 [1 + (1 - t^2)^{1/2}]^{-1})^\mathscr{k}$$
$$\times \mathbf{a}^+(\delta_L e_\mathscr{k})\mathbf{a}(\delta_{-L} e_\mathscr{k}), \tag{4.34}$$

where $s^\mathscr{k} = s_1{}^{k_1} \cdots s_n{}^{k_n}$, represents the normal form of S_L on $\Gamma_1\mathscr{H}$.

D. Coherent vector representations

Theorem 4.6: Given a bounded linear operator R in $\mathscr{H}, \langle,\rangle$, we have on $\Gamma_1\mathscr{H}$ the following identity:

$$\Gamma(I+R)=\int\gamma_{\mathscr{H}}^{1/2}(dx)\mathbf{a}^{+}(\exp Rx)\mathbf{a}(\exp x), \tag{4.35}$$

where on every $f\in\Gamma_1\mathscr{H}$ the right side is the usual Pettis integral.

Proof: Take an $f\in\Gamma_1\mathscr{H}$ and $x\in\mathscr{H}$. We have

$$\begin{aligned}\left(\int\gamma_{\mathscr{H}}^{1/2}(dx)\mathbf{a}^{+}(\exp Rx)\mathbf{a}(\exp x)f\right)[z]&\\ =\int\gamma_{\mathscr{H}}^{1/2}(dx)\langle\exp z,(\exp Rx)\mathbf{a}(\exp x)f\rangle&\\ \times\int\gamma_{\mathscr{H}}^{1/2}(dx)\ \overline{(\exp R^{*}z)[x]}(\mathbf{a}(\exp x)f)[z],&\end{aligned} \tag{4.36}$$

and using (2.12) we arrive to the desired result. (■)

Take $L\in\mathscr{L}(\mathscr{H})$ and consider its representation given by (4.1),

$$L=\sum_{n=1}^{\infty}t_n\langle\cdot,e_n\rangle e_n,$$

where $\{e_n\}$ are chosen such as to make all t_n real. Let us define a conjugate $^{-}$ in $\mathscr{H}$ setting

$$\bar{x}=\sum_{n=1}^{\infty}\langle x,e_n\rangle e_n, \tag{4.37}$$

so that

$$L=\ ^{-}T=T^{-}, \tag{4.38}$$

where

$$T=\sum_{n=1}^{\infty}t_n\langle e_n,\cdot\rangle e_n \tag{4.39}$$

is a Hilbert–Schmidt operator in $\mathscr{H}$.

Let us write H for the real part of $\mathscr{H}$ relative the conjugation $^{-}$,

$$H=\{x:\ x=\bar{x}\} \tag{4.40}$$

We shall need a Gaussian measure γ_H sitting on Hilbert– Schmidt enlargements of H and such that writing γ_K for the projection of γ_H on an arbitrary finite-dimensional linear subset $K\subset H$ we have

$$\gamma_K(dx)=(2\pi)^{-1/2\dim K}\exp-\tfrac{1}{2}|x|^2\,dx. \tag{4.41}$$

Theorem 4.7: Consider a trace class $L\in\mathscr{L}(\mathscr{H})$ and choose a conjugation $^{-}$ in $\mathscr{H}$ such that $T=\ ^{-}L=L^{-}$, is a self-adjoint operator in H. Then

$$\begin{aligned}\delta_L=\det(I-T)^{-1/2}\int\gamma_H(dx)\exp-\tfrac{1}{2}\langle x,T(I-T)^{-1}x\rangle&\\ \times(\exp iT^{1/2}(I-T)^{1/2}x),&\end{aligned} \tag{4.42}$$

where on the right side we have the Pettis integral and the integration takes place on a Hilbert–Schmidt enlargement of H.

Proof: We assume first that L is finite dimensional. By explicit integration we get for an arbitrary $y\in\mathscr{H}$:

$$\begin{aligned}\int\gamma_H(dx)\exp-\tfrac{1}{2}\langle x,T(I-T)^{-1}x\rangle&\\ \times(\exp iT^{1/2}(I-T)^{1/2}x)[y]&\\ =\int\gamma_H(dx)\exp(-\tfrac{1}{2}\langle x,T(I-T)^{-1}x\rangle&\\ +\langle y,iT^{1/2}(I-T)^{1/2}x\rangle)&\\ =\det(I-T)^{1/2}\delta_L[y]=\det(I-T)^{1/2}\exp-\tfrac{1}{2}\langle T\bar{y},y\rangle,&\end{aligned} \tag{4.43}$$

where the continuous extension over a suitable Hilbert–Schmidt enlargement of H is implicitly anticipated under the integral sign. Using the dominated convergence theorem on the right side, we extend our result over arbitrary trace class $L\in\mathscr{L}(\mathscr{H})$. (■)

