Defining equations for supergroup orbits in super Clifford modules

M. J. Bergvelt
Max Planck Institut für Mathematik, Gottfried-Claren-strasse 26, 5300 Bonn 03, Federal Republic of Germany

(Received 27 September 1988; accepted for publication 6 December 1989)

The defining equations for group orbits of a maximal subgroup of $\mathrm{Gl}_{p|q}$ in certain highest weight representations of the Lie super algebra $\mathrm{gl}_{p|q}$ are discussed.

I. INTRODUCTION

The direct sum $\oplus_{k=1}^{n}\Lambda^k(C^n)$ of fundamental representations of $\mathrm{gl}(n,C)$ is a module over the Clifford algebra generated by 1 and fermionic creation and annihilation operators ψ_i,ψ_i^*, $i\in\{1,...,n\}$. It is easy to show that the $v\in\Lambda^k(C^n)$ that satisfy the *Plücker equation*

$$\sum_{i=1}^{n}\psi_i v\otimes\psi_i^* v=0 \tag{1.1}$$

form precisely the $\mathrm{Gl}(n,C)$ orbit of the highest weight vector in $\Lambda^k(C^n)$ (cf. Peterson and Kac[1]).

In a similar manner one can use a super Clifford module to construct representations $M_{k,m}$ of the Lie super algebra $\mathrm{gl}_{p|q}$. In this paper we show that the analog of Eq. (1.1) (introduced by Kac and van de Leur[2] in connection with the super K.P. equation) describes a group orbit of the maximal subgroup of the supergroup $\mathrm{GL}_{p|q}$ that acts on the representations $M_{k,m}$. (Since these representations are not integrable not the whole group $\mathrm{Gl}_{p|q}$ acts.) The extension of these results to infinite-dimensional Lie super algebras and the connection with super Grassmannians and super K.P. equation will be treated elsewhere.

II. SUPER CLIFFORD ALGEBRAS AND MODULES OVER $\mathrm{gl}_{p/q}$

Let p,q be non-negative integers. For pairs (k,m) of integers with $0\leqslant k\leqslant p$, $0\leqslant m\leqslant q$ we define indexing sets

$$S_{k,m}:=\{i\in Z\,|\,1\leqslant i\leqslant k\}\cup\{j\in Z\,|\,p+1\leqslant j\leqslant p+m\}. \tag{2.1}$$

Let $\mathrm{SCl}_{p|q}$ be the super Clifford algebra generated by 1 and ψ_i,ψ_j^*, $i,j\in S_{p,q}$, satisfying the defining relations

$$\begin{aligned}&\psi_i\psi_j^*+(-1)^{\bar i\bar j}\psi_j^*\psi_i=\delta_{ij},\\&\psi_i\psi_j+(-1)^{\bar i\bar j}\psi_j\psi_i=0,\\&\psi_i^*\psi_j^*+(-1)^{\bar i\bar j}\psi_j^*\psi_i^*=0,\end{aligned} \tag{2.2}$$

where the parity of a generator is given by $p(\psi_i)=p(\psi_i^*)=\bar 0$ if $1\leqslant i\leqslant p$ and $p(\psi_i)=p(\psi_i^*)=\bar 1$ if $p+1\leqslant i\leqslant p+q$.

Now consider the irreducible $\mathrm{SCl}_{p|q}$ module $M_{k,m}$ with a nonzero even element $|k,m\rangle$ such that

$$\begin{aligned}&\psi_i|k,m\rangle=0,\quad\text{for }i\in S_{k,m},\\&\psi_j^*|k,m\rangle=0,\quad\text{for }j\notin S_{k,m}.\end{aligned} \tag{2.3}$$

A basis for $M_{k,m}$ is given by the following elements:

$$\begin{aligned}H_I=\;&\psi_1^{*I_1}\cdots\psi_k^{*I_k}\psi_{p+1}^{*I_{p+1}}\cdots\psi_{p+m}^{*I_{p+m}}\psi_{k+1}^{I_{k+1}}\cdots\\&\times\psi_p^{I_p}\psi_{p+m+1}^{I_{p+m+1}}\cdots\psi_{p+q}^{I_{p+q}}|k,m\rangle,\end{aligned} \tag{2.4}$$

where $I=\Sigma_{j=1}^{p+q}I_j\delta_j\in\mathbf{Z}^{p+q}$, δ_j is the unit vector in $\mathbf{Z}^{p+q}$ with 1 in the jth place and zeroes elsewhere, and $I_j=0,1$ for $j\leqslant p$ and nonnegative if $j>p$.

On $\mathrm{SCl}_{p|q}$ we have an action of the Lie super algebra $\mathrm{gl}_{p|q}$ given by

$$E_{ij}=(-1)^{\bar j}\psi_i\psi_j^*,\quad i,j\in S_{p,q}. \tag{2.5}$$

Define elements of the dual of the space of diagonal matrices of $\mathrm{gl}_{p|q}$ by

$$\langle\epsilon_i,E_{jj}\rangle=\delta_{i,j},\quad i,j\in S_{p,q}. \tag{2.6}$$

Then the weights of ψ_i,ψ_j^* are ϵ_i and $-\epsilon_j$, respectively, and the weight of the vacuum $|k,m\rangle$ is $\Sigma_{i\in Sk,m}(-1)^{\bar i}\epsilon_i$. The basis (2.4) consists of weight vectors. We define an ordering on the weights by

$$\begin{aligned}&\epsilon_1>\epsilon_2>\cdots>\epsilon_k>\epsilon_{p+1}>\cdots>\epsilon_{p+m}>\epsilon_{k+1}>\cdots\epsilon_p\\&\quad>\epsilon_{p+m+1}>\cdots>\epsilon_{p+q}.\end{aligned} \tag{2.7}$$

We will call a root vector E_{ij} is positive if the corresponding weight $\epsilon_i-\epsilon_j$ is positive, i.e., if $\epsilon_i>\epsilon_j$. The vacuum $|k,m\rangle$ is a highest weight vector for this choice of positive root vectors.

Introduce on $M_{k,m}$ a grading by defining $\deg|k,m\rangle=0$ and $\deg(\psi_i)=1=-\deg(\psi_i^*),\forall i\in S_{p,q}$. Then $M_{k,m}$ decomposes:

$$M_{k,m}=\oplus_{s\in Z}M_{k,m}^{(s)}, \tag{2.8}$$

where $M_{k,m}^{(s)}$ has as basis of the H_I of (2.4) of degree s (i.e., there occur s more ψ_i's than ψ_i^*'s in H_I). The $M_{k,m}^{(s)}$ are irreducible under the action of $\mathrm{gl}_{p/q}$. The highest weight vectors are

$$|k,m,s\rangle=\begin{cases}\psi_{k+1}\cdots\psi_p\psi_{p+m+1}^{s-(p-k)}|k,m\rangle, & p-k<s,\\ \psi_{k+1}\cdots\psi_{k+s}|k,m\rangle, & 0<s\leqslant p-k,\\ |k,m\rangle, & s=0,\\ \psi_{p+m}^{*s}|k,m\rangle, & s<0.\end{cases} \tag{2.9}$$

III. SUPER PLÜCKER EQUATIONS

Let A be an arbitrary (but fixed) Grassman algebra. Define $M_{k,m}(A):=A\otimes M_{k,m}$ and put $\mathrm{gl}_{p|q}(A)=(A\otimes\mathrm{gl}_{p|q})_{\bar 0}$. Introduce the operator

$$S:=\sum_{i=1}^{p+q}(-1)^{\bar i}\psi_i\otimes\psi_i^*. \tag{3.1}$$

On the $\mathrm{gl}_{p|q}(A)$ modules $M_{k,m}(A)$ we consider the *super Plücker equation*:

0022-2488/90/040773-03$03.00

$$S(v\otimes v)=0, \tag{3.2}$$

for pure $v\in M^{(s)}_{k,m}(A)$ such that the body of v is nonzero (pure means that v is purely even or odd). (To avoid confusion let us point out that the Eq. (3.2) does not describe the embedding of a super Grassmannian in super projective space [as the purely even Plücker Eq. (1.1) does for Grassmannians]. In fact super Grassmannians are not (in general) projective, see, e.g., Manin[3].)

The supergroup associated to $\mathrm{gl}_{p|q}(A)$ is $\mathrm{Gl}_{p|q}(A)$, the group of even invertible matrices over A of size $p|q$. This group does not act in $M^{(s)}_{k,m}(A)$, but the subgroup G does, where G is the group of even invertible transformations of $M^{(s)}_{k,m}(A)$ generated by

(i) $\exp(xE_{ij}),\ x\in A_{\bar 0},\ p(\bar i)=p(\bar j)=\bar 0,\ i\neq j,$

(ii) $\exp(\theta E_{ij}),\ \theta\in A_{\bar 1},\ p(\bar i)=\bar 1, p(\bar j)=0,$

or $p(\bar i)=\bar 0,\ p(\bar j)=\bar 1,\ i\neq j,$

(iii) $\exp(xE_{ij}),\ x\in A_{\bar 0},\ p(\bar i)=p(\bar j)=\bar 1,$

$\epsilon_i>\epsilon_j.$

Note that the generators of the Lie superalgebra appearing in (i) and (ii) are nilpotent on $M_{k,m}(A)$. Although in (iii) the generator is not nilpotent it is locally nilpotent (as we are working in a highest weight module) and hence the operator in (iii) is well defined [in contrast with $\exp(xE_{ij})$ with $\epsilon_i<\epsilon_j$ and x not nilpotent].

Lemma 3.1: If $v\in M_{k,m}(A)$ is a solution of the super Plücker equation then also $g\cdot v$ is, for any $g\in G$.

A solution of (3.2) can be expanded into homogeneous elements with respect to the weight space decomposition:

$$v=\sum v_M H_M. \tag{3.3}$$

In the sequel we will denote by ϵ_M the weight of H_M.

Lemma 3.2: The homogeneous solutions of the super Plücker equation are the H_I with $i_j=0$ if $j>p$.

Proof of Lemma 3.2: For a homogeneous solution H_I all terms in (3.2) must be individually zero:

$$\psi_{\mathrm{j}}H_I\otimes\psi_j^*H_I=0. \tag{3.4}$$

The action of the generators of $\mathrm{SCl}_{p|q}$ on the homogeneous basis elements H_I is given by

$$\left.\begin{aligned}\psi_jH_I&=\begin{cases}\delta_{0I_j}H_{I+\delta_j}, & j\notin S_{k,m}\\ \delta_{1I_j}H_{I-\delta_j}, & j\in S_{k,m}\end{cases}\\ \psi_j^*H_I&=\begin{cases}\delta_{1I_j}H_{I-\delta_j}, & j\notin S_{k,m}\\ \delta_{0I_j}H_{I+\delta_j}, & j\in S_{k,m}\end{cases}\end{aligned}\right\},\quad j\leqslant p, \tag{3.5}$$

and

$$\left.\begin{aligned}\psi_jH_I&=\begin{cases}H_{I+\delta_j}, & j\notin S_{k,m}\\ I_jH_{I-\delta_j}, & j\in S_{k,m}\end{cases}\\ \psi_j^*H_I&=\begin{cases}-I_jH_{I-\delta_j}, & j\notin S_{k,m}\\ \delta_{oi_j}H_{I+\delta_j}, & j\in S_{k,m}\end{cases}\end{aligned}\right\},\quad j>p. \tag{3.6}$$

We see from this that for $j\leqslant p$ Eq. (3.4) always holds and for $j>p$ only if $i_j=0$. ■

The lemma shows that the homogeneous solutions of the super Plücker equations only occur in the spaces $M^{(s)}_{k,m}(A)$, where $-k\leqslant s\leqslant p-k$. These solutions can be identified with the weight vectors [with respect to the action $\mathrm{gl}_p(C)$] $e_{i_1}\wedge\cdots e_{i_{k+s}}$ of $\Lambda^{k+s}(C^p)$, (where $\{e_i\}_{i=1}^p$ is the standard basis of C^p and $|k,m\rangle$ is identified with $e_1\wedge e_2\wedge\cdots\wedge e_k$). From this it follows that all homogeneous solutions of (3.2) in $M^{(s)}_{k,m}(A)$ are mapped into each other by the action of elements of G. For $0\leqslant s\leqslant p-k$ the highest weight vector $|k,m,s\rangle$ is a solution and hence of these values of s all homogeneous solutions in $M^{(s)}_{k,m}(A)$ lie on the G orbit through $|k,m,s\rangle$.

Now we turn to nonhomogeneous solutions (with respect to the weight decomposition) v of (3.2). Because we demanded v to have nonvanishing body there is a component H_{M_0} of lowest weight in the decomposition (3.3) such that v_{M_0} is invertible. We may just as well assume that v_{M_0} is 1.

Lemma 3.3: H_{M_0} is a solution of the super Plücker equation.

Proof of Lemma 3.3: Consider in (3.2) the component of weight $2\epsilon_{M_0}$. In general it will look like

$$S\left(H_{M_0}\otimes H_{M_0}+\sum_{\epsilon_K+\epsilon_M=2\epsilon_{M_0}}v_KH_K\otimes v_MH_M\right)=0. \tag{3.7}$$

In the summation terms appear with either $\epsilon_K<\epsilon_{M_0}$ or $\epsilon_M<\epsilon_{M_0}$. This means that either v_K or v_M is nilpotent and has degree greater than zero in the gradation of A where the generators have degree one. Then in the gradation of the tensor square of $M_{k,m}(A)$ induced by this gradation of A the only term of degree zero in (3.7) is $S(H_{I_{j_0}}\otimes H_{I_{j_0}})=0$. ■

From Lemma 3.3 it follows that all pure solutions of (3.2) (with nonvanishing body) must be even. Moreover these purely even solutions occur only in $M^{(s)}_{k,m}(A)$ for $-k\leqslant s\leqslant p-k$.

Lemma 3.4: Let v and M_0 be as above. Then there exists an element g of G such that

$$g\cdot v=H_{M_0}+\sum v_KH_K,$$

with all v_K nilpotent.

Proof of Lemma 3.4: Let $H_J\neq H_{M_0}$ be the component of lowest weight in the decomposition (3.3) such that v_J is invertible. (If such a J does not exist, we can take g to be the identity and the lemma is proven for this v.) Consider the component of the super Plücker equation of weight $\epsilon=\epsilon_{M_0}+\epsilon_J$. It reads

$$S\left(H_{M_0}\otimes v_JH_J+v_JH_J\otimes H_{M_0}\right.$$
$$\left.+\sum_{\epsilon=\epsilon_K+\epsilon_M}v_KH_K\otimes v_MH_M\right)=0. \tag{3.8}$$

As before the terms in the summation have all degree greater than zero (in the A gradation). So taking the zero degree component of (3.8) we find after dividing by v_J

$$S(H_{M_0}\otimes H_J+H_J\otimes H_{M_0})=0. \tag{3.9}$$

Sublemma: For every $H_I, H_J\in M^{(s)}_{k,m}$, with $I\neq J$ there is an index l in $S_{p,q}$ such that $\psi_lH_I\neq 0$, $\psi_l^*H_J\neq 0$.

Proof of Sublemma: Introduce N_I, N_I^* as the numbers of ψ_i's, ψ_i^*'s in H_I:

$$N_I = \sum_{t \notin S_{k,m}} I_t, \quad N_I^* = \sum_{t \in S_{k,m}} I_t. \tag{3.10}$$

Then since H_I, H_J belong to $M_{k,m}^{(s)}$ we have $s = N_I - N_I^* = N_J - N_J^*$. Now if there is a $t \notin S_{k,m}$ such that $I_t < J_t$, then by (3.5) and (3.6) we have $\psi_t H_I \neq 0$, $\psi_t^* H_J \neq 0$, (for $J_t \geqslant 1$ and if $t \leqslant p$, $I_t = 0$, $J_t = 1$). So in this case the sublemma is proved. Suppose therefore that $I_t \geqslant J_t$, $t \notin S_{k,m}$. This implies $N_I \geqslant N_J$ and also $N_I^* \geqslant N_J^*$. If there is an $r \in S_{k,m}$ such that $I_r > J_r$ then again we have $\psi_r H_I \neq 0$, $\psi_r^* H_J \neq 0$ and we are done. So we may additionally suppose that for all $r \in S_{k,m}$, $I_r \leqslant J_r$. This implies $N_I^* \leqslant N_J^*$. But then we have $N_I = N_J$, and $N_I^* = N_J^*$. Combining $I_r \leqslant J_r$ with $N_I^* = N_J^*$ we find $I_r = J_r$, $r \in S_{k,m}$, similarly we also have $I_t = J_t$, $t \notin S_{k,m}$ and hence $I = J$ in contradiction with our assumption. ■

We continue the proof of Lemma 3.4. According to the sublemma there is in (3.9) a nonzero term $(-1)^{\bar{i}} \psi_i H_{M_0} \otimes \psi_i^* H_J$.Looking at the weights in (3.9) we see that this term can only be cancelled by a term $(-1)^{\bar{j}} \psi_j H_J \otimes \psi_j^* H_{M_0}$ (this counterterm is unique). Therefore

$$\epsilon_J = \epsilon_{M_0} + \epsilon_i - \epsilon_j. \tag{3.11}$$

This means that $E_{ij} H_{M_0} = \alpha H_J$, $\alpha = \pm 1$. Since $\epsilon_{M_0} < \epsilon_J$ the root vector E_{ij} is positive. Define now the following element of G: $g^{(1)} = \exp(-\alpha v_J Eij)$. Then $v' g^{(1)} \cdot v$ has H_{M_0} as lowest component with invertible coefficient, does not contain H_J, and other components with invertible coefficient (if any) will have higher weight than H_J. Repeating this we eliminate all components with invertible coefficients. The process terminates since we are in a highest weight module. ■

Lemma 3.5: Let $v = H_{M_0} + \Sigma v_K H_K$ be a solution of the super Plücker equation with all v_K nilpotent. Then there is a g in G such that $g \cdot v = H_{M_0}$.

Proof: Let $H_L \neq H_{M_0}$ be some weight vector occurring in v. Consider the component of (3.2) of weight $\epsilon = \epsilon_L + \epsilon_{M_0}$:

$$S\Big(H_{M_0} \otimes v_L H_L + v_L H_L \otimes H_{M_0} + \sum_{\epsilon = \epsilon_k + \epsilon_M} v_K H_K \otimes v_M H_M\Big) = 0. \tag{3.12}$$

By the sublemma there is an index l such that $((-1)^l \psi_l \otimes \psi_l^*)(H_{M_0} \otimes v_L H_L) \neq 0$. This term must be compensated by some term $((-1)^k \psi_k \otimes \psi_k^*) \times (v_L H_l \otimes H_{M_0})$ or $((-1)^k \psi_k \otimes \psi_k^*)(v_K H_K \otimes v_M H_M)$. As in the proof of Lemma 3.4 this means that there is in v a component H_P with $\epsilon_P = \epsilon_{M_0} + \epsilon_l - \epsilon_k$ and $E_{lk} H_{M_0} = \alpha H_P, \alpha = \pm 1$. Define $g^{(1)} = \exp(-\alpha v_P E_{lk})$. It belongs to G since v_P is nilpotent. Then $v' = g^{(1)} \cdot v$ does not contain a component along H_P. By repeating this we eliminate from v components with nilpotent coefficients. It may seem that we will introduce in this way many new components (of possibly lower degree). This is true, but we must then remember that all v_k are nilpotent. So if there are say x different v_K's, then only products of up to x v_K's are potentially nonzero and our process stops after a finite number of steps. ■

We can summarize the conclusion of this paper in the following theorem.

Theorem: A pure element v of $M_{k,m}^{(s)}(A)$ with nonvanishing body solves the super Plücker equation

$$S(v \otimes v) = 0$$

if and only if v is even, lies on the G orbit through the homogeneous solutions and $-k \leqslant s \leqslant p - k$. If s is larger or equal than zero v lies on the G orbit through the vacuum $|k,m,s\rangle$.

ACKNOWLEDGMENTS

It is a pleasure to thank T. Shiota, J. van de Leur, K. Ueno, and V. G. Kac for interesting conversations.

[1] D. H. Peterson and V. G. Kac, Proc. Natl. Acad. Sci. USA **80**, 1778–1782 (1983).

[2] V. G. Kac and J. W. van de Leur, Ann. Inst. Fourier, Grenoble **37**, 99–137 (1987).

[3] Y. I. Manin, *Gauge Field Theory and Complex Geometry* (Springer, Berlin, 1988).

Some aspects of quantum groups and supergroups

E. Corrigan, D. B. Fairlie, and P. Fletcher
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, England
R. Sasaki
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, England and Research Institute for Theoretical Physics, Hiroshima University, Takehara, Hiroshima 725, Japan

(Received 20 July 1989; accepted for publication 8 November 1989)

Some features of Manin's construction of quantum groups are developed and extended to supergroups.

I. INTRODUCTION

In recent years a great deal of activity has been directed toward the exploration of quantum groups and algebras.[1,2] These structures may be thought of as matrix groups in which the elements are themselves noncommutative, obeying sets of bilinear product relations, and as a deformation of ordinary Lie algebras, respectively. They arise in quantum inverse scattering theory and as representations of transfer matrices in statistical mechanics. In these cases the sufficient condition for associativity of the algebra turns out to be the Yang–Baxter[3] relation, the analog of the Jacobi identity for quantum groups. In this paper we shall concentrate upon certain algebraic aspects of the theory and develop ideas arising principally from the viewpoint of Manin,[4] who considers a quantum group as effecting linear transformations upon a space whose elements, or coordinates, are noncommutative. The conditions for such a mapping to be an endomorphism constitute the quantum group relations. In fact, since this idea underlies the classical transformation groups, it is a very natural, though at first sight unfamiliar, approach to the deformation of classical groups.

We recount the idea of Manin for the simplest example of $\mathrm{GL}_q(2)$ and develop properties of this quantum group. A natural extension leads to the definition of the dual group $\mathrm{GL}_q(2)$ whose elements are Grassmannian, and $\mathrm{GL}_q(1|1)$ which is connected with the quantum extension of the superalgebra $\mathrm{SU}(1|1)$, just as $\mathrm{GL}_q(2)$ is connected with the quantum extension of $\mathrm{SU}(2)$. Quantum superalgebras and groups have been studied before.[5] These groups are then displayed in a more familiar way as bilinear relations specified by an R matrix satisfying the Yang–Baxter equation

$$R_{12}R_{13}R_{23}=R_{23}R_{13}R_{12}$$

as a guarantee of associativity, and the generalization to $\mathrm{GL}_q(N)$, $\mathrm{GL}_q(N|M)$ is made. The minimal set of relations imposed by the R-matrix relations is shown to be equivalent to those imposed by Manin's construction, and in fact Manin's construction can be used to infer the structure of the R matrix for the classical groups.

II. MANIN'S CONSTRUCTION

Manin introduces what he calls the quantum plane $R_q[2,0]$, whose elements are pairs $x=(x,y)$, whose components x,y are assumed to satisfy the algebraic relation

$$xy=q^{-1}yx, \tag{2.1}$$

where q is a complex number. The components neither commute nor anticommute unless $q=\pm 1$, respectively. A Grassmannian quantum plane $R_q[0,2]$ dual to the (x,y) plane is also introduced, with elements $\xi=(\xi,\eta)$ which are required to satisfy

$$\xi^2=0,\quad \eta^2=0,\quad \xi\eta+q\eta\xi=0. \tag{2.2}$$

Now consider a matrix

$$M=\begin{pmatrix} a & b\\ c & d\end{pmatrix}\in\mathrm{GL}_q(2),$$

which effects simultaneously linear transformations of the quantum plane and its dual,

$$x'=Mx\in R_q[2,0],$$
$$\xi'=M\xi\in R_q[0,2]. \tag{2.3}$$

The images x', ξ' are supposed to lie in the appropriate planes, i.e., their components satisfy (2.1) and (2.2). (The elements of M are supposed to commute with x,y,ξ,η.) This condition imposes restrictions upon M, giving the $\mathrm{GL}_q(2)$ relations

$$\begin{aligned} &ab=q^{-1}ba,\quad cd=q^{-1}dc,\\ &ac=q^{-1}ca,\quad bc=cb,\\ &bd=q^{-1}db,\quad ad-da=(q^{-1}-q)bc.\end{aligned} \tag{2.4}$$

Using these relations, it is easy to show that $\mathrm{Det}_q\, M=ad-q^{-1}bc$ commutes with all the elements a,b,c,d and thus may be considered as a number, the "quantum determinant." The choice $\mathrm{Det}_q\, M=1$ restricts the quantum "group" to $\mathrm{SL}_q(2)$ by analogy with the classical restriction to the special linear group. Because $\mathrm{Det}_q\, M$ commutes with the elements of M there exists an inverse

$$M^{-1}=(\mathrm{Det}_q\, M)^{-1}\begin{pmatrix} d & -qb\\ -q^{-1}c & a\end{pmatrix}, \tag{2.5}$$

which is both a left and right inverse for M. Note that M^{-1} is a member of $\mathrm{GL}_{q^{-1}}(2)$ rather than $\mathrm{GL}_q(2)$, and thus $\mathrm{GL}_q(2)$ is not, strictly speaking, a group. Furthermore, it is clear that if

0022-2488/90/040776-05$03.00

$$M=\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad M'=\begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \in \mathrm{GL}_q(2),$$

and (a,b,c,d) pairwise commute with (a',b',c',d') then MM' and $M'M$ are both $\mathrm{GL}_q(2)$ matrices. Also

$$\mathrm{Det}_q(MM') = \mathrm{Det}_q(M'M) = (\mathrm{Det}_q\, M)(\mathrm{Det}_q\, M'), \tag{2.6}$$

reinforcing the identification with a determinant.

The algebra (2.4) is associative under multiplication and the relations may be reexpressed in a tensor product form

$$R_{ijkl}M_{km}M_{ln} = M_{jl}M_{ik}R_{klmn}, \tag{2.7}$$

where R_{ijkl} is a matrix, whose explicit form is given by

$$\begin{array}{cc} & (k,l) \\ (i,j) & \begin{pmatrix} q^{-1} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & q^{-1}-q & 1 & 0 \\ 0 & 0 & 0 & q^{-1} \end{pmatrix}, \end{array} \tag{2.8}$$

where the rows are all pairs (i,j), $i,j = 1,2$ in natural order, and similarly the columns are pairs (k,l). Expression (2.8) is a member of a general class of R matrices, each labeled by an additional parameter x, and each associated with one of the classical affine Lie algebras.[2] An explicit form of the R matrices for the classical series is given by Jimbo. For $\hat{A}_n$ it is

$$R(x) = (q^{-1}-xq)\sum E_{\alpha\alpha}\otimes E_{\alpha\alpha} + (1-x)\sum_{\alpha\neq\beta} E_{\alpha\alpha}\otimes E_{\beta\beta} + (q^{-1}-q)\Big(\sum_{\alpha<\beta} + x\sum_{\alpha>\beta}\Big)E_{\alpha\beta}\otimes E_{\beta\alpha}. \tag{2.9}$$

In this expression, the indices i, j, k, l have been suppressed for the sake of clarity. The i,jth element of the matrix $E_{\alpha\beta}$ is given by

$$(E_{\alpha\beta})_{ij} = \delta_{i\alpha}\delta_{j\beta}.$$

For $\hat{A}_1$, and $x=0$, the matrix (2.8) is recovered. The R-matrix (2.9) satisfies the well-known Yang–Baxter relation

$$R_{12}(x)R_{13}(xy)R_{23}(y) = R_{23}(y)R_{13}(xy)R_{12}(x), \tag{2.10}$$

which for $x,y=0$ is a sufficient condition for the associativity of the quantum matrices.

As Manin's construction ensures associativity, it might thus be employed as a method for the construction of Yang–Baxter R matrices satisfying (2.10) by reexpressing the quantum group relations in the form (2.7) and identifying R. However, even then (2.10) is not guaranteed, as we shall see later.

There is also a second curious property applicable to 2×2 quantum groups. It asserts that if $M\in\mathrm{GL}_q(2)$ then $M^n \in \mathrm{GL}_{q^n}(2)$. It is elegant but appears neither to generalize nor to fit into a proper algebraic scheme. [The product is not to be confused with comultiplication[1] which preserves (2.4).] It is proved in the Appendix.[6]

III. QUANTUM SUPERGROUPS

Returning to the quantum plane (x,y) and its dual (ξ,η), suppose we postulate a linear transformation $\hat{M}$ which maps the plane into its dual and vice versa, i.e.,

$$\xi' = \hat{M}x, \quad x' = \hat{M}\xi, \tag{3.1}$$

and again impose the quantum plane conditions upon (ξ',η') and (x',y'). If the elements of $\hat{M}$ are designated by

$$\hat{M} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix},$$

then the constraints are ten in number:

$$\begin{aligned} &\alpha\beta + q\beta\alpha = 0, \quad \alpha\gamma + q\gamma\alpha = 0, \\ &\beta\delta + q\delta\beta = 0, \quad \gamma\delta + q\delta\gamma = 0, \\ &\alpha\delta + \delta\alpha = 0, \\ &\beta\gamma + \gamma\beta + (q-q^{-1})\delta\alpha = 0, \\ &\alpha^2 = \beta^2 = \gamma^2 = \delta^2 = 0. \end{aligned} \tag{3.2}$$

These relations may be considered as a deformation of a Grassmann algebra on four elements $(\alpha,\beta,\gamma,\delta)$. As with the quantum matrix, they may be expressed in terms of an $\hat{R}$ matrix in the form (2.7),

$$\hat{R}\hat{M}\hat{M} = -\hat{M}\hat{M}\hat{R}, \tag{3.3}$$

where

$$\hat{R} = \begin{pmatrix} q+q^{-1} & 0 & 0 & 0 \\ 0 & 2 & q-q^{-1} & 0 \\ 0 & -(q-q^{-1}) & 2 & 0 \\ 0 & 0 & 0 & q+q^{-1} \end{pmatrix}. \tag{3.4}$$

Note that in the classical limit (i.e., $q\rightarrow 1$) $\hat{R}$ becomes twice the identity matrix. This matrix $\hat{R}$ is (2.9) evaluated at $x=-1$. Notice also that although the algebra (3.2) is an associative algebra of the matrix elements of $\hat{M}$, $\hat{R}$ does not satisfy the Yang–Baxter equation (2.10), thus demonstrating that the Yang–Baxter relation is not a necessary condition for associativity.

Since $\hat{M}$ is entirely Grassmannian, an inverse proper cannot exist. However, the analog of left and right adjugate matrices can be constructed, giving

$$\begin{pmatrix} q\delta & \beta \\ -\gamma & -q^{-1}\alpha \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = (\beta\gamma + q\delta\alpha)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.5}$$

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}\begin{pmatrix} -q^{-1}\delta & \beta \\ -\gamma & q\alpha \end{pmatrix} = (\gamma\beta + q\delta\alpha)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.6}$$

The combination $\beta\gamma + q\delta\alpha$ may be thought of as a left quantum determinant and Δ_L and $\gamma\beta + q\delta\alpha$ as a right quantum determinant Δ_R. The expressions Δ_L, Δ_R satisfy the relation

$$\Delta_L\begin{pmatrix} -q^{-1}\delta & \beta \\ -\gamma & q\alpha \end{pmatrix} = \begin{pmatrix} q\delta & \beta \\ -\gamma & -q^{-1}\alpha \end{pmatrix}\Delta_R, \tag{3.7}$$

which is a consequence of (3.5) and (3.6) and associativity.

In a similar manner one can construct the quantum analog of $\mathrm{GL}(1|1)$, which we call $\mathrm{GL}_q(1|1)$, the group of linear transformations acting upon a quantum superplane with one bosonic and one fermionic coordinate. (We use the convention of roman script for bosonic quantities, greek for fermionic.) Consider a quantum superplane and its dual;

$$\begin{pmatrix}x\\ \xi\end{pmatrix}, \quad \begin{pmatrix}\eta\\ y\end{pmatrix},$$

satisfying

$$\begin{matrix} x\xi - q^{-1}\xi x = 0, \\ \xi^2 = 0, \end{matrix} \quad \text{and} \quad \begin{matrix} \eta^2 = 0, \\ \eta y - qy\eta = 0. \end{matrix} \tag{3.8}$$

Define a $\mathrm{GL}_q(1|1)$ matrix,

$$\mathscr{M} = \begin{pmatrix} a & \beta\\ \gamma & d \end{pmatrix},$$

and require

$$\mathscr{M} = \begin{pmatrix}x\\ \xi\end{pmatrix} = \begin{pmatrix}x'\\ \xi'\end{pmatrix}, \quad \mathscr{M}\begin{pmatrix}\eta\\ y\end{pmatrix} = \begin{pmatrix}\eta'\\ y'\end{pmatrix},$$

and impose (3.8) once again on the transformed variables. We assume that β and γ anticommute with ξ and η. Then we obtain eight relations

$$\begin{aligned} &a\beta = q^{-1}\beta a, \quad a\gamma = q^{-1}\gamma a,\\ &d\beta = q^{-1}\beta d, \quad d\gamma = q^{-1}\gamma d,\\ &\beta^2 = 0, \quad \gamma^2 = 0, \quad \beta\gamma + \gamma\beta = 0,\\ &ad - da + q^{-1}\beta\gamma + q\gamma\beta = 0. \end{aligned} \tag{3.9}$$

In this case the left and right inverses may be defined and are equal;

$$\begin{aligned} \mathscr{M}_L^{-1} &= \begin{pmatrix} \Delta_1^{-1}d & -\Delta_1^{-1}q^{-1}\beta\\ -\Delta_2^{-1}q^{-1}\gamma & \Delta_2^{-1}a \end{pmatrix}\\ &= \begin{pmatrix} d\Delta_1^{-1} & -q\beta\Delta_2^{-1}\\ -q\gamma\Delta_1^{-1} & a\Delta_2^{-1} \end{pmatrix} = \mathscr{M}_R^{-1}, \end{aligned}$$

where $\Delta_1 = ad - q\beta\gamma$ and $\Delta_2 = da - q\gamma\beta$. The theorems in Sec. II also apply to $\mathrm{GL}_q(1|1)$. In particular, if $\mathscr{M}\in\mathrm{GL}_q(1|1)$ then $\mathscr{M}^n\in\mathrm{GL}_{q_n}(1|1)$. Similar results may be deduced for the dual matrix

$$\hat{\mathscr{M}} = \begin{pmatrix} \alpha & b\\ c & \delta \end{pmatrix},$$

which transforms the superplane into its dual.

IV. GENERALIZATIONS TO GL$_q$(*N*) AND GL$_q$(*N*|*M*)

It is obviously desirable to extend the analysis to the quantum analogs of linear transformations in higher dimensional spaces. Consider first $\mathrm{GL}_q(N)$. Instead of the quantum two-plane, take a vector

$$x = \begin{pmatrix} x_1\\ x_2\\ \vdots\\ x_N \end{pmatrix} \in R_q[N,0],$$

and impose the relations

$$x_ix_j - q^{-1}x_jx_i = 0 \quad \text{for } i<j. \tag{4.1}$$

Adjoin a dual quantum space

$$\xi = \begin{pmatrix} \xi_1\\ \xi_2\\ \vdots\\ \xi_N \end{pmatrix} \in R_q[0,N],$$

with the relations

$$\xi_i^2 = 0, \quad \xi_i\xi_j + q\xi_j\xi_i = 0 \quad \text{for } i<j. \tag{4.2}$$

The relations (4.1) and (4.2) can be written in the form

$$x^TG_{kl}x = 0, \quad \xi^TF_{kl}\xi = 0,$$

where G_{kl} is a matrix whose entries are all zero except for the klth and the lkth, i.e.,

$$(G_{kl})_{rs} = \frac{\sqrt{q}}{\sqrt{q+q^{-1}}}\delta_{rk}\delta_{sl} - \frac{\sqrt{q^{-1}}}{\sqrt{q+q^{-1}}}\delta_{rl}\delta_{sk}, \quad k<l. \tag{4.3}$$

Similarly,

$$\begin{aligned} (F_{kl})_{rs} &= \frac{\sqrt{q^{-1}}}{\sqrt{q+q^{-1}}}\delta_{rk}\delta_{sl} + \frac{\sqrt{q}}{\sqrt{q+q^{-1}}}\delta_{rl}\delta_{sk}, \quad k<l,\\ (F_{kk})_{rs} &= \delta_{rk}\delta_{sk}. \end{aligned} \tag{4.4}$$

Now

$$(G_{ij})_{rs}(F_{kl})_{rs} = \mathrm{Tr}(G_{ij}F^T_{kl}) = 0 \tag{4.5}$$

by construction. This enables us to write the quantum matrix condition very succinctly. Suppose the matrix of linear transformations is given by M, i.e.,

$$x' = Mx, \quad \xi' = M\xi. \tag{4.6}$$

Then, $x^TM^TG_{ij}Mx = 0$ implies that $M^TG_{ij}M$ is a linear combinations of G's, i.e.,

$$M^TG_{ij}M = \sum_{k,l} A_{ijkl}G_{kl}, \tag{4.7}$$

and similarly, $\xi^TM^TF_{ij}M\xi = 0$ implies

$$M^TF_{ij}M = \sum_{k,l} B_{ijkl}F_{kl}. \tag{4.8}$$

Due to orthogonality (4.5) we have sets of relations

$$\mathrm{Tr}(M^TG_{ij}MF^T_{kl}) = 0, \quad \mathrm{Tr}(M^TF_{ij}MG^T_{kl}) = 0. \tag{4.9}$$

The number of relations of the first kind is simply the number of independent G's, $\frac{1}{2}N(N-1)$ multiplied by the number of independent F's, $\frac{1}{2}N(N+1)$ giving $\frac{1}{4}(N^4 - N^2)$, and similarly for the second kind, resulting in $\frac{1}{2}N^2(N^2-1)$ relations, the full set for $\mathrm{GL}_q(N)$. Notice that the relations (4.9) imply also that M^T is a quantum matrix, as it satisfies the same bilinear algebra. In fact, we can dispense with the dual (Grassmannian) plane in setting up the quantum group conditions; we can take simply

$$\begin{aligned} x' &= Mx\in R_q[N,0],\\ x'' &= M^Tx\in R_q[N,0]. \end{aligned} \tag{4.10}$$

In the classical case $q = 1$, and G_{ij} spans the space of antisymmetric matrices, while the F_{ij} spans the symmetric ones. We can refer to the quantum case of G_{ij} as q-antisymmetric, and F as q-symmetric matrices.

It is now relatively easy to construct an R matrix, and to exhibit these relations in the form of Eq. (2.7).

Define the $N^2\times N^2$ matrix

$$R(\mu,\nu) = \mu\sum_{i,j}(G_{ij})^TG_{ij} + \nu\sum_{k,l}(F_{kl})^TF_{kl}, \tag{4.11}$$

where μ and ν are arbitrary parameters. Then, on account of the orthogonality relations (4.5) together with the additional orthonormality conditions

$$\mathrm{Tr}(G^T_{ij}G_{kl}) = \delta_{ik}\delta_{jl}, \tag{4.12}$$

$$\mathrm{Tr}(F^T_{ij}F_{kl}) = \delta_{ik}\delta_{jl}, \tag{4.13}$$

Eq. (4.11), written with explicit indices as

$$R_{st,uv}(\mu,\nu) = \mu \sum_{i,j} (G_{ij})_{ts}(G_{ij})_{uv} + \nu \sum_{i,j} (F_{ij})_{ts}(F_{ij})_{uv} \tag{4.14}$$

is just the eigenvalue expansion of an $N^2 \times N^2$ matrix with two degenerate eigenvalues with degeneracies $\frac{1}{2}N(N-1)$ and $\frac{1}{2}N(N+1)$. The sets of quantities $(G_{ij})_{st}$, $(F_{ij})_{st}$ are eigenvectors in the sense that

$$\begin{aligned}(G_{ij})_{ts}R_{st,uv} &= \mu(G_{ij})_{uv}, \\ (F_{ij})_{ts}R_{st,uv} &= \nu(F_{ij})_{uv},\end{aligned} \quad \text{or} \quad \begin{aligned}G^T_{ij}R &= \mu G_{ij}, \\ F^T_{ij}R &= \nu F_{ij}.\end{aligned}$$

Imposing the conditions

$$R_{\sigma\tau pq}(\mu,\nu)M_{pu}M_{qv} = M_{\tau q}M_{\sigma p}R_{pquv}(\mu,\nu) \tag{4.15}$$

produces a set of equations whose content is just (4.9), as may be readily derived by taking the trace of (4.15) with $(G_{ij})^T F_{kl}$ and $(F_{ij})^T G_{kl}$. The orthogonality properties (4.5), (4.12), and (4.13) ensure that G_{ij} and F_{ij} are eigenvectors of R, and since the eigenvalues differ, the equations (4.9) are a consequence of (4.15). Note, however, that the relations (4.15) are not all necessarily independent, while (4.9) are, by construction. No further conditions result from taking the trace of (4.15) with the combinations $(F_{ij})^T F_{kl}$ and $(G_{ij})^T G_{kl}$. It is easy to see that (4.11) gives $R(x)$, (2.9) for $\mu = -q + xq^{-1}$, $\nu = q^{-1} - xq$.

The extension for the dual Grassmann matrix $\widehat{M}$ is very much the same. Postulate a similar ansatz for the $\widehat{R}$ matrix, but with different eigenvalues, μ,ν. Then impose

$$\widehat{R}_{\sigma\tau pq}(\mu,\nu)\widehat{M}_{pu}\widehat{M}_{qv} = -\widehat{M}_{\tau q}\widehat{M}_{\sigma p}\widehat{R}_{pquv}(\mu,\nu). \tag{4.16}$$

The eigenvalues of $\widehat{R}$ are $\pm(q+q^{-1})$, i.e., $x = +1$ in (2.9). This fact has the consequence that this time the matrix elements of this relation that do not vanish are those of the trace with $G_{ij}(G_{kl})^T$ and $F_{ij}(F_{kl})^T$, while those with a mixed G and F are automatically satisfied, thus giving $\frac{1}{2}N^2(N^2+1)$ independent relations for the quantum Grassmann group

$$\mathrm{Tr}(\widehat{M}^T G_{ij}\widehat{M}G^T{}_{kl}) = 0, \tag{4.17}$$

$$\mathrm{Tr}(\widehat{M}^T F_{ij}\widehat{M}F^T{}_{kl}) = 0. \tag{4.18}$$

These provide the generalization of (3.2) to arbitrary N.

This generalization gives the class of R matrices associated with the Lie groups of the A_n series.[2] We might also enquire about the corresponding extension to other series, e.g., the C_n series. What we must do to obtain the C_n, i.e., the Sp$(2n)$ series, is to adjoin to the quantum plane conditions (4.6) an additional symplectic requirement,

$$M^T\epsilon M = \lambda\epsilon, \tag{4.19}$$

where ϵ is an $N \times N$ matrix ($N = 2n$), with nonvanishing elements only for $i+j = N+1$, i.e., on the antidiagonal, where they are

$$q^N,\ldots,q^{(N/2)+1},q^{(N/2)-1},\ldots,q,1.$$

We can write the quantum "group" condition in a form analogous to (4.9) after redefining matrices

$$G'_{ij} = G_{ij} - [2\,\mathrm{Tr}(\epsilon^T G_{ij})/Nq^N]\epsilon, \tag{4.20}$$

$$F'_{ij} = F_{ij} - [2\,\mathrm{Tr}(\epsilon^T F_{ij})/Nq^N]\epsilon \tag{4.21}$$

so that they are orthogonal to ϵ in the sense of (4.17) and (4.18). Then the quantum conditions can be written as

$$\begin{aligned}\mathrm{Tr}(M^T G'_{ij} M F'_{kl}{}^T) &= \mathrm{Tr}(M^T F'_{ij} M G'_{kl}{}^T) = 0, \\ \mathrm{Tr}(M^T G'_{ij} M\epsilon^T) &= \mathrm{Tr}(M^T F'_{ij} M\epsilon^T) = 0, \\ \mathrm{Tr}(M^T\epsilon M G'_{ij}{}^T) &= \mathrm{Tr}(M^T\epsilon M F'_{ij}{}^T) = 0.\end{aligned} \tag{4.22}$$

The number of such relations is

$$\tfrac{1}{2}N(N-1)(N^2+N+2) - 2. \tag{4.23}$$

For $N = 2$ this gives 6, as before, and for $N = 4$ it gives 130, a number which agrees with computer calculations in REDUCE, using the Sp(2) R matrix of Jimbo to define quantum group conditions via (2.7).

In an analogous fashion the dual group relations can be found by replacing (4.22) by

$$\begin{aligned}\mathrm{Tr}(\widehat{M}^T G'_{ij}\widehat{M}G'_{kl}{}^T) &= \mathrm{Tr}(\widehat{M}^T F'_{ij}\widehat{M}F'_{kl}{}^T) \\ &= \mathrm{Tr}(\widehat{M}^T\epsilon^T\widehat{M}\epsilon) = 0,\end{aligned} \tag{4.24}$$

the number of relations being

$$\tfrac{1}{2}N^4 - \tfrac{1}{2}N^2 + N + 2. \tag{4.25}$$

As is to be expected, this is complementary to the previous calculation; the sum of (4.23) and (4.25) is N^4.

V. FURTHER GENERALIZATIONS

The assumptions made for the quantum hyperplane conditions (4.1) and (4.2) need not be the only viable structures. In fact, there is a natural generalization of the Clifford sequence. Start with a quantum plane (x,y) and its dual (ξ,η). Then construct the quantum matrix M and its dual $\widehat{M}$. Now view the elements of M as constituting the coordinates a, b, c, d in a quantum hyperplane, with $\widehat{M}$ furnishing the dual coordinates, and take the relations (2.4) and (3.2) as those to be preserved by linear transformations M', $\widehat{M}'$ acting upon the quantum hyperplanes. This leads to conditions on the 16 elements of M' and those of $\widehat{M}'$, which in turn can be thought of as the requirements for a 16 dimensional hyperplane, subject to a linear transformation M'' etc. This sequence will generate a quantum Clifford sequence.

This approach to quantum groups raises the obvious question of the representation of the elements of the quantum plane, and of the quantum matrix itself, by finite-dimensional matrices whose elements themselves commute. That such representations do exist with q an nth root of unity is demonstrated by setting $x = g$ and $y = h$, where g,h are $n \times n$ matrices given by

$$g=\begin{pmatrix}1 & 0 & 0 & \cdots & 0\\ 0 & \omega & 0 & \cdots & 0\\ 0 & 0 & \omega^2 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \omega^{n-1}\end{pmatrix},$$

$$h=\begin{pmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ 1 & 0 & 0 & \cdots & 0\end{pmatrix},$$

$$g^n=h^n=I,\quad \omega^n=1.$$

It is easy to verify that $gh=\omega^{-1}hg$, the quantum plane condition.

It is difficult to find representations of M with $\mathrm{Det}_q(M)\neq 0$ and $q\neq\pm 1$. A specific example for A_1 is

$$M=\begin{pmatrix}g^2 & \omega h\\ \omega h^4 & g^4(1+h^5)\end{pmatrix},$$

where $n=6$ and $q=\omega^2$.

It is also possible to create infinite-dimensional representations, though degenerate, i.e., with a vanishing $\mathrm{Det}_q(M)$. One such representation is

$$M=\begin{pmatrix}e^{ia(p_1+p_2)} & e^{ia(x_1+p_2)}\\ e^{ia(p_1+x_2)} & e^{ia(x_1+x_1)}\end{pmatrix},$$

where p_1, p_2 and x_1,x_2 satisfy the commutation rules appropriate to canonically conjugate variables and $q=e^{-ia^2}$. We have not found any such example which is not degenerate.

For A_n quantum groups a representation of the quantum hyperplane (4.1) is given by

$$\begin{aligned}
x_1&=x\otimes x\otimes x\otimes\cdots\otimes x,\\
x_2&=y\otimes x^2\otimes x^2\otimes\cdots\otimes x^2,\\
x_3&=x\otimes y\otimes x^2\otimes\cdots x^2,\\
\vdots&\quad\vdots\\
x_{n+1}&=x\otimes x\otimes\cdots\otimes x\otimes y.
\end{aligned}$$

ACKNOWLEDGMENTS

We are grateful to C. Tunstall, J. Uretsky, and C. K. Zachos for useful comments. R. S. thanks the Department of Mathematical Sciences at Durham for hospitality.

P. F. wishes to thank the University of Durham for a Research Studentship.

APPENDIX

To provide a proof that if $M\in\mathrm{GL}_q(2)$ then $M^n\in\mathrm{GL}_{q^n}(2)$, we may proceed as follows. The quantum group relations (2.4) can be rewritten suggestively as

$$M\epsilon M^{T_q}\propto\epsilon,\quad M^{T_q}\epsilon M\propto\epsilon, \tag{A1}$$

where

$$\epsilon=\begin{pmatrix}0 & +1\\ -1 & 0\end{pmatrix}$$

and T_q is an operation similar to transportation, defined by

$$M^{T_q}=\sigma_q{}^{-1}M^T\sigma_q,\quad \sigma_q=\begin{pmatrix}\sqrt{q} & 0\\ 0 & 1/\sqrt{q}\end{pmatrix}.$$

Actually, (A1) provided the motivation for a number of the results in this article, and looks suspiciously like a "quantum" version of Sp(2). To prove the assertion concerning quantum $\mathrm{GL}_q(2)$ matrices, it is clearly enough to show that

$$(M^n)^{T_{q^n}}=(M^{T_q})^n. \tag{A2}$$

The following elegant inductive argument is due to C. Tunstall. Let

$$M_n=\begin{pmatrix}g_n(a,d,bc,q) & bf_n(d,a,bc,q^{-1})\\ cf_n(a,d,bc,q) & g_n(d,a,bc,q^{-1})\end{pmatrix}, \tag{A3}$$

where $g_0=1, f_0=0$, and

$$f_n(a,d,bc,q)=q^{1-n}f_n(d,a,bc,q^{-1}),$$

$$(dq-a)f_n(a,d,bc,q)-q^ng_n(d,a,bc,q^{-1})$$
$$\quad+g_n(a,d,bc,q)=0.$$

Then the induction steps are easily checked and

$$M_{n+1}=MM_n. \tag{A4}$$

We now use this to observe that M_n can be identified with M^n. Introducing the abbreviations

$$\begin{aligned}g_n&=g_n(a,d,bc,q), & & & \tilde{g}_n&=g_n(d,a,bc,q^{-1}),\\ f_n&=f_n(a,d,bc,q), & &\text{and} & \tilde{f}_n&=f_n(d,a,bc,q^{-1}).\end{aligned}$$

and using the above definitions, we find

$$(M^n)^{T_{q^n}}=(\sigma_{q^n})^{-1}(M^n)^T(\sigma_{q^n})=\begin{pmatrix}g_n & (c/q)\tilde{f}_n\\ qbf_n & \tilde{g}_n\end{pmatrix}$$

and

$$(M^{T_q})^n=\begin{pmatrix}a & c/q\\ qb & d\end{pmatrix}^n=\begin{pmatrix}g_n & (c/q)\tilde{f}_n\\ qbf_n & \tilde{g}_n\end{pmatrix}, \tag{A5}$$

yielding the desired result.

[1] P. P. Kulish and E. K. Skylanin, in *Integrable Field Theories*, edited by J. Hietarinta and C. Montonen, *Lecture Notes in Physics*, Vol. 151 (Springer, Berlin, 1981); L. D. Faddeev, N. Reshetikhin, and L. A. Takhtajan, "Quantisation of Lie groups and Lie algebras," LOMI preprint E-14-87, Leningrad (1987); V. Drinfeld, "Proceedings ICM," Berkeley (1986); L. A. Takhtajan, "Quantum Groups and Integrable Models," Proceedings of the Taniguchi Symposium, Kyoto (1988).

[2] M. Jimbo, Commun. Math. Phys. **102**, 537 (1986); V. Drinfeld, Sov. Math. Dokl. **32**, 254 (1985).

[3] C. N. Yang, Phys. Rev. **19**, 1312 (1967); R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic, London, 1982).

[4] Yu. I. Manin, Ann. Inst. Fourier (Grenoble) **37**, 191 (1987).

[5] P. P. Kulish, "Quantum superalgebra osp(2|1)," Kyoto preprint RIMS-615 (1988); Yu. I. Manin, Commun. Math. Phys. **123**, 163 (1989); C. Devchand, "A q-analogue of the Lie superalgebra osp(2|1) and its metaplectic representation," Frieburg preprint (1989).

[6] This observation has also been noted by Kobyzev (Yu. I. Manin, private communication) and by Zumino (private communication); S. Vokos, B. Zumino, and J. Wess, "Properties of quantum 2×2 matrices," LAPP-TH-253/89 (1989).

Classification of unitary highest weight representations for noncompact real forms

Juan Garcia-Escudero[a] and Miguel Lorente
Departamento de Física, Universidad de Oviedo, 33007 Oviedo, Spain

(Received 25 July 1989; accepted for publication 6 December 1989)

Using Jakobsen theorems, unitarizability in Hermitian symmetric spaces is discussed. The set of all missing highest weights is explicitly calculated and the construction of their corresponding highest weights vectors is studied.

I. INTRODUCTION

One of the new methods for the construction of representations for semisimple Lie algebras is based on enveloping algebras.

Irreducible representations of simple Lie Algebras arise if we take the quotient space of Verma modules with respect to some invariant subspaces generated by highest weight vectors.[1] When a scalar product is induced in these Verma modules infinitesimally unitary representations can be defined.[2]

In this paper we discuss the unitarizability on noncompact real forms following the Jakobsen method.[3–5] He uses the Bernshtein, Gel'fand, and Gel'fand theorem and the scalar product induced by a sesquilinear form introduced by Harish–Chandra in Refs. 6 and 7. We use his method to obtain a complete and explicit classification of the highest weights that we must exclude in order to unitarize when the reduction level is strictly higher than one. In the examples and when the expressions are not extremely long, we illustrate the procedure by writing the corresponding highest weight vectors that generate the invariant subspaces in which the sesquilinear form vanishes.

There exists other methods in the literature[8–11] following different paths but arriving at the same final results.

Indecomposable representations have found application in physics for a long time.[12] Therefore, certain types of indecomposable representations are associated with the Poincaré algebra, the algebra of the Euclidean group, and others.[13] For application of the algebras treated here see Refs. 14 and 15.

In this work we consider Hermitian symmetric spaces for which the reduction level may be higher than one: $su(p,q)$, $sp(n,\mathbb{R})$, $so^*(2n)$, e_6, and e_7. In Sec. II we give some concepts that will be needed. In Sec. III we describe the Jakobsen method by means of a step series and we state how to construct the highest weight vectors. In Sec. IV we introduce the concept of height and we give a notation that allow us to easily localize the noncompact roots in Jakobsen diagrams. In addition, we use the Jakobsen diagrams to obtain, in a very simple way, the split rank that is useful for the calculation of the λ_s parameter. Finally in Sec. V we first study general cases giving the sets of all highest weights that will be missing and then apply the method to some examples.

II. PRELIMINARIES

Let g be a semisimple Lie algebra over $\mathscr{R}$ and let g^c be its complexification. Let $B(X,Y) = \mathrm{tr}(\mathrm{ad}\, X\ \mathrm{ad}\, Y)$; $X,Y \in g^c$ be the Killing form. A real form g_0 of g^c is called compact if $B(X,X) < 0$ for each $X \in g_0$ and an automorphism θ of g^c exists such that

$$\theta g_0 \subset g_0, \quad \theta g \subset g,$$

and

$$g = k + p, \quad g_0 = k + ip,$$

where $i = \sqrt{-1}$, k is the set of all $X \in g$ such that $\theta X = X$, and p is the set of all $Y \in g$ such that $\theta Y = -Y$.

Let k^c and p^c be the subspaces of g^c spanned by k,p, respectively, over $\mathcal{C}$. It holds

$$[k^c,k^c] \subset k^c, \quad [k^c,p^c] \subset p^c, \quad [p^c,p^c] \subset k^c.$$

Let h be a Cartan subalgebra of g and h^c the complexification of h. Then h^c is a Cartan subalgebra of g^c and, for the cases considered here (Hermitian symmetric spaces of noncompact type), holds

$$[h^c,k^c] \subset k^c, \quad [h^c,p^c] \subset p^c.$$

For given g^c, h^c, let Δ be the root system of g^c and Δ^+ the system of positive roots. We say that α is compact if $E_\alpha \in k^c$ and noncompact if $E_\alpha \in p^c$. The set of compact and noncompact roots of g^c with respect to h^c are denoted by Δ_c and Δ_n, respectively. The set of compact simple roots is denoted by Σ_c, β is the only noncompact simple root, and γ_r is the highest root (which is a noncompact positive root).

Let $k_1 = [k,k]$ and assume that k has a nonempty center η of dimension one. Then $k = k_1 \oplus \eta$ and $h = (h \cap k_1) \oplus \eta$. On the other hand, $h^c = (h \cap k_1)^c \oplus \eta^c$ is an orthogonal direct sum with respect to the Killing form: for if $H_\mu \in (h \cap k_1)^c$ and $H_0 \in \eta^c$

$$(H_\mu,H_0) = ([E_\mu,E_{-\mu}],H_0) = (E_\mu,[E_{-\mu},H_0]) = 0.$$

For $\gamma_1,\gamma_2 \in \Delta$ we use the notation

$$\langle \gamma_1,\gamma_2 \rangle = 2(\gamma_1,\gamma_2)/(\gamma_2,\gamma_2) = \gamma_1(H_{\gamma_2}),$$

where $(.,.)$ is the bilinear form on $(h^c)^*$ induced by the Killing form on g^c.

Let $u(g^c)$ be the universal enveloping algebra of g^c, $\Lambda \in (h^c)^*$, and $R = \frac{1}{2}\Sigma_{\alpha \in \Delta^+} \alpha$. The Verma module M_Λ of the

[a] This work contains a part of the Doctoral Thesis written by one of us (J.G.E.).

0022-2488/90/040781-10$03.00

highest weight Λ is defined to be $M_\Lambda = u(g^c)/I_\Lambda$, where I_Λ is the left ideal generated by the elements $(H - \Lambda(H))$, $H\in h^c$, and the set of generators X_γ with $\gamma\in\Delta^+$. To fix a basis on $(h^c)^*$ we choose the set of compact simple roots Σ_c for the space $((h\cap k_1)^c)^*$ and one element $\epsilon\in(\eta^c)^*$ for which

$$\langle\epsilon,\mu\rangle = 0, \quad \forall\mu\in\Sigma_c \text{ and } \langle\epsilon,\gamma_r\rangle = 1,$$

then each $\Lambda\in(h^c)^*$ may be written as $\Lambda = \Lambda_0 + \lambda\epsilon$, where Λ_0 satisfies $\langle\Lambda,\mu_i\rangle = \langle\Lambda_0,\mu_i\rangle\forall\mu_i\in\Sigma_c$. If we choose a normalization for Λ_0 of the type $\langle\Lambda_0,\gamma_r\rangle = 0$, from the last decomposition of Λ we conclude that $\langle\Lambda,\gamma_r\rangle = \lambda$. The relations $\langle\Lambda,\mu_i\rangle = \langle\Lambda_0,\mu_i\rangle$ and $\langle\Lambda_0,\gamma_r\rangle = 0$ fix Λ_0 uniquely. In the following we consider Λ_0 to be k_1 dominant and integral, that is, $\langle\Lambda_0,\mu_i\rangle = n_i$, where n_i are non-negative integers.

Now, if M_Λ is a Verma module, $\widetilde{L}_\Lambda$ an invariant submodule, and $L_\Lambda = M_\Lambda/\widetilde{L}_\Lambda$ a quotient module and if $\rho_\Lambda = M_\Lambda$, $\widetilde{L}_\Lambda$, L_Λ is irreducible then we say that ρ_Λ is infinitesimally unitary if there exists a scalar product (,) on the carrier space V of ρ_Λ such that

$$(u,\rho_\Lambda(X)w) = -(\rho_\Lambda(X)u,w),$$

for all $X\in g$ and $u,w\in V$. The above condition is called g invariance.

In a Verma module this scalar product is induced by a sesquilinear form. For definition and construction of this form see Ref. 2.

In the following we are going to reformulate the Jakobsen method to calculate the modules M_Λ that are unitarizable by using a diagramatic representation of Δ_n^+.

III. JAKOBSEN METHOD

The modules M_Λ are determinated by Λ_0 and λ where Λ_0 is k_1 dominant and integral and $\lambda\in\mathscr{R}$.

There exists a way to represent the set Δ_n^+ by means of bidimensional diagrams in the following way: one begins with β and draws an arrow originating at β for each compact simple root μ_i such that $\beta + \mu_i\in\Delta_n^+$.

Lemma 4.1 of Ref. 3 shows that $i\leqslant 2$. We suppose for simplicity that $i = 2$. Then one draws two arrows: one originating at $\beta + \mu_1$ and parallel to μ_2 and another originating at $\beta + \mu_2$ and parallel to μ_1, both arrows point toward $\beta + \mu_1 + \mu_2$ which is also a root. The next step would be to add compact simple roots to the noncompact roots previously obtained by keeping those that are noncompact roots. Continuing along these lines the diagram may be completed.

For the description of the possible places for unitarity, Jakobsen uses the Bernshtein, Gel'fand, and Gel'fand theorem. This theorem describes the circumstances under which the irreducible quotient L_ξ of a highest weight module can occur in the Jordan–Hölder series $\mathrm{JH}(M_\Lambda)$ of another.

Definition: Let ξ, $\Lambda\in(h^c)^*$. A sequence of roots $\alpha_1,\ldots,\alpha_k\in\Delta^+$ is said to satisfy condition (A) for the pair $(\xi + R, \Lambda + R)$ if

(a) $\xi + R = \sigma_{\alpha_k}\cdots\sigma_{\alpha_1}(\Lambda + R)$, where σ_{α_i} is the Weyl reflexion with respect to α_i,

(b) Take $\xi_0\equiv\Lambda$, $\xi_i + R = \sigma_{\alpha_i}\cdots\sigma_{\alpha_1}(\Lambda + R)$,
Then $\xi_{i-1} - \xi_i = n_i\alpha_i$, $n_i\in\mathscr{N}$.

Theorem: (Bernshtein, Gel'fand, and Gel'fand); Let $\xi,\Lambda\in(h^c)^*$ and let L_ξ, M_Λ be two Verma modules. Then $L_\xi\in\mathrm{JH}(M_\Lambda)$ if and only if there exists a sequence $\alpha_1,\ldots,\alpha_k\in\Delta^+$ satisfying condition (A) for the pair $(\xi + R,\Lambda + R)$.

On the other hand, under some conditions the α_i's may be considered as noncompact ones.

Proposition: Let ξ, $\Lambda\in(h^c)^*$ and assume that the sequence $\alpha_1,\ldots,\alpha_k$ satisfies condition (A) for the pair $(\xi + R,\Lambda + R)$. If ξ is k_1 dominant we may assume that $\alpha_i\in\Delta_n^+, i = 1,\ldots,k$.

Let V_{Λ_0} be an irreducible finite-dimensional $u(k_1^c)$ module with highest weight Λ_0. We first consider the $u(k_1^c)$ module $p^-\otimes V_{\Lambda_0}$. The highest weights on $p^-\otimes V_{\Lambda_0}$ are of the form $\Lambda_0 - \alpha$ for certain $\alpha\in\Delta_n^+$ that we will describe in terms of the Jakobsen diagrams.

We now describe the method:

(i) Let $\alpha\in\Delta_n^+$ and assume $\alpha - \mu_j\in\Delta_n^+$ for $\mu_j\in\Sigma_c$, $j = 1,\ldots,i$ and $i\leqslant 2$.
Then $\Lambda_0 - \alpha$ is a highest weight for the $u(k_1^c)$ module $p^-\otimes V_{\Lambda_0}$ if and only if for all $j = 1,\ldots,i$,

$$\Lambda_0(H_{\mu_j})\equiv\langle\Lambda_0,\mu_j\rangle\geqslant\max\{1,\langle\alpha,\mu_j\rangle\}.$$

Recall that Λ_0 is fixed by given integers $\langle\Lambda_0,\mu_i\rangle$, $\mu_i\in\Sigma_c$ and $\langle\Lambda_0,\gamma_r\rangle = 0$.

(ii) For those $\alpha\in\Delta_n^+$ of step (i) let $\lambda_\alpha\in\mathscr{R}$ be determined by the equation

$$\langle\Lambda + R,\alpha\rangle = (\Lambda_0 + \lambda_\alpha\epsilon + R)(H_\alpha) = 1.$$

Let λ_0 denote the smallest among those λ_α's, and let α_0 denote the corresponding element of Δ_n^+. We now define the following sets:

$$C_{\alpha_0}^+ = \{\alpha\in\Delta_n^+/\alpha\geqslant\alpha_0\} \text{ and } C_{\alpha_0}^- = \{\alpha\in\Delta_n^+/\alpha\leqslant\alpha_0\}.$$

The way in which those sets appear in the diagram of Δ_n^+ suggest that we can call $C_{\alpha_0}^+$ and $C_{\alpha_0}^-$ the forward and backward cone, respectively, at α_0.

(iii) Let $\omega_q = n_1\alpha_1 + \cdots + n_r\alpha_r, n_i\in\mathscr{N}$. If $\alpha_1,\ldots,\alpha_r\in\Delta_n^+$ satisfies condition (A) for the pair $(\Lambda - \omega_q + R,\Lambda + R)$, where $\Lambda = \Lambda_0 + \lambda_q\epsilon$ and $\Lambda_0 - \omega_q$ is the weight of a highest weight vector q in the $u(k_1^c)$ module $u(p^-)\otimes V_{\Lambda_0}$ and $\lambda_q < \lambda_0$, then

$$\alpha_i\in C_{\alpha_0}^+, \quad \forall i = 1\cdots r.$$

(iv) The α_i's appearing in ω_q must satisfy certain conditions that we now describe.

Because inner products between positive noncompact roots are non-negative and because $\lambda_q < \lambda_0$, it follows that for $\alpha_i\in\Delta_n^+$

$$\langle\Lambda_0 + \lambda_0\epsilon + R,\alpha_i\rangle > \langle\Lambda_0 + \lambda_q\epsilon + R,\alpha_i\rangle > 0.$$

On the other hand, to check the k_1 dominance of $\Lambda_0 - \omega_q$ i.e.,

$$\langle\Lambda_0 - \omega_q,\mu_j\rangle\geqslant 0, \quad \forall\mu_j\in\Sigma_c,$$

it is useful to have in mind that if a compact simple root μ is pointing toward a noncompact positive root α in the diagram then $\langle\alpha,\mu\rangle > 0$ and if μ arises outwards α and at the same time μ is not pointing toward α then $\langle\alpha,\mu\rangle < 0$.

(v) M_Λ with $\Lambda = \Lambda_0 + \lambda_0\epsilon$ is unitarizable. The value $\lambda = \lambda_0$ is called the last possible place for unitarity because for $\lambda > \lambda_0$ there is no unitarity. The description of the general situation follows by forming tensor products of M_Λ with the unitary module $M_{\lambda,\epsilon}$ corresponding to $\Lambda_0 = 0$. The restriction of $M_\Lambda \otimes M_{\lambda,\epsilon}$ to the diagonal is the unitarizable module $M_{\Lambda'}$ with $\Lambda' = \Lambda_0 + (\lambda_0 + \lambda_s)\epsilon$.

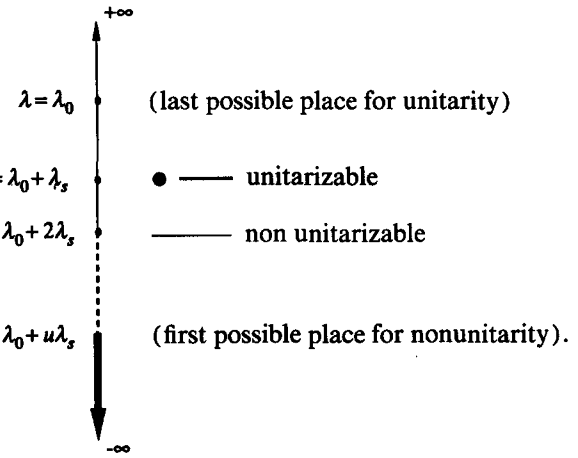
This means that if we want to unitarize we must take the quotient space with respect to the invariant subspace generated by the highest weight vector corresponding to $\lambda_0 + \lambda_s$, which is a second-order polynomial: this polynomial will be missing.

The modules $M_{\Lambda''}$ with $\Lambda'' = \Lambda_0 + \lambda\epsilon, \lambda_0 + \lambda_s < \lambda < \lambda_0$ are not unitarizable.

For $\lambda_0 + 2\lambda_s$ we may have unitarity and there will be a third-order missing polynomial, while there is no unitarity for $\lambda_0 + 2\lambda_s < \lambda < \lambda_0 + \lambda_s$.

Continuing along these lines we arrive at the first possible place for nonunitarity that corresponds to $\lambda = \lambda_0 + u\lambda_s$ (we call $u+1$ the reduction level) and all representations with $\lambda < \lambda_0 + u\lambda_s$ are unitary.

The following diagram illustrates the possible places for unitarity (in the next section we will see that $\lambda_s < 0$).



+∞
λ=λ0
(last possible place for unitarity)
λ=λ0+λs
● —— unitarizable
λ=λ0+2λs
——— non unitarizable
λ=λ0+uλs
(first possible place for nonunitarity).
-∞

In order to construct the highest weight vector corresponding to the highest weight known, say $\Lambda + R - m\beta - \Sigma_i m_i\mu_i$, we start by finding a set of simple roots $\alpha_1, \alpha_2, \ldots, \alpha_s$ such that

$$\Lambda + R - m\beta - \sum_i m_i\mu_i = \sigma_{\alpha_1}\sigma_{\alpha_2}\cdots\sigma_{\alpha_s}(\Lambda + R),$$

then we make use of the method outlined in Ref. 1. As can be seen in the examples the exponents are not always positives, and even in some cases (see $\mathrm{sp}(n,\mathscr{R})$) they are not integers.

The expressions obtained are only formally valid and use has to be made of Taylor series for this powers of the generators and then to apply the commutation relations $[E_{\alpha_i}, E^m_{\alpha_j}]$ $(m\in\mathscr{N})$ to each term in the series. For instance, in su(2,2) from

$$[E_\beta, E^m_{\mu_1}] = mE^{m-1}_{\mu_1}E_{\alpha_1}$$

[where $\beta = (0,1,-1,0)$, $\mu_1 = (0,0,1,-1)$ and $\alpha_1 = (0,1,0,-1)$] it follows that $[E_\beta, f(E_{\mu_1})] = f'(E_{\mu_1})E_{\alpha_1}$ for any analytic function of the operator E_{μ_1}.

IV. SOME DEFINITIONS AND NOTATIONS

From the Jakobsen diagrams (see the examples in Sec. V) we observe that all positive noncompact roots can be expressed as

$$\alpha = \beta + \mu_{i_1} + \cdots + \mu_{i_k}, \quad \mu_{i_m} \in \Sigma_c, \quad m = 1,\ldots,k.$$

Thus, given α in this way we define its "height" as $k+1$ (the roots μ_{i_m} may be repeated).

On the other hand, given the decomposition $\Lambda = \Lambda_0 + \lambda\epsilon$ we may relate the products $\langle\Lambda,\alpha\rangle$ and $\langle\Lambda_0,\alpha\rangle$, $\alpha\in\Delta_n^+$, in the following way:

$$\langle\Lambda,\alpha\rangle = \langle\Lambda_0,\alpha\rangle + \lambda\,[(\gamma_r,\gamma_r)/(\alpha,\alpha)].$$

In fact

$$\langle\Lambda,\alpha\rangle = \langle\Lambda_0,\alpha\rangle + \lambda\,\langle\epsilon,\alpha\rangle,$$

and, decomposing

$$\alpha = \gamma_r - \sum_{\mu_i\in\Sigma_c} \mu_i$$

(see Jakobsen diagrams) then

$$\langle\epsilon,\alpha\rangle = 2(\epsilon,\gamma_r)/(\alpha,\alpha) = (\gamma_r,\gamma_r)/(\alpha,\alpha),$$

where we use the fact that $\langle\epsilon,\gamma_r\rangle = 1$ and $\langle\epsilon,\mu\rangle = 0\ \forall\mu\in\Sigma_c$.

Be means of a direct calculation we obtain the following useful expressions for the products $\langle R,\alpha\rangle$ and $\langle\Lambda,\alpha\rangle$ that will be needed in the next section:

(a) $\mathrm{su}(p,q)$, $\mathrm{so}^*(2n)$, e_6 and e_7
$\langle R,\alpha\rangle =$ height of α,
$\langle\Lambda,\alpha\rangle = (\Lambda_0,\alpha) + \lambda$,

(b) $\mathrm{sp}(n,\mathscr{R})$
if α is short $\quad\langle R,\alpha\rangle =$ height $+ 1$,
$\langle\Lambda,\alpha\rangle = (\Lambda_0,\alpha) + 2\lambda$,
if α is long $\quad\langle R,\alpha\rangle = \frac{1}{2}\{$height $+ 1\}$,
$\langle\Lambda,\alpha\rangle = \frac{1}{2}(\Lambda_0,\alpha) + \lambda$.

From the Jakobsen diagrams we see that all roots of the same height are in an horizontal line.

In Fig. 1 such roots are inside a little circle and we may localize them by means of a subindex j that is equal to the height of the root and an ordenation superindex i that is equal to one, for the root placed at the right branch of the cone generated by α_0, and it increases from unit to unit when we are going toward the left branch. In this way we will write α^i_j.

In order to calculate the parameter λ_s in step (v) we make use of the following definition.

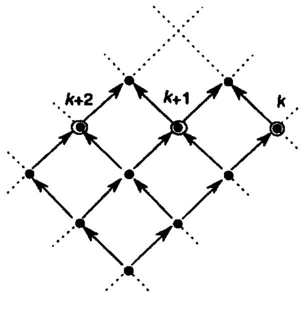


FIG. 1. The ordenation superindex in a Jakobsen diagram.

Definition (Harish-Chandra)*:* Let γ_1 be the smallest element of Δ_n^+ and, inductively, let γ_k be the smallest element of Δ_n^+ that is orthogonal to $\gamma_1,...,\gamma_{k-1}$. Let $\gamma_1,...,\gamma_t$ be the maximal collection obtained. Then t is the split rank of g (Ref.4).

With our notation $\gamma_1 \equiv \beta$. We use the Jakobsen diagrams to obtain the split rank.

su(p,q)

The collection $\gamma_1,...,\gamma_t$ follows by drawing a line from β as is indicated in Fig. 2. The roots founded are those which are on the line:

If $p \leqslant q$: $e_p - e_{p+1},\quad e_{p-1} - e_{p+2},...,\quad e_1 - e_{2p}$,

If $p \geqslant q$: $e_p - e_{p+1},\quad e_{p-1} - e_{p+2},...,\quad e_{p-(q-1)} - e_{p+q}$.

So, if $p \leqslant q$ the split rank is p and if $p \geqslant q$ the split rank is q, therefore

$$\text{Split rank su}(p,q) = \min\{p,q\}.$$

sp(n,$\mathscr{R}$)

In the same way as in su(p,q) the collection obtained here is

$$2e_n, 2e_{n-1},...,2e_1.$$

Thus

$$\text{Split rank sp}(n,\mathscr{R}) = n.$$

so*(2n)

The collection is, in this case

$$e_{n-1} + e_n,...,e_1 + e_2, \quad \text{if } n \text{ is even},$$

$$e_{n-1} + e_n,...,e_1 + e_3, \quad \text{if } n \text{ is odd}.$$

Then the split rank is $n/2$ if n is even and $(n-1)/2$ if n is odd:

$$\text{Split rank so*}(2n) = [n/2],$$

where $[x]$ denotes the largest integer $\leqslant x$.

so(2n−1,2), so(2n−2,2)

The split rank is, in both cases, equal to two. The collection is, in this case $e_1 - e_2$, $e_1 + e_2$.

e_6,e_7

The collection obtained is now

$$e_6{:}\{\tfrac{1}{2}(e_1 - e_2 - e_3 - e_4 - e_5 - e_6 - e_7 + e_8),$$

$$\tfrac{1}{2}(-e_1 + e_2 + e_3 + e_4 - e_5 - e_6 - e_7 + e_8)\},$$

then the split rank of e_6 is equal to two

$$e_7{:}\{e_6 - e_5, e_6 + e_5, e_8 - e_7\},$$

thus the split rank of e_7 is equal to three.

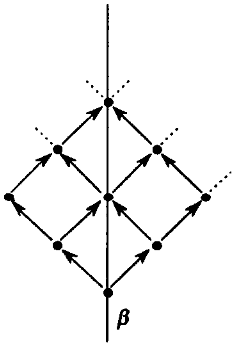


FIG. 2. Diagram for the calculation of the split rank.

Now let $h^- = \Sigma_{i=1}^t \not{c}H_\gamma$ and, for $1 \leqslant j \leqslant t$,$t$ being the split rank, let c_j be the number of compact positive roots μ such that $\mu|_{h^-} = \frac{1}{2}(\gamma_j - \gamma_i)$, $i<j$. Then, if we consider the most singular nontrivial unitary module corresponding to $\Lambda_0 = 0$, according to Theorem 5.10 in Ref. 4, $\lambda_q = -\frac{1}{2}c_j$. A straightforward calculation case by case shows that

$$\lambda_q = (j-1)\lambda_s, \quad 1 \leqslant j \leqslant t,$$

with λ_s given in the following table:

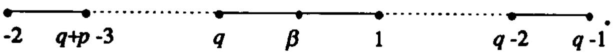
	su(p,q)	sp$(n,\Re)$	so*$(2n)$	e_6	e_7
λ_s	-1	$-\frac{1}{2}$	-2	-3	-4

V. POSSIBLE PLACES FOR UNITARITY

We consider here those cases for which the reduction level is strictly higher than one.

su(p,q)

Dynkin diagram



Let M_Λ be a representation for su(p,q) with $\Lambda = (\Lambda_1,\Lambda_2,...,\Lambda_{p+q})$ with respect to the standard orthonormal basis of $\mathscr{R}^n (n = p+q)$, satisfying the following conditions on its components:

$$\Lambda_1 = \Lambda_2 = \cdots = \Lambda_i \neq \Lambda_{i+1},$$

$$\Lambda_n = \Lambda_{n-1} = \cdots = \Lambda_{n-j+1} \neq \Lambda_{n-j}.$$

If we put $\Lambda = \Lambda_0 + \epsilon\lambda$ these conditions are equivalent to the following ones:

$$\langle \Lambda_0,\mu_{q-1} \rangle = \langle \Lambda_0,\mu_{q-2} \rangle = \cdots = \langle \Lambda_0,\mu_{t+1} \rangle = 0,$$

$$\langle \Lambda_0,\mu_t \rangle \neq 0,$$

$$\langle \Lambda_0,\mu_{n-2} \rangle = \langle \Lambda_0,\mu_{n-3} \rangle = \cdots = \langle \Lambda_0,\mu_{s+1} \rangle = 0,$$

$$\langle \Lambda_0,\mu_s \rangle \neq 0,$$

with $t = q - j$ and $s = n - i - 1$.

Applying steps (i) and (ii) we obtain

$$\alpha_0 = \beta + \mu_1 + \cdots + \mu_t + \mu_q + \mu_{q+1} + \cdots + \mu_s,$$

with height $t + s - q + 2$. Then $\lambda_0 = q - t - s - 1$.