[1] P. A. M. Dirac, "Mathematical Foundations of Quantum Theory," in *Proceedings of a Conference on Mathematical Foundations of Quantum Theory*, edited by A. R. Marlow (1978), pp. 1–8.
[2] E. Nelson, J. Funct. Anal. **12**, 211–227 (1973).
[3] H. Araki, Pub. RIMS **18**, 283–338 (1982).
[4] M. Fannes, Comm. Math. Phys. **51**, 55–66 (1976).
[5] W. Slowikowski, Adv. Appl. Math. **9**, 377–427 (1988).
[6] M. Vergne, C. R. Acad. Sci. Paris, Ser. A **285**, 191–194 (1977).
[7] K. Mølmer and W. Slowikowski, J. Phys. A. Math. Gen. **21**, 2565–2571 (1988).
[8] D. F. Walls and H. J. Kimble, J. Opt. Soc. Am. B **4**, 1453–1741 (1987).
[9] H. P. Yuen, Phys. Rev. A **13**, 2226–2242 (1976).
[10] B. Yurke, L. McCall, and J. R. Klauder, Phys. Rev. A **33**, 4033–4054 (1986).
[11] R. J. Glauber, Phys. Rev. **131**, 2766–2788 (1963).
[12] V. Bargmann, Comm. Pure Appl. Math. **14**, 187–214 (1961).
[13] I. Segal, Adv. Math. Suppl. Stud. **3**, 321–345 (1978).
[14] W. Feller, Proc. Natl. Acad. Sci. USA **48**, 2204 (1962).
[15] P. Kristensen, L. Mejlbo, and E. T. Poulsen, Commun. Math. Phys. **6**, 29–48 (1967).
[16] B. L. Schumaker, Phys. Rep. **135**, 317–408 (1986).
[17] N. Dunford, and J. T. Schwartz, *Linear Operators* (Interscience, New York, 1958).
[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, II: Fourier Analysis, Selfadjointness* (Academic, New York, 1975).

Erratum: On the Cauchy problem for Yang–Mills equations with external current [J. Math. Phys. 30, 1699 (1989)]

Z. Swierczynski
Institute of Physics, Jagellonian University, Reymonta 4, 30-059 Krakow, Poland

(Received 4 January 1990; accepted for publication 28 February 1990)

The gauge group we would like to consider is $G = \mathrm{SO}(n)$ or $G = \mathrm{SU}(n)$ [not $G = \mathrm{Gl}_R(n)$ or $G = \mathrm{Gl}_C(n)$ as we wrote on p. 1700].

We replace the space $\mathscr{H}_r$ with

$$\mathscr{H}_r^q = \{(A,E,j^0)\colon A_k \in H_{r+1},\ E_k \in H_r,\ j^0 \in H_r \cap W_{6/5}^q\},$$

where

$$W_p^q = \left\{ f\colon \left(\sum_{|\kappa|<q} \int (\mathrm{Tr}(\partial_\kappa f)^2)^{p/2}\, d^3x \right)^{1/p} < \infty \right\}$$

$$0 \leqslant q \leqslant r-1$$

[this is necessary since the function J defined by formula (9c) does not map $\mathscr{H}_r$ into $\mathscr{H}_r$]. Now, if $0 \leqslant q \leqslant r-1$, $r = 1, 2, \ldots$, and the components of the external current j_k treated as functions of time with values in the space $H_{r+1} \cap W_{6/5}^{q+1}$ are continuous, then the function J is also continuous. With suitably changing considerations contained in the paper [the spaces $\mathscr{H}_{s+2}^{s}$ are needed in considerations concerning the regularity of the solution; in order to prove the global existence of the solution it should be shown that the $\mathscr{H}_1^0$ norm of the solution is bounded on each finite time interval, the $L_{6/5}$ estimation follows from Eq. (1e) and the inequality $\|fg\|_{L_{6/5}} \leqslant \|f\|_{L_{6/5}} \|g\|_{L_\infty}$], one can prove that for any initial data satisfying the conditions $A_k(0,\mathbf{x}) \in H_{r+3}^{\mathrm{loc}}$, $E_k(0,\mathbf{x}) \in H_{r+2}^{\mathrm{loc}}$, $D_k E_k(0,\mathbf{x}) \in H_{r+2}^{\mathrm{loc}}$, $r = 1, 2, \ldots$, and the external current being a C^s function of time with values in the space H_{r+3}^{loc} for $s = 0, 1, \ldots, r$, there exists a global solution. The potentials and electric field treated as functions of the variables t, x^1, x^2, x^3 are of C^{r+1} and C^r class, respectively.

Formulas (9b), (23), (33), (42), and (44) should read:

$$\mathscr{A} = \begin{pmatrix} E_k^T \\ \partial_1\partial_1 A_k - \partial_k \partial_1 A_1 \\ 0 \end{pmatrix}, \tag{9b}$$

$$I(p) = \int_{K_p} \frac{1}{2}\,\mathrm{Tr}\left\{ \frac{1}{4}\,(F_{\mu\nu} 1^\mu m^\nu)^2 + \sum_{A=1}^{2} (F_{\mu\nu} 1^\mu e_A^\nu)^2 + (F_{\mu\nu} e_1^\mu e_2^\nu)^2 \right\} d^3\mathbf{x}, \tag{23}$$

$$\left| \int_{K_p} r\,dr\,d\Omega \int d\lambda\, \lambda^2 [x^\nu j_\gamma(\lambda x), F_{\alpha\beta}(x)] \right| \leqslant \left(C_3 \|j_0(0)\|_{L_2}^2 + C_4 r_0 \int_0^{r_0} \|\mathbf{j}(s)\|_{H_2}^2\, ds + C_5 \left(\|A(0)\|_{L_2}^2 + 2r_0 \int_0^{r_0} \overline{E}(s)\, ds \right) r_0 \int_0^{r_0} \|\mathbf{j}(s)\|_{L_\infty}^2\, ds + C_6 r_0^2 \int_0^{r_0} \|\mathbf{j}(s)\|_{L_\infty}^2\, \mathrm{ds} \right) \left(r_0^2 \int_0^{r_0} \|F(s)\|_{L_\infty}^2\, ds \right)^{1/2}, \tag{33}$$

$$\|F(t)\|_{L_\infty}^2 \leqslant f_1'(t) + f_2'(t) \int_0^t \|F(s)\|_{L_\infty}^2\, ds + f_3'(t) \int_0^t \|j_0(s)\|_{L_\infty}^2\, ds, \tag{42}$$

$$N(t) = \|F(t)\|_{L_\infty}^2 + \|j_0(t)\|_{L_\infty}^2. \tag{44}$$

 0022-2488/90/092334-01$03.00