Then a first-order polynomial will be missing with highest weight

$$\Lambda_0 + (q - t - s - 1)\epsilon - \alpha_0,$$

where, in this case, $\epsilon = (q/n,q/n,...,q/n, -p/n,..., -p/n)$ with p copies of q/n and q of $-p/n$.

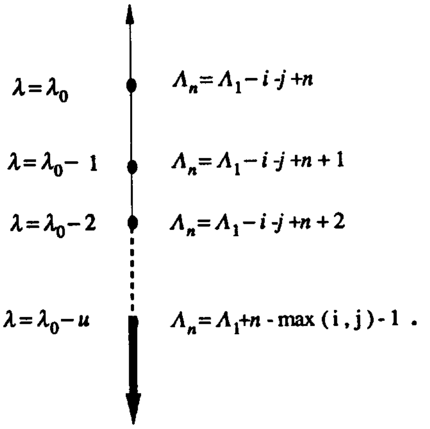
For $\lambda_q = \lambda_0 + \lambda_s = \lambda_0 - 1$ we obtain from steps (iii) and (iv) a second-order polynomial that will be missing with heighest weight

$$\Lambda_0 + (\lambda_0 - 2)\epsilon - \alpha^1_{2-\lambda_0} - \alpha^2_{2-\lambda_0}.$$

The next case is $\lambda_q = \lambda_0 - 2$ where a third-order polynomial with highest weight $\Lambda_0 + (\lambda_0 - 2)\epsilon - \alpha^1_{3-\alpha_0} - \alpha^2_{3-\lambda_0} - \alpha^3_{3-\lambda_0}$ will be missing. Continuing in the same way we arrive at $\lambda_q = \lambda_0 - u$, $u = \min(q - t - 1, n - s - 2)$ where a polynomial of order

$u+1=\min(i,j)$ will be missing. For $\lambda<\lambda_0-u$ it is impossible to find a polynomial of order strictly higher than $\min(i,j)$ because the k_1 dominance is violated (see step iv).

Thus the reduction level is $\min(i,j)$. On the other hand, $\lambda=\langle\Lambda,\gamma_r\rangle$ or, equivalently, $\Lambda_n=\Lambda_1-\lambda$. Then, taking into account the possible values of λ we obtain the following diagram:


$\lambda=\lambda_0$
$\Lambda_n=\Lambda_1-i\text{-}j+n$
$\lambda=\lambda_0-1$
$\Lambda_n=\Lambda_1-i\text{-}j+n+1$
$\lambda=\lambda_0-2$
$\Lambda_n=\Lambda_1-i\text{-}j+n+2$
$\lambda=\lambda_0-u$
$\Lambda_n=\Lambda_1+n-\max(i,j)-1$.

Example: su(5,8)

Assume that

$\langle\Lambda_0,\mu_7\rangle=\langle\Lambda_0,\mu_6\rangle=0,$

$\langle\Lambda_0,\mu_5\rangle\neq0,$

$\langle\Lambda_0,\mu_{11}\rangle=\langle\Lambda_0,\mu_{10}\rangle=\langle\Lambda_0,\mu_9\rangle=0,$

$\langle\Lambda_0,\mu_8\rangle\neq0,$

$\langle\Lambda_0,\mu_i\rangle=n_i,\quad 1\leqslant i\leqslant5$ or $i=8.$

With those conditions we obtain $\alpha_0=\beta+\mu_1+\mu_2+\mu_3+\mu_4+\mu_5+\mu_8$ the height of which is 7, then

$$1=\langle\Lambda,\alpha_0\rangle+\langle R,\alpha_0\rangle=\langle\Lambda,\alpha_0\rangle+7=\lambda_0+7\quad\text{or}\quad\lambda_0=-6.$$

For $\lambda_q=\lambda_0+\lambda_s=-7$ and having in mind that now $\Lambda'=\Lambda_0-7\epsilon$:

$\langle\Lambda'+R,\alpha_8^1\rangle=-7+\langle R,\alpha_8^1\rangle=1,$

$\langle\Lambda'+R,\alpha_8^2\rangle=1,$

then a second-order polynomial will be missing with highest weight $\Lambda'-\alpha_8^1-\alpha_8^2$.

For $\lambda_q=-8$, $\Lambda''=\Lambda_0-8\epsilon$ and we have

$$\langle\Lambda''+R,\alpha_9^1\rangle=\langle\Lambda''+R,\alpha_9^2\rangle=\langle\Lambda''+R,\alpha_9^3\rangle=1,$$

then a third-order polynomial will be missing with highest weight $\Lambda''-\alpha_9^1-\alpha_9^2-\alpha_9^3$.

For $\lambda_q<-8$ only a set of roots belonging to $C^+_{\alpha_0}$ and without a coefficient equal to one in Fig. 3 could satisfy condition (A). But, for example, α_{10}^1 could not belong to this set because of μ_9 which point towards it and because there is no root between those from which arises μ_9, the result would not be k_1 dominant. The same thing occurs with $\alpha_{10}^2,\alpha_{11}^1$ because of μ_{10} and with $\alpha_{10}^3,\alpha_{11}^2,\gamma_r$ because of μ_{11}. There is unitarity for $\lambda_q<-8$.

The highest weight vectors which we must eliminate (missing polynomials) in order to obtain unitarity are, formally, the following:

(i) Height 7:

$$E^{-n_5}_{-\mu_5}E^{-n_4-n_5-1}_{-\mu_4}E^{-n_3-n_4-n_5-2}_{-\mu_3}E^{-n_2-n_3-n_4-n_5-3}_{-\mu_2}$$
$$\times E^{-n_1-n_2-n_3-n_4-n_5-4}_{-\mu_1}E^{-n_8}_{-\mu_8}E_{-\beta}E^{n_8+1}_{-\mu_8}$$
$$\times E^{n_1+n_2+n_3+n_4+n_5+5}_{-\mu_1}E^{n_2+n_3+n_4+n_5+4}_{-\mu_2}E^{n_3+n_4+n_5+3}_{-\mu_3}$$
$$\times E^{n_4+n_5+2}_{-\mu_4}E^{n_5+1}_{-\mu_5},$$

with $\Lambda=\Lambda_0-6\epsilon-\beta-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5-\mu_8$.

(ii) Height 8:

$$E^{-n_5}_{-\mu_5}E^{-n_4-n_5-1}_{-\mu_4}E^{-n_3-n_4-n_5-2}_{-\mu_3}E^{-n_2-n_3-n_4-n_5-3}_{-\mu_2}$$
$$\times E^{-n_1-n_2-n_3-n_4-n_5-4}_{-\mu_1}E^{-n_8}_{-\mu_8}E^2_{-\beta}E^{n_8+2}_{-\mu_8}E_{-\mu_9}$$
$$\times E^{n_1+n_2+n_3+n_4+n_5+6}_{-\mu_1}E^{n_2+n_3+n_4+n_5+5}_{-\mu_2}E^{n_3+n_4+n_5+4}_{-\mu_3}$$
$$\times E^{n_4+n_5+3}_{-\mu_4}E^{n_5+2}_{-\mu_5}E_{-\mu_6},$$

with

$$\Lambda=\Lambda_0-7\epsilon-2\beta-2\mu_1-2\mu_2-2\mu_3-2\mu_4-2\mu_5$$
$$-\mu_6-2\mu_8-\mu_9.$$

(iii) Height 9:

$$E^{-n_5}_{-\mu_5}E^{-n_5-n_4-1}_{-\mu_4}E^{-n_3-n_4-n_5-2}_{-\mu_3}E^{-n_2-n_3-n_4-n_5-3}_{-\mu_2}$$
$$\times E^{-n_1-n_2-n_3-n_4-n_5-4}_{-\mu_1}E^{-n_8}_{-\mu_8}E^3_{-\beta}E^{n_8+3}_{-\mu_8}E^2_{-\mu_9}$$
$$\times E_{-\mu_{10}}E^{n_1+n_2+n_3+n_4+n_5+7}_{-\mu_1}E^{n_2+n_3+n_4+n_5+6}_{-\mu_2}$$
$$\times E^{n_3+n_4+n_5+5}_{-\mu_3}E^{n_4+n_5+4}_{-\mu_4}E^{n_5+3}_{-\mu_5}E^2_{-\mu_6}E_{-\mu_7}$$

with

$$\Lambda=\Lambda_0-8\epsilon-3\beta-3\mu_1-3\mu_2-3\mu_3-3\mu_4-3\mu_5$$
$$-2\mu_6-\mu_7-3\mu_8-2\mu_9-\mu_{10}.$$

In the three cases, in order to cancel the negative powers of the generators, appropriate commutation relations are to be applied, as stated before.

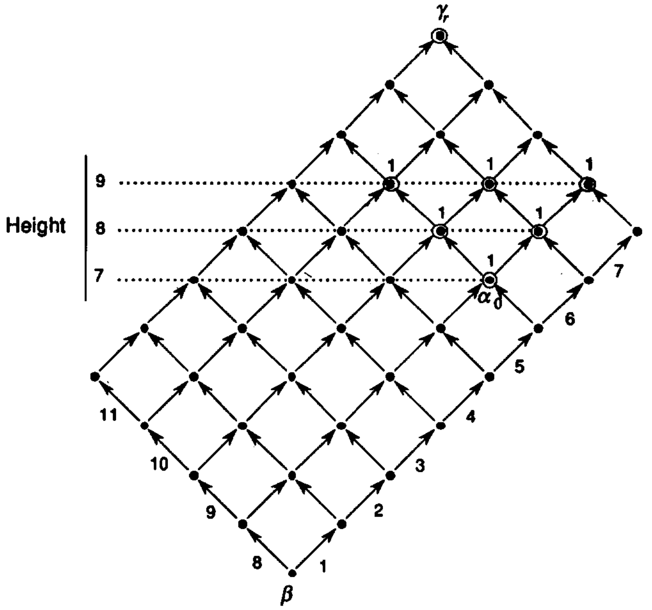
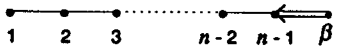


FIG. 3. Missing polynomials in the example for su(5,8).

sp(n,$\mathscr{R}$)

Dynkin diagram  .

Let M_Λ be a representation for sp$(n,\mathscr{R})$ with $\Lambda=(\Lambda_1,\ldots,\Lambda_n)$ we put $\Lambda=\Lambda_0+\lambda\epsilon$ where $\epsilon=(1,1,\ldots,1)$. We consider two cases.

Case I.

The weight Λ satisfy the following conditions on its components:

$$\Lambda_1=\Lambda_2=\cdots=\Lambda_i\geqslant\Lambda_{i+1}+2,$$

or, equivalently

$$\langle\Lambda_0,\mu_1\rangle=\langle\Lambda_0,\mu_2\rangle=\cdots=\langle\Lambda_0,\mu_{i-1}\rangle=0,$$
$$\langle\Lambda_0,\mu_i\rangle=n\geqslant 2.$$

Applying the Jakobsen method we obtain

$$\alpha_0=\beta+2\mu_{n-1}+2\mu_{n-2}+\cdots+2\mu_{n-(n-i)},$$

with height $2(n-i)+1$. As α_0 is a long root, the condition $\langle\Lambda+R,\alpha_0\rangle=1$ implies

$$\tfrac{1}{2}(\Lambda_0,\alpha_0)+\lambda_0+n-i+1=1,\quad \lambda_0=i-n,$$

then a first-order polynomial with highest weight $\Lambda_0+(i-n)\epsilon-\alpha_0$ will be missing when we unitarize.

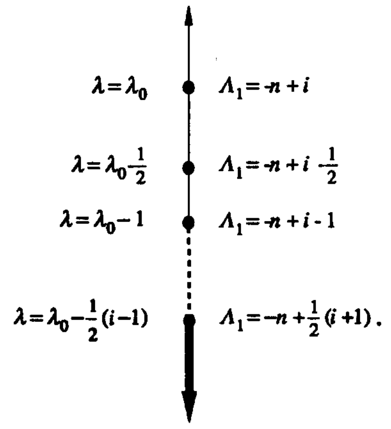
For $\lambda=\lambda_0+\lambda_s=i-n-\frac{1}{2}$ we obtain a second-order polynomial which will be missing with highest weight

$$\Lambda_0+(i-n-\tfrac{1}{2})\epsilon-2\alpha^1_{2(n-i)+2}.$$

For $\lambda=\lambda_0+2\lambda_s$ the third-order missing polynomial has highest weight

$$\Lambda_0+(i-n-1)\epsilon-2\alpha^1_{2(n-i)+3}-\alpha^2_{2(n-i)+3}.$$

Following along these lines we arrive at $\lambda=\lambda_0+(i-1)\lambda_s=\frac{1}{2}(i+1)-n$, where a ith-order polynomial will be missing. For $\lambda<\frac{1}{2}(i+1)-n$ it is impossible to obtain polynomials of order strictly higher than i because there would not be k_1 dominance, therefore the reduction level is i. On the other hand, from the condition $\lambda=\langle\Lambda,\gamma_r\rangle$ it follows that $\Lambda_1=\lambda$. Thus, for the different values of λ we obtain the following diagram which give us the possible values of Λ_1 for unitarity:



Case II.

We consider in this case the following conditions:

$$\Lambda_1=\Lambda_2=\cdots=\Lambda_i,\quad \Lambda_i-\Lambda_{i+1}=1,$$
$$\Lambda_{i+1}=\Lambda_{i+2}=\cdots=\Lambda_{i+j}\neq\Lambda_{i+j+1},$$

which are equivalent to the following ones:

$$\langle\Lambda_0,\mu_1\rangle=\langle\Lambda_0,\mu_2\rangle=\cdots=\langle\Lambda_0,\mu_{i-1}\rangle=0,$$
$$\langle\Lambda_0,\mu_i\rangle=1,$$
$$\langle\Lambda_0,\mu_{i+1}\rangle=\langle\Lambda_0,\mu_{i+2}\rangle=\cdots=\langle\Lambda_0,\mu_{i+j-1}\rangle=0,$$
$$\langle\Lambda_0,\mu_{i+j}\rangle=n\geqslant 1.$$

In this case

$$\alpha_0=\beta+2(\mu_{n-1}+\mu_{n-2}+\cdots+\mu_{n-(n-i-j)})$$
$$+\mu_{n-(n-i-j)}+\cdots+\mu_{n-(n-i)},$$

the height of which is $2(n-i)-j+1$. As α_0 is a short root, the condition $\langle\Lambda+R,\alpha_0\rangle=1$ implies

$$(\Lambda_0,\alpha_0)+2\lambda_0+2(n-i+1)-j=1,$$

and, having in mind that we can also state

$$\alpha_0=\gamma_r-2(\mu_1+\cdots+\mu_{i-1})-\mu_i-\mu_{i+1}$$
$$-\cdots-\mu_{i+j-1},$$

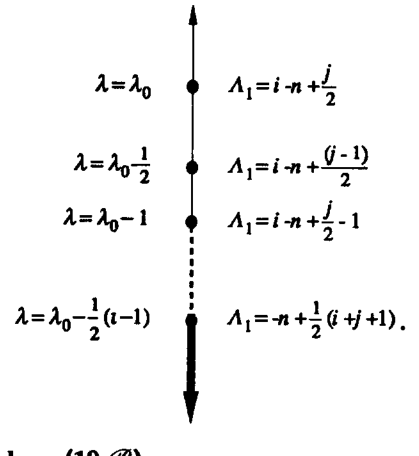
we have

$$\lambda_0=i-n+j/2.$$

For $\lambda=\lambda_0+\lambda_s=i-n+(j-1)/2$ we obtain a second-order polynomial that will be missing with highest weight

$$\Lambda_0+(i-n+(j-1)/2)\epsilon-\alpha^1_{2(n-i+1)-j}$$
$$-\alpha^2_{2(n-i+1)-j}.$$

Continuing as in case I we arrive at $\lambda=\lambda_0+(i-1)\lambda_s=-n+\frac{1}{2}(i+j+1)$ where an ith-order polynomial will be missing. As for $\lambda<-n+\frac{1}{2}(i+j+1)$ there is no k_1 dominance the reduction level is i. The diagram in this case is the following:



Example: sp(10,$\mathscr{R}$)

Case I.

Let the conditions on Λ_0 be

$$\langle\Lambda_0,\mu_1\rangle=\langle\Lambda_0,\mu_2\rangle=\langle\Lambda_0,\mu_3\rangle=\langle\Lambda_0,\mu_4\rangle=0,$$
$$\langle\Lambda_0,\mu_5\rangle=n_5\geqslant 2,\quad \langle\Lambda_0,\mu_i\rangle=n_i,\quad 6\leqslant i\leqslant 9.$$

In this case

$$\alpha_0=\beta+2\mu_9+2\mu_8+2\mu_7+2\mu_6+2\mu_5$$

with height 11. As α_0 is a long root, the condition $\langle\Lambda+R,\alpha_0\rangle=1$ implies

$$\tfrac{1}{2}(\Lambda_0,\alpha_0)+\lambda_0+6=1,\quad \lambda_0=-5.$$

For $\lambda_q = \lambda_0 + \lambda_s = -\frac{11}{2}$,

$\langle \Lambda + R, \alpha_{12}^1 \rangle = 2\lambda_q + 13 = 2$ because α_{12}^1 is short.

Then there is a second-order polynomial missing for $\lambda_q = -11/2$ with highest weight $\Lambda_0 - (11/2)\epsilon - 2\alpha_{12}^1$.

For $\lambda_q = \lambda_0 + 2\lambda_s = -6$,

$\langle \Lambda + R, \alpha_{13}^1 \rangle = 2\lambda_q + 14 = 2$ because α_{13}^1 is short,

$\langle \Lambda + R, \alpha_{13}^2 \rangle = \lambda_q + 7 = 1$ because α_{13}^2 is long,

and we will have a third-order missing polynomial with highest weight $\Lambda_0 - 6\epsilon - 2\alpha_{13}^1 - \alpha_{13}^2$.

For $\lambda_q = -\frac{13}{2}$,

$\langle \Lambda + R, \alpha_{14}^1 \rangle = \langle \Lambda + R, \alpha_{14}^2 \rangle = 2\lambda_q + 15 = 2$,

a fourth-order polynomial will be missing with highest weight $\Lambda_0 - \frac{13}{2}\epsilon - 2\alpha_{14}^1 - 2\alpha_{14}^2$.

For $\lambda_q = -7$,

$\langle \Lambda + R, \alpha_{15}^1 \rangle = \langle \Lambda + R, \alpha_{15}^2 \rangle = 2\lambda_q + 16 = 2$,

$\langle \Lambda + R, \alpha_{15}^3 \rangle = \lambda_q + 8 = 1$,

and there will be a fifth-order polynomial missing with highest weight

$$\Lambda_0 - 7\epsilon - 2\alpha_{15}^1 - 2\alpha_{15}^2 - \alpha_{15}^3.$$

For $\lambda_q < -7$ the roots which we must consider are those belonging to $C_{\alpha_0}^+$ without a coefficient in Fig. 4.

By an argument along the lines of the example for su(5,8) we see that for those λ_q there is no k_1 dominance. Then the first possible place for nonunitarity is $\lambda_q = -7$.

We state in the following the highest weight vectors for the two first heights.

(i) Height 11:

$$E_{-\mu_5}^{-n_5+1} E_{-\mu_6}^{-n_5-n_6} E_{-\mu_7}^{-n_5-n_6-n_7-1} E_{-\mu_8}^{-n_5-n_6-n_7-n_8-2}$$
$$\times E_{-\mu_9}^{-n_5-n_6-n_7-n_8-n_9-3} E_{-\beta} E_{-\mu_9}^{n_5+n_6+n_7+n_8+n_9+5}$$
$$\times E_{-\mu_8}^{n_5+n_6+n_7+n_8+4} E_{-\mu_7}^{n_5+n_6+n_7+3} E_{-\mu_6}^{n_5+n_6+2} E_{-\mu_5}^{n_5+1},$$

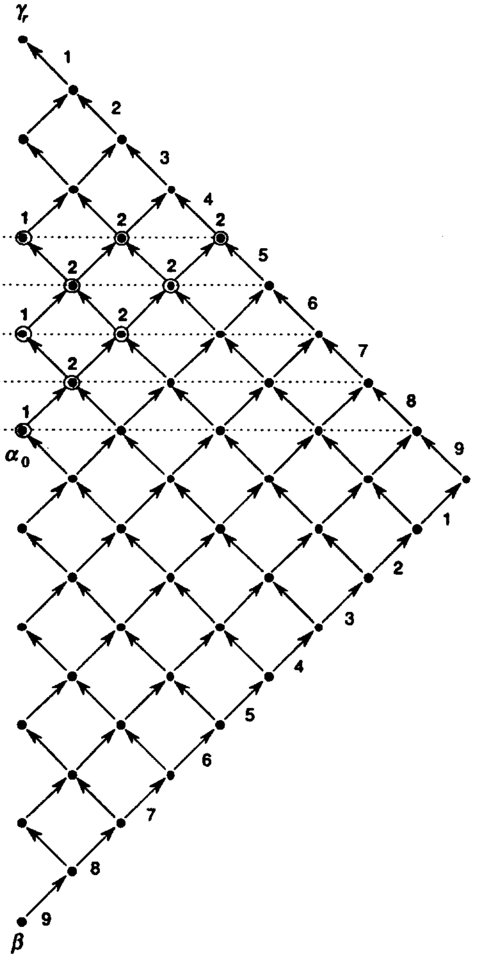


FIG. 4. Mising polynomials in the sp(10,R) example, case I.

with $\Lambda = \Lambda_0 - 5\epsilon - \beta - 2\mu_5 - 2\mu_6 - 2\mu_7 - 2\mu_8 - 2\mu_9$.

(ii) Height 12:

$$E_{-\mu_5}^{-n_5+1} E_{-\mu_6}^{-n_5-n_6} E_{-\mu_7}^{-n_5-n_6-n_7-1} E_{-\mu_8}^{-n_5-n_6-n_7-n_8-2} E_{-\mu_9}^{-n_5-n_6-n_7-n_8-n_9-3} E_{-\beta}^{3/2} E_{-\mu_9}^{n_5+n_6+n_7+n_8+n_9+6} E_{-\mu_8}^{n_5+n_6+n_7+n_8+5}$$
$$\times E_{-\mu_7}^{n_5+n_6+n_7+4} E_{-\mu_6}^{n_5+n_6+3} E_{-\mu_5}^{n_5+2} E_{-\mu_4}^{2} E_{-\mu_5}^{-n_5} E_{-\mu_6}^{-n_5-n_6-1} E_{-\mu_7}^{-n_5-n_6-n_7-2} E_{-\mu_8}^{-n_5-n_6-n_7-n_8-3} E_{-\mu_9}^{-n_5-n_6-n_7-n_8-n_9-4}$$
$$\times E_{-\beta}^{1/2} E_{-\mu_9}^{n_5+n_6+n_7+n_8+n_9+5} E_{-\mu_8}^{n_5+n_6+n_7+n_8+4} E_{-\mu_7}^{n_5+n_6+n_7+3} E_{-\mu_6}^{n_5+n_6+2} E_{-\mu_5}^{n_5+1},$$

with

$$\Lambda = \Lambda_0 - \tfrac{11}{2}\epsilon - 2\beta - 2\mu_4 - 4\mu_5 - 4\mu_6 - 4\mu_7 - 4\mu_8 - 4\mu_9.$$

Case II.

Now let there be the following conditions in Λ_0:

$$\langle \Lambda_0, \mu_1 \rangle = \langle \Lambda_0, \mu_2 \rangle = 0, \quad \langle \Lambda_0, \mu_3 \rangle = \langle \Lambda_0, \mu_4 \rangle = 1,$$
$$\langle \Lambda_0, \mu_5 \rangle = \langle \Lambda_0, \mu_6 \rangle = 0, \quad \langle \Lambda_0, \mu_7 \rangle = n \geqslant 1.$$

In this case

$$\alpha_0 = \beta + 2\mu_9 + 2\mu_8 + 2\mu_7 + 2\mu_6 + 2\mu_5 + 2\mu_4 + \mu_3$$

with height 14 (Fig. 5). As α_0 is a short root:

$$1 = \langle \Lambda + R, \alpha_0 \rangle = (\Lambda_0, \gamma_r - 2\mu_1 - 2\mu_2 - \mu_3) + 2\lambda_0 + 15,$$

$\lambda_0 = -\frac{13}{2}$.

For $\lambda_q = \lambda_0 + \lambda_s = -7$,

$\langle \Lambda + R, \alpha_{15}^1 \rangle = 2\lambda_q + 15 = 1$,

$\langle \Lambda + R, \alpha_{15}^2 \rangle = \lambda_q + 8 = 1$,

then a second-order polynomial with highest weight $\Lambda_0 - 7\epsilon - \alpha_{15}^1 - \alpha_{15}^2$ will be missing.

For $\lambda_q = \lambda_0 + 2\lambda_s = -\frac{15}{2}$,

$\langle \Lambda + R, \alpha_{16}^1 \rangle = 2\lambda_q + 16 = 1$,

$\langle \Lambda + R, \alpha_{16}^2 \rangle = 2\lambda_q + 17 = 2$,

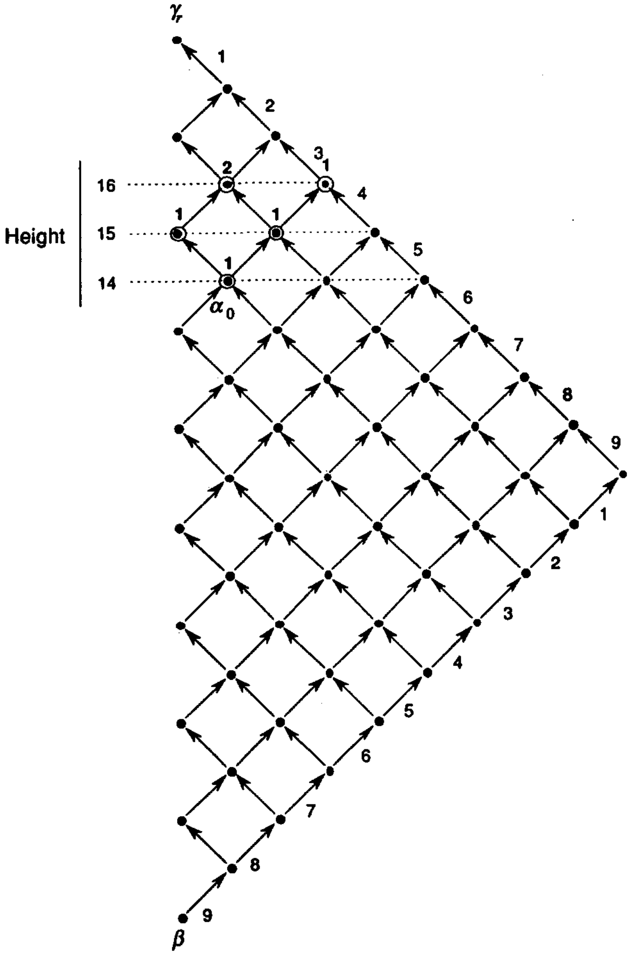
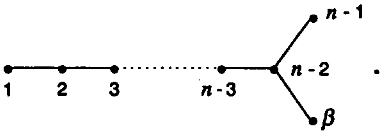


FIG. 5. Missing polynomials in the sp(10,R) example, case II.

then a third-order polynomial will be missing with highest weight $\Lambda_0-(15/2)\epsilon-\alpha^1_{16}-2\alpha^2_{16}$.

For $\lambda_q<-15/2$ there exists unitarity.

so*(2n)

Dynkin diagram 

.

Let now M_Λ be a representation for so*($2n$) with $\Lambda=(\Lambda_1,...,\Lambda_n)$. In this case given $\Lambda=\Lambda_0+\lambda\epsilon$ we have $\epsilon=(\frac{1}{2},\frac{1}{2},...,\frac{1}{2})$. We consider the following conditions on its components

$$\Lambda_1=\Lambda_2=\cdots=\Lambda_i>\Lambda_{i+1}+1,\quad i\neq 1,$$

or, equivalently

$$\langle\Lambda_0,\mu_1\rangle=\langle\Lambda_0,\mu_2\rangle=\cdots=\langle\Lambda_0,\mu_{i-1}\rangle=0,$$
$$\langle\Lambda_0,\mu_i\rangle=n>1.$$

From the Jakobsen method we have

$$\alpha_0=\beta+(\mu_{n-1}+\mu_{n-2})+(\mu_{n-2}+\mu_{n-3})+\cdots+(\mu_{n-(n-i)}+\mu_{n-(n-i)-1}),$$

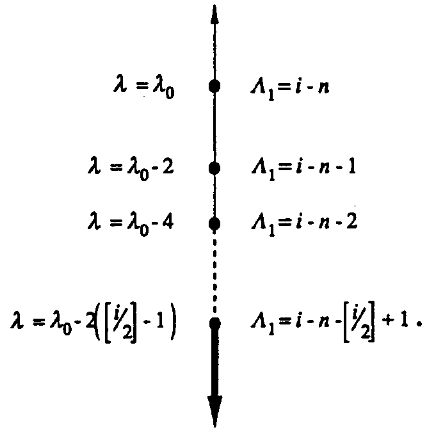
the height of which is $2(n-i)+1$. The condition $\langle\Lambda+R,\alpha_0\rangle=1$ implies in this case

$$(\Lambda_0,\alpha_0)+\lambda_0+2(n-i)+1=1,\quad \lambda_0=2(i-n).$$

For $\lambda=\lambda_0+\lambda_s=2(i-n-1)$ we obtain a second-order polynomial which will be missing with highest weight

$$\Lambda_0+2(i-n-1)\epsilon-\alpha^1_{2(n-i)+3}-\alpha^2_{2(n-i)+3}.$$

Following along these lines we arrive at $\lambda=\lambda_0+\{[i/2]-1\}\lambda_s$ where a polynomial with order $[i/2]$ is missing. For $\lambda<2(i-n-\{[i/2]-1\})$ there is impossible to obtain missing polynomials the order of which is strictly higher than $[i/2]$ because in those places the weights are not k_1 dominants then the reduction level is $[i/2]$. From the condition $\lambda=\langle\Lambda,\gamma_r\rangle$ we obtain $\Lambda_1=\lambda/2$. In this way we obtain the following diagram.



Example: so*(16)

We consider the following conditions on Λ_0:

$$\langle\Lambda_0,\mu_i\rangle=0,\quad\text{for } 1\leqslant i\leqslant 5,$$
$$\langle\Lambda_0,\mu_6\rangle=n_6>1,\quad \langle\Lambda_0,\mu_7\rangle=n_7.$$

Then

$$\alpha_0=\beta+2\mu_6+\mu_7+\mu_5,$$

with height 5 (see Fig. 6).

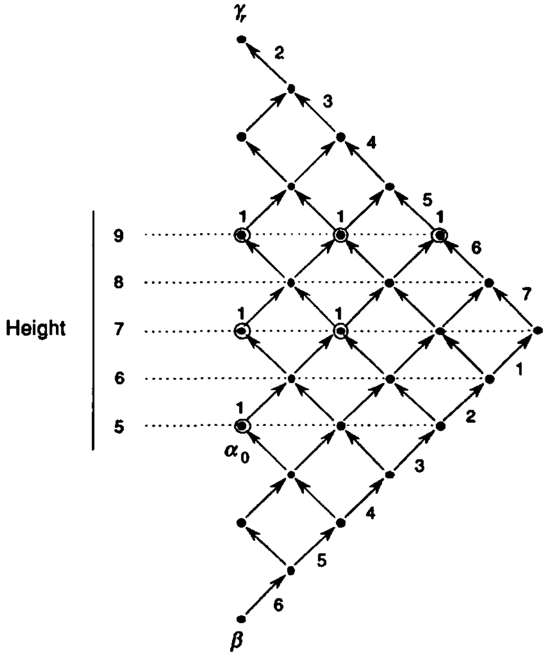


FIG. 6. Missing polynomials in the So*(16) example.

The condition $\langle \Lambda + R,\alpha_0\rangle = 1$ implies

$(\Lambda_0,\alpha_0) + \lambda_0 + 5 = 1, \quad \lambda_0 = -4.$

For $\lambda_q = \lambda_0 + \lambda_s = -6$ we have $\Lambda' = \Lambda_0 - 6\epsilon$ and

$\langle \Lambda' + R,\alpha_6^1\rangle = \lambda_q + 6 = 0,$

then it is not a valid root. However for the roots the height of which is 7 we have

$$\langle \Lambda' + R,\alpha_7^1\rangle = \langle \Lambda' + R,\alpha_7^2\rangle = 1,$$

and there will be a second-order polynomial that will be missing with highest weight:

$$\Lambda_0 - 6\epsilon - \alpha_7^1 - \alpha_7^2.$$

For $\lambda_q = \lambda_0 + 2\lambda_s = -8$, the roots are those with height 9:

$$\langle \Lambda + R,\alpha_9^1\rangle = \langle \Lambda + R,\alpha_9^2\rangle = \langle \Lambda + R,\alpha_9^3\rangle = \lambda_q + 9 = 1,$$

and a third-order polynomial will be missing with highest weight

$$\Lambda_0 - 8\epsilon - \alpha_9^1 - \alpha_9^2 - \alpha_9^3.$$

For $\lambda_q < -8$ there is no roots for which there exists k_1 dominance then the first possible place for nonunitarity is $\lambda_q = -8$.

The highest weight vectors are in this case, formally, the following:

(i) Height 5:

$$E^{-n_7-1}_{-\mu_7}E^{-n_6-n_7-1}_{-\mu_6}E^{-n_6-n_7-2}_{-\mu_5}E^{-n_6}_{-\mu_7}E^{-n_6-1}_{-\mu_6}E_{-\beta}$$
$$\times E^{n_6+2}_{-\mu_6}E^{n_6+1}_{-\mu_7}E^{n_6+n_7+3}_{-\mu_5}E^{n_6+n_7+2}_{-\mu_6}E^{n_7+1}_{-\mu_7},$$

with $\Lambda = \Lambda_0 - 4\epsilon - \beta - \mu_5 - 2\mu_6 - \mu_7$.

(ii) Height 7:

$$E^{-1}_{-\mu_4}E^{-n_7-1}_{-\mu_7}E^{-n_6-n_7-1}_{-\mu_6}E^{-n_6-n_7-3}_{-\mu_5}E^{-n_6}_{-\mu_7}E^{-n_6-2}_{-\mu_6}E^{2}_{-\beta}$$
$$\times E^{n_6+4}_{-\mu_6}E^{n_6+2}_{-\mu_7}E^{n_6+n_7+5}_{-\mu_5}E^{n_6+n_7+3}_{-\mu_6}E^{n_7+1}_{-\mu_7}$$
$$E^{3}_{-\mu_4}E_{-\mu_3}E_{-\mu_5},$$

with $\Lambda = \Lambda_0 - 6\epsilon - 2\beta - \mu_3 - 2\mu_4 - 3\mu_5 - 4\mu_6 - 2\mu_7$.

(iii) Height 9:

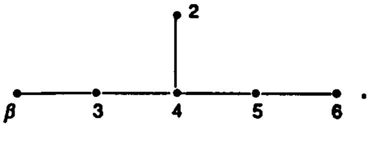
$$E^{-1}_{-\mu_3}E^{-2}_{-\mu_4}E^{-n_7-1}_{-\mu_7}E^{-n_6-n_7-1}_{-\mu_6}E^{-n_6-n_7-4}_{-\mu_5}E^{-n_6}_{-\mu_7}$$
$$\times E^{-n_6-3}_{-\mu_6}E^{3}_{-\beta}E^{n_6+6}_{-\mu_6}E^{n_6+3}_{-\mu_7}E^{n_6+n_7+7}_{-\mu_5}E^{n_6+n_7+4}_{-\mu_6}$$
$$\times E^{n_7+1}_{-\mu_7}E^{5}_{-\mu_4}E^{4}_{-\mu_3}E^{2}_{-\mu_5}E^{2}_{-\mu_2}E_{-\mu_1}E_{-\mu_4},$$

with

$$\Lambda = \Lambda_0 - 8\epsilon - 3\beta - \mu_1 - 2\mu_2 - 3\mu_3 - 4\mu_4 - 5\mu_5 - 6\mu_6 - 3\mu_7.$$

e₆

Dynkin diagram



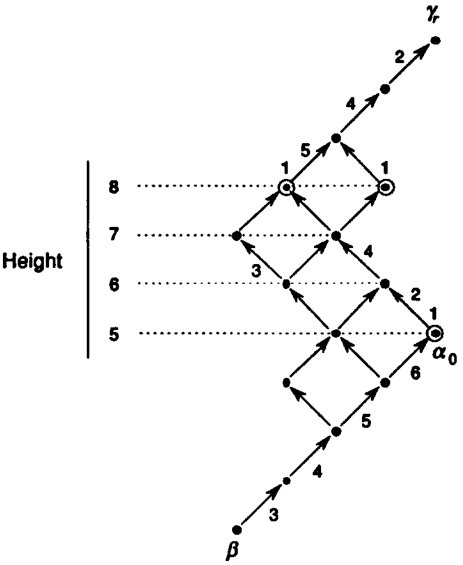


FIG. 7. Missing polynomials in e_6.

Here $\Lambda = \Lambda_0 + \lambda\epsilon$ with $\epsilon = (0,0,0,0,0,-\frac{2}{3},-\frac{2}{3},\frac{2}{3})$. For $\Lambda_0 \neq 0$ the only case for which the reduction level is strictly higher than one (all the rest are excluded for k_1 dominance arguments) is the following:

$$\langle \Lambda_0,\mu_i\rangle = 0, \quad 2 \leqslant i \leqslant 5,$$
$$\langle \Lambda_0,\mu_6\rangle = n > 0.$$

Applying Jakobsen method we obtain

$$\alpha_0 = \beta + \mu_3 + \mu_4 + \mu_5 + \mu_6,$$

with height 5 (see Fig. 7). From the condition $\langle \Lambda + R,\alpha_0\rangle = 1$ we obtain

$$1 = \langle \Lambda,\alpha_0\rangle + \langle R,\alpha_0\rangle = \langle \Lambda,\gamma_r\rangle + 5 = \lambda_0 + 5,$$
$$\lambda_0 = -4.$$

For $\lambda_q = \lambda_0 + \lambda_s = -7$,

$\langle \Lambda + R,\alpha_8^1\rangle = \langle \Lambda + R,\alpha_8^2\rangle = 1,$

then a second-order polynomial will be missing with highest weight:

$$\Lambda_0 - 7\epsilon - \alpha_8^1 - \alpha_8^2.$$

The value $\lambda_q = -7$ is the first possible place for nonunitarity.

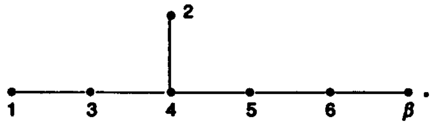
The highest weight vector corresponding to the highest weight

$\Lambda_0 - 4\epsilon - \beta - \mu_3 - \mu_4 - \mu_5 - \mu_6$ is, formally;

$$E^{-n}_{-\mu_6}E^{-n-1}_{-\mu_5}E^{-n-2}_{-\mu_4}E^{-n-3}_{-\mu_3}E_{-\beta}$$
$$\times E^{n+4}_{-\mu_3}E^{n+3}_{-\mu_4}E^{n+2}_{-\mu_5}E^{n+1}_{-\mu_6}.$$

e₇

Dynkin diagram



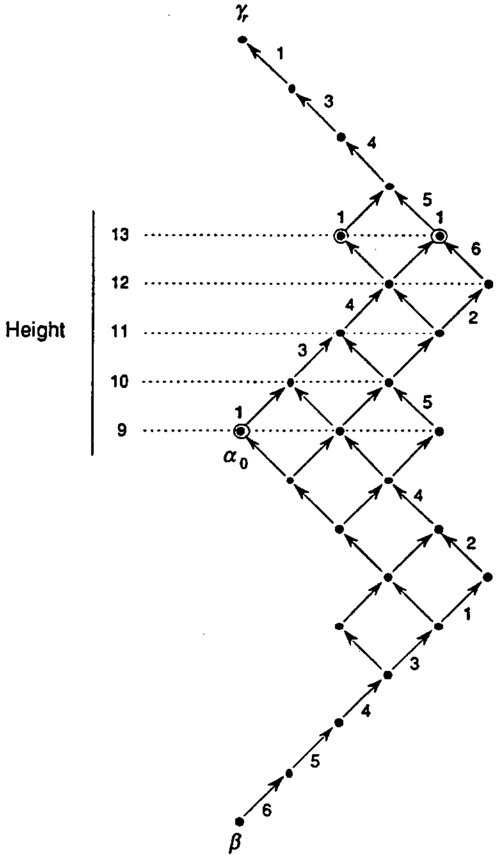


FIG. 8. Missing polynomials in e_7.

For this case $\epsilon = (0,0,0,0,0,1,-\frac{1}{2},\frac{1}{2})$. As in e_6 we must only consider one case:

$$\langle\Lambda_0,\mu_i\rangle = 0, \quad 1 \leqslant i \leqslant 5,$$
$$\langle\Lambda_0,\mu_6\rangle = n > 0,$$

with those conditions

$$\alpha_0 = \beta + 2\mu_6 + 2\mu_5 + 2\mu_4 + \mu_3 + \mu_2$$

with height 9 (see Fig. 8), then

$$1 = \langle\Lambda,\alpha_0\rangle + \langle R,\alpha_0\rangle = \langle\Lambda,\alpha_0\rangle + 9 = \lambda_0 + 9,$$
$$\lambda_0 = -8.$$

For $\lambda_q = -12$;

$$\langle\Lambda + R,\alpha_{13}^1\rangle = \langle\Lambda + R,\alpha_{13}^2\rangle = 1$$

and we obtain, for $\lambda_q = -12$, a missing second-order polynomial with highest weight

$$\Lambda_0 - 12\epsilon - \alpha_{13}^1 - \alpha_{13}^2.$$

The highest weight vector in height 9 is, formally:

$$\begin{aligned}
&E^{-1}_{-\mu_2}E^{-1}_{-\mu_3}E^{-3}_{-\mu_4}E^{-4}_{-\mu_5}E^{-n-4}_{-\mu_6}E^{-2}_{-\mu_2}E^{-2}_{-\mu_3}E^{-5}_{-\mu_4}E^{-n-5}_{-\mu_5}E^{-1}_{-\mu_6}E^{-3}_{-\mu_2}E^{-3}_{-\mu_3}E^{-n-6}_{-\mu_4}E^{-2}_{-\mu_5}E^{-1}_{-\mu_6}\\
&\quad\times E^{-n-3}_{-\mu_2}E^{-n-3}_{-\mu_3}E^{-n-2}_{-\mu_4}E^{-n-1}_{-\mu_5}E^{-n}_{-\mu_6}E_{-\beta}E^{n+1}_{-\mu_6}E^{n+2}_{-\mu_5}E^{n+3}_{-\mu_4}E^{n+4}_{-\mu_3}E^{n+4}_{-\mu_2}\\
&\quad\times E_{-\mu_6}E^{2}_{-\mu_5}E^{n+7}_{-\mu_4}E^{3}_{-\mu_3}E^{3}_{-\mu_2}E_{-\mu_6}E^{n+6}_{-\mu_5}E^{5}_{-\mu_4}E^{2}_{-\mu_3}E^{2}_{-\mu_2}E^{n+5}_{-\mu_6}E^{4}_{-\mu_5}E^{3}_{-\mu_4}E_{-\mu_3}E_{-\mu_2},
\end{aligned}$$

with highest weight

$$\Lambda_0 - 8\epsilon - \beta - \mu_2 - \mu_3 - 2\mu_4 - 2\mu_5 - 2\mu_6.$$

[1] M. Lorente and B. Gruber, J. Math. Phys. **25**, 1674–1681 (1984).

[2] B. Gruber, R.Lenczewski, and M. Lorente, accepted for publication in J. Math. Phys.

[3] H. Jakobsen, J. Funct. Anal. **52**, 385–412 (1983).

[4] N. R. Wallach, Trans. Am. Math. Soc. **251**, 1–17, 19–37 (1979).

[5] H. Jakobsen, Math. Ann. **256**, 439–447 (1981).

[6] Harish-Chandra, Am. J. Math. **77**, 743–777 (1955).

[7] N. Shapovalov, Funct. Ann. **6**, 307–312 (1972).

[8] R. Parthasarathy, Proc. Indian Acad. Sci. **89**, 1–24 (1980).

[9] T. Enright and R. Parthasarathy, *A Proof of a Conjecture of Kashiwara and Vergne*, Lecture Notes in Mathematics, Vol. 880 (Springer, Berlin, 1981).

[10] T. Enright, R. Howe, and N. Wallach, "A classification of unitary highest weight modules," in *Proceedings of the University of Utah Conference* 1982, edited by P. C. Tombi, Progress in Mathematics (Birkhäuser, Basel, 1983),

[11] J. A. Wolf, *Group Theoretical Methods in Physics*, edited by H. D. Doebner, J. D. Hennig, and T. D. Palev; Lecture Notes in Physics, Vol. 313 (Springer, Berlin, 1988).

[12] B. Gruber, A. U. Klimyk, and Y. F. Smirnov, Nuovo Cimento A **69** 97–127 (1982). (See also B. Gruber and M. Lorente, in *Proceedings of the Symposium on Indecomposable Representations of Lie Groups and its Physical Applications*, edited by V. Cantoni and A. O. Barut (Instituto di Alta Matematica, Universitá di Roma, Rome (to be published).

[13] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (World Scientific, Singapore, 1986).

[14] W. F Heidenreich and M. Lorente, J. Math. Phys. **29**, 1698–1704 (1988).

[15] M. Lorente, *Conformal Groups and Related Symmetries*, Lecture Notes in Physics, Vol. 261, edited by A. O. Barut and H. D. Doebner (Springer, Berlin, 1986), p. 101.

Maximum entropy formalism and analytic extrapolation

J. Antolín[a)]
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50009 Zaragoza, Spain

(Received 9 May 1989; accepted for publication 15 November 1989)

The maximum entropy principle is used to obtain a new analytic extrapolation method just complementary to the Padé-type method which leads to rigorous upper and lower bounds on the extrapolated function on the cut complex plane. Among the large class of functions that could equally well represent an analytical function in the experimental region a choice is made of the unique function that maximizes the entropy functional associated to this set of functions. The result is the least biased function compatible with the actual experimental data. This extrapolation method is applied to kaon–nucleon experimental data in order to obtain the most reasonable values for the KNY coupling constants compatible with the available experimental data and analytical constraints.

I. INTRODUCTION

Analytic continuation is perhaps the most clear-cut example of an undetermined inverse problem. The direct problem, restriction of a holomorphic function $f(z)$ to its values $f(t)$ $(t \in \Gamma)$, along some one-dimensional continuum Γ, follows a very smooth law; however, the inverse problem, which means "analytic continuation from Γ to the whole complex plane," is known to have a unique, but highly unstable, solution.

The practical (physical) situation is even worse because one never knows the function in a continuous set but in a discrete set of points and these known values are affected by errors coming from experimental measurements or from theoretical calculations. These facts make the task of analytic continuation, using experimental data, impossible without further assumptions. Ciulli has shown that, without further constraints, even the smallest uncertainties in the initial data lead, after extrapolation, to results differing by arbitrary amounts even in regions very close to the data region.[1]

Consequently there is a large class of functions that could equally well represent the analytic function in the experimental region, giving a good χ^2 fit, but which, when extrapolated, will give widely varying results.

The problem is then to search for other properties of the functions besides analyticity which act as stabilizers of the analytic extrapolation limiting the admissible number of parametrizations fitting the data or choosing among all these candidates the one most reasonable in a certain sense.

The nonuniqueness and instability of the analytic extrapolation of a function known with errors also forces the search for not only one, but various alternative solutions obtained with each method and also study the advantages and difficulties of each technique.

The aim of this paper is to show how the use of a very general physical principle, the maximum entropy (ME) principle, provides the analytic extrapolation method just complementary to the Padé-type method presented in a previous paper.[2] The latter provides rigorous upper and lower bounds on the values of the extrapolated function and the ME method chooses among all the admissible functions (which satisfy these bounds) the one most reasonable in the sense that it is the least biased function compatible with the information we actually have. The choice is made by maximizing the entropy functional associated to the set of functions compatible with the experimental data or theoretical calculations.

In Sec. II we briefly review the type of functions we are going to deal with and the use of positivity as a stabilizer of the analytic extrapolation. In Sec. III we introduce the ME formalism and apply it to our analytic extrapolation problem.

Finally in Sec. IV we apply the method to an extrapolation problem in particle physics, obtaining values for the KNY coupling constants compatible with the most recent determinations of these parameters by using different analytic extrapolation or model-dependent methods. The actual aim of this section is not so much to obtain another set of results on the controverted kaon–nucleon amplitudes as to show, in a concrete case, the complementary of the Stieltjes–Padé and maximum entropy methods in performing analytical extrapolations. The applications of both methods to deep inelastic structure functions,[3] to recent pion form factor measurements,[4] and to density-dependent quantities of many fermion systems[5] are in progress.

II. STIELTJES FUNCTIONS AND STABILIZATION OF THE ANALYTIC EXTRAPOLATION

We want to study functions, like scattering amplitudes, form factors, structure functions, or discrepancy functions, that are analytic in the complex plane except real cuts and perhaps some poles:[6]

$$\Delta(\omega)=\frac{1}{\pi}\int_a^b \frac{\operatorname{Im}\Delta(\omega')d\omega'}{\omega-\omega'}+\sum_{j=1}^{P}\frac{X_j}{\omega-\omega_j}. \qquad (2.1)$$

These functions satisfy, in general, the Schwarz reality condition, are supposed to be asymptotically polynomially

[a)] Permanent address: Departamento de Física Aplicada, EUITIZ, 50009 Zaragoza, Spain bounded, and are known with errors, coming from experimental measurements or from theoretical calculations in a set of discrete points on the real axis. We also have the very important positivity condition (or hypothesis) on the unknown imaginary part of the function $\Delta(\omega)$ on the cut, which acts as a stabilizer of the analytic extrapolation.

Our problem is to extrapolate the data to other regions where we do not know this function: pole positions in order to obtain their residues X_j, positions of complex or real zeros of the function, and the values of the function on the cuts.

By means of some transformations in the integration and evaluation variables and using absorption processes to avoid having the pole terms we turn the original data

$$\Delta(\omega_i)\pm\Delta^e(\omega_i), \quad i=1,\ldots,M, \tag{2.2}$$

into a new set of data on a Stieltjes function,[2,7] where the index e stands for error,

$$H(z(\omega))=\int_0^1 \frac{\chi(x(\omega))dx}{1+xz}, \tag{2.3}$$

$$\chi(x)=\frac{1}{\pi}\,\mathrm{Im}\,\Delta(\omega'(x))\prod^{P}\frac{1-x\varepsilon_j}{1+xz_j}\geqslant 0, \quad x\in[0,1], \tag{2.4}$$

where ε_j, related to the pole positions ω_j, and z_j, related to the absorption experimental points, are known parameters satisfying $0<\varepsilon_j<1$, $z_j>-1$.

The new data are the first $N+1$ moments of $\chi(x)$, which are, in turn, the first $N+1$ formal expansion coefficients of $H(z)$:

$$h_0\pm h_0^e,\ h_1\pm h_1^e,\ \ldots\ ,h_N\pm h_N^e, \tag{2.5}$$

where

$$H(z)=\sum_{n=0}^{\infty} h_n(-z)^n, \quad h_n=\int_0^1 \chi(x)x^n\,dx. \tag{2.6}$$

These coefficients, which constitute a totally monotonic (TM) sequence, must satisfy the following constraints[8,9]:

$$\{h_n\}\in\mathrm{TM}\Leftrightarrow\Delta^k h_n=\sum_{m=0}^{k}(-1)^m\binom{k}{m}h_{n+m}\geqslant 0, \quad n,k=0,1,2,\ldots, \tag{2.7}$$

which act as stabilizers of the analytic extrapolation. The Padé approximants (PA) constructed with these coefficients rigorously bound the function $H(z)$, and therefore $\Delta(\omega)$, in the cut complex plane.[8,10,11]

The convergence properties of PA fail on the real cut, but we can use the orthogonality properties of Padé denominators and the positivity of their residues to obtain an approximation to the weight function $\chi(x)$ in the form[12]

$$\chi_S^{(n)}(x)dx=d\psi_S^{(n)}(x)=\sum_{i=1}^{n} f_i^{(n)}\delta(x-\varepsilon_i^{(n)})dx, \tag{2.8}$$

where $N+1=2n$ is the number of known moments and

$$[n-1/n]=\sum_{i=1}^{n}\frac{f_i^{(n)}}{1+z\varepsilon_i^{(n)}}, \quad \varepsilon_i^{(n)}\in[0,1], \quad f_i^{(n)}>0, \tag{2.9}$$

is the PA to $H(z)$.

The main advantage is that we get rigorous upper and lower bounds on the distribution $\psi(x)$. On the other hand, the main difficulty is that we have discontinuous approximations to the distribution and a smoothing procedure is needed in order to have continuous approximations to the weight $\chi(x)$.[12,13]

As can be seen, a unique reconstruction of $\chi(x)$, and therefore of $\Delta(\omega)$, on the cuts is impossible in view of the limited information we have—the first $N+1$ moments of a positive function $\chi(x)$, affected by errors.

III. THE MAXIMUM ENTROPY METHOD

The nonuniqueness of the solution of the previous reduced moment problem forces the search for other alternative methods in order to compare the different solutions. In view of the main difficulty with the Padé method, the discontinuity of the approximations on the cuts, we are going to try and get approximations that automatically have positivity and continuity.

The ME method is based on a very general principle, which is one of the foundations of statistical mechanics and has recently had a large number of successful applications in other inverse problems including image reconstruction, data analysis, and information theory.[14–16]

For many years it has been recognized that entropy acts as a kind of measure in the space of probability distributions, in such a way that those distributions of high entropy are in some sense favored over others. Nature prefers distributions of maximum entropy because distributions of higher entropy are more likely than others.

We can state briefly the principle in this way[14]: When we make inferences based on incomplete information we should draw them from that probability distribution having the maximum entropy permitted by the information we do have.

The incomplete information we have now is the set of $N+1$ moments of a function and the ME principle says that we have to choose between all the weight functions compatible with the constraints imposed by their first $N+1$ moments the one that maximizes the entropy functional of the weight function:

$$S(\chi)=-\int_0^1\chi(x)\ln\chi(x)dx+\sum_{n=0}^{N}\lambda_n\left(h_n-\int_0^1\chi(x)x^n\,dx\right). \tag{3.1}$$

The ME choice is the least biased choice we can make taking into account the information we actually have. We are going to see how this general physical principle leads to sequences of approximations that have many interesting and concrete mathematical properties.

To calculate the ME solution to our problem, we have to solve this Lagrange multiplier problem: Find the maximum of $S(\chi)$ permitted by the constraints. Functional variation with respect to the unknown $\chi(x)$ gives the expression for the ME solution

$$\chi_E^{(N)}=\exp\left(\sum_{n=0}^{N}\lambda_n x^n\right), \tag{3.2}$$

supplemented by the conditions

$$h_n=\int_0^1 x^n\chi_E^{(N)}(x)dx, \quad n=0,1,\ldots,N. \tag{3.3}$$

We can see how these approximations automatically incorporate positivity and continuity. In order to get maximum entropy solutions we have to solve a nonlinear system of $N+1$ equations with $N+1$ unknown Lagrange multipliers. The system cannot be solved analytically except for $N=1$.

After normalization we have the following relation between λ_0 and the remaining Lagrange multipliers:

$$Z\equiv\exp(-\lambda_0)=\int_0^1\exp\left(\sum_{n=1}^{N}\lambda_n x^n\right). \tag{3.4}$$

Therefore we have to solve this system of N equations:

$$\langle x^n\rangle=h_n, \quad n=1,2,\ldots,N, \tag{3.5}$$

where

$$\langle x^k\rangle=\frac{1}{Z}\int_0^1 dx\, x^k\exp\left(-\sum_{n=1}^{N}\lambda_n x^n\right). \tag{3.6}$$

Now we introduce a potential function

$$U(\lambda_1,\lambda_2,\ldots,\lambda_n)\equiv\ln Z+\sum_{n=1}^{N}\lambda_n h_n, \tag{3.7}$$

whose stationary points are also the solutions of maximum entropy

$$\frac{\partial U}{\partial\lambda_n}=0\Rightarrow\langle x^n\rangle=h_n, \quad n=0,1,\ldots,N. \tag{3.8}$$

There are some properties concerning the solutions of ME and the potential $U(\lambda_1,\lambda_2,\ldots,\lambda_n)$.[16]

First it can be proved that the potential U is everywhere convex. This means that if a stationary point is found it must be a unique absolute minimum. Conversely convexity alone does not guarantee the existence of the minimum. The existence of a ME solution depends on the sequence of known moments, as we can easily see in the analytic case $N=1$:

$$Z=\int_0^1 dx\exp(-\lambda_1 x)=\frac{1-\exp(-\lambda_1)}{\lambda_1}, \tag{3.9}$$

$$U(\lambda_1)=\ln((1-\exp(-\lambda_1))/\lambda_1+\lambda_1)h_1. \tag{3.10}$$

Thus $U(\lambda_1)$ is a convex function but possesses a minimum at some finite λ_1 only if $h_1<1=h_0$.

The conditions the sequence of moments $\{h_n\}_{n>0}$ must satisfy to guarantee the existence of a ME solution are given by the following theorem.[17]

Theorem 1: A necessary and sufficient condition for the potential $U(\lambda_1,\lambda_2,\ldots,\lambda_n)$ to have a unique minimum at some finite values of λ's, for any N, is that the moment sequence $\{h_n\}_{n>0}$ should be a totally monotonic sequence:

$$\{h_n\}_{n>0}\in\text{TM}. \tag{3.11}$$

This theorem guarantees the existence of a ME solution $\chi_E^{(N)}(x)$, for any N. The solution is non-negative, absolutely continuous, and satisfies the reduced moment problem.

Given a finite set of moments $\{h_0,h_1,\ldots,h_N\}$ of an unknown weight function $\chi(x)$, moment theory provides rigorous lower and upper bounds for a functional of $\chi(x)$, like $H(z)$, for instance, by using the approximation $\chi_S^{(n)}(x)$, for $\chi(x)$ [the PA of $H(z)$]. Besides the ME formalism selects from this acceptable set of approximations the least biased one for the unmeasured characteristics of the experimental data [the one associated with $\chi_E^{(N)}(x)$, $H_E(z)$].

The following inequalities are then verified:

$$\begin{aligned}&0\leqslant[0/1]\leqslant\cdots\leqslant[n-1/n]\leqslant H(z)\sim H_E(z)\\&=\int_0^1\frac{\chi_E^{(N)}(x)dx}{1+xz}\leqslant[n-1/n-1]\leqslant\cdots\leqslant[0/0],\\&\text{for } z\leqslant 0,\quad N=2n-1,\\[1ex]&0\leqslant[0/0]\leqslant[0/1]\leqslant\cdots\leqslant[n-1/n-1]\leqslant[n-1/n]\\&\leqslant H(z)\sim H_E(z),\quad\text{for } -1<z<0,\quad N=2n-1,\end{aligned} \tag{3.12}$$

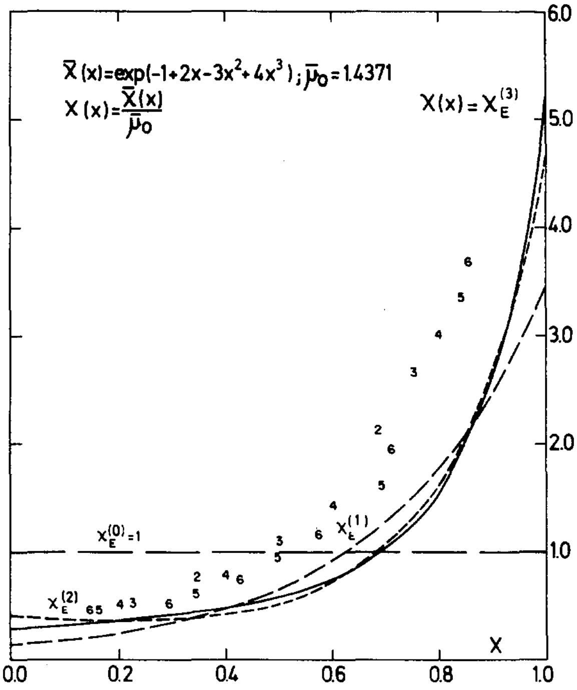


FIG. 1. Maximum entropy weights for an exponential of a polynomial. The fourth approximation is exactly the weight function. ME approximations are exact for these kind of functions. Figures in the plot show the Bernstein averages of corresponding order.

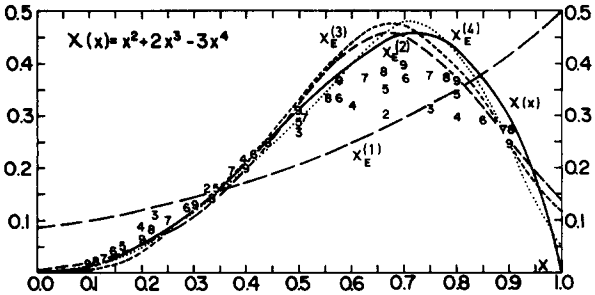


FIG. 2. Comparison between ME approximations to a polynomic model function obtained with $N+1=2, 3, 4, 5$ moments, $\chi_E^{(N)}$, and Bernstein averages at points $x_{N,K}$ obtained with 2, 3,...,9 moments. Figures in the plot show the Bernstein averages of corresponding order.

where $[l/m]$ is the PA for $H(z)$ constructed with the first $l+m+1$ moments. There are also other inequalities obtained from the modified PA $[n-1/n]^c$ and $[n/n]^c$ constructed by using the nonzero radius of convergence $(R=1)$ of the series expansion of $H(z)$, which are just complementary to the previous relations.[2]

For complex z one has inclusion regions for $H(z)$ and its ME approximation $H_E(z)$. Average convergence for ME approximations has also been shown.[17]

Theorem 2: A ME sequence $\chi_E^{(N)}$ converges in the following sense:

$$\lim_{N\to\infty}\int_0^1 \chi_E^{(N)}(x)F(x)dx=\int_0^1 \chi(x)F(x)dx\equiv\langle F\rangle, \tag{3.13}$$

where $F(x)$ is some continuous function. In particular we have, as with PA, convergence for Stieltjes functions, i.e., $F(x,z)=1/(1+xz)$, real $z, z>-1$.

As PA are exact approximations to rational functions, i.e., when the weight function is a finite sum of delta functions, the ME method is also exact when we have the moments of an exponential of polynomials. Figure 1 shows the first three approximations to an exponential polynomial of degree 3. In fact, the fourth approximation is just the exact weight. If we insist and try higher polynomials the new λ's are always zero.

The advantage the ME distribution $\psi_E^{(N)}(x)$ has among all the admissible distributions having the same first $N+1$ moments is that it is maximally noncommital about the unknown moments of the true distribution and in this sense is the best choice we can take.

This fact has been checked by comparing the ME method to another, more classical, method to invert moments: the Bernstein polynomial method used in Ref. 3.

In this method one constructs averages for $\chi(x)$ in the form

$$\tilde{\chi}(x_{N,K})\equiv\int_0^1 B^{(N,K)}(x)\chi(x)dx,$$
$$K=0,1,\ldots,N,\quad N=1,2,\ldots, \tag{3.14}$$

where $B^{(N,K)}$ is the normalized Bernstein polynomial

$$B^{(N,K)}(x)=\frac{x^K(1-x)^{N-K}}{\int_0^1 x^K(1-x)^{N-K}dx}, \tag{3.15}$$

which acts as weight over the mean points

$$x_{N,K}=\int_0^1 B^{(N,K)}(x)x\,dx=\frac{K+1}{N+2}\in[0,1],$$
$$K=0,1,\ldots,N,\quad N=1,2,\ldots. \tag{3.16}$$

Using (3.15) in (3.16) the averages can be easily calculated as

$$\tilde{\chi}(x_{N,K})=\frac{(N+1)!}{K!}\sum_{L=0}^{N-K}\frac{(-1)^L}{L!(N-K-L)!}h_{K+L},$$
$$K=0,1,\ldots,N,\quad N=1,2,\ldots, \tag{3.17}$$

where h_{K+L} are the known moments of $\chi(x)$.

The main advantage of this method is that one has a very fast and easy estimation of the function $\chi(x)$ at a set of points $x_{N,K}$. The Bernstein weights are suitable weights because they become more and more peaked around $x_{N,K}$ as N increases.

Figure 2 shows a comparison between the ME method and the Bernstein method showing how the first four ME approximations to a polynomic model function are better than the Bernstein averages obtained by using many more moments.

The stability of the ME extrapolation, outside the cut, i.e., for averages as in (3.13), is guaranteed owing to the constraints the moments of $\chi(x)$ must fulfill in order to be a TM sequence (2.7). Stability is guaranteed by positivity.

In practice the results of ME extrapolations outside the cut are bounded by the PA extrapolations [see (3.12)]. We have shown by using model functions[11] the stability of the latter, not only when the experimental errors tend to zero but when realistic errors associated to the physical processes are taken into account. We have tested this fact in Sec. IV.

IV. APPLICATION TO KN AMPLITUDES

In a set of recent papers we have analyzed the $K^{\pm}p$ and $K^{\pm}n$ forward elastic amplitudes by using the available experimental data, analytical properties of these complex functions, and the analytical extrapolation method briefly described in Sec. II.

TABLE I. Results of a three parameter fit for $K^{\pm}p$ and $K^{\pm}n$ amplitudes. Here ε_1 is the position of the pole and z_1 the experimental point used to absorb the pole term.

	$K^{\pm}p$	$K^{\pm}n$
ε_1	0.650	0.807
z_1	0.245	9.640
h_0	1.44±0.21	0.205±0.018
h_1	0.39±0.22	0.047±0.010
h_2	0.134±0.09	0.017±0.005
χ_R^2	0.36	0.82

TABLE II. Upper and lower sequences of the moments allowed by positivity.

	h_0	h_1	h_2	h_3	h_4	h_5	h_6	h_7
$K^{\pm}p$	0.144	0.39	0.134	0.051 5 0.048 7	0.022 5 0.019 5	0.011 0 0.008 3	0.007 6 0.003 9	0.005 7 0.002 0
$K^{\pm}n$	0.205	0.047	0.017	0.007 64 0.005 92	0.003 94 0.003 21	0.002 28 0.001 74	0.001 45 0.000 98	0.001 00 0.000 62

We have applied the ME method to the same experimental data in order to compare both methods and also to obtain the ME extrapolations for the kaon–nucleon–hyperon coupling constants which must be, according to the ME principle, the unique values of the coupling constants that take into account only the available experimental data and the analytical properties of the amplitudes.

It is clear that dispersion relations for the $K^{\pm}p$ and $K^{\pm}n$ amplitude[18] allow one to obtain transformed functions like (2.1). In this case, ω is the laboratory kaon energy, the X_j are related to the coupling constants, $\Delta(\omega)$ is the kaon–nucleon discrepancy function, and the integral is related to the nonphysical cut (experimentally inaccessible) dominated by the positive contribution of the Y^*_{1405} resonance.[19] The Gronwall transformation allows the determination of the first moments of the related Stieltjes

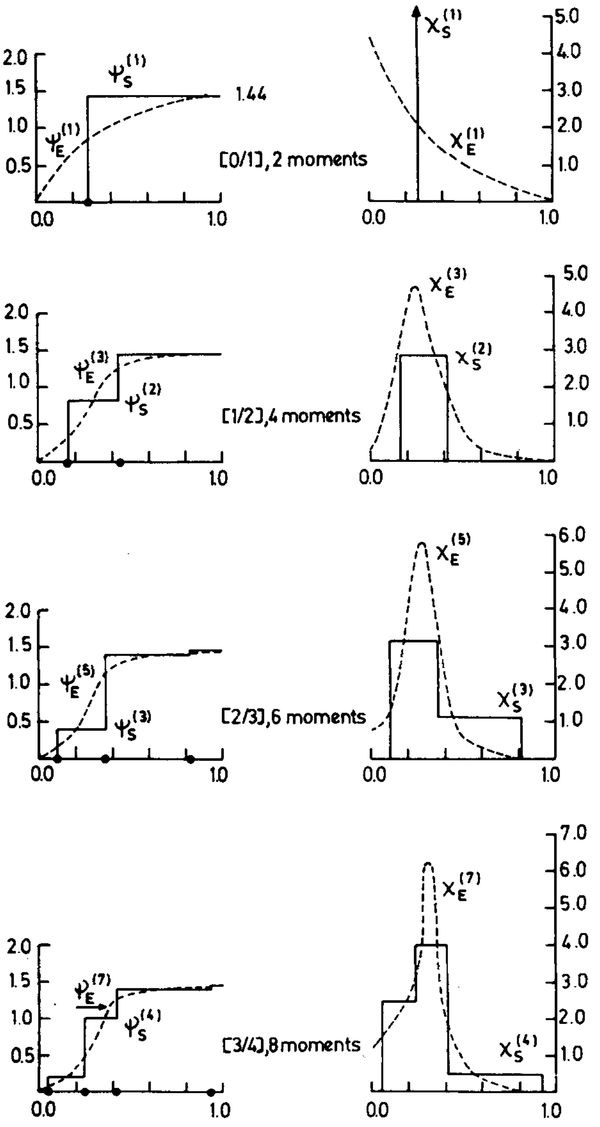


FIG. 3. Comparison between Stieltjes ($\psi_S^{(N)}$, $\chi_S^{(n)}$) and ME ($\psi_E^{(N)}$, $\chi_E^{(N)}$) extrapolations for the $K^{\pm}p$ unknown weight, obtained with 2, 4, 6, and 8 moments. The ME distributions keep inside the bounds imposed by the Chebyshev inequalities. $N=2n-1$.

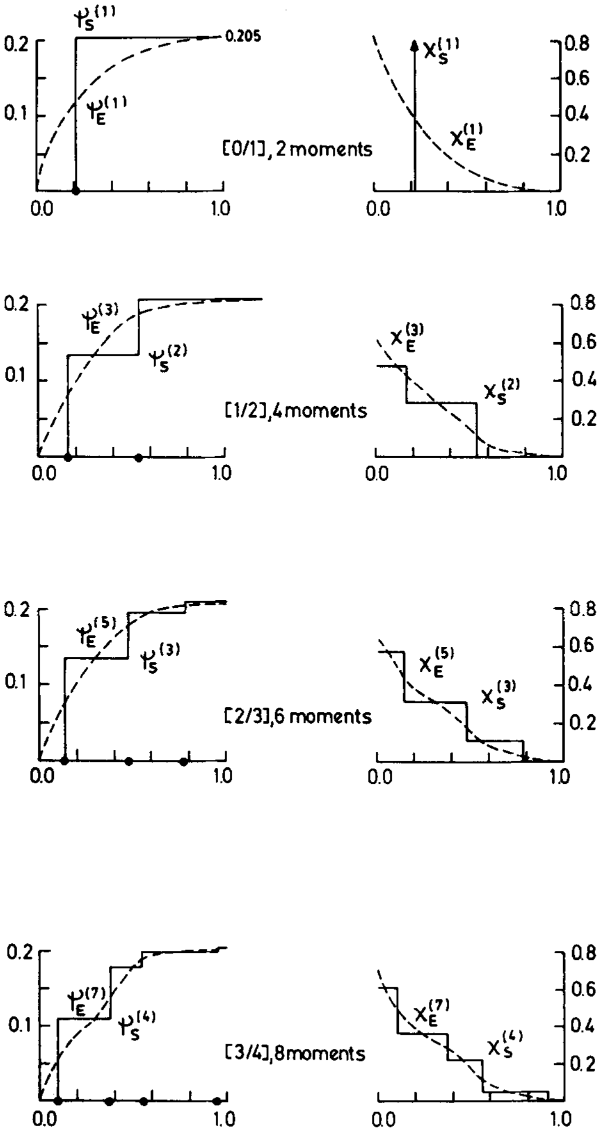


FIG. 4. Comparison between Stieltjes and ME extrapolations for the $K^{\pm}n$ unknown weight.

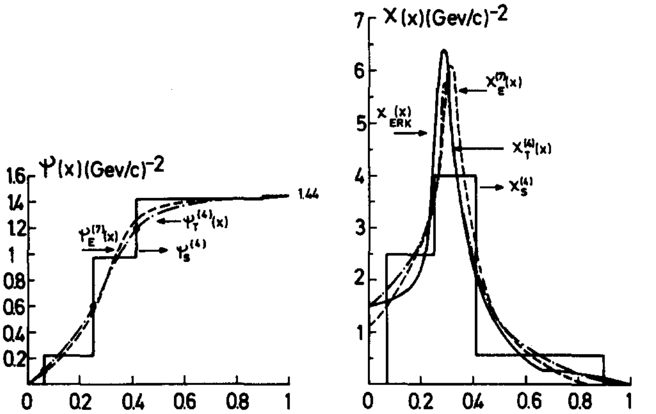


FIG. 5. Comparison between Stieltjes (S), Chebyshev (T), maximum entropy (E), and a model-dependent extrapolation (ERK) for the $K^{\pm}p$ distribution (left) and weight (right). $N=2n-1$.

function (2.3) from total KN cross sections and the measured real parts of the amplitudes. More physical details can be found in Refs. 11 and 20.

We have used this finite set of moments to calculate the associated ME weights and distributions $[\chi_E^{(N)}(x), \psi_E^{(N)}(x)]$ and afterwards the ME coupling constants.

We have done the following ME analyses of data:

(A1) $K^{\pm}p$ amplitudes: $F_p^{\pm}(\omega) = D_p^{\pm}(\omega) + iA_p^{\pm}(\omega)$

Experimental data:

(i) total cross sections $K^{\pm}p \rightarrow K^{\pm}p$ (see Ref. 21);

(ii) the 218 measurements of $D_p^{\pm}(\omega)$ (see Ref. 22).

Hypotheses:

(i) positivity and unimodality of $A_p(\omega)$ on the unphysical cut[18,19];

(iii) only one reduced pole to simulate the closeness of Λ and Σ.

Most interesting parameter:

$G^2 = G_{KN\Lambda}^2 + 0.9\, G_{KN\Sigma}^2$.

(A2) $K^{\pm}n$ amplitudes: $F_n^{\pm}(\omega) = D_n^{\pm}(\omega) + iA_n^{\pm}(\omega)$

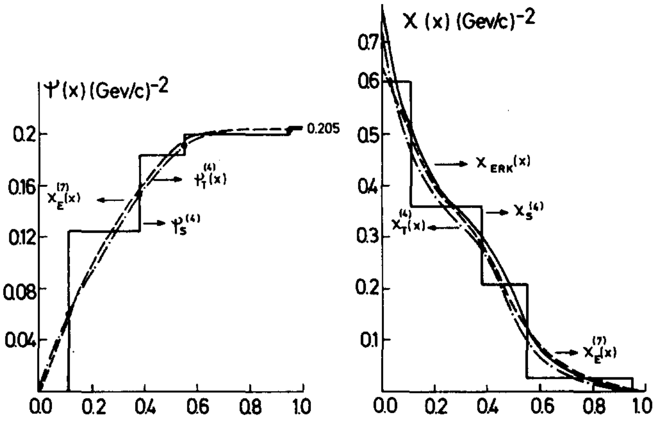


FIG. 6. Comparison between Stieltjes (S), Chebyshev (T), maximum entropy (E), and a model-dependent extrapolation (EKR) for the $K^{\pm}n$ distribution (left) and weight (right).

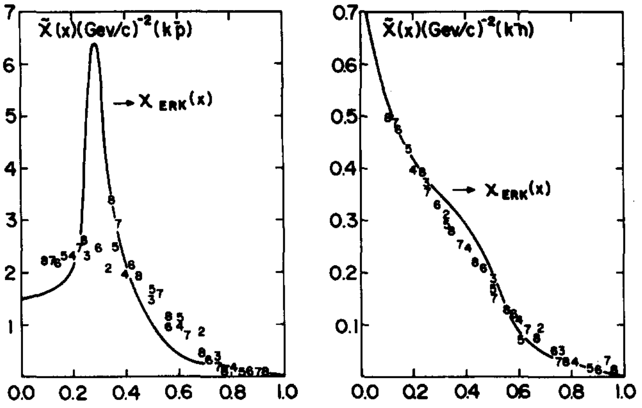


FIG. 7. Bernstein averages obtained by using the same information we have used in Fig. 5 and 6 for the K^-p and K^-n case. Figures in the plot show the Bernstein averages of corresponding order. The continuous line is the (model-dependent) effective range K-matrix result.

Experimental data:

(i) total cross sections $K^{\pm}n \rightarrow K^{\pm}n$ (see Ref. 21);

(ii) the five measurements of $D_n^{\pm}(\omega)$ (see Ref. 23);

(iii) the 115 measurements of the charge-exchange reaction[24]

$$\left.\frac{d\sigma}{d\Omega}\right|_{\theta=0^\circ}(K^-p \rightarrow \bar{K}^0 n) = |D_p^- - D_n^-|^2 + |A_p^- - A_n^-|^2. \qquad (4.1)$$

Hypotheses:

(i) positivity of $A_n(\omega)$ on the unphysical cut;

(ii) a parametrization for the relatively well known values of $D_p^-(\omega)$.

Most interesting parameter: $G_{KN\Sigma}^2$.

Table I shows a result for the moments of the unknown weights $\chi^n(x)$ and $\chi^P(x)$, corresponding to $K^{\pm}p$ and $K^{\pm}n$ amplitudes, respectively, by using the above data and hypotheses. The points used in the pole absorption, z_1, were varied systematically in the energy region obtaining similar results. It is worth remarking that statistical errors quoted in this table can be reduced by using positivity. In other words, (2.7) does not allow the full error bars in the calculated moments.

Using the central values for the moments, we have also calculated the bounds for the successive moments permitted by the minimum and maximum allowed values of h_3 obtaining upper and lower moment sequences compatible with positivity (Table II).

We have tested the stability of the ME method by taking different sets of moments inside the error bars allowed by positivity. We have also checked that results are practically independent of whether eight or ten moments are used.

Figures 3 and 4 show comparisons between the Stieltjes–Padé and ME approximations to the weights $\chi(x)$ [(2.4)] related to $A^-(\omega)$ on the cut. The Stieltjes histograms for the distribution function $\psi_S(x)$ constructed by means of the poles and residues of the PA $[\epsilon_j^{(n)}$ and $f_j^{(n)}$ in (2.9)] can be seen. The interlacing properties of the zeros of the denominators, which are orthogonal polynomials with respect to the weight, are apparent. We can also see how the ME distributions $\psi_E(x)$ keep inside the bounds imposed by the Chebyshev inequalities.[13] On the right-hand side, we have the corresponding weight functions $\chi_S(x)$ obtained from the slopes of the segments joining the midpoints of the discontinuities of $\psi_S(x)$ and also the ME solutions.

Figures 5 and 6 show comparisons between a Stieltjes–Chebyshev approximation obtained by using quasiorthogonal polynomials,[12,13] our ME weight, and a model-dependent result (ERK matrix analysis)[19] for the same weight. Despite using very different methods and parametrizations, the results are quite similar.

We have used these two weights in (2.3) to reconstruct the ME approximation to the $K^{\pm}N$ discrepancy functions (2.1), and therefore the ME values for the related coupling constants. The final result is

$$G^2=15.80\pm1.1,\quad G^2_{KN\Sigma}=2.35\pm1.4. \tag{4.2}$$

The accuracy in these calculations has been estimated by using the errors in the moments as well as the constraints imposed by positivity. As expected the values for the coupling constants lie inside the allowed corridor calculated in Refs. 11 and 20.

Figure 7 shows the Bernstein averages obtained with 2, 3,...,8, moments for the K^-p and K^-n unknown weights. Although these averages are in agreement with the ME approximation in the K^-n case, the difficulty of the Bernstein method to simulate the maximum of the K^-p case is clear. On the other hand, the ME approximations obtained by using only four moments (see Figs. 3 and 4) are already very good despite the little information we used.

This application illustrates our double general conclusion in performing analytical extrapolations.

(i) We need various alternative methods owing to the instability and nonuniqueness of the solution.

(ii) We should use these methods not as competitors but as complementary methods, in order to check the different solutions.

This is the case of Stieltjes–Padé and ME methods.

ACKNOWLEDGMENTS

I thank Dr. A. Cruz and Dr. N. M. Queen for their interest in this work.

I acknowledge the financial support of Comisión Asesora de Investigación Científica y Técnica (Plan movilizador de la Física de Altas Energías).

[1] S. Ciulli, C. Pomponiu, and I. Sabba-Stefanescu, Phys. Rep. **17**, 133 (1975).

[2] J. Antolín and A. Cruz, J. Math. Phys. **27**, 104 (1986).

[3] F. J. Yndurain, Phys. Lett. B **74**, 68 (1978).

[4] J. A. Alonso and J. A. Casas, Phys. Lett. B **197**, 239 (1987).

[5] F. W. King and K. J. Dykema, J. Phys. B **16**, 2071 (1983); F. J. Galvez and J. S. Dehesa, Phys. Rev. A **37**, 3154 (1988).

[6] J. A. Casas, C. López, and F. J. Yndurain, Phys. Rev. D **32**, 736 (1985); S. Ciulli and T. D. Spearman, *ibid.* **27**, 1580 (1983).

[7] T. H. Gronwall, Ann. Math. **33** (2), 101 (1932).

[8] J. Gilewicz, "Approximants de Padé," in *Lecture Notes in Mathematics*, Vol. 667 (Springer, Berlin, 1978).

[9] C. Brezinski, *Padé-Type Approximation and General Orthogonal Polynomials* (Birkhauser, Basel, 1980).

[10] G. A. Baker, *Essentials of Padé Approximants* (Academic, London, 1975).

[11] J. Antolín and A. Cruz, J. Phys. G **12**, 297 (1986).

[12] J. Antolín and A. Cruz, J. Phys. G **12**, 947 (1986).

[13] T. S. Chihara, *An Introduction to Orthogonal Polynomials* (Gordon and Breach, New York, 1978).

[14] E. T. Jaynes, Proc. IEEE **70**, 939 (1982).

[15] S. F. Gull and J. Skilling, IEE Proc., **131**, 646 (1984).

[16] R. D. Levine, J. Phys. A: Math. Gen. **13**, 91 (1980); N. Agmon, Y. Alhassid, and R. D. Levine, J. Comp. Phys. **30**, 250 (1979).

[17] L. R. Mead and N. Papanicolau, J. Math. Phys. **25**, 2404 (1984).

[18] B. di Claudio, G. Violini, and N. M. Queen, Nucl. Phys. B **161**, 238 (1979); G. K. Atkin, J. E. Bowcock, and N. M. Queen, J. Phys. G: Nucl. Phys. **7**, 613 (1981).

[19] A. D. Martin, Phys. Lett. B **65**, 346 (1976).

[20] J. Antolín, Phys. Rev. D **35**, 122 (1987).

[21] V. Flaminio, I. F. Graf, J. D. Hansen, W. G. Moorhead, and D. R. O. Morrison, CERN-HERA 79-02, 1979.

[22] N. M. Queen, University of Birmingham Report UB-K_p-1-78, 1978.

[23] P. Baillon, Y. Declais, M. Ferro-Luzzi, P. Jenni, J. M. Perreau, J. Seguinot, and T. Ypsilantis, Nucl. Phys. B **134**, 31 (1978); P. Jenni, P. Baillon, C. Brieman, M. Ferro-Luzzi, J. M. Perreau, R. D. Tripp, and T. Ypsilantis, *ibid.* **105**, 1 (1976).

[24] M. Alston-Garnjost, R. W. Kenney, D. L. Pollard, R. R. Ross, and R. D. Tripp, Phys. Rev. D **17**, 2266 (1978).

A simple Lagrangian for integrable systems

M. A. de Almeida da Silva and Ashok Das[a]

Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ 21945, Brazil

(Received 22 September 1989; accepted for publication 29 November 1989)

A simple Lagrangian for integrable models that can describe the entire hierachy of equations associated with the system is proposed. In addition, the Lax equation, the recursion relation between the conserved quantities, as well as the vanishing of the Nijenhuis torsion tensor can also be derived from the system of equations following from this Lagrangian.

The study of integrable systems has led to many interesting concepts (for reviews see Refs. 1 and 2). Among them, let us note that every integrable system has associated with it a Lax pair and a Lax equation.[3] Given a Lax equation, one can construct the conserved quantities of the theory quite easily. The involution of the conserved quantities, which is necessary to prove integrability, follows if they satisfy some recursion relation. In an integrable system, each of the conserved quantities can be thought of as a Hamiltonian and all such Hamiltonians would have commuting flows. In simple language, this means that with every integrable equation is associated a hierarchy of equations that are all integrable.

Integrable systems have been studied from various points of view.[1,2] Recently, the integrability of such systems was analyzed from a study of their phase-space geometry.[4,5] Let us note that a very special feature of integrable systems is the existence of more than one Poisson bracket structure in these theories.[6] Consequently, the phase space of such systems corresponds to a special class of symplectic manifolds. Assuming that a dual Poisson bracket structure exists for such theories, the geometrical meaning of various concepts such as the Lax equation was derived in Refs. 4 and 5 and it was also shown that a sufficient condition for integrability corresponds to the vanishing of the appropriate Nijenhuis torsion tensor on this manifold.[7] In this article, we propose a simple Lagrangian that would generate the hierachy of equations of an integrable system. The Lax equation and, therefore, the conserved quantities, as well as the recursion relation between them can also be derived quite easily, leading to the fact that such quantities are in involution. This can also be shown to be equivalent to the vanishing of the Nijenhuis torsion tensor. In what follows, we will use the notation of Ref. 4 and describe the Lagrangian and the resulting consequences for a finite-dimensional system. We would conclude with the Korteweg–de Vries (KdV) equation as an example.

Let us consider an integrable system whose phase space is parametrized by the generalized coordinates y^μ, $\mu = 1,2,...,2N$. Let the two distinct Poisson bracket structures be $f^{\mu\nu}(y)$ and $F^{\mu\nu}(y)$. The symplectic structures of the manifold are defined by the inverses of these antisymmetric tensors, namely,

$$f_{\mu\lambda}(y)\, f^{\lambda\nu}(y) = \delta_\mu{}^\nu = F_{\mu\lambda}(y)\, F^{\lambda\nu}(y)\ . \tag{1}$$

From these antisymmetric tensors, one can construct a natural (1,1) tensor on the manifold as

$$S_\mu{}^\nu(y) = F_{\mu\lambda}(y)\, f^{\lambda\nu}(y)\ . \tag{2}$$

Let us next consider the following Lagrangian:

$$L = \theta_\mu(y)\dot{y}^\mu - (B(\lambda,y) - \lambda A(\lambda,y)) + \xi^\mu(S_\mu{}^\nu\, \partial_\nu A - \partial_\mu B), \tag{3}$$

where $\partial_\mu = \partial/\partial y^\mu$ and $\theta_\mu(y)$ defines the symplectic structure $F_{\mu\nu}(y)$ as

$$F_{\mu\nu}(y) = \partial_\mu\, \theta_\nu(y) - \partial_\nu\, \theta_\mu(y)\ . \tag{4}$$

Furthermore, ξ^μ is a Lagrange multiplier and the functions A and B depend on the coordinates y^μ, as well as on an arbitrary parameter λ. The Euler–Lagrange equations are now easily derived. The Lagrange multiplier merely enforces the constraint

$$S_\mu{}^\nu(y)\partial_\nu\, A(\lambda,y) - \partial_\mu\, B(\lambda,y) = 0. \tag{5}$$

Using Eq. (5), as well as its consistency, the dynamical equation now takes the form

$$\begin{aligned} F_{\mu\nu}(y)\dot{y}^\nu &= \partial_\mu\, B(\lambda,y) - \lambda\, \partial_\mu\, A(\lambda,y) \\ &= S_\mu{}^\nu(y)\partial_\nu\, A(\lambda,y) - \lambda\, \partial_\mu A(\lambda,y)\ . \end{aligned} \tag{6}$$

Equivalently, we can write

$$\dot{y}^\mu = f^{\mu\nu}(y)\partial_\nu\, A(\lambda,y) - \lambda F^{\mu\nu}(y)\, \partial_\nu\, A(\lambda,y)\ . \tag{7}$$

Let us next make the power series expansion

$$A(\lambda,y) = \sum_{j=0}^{n} \lambda^{\,n-j} A_j(y)\ , \tag{8}$$

where the positive integer n satisfies $n \leqslant N$. Substituting expansion (8) into Eq. (7), we note that since y^ν as well as the Poisson bracket structures are independent of the arbitrary parameter λ, consistency of Eq. (7) would lead to

$$f^{\mu\nu}(y)\partial_\nu\, A_j = F^{\mu\nu}(y)\partial_\nu\, A_{j+1},\quad j = 0,1,...,(n-1). \tag{9}$$

The λ-independent dynamical equation would then take the form

$$\dot{y}^\mu = f^{\mu\nu}(y)\partial_\nu\, A_n. \tag{10}$$

Consequently, A_n can be thought of as the true Hamiltonian of the system. Equation (10) also brings out the interplay of the two Poisson bracket structures of the system, namely, whereas the naive symplectic structure obtained from the Lagrangian would correspond to $F_{\mu\nu}(y)$, the final Poisson bracket structure in Eq. (10) is $f^{\mu\nu}(y)$. Let us note here that the recursion relation in Eq. (9) implies that all the coefficients in the power series expansion would be in involution (for details see Ref. 4): Also, since they would be in involution with A_n, they would also be conserved. Thus, in fact, we can identify the A_j's with the different Hamiltonians of the system and Eq. (10) would then describe the hierarchy of

[a] Permanent address: Department of Physics and Astronomy, University of Rochester, Rochester, NY 14627.

0022-2488/90/040798-03$03.00

equations of the system for different values of n.

Let us next derive the Lax equation for the system. Note that the recursion relation in Eq. (9) can also be written as

$$\partial_\mu A_{j+1} - S_\mu{}^\nu(y)\partial_\nu A_j = 0. \tag{11}$$

Relation (11) must hold for all times since otherwise Eq. (7) would become inconsistent. The time independence gives

$$\partial_\nu(\partial_\mu A_{j+1})\dot{y}^\nu - \frac{dS_\mu{}^\nu}{dt}\partial_\nu A_j - S_\mu{}^\lambda\partial_\nu(\partial_\lambda A_j)\dot{y}^\nu = 0 .$$

Using the commutativity of the derivatives, the above equation can also be written as

$$\partial_\mu\left[\frac{dA_{j+1}}{dt}\right] - \partial_\nu A_{j+1}\,\partial_\mu \dot{y}^\nu - \frac{dS_\mu{}^\nu}{dt}\partial_\nu A_j$$
$$- S_\mu{}^\lambda\partial_\lambda\left[\frac{dA_j}{dt}\right] + S_\mu{}^\lambda\partial_\lambda\,\dot{y}^\nu\,\partial_\nu A_j = 0 . \tag{12}$$

Let us define

$$U_\mu{}^\nu(y) = \partial_\mu\,\dot{y}^\nu \tag{13}$$

and note that since the A_j's are conserved, Eq. (12) can, in fact, be written [using Eq. (11)] as

$$\left[\frac{dS_\mu{}^\nu}{dt} - S_\mu{}^\lambda U_\lambda{}^\nu + U_\mu{}^\nu S_\lambda{}^\nu\right]\partial_\nu A_j = 0. \tag{14}$$

Equation (14) must hold for all the A_j's and, consequently, we must have

$$\frac{dS_\mu{}^\nu}{dt} = S_\mu{}^\lambda U_\lambda{}^\nu - U_\mu{}^\lambda S_\lambda{}^\nu ,$$

which in a matrix notation can be written as

$$\frac{dS}{dt} = [S,U] . \tag{15}$$

Equation (15) is, in fact, the Lax equation[3]; it follows from this that the quantities (see Ref. 4)

$$K_n = (1/n)\mathrm{Tr}\,S^n, \quad n = \pm 1, \pm 2,..., \tag{16}$$
$$K_0 = \log|\det S|$$

would be conserved under the evolution of the dynamical system. The Hamiltonians, A_j's, can then be identified with the appropriate linearly independent conserved quantities.

Let us next show that the vanishing of the Nijenhuis torsion tensor also follows from the equations of the system. Let us note that the recursion relation (11) implies that

$$\partial_\lambda\,\partial_\alpha A_{j+1} = \partial_\lambda S_\alpha{}^\mu\,\partial_\mu A_j + S_\alpha{}^\mu\,\partial_\lambda\partial_\mu A_j . \tag{17}$$

Consequently,

$$S_\alpha{}^\lambda\partial_\lambda\,\partial_\beta A_{j+1} = S_\alpha{}^\lambda\partial_\lambda S_\beta{}^\mu\,\partial_\mu A_j + S_\alpha{}^\lambda S_\beta{}^\mu\,\partial_\lambda\,\partial_\mu A_j , \tag{18}$$

$$S_\beta{}^\lambda\partial_\lambda\,\partial_\alpha A_{j+1} = S_\beta{}^\lambda\partial_\lambda S_\alpha{}^\mu\,\partial_\mu A_j + S_\beta{}^\lambda S_\alpha{}^\mu\,\partial_\lambda\partial_\mu A_j . \tag{19}$$

Subtracting Eq. (19) from Eq. (18), we obtain

$$[S_\alpha{}^\lambda\partial_\lambda S_\beta{}^\mu - S_\beta{}^\lambda\partial_\lambda S_\alpha{}^\mu]\partial_\mu A_j$$
$$= S_\alpha{}^\lambda\partial_\beta\,\partial_\lambda A_{j+1} - S_\beta{}^\lambda\partial_\alpha\,\partial_\lambda A_{j+1}$$
$$= \partial_\beta[S_\alpha{}^\lambda\partial_\lambda A_{j+1}] - \partial_\beta S_\alpha{}^\lambda\partial_\lambda A_{j+1}$$
$$- \partial_\alpha[S_\beta{}^\lambda\partial_\lambda A_{j+1}] + \partial_\alpha S_\beta{}^\lambda\partial_\lambda A_{j+1} . \tag{20}$$

If we now use the recursion relation in Eq. (11), we obtain

$$[S_\alpha{}^\lambda\partial_\lambda S_\beta{}^\mu - S_\beta{}^\lambda\partial_\lambda S_\alpha{}^\mu]\partial_\mu A_{j'}$$
$$= S_\lambda{}^\mu[\partial_\alpha S_\beta{}^\lambda - \partial_\beta S_\alpha{}^\lambda]\partial_\mu A_j . \tag{21}$$

Since (21) must hold for all the A_j's, it follows that

$$N^\mu_{\alpha\beta} = S_\alpha{}^\lambda\partial_\lambda S_\beta{}^\mu - S_\beta{}^\lambda\partial_\lambda S_\alpha{}^\mu$$
$$- S_\lambda{}^\mu[\partial_\alpha S_\beta{}^\lambda - \partial_\beta S_\alpha{}^\lambda] = 0 . \tag{22}$$

The lhs of Eq. (22) defines the Nijenhuis torsion tensor[8] associated with the (1,1) tensor $S_\mu{}^\nu(y)$ and its vanishing follows from the equations of the system. Thus the simple Lagrangian of Eq. (3) gives all the essential features of an integrable system.

We now conclude with the specific example of the KdV equation.[9] This is a (1 + 1)-dimensional continuum system with the dynamical variables described by $u(x,t)$. [Thus $y^\mu \to u(x,t)$.] In this case the two Poisson bracket structures correspond to

$$F^{\mu\nu}(y) \to D\delta(x-y) = \frac{\partial}{\partial x}\delta(x-y) ,$$
$$f^{\mu\nu}(y) \to M\delta(x-y) \tag{23}$$
$$= \left[\frac{\partial^3}{\partial x^3} + \frac{1}{3}\left[\frac{\partial}{\partial x}u(x) + u(x)\frac{\partial}{\partial x}\right]\right]\delta(x-y) .$$

Let us next consider the Lagrangian

$$L_{\mathrm{KdV}} = \frac{1}{2}\int dx\,\dot{u}(x)D^{-1}u(x) - \Big[B[\lambda,u] - \lambda A[\lambda,u]\Big]$$
$$+ \int dx\,\xi(x)\left[D^{-1}M\frac{\delta A}{\delta u(x)} - \frac{\delta B}{\delta u(x)}\right], \tag{24}$$

where D^{-1} is the inverse of the derivative operator with the appropriate asymptotic condition. Now A and B are functionals of the dynamical variable $u(x,t)$ and depend on the arbitrary parameter λ as well. The Euler–Lagrange equations following from Eq. (24) are [compare with Eqs. (5) and (6)]

$$D^{-1}M\frac{\delta A}{\delta u(x)} - \frac{\delta B}{\delta u(x)} = 0 , \tag{25}$$
$$D^{-1}\dot{u} = \frac{\delta B}{\delta u(x)} - \lambda\frac{\lambda A}{\delta u(x)}$$
$$= D^{-1}M\frac{\delta A}{\delta u(x)} - \lambda\frac{\delta A}{\delta u(x)} ,$$

or,

$$\frac{\partial u}{\partial t} = M\frac{\partial A}{\delta u(x)} - \lambda D\frac{\delta A}{\delta u(x)} . \tag{26}$$

The last of Eqs. (26) is, indeed, what was obtained by Lenard,[10] as well as by Chern and Peng[11] in their study of the KdV equation and corresponds to a zero-curvature condition. If we now make the power series expansion

$$A[\lambda,u] = \sum_{j=0}^{n}\lambda^{n-j}A_j[u] ,$$

then Eq. (26) reduces to

$$M\frac{\delta A_j}{\delta u(x)} = D\frac{\delta A_{j+1}}{\delta u(x)} \tag{27}$$

and

$$\frac{\partial u}{\partial t} = M\frac{\delta A_n}{\delta u(x)} = \left(D^3 + \frac{1}{3}(Du + uD)\right)\frac{\delta A_n}{\delta u(x)}. \tag{28}$$

The recursion relation in Eq. (27) is precisely the one satisfied by the conserved quantities of the KdV equation[6] and, consequently, the A_j's can be identified with the KdV Hamiltonians. Equation (28) then describes the nth equation in the KdV hierarchy. All our earlier analysis can now be applied to the KdV equation in a straightforward manner, taking the continuum nature into account.

In conclusion, we have presented a simple Lagrangian for integrable models that describes the entire hierarchy of equations. The Lax equation, the recursion relation between the conserved quantities, and the vanishing of the Nijenhuis torsion tensor also follow from the system of equations derived from this Lagrangian.

Note added: Since $S_\mu{}^\nu$ is a $2N \times 2N$ matrix and there can only be N conserved quantities in an integrable system, Eq. (15) cannot rigorously be derived from Eq. (14). However, from the existence of two Poisson bracket structures and of the recursion relation in Eq. (9), we can derive the Lax equation simply following the methods of Refs. 4 and 5. Similarly, it does not rigorously follow from Eq. (20) that the Nijenhuis torsion tensor vanishes. While the Nijenhuis torsion tensor may vanish for some integrable models, there exist integrable systems for which it does not vanish. We would like to thank Professor S. Okubo for pointing this out to us.

ACKNOWLEDGMENTS

One of us (AD) would like to thank the Instituto de Física at UFRJ for the hospitality and the Brazilian agencies CNPq and FAPERJ for the financial support during the visit when this work was done.

This work was supported in part by the US Department of Energy Contract No. DE-AC02-76ER13065.

[1] L. D. Faddeev and L. A. Takhtadjan, *Hamiltonian Methods in the Theory of Solitons* (Springer, Berlin, 1987).

[2] A. Das, *Integrable Models* (World Scientific, Singapore, 1989).

[3] P. D. Lax, Comm. Pure Appl. Math. **21**, 467 (1968); Comm. Pure Appl. Math. **28**, 141 (1975).

[4] S. Okubo and A. Das, Phys. Lett. B **209**, 311 (1988).

[5] A. Das and S. Okubo, Ann. Phys. **190**, 215 (1989).

[6] F. Magri, J. Math. Phys. **19**, 1156 (1978).

[7] See, also, G. Marmo, in *Proceedings of the International Meeting on Geometry and Physics, Florence, 1982,* edited by M. Modugno (Pitagora Editrice, Bologna, 1983).

[8] A. Nijenhuis, Indag. Math. **13**, 200 (1951).

[9] D. J. Korteweg and G. de Vries, Philos. Mag. **39**, 422 (1895).

[10] A. Lenard (unpublished), as reported in C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Comm. Pure Appl. Math. **27**, 97 (1974).

[11] S-S. Chern and C-K. Peng, Manuscr. Math. **28**, 207 (1979).

Generating functions, bi-Hamiltonian systems, and the quadratic-Hamiltonian theorem

José F. Cariñena and Manuel F. Rañada
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, 50009 Zaragoza, Spain

(Received 21 February 1989; accepted for publication 25 October 1989)

The concept of bi-Hamiltonian systems and its connection with the theory of canonoid transformations is shown from a geometrical viewpoint. The relations between symplectic and canonoid diffeomorphisms are studied using an approach based on the theory of generating functions. These results are used for obtaining a new theorem which will represent an intrinsic and coordinate-free generalization of the "quadratic Hamiltonian theorem."

I. INTRODUCTION

A dynamical system characterized by an even number of real variables x_i, y_i, where $i = 1,...,n$, with a time evolution represented by the first-order equations

$$\begin{cases} \dfrac{dx_i}{dt} = f_i(x_k,y_k), \\ \dfrac{dy_i}{dt} = g_i(x_k,y_k), \end{cases}$$

is called Hamiltonian if there is a function $H = H(x_k,y_k)$ such that $f_i(x_k,y_k) = \partial H/\partial y_i$ and $g_i(x_k,y_k) = -\partial H/\partial x_i$.

If the change of (x_i,y_i) for (X_i,Y_i) represents a transformation of the phase space, then the above equations of motion become

$$\begin{cases} \dfrac{dX_i}{dt} = F_i(X_k,Y_k), \\ \dfrac{dY_i}{dt} = G_i(X_k,Y_k), \end{cases}$$

and thus the transformed system will be Hamiltonian if there is a real function $K = K(X_k,Y_k)$ satisfying $F_i(X_k,Y_k) = \partial K/\partial Y_i$ and $G_i(X_k,Y_k) = -\partial K/\partial X_i$.

According to this, two different situations arise. Those transformations that preserve the Hamiltonian character of a concrete Hamiltonian dynamical system are called canonoid with respect to this particular system[1] and those preserving the form of such equations, whatever the original Hamiltonian function is, are called canonical.

The relation between canonical and canonoid transformations was investigated by Currie and Saletan,[2] who proved the so-called "canonical transformation theorem"[1] or "quadratic-Hamiltonian theorem,"[3] according to which a transformation is canonoid with respect to all quadratic Hamiltonians of the form

$$H = c_\alpha \xi^\alpha + d_{\alpha\beta}\xi^\alpha\xi^\beta,$$

where $\alpha = 1,...,2n$; c_α and $d_{\alpha\beta}$ are constant; and $\{\xi^1,...,\xi^n\}$ represent the positions and $\{\xi^{n+1},...,\xi^{2n}\}$ the momenta if and only if the Poisson brackets are invariant up to a nonzero multiplicative constant, i.e., it is a generalized or extended canonical transformation. This theorem has been considered by some authors as one of the fundamental theorems[1] of theoretical mechanics.

In recent years the geometrical approach to the theory of dynamical systems has deserved great attention. Thus a general dynamical system is interpreted as a vector field on a manifold M and dynamical systems whose evolution is described by Hamiltonian equations are interpreted as a special class of vector fields on symplectic manifolds (M,ω), i.e., those with flows preserving the symplectic form. Therefore, the different theorems and properties of the Lagrangian and Hamiltonian dynamics have been translated to this geometric setting[4] and in this way we now know that the momentum phase space is represented by a symplectic manifold, which turns out to be the cotangent bundle T^*Q of the configuration space Q endowed with its natural symplectic structure $\omega_0 = -d\theta_0$; univalent canonical transformations are represented by symplectomorphisms; and Poisson brackets of dynamical variables are represented by the action of ω_0 on their associated vector fields; etc. This approach is considered more fundamental than the traditional one mainly for three reasons: (i) it is valid for topologically nontrivial configuration spaces, (ii) the theorems and propositions are proved using an intrinsic or coordinate-free formulation and are then ready for a possible generalization to the infinite-dimensional case; and finally, (iii) properties previously known in the traditional approach appear here as particular cases of these new and more general results.

A canonical change of coordinates will preserve the Hamiltonian structure of the equations, but not the linear or quadratic character of the Hamiltonian. Consequently, the statement of the "quadratic-Hamiltonian theorem" is clearly a coordinate-dependent one.

The properties of canonoid transformations have recently been investigated[5,6] using the tools of modern differential geometry. Thus some interesting properties were obtained in an intrinsic way (that is, without reference to a concrete type of coordinates) and particularly the existence of "generating functions" was proved (in a similar way to the better known canonical case). Another approach is due to Marmo *et al.*,[3] who presented a geometrical study of this matter mainly based on symplectic actions of Lie algebras.[7]

In this paper we first aim to study some aspects of the theory of "generating functions" of canonoid transforma-

0022-2488/90/040801-07$03.00

tions and then to apply it to deal with the problem of relating canonical with canonoid transformations. These results will permit us to present a new theorem which will represent an intrinsic and coordinate-free generalization of the "quadratic-Hamiltonian theorem."

The paper is organized as follows. In Sec. II we study the theory of bi-Hamiltonian systems and its relation with the theory of canonoid transformations from a geometric perspective. It has been proved that canonoid transformations imply the existence of "generating functions" and conversely, that it is possible to generate canonoid transformations starting from these functions. Section III is mainly devoted to the study of properties of these "generating functions." In Sec. IV a theorem concerning the relations between canonical and canonoid transformations is proved: As stated above, it represents a generalization of the "quadratic-Hamiltonian theorem." Finally, in Sec. V we will use coordinate expressions and consider the particular case of the configuration space Q being $\mathbb{R}^n$. In this way we will obtain a new version of the "quadratic-Hamiltonian theorem."

II. BI-HAMILTONIAN SYSTEMS

Let (M,ω) be a symplectic manifold of dimension $\dim M = 2n$. We will denote by $\mathfrak{X}(M)$ the set of all the C^∞ vector fields defined in M and by $\mathfrak{X}_H(M,\omega)$ and $\mathfrak{X}_{LH}(M,\omega)$ the set of all those vector fields that are Hamiltonian and locally Hamiltonian, respectively, that is $\Gamma\in\mathfrak{X}_H(M,\omega)$ means that $i(\Gamma)\omega$ is an exact one-form, while $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$ means that $i(\Gamma)\omega$ is closed, i.e., $\mathscr{L}_\Gamma\omega = 0$, where $\mathscr{L}_\Gamma$ is the Lie derivative with respect to the vector field Γ.

A diffeomorphism Φ on M is called symplectic if it preserves the two-form ω, i.e., $\Phi^*(\omega) = \omega$. We remark that the symplectic character of a particular Φ depends just on its relation with the symplectic form ω and does not involve any other mathematical object. Contrary to this, there is also a related concept in the formalism of the symplectic geometry which has to do not only with ω, but also with a given vector field.[5] Thus given a locally Hamiltonian vector field $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$, a transformation $\Phi\in\mathrm{Diff}(M)$ is said to be canonoid with respect to Γ if the transformed field $\Phi_*\Gamma$ is also locally Hamiltonian (that is, $\mathscr{L}_{\Phi_*\Gamma}\omega = 0$). In the language of traditional classical mechanics a general canonoid transformation "preserves the canonical formalism not for all the Hamiltonians but for only some Hamiltonians or perhaps only one."[8] In a similar way, in this symplectic setting if Φ is canonoid with respect to Γ, probably this property will not hold for another vector field Γ' [that is, $\Phi_*\Gamma'\notin\mathfrak{X}_{LH}(M,\omega)$].

Given two different symplectic structures ω_1 and ω_2 in the same manifold M, vector fields that are simultaneously locally Hamiltonian with respect to both ω_1 and ω_2 are called bi-Hamiltonian dynamical systems. We will denote by $\mathfrak{X}_{LH}(M;\omega_1,\omega_2)$ the set of such vector fields; if there is no danger of confusion we write $\mathfrak{X}_{LH}(\omega_1,\omega_2)$.

If a nonsymplectic transformation Φ is a canonoid transformation for Γ, then $\mathscr{L}_{\Phi_*\Gamma}\omega = 0$ and consequently, also $\mathscr{L}_\Gamma\Phi^*\omega = 0$. Therefore, this vector field will be locally Hamiltonian with respect to two different symplectic structures: the primitive ω and the new and different $\Phi^*\omega$.

Thus if Φ is canonoid with respect to Γ, then $\Gamma\in\mathfrak{X}_{LH}(\omega,\Phi^*\omega)$, but we remark that the converse is also true. Indeed, let Γ be such that $\Gamma\in\mathfrak{X}_{LH}(\omega,\Phi^*\omega)$; then $\mathscr{L}_\Gamma\Phi^*\omega = 0$ and therefore, $\Phi^*(\mathscr{L}_{\Phi_*\Gamma}\omega) = 0$. Hence, Φ is a canonoid transformation with respect to Γ. Consequently, a diffeomorphism $\Phi\in\mathrm{Diff}(M)$ will be a canonoid transformation with respect to any vector field $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$ such that $\Gamma\in\mathfrak{X}_{LH}(\omega,\Phi^*w)$.

Finally, from

$$\mathscr{L}_{[X,Y]}\omega_i = \mathscr{L}_X(\mathscr{L}_Y\omega_i) - \mathscr{L}_Y(\mathscr{L}_X\omega_i), \quad i = 1,2,$$

we see that the set $\mathfrak{X}_{LH}(M;\omega_1,\omega_2)$ is a Lie algebra.

III. GENERATING FUNCTIONS OF CANONOID TRANSFORMATIONS

Let (M,ω) be a symplectic manifold. We recall that this manifold is called an exact symplectic manifold if ω is exact, i.e., there exists a one-form θ such that the two-form ω is given by $\omega = -d\theta$ (for instance, when M is the cotangent bundle T^*Q of a manifold Q). In this case, a diffeomorphism Φ of M, $\Phi\in\mathrm{Diff}(M)$ is symplectic if and only if (see Ref. 4) $\theta - \Phi^*(\theta)$ is closed. This means the existence of a locally defined function F such that

$$\theta - \Phi^*(\theta) = dF.$$

This function F is called a generating function for the symplectic transformation Φ. This property will also be true for a general symplectic manifold, where the role of θ is played by a local one-form for ω (if M is T^*Q, then θ is the Liouville one-form θ_0, but in the more general case the function F will depend on the choice of the local one-form for ω). We now make a final introductory point concerning this function. Indeed, the generating function is not defined in the manifold M, but on the graph G_Φ of Φ, $G_\Phi \subset M\times M$ [in the more general case of $\Phi:(M_1\omega_1)\to(M_2,\omega_2)$; then $G_\Phi\subset M_1\times M_2$] because G_Φ is a Lagrangian submanifold of $M\times M$. However, since G_Φ is diffeomorphic to M, we will consider the associated F defined in M.

It was proved in Ref. 5 that the canonoid character of a transformation Φ with respect to a vector field Γ is also related to the existence of an associated function. Now what must be a closed form is not the difference between the one-form θ and its pullback $\Phi^*\theta$, but the difference between the corresponding Lie derivatives with respect to Γ. Thus $\mathscr{L}_\Gamma\theta - \mathscr{L}_\Gamma\Phi^*\theta$ being closed is equivalent to Φ being canonoid with respect to Γ. Locally, by the Poincaré lemma, there exists a function S such that

$$dS = \mathscr{L}_\Gamma\theta - \mathscr{L}_\Gamma\Phi^*\theta.$$

We call the function S a generating function for the canonoid diffeomorphism Φ and remark that in this case this function S depends not only on the transformation Φ, but also on the vector field Γ.

Every symplectic transformation uniquely determines, up to an additive constant and via the one-form θ, a function F. The converse is not true because two different symplectic transformations may yield the same function F. That is, if Φ and Ψ are symplectomorphisms such that $\Phi^*\theta = \Psi^*\theta$, then

$F^\Phi = F^\Psi$ (up to an additive constant). This fact can also be seen in Darboux coordinates $\{q^i,p_i,i=1,\dots,n\}$ since then, given a function F, the solution $Q^i=Q^i(q^k,p^k)$, $P_i = P_i(q^k,p^k)$, where $i,k=1,\dots,n$, of $p_i\,dq^i - P_i\,dQ^i = dF$ may be nonunique.

Let λ denote the map

$$\lambda:\mathrm{Diff}(M)\times\mathfrak{X}_{LH}(M,\omega)\to\Lambda^1(M)$$

defined by

$$\lambda(\Phi,\Gamma)=\mathscr{L}_\Gamma\theta-\mathscr{L}_\Gamma\Phi^*(\theta).$$

If Φ is canonoid with respect to Γ, then $\lambda(\Phi,\Gamma)$ is in the subset of closed forms, $\lambda(\Phi,\Gamma)\in Z^1(M)\subset\Lambda^1(M)$. Therefore, if we denote by λ_Φ the map $\lambda_\Phi:\mathfrak{X}_{LH}(M,\omega)\to\Lambda^1(M)$ defined by

$$\lambda_\Phi(\Gamma)=\lambda(\Phi,\Gamma),$$

we have $\mathfrak{X}_{LH}(\omega,\Phi^*\omega)=\lambda_\Phi^{-1}[Z^1(M)]$. Also, if we denote by λ_Γ the map λ_Γ: $\mathrm{Diff}(M)\to\Lambda^1(M)$ defined by $\lambda_\Gamma(\Phi) = \lambda(\Phi,\Gamma)$, we have that $\lambda_\Gamma^{-1}[Z^1(M)]$ represents the set of all the diffeomorphisms which are canonoid with respect to Γ. Notice that $\mathrm{Sp}(M,\omega)\subset\lambda_\Gamma^{-1}[Z^1(M)]$, where $\mathrm{Sp}(M,\omega)$ denotes the group of symplectic transformations.

If $\lambda(\Phi,\Gamma)\in Z^1(M)$, then for any point $p\in M$ there exists a connected coordinate neighborhood such that the restriction $\lambda(\Phi,\Gamma)_{|U}$ is exact, $\lambda(\Phi,\Gamma)_{|U}\in B^1(U)$. Since $B^1(U)$ is diffeomorphic to $C^\infty(U)/\mathbb{R}$, we see that if Φ is canonoid with respect to the vector field Γ, the Φ and Γ uniquely determine the function S up to an additive constant.

Concerning the inverse approach, notice that it is possible to have the existence of more than one element in $\lambda_\Gamma^{-1}(dS)$: This will mean that the function S will generate several canonoid transformations with respect to Γ. This situation is more complicated than in the symplectic case because besides the previous case of $\Phi\neq\Psi$, but $\Phi^*(\theta)=\Psi^*(\theta)$, we also have the possibility $\Phi^*(\theta)\neq\Psi^*(\theta)$, but $\mathscr{L}_\Gamma\Phi^*(\theta)=\mathscr{L}_\Gamma\Psi^*(\theta)$. In local coordinates, given a function S, the transformation $Q^i = Q^i(q,p)$, $P_i=P_i(q,p)$ generated by S with respect to Γ must be obtained by solving second-order differential equations. In some cases the system of equations can admit more than one solution.

Let us now suppose that the diffeomorphism Φ is a transformation which is canonoid not just for one vector field Γ, but for a family $\{\Gamma_i, i=1,\dots,k\}$ of k locally Hamiltonian vector fields. Then (i) the different $\Gamma_i, i=1,\dots,k$, satisfy

$$\Gamma_i\in\mathfrak{X}_{LH}(M;\omega,\Phi^*\omega)$$

and (ii) there will be k different functions $S_i, i=1,\dots,k$, such that

$$\mathscr{L}_{\Gamma_i}\theta-\mathscr{L}_{\Gamma_i}\Phi^*(\theta)=dS_i.$$

Notice that property (i) is a direct consequence of $\mathscr{L}_{\Phi_*\Gamma_i}\omega=0$ and property (ii) emphasizes the dependence on the vector fields of the generating functions. That is, there are as many functions S_i as vector fields Γ_i. We remark that these functions are different, but all are generating functions for the same diffeomorphism.

We are interested in studying not only the relation between every function S_i with its associated vector field Γ_i, but also the relation between the different S_i's.

We remarked that there is a different generating function of Φ for every vector field with respect to which Φ is a canonoid transformation. Thus we can guarantee the existence of an S^Γ for every vector field Γ in $\mathfrak{X}_{LH}(\omega,\Phi^*\omega)$.

The two extreme cases are when $\mathfrak{X}_{LH}(\omega,\Phi^*\omega)$ reduces to the null vector field, which corresponds to a noncanonoid transformation for any dynamics, and when $\mathfrak{X}_{LH}(\omega,\Phi^*\omega)$ is the full set $\mathfrak{X}_{LH}(M,\omega)$, corresponding to $\Phi^*(\omega)=c\omega$, where c denotes a constant real number.[9,10] In this last case we will distinguish the two following situations.

(i) Here Φ is itself a symplectomorphism and thus $\Phi^*(\omega)=\omega$. In the more traditional language of mechanics Φ is then called a restricted or univalent canonical or even, if there is no danger of confusion, just a canonical transformation.

(ii) The more general case of Φ is such that $\Phi^*(\omega)=c\omega$. In the mechanical language Φ is usually called a generalized or extended canonical transformation and c is said to be the valence of the transformation. See, e.g., Ref. 11, where some applications to the virial theorem and Toda lattice are given.

The two following propositions study these two situations.

Proposition 3.1: Let (M,ω) be a symplectic manifold and Φ be a diffeomorphism. Then Φ is a symplectomorphism if and only if Φ is a canonoid transformation with respect to every locally Hamiltonian vector field Γ and there is a Γ-independent function F such that $S^\Gamma=\Gamma(F)$.

Proof: Let Φ be a symplectomorphism. Then there is a function F such that $\theta-\Phi^*\theta=dF$. Thus for any $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$ we have

$$\mathscr{L}_\Gamma\theta-\mathscr{L}_\Gamma\Phi^*(\theta)=d\,[\Gamma(F)].$$

Hence, Φ is canonoid with respect to Γ and the associated function S^Γ is $S^\Gamma=\Gamma(F)$. Conversely, let Φ be canonoid for every $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$. Then $\mathscr{L}_\Gamma\theta-\mathscr{L}_\Gamma\Phi^*(\theta)=dS^\Gamma$. Thus if every S^Γ is of the form $S^\Gamma=\Gamma(F)$, we obtain $\mathscr{L}_\Gamma\theta-\mathscr{L}_\Gamma\Phi^*(\theta)=\mathscr{L}_\Gamma(dF)$ for any $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$. Since a local basis of $\mathfrak{X}(M)$ can be built from locally Hamiltonian vector fields, then $\theta-\Phi^*\theta=dF$ and Φ will be a symplectomorphism.

Proposition 3.2: Let (M,ω) be a symplectic manifold and Φ a diffeomorphism such that $\Phi^*\omega=c\omega$, where c is a nonzero constant. Then (i) Φ is a canonoid transformation of M with respect to every locally Hamiltonian vector field Γ and (ii) there is an Γ-independent function F such that the generating function S^Γ associated to every vector field Γ can be written as $S^\Gamma=c\Gamma(F)+(1-c)\{i(\Gamma)\theta-H\}$, with H a (perhaps only locally defined) Hamiltonian for Γ.

Proof: (i) Assume that $\Phi^*\omega=c\omega$; then if $\mathscr{L}_\Gamma\omega=0$ we have

$$\mathscr{L}_\Gamma\omega=(1/c)\mathscr{L}_\Gamma\Phi^*\omega=(1/c)\Phi^*(\mathscr{L}_{\Phi_*\Gamma}\omega)=0$$

and therefore, $\Phi_*\Gamma\in\mathfrak{X}_{LH}(M,\omega)$. Consequently, Φ will be a canonoid transformation for any $\Gamma\in\mathfrak{X}_{LH}(M,\omega)$.

(ii) Notice that

$$\omega - (1/c)\Phi^*(\omega) = -d[\theta - (1/c)\Phi^*(\theta)].$$

Therefore, Φ is an extended canonical transformation of (M,ω) with valence c, if and only if $\theta - (1/c)\Phi^*(\theta)$ is closed. Thus this means the local existence of a function F such that

$$\theta - (1/c)\Phi^*(\theta) = dF.$$

Using this function F we obtain

$$\begin{aligned}\mathscr{L}_\Gamma\theta - \mathscr{L}_\Gamma\Phi^*(\theta) &= \mathscr{L}_\Gamma\theta - \mathscr{L}_\Gamma c\theta + c\mathscr{L}_\Gamma\, dF\\ &= c\mathscr{L}_\Gamma\, dF + (1-c)d\{i(\Gamma)\theta - H\}\\ &= d\{c\Gamma(F) + (1-c)(i(\Gamma)\theta - H)\},\end{aligned}$$

where H is a local Hamiltonian for Γ, i.e., $i(\Gamma)d\theta = -dH$. Hence, we find that the function S^Γ is

$$S^\Gamma = c\Gamma(F) + (1-c)\{i(\Gamma)\theta - H\}.$$

Finally, even if the function H is only locally defined, this fact does not represent any difficulty because so is S^Γ. Notice, also, that when $c=1$ we recover the result of Proposition 3.2.

Propositions 3.1 and 3.2 can be used for proving when a given transformation that is presented at first as a canonoid is in fact a (restricted or extended) canonical transformation. Thus in this new approach what we really will analyze is not the transformation Φ itself, but its associated set of generating functions.

We have previously stated that the set $\mathfrak{X}_{LH}(M;\omega_1,\omega_2)$ is a Lie algebra. The following proposition studies the relation between the Lie algebra structure of the set of vector fields with respect to which Φ is canonoid and the corresponding set of associated functions.

Proposition 3.3: Let Φ be a diffeomorphism of (M,ω) which is canonoid with respect to two different locally Hamiltonian vector fields Γ_i, $i=1,2$ with the associated generating functions $S_i, i=1,2$. Then the generating function S_{12} of Φ with respect to the field $[\Gamma_1,\Gamma_2]$ is $S_{12} = \Gamma_1(S_2) - \Gamma_2(S_1)$.

Proof: We have

$$\begin{aligned}&\mathscr{L}_{[\Gamma_1,\Gamma_2]}\{\theta - \Phi^*(\theta)\}\\ &\quad= \mathscr{L}_{\Gamma_1}\{\mathscr{L}_{\Gamma_2}(\theta - \Phi^*\theta)\} + \mathscr{L}_{\Gamma_2}\{\mathscr{L}_{\Gamma_1}(\theta - \Phi^*\theta)\}\\ &\quad= \mathscr{L}_{\Gamma_1}(dS_2) - \mathscr{L}_{\Gamma_2}(dS_1) = d\,[\Gamma_1(S_2) - \Gamma_2(S_1)]\end{aligned}$$

and the proposition is proved.

Next, we give a proposition which makes use of the expression obtained for S_{12}, together with the properties of a particular Lie algebra structure.

Proposition 3.4: Let Φ be a diffeomorphism canonoid with respect to a set of locally Hamiltonian vector fields $\{\Gamma_\alpha, \alpha = 1,\ldots,2n\}$, which constitute an effective realization of the $2n$-dimensional Abelian Lie algebra. Then (i) there is a Γ_α-independent function G such that all the $2n$ generating functions S_α depend on G and (ii) the expression of the transformed form $\Phi^*(\omega)$ is constant in Darboux coordinates for ω.

Proof: (i) If $\{\Gamma_\alpha, \alpha = 1,\ldots,2n\}$ is an effective realization of the $2n$-dimensional Abelian Lie algebra, we have

$$[\Gamma_\alpha,\Gamma_\beta] = 0, \quad \alpha,\beta = 1,\ldots,2n;$$

consequently, the generating functions $S_{\alpha\beta}$ that according to Proposition 3.3 are given by $S_{\alpha\beta} = \Gamma_\alpha(S_\beta) - \Gamma_\beta(S_\alpha)$ must be constant. We know that if $[\Gamma_\alpha,\Gamma_\beta] = 0$ there exist local coordinates (ξ^α) such that the vector fields Γ_α take the form[12] $\Gamma_\alpha = \partial/\partial\xi^\alpha$. Therefore, in these coordinates we will have

$$\frac{\partial S_\beta}{\partial\xi^\alpha} - \frac{\partial S_\alpha}{\partial\xi^\beta} = c_{\alpha\beta},$$

where $c_{\alpha\beta}$ are real constants such that $c_{\alpha\beta} = -c_{\beta\alpha}$. Let $\zeta(S)$ be defined by $\zeta(S) = S_\beta\, d\xi^\beta$ and let $\zeta'(S)$ denote the one-form $\zeta'(S) = \zeta(S) - \frac{1}{2}c_{\mu\beta}\xi^\mu\, d\xi^\beta$. The meaning of the above system of equations is that $d\zeta'(S) = 0$; this is equivalent to the existence of a function $G = G(\xi^\beta)$ such that $\zeta'(S) = dG$. Hence,

$$S_\beta = \frac{\partial G}{\partial\xi^\beta} + \frac{1}{2}c_{\mu\beta}\xi^\mu.$$

(ii) Let $t_{\alpha\beta}$ denote the components of $\Phi^*(\omega)$ with respect to the system of coordinates $\{\xi^\alpha, \alpha = 1,\ldots,2n\}$, namely,

$$\Phi^*(\omega) = t_{\alpha\beta}\, d\xi^\alpha \wedge d\xi^\beta.$$

Then since Φ is canonoid with respect to the $\Gamma_\alpha = \partial/\partial\xi^\alpha$, we have

$$\mathscr{L}_{\Gamma_\mu}(t_{\alpha\beta}\, d\xi^\alpha \wedge d\xi^\beta) = 0;$$

this means that $\Gamma_\mu(t_{\alpha\beta}) = 0$, $\mu,\alpha,\beta = 1,\ldots,2n$. Hence, the components $t_{\alpha\beta}$ are constant. Finally, Darboux coordinates $\{q^i,p_i, i = 1,\ldots,n\}$, for ω can be obtained from linear combinations (with constant coefficients) of the ξ^α's. Hence, the transformed two-form $\Phi^*(\omega)$ will also be constant when expressed in the $\{q^i,p_i, i = 1,\ldots,n\}$ coordinate system.

If Φ is such that $\Phi^*(\omega) = \omega$, then the Poisson brackets are preserved; if $\Phi^*(\omega) = c\omega$, then they are preserved up to a multiplicative constant. Since the matrix elements of the new form $\Phi^*(\omega)$ are the Lagrange brackets $[q^i,q^k]$, $[p_i,p_k]$, and $[q^i,p_k]$, we see that if Φ satisfies the hypothesis of Proposition 3.4, then the new fundamental Poisson brackets (that is, $\{Q^i,Q^k\}$, $\{P_i,P_k\}$, and $\{Q^i,P_k\}$) are constant.

These properties show that there is a direct relation between the structure and size of the subset of $\mathfrak{X}_{LH}(M,\omega)$ with respect to which Φ is canonoid, the form in which the associated function S^Γ depends on Γ, and the "modification" or "lack of symplecticity" that Φ produces in the symplectic structure.

A particularly interesting case of a transformation that is canonoid for two different vector fields will be when Φ is canonoid with respect to two $C^\infty(M)$-proportional locally Hamiltonian vector fields Γ_1 and $\Gamma_2 = f\Gamma_1, f \in C^\infty(M)$. Notice first that in order for the two vector fields to be locally Hamiltonian it is required that $df \wedge i(\Gamma_1)\omega = 0$, as the relation

$$\mathscr{L}_{\Gamma_2}\omega = f\mathscr{L}_{\Gamma_1}\omega + df \wedge i(\Gamma_1)\omega$$

shows. We remark that in the particular case of Γ_1 being a Hamiltonian field, then there will be a function H such that $i(\Gamma_1)\omega = dH$; the above equation becomes $df \wedge dH = 0$ and implies that the function f must be of the form $f = f(H)$. If Γ_1 is only locally Hamiltonian the assertion is only locally true.

Proposition 3.5: Let $\Phi \in \mathrm{Diff}(M)$ be a canonoid transformation with respect to the two locally Hamiltonian vector fields $\Gamma_1 = \Gamma$ and $\Gamma_2 = f\Gamma_1, f \in C^\infty(M)$. Then the generating function S_2 associated to Γ_2 takes the form $S_2 = fS_1 + u(f)$, where S_1 denotes the generating function of Φ associated to Γ_1 and $u(f)$ is a function of f.

Proof: If Φ is canonoid with respect to Γ_2 we have

$$\mathscr{L}_{\Gamma_2}\{\theta - \Phi^*\theta\} = dS_2.$$

On the other hand,

$$\mathscr{L}_{\Gamma_2}\{\theta - \Phi^*\theta\} = f\mathscr{L}_\Gamma\{\theta - \Phi^*\theta\} + \{i(\Gamma)(\theta - \Phi^*\theta)\}df$$
$$= f\,dS_1 + \{i(\Gamma)(\theta - \Phi^*\theta)\}df$$

since $\Gamma_2 = f\Gamma$. Thus combining the two expressions we find that $dS_2 = d(fS_1) + [i(\Gamma)(\theta - \Phi^*\theta) - S_1]df$. Therefore, the last term must be an exact form and then the expression between the square brackets must be a function of f. Thus we find that S_2 is related to S_1 by $S_2 = fS_1 + u$, where $u = u(f)$ is a function of f such that its derivative $u'(f)$ takes the value $u'(f) = i(\Gamma)(\theta - \Phi^*\theta) - S_1$.

Notice that in Proposition 3.5 we have not only related S_2 with S_1, but also obtained an expression for S_1 as a function of Γ, θ, and f. Nevertheless, this expression involves an undetermined function u.

IV. SYMPLECTOMORPHISMS VERSUS CANONOID TRANSFORMATIONS

We have previously proved that if Φ is such that $\Phi^*\omega = c\omega$, then the S^Γ associated to every Γ takes the form $S^\Gamma = c\Gamma(F) + (1-c)\{i(\Gamma)\theta - H\}$. The following proposition gives a new relation between Φ, Γ, and F which will be used for obtaining a new expression for S^Γ.

Proposition 4.1: Let Φ be such that $\Phi^*\omega = c\omega$. Then

$$i(\Gamma)\{\theta - \Phi^*\theta\} = c\Gamma(F) + (1-c)\{i(\Gamma)\theta\}.$$

Proof: If $\Phi^*\omega = c\omega$, then $\theta - (1/c)\Phi^*\theta = dF$ and thus

$$i(\Gamma)\{\theta - \Phi^*\theta\} = i(\Gamma)\{\theta - c[\theta - dF]\}$$
$$= c\Gamma(F) + (1-c)\{i(\Gamma)\theta\}.$$

Thus using this equality, the function S^Γ given by Proposition 3.2 can be alternatively written as follows:

$$S^\Gamma = i(\Gamma)\{\theta - \Phi^*\theta\} + (c-1)H.$$

This new expression of S^Γ will be of great use later on.

It is known[9,10] that if Φ is canonoid with respect to every locally Hamiltonian vector field (that is, $\Phi_*[\mathfrak{X}_{LH}(M,\omega)] \subset \mathfrak{X}_{LH}(M,\omega)$), then $\Phi^*(\omega) = c\omega$. The following theorem proves that there exist inside $\mathfrak{X}_{LH}(M,\omega)$ some smaller subsets such that in order to guarantee the property $\Phi^*(\omega) = c\omega$ it will be sufficient for Φ to be canonoid only with respect to one of those subsets.

Usually, canonoid transformations are considered in the particular case of transformations of a symplectic manifold (M,ω), but as when studying canonical transformations, we first consider the maps $\Phi:(M_1,\omega_1) \to (M_2,\omega_2)$ relating ω_1 to ω_2, $\Phi^*(\omega_2) = \omega_1$ and then obtain symplectomorphisms of (M,ω) as a particular case. Here we will also consider maps relating two different symplectic manifolds.

Theorem 4.2: Let (M_1,ω_1) and (M_2,ω) be $2n$-dimensional symplectic manifolds and $\{X_1, X_2, \ldots, X_N\}$, $N \geqslant 2n$ a set of globally defined vector fields in M_1 which are locally ω_1 Hamiltonian and span $\mathfrak{X}(M_1)$. Then a diffeomorphism $\Phi \in \mathrm{Diff}(M_1,M_2)$ is a generalized canonical transformation if and only if the following properties hold.

(i) Here Φ is canonoid with respect to every one of the fields X_k, $k = 1,\ldots,N$.

(ii) There exist N nonconstant functions $f_k \in C^\infty(M_1)$ such that Φ is canonoid with respect to the new set of locally Hamiltonian fields $Y_k = f_k X_k$.

(iii) When $i(X_j)i(X_k)\omega_1 = 0$ there is a locally Hamiltonian vector field $X_{jk} = z_j X_k + z_k X_j$ with $z_j, z_k \in C^\infty(M_1)$ for which Φ is canonoid.

Proof: First notice that the number of globally defined vector fields required to span $\mathfrak{X}(M_1)$ must be equal to or greater than the dimension of M_1 since $2n$ vector fields that are linearly independent and span $\mathfrak{X}(M_1)$ do not exist for a general M_1. Nevertheless, since generating functions are locally defined and coordinate neighborhoods are parallelizable, in every symplectic chart for ω_1 there will be $2n$ vector fields X_a, $a = 1,\ldots,2n$ in the set $\{X_k\}$, $k = 1,\ldots,N$ such that for every point m in such a chart the $2n$ vectors $X_a(m)$ will form a basis of $T_m(M_1)$.

As far as (iii) of Theorem 4.2 is concerned, let us remark that if X_j and X_k are locally Hamiltonian, then besides the trivial case of the linear combinations $a_j X_j + a_k X_k$ with the coefficients a_j, a_k only depending on the respective local Hamiltonians H_j, H_k, there will exist other fields $X_{jk} = z_k X_j + z_j X_k \in \mathfrak{X}_{LH}(M,\omega_1)$ with $z_i \in C^\infty(M_1)$, for instance, $H_k X_j + H_j X_k$. Therefore, property (iii) assumes the existence of a nontrivial linear combination $X_{jk} = z_j X_k + z_k X_j \in \mathfrak{X}_{LH}(M,\omega_1)$ for which Φ is canonoid.

(i) That $\Phi^*\omega_2 = c\omega_1$ implies properties (i), (ii), and (iii) of Theorem 4.2 follows immediately since these diffeomorphisms are precisely characterized by $\Phi_*(\mathfrak{X}_{LH}(M_1\omega_1)) \subset \mathfrak{X}_{LH}(M_2\omega_2)$ and therefore since they are canonoid for any locally Hamiltonian field so will they be for the X_k, Y_k, and X_{jk}.

(ii) Assume now that Φ is a canonoid transformation simultaneously for one of the X_k and its proportional vector field $Y_k = f_k X_k$. Then if S_k denotes the generating function of Φ associated to X_k we have

$$\mathscr{L}_{X_k}\{\theta_1 - \Phi^*\theta_2\} = dS_k,$$

but since Φ is also canonoid with respect to Y_k, the function S_k takes the form

$$S_k = i(X_k)\{\theta_1 - \Phi^*\theta_2\} - u_k',$$

with u_k' an undetermined function of f_k. Thus we obtain

$$i(X_k)\{d(\theta_1 - \Phi^*\theta_2)\} + du_k' = 0$$

and therefore, u_k' satisfies

$$i(X_k)\omega_1 = i(X_k)\Phi^*\omega_2 + du_k'.$$

The meaning of the function u_k' is now clear; its differential represents what we can call the "lack of symplecticity" of Φ along the integral curves of X_k.

Relating this expression with that obtained for X_j it follows that

$$i(X_j)du'_k + i(X_k)du'_j = 0.$$

Moreover, since $Y_k \in \mathfrak{X}_{LH}(M,\omega_1)$, the function f_k satisfies $df_k \wedge i(X_k)\omega_1 = 0$. Also, according to Proposition (3.5), $du'_k \wedge df_k = 0$. Therefore, there will be a function m_k such that $du'_k = m_k\{i(X_k)\omega_1\}$. Using this, we obtain

$$(m_k - m_j)\{i(X_j)i(X_k)\omega_1\} = 0.$$

Since ω_1 is nondegenerate and we can choose a local basis from the total set $\{X_k\}$ in every local chart, for every index j there will be at least an index k such that $i(X_j)i(X_k)\omega_1 \neq 0$. Consequently, the functions m_k and m_j associated to such vector fields satisfy $m_k = m_j$.

If $i(X_j)i(X_k)\omega_1 = 0$, then according to property (iii) of Theorem 4.2 Φ is also canonoid with respect to the vector field X_{jk} and consequently, $\mathscr{L}_{X_{jk}}\{\theta_1 - \Phi^*(\theta_2)\}$ as given by

$$\begin{aligned}\mathscr{L}_{X_{jk}}&\{\theta_1 - \Phi^*\theta_2\}\\ &= z_j\, dS_k + z_k\, dS_j + [i(X_k)\{\theta_1 - \Phi^*\omega_2\}]dz_j\\ &\quad + [i(X_j)\{\theta_1 - \Phi^*\theta_2\}]dz_k\\ &= d\,[z_j i(X_k)\{\theta_1 - \Phi^*\theta_2\} + z_k i(X_j)\{\theta_1 - \Phi^*\theta_2\}]\\ &\quad - z_j\, du'_k - z_k\, du'_j\end{aligned}$$

must be closed. If for du'_j and du'_k we use the expressions obtained above, differentiation gives

$$m_k\, dz_j \wedge i(X_k)\omega_1 + m_j\, dz_k \wedge i(X_j)\omega_1 = 0.$$

However, if X_{jk} is locally Hamiltonian, then the functions z_j and z_k must satisfy

$$dz_j \wedge i(X_k)\omega_1 + dz_k \wedge i(X_j)\omega_1 = 0$$

and therefore, the condition of Φ canonoid with respect to X_{jk} becomes

$$(m_k - m_j)\{dz_j \wedge i(X_k)\omega_1\} = 0.$$

Consequently, we find that even when $i(X_j)i(X_k)\omega_1 = 0$ the coefficients m_k and m_j must satisfy

$$m_k = m_j = m.$$

Notice that the functions m_k were introduced by $du'_k = m_k\{i(X_k)\omega_1\}$ and then $dm_k \wedge \{i(X_k)\omega_1\} = 0$. This means that in any local chart every m_k must be a function of the corresponding local Hamiltonian for X_k. Consequently, the property $m_k = m_j = m$ for $k,j = 1,\ldots,N$ is only possible if the common value m is a constant function.

Thus we finally obtain

$$i(X_k)\Phi^*\omega_2 = c\{i(X_k)\omega_1\}$$

and

$$S_k = i(X_k)\{\theta_1 - \Phi^*\theta_2\} + (c-1)H_k,$$

with H_k locally defined by $dH_k = i(X_k)\omega_1$ and the constant c defined by $c = 1 - m$.

In short, we have proved that a transformation is generalized canonical if and only if it is canonoid with respect to the set $\{X_k, Y_k, X_{jk}\}$. We remark that the minimum number N of vector fields X_k necessary for Theorem 4.2 depends on the topological characteristics of M_1. Usually M_1 is the cotangent bundle T^*Q of an n-dimensional manifold Q. In such a case $N = 2N_Q$, where N_Q denotes the minimum number of globally defined vector fields in Q necessary for the span $\mathfrak{X}(Q)$. Only if Q is parallelizable do we have $N_Q = n$. Obviously, the simplest case will be when it is possible to obtain a set of commuting X_k, but in any case, Theorem 4.2 does not impose any restriction concerning the algebraic structure of the X_k.

V. A PARTICULAR CASE: THE QUADRATIC-HAMILTONIAN THEOREM

The traditional approach to mechanical systems corresponds to topologically trivial phase spaces. In these cases M is $\mathbb{R}^n \times \mathbb{R}^n$, locally Hamiltonian vector fields become globally Hamiltonian, and the $2n$ coordinates are also globally defined. Consequently, in these cases, if Φ is canonoid with respect to a Hamiltonian H, then the function S will be globally defined by equations of the form

$$\frac{\partial^2 H}{\partial q^J \partial p_k} p_k - \frac{\partial H}{\partial q^j} + P_k \frac{\partial}{\partial q^j}\{H,Q^k\} + \{H,P_k\}\frac{\partial Q^k}{\partial q^j} = \frac{\partial S}{\partial q^j},$$

$$\frac{\partial^2 H}{\partial p_j \partial p_k} p_k + P_k \frac{\partial}{\partial p_j}\{H,Q^k\} + \{H,P_k\}\frac{\partial Q^k}{\partial p_j} = \frac{\partial S}{\partial p_j}.$$

Theorem 5.1: Let $F_i = F_i(q^i)$, $G_i = G_i(p_i)$, and $K_i = K_i(q^i q^{i+1})$ be $3n$ arbitrary functions with the only restriction that $F_i = F_i(q^i)$, $G_i = G_i(p_i)$ have nonvanishing second derivatives. If Φ is a diffeomorphism of the phase space $\mathbb{R}^n \times \mathbb{R}^n$ that is canonoid with respect to all the Hamiltonians of the form

$$H = a_i q^i + b_i p_i + c_i F_i(q^i) + d_i G_i(p_i) + e_i K_i(q^i q^{i+1}),$$

where a_i, b_i, c_i, d_i, and e_i, $i = 1,\ldots,n$ are constant, then Φ is an extended canonical transformation.

Proof: To begin, first consider the case of Φ canonoid with respect to the family of Hamiltonians $H_{1i} = q^i$, $i = 1,\ldots,n$. Then if we denote by S_{1i} the associate family of generating functions, we have

$$\frac{\partial S_{1i}}{\partial q^j} = \frac{\partial}{\partial p_i}\left\{P_k \frac{\partial Q^k}{\partial q^j} - p_j\right\}, \quad \frac{\partial S_{1i}}{\partial p_j} = \frac{\partial}{\partial p_i}\left\{P_k \frac{\partial Q^k}{\partial p_j}\right\}.$$

Assume now that Φ is also canonoid with respect to the family $H_{2i} = F_i(q^i)$ (the notation means that every one of the F_i is a function of only the corresponding coordinate q^i) and denote by S_{2i}, $i = 1,\ldots,n$ its associate set of generating functions: Then their derivatives take the values

$$\frac{\partial S_{2i}}{\partial q^j} = \delta_{ij}\, F''_i P_k \frac{\partial Q^k}{\partial p_j} + F'_i\left\{P_k \frac{\partial^2 Q^k}{\partial q^j \partial p_i} - \delta_{ij} + \frac{\partial P_k}{\partial p_i}\frac{\partial Q^k}{\partial q^j}\right\}$$

$$\frac{\partial S_{2i}}{\partial p_j} = F'_i \frac{\partial^2 Q^k}{\partial p_j \partial p_i} P_k + F'_i \frac{\partial P_k}{\partial p_i}\frac{\partial Q^k}{\partial p_j},$$

which can be rewritten in the form

$$\frac{\partial S_{2i}}{\partial q^j} = F'_i \frac{\partial S_{1i}}{\partial q^j} + \delta_{ij}\, F''_i P_k \frac{\partial Q^k}{\partial p_j}, \quad \frac{\partial S_{2i}}{\partial p_j} = F'_i \frac{\partial S_{1i}}{\partial p_j};$$

from here we find that $F''_i S_{1i} + u'_i = F''_i P_k(\partial Q^k/\partial p_i)$, with the $u_i = u_i(q^i)$ such that $S_{2i} = q^i S_{1i} + u_i$.

We have obtained an expression for S_{1i}; therefore, we can obtain new values for $\partial S_{1i}/\partial q^j$ and $\partial S_{1i}/\partial p_j$ and relate them with those previously known. If we do so, we arrive at

$$[p_j, p_i] = 0$$

and

$$[q^j,p_i] = \delta^j_i\left\{1-\frac{\partial}{\partial q^j}\left(\frac{u'_i}{F''_i}\right)\right\}, \quad i,j=1,...,n,$$

where [.,.] denote the Lagrange brackets of the (q^i,p_j) with respect to the (Q^i,P_j). We remark that the expression obtained for $[q^j,p_i]$ is well defined since F_i was assumed nonlinear and in the points where $F''_i=0$, then, also $u'_i=0$ in such a way that the term u'_i/F''_i is well defined. Repeating the argument, but first changing q^i for p_i, then $F_i(q^i)$ for $G_i(p_i)$, and denoting by v_i the unknown functions obtained when integrating and which now will be of the form $v_i=v_i(p_i)$, we will arrive at

$$[q^i,q^j]=0$$

and

$$[q^i,p_j] = \delta^i_j\left\{1-\frac{\partial}{\partial p_j}\left(\frac{v'_i}{G''_i}\right)\right\}, \quad i,j=1,...,n.$$

These two results must be compatible; this implies that

$$\frac{\partial}{\partial q^i}\left(\frac{u'_i}{F''_i}\right)=\frac{\partial}{\partial P_i}\left(\frac{v'_i}{G''_i}\right)=\text{const.}$$

Hence, we can conclude that

$$[q^i,p_j]=\delta^i_j\, c_i.$$

Finally, we will prove that the n constants $c_i,i=1,...,n$, are really the same. Let us consider as a Hamiltonian the function $K(q^aq^b)$. Then if $S(K)$ denotes the associate generating function, we have

$$\frac{\partial S(K)}{\partial q^j}=K''\left[q^a\frac{\partial Q^k}{\partial p_b}+q^b\frac{\partial Q^k}{\partial p_a}\right](\delta^a_j\,q^b+\delta^b_j\,q^a)$$
$$+K'\left[\delta^a_j\frac{\partial Q^k}{\partial p_b}+q^a\frac{\partial^2Q^k}{\partial q^j\,\partial p_b}+\delta^b_j\frac{\partial Q^k}{\partial p_a}\right.$$
$$\left.+q^b\frac{\partial^2Q^k}{\partial q^j\,\partial p_a}\right]P_k+K'\left[q^a\frac{\partial^2Q^k}{\partial p_j\,\partial p_b}+\frac{\partial^2Q^k}{\partial p_j\,\partial p_a}\right]$$
$$\times\frac{\partial Q^k}{\partial q^j}-(\delta^a_j\,q^b+\delta^b_j\,q^a)K',$$

$$\frac{\partial S(K)}{\partial p_j}=K'\left[q^a\frac{\partial^2Q^k}{\partial p_j\,\partial p_b}+q^b\frac{\partial^2Q^k}{\partial p_j\,\partial p_a}\right]P_k$$
$$+K'\left[q^a\frac{\partial P_k}{\partial p_b}+q^b\frac{\partial P_k}{\partial p_a}\right]\frac{\partial Q^k}{\partial p_j}.$$

Integrating the second equation we find

$$S(K)=K'\left[q^aP_k\frac{\partial Q^k}{\partial p_b}+q^bP_k\frac{\partial Q^k}{\partial p_a}\right]+u_{ab}(q^1,...,q^n),$$

where the function u_{ab} must be determined by substituting into the first equation. In this way we obtain

$$\frac{\partial u_{ab}}{\partial q^j}=(\delta^a_j(c_a-1)q^b+\delta^b_j(c_b-1)q^a)K'.$$

Consequently,

$$\frac{\partial}{\partial q^b}\left(\frac{\partial u_{ab}}{\partial q^a}\right)=(c_a-1)K'+(c_a-1)q^aq^bK'',$$
$$\frac{\partial}{\partial q^a}\left(\frac{\partial u_{ab}}{\partial q^b}\right)=(c_b-1)K'+(c_b-1)q^aq^bK'',$$

and we conclude that $c_a=c_b$.

Therefore, if the transformation is canonoid with respect to the Hamiltonians $K_1(q^1q^2)$, $K_2(q^2q^3),...,K_{n-1}(q^{n-1}q^n)$ we then obtain that $c_1=c_2=\cdots=c_n$ and Theorem 5.1 is proved.

Notice that in the particular case of taking as functions F_i, G_i, and K_i precisely the quadratic functions $(q^i)^2$, $(p_i)^2$, and $q^iq^{i+1},i=1,...,n$, we will recover the quadratic-Hamiltonian theorem of Currie and Saletan.[2] Moreover, it must be remarked that the hypotheses are weaker in this new approach since we have not needed to make use of mixed quadratic terms of the form q^iq^j, with $j\neq i+1$, p_ip_j and q^ip_j, $i,j=1,...,n$.

Consequently, the quadratic-Hamiltonian theorem turns out to be a particular case of Theorem 5.1. Indeed, if we give different values to the $3n$ different functions F_i, G_i, and K_i we will obtain different families of $5n$ independent Hamiltonians (they are really $5n-1$ since the K_n is superfluous because we do not need the term q^nq^1) in such a form that every one will be a sufficient set in order to assure the extended canonicity of the transformation. Obviously, the simplest case will correspond to the aforementioned $F_i=(q^i)^2$, $G_i=(p_i)^2$, and $K_i=q^iq^{i+1}$.

ACKNOWLEDGMENTS

We thank C. López for valuable discussions and the referee for his comments.

Partial financial support by CICYT (Comisión Interministerial Científica y Técnica) is acknowledged.

[1] E. J. Saletan and A. H. Cromer, *Theoretical Mechanics* (Wiley, New York, 1971).

[2] D. G. Currie and E. J. Saletan, Nuovo Cimento B **9**, 143 (1972).

[3] G. Marmo, E. J. Saletan, R. Schmid, and A. Simoni, Nuovo Cimento B **100**, 297 (1987).

[4] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin, New York, 1978).

[5] J. F. Cariñena and M. F. Rañada, J. Math. Phys. **29**, 2181 (1988).

[6] J. F. Cariñena and M. F. Rañada, J. Math. Phys. **30**, 2258 (1989).

[7] G. Marmo, E. J. Saletan, A. Simoni, and B. Vitale, *Dynamical Systems, A Differential Geometric Approach* (Wiley, Chichester, 1985).

[8] Y. Gelman and E. J. Saletan, Nuovo Cimento B **18**, 53 (1973).

[9] L. H. Chung, Proc. R. Soc. London Ser. A **LXII**, 237 (1945).

[10] J. Gomis, J. Llosa, and N. Román, J. Math. Phys. **25**, 1348 (1984).

[11] B. Nachtergaele and A. Verbeure, J. Geom. Phys. **3**, 315 (1986).

[12] M. Crampin and F. A. E. Pirani, *Applicable Differential Geometry* (Cambridge U.P., Cambridge, 1986).

Conformal transformations and the effective action in the presence of boundaries

J. S. Dowker and J. P. Schofield
Department of Theoretical Physics, The University of Manchester, Manchester, England

(Received 11 October 1989; accepted for publication 29 November 1989)

The conformal properties of the heat kernel expansion are used to determine the local form of the coefficients in a manifold with boundary. The conformal transformation of the effective action is obtained. A novel derivation of the boundary term in the Gauss–Bonnet–Chern theorem is detailed.

I. INTRODUCTION

This is a further[1,2] installment of our work concerning some aspects of conformal transformations in conventional field theory. It should be regarded as an immediate continuation of Ref. 2., where, among other things, we considered the relation between dimensional and ζ-function regularization. In our explicit calculations[3,4] using conformal transformations, we employed the ζ-function method. The intention here is to give a closer look at the dimensional approach[5] and also to include boundary effects, which have so far been excluded. We wish to determine the boundary contribution to the conformal transformation of the effective action in three and four dimensions.

The system is a scalar field satisfying Dirichlet conditions. For mathematical convenience, the metric of spacetime is taken with a positive definite signature.

In an extensive Appendix we have presented details of our conventions regarding boundary quantities (such as the extrinsic curvature). We have chosen to do this while giving a novel derivation of the boundary term in the Gauss–Bonnet–Chern theorem. Our justification for such an excursion is that a similar treatment, with due and explicit regard to signs, does not seem to be available. We hope it will prove of value since these topics are of increasing interest to physicists.

II. THE C_1 COEFFICIENT

To illustrate the use of the results of Ref. 2 we derive a slight generalization of an expression first given by Lüscher *et al.*[6] and later by Alvarez[7] for the integral of the local Minakshisundaram coefficient $c_1^{(n)}(y)$ against an arbitrary test function $f(y)$. The general definition is

$$C_k^{(n)}[g;f] \equiv \int_{\mathscr{M}} c_k^{(n)}(g)(y)f(y)dV,$$

where $dV = g^{1/2}\, d_y^n$. Since f can have support on the boundary, this integral will contain volume and surface parts: It is a convenient way of dealing with the local coefficient.

The recent paper of Branson and Gilkey[8] (received after the present work was mostly completed) is also concerned with such integrals.

It is logical to define a corresponding ζ function

$$\zeta\,[s,g;f] \equiv \int \zeta(s,g)(y,y)f(y)dV,$$

related by a Mellin transform to the averaged heat kernel

$$K\,[t;f] \equiv \int K(y,y,t)f(y)dV.$$

All formal results valid for $f = 1$ can be lifted to the general case by the appropriate notational changes. For example, if p is a non-negative integer,

$$\zeta\,[-p,g;f] = (-1)^p[p!/(4\pi)^{n/2}]C_{n/2+p}^{(n)}[g;f].$$

It is sufficient for our purposes to determine the behavior of $\zeta[s,g;f]$ under conformal transformations when $f = 1$.

As in Ref. 2, under a conformal change of g we have

$$\delta\zeta\,[s,\lambda^2 g;1]|_{\lambda=1} = 2s\zeta\,[s,g;\delta\lambda\,] + 2s\zeta\,[s+1,g;\mathbf{J}\,\delta\lambda\,],$$

where

$$\mathbf{J} = (m^2 + (n-1)(\xi - \xi(n))\Delta_2)$$

with

$$\xi(n) = (n-2)/4(n-1).$$

The restriction to $\lambda = 1$ is inessential, but convenient.

If s is set equal to $-p$ it follows that

$$\delta C_k^{(n)}[\lambda^2 g;1]|_{\lambda=1} = (n-2k)C_k^{(n)}[g;\delta\lambda\,] + 2C_{k-1}^{(n)}[g;\mathbf{J}\,\delta\lambda\,], \tag{1}$$

which is our basic equation.[2] We are going to use (1) to derive $C_k^{(n)}[g;f]$ given $C_k^{(n)}[g;1]$, which from now on is sometimes denoted by $C_k^{(n)}[g]$.

Equation (1) can be rewritten as

$$C_k^{(n)}[g;\delta\lambda\,] = [1/(n-2k)][\delta C_k^{(n)}[\lambda^2 g]|_{\lambda=1} - 2C_{k-1}^{(n)}[g;\mathbf{J}\,\delta\lambda\,]] \tag{2}$$

in order to emphasize the structure of the calculation. As an example we take $k = 1$.

The expression for the standard integrated coefficient $C_1^{(n)}[g]$ corresponding to the equation of motion

$$(\Delta_2 - \xi R + m^2)\phi = 0$$

is

$$C_k^{(n)}[g] = m^2|\mathscr{M}| + \left[\frac{1}{6} - \xi\right]\int_{\mathscr{M}} R\,dV + \frac{1}{3}\int_{\partial\mathscr{M}} \kappa\,dS, \tag{3}$$

where dS is the covariant surface area element on the smooth boundary $\partial\mathscr{M}$, i.e., $dS = h^{1/2}\,d^{\,n-1}x$, where h is the induced metric on $\partial\mathscr{M}$. Here κ is the trace of the extrinsic curvature κ_{ij} of $\partial\mathscr{M}$ in $\mathscr{M}$.

0022-2488/90/040808-11$03.00

The first term in the bracket in (2), for $k=1$, can be determined directly from (3) and the known conformal transformations of R and κ.[9,10] One finds

$$\delta C_1^{(n)}[\lambda^2 g]|_{\lambda=1}$$
$$= (n-2)\left[\left(\frac{1}{6}-\xi\right)\int_{\mathscr{M}} R\,\delta\lambda\,dV + \frac{1}{3}\int_{\partial\mathscr{M}} \kappa\,\delta\lambda\,dS\right]$$
$$+ nm^2\int_{\mathscr{M}} \delta\lambda\,dV - 2\xi(n-1)\int_{\partial\mathscr{M}} n^\mu\,\partial_\mu\,\delta\lambda\,dS, \quad (4)$$

where n^μ is the *inward* unit normal to the boundary $\partial\mathscr{M}$.

If (4) is substituted in the rhs of (2), the last two terms of (4) combine with the second term in the bracket of (2) to give a factor of $(n-2)$, which can be cancelled, leaving, if $\delta\lambda(y)$ is replaced by $f(y)$,

$$C_1^{(n)}[g;f] = m^2\int_{\mathscr{M}} f\,dV + \left(\frac{1}{6}-\xi\right)\int_{\mathscr{M}} Rf\,dV$$
$$+\frac{1}{3}\int_{\partial\mathscr{M}} \kappa f\,dS - \frac{1}{2}\int_{\partial\mathscr{M}} n^\mu\,\partial_\mu f\,dS. \quad (5)$$

The generalization over the result of Refs. 6 and 7 consists of working with a more general equation of motion in n dimensions.

The last term in (5) has been "generated" by the conformal transformation. A direct calculation from the heat equation is more messy, although, of course, we have assumed that the form of $C_1^{(n)}[g;1]$ is already known. However, Branson and Gilkey[8] have developed a systematic method of finding *all* the numerical coefficients based just on functorial properties of the $C_k^{(n)}[g;f]$ similar to (1).

III. EFFECTIVE ACTION I

In this section attention is restricted to the conformally invariant situation, i.e., $m=0$, $\xi=\xi(n)$.

There are two equivalent[2] ways of evaluating the finite effective action $W_R[g]$.

One method consists of working in the actual dimension of the space-time and integrating the anomaly, which, up to a factor, is just the variation $\delta W_R[\lambda^2 g]/\delta\lambda$. In two dimensions this amounts to integrating $C_1^{(2)}[g;\delta\lambda\ ]$ [see (5)] with respect to λ and yields a well-known result.[6,7,11] The same method has been applied to four dimensions[12,1] (in the case of an empty boundary).

The other method uses the result that

$$W_R[\lambda^2 g] - W_R[g]$$
$$= \lim_{n\to m}(4\pi)^{-n/2}\left[\frac{C_{m/2}^{(n)}[\lambda^2 g] - C_{m/2}^{(n)}[g]}{n-m}\right], \quad (6)$$

where m is the dimension of space-time and n is an arbitrary dimension. For $m=4$ this approach has been used by Brown and Ottewill.[5] In some ways this approach is more convenient since it deals directly with W and not its derivative, thus avoiding the function integration. Such a method has been used by Melmed[16] when $\partial\mathscr{M}\neq\emptyset$.

In order to discuss the case when $\partial\mathscr{M}$ is not empty, we need the complete coefficient $C_{m/2}^{(n)}[g]$, including boundary contributions; thus before returning to (6), we consider the $m=4$ and $m=3$ cases more closely.

IV. THE C_2 COEFFICIENT

We revert to the more general equation of motion and recall that the form of $C_k^{(n)}[g]$ is

$$C_k^{(n)}[g] = \int_{\mathscr{M}} a_k^{(n)}(g)(y)dV + \int_{\partial\mathscr{M}} b_k^{(n)}(g)(x)dS. \quad (7)$$

The volume density for $k=2$ has been known for some time[13,14] and is

$$a_2^{(n)}(g)(y) = (1/180)[|\text{Riem}|^2 - |\text{Ric}|^2 + \Delta_2 R\ ]$$
$$-\tfrac{1}{36}(6\xi-1)\Delta_2 R + \tfrac{1}{72}((6\xi-1)R - 6m^2)^2, \quad (8)$$

where $|\text{Riem}|^2 = R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma}$ and $|\text{Ric}|^2 = R_{\mu\nu}R^{\mu\nu}$. We note that the numerical coefficients are independent of the dimension n.

The boundary contribution $b_2^{(4)}$ has recently been elucidated by Moss[15] using the work of Melmed[16] on the conformal case, which itself is based on an earlier calculation[17] by Kennedy for a flat embedding space $\mathscr{M}-\partial\mathscr{M}$.

The general form of $b_2^{(n)}$ was written by Kennedy[18] as a linear combination of terms constructed from the extrinsic curvature of $\partial\mathscr{M}$ and the curvature $R_{\mu\nu\rho\sigma}$ of the embedding space; we repeat it here:

$$b_2^{(n)}(x) = b_{2,1}\kappa + b_{2,2}\kappa\,\text{tr}(\kappa^2) + b_{2,3}\ \text{tr}(\kappa^3)$$
$$+ b_{2,4}\kappa R + b_{2,5}\kappa R_{\mu\nu}n^\mu n^\nu + b_{2,6}R_{\mu\nu}\chi^{\mu\nu}$$
$$+ b_{2,7}R_{\mu\nu\rho\sigma}\chi^{\mu\rho}n^\nu n^\sigma + b_{2,8}n^\mu\,\partial_\mu R.$$

We have defined

$$\chi^{\mu\nu} = y^\mu{}_{,i}y^\nu{}_{,j}\kappa^{ij},$$

where $y^\mu(x)$ define the boundary. (See the Appendix for an explanation of the notation used here and later.)

Some of the coefficients, which we again note are independent of the dimension n, were determined by consideration of special cases such as disks and hemispheres. Thus for Dirichlet conditions and $m=0$,

$$b_{2,1} = \tfrac{1}{189},\quad b_{2,2} = -\tfrac{11}{315},\quad b_{2,3} = \tfrac{8}{189},\quad b_{2,4} = \tfrac{1}{18}(1-6\xi). \quad (9)$$

Moss[15] employs a similar approach, but is able to get further in four dimensions by using Melmed's[16] results, as we now discuss, restricting ourselves to Dirichlet conditions for simplicity.

Kennedy[17,18] shows that when $R_{\mu\nu\rho\sigma}$ is zero, $b_2^{(n)}$ is given by

$$b_2^{(n)} = \tfrac{1}{945}f(\kappa)$$
$$\equiv \tfrac{1}{945}\,[40\,\text{tr}(\kappa^3) - 33\,\text{tr}(\kappa)\text{tr}(\kappa^2) + 5(\text{tr}(\kappa))^3], \quad (10)$$

where, for example, $\text{tr}(\kappa^2) = \kappa^i{}_j\kappa^j{}_i$ and the indices are raised and lowered by the intrinsic metric γ_{ij} on $\partial\mathscr{M}$. [This result is given by the first three coefficients in (9).]

To set our conventions some very standard equations are now presented. Under the conformal transformation $g_{\mu\nu}\to\lambda^2 g_{\mu\nu} = e^{-2\omega}g_{\mu\nu}$ we have the following scalings:

$$\gamma_{ij}\to e^{-2\omega}\gamma_{ij},\quad n^\mu\to e^\omega n^\mu,$$
$$\kappa_{ij}\to\kappa_{ij}(\lambda^2 g) = e^{-\omega}(\kappa_{ij} + \gamma_{ij}n^\mu\omega_\mu),$$

so that

$$\kappa\to\kappa(\lambda^2 g)=e^{\omega}(\kappa+(n-1)n^{\mu}\omega_{\mu}).$$

For the embedding scalar curvature,

$$R\to R(\lambda^2 g)$$
$$=e^{2\omega}(R(g)+2(n-1)\Box\omega+(n-1)(2-n)\omega^{\mu}\omega_{\mu})$$

(where $\omega_{\mu}\equiv\partial_{\mu}\omega$ and, for later use, $\omega_{\mu\nu}\equiv\nabla_{\nu}\,\partial_{\mu}\omega$).

Define the trace-free part of κ by

$$P_{ij}\equiv\kappa_{ij}-[1/(n-1)]\kappa\gamma_{ij}.$$

Then

$$P^i{}_j\to P^i{}_j(\lambda^2 g)=e^{\omega}P^i{}_j.$$

As pointed out by Melmed,[16] this simple transformation law is the key to extending Kennedy's[17,18] flat embedding space result since one can easily construct conformally invariant polynomials in κ_{ij}. A traced string of s $P^i{}_j$ tensors scales by $e^{s\omega}$, so that it is only necessary to find a product of such strings which scales as $e^{\omega(n-1)}$ in order to give a conformal invariant when integrated over the boundary.

We now return to the coefficient (7) with $k=2$ and, for the moment, set $n=4$: Since the numerical coefficients are universal this involves no loss of generality.

The aim is to use the conformal properties of $C_2^{(4)}$ in order to find these numerical coefficients and hence the exact form of $C_2^{(n)}[g]$, from which $W_R[\lambda^2 g]-W_R[g]$ can be found via (6). Further use of (2) then allows $C_2^{(n)}[g;f]$ to be determined.

Since we are interested in variations it is convenient to rewrite the volume density $a_2^{(4)}$ [see (8)] as

$$a_2^{(4)}(g)(y)=(1/120)|\mathrm{Weyl}|^2-(4\pi^2/45)\chi_V^{(4)}(y)$$
$$-\tfrac{1}{30}(5\xi-1)\Delta_2 R+\tfrac{1}{72}((6\xi-1)R-6m^2)^2, \tag{11}$$

where $\chi_V^{(4)}(y)$ is the Euler characteristic volume density and $|\mathrm{Weyl}|^2=C_{\mu\nu\rho\sigma}C^{\mu\nu\rho\sigma}$, with $C^{\mu\nu\rho\sigma}$ the Weyl conformal tensor.

We proceed in two stages. First, the conformally invariant ($m=0$, $\xi=\frac{1}{6}$) case is considered, following Melmed[16]; this is then extended to the more general equation of motion using (1).

If (11) is substituted into (7) one finds, if $m=0$ and $\xi=\frac{1}{6}$,

$$C_{2CF}^{(4)}[g]=\frac{1}{120}\int_{\mathcal{M}}|\mathrm{Weyl}|^2\,dV$$
$$-\frac{4\pi^2}{45}\chi^{(4)}+\frac{1}{945}\int_{\partial\mathcal{M}}f(\kappa)dS$$
$$+\frac{4\pi^2}{45}\int_{\partial\mathcal{M}}\chi_B^{(4)}\,dS+\int_{\partial\mathcal{M}}U\,dS. \tag{12}$$

A number of things have been done to obtain this (intermediate) form. First, the boundary contribution to the Euler characteristic has been added and subtracted so that the full Euler invariant $\chi^{(4)}$ is exhibited. Second, Kennedy's[17,18] flat space result (10) has been included. The U term is the unknown remainder, which must vanish when the embedding space is flat, $R^{\mu\nu\rho\sigma}=0$. The $\Delta_2 R$ term has been integrated by parts to give a boundary term that has been absorbed into U. We have also put into U those terms in the boundary contribution to $\chi^{(4)}$ that vanish with $R^{\mu\nu\rho\sigma}$. Thus explicitly, $\chi_B^{(4)}$ in (12) is

$$\chi_B^{(4)}=(1/12\pi^2)[2\,\mathrm{tr}(\kappa^3)-3\,\mathrm{tr}(\kappa)\mathrm{tr}(\kappa^2)$$
$$+(\mathrm{tr}(\kappa))^3]\equiv(1/32\pi^2)p(\kappa).$$

We now note the essential relation

$$8f(\kappa)+21p(\kappa)=168G(\kappa), \tag{13}$$

where $G(\kappa)$ is the conformally covariant polynomial[16]

$$G(\kappa)=\mathrm{tr}(P^3)=\mathrm{tr}(\kappa^3)-\mathrm{tr}(\kappa)\mathrm{tr}(\kappa^2)+\tfrac{2}{9}(\mathrm{tr}(\kappa))^3. \tag{14}$$

Thus (12) becomes

$$C_{2CF}^{(4)}[g]=\frac{1}{120}\int_{\mathcal{M}}|\mathrm{Weyl}|^2\,dV-\frac{4\pi^2}{45}\chi^{(4)}$$
$$+\frac{2}{35}\int_{\partial\mathcal{M}}G\,dS+\int_{\partial\mathcal{M}}U\,dS.$$

The important fact that $C_{2CF}^{(4)}$ is conformally invariant[19] is used to determine the unknown expression U. The first three terms in the above equation are manifestly conformally invariant and the claim[16] has been made that there are no terms constructed from $R^{\mu\nu\rho\sigma}$ and κ_{ij} which make the last U term also invariant. In fact, it is simple to show that the expression

$$\int_{\partial\mathcal{M}}C_{\mu\nu\rho\sigma}n^{\nu}n^{\sigma}P^{\mu\rho}\,dS$$
$$=\int_{\partial\mathcal{M}}\left(R_{\mu\nu\rho\sigma}n^{\nu}n^{\sigma}\chi^{\mu\rho}+\frac{\kappa}{2}R_{\mu\nu}n^{\mu}n^{\nu}\right.$$
$$\left.+\frac{1}{2}R_{\mu\nu}\chi^{\mu\nu}-\frac{1}{6}R\kappa\right)dS$$

is conformally invariant and, also, that this is the only possible such term. The complete conformal coefficient is thus (see Moss and Dowker[20])

$$C_{2CF}^{(4)}[g]=\frac{1}{120}\int_{\mathcal{M}}|\mathrm{Weyl}|^2\,dV-\frac{4\pi^2}{45}\chi^{(4)}$$
$$+\frac{2}{35}\int_{\partial\mathcal{M}}G\,dS-\frac{1}{15}\int_{\partial\mathcal{M}}C_{\mu\nu\rho\sigma}n^{\nu}n^{\sigma}P^{\mu\rho}\,dS. \tag{15}$$

The coefficient $\frac{1}{15}$ can be determined from Kennedy's[17,18] values [see (9)] for $\xi=\frac{1}{6}$ since in this case there is no term like κR. [Kennedy obtained the last coefficient in (9) by looking at product spaces of the form $\mathcal{M}_1\times\mathcal{M}_2$, where $\mathcal{M}_1$ is curved, but has no boundary (e.g., a sphere) and $\mathcal{M}_2$ is flat with a boundary (e.g., a disk). It can be checked that (15) gives the correct constant term in the heat kernel expansions on these spaces.]

We now reinstate the nonconformally invariant terms ($m\neq 0$, $\xi\neq\frac{1}{6}$) in (11) and construct the integrated coefficient (7). Again, the $\Delta_2 R$ term is integrated by parts and relegated to the boundary. Using (15) we can write the complete coefficient as

$$C_2^{(4)}[g]=C_{2CF}^{(4)}[g]+\frac{1}{72}\int_{\mathscr{M}}((6\xi-1)R-6m^2)^2\,dV+\int_{\partial\mathscr{M}}Y\,dS, \tag{16}$$

where Y is to be found.

From the structure of (16) and the fact that Y must vanish in the conformal case (up to a divergence), Y takes the general form (cf. Moss[15])

$$Y=\alpha((1-6\xi)R+6m^2)\kappa+\beta(1-6\xi)n^\mu\partial_\mu R. \tag{17}$$

Moss[15] fixes α and β [as well as the coefficient of G in (15)] by special case calculations. Our approach is to use the general conformal property (1) for the values $n=4$ and $k=2$, so that (1) reduces to the simpler form

$$\delta C_2^{(4)}[\lambda^2 g]|_{\lambda=1}=2\int_{\mathscr{M}}c_1^{(4)}(g)(y)\left(m^2+3\left(\xi-\frac{1}{6}\right)\Delta_2\right)\delta\lambda(y)\,dV. \tag{18}$$

The rhs of (18) can be found from (5) and the lhs is found from (16). We note that only the two integrals in (16) will contribute.

From the known conformal behavior of R, κ, etc. it is easily established that $\alpha=\frac{1}{18}$ and $\beta=-\frac{1}{12}$, which are the same values at which Moss arrives. Incidentally, the value of α also follows directly from Kennedy's[18] last coefficient in (9)!

As a technical point, when the conformal variation in (18) is being evaluated, the condition $\lambda=1$ means that one need work only to linear order in ω ($\lambda=e^{-\omega}$), which is very convenient.

Incidentally, Eq. (18) also determines the second term in (16). That is, if we assume the extra, nonconformally invariant volume density to be a general linear combination of m^4, m^2R, and R^2, Eq. (18) uniquely leads to the combination in (16).

Having now found $C_2^{(4)}[g]$ we can determine $C_2^{(n)}$ simply by rewriting everything in terms of $R^{\mu\nu\rho\sigma}$ and κ_{ij} and using the universal nature of the numerical coefficients.

Thus from (16) with (15), or (12), we find the explicit form (Moss and Dowker[20])

$$\begin{aligned}C_2^{(n)}[g]=&\int_{\mathscr{M}}\left[\frac{1}{180}|\mathrm{Riem}|^2-\frac{1}{180}|\mathrm{Ric}|^2+\frac{1}{72}((6\xi-1)R-6m^2)^2\right]dV+\frac{2}{35}\int_{\partial\mathscr{M}}G(\kappa)\,dS-\frac{1}{360}\int_{\partial\mathscr{M}}q(\kappa)\,dS\\&-\frac{1}{15}\int_{\partial\mathscr{M}}\left(R_{\mu\nu\rho\sigma}n^\nu n^\sigma\chi^{\mu\rho}+\frac{\kappa}{2}R_{\mu\nu}n^\mu n^\nu+\frac{1}{2}R_{\mu\nu}\chi^{\mu\nu}-\frac{1}{6}R\kappa\right)dS\\&+\frac{1}{6}\int_{\partial\mathscr{M}}\left[\frac{1}{3}((1-6\xi)R+6m^2)\kappa-\frac{1}{2}(1-6\xi)n^\mu\partial_\mu R\right]dS.\end{aligned} \tag{19}$$

Here $\frac{1}{32}\pi^2 q(\kappa)$ is the integrand of the boundary term in the Gauss–Bonnet–Chern expression for the Euler number. A derivation is given in the Appendix of the standard expression

$$q(\kappa)=-8\det(\kappa)+4\kappa\hat{R}-8\,\mathrm{tr}(\kappa\hat{R}), \tag{20}$$

with

$$\det(\kappa)=\tfrac{1}{16}p(\kappa).$$

Here $\hat{R}_{ijkl}$ is the intrinsic curvature of $\partial\mathscr{M}$ and is related to the embedding curvature $R_{\mu\nu\rho\sigma}$ by the Gauss–Codazzi equations. Because the dimension n does not appear in the Gauss–Codazzi equations, the coefficients are universal whichever curvature is used.

Since $C_2^{(n)}[g;1]$ is now known, the object is to employ (2) for $k=2$, but *any* n, in order to determine $C_2^{(n)}[g;f]$, in particular, $C_2^{(4)}[g;f]$. [Note that one needs Eq. (5) in order to evaluate the last term of (2).]

Working to first order in ω it is found, after some calculation, that

$$\begin{aligned}C_2^{(n)}[g;f]=&\int_{\mathscr{M}}\left[\frac{1}{180}[|\mathrm{Riem}|^2-|\mathrm{Ric}|^2+\Delta_2R]+\frac{1}{2}A^2\right]f\,dV+\frac{1}{6}\int_{\mathscr{M}}A\Delta_2f\,dV\\&+\frac{1}{360}\int_{\partial\mathscr{M}}\left[\left(\frac{320}{21}\mathrm{tr}(\kappa^3)-\frac{88}{7}\kappa\,\mathrm{tr}(\kappa^2)+\frac{40}{21}\kappa^3-4R_{\mu\nu}\chi^{\mu\nu}-4\kappa R_{\mu\nu}n^\mu n^\nu\right.\right.\\&\left.\left.+16R_{\mu\nu\rho\sigma}n^\mu n^\rho\chi^{\nu\sigma}\right)f+\frac{60}{7}\left(14A+\frac{1}{5}\kappa^2-\mathrm{tr}(\kappa^2)\right)n^\mu\partial_\mu f\right]dS\\&+\frac{1}{3}\int_{\partial\mathscr{M}}\left[A\kappa f+\frac{7-45\xi}{30}n^\mu\partial_\mu Rf+\frac{1}{20}(4\kappa+5n^\mu\partial_\mu)\Delta_2f\right]dS,\end{aligned} \tag{21}$$

where $A=(m^2+(\frac{1}{6}-\xi)R)$.

An integration by parts has been performed to introduce the conventional Δ_2R term in the volume part, although this is not necessarily convenient for further manipulation.

Equation (21), the conclusion of this section, agrees, up to intrinsic divergences, with the expression given by Branson and Gilkey,[8] derived by a method that also involves functorial relations between the Minakshisundaram coefficients, but which differs in detail and general approach.

To achieve the agreement it is necessary to use the relation between the embedding and intrinsic Laplacians:

$$\Delta_2\omega + \kappa(n^\mu\omega_\mu) = \hat{\Delta}_2\omega + n^\mu n^\nu\omega_{\mu\nu}.$$

V. THE $C_{3/2}$ COEFFICIENT

In three dimensions the relevant object is $C^{(n)}_{3/2}[g]$, which has been calculated in Ref. 21: It is conveniently written as, for $m = 0$,

$$C^{(n)}_{3/2}[g] = \frac{\sqrt{\pi}}{192}\int_{\partial\mathscr{M}}(-3\kappa^2 + 6\,\mathrm{tr}(\kappa^2) - 4\hat{R} + 12(8\xi - 1)R)dS.$$

A straightforward application of (2) yields

$$C^{(n)}_{3/2}[g;f] = \frac{\sqrt{\pi}}{192}\int_{\partial\mathscr{M}}[(-3\kappa^2 + 6\,\mathrm{tr}(\kappa^2) - 4\hat{R} + 12(8\xi - 1)R)f + 30\kappa n^\mu\,\partial_\mu f - 24n^\mu n^\nu\nabla_\mu\,\partial_\nu f]dS, \tag{22}$$

which agrees with Branson and Gilkey.[8]

VI. EFFECTIVE ACTION II

Using (21), with $f = 1$, $m = 0$, and $\xi = \xi(n)$, Eq. (6) for the effective action in four dimensions yields, after some calculation,

$$\begin{aligned}
&W_R[e^{-2\omega}g] - W_R[g]\\
&= -\frac{1}{2880\pi^2}\int_{\mathscr{M}}[(|\mathrm{Riem}|^2 - |\mathrm{Ric}|^2 + \Delta_2R)\omega - 2R_{\mu\nu}\omega^\mu\omega^\nu - 4\omega^\mu\omega_\mu\Delta_2\omega + 2(\omega^\mu\omega_\mu)^2 + 3(\Delta_2\omega)^2]dV\\
&\quad - \frac{1}{5760\pi^2}\int_{\partial\mathscr{M}}\bigg[\bigg(\frac{320}{21}\mathrm{tr}(\kappa^3) - \frac{88}{7}\kappa\,\mathrm{tr}(\kappa^2) + \frac{40}{21}\kappa^3 - 4R_{\mu\nu}\chi^{\mu\nu}\\
&\quad - 4\kappa R_{\mu\nu}n^\mu n^\nu + 16R_{\mu\nu\rho\sigma}n^\mu n^\rho\chi^{\nu\sigma} + 2n^\mu\,\partial_\mu R\bigg)\omega - N\bigg(\frac{12}{7}\kappa^2 - \frac{60}{7}\mathrm{tr}(\kappa^2) - 12\Delta_2\omega + 8\omega^\mu\omega_\mu\bigg)\\
&\quad - \frac{4}{7}N^2\kappa - \frac{16}{21}N^3 + 24\kappa\Delta_2\omega - 4\chi^{\mu\nu}\omega_\mu\omega_\nu - 20\kappa\omega^\mu\omega_\mu + 30n^\mu\,\partial_\mu(\Delta_2\omega - \omega^\nu\omega_\nu)\bigg]dS,
\end{aligned} \tag{23}$$

where $N \equiv n^\mu\omega_\mu$. The alternative method of integrating the trace anomaly, used in earlier works,[5,12] will now be outlined.

The standard anomaly equation is

$$\delta W_R[\lambda^2 g] = -(4\pi)^{-2}C^{(4)}_2[\lambda^2 g;\delta\omega] \quad (\lambda = \exp(-\omega)),$$

so that if we set $\delta\omega = \omega\,dt$,

$$W_R[\lambda^2 g] - W_R[g] = -(4\pi)^{-2}\int_0^1 C^{(4)}_2[\exp(-2\omega t)g;\omega]dt. \tag{24}$$

For the integrand we can use (21), with $m = 0$ and $\xi = \frac{1}{6}$. In four dimensions the expression simplifies a little:

$$\begin{aligned}
C^{(4)}_2[g;f] &= \frac{1}{180}\int_{\mathscr{M}}[|\mathrm{Riem}|^2 - |\mathrm{Ric}|^2 + \Delta_2R\,]f\,dV + \frac{2}{35}\int_{\partial\mathscr{M}}G(\kappa)f\,dS - \frac{1}{360}\int_{\partial\mathscr{M}}qf\,dS + \frac{1}{180}\int_{\partial\mathscr{M}}n^\mu\,\partial_\mu\,Rf\,dS\\
&\quad + \frac{1}{60}\int_{\partial\mathscr{M}}(4\kappa - 5n^\mu\,\partial_\mu)\Delta_2 f\,dS - \frac{1}{42}\int_{\partial\mathscr{M}}\bigg(\frac{1}{5}\kappa^2 - \mathrm{tr}(\kappa^2)\bigg)n^\mu\,\partial_\mu\,f\,dS - \frac{1}{15}\int_{\partial\mathscr{M}}C_{\mu\nu\rho\sigma}n^\nu n^\sigma P^{\mu\rho}f\,dS.
\end{aligned} \tag{25}$$

If (25) is substituted into (24) and the integration performed it is found that we regain (23), as we should.

We do not expand on the details except to remark that they are somewhat simpler than in the dimensional method. The reason is that in the latter, although one deals only with $C^{(n)}_2[g;1]$, a conformal transformation has to be performed to all orders in ω and then an overall factor of $(n - 4)$ extracted. In the integration method it is true that one first has to find $C^{(n)}_2[g;f]$, but this involves just a first-order transformation, which is much simpler. A complete transformation is next necessary, but now one uses $C^{(4)}_2$ and no factor need be extracted.

Formula (23) reduces to the known one[5,12] when the boundary is empty. (We draw attention to the fact that the corresponding quantity in Ref. 3 suffers from algebraic errors.) Melmed has also derived an expression for W using the dimensional method,[22] but his starting expression for the $C^{(4)}_2$ coefficient is incomplete.

From (22) we can similarly evaluate the effective potential in three dimensions (for $m=0$ and $\xi=\frac{1}{8}$). It is found that

$$W_R[\lambda^2 g]-W_R[g]$$
$$=\frac{1}{1536\pi}\int_{\partial\mathscr{M}}[(7\kappa^2-10\,\mathrm{tr}(\kappa^2)-2N^2+4\kappa N$$
$$+4R+4\Delta_2\omega)\omega-6\kappa N-42N^2-4\Delta_2\omega]dS. \quad (26)$$

VII. CONCLUSION

We do not have any immediate application of Eqs. (23) and (26) in mind. However, the first coefficient $C_1^{(n)}$ occurs in string theory and the higher ones occur in the analysis of membrane vacuum energy.[23] It is thus possible that our results will prove useful in this area. Different boundary conditions will have to be investigated.

One possible use would be to extend the analysis of high temperature expansions of the free energy of quantum fields in static space-times[3,4] to the case when the (three dimensional) spatial section has a boundary. At the moment this seems to have only formal interest, but will be considered at another time.

A notable feature of the method using relation (1) is that volume and boundary nonconformally invariant pieces are much easier to evaluate than conformally invariant pieces. In fact, it is the calculation of the latter that involves the most effort. Our derivation of $C_2^{(n)}$ (see Ref. 20 and above) uses Kennedy's[17] result [see (10)], which is where a great deal of the computation is hidden. However, the calculations of Branson and Gilkey[8] prove that it is possible to determine all the terms in $C_k^{(n)}$ from consideration of their functorial properties alone. Despite the numerical work, this is an important advance in the evaluation of these coefficients.

ACKNOWLEDGMENTS

We thank P. B. Gilkey for providing a preliminary version of Ref. 8.

One of us (JPS) would like to thank the Science and Engineering Research Council for financial support.

APPENDIX: BOUNDARY GEOMETRY AND THE GAUSS–BONNET–CHERN THEOREM

For the possible convenience of the reader we describe some mathematical results involving boundary effects. To a large extent, but not entirely, the discussion will be a reworking of standard material, so that it is frankly pedagogical. However, the discussion does serve to fix and explain some notation used earlier. A few comments on the literature are also given to aid the compulsive checker.

Attention is first turned to Chern's intrinsic proof of the generalized Gauss–Bonnet theorem. A useful, but rather abbreviated, summary is given by Kobayashi and Nomizu,[24] which is roughly followed below.

The proof involves an essential result due to Hopf. Consider an n- (even) dimensional manifold $\mathscr{M}$ with boundary $\partial\mathscr{M}$ and imagine a vector field X to exist on $\mathscr{M}$. Hopf showed that the Euler number of $\mathscr{M}$ (defined either strictly topologically in terms of a simplicial decomposition or as the alternating sum of the dimensions of the cohomology groups of $\mathscr{M}$) equals the sum of the indices of X at its zeros, or at its singularities if we take X to be a unit vector field, as we do from now on. We shall not prove this result here.

Without loss of generality, assume that there is only one singular point p_0 which does not lie on $\partial\mathscr{M}$; X is otherwise arbitrary.

In global language, X is a section of the unit tangent bundle of $\mathscr{M}$, i.e., of the sphere bundle $S(\mathscr{M})$; X is a mapping from $\mathscr{M}$ to $S(\mathscr{M})$.

The local coordinates on $S(\mathscr{M})$ can be taken to be the local coordinates y^μ on $\mathscr{M}$, together with the tangent-space vector components u^ν and subject to the normalization $u^\nu u_\nu=1$. Thus $S(\mathscr{M})$ is a manifold of $2n-1$ dimensions. Chern's important idea was to work on $S(\mathscr{M})$ rather than on $\mathscr{M}$.

If the u^ν coordinates of the point (y^μ, u^ν) of $S(\mathscr{M})$ are forgotten we obtain the point $\{y^\mu\}$ of $\mathscr{M}$ that lies "below" it. This projection will be denoted by $\pi{:}S(\mathscr{M})\to\mathscr{M}$ or $\pi{:}(y,u)\to\{y\}$. Thus $\pi^{-1}(P)$ is all that part of $S(\mathscr{M})$ lying above the point P of $\mathscr{M}$. This *fiber* is isometric to the $(n-1)$ sphere.

As a submanifold of $S(\mathscr{M})$, X has a boundary consisting of the union of those points that lie above the singularity point p_0 with those above the geometric boundary of $\mathscr{M}$, $\partial\mathscr{M}$. In symbols,

$$\partial X(\mathscr{M})=X(p_0)\cup X(\partial\mathscr{M}). \quad \text{(A1)}$$

The important fact (a restatement of Hopf's result) is that

$$X(p_0)=-\chi\pi^{-1}(p_0). \quad \text{(A2)}$$

In words, (A2) says that the part of the boundary of $X(\mathscr{M})$ that lies above the singularity point consists of the fiber at that point taken, negatively, as many times as the Euler number χ of $\mathscr{M}$. The minus sign is an orientation effect.

The generalization of the Euler number density $R/4\pi$ in two dimensions is now introduced as an n form Λ (detailed later) on $\mathscr{M}$ such that, *when pulled back to* $S(\mathscr{M})$, it is exact:

$$\pi^*\Lambda=-d\Pi, \quad \text{(A3)}$$

where Π is an $(n-1)$-form on $S(\mathscr{M})$, which Chern constructs.

We now note that π^* maps the cohomology of $\mathscr{M}$ into that of $S(\mathscr{M})$, while X^* performs the inverse operation. Thus $X^*\pi^*$ amounts to the identity, so that

$$\int_{\mathscr{M}}\Lambda=\int_{\mathscr{M}}X^*\pi^*(\Lambda)=\int_{X(\mathscr{M})}\pi^*\Lambda.$$

The integral of the Euler density on $\mathscr{M}$ has been turned into an integral of its pullback over the image of $\mathscr{M}$ in $S(\mathscr{M})$ via the vector field X. (Chern actually makes no formal distinction between these two integrals.)

Because of (A3), Stokes' theorem can be invoked on $S(\mathscr{M})$ to give, using (A1) and (A2),

$$\int_{\mathscr{M}}\Lambda=-\int_{X(\mathscr{M})}d\Pi=-\int_{\partial X(\mathscr{M})}\Pi$$

$$= \chi \int_{\pi^{-1}(p_0)} \Pi - \int_{X(\partial\mathscr{M})} \Pi. \quad \text{(A4)}$$

The integral of Π over the fiber S^{n-1} is unity, so that the first term on the rhs of (A4) is just χ. (We deal with the second, geometrical boundary term later.) The result is an elegant, geometric derivation of the Gauss–Bonnet theorem.

Apart from Chern's original papers[25,26] and lectures,[27] the authors know of no other detailed derivation of the statement just made than the refined, updated treatment by Greub *et al.*[28] Thus we feel that a lower level elaboration would not be out of order here. In any case the geometrical details are helpful, even if somewhat standard.

Cartan's moving frame method is employed with the orthonormal tangent basis $\{\mathbf{e}_a\}$ and the dual-form basis $\{\omega^a\}$. The structure equations are

$$d\omega^a = \omega^b \wedge \omega_b{}^a, \quad d\omega_a{}^b = \omega_a{}^c \wedge \omega_c{}^b + \Omega_a{}^b.$$

The curvature two-form $\Omega_a{}^b$ is given in terms of the curvature tensor by

$$\Omega_a{}^b = \tfrac{1}{2} R_a{}^b{}_{cd}\, \omega^c \wedge \omega^d.$$

Our conventions are those of Boothby[29] and seem to agree with those of Chern.

It might be useful to remark here that our curvature tensor has the *opposite* sign to that in Hawking and Ellis,[30] for example. This difference is occasioned by the arrangement of the structure equations. Hicks[31] makes a choice that amounts to switching the indices on ω_{ab} and Ω_{ab}. The latter sign accounts for the lack of a factor of $(-1)^p$ in Hicks' definition of $\pi^*\Lambda$. (Reference 31, p. 114 denotes the corresponding quantity Q and, like Chern, does not distinguish between Λ and $\pi^*\Lambda$.) Hick's curvature tensor agrees with that of Hawking and Ellis. (The same remarks also apply to Spivak[32] and Eguchi *et al.*[33] Note, also, that Eisenhart's definition[9] of the curvature is the same as that of Hawking and Ellis, but that his Ricci and scalar curvatures have the opposite sign. Our $R_{\mu\nu}$ and R agree with those of Hawking and Ellis.)

Curiously, Kobayashi and Nomizu[24] appear to use the same conventions as Hicks,[31] but have a $(-1)^p$ in their expression for $\pi^*\Lambda$.

Another curiosity is that Gilkey[34] (pp. 338–339) follows Chern's formulas exactly, apart from omitting the minus sign in (A3). It seems that the second integral in the equation for χ on p. 339 should have the opposite sign.

The present authors admit to continuing difficulty in chasing through the minus signs in this topic, but hope those here are correct.

The raising and lowering of indices is purely cosmetic if the frames are orthonormal and the signature Euclidean.

If $\mathbf{u}$ is a tangent vector, then $\mathbf{u} = u^a \mathbf{e}_a$ and

$$d\mathbf{u} = \theta^a \mathbf{e}_a,$$

where the one-form θ is

$$\theta^a = du^a + u^b \omega^a.$$

A basis for forms on $S(\mathscr{M})$ is provided by the set $\{\omega^a, \theta^a\}$.

The explicit expression for the form $\pi^*\Lambda$ is, if n is even $(n = 2p)$, the Pfaffian of the curvature matrix:

$$\pi^*\Lambda = [(-1)^p/2^n \pi^p p!]\, \epsilon_{a_1 \cdots a_n} \Omega^{a_1 a_2} \wedge \Omega^{a_3 a_4} \wedge \cdots \wedge \Omega^{a_{n-1} a_n}. \quad \text{(A5)}$$

Using a recursion method, Chern showed that Π is given by the sum

$$\Pi = \frac{1}{\pi^p} \sum_{k=0}^{p-1} (-1)^k \frac{2^{-p-k}}{1.3 \cdots (2p-2k-1) k!} \Phi_k, \quad \text{(A6)}$$

where the Φ_k are $(n-1)$ forms on $S(\mathscr{M})$ given by

$$\Phi_k = \epsilon_{a_1 \cdots a_n} u^{a_1} \theta^{a_2} \wedge \cdots \wedge \theta^{a_{n-2k}} \wedge \Omega^{a_{n-2k+1} a_{n-2k+2}} \wedge \cdots \wedge \Omega^{a_{n-1} a_n}. \quad \text{(A7)}$$

We shall derive (A6) later by a simplified method.

The forms Λ and Φ_k are intrinsically defined.

We remark here that we are following the algebra of Chern's first paper[25] except for the sign of Λ, which is as in his second paper[26] (and later works).

Substitute (A6) into the first integral on the rhs of (A4). The integration domain is the fiber over the point p_0. On this domain all terms in Π that contain an Ω^{ab} will vanish because Ω^{ab} involves the ω^a, which are one-forms in dy^μ and thus are zero on the fiber. Hence only the Φ_0 term survives: Its form is

$$\Phi_0 = \epsilon_{a_1 \cdots a_n} u^{a_1} \theta^{a_2} \wedge \cdots \wedge \theta^{a_n}.$$

For the same reason as before, the connection forms $\omega_a{}^b$ are zero on the fiber, so that θ^a can be replaced by du^a. Up to a factor, Φ_0 can then be recognized as the volume form on the sphere S^{n-1} expressed in the Cartesian coordinates u^a. The coefficients are such that the integral of Π over the fiber is unity, which is the conclusion of this part of the calculation.

Equation (A4) can then be written as

$$\chi = \int_{\mathscr{M}} \Lambda + \int_{X(\partial\mathscr{M})} \Pi = \int_{\mathscr{M}} \Lambda + \int_{\partial\mathscr{M}} X^*\Pi. \quad \text{(A8)}$$

The second integral is now looked at more closely.

We choose X to be any extension of the normal vector field $\mathbf{n}$ on the boundary $\partial\mathscr{M}$ and rewrite (A8) as

$$\chi = \int_{\mathscr{M}} \Lambda + \int_{\partial\mathscr{M}} n^*\Pi, \quad \text{(A9)}$$

which is the final, formal statement of the generalized Gauss–Bonnet–Chern theorem.

For definiteness we choose $\mathbf{n}$ to be the *inward* vector. [In even dimensions it is easily checked from the explicit formulas that $\mathbf{n}^*\Pi = (-\mathbf{n})^*\Pi$. In contrast, for odd dimensions, there is a sign change.] The effect of $\mathbf{n}^*$ on the Φ_k of (A7) is to replace u^a by n^a, the components of $\mathbf{n}$, on $\partial\mathscr{M}$. (We could set $\mathbf{u} = -\mathbf{n}$ with no change, in the end.)

At this point, and possibly earlier, it is advantageous to work with boundary adapted frames, which are such that on $\partial\mathscr{M}$, $\mathbf{e}_n = \mathbf{n}$ and $\{\mathbf{e}_a\}$ for $a = 1$ to $n-1$ form a basis for the tangent space to $\partial\mathscr{M}$. (See Ref. 31, p. 81.)

Then on $\partial\mathscr{M}$, $\omega^n = 0$ and the components n^a and θ^a are given by

$$n^a = \delta^{an}, \quad \theta^a = \omega_n{}^a. \quad \text{(A10)}$$

The last of Eqs. (A10) follows from

$$d\mathbf{n} = d\mathbf{e}_n = \omega_n{}^a \mathbf{e}_a, \quad \text{(A11)}$$

which shows that the form $\omega_n{}^a$ is related to the extrinsic curvature of $\partial\mathscr{M}$.

To find the exact relation, some ordinary Riemannian geometry is reproduced.

The intrinsic metric γ_{ij} on the boundary hypersurface $\partial\mathscr{M}$ is given in terms of the metric $g_{\mu\nu}$ on $\mathscr{M}$ by (Eisenhart,[9] p. 146)

$$\gamma_{ij} = g_{\mu\nu} y^\mu{}_{,i} y^\nu{}_{,j} \quad (1 \leqslant i, j \leqslant n-1).$$

Here $\partial\mathscr{M}$ is defined by $y^\mu = y^\mu(x)$ in terms of the functions $y^\mu(x)$, where x^i are the coordinates on $\partial\mathscr{M}$ and y^μ are those on $\mathscr{M}$.

The inversion of the above equation yields

$$g^{\mu\nu} = h^{\mu\nu} + n^\mu n^\nu,$$

where

$$h^{\mu\nu} = y^\mu{}_{,i} y^\nu{}_{,j} \gamma^{ij}.$$

The tensor $h^{\mu\nu}$ is equivalent to γ^{ij} and can be considered to be the induced metric on $\partial\mathscr{M}$. (See Hawking and Ellis,[30] Sec. 2.7). The tensor $h^\mu{}_\nu$ is a projection operator.

The extrinsic curvature $\chi_{\mu\nu}$ is defined by the projection of the derivative of an extension of the *outward* normal vector field $-\mathbf{n}$:

$$\chi_{\mu\nu} = -n_{\rho;\sigma} h^\rho{}_\mu h^\sigma{}_\nu = \chi_{\nu\mu}. \tag{A12}$$

[The sign here is that used by Eisenhart[9] (his ξ is our $\mathbf{n}$) and McKean and Singer.[35] The sign is such that $\kappa = \operatorname{tr}\chi$ is positive for a disk. Equation (A12) should be compared with the definition of the second fundamental form $V(X,Y)$ in terms of the difference of the covariant derivatives in the embedding and embedded spaces (Hicks,[31] p. 75):

$$D_X Y - \widehat{D}_X Y = V(X,Y) = -\langle LX, Y\rangle \mathbf{n},$$

where $\widehat{D}$ is the hypersurface derivative and $L(X) \equiv D_X \mathbf{n}$ is the Weingarten map.]

We note the condition

$$n^\mu \chi_{\mu\nu} = 0.$$

The intrinsic components of κ are given by

$$\kappa_{ij} = \chi_{\mu\nu} y^\mu{}_{,i} y^\nu{}_{,j}.$$

Similarly, we can define the traceless $\chi_{\mu\nu}$ by

$$P^{\mu\nu} = P^{ij} y^\mu{}_{,i} y^\nu{}_{,j},$$

so that

$$P_{\mu\nu} = \chi_{\mu\nu} - [1/(n-1)]\chi_\rho{}^\rho h_{\mu\nu}.$$

These general equations can be written with respect to boundary adapted frames. To emphasize this special choice, the indices μ, ν, ρ, etc. are changed to a, b, c, etc. The projection $h^a{}_b$ then takes the form, expressed as a matrix,

$$\mathbf{h} = \mathbf{1}_n - \mathbf{n}\times\mathbf{n} = \begin{pmatrix} \mathbf{1}_{n-1} & 0 \\ 0 & 0 \end{pmatrix},$$

so that the effect of a projection with $\mathbf{h}$ is simply to restrict the index range to $1 \to (n-1)$.

For example, (A12) becomes

$$\chi_{ab} = \begin{cases} -n_{a;b}, & \text{if } 1 \leqslant a,b \leqslant n-1, \\ 0, & \text{otherwise,} \end{cases}$$

and comparison with (A11) shows that

$$\omega_n{}^a = -\chi^a{}_b \omega^b,$$

the sought-after relation.

We now return to the expression for $n^*\Pi$ on $\partial\mathscr{M}$.

Setting u^a equal to n^a with boundary adapted frames means that $\mathbf{n}^*\Phi_k$ reduces on $\partial\mathscr{M}$ to

$$\mathbf{n}^*\Phi_k = -\epsilon_{a_2\cdots a_n} \omega_n{}^{a_2} \wedge \cdots \wedge \omega_n{}^{a_{n-2k}} \wedge \Omega^{a_{n-2k+1} a_{n-2k+2}} \wedge \cdots \wedge \Omega^{a_{n-1} a_n}, \tag{A13}$$

with all indices restricted to the range $1 \to (n-1)$.

From the structures of θ and Ω (remember $\omega^n = 0$ on $\partial\mathscr{M}$), it follows that $n^*\Phi_k$ is proportional to the volume element dS on $\partial\mathscr{M}$ (cf. Spivak,[32] p. 573).

As an explicit example we look at the case of $n = 4$, which is relevant for the discussion in Sec. IV.

Here Π is given by

$$\Pi = (1/24\pi^2)[2\Phi_0 - 3\Phi_1]$$

and $\mathbf{n}^*\Phi_0$ is

$$\begin{aligned} \mathbf{n}^*\Phi_0 &= -\epsilon_{abc}\omega_n{}^a \wedge \omega_n{}^b \wedge \omega_n{}^c \\ &= \epsilon_{abc}\chi^a{}_d \chi^b{}_e \chi^c{}_f \omega^d \wedge \omega^e \wedge \omega^f \\ &= \epsilon_{abc}\epsilon^{def}\chi^a{}_d \chi^b{}_e \chi^c{}_f \, dS = 3!\det(\chi) dS. \end{aligned}$$

It is here that we use the choice of $\mathbf{e}_n$ as the *inward* normal $\mathbf{n}$. Stokes' theorem takes $\partial\mathscr{M}$ with its induced orientation; $\omega^1 \wedge \omega^2 \wedge \ldots \omega^{n-1}$ is the volume form on $\partial\mathscr{M}$ only if $\mathbf{e}_n = n$ because then, in boundary adapted local coordinates, the coordinate of $\mathscr{M}$ normal to $\partial\mathscr{M}$ is *positive* and $dS > 0$. (See Bott and Tu[36] and Chern,[27] pp. 72–74.)

The form $\mathbf{n}^*\Phi_1$ is likewise evaluated:

$$\mathbf{n}^*\Phi_1 = -\epsilon_{abc}\omega_n{}^a \wedge \Omega^{bc} = \tfrac{1}{2}\epsilon_{abc}\epsilon^{def}\chi^a{}_d R^{bc}{}_{ef} \, dS. \tag{A14}$$

In Eq. (A14) R_{bcef} is the curvature of the embedding manifold $\mathscr{M}$, but the indices run over 1, 2, and 3 only. It is better to rewrite Π in terms of the curvature $\widehat{R}$ intrinsic to the hypersurface. This is accomplished by the Gauss–Codazzi equations, which read, in form language and boundary adapted frames (Hicks,[31] p. 82), as

$$\widehat{\Omega}^{ab} = -\omega_n{}^a \wedge \omega_n{}^b + \Omega^{ab} \quad (1 \leqslant a,b \leqslant n-1), \tag{A15}$$

so that $\mathbf{n}^*\Pi$ is given by

$$\mathbf{n}^*\Pi = -(1/24\pi^2)[\mathbf{n}^*\Phi_0 + 3\widehat{\Phi}_1],$$

where $\widehat{\Phi}_1$ is

$$\begin{aligned} \widehat{\Phi}_1 &= -\epsilon_{abc}\omega_n{}^a \wedge \widehat{\Omega}^{bc} \\ &= \tfrac{1}{2}\epsilon_{abc}\epsilon^{def}\chi^a{}_d \widehat{R}^{bc}{}_{ef} \, dS \\ &= [-\chi^a{}_a \widehat{R} + 2\chi_{ab}\widehat{R}^{ab}] \, dS. \end{aligned}$$

In convariant equations such as these, the indices $a, b, \ldots$ can be replaced by the general ones $\mu, \nu, \ldots$ if desired. Furthermore, κ_{ij} can be used instead of χ_{ab}, or $\chi_{\mu\nu}$. For example,

$$\chi^a{}_a = \chi_\mu{}^\mu = \kappa_i{}^i \equiv \kappa,$$

$$\operatorname{tr}(\chi^2) = \chi_{ab}\chi^{ba} = \chi_{\mu\nu}\chi^{\nu\mu} = \kappa_{ij}\kappa^{ji} = \operatorname{tr}(\kappa^2),$$

and

$$\det(\chi) = \det(\kappa).$$

For the intrinsic curvature $\widehat{R}_{ijkl}$ obtained from the me-ric γ_{ij}, we have the relation

$$\widehat{R}^{\mu\nu\rho\sigma} = \widehat{R}^{ijkl} y^{\mu}{}_{,i} y^{\nu}{}_{j} y^{\rho}{}_{,k} y^{\sigma}{}_{,l}.$$

Note that we are using the same kernel symbol for the two forms of the intrinsic curvature. For those who wish to make comparisons, $\widehat{R}_{\mu\nu\rho\sigma}$ is what Hawking and Ellis[30] call R'_{abcd}, not forgetting the difference in conventions.

The coordinate forms of the Gauss–Codazzi equations are sometimes needed. They are (Eisenhart,[9] p. 149, taking the different conventions into account)

$$\widehat{R}_{ijkl} = \kappa_{il}\kappa_{jk} - \kappa_{ik}\kappa_{jl} + R_{\mu\nu\rho\sigma} y^{\mu}{}_{,i} y^{\nu}{}_{j} y^{\rho}{}_{,k} y^{\sigma}{}_{,l}$$

and

$$\kappa_{ij;k} - \kappa_{ik;j} = -R_{\mu\nu\rho\sigma} y^{\mu}{}_{,i} y^{\rho}{}_{j} y^{\sigma}{}_{,k} n^{\nu}.$$

The boundary term in the Gauss–Bonnet–Chern formula (A9) is

$$\int_{\partial\mathscr{M}} \mathbf{n}^*\Pi = \frac{1}{32\pi^2} \int_{\partial\mathscr{M}} q(\kappa)\, dS,$$

with

$$q(\kappa) = -8 \det(\kappa) + 4\kappa\widehat{R} - 8\kappa_{ij}\widehat{R}^{ij},$$

which is the expression quoted in the text.

If the Gauss–Codazzi equations are used to bring back the embedding curvature, it is found that $\mathbf{n}^*\Phi_1$, (A14), is given by

$$\mathbf{n}^*\Phi_1 = -\left[\kappa R - 2R_{\mu\nu}(\kappa n^{\mu} n^{\nu} + \chi^{\mu\nu}) - 2R_{\mu\nu\rho\sigma}\chi^{\nu\sigma} n^{\rho} n^{\mu}\right] dS. \tag{A16}$$

For completeness, the expression for the volume density Λ will be given and some further, relevant information outlined.

If the form of $\Omega_a{}^b$ is substituted into (A5) there results

$$\Lambda = [(-1)^p/2^n\pi^p p!]\epsilon_{a_1\cdots a_n}\epsilon^{b_1\ldots b_n} \times R^{a_1a_2}{}_{b_1b_2}\cdots R^{a_{n-1}a_n}{}_{b_{n-1}b_n}\, dV \equiv K_n\, dV.$$

(In a *coordinate* frame the indices a, b can be replaced by μ, ν if it is remembered that $\epsilon^{\mu\ldots} = g^{-1}\epsilon_{\mu\ldots}$, with $\epsilon_{\mu\ldots}$ the standard epsilon symbol.)

We refer to Spivak[32] (pp. 385–388) for some historical comments.

The essential statement of the previous discussion is that Λ is not exact on $\mathscr{M}$, but it *is* exact on the bundle $S(\mathscr{M})$, i.e., when pulled back to $S(\mathscr{M})$. However, Λ is closed on $\mathscr{M}$, $d\Lambda = 0$ by virtue of the Bianchi identity (e.g., Spivak,[32] p. 433). [This also follows from (A3).] Thus Λ determines a cohomology class belonging to $H^n(\mathscr{M})$. Further, the class is independent of the connection. This is equivalent to saying that the integral $\int\Lambda$ (taken over a closed manifold) is a topological invariant and follows from the fact that the difference $\Lambda_1 - \Lambda_0$ of two forms computed using two connections ω_1 and ω_0 is exact on $\mathscr{M}$:

$$\Lambda_1 - \Lambda_0 = d\Psi, \tag{A17}$$

as a direct calculation shows (e.g., Spivak,[32] p. 434). This is an example of a general result in the theory of characteristic classes.

The infinitesimal form of Eq. (A17) is easily derived. From (A5),

$$\delta\Lambda = [(-1)^p/2^n\pi^p(p-1)!]\epsilon_{a_1\ldots a_n} \times \delta\Omega^{a_1a_2}\wedge\Omega^{a_3a_4}\wedge\cdots\wedge\Omega^{a_{n-1}a_n};$$

using the structure equations, this becomes

$$\delta\Lambda = [(-1)^p/2^n\pi^p(p-1)!]\epsilon_{a_1\cdots a_n} \times D\delta\omega^{a_1a_2}\wedge\Omega^{a_3a_4}\wedge\cdots\wedge\Omega^{a_{n-1}a_n},$$

where D is the covariant derivative defined by (to check conventions)

$$DF_a{}^b \equiv dF_a{}^b - F_a{}^c\wedge\omega_c{}^b + (-1)^p\omega_a{}^c\wedge F_c{}^b$$

for a matrix p form $F_a{}^b$.

Use of the Bianchi identity $D\Omega^{ab} = 0$ easily produces the infinitesimal exactness condition

$$\delta\Lambda = [(-1)^p/2^n\pi^p(p-1)!]d\left[\epsilon_{a_1\cdots a_n} \times \delta\omega^{a_1a_2}\wedge\Omega^{a_3a_4}\wedge\cdots\wedge\Omega^{a_{n-1}a_n}\right]. \tag{A18}$$

The functional constancy of $\int K_n\, dV$ has been discussed from the conventional coordinate viewpoint by Buchdahl.[37]

Of course, Eq. (A17) is guaranteed by the Gauss–Bonnet theorem, (A9). In fact, the form of Π can be deduced from the explicit construction of Ψ from (A17). This program has been carried out by Spivak[32] using the full apparatus of connections on principle fiber bundles, which we have tried to avoid. [If B is the frame bundle and $\omega: B\to\mathscr{M}$ is the projection, $\omega^*\Lambda$ will be given by an equation identical to (A5) except that $\Omega_a{}^b$ now stands for the curvature on B. A particular orthonormal moving frame $\mathbf{E} = \{\mathbf{e}_a\}$ is a section of B and what we have been calling $\Omega_a{}^b$ equals $E^*\Omega_a{}^b$. Further commentary would be out of place here.]

We shall present an alternative derivation of Π based on the infinitesimal variation (A18). We have not seen this elsewhere. (The corresponding discussion in Ref. 28 is a modern version of Chern's original method.)

Equation (A9) implies that on $\partial\mathscr{M}$,

$$\delta\Theta|_{\partial\mathscr{M}} = \hat{d}\Xi - \delta(\mathbf{n}^*\Pi), \tag{A19}$$

where $\delta\Lambda = d(\delta\Theta)$ [see (A18)] and we have emphasized the fact that we now have to work intrinsically on $\partial\mathscr{M}$ by placing a caret over the exterior derivative.

We force $\delta\Theta$ into the form (A19) to extract information about Π.

First, $\delta\Theta$ is rearranged so that all summed indices range over 1–$n-1$. For the moment we forget the numerical coefficient in (A18) (without change of notation); then

$$\delta\Theta = -2\epsilon_{a_2\cdots a_n}\,\delta\omega_n{}^{a_2}\wedge\Omega^{a_3a_4}\wedge\cdots\wedge\Omega^{a_{n-1}a_n} - (n-2)\epsilon_{a_1\cdots a_{n-1}}\,\delta\omega^{a_1a_2}\wedge\Omega^{a_3a_4}\wedge\cdots\wedge\Omega_n{}^{a_{n-1}}. \tag{A20}$$

The aim is to write $\delta\Theta$ on $\partial\mathscr{M}$ in terms of intrinsic quantities. To do this note that on $\partial\mathscr{M}$ an index equal to n is a "dead" index, so that, for example, $\omega_n{}^a$ is a vector one-form and $\Omega_n{}^a$ is a vector two-form. From the structure equations,

$$\Omega_n{}^a = d\omega_n{}^a - \omega_n{}^c\wedge\omega_c{}^a = \widehat{D}\omega_n{}^a$$

and

$$\delta\widehat{\Omega}_a{}^b = \widehat{D}\delta\omega_a{}^b \tag{A21}$$

in terms of the intrinsic covariant exterior derivative $\hat{D}$. The $\omega_a{}^b$ on $\partial\mathscr{M}$ for $1\leqslant a, b\leqslant n-1$ form a set of intrinsic connection forms (Hicks,[31] p. 81).

We now eliminate $\Omega_a{}^b$ from (A20) in favor of $\hat{\Omega}_a{}^b$ by using the Gauss–Codazzi equation (A15). Because of the ϵ symbol there are no complicated combinatorial problems. A simple binomial expansion is all that is required!

Using (A21) slight manipulation yields

$$\delta\Theta|_{\partial\mathscr{M}} = -\sum_{k=0}^{p-1}\binom{p-1}{k}\frac{2}{n-1-2k}\epsilon_{a_2\cdots a_n}\,\delta[\omega_n{}^{a_2}\wedge\cdots\wedge\omega_n{}^{a_{n-2k}}]\wedge\hat{\Omega}^{a_{n-2k+1}a_{n-2k+2}}\wedge\cdots\wedge\hat{\Omega}^{a_{n-1}a_n}$$

$$-\sum_{k=0}^{p-2}\binom{p-2}{k}\frac{2(p-1)}{n-3-2k}\epsilon_{a_2\cdots a_n}\,\delta\omega^{a_2a_3}\wedge\hat{D}\,[\omega_n{}^{a_4}\wedge\cdots\wedge\omega_n{}^{a_{n-2k}}]\wedge\hat{\Omega}^{a_{n-2k+1}a_{n-2k+2}}\wedge\cdots\wedge\hat{\Omega}^{a_{n-1}a_n}.$$

The *intrinsic* Bianchi identity on $\partial\mathscr{M}$, $\hat{D}\hat{\Omega}=0$ allows the second term to be written as

$$\hat{d}\left[\sum_{k=0}^{p-2}\binom{p-2}{k}\frac{2(p-1)}{n-3-2k}\epsilon_{a_2\cdots a_n}\,\delta\omega^{a_2a_3}\wedge\omega_n{}^{a_4}\wedge\cdots\wedge\omega_n{}^{a_{n-2k}}\wedge\hat{\Omega}^{a_{n-2k+1}a_{n-2k+2}}\wedge\cdots\wedge\hat{\Omega}^{a_{n-1}a_n}\right]$$

$$-\sum_{k=0}^{p-1}\binom{p-1}{k}\frac{2}{n-1-2k}\epsilon_{a_2\cdots a_n}\,\omega_n{}^{a_2}\wedge\cdots\wedge\omega_n a^{n-2k}\wedge\delta[\hat{\Omega}^{a_{n-2k+1}a_{n-2k+2}}\wedge\cdots\wedge\hat{\Omega}^{a_{n-1}a_n}], \qquad \text{(A22)}$$

where we have again employed (A21) and extended the summation range to $k=0$ in the last term, where the summand is zero for this value.

Putting the above expressions together we obtain

$$\delta\Theta|_{\partial\mathscr{M}} = -\delta\left[\sum_{k=0}^{p-1}\binom{p-1}{k}\frac{2}{n-1-2k}\epsilon_{a_2\cdots a_n}\,\omega_n{}^{a_2}\wedge\cdots\wedge\omega_n{}^{a_{n-2k}}\wedge\hat{\Omega}^{a_{n-2k+1}a_{n-2k+2}}\wedge\cdots\wedge\hat{\Omega}^{a_{n-1}a_n}\right]+\hat{d}\Xi,$$

where Ξ is the expression in square brackets in the first term in (A22).

Comparing with (A19) we now have an explicit form for $\mathbf{n}^*\Pi$:

$$\mathbf{n}^*\Pi = -\sum_{k=0}^{p-1}\binom{p-1}{k}\frac{2}{n-1-2k}\hat{\Phi}_k,$$

where $\hat{\Phi}_k$ is given by (A13) with $\hat{\Omega}$ for Ω.

In order to regain Chern's expression (A6) we have to reinstate the embedding curvature Ω. Again, a simple binomial expansion and a resummation using the formula

$$\sum_{m=0}^{N}\frac{(-1)^m}{m!(N-m)!(2N-2m+1)} = (-1)^N 2^{2N+1}\frac{(N+1)!}{(2N+2)!},$$

produces the answer (if the overall numerical coefficient is restored):

$$\mathbf{n}^*\Pi = \frac{1}{\pi^p}\sum_{k=0}^{p-1}(-1)^k\frac{2^{-p-k}}{1.3...(2p-2k-1)k!}\mathbf{n}^*\Phi_k.$$

As far as this calculation goes, $\partial\mathscr{M}$ is an *arbitrary*, closed hypersurface (boundary) of $\mathscr{M}$. Thus after removal of the $\mathbf{n}^*$ and reversion to a general frame, we obtain Chern's formula (A6). In the above derivation the forms Φ_k arise naturally and there is no need for any recursion relations.

The advantage of this method over the one outlined by Spivak[32] is that it is conceptually and algebraically simpler, although a certain amount of exterior analysis is needed. Further, Spivak resorts to special case calculation to evaluate the coefficients.

The case when $\mathscr{M}$ has odd dimension n, $= m+1 = 2q+1$ can be treated in a similar fashion because general topological arguments show that the Euler number of $\mathscr{M}$ is half that of the boundary $\partial\mathscr{M}$ and we can apply the previous formulas to this even-dimensional manifold. Thus

$$\chi(\mathscr{M}) = \int_{\partial\mathscr{M}}\mathbf{n}^*\Pi = \frac{1}{2}\chi(\partial\mathscr{M}) = \frac{1}{2}\int_{\partial\mathscr{M}}\Lambda_{2q},$$

where

$$\tfrac{1}{2}\Lambda_{2q} = (1/2^n\pi_q q!)\epsilon_{a_1\cdots a_m}\hat{\Omega}^{a_1a_2}\wedge\cdots\wedge\hat{\Omega}^{a_{m-1}a_m}$$
$$= -(1/2^n\pi^q q!)\hat{\Phi}_q.$$

Eliminating $\hat{\Omega}$ in favor of Ω we find the simple binomial form

$$\mathbf{n}^*\Pi = -\frac{1}{2^n\pi^q q!}\sum_{k=0}^{q}(-1)^k\binom{q}{k}\mathbf{n}^*\Phi_k.$$

Removal of $\mathbf{n}^*$ gives the expression quoted by Chern.[26]

[1] J. S. Dowker, Phys. Rev. D **33**, 3150 (1986).

[2] J. S. Dowker, Phys. Rev. D **39**, 1235 (1989).

[3] J. S. Dowker and J. P. Schofield, Phys. Rev. D **38**, 3327 (1988).

[4] J. S. Dowker and J. P. Schofield, Nucl. Phys. B **327**, 267 (1989).

[5] M. R. Brown and A. C. Ottewill, Phys. Rev. D **21**, 2514 (1985).

[6] M. Lüscher, K. Symanzik, and P. Weisz, Nucl. Phys. B **173**, 365 (1980).

[7] O. Alvarez, Nucl. Phys. B **216**, 125 (1983).

[8] T. P. Branson and P. B. Gilkey, University of Oregon report, July 1989.

[9] L. P. Eisenhart, *Riemannian Geometry* (Princeton U. P., Princeton, NJ, 1926).

[10] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[11] A. Polyakov, Phys. Lett. B **103**, 207 (1981).

[12] M. R. Brown, J. Math. Phys. **25**, 136 (1984); R. J. Riegert, Phys. Lett. B **134**, 56 (1984); L. Bukhbinder, V. P. Gusynin, and P. I. Fomin, Yad. Phys. **44**, 828 (1986) [Sov. J. Nucl. Phys. **44**, 534 (1986)]; E. S. Fradkin and T. Tseytlin, Phys. Lett. B **134**, 187 (1984); L. Bukhbinder, S. Odintsov, and A. Shapiro, Phys. Lett. B **162**, 92 (1985).

[13] B. S. DeWitt, *Dynamical Theory of Groups and Fields* (Blackie, London, 1965).

[14] M. Berger, P. Gauduchon, and E. Mazet, *Lecture Notes in Mathematics*, Vol. 194 (Springer, Berlin, 1971).

[15] I. G. Moss, Class. Quant. Grav. **6**, 659 (1989).

[16] J. Melmed, J. Phys. A **21**, L1131 (1989).
[17] G. Kennedy, J. Phys. A **11**, L 1739 (1978).
[18] G. Kennedy, Ph. D. thesis, University of Manchester, 1979.
[19] J. S. Dowker and G. Kennedy, J. Phys. A **11**, 895 (1978).
[20] I. G. Moss and J. S. Dowker, Phys. Lett. B (in press).
[21] G. Kennedy, R. Critchley, and J. S. Dowker, Ann. Phys. **125**, 346 (1980).
[22] B. F. Whiting (private communication, 1988).
[23] H. Luckock and I. G. Moss, Class Quant. Grav., (to be published).
[24] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1969), Vol. II.
[25] S-S. Chern, Ann. Math. **45**, 747 (1944).
[26] S-S. Chern, Ann. Math. **46**, 674 (1945).
[27] S-S. Chern, *Differentiable Manifolds,* Lecture Notes (University of Chicago, Chicago, 1959).
[28] W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvature and Cohomology* (Academic, New York, 1973), Vol. II, Chap. X.
[29] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry* (Academic, New York, 1975).
[30] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge U. P., Cambridge, 1973).
[31] N. J. Hicks, *Notes on Differential Geometry* (Van Nostrand, New York, 1965).
[32] M. Spivak, *Differential Geometry* (Publish or Perish, Boston, 1975), Vol. 5, Chap. 13.
[33] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 214 (1980).
[34] P. B. Gilkey, Adv. Math. **15**, 334 (1975).
[35] H. P. McKean and I. M. Singer, J. Dif. Geom. **1**, 43 (1967).
[36] R. Bott and L. W. Tu, *Differential Forms in Algebraic Topology* (Springer, Berlin, 1982), p. 31.
[37] H. A. Buchdahl, Proc. Camb. Philos. Soc. **68**, 179 (1970).

Sectional curvature and tidal accelerations

John K. Beem
Mathematics Department, University of Missouri at Columbia, Columbia, Missouri 65211
Phillip E. Parker
Mathematics Department, Wichita State University, Wichita, Kansas 67208

(Received 28 April 1989; accepted for publication 6 December 1989)

Sectional curvature is related to tidal accelerations for small objects of nonzero rest mass. Generically, the magnification of tidal accelerations due to high speed goes as the square of the magnification of energy. However, some space-times have directions with bounded increases in tidal accelerations for relativistic speeds. These investigations also yield a characterization of null directions that fail to satisfy the generic condition used in singularity theorems. For Ricci flat four-dimensional space-times, tidally nondestructive directions are characterized as repeated principal null directions.

I. INTRODUCTION

In special relativity, objects with nonzero rest mass are restricted to speeds strictly less than c because of the unbounded increase in energy as their speed approaches c. On the other hand, these objects do not experience tidal accelerations as their speed increases. In general relativity, one has the corresponding energy increase as the speed approaches c and we show that one also has, generically, an unbounded increase in tidal accelerations. In other words, one has (generically) that high speeds have the effect of magnifying tidal forces. Geometrically, the unbounded increases in tidal accelerations correspond to the sectional curvature becoming generically unbounded near degenerate (i.e., null) sections. On the other hand, in some space-times there are certain directions such that (presumably small) objects may approach speed c in these directions and yet unbounded tidal accelerations are not experienced. The inward and outward radial directions in Schwarzschild space-time give examples of such "tidally nondestructive" directions. These nondestructive directions correspond to null directions that fail to satisfy the generic condition, but not all null directions that fail to satisfy the generic condition correspond to nondestructive directions. Null directions that fail to satisfy the generic condition are characterized in terms of indeterminate null planes. Nondestructive directions are characterized in terms of null vectors that have constant sectional curvature for all nondegenerate sections containing them. In the Ricci flat case, a direction is fully indeterminate iff it is a principal null direction. Furthermore, it is nondestructive iff it is a repeated principal null direction. In this paper all observers traverse geodesics and are thus unaccelerated. For a general investigation of accelerating observers, see Retzloff *et al.*[1,2]

There is an interesting directional paradox. A given spatial direction for a fixed frame may correspond to a nondestructive direction and yet the spatially opposite direction may be destructive. This apparent paradox is explained in terms of the fact that objects moving in opposite spatial directions have world velocity vectors that are neither parallel nor antiparallel.

On the other hand, let the Ricci flat four-dimensional space-time (M,g) have two nondestructive null directions at each point. In this case, (M,g) must be of Petrov type D and each of these two null directions is a double principal null direction. For such a space-time, one may always choose a frame at a fixed point p of M such that the corresponding spatial directions appear as direct opposites.

A number of authors have investigated problems associated with the sectional curvature near null sections. Thorpe[3] showed that sectional curvature can be continuously extended to null sections only in the case of constant curvature. Various boundedness conditions which imply constant curvature have been obtained by Kulkarni,[4] Dajczer and Nomizu,[5] Harris,[6] and Nomizu.[7] A related function called null sectional curvature has been studied by Harris[8] and Koch.[9] The values of the sectional curvature function have been investigated by Beem and Parker.[10] An extensive investigation of sectional curvature has been done by Hall,[11,12] Cormack and Hall,[13] Hall and Rendall,[14] Rendall,[15] Kulkarni,[16,17] and Ruh.[18] Among other things, these authors have investigated to what extent the curvature determines the metric. In a generic sense, the sectional curvature determines the metric and the curvature tensor determines the metric up to a constant factor. Examples show that this determination of the metric tensor is only true in a generic sense; see Yau.[19] In certain cases, there are techniques for calculating the metric from the curvature; see Quevedo.[20]

II. SECTIONAL CURVATURE AND TIDAL ACCELERATION

Let (M,g) be a space-time of signature $(+,-,\ldots,-)$ and dimension $n>2$. If u and v are tangent vectors at some point p of M, and if they span a nondegenerate two-dimensional section of T_pM, the sectional curvature K for this section is given by[21]

$$K(u,v)=\frac{g(R(u,v)v,u)}{g(u,u)g(v,v)-[g(u,v)]^2}. \tag{2.1}$$

The numerator and denominator of K are each homogeneous polynomials of degree 4 in $2n$ variables. We will denote these polynomials by P_1 and P_2, respectively:

0022-2488/90/040819-09$03.00

$$P_1(u,v) = g(R(u,v)v,u),$$

$$P_2(u,v) = g(u,u)g(v,v) - [g(u,v)]^2. \qquad (2.2)$$

The two-dimensional section determined by u and v is *null* or *degenerate* exactly when $P_2(u,v) = 0$. A null section is a *pole* if $P_1(u,v) \neq 0$ and is *indeterminate* if $P_1(u,v) = 0$. In dimension three, a null vector determines exactly one degenerate section which is then either a pole or indeterminate. In higher dimensions, a null vector determines an $(n-3)$-dimensional family of degenerate planes. For higher dimensions, we will use the following definition to help subclassify the behavior of the sectional curvature function at null directions.

Definition 2.1: Let w be a null vector. If all null planes containing w are poles, then w will be called a *complete pole.* If at least one null plane containing w is a pole and at least one is indeterminate, then w will be called *partially indeterminate.* If all planes containing w are indeterminate, then w is *fully indeterminate.*

In Sec. V we will show that a null direction is fully indeterminate iff it fails to satisfy the generic condition used in the proof of singularity theorems.

To investigate tidal forces, one starts with a unit speed timelike geodesic γ. The *Jacobi equation,*[22,23]

$$J'' + R(J,\gamma')\gamma' = 0, \qquad (2.3)$$

measures the divergence of nearby geodesics. Consequently, the Jacobi field J measures the tidal acceleration of nearby test particles.[23,24] If one takes γ' to be $E_0 = \partial/\partial x_0$ and takes a parallel orthonormal basis $E_0,...,E_{n-1}$ along the geodesic, then in local coordinates $(J^i)'' = R_{i0k0}J^k$, where J^i represents the component of J in the E_i direction. Here J is assumed to be orthogonal to the geodesic and thus $J^0 = 0$. In this paper J will denote both the vector field along the geodesic and the corresponding column with components J^i for i running from 1 to $n-1$. Letting B be the $(n-1)\times(n-1)$ matrix with (i,k) component R_{i0k0}, the Jacobi equation becomes $J'' = BJ$, where B is a real symmetric matrix which we shall call the *tidal acceleration matrix.* For a given Jacobi field J the magnitude of the tidal acceleration is

$$[-g(J'',J'')]^{1/2} = [R_{i0k0}R_{i0m0}J^kJ^m]^{1/2},$$

where the summation is over all three indices i, k, and m. The radial component of the tidal acceleration is given by

$$\begin{aligned} &-g(J'',J)/[-g(J,J)]^{1/2} \\ &= -g(-R(J,\gamma')\gamma',J)/[-g(J,J)]^{1/2} \\ &= -K(J,\gamma')[-g(J,J)]^{1/2}. \end{aligned}$$

The principal axis theorem guarantees that the eigenvalues of the tidal acceleration matrix B are real and that for fixed t there is an orthogonal matrix which diagonalizes B. Let b_i be the ith eigenvalue of B and let the corresponding unit eigenvector be denoted by e_i. If J is equal to e_i, then the tidal acceleration is $|b_i|$ and the radial tidal acceleration is b_i. Furthermore, $-b_i$ is equal to the sectional curvature $K(e_i,\gamma')$ of the timelike plane of e_i and γ'. Diagonalizing B, one obtains that for unit length Jacobi fields at a fixed point p of M, the largest tidal acceleration is equal to the maximum of $|b_i|$, where b_i are the eigenvalues of B. Furthermore, this largest value is equal to the maximum of the absolute values of the sectional curvatures of all timelike planes that contain the vector γ'. This value will be denoted by $\mathrm{Max}(\gamma')$.

Definition 2.2: If u is a timelike vector at p, then Max(u) will be the maximum of $|K(u,v)|$, where this maximum is taken over all nonzero spacelike vectors v at p.

An object of nonzero mass and nonzero volume can only withstand a limited tidal acceleration in each direction. Assume for simplicity that one has a spherical object of uniform construction and that the centroid of this object traverses a geodesic γ. If $\mathrm{Max}(\gamma')$ exceeds a certain value, then the integrity of the object is lost. We now define as nondestructive those indeterminate null directions where some nearby timelike directions have bounded values for Max(u). Here two nontrivial vectors determine the same direction at p if each is a positive scalar multiple of the other. A topology on the set of directions may be obtained by taking a positive definite auxiliary metric on M and identifying each direction with a point of the unit sphere bundle over M using this auxiliary metric.

Definition 2.3: The null vector w at p is said to define a *nondestructive* (world) *direction* if there is a continuous curve $X(t)$ of timelike vectors at p defined for all t in $[0,1)$ such that (1) the direction determined by $X(t)$ approaches the direction of w as t approaches 1, and (2) $\mathrm{Max}(X(t))$ is uniformly bounded on $[0,1)$.

The above definition requires that there exists a one parameter family of observers at p going arbitrarily close to the speed of light such that all members of this family have uniformly bounded tidal accelerations. In other words, at the point p it is possible in some theoretical sense for a small object to go arbitrarily close to c and not face unbounded tidal forces. We have stated the above definition in terms of directions since if all $X(t)$ are taken as unit vectors, then the vectors $X(t)$ cannot converge to any fixed null vector.

A null direction which fails to be nondestructive will be called *destructive.* The next proposition shows that if the null vector w determines a nondestructive direction, then the sectional curvatures of planes containing w are uniformly bounded.

Proposition 2.4: If the null vector w determines a nondestructive direction, then there is some number L such that every nondegenerate plane containing w has the absolute value of its sectional curvature less than L.

Proof: Let $X(t)$ be a continuous family of timelike vectors with the direction of $X(t)$ converging to the direction of w as $t \to 1$. Assuming that no such number L exists, then for each fixed integer N there is some vector u at p such that the plane $[w,u]$ spanned by w and u is nondegenerate and $|K(w,u)| > N$. The sectional curvature function is continuous on nondegenerate planes and thus for all t sufficiently close to one, the inequality $|K(X(t),u)| > N$ must hold. Since N was arbitrary, this implies that the null vector w determines a destructive direction that contradicts the hypothesis. ■

Proposition 2.5: Let w be a null vector at some p in (M,g) and assume there is some number L such that every nondegenerate plane containing w has the absolute value of its sectional curvature less than L. Then w must be fully indeterminate.

Proof: Assume w is not fully indeterminate. Let $P_1(w,v)\neq 0$ and $P_2(w,v) = 0$. There is some sequence of vectors $\{u_k\}$ at p, converging to v, such that the planes $[w,u_k]$ are all nondegenerate. Furthermore, $P_1(w,u_k)\to P_1(w,v)$ and $P_2(w,u_k)\to 0$ as $t\to 1$. Hence, $|K(w,u_k)|\to +\infty$, in contradiction to the hypothesis. ■

Note that Propositions 2.4 and 2.5 imply that fully indeterminate is a necessary condition for nondestructive. On the other hand, fully indeterminate is not a sufficient condition for a null direction to be nondestructive.

In order to study the generic nature of destructive directions, we assume the dimension of M is at least three and use the *coarse* C^2 topology[25] on the space $Lor(M)$ of all Lorentzian metrics on M. Note that if $E_0,...,E_{n-1}$ form an orthonormal frame at p as above, then for all metrics g' sufficiently close to g in this topology the vector E_0 will continue to be timelike. If g' is a metric near g, then we shall use $E_0(g'),...,E_{n-1}(g')$ to denote the orthonormal basis for g' obtained by using Gram–Schmidt orthogonalization on the original basis. We require that $E_0(g')$ be parallel to the original E_0 and that the plane $[E_0,E_1]$ be equal to the plane $[E_0(g'),E_1(g')]$.

The next result shows that, generically, extended objects with nonzero rest mass cannot go close to the speed of light without facing unbounded tidal accelerations. We omit the proof.

Proposition 2.6: Assume that M is a manifold of dimension at least three. Let p be a point of M and let u, v be two linearly independent tangent vectors at p. Let Z be the subset of $\mathrm{Lor}(M)$ such that u is timelike. For each g in Z let $E_0(g)$ be a unit timelike vector in the direction of u and let $E_1(g)$ be the unit spacelike vector that is orthogonal to $E_0(g)$, that lies in the plane of u and v, and that has a positive component in the v direction. Let U be the set of all g in Z such that $E_0(g) + E_1(g)$ is either a complete pole or partially indeterminate. Then U is open and dense in Z using the coarse C^2 topology.

The *Whitney* or *fine* C^2 topology is an alternative topology on $\mathrm{Lor}(M)$ which is always at least as fine as the coarse C^2 topology. We remark that Proposition 2.6 remains valid if the word coarse is replaced with the word fine. It should be mentioned that the entire space $\mathrm{Lor}(M)$ may be empty since some compact manifolds do not admit a Lorentzian metric. Thus the set Z in the last proposition is open but may also be empty.

One usually regards the sectional curvature as a ratio of two quartics. However, by changing the domain, one may regard the sectional curvature as a ratio of quadratics. Let $G_2(n)$ denote the Grassmann manifold of two-dimensional linear subspaces of R^n and let Λ^2R^n be the second exterior power of R^n. An element of Λ^2R^n is called decomposable if it is a simple product of the form $u\wedge w$. A nontrivial decomposable element $u\wedge w$ corresponds to the element of $G_2(n)$ with basis u, w. Clearly, all nonzero scalar multiples of $u\wedge w$ yield the same element of the Grassmann manifold. Furthermore, each element of $G_2(n)$ corresponds to a decomposable two-vector of Λ^2R^n and also to all nonzero scalar multiples of this two-vector. If $E_1,...,E_n$ is a basis of R^n, then $E_i\wedge E_j$ $(1\leqslant i<j\leqslant n)$ is a basis for Λ^2R^n and the dimension of this last space is $d=n(n-1)/2$. Deleting the zero two-vector and identifying the elements of Λ^2R^n that differ by a nonzero scalar multiple, we obtain the real projective space RP^{d-1} of dimension $d-1$. Thus we may regard $G_2(n)$ as a subset of RP^{d-1} via this Plücker embedding. Furthermore, the above-mentioned basis yields coordinates for Λ^2R^n and homogeneous coordinates for RP^{d-1}.

Fix a point p of M and, for convenience, an orthonormal basis $E_0,...,E_{n-1}$ at p. The tangent space T_pM is a copy of R^n and the sectional curvature at p is a function defined on the (nondegenerate) elements of $G_2(n)$. The function $g(R(u,v)v,u)$ determines a quadratic function on Λ^2R^n which will be denoted by Q_1 and the function $\Lambda^2g(u\wedge w) = g(u,u)g(v,v) - [g(u,v)]^2$ determines a second quadratic function which will be denoted by Q_2. In terms of the basis $E_\alpha\wedge E_\beta$, the form Q_1 is represented by a symmetric matrix of size $d\times d$ with elements $R_{\alpha\beta\delta\sigma}$, where the pair α, β $(\alpha<\beta)$ is thought of as a single index and the pair δ, σ $(\delta<\sigma)$ is also. Similarly, the form Q_2 is represented by a $d\times d$ symmetric matrix whose elements are the 2×2 subdeterminants of the matrix g. It follows that the sectional curvature defined on $G_2(n)$ regarded as a subset of RP^{d-1} extends to a ratio of quadratics on RP^{d-1}. We shall abuse notation slightly and use $K = Q_1/Q_2$ for this ratio on RP^{d-1} as well. The *null variety* is the projective variety $N = \{z|Q_2(z) = 0\}$ and the *homaloidal variety* is $H = \{z|Q_1(z) = 0\}$. When convenient, we may use the same letters to denote their intersections with $G_2(n)$; these intersections are the *null locus N* and *homaloidal locus H*, respectively. In the next section of this paper we consider the special case of $n=3$.

III. SOME THREE-DIMENSIONAL RESULTS AND APPLICATIONS

We now consider three-dimensional space-times. This is one dimension too low from a physical viewpoint, but it is an important dimension because results from dimension three can be used to study higher dimensions. In particular, if q is a fixed point of four-dimensional space-time M_0, and L is a three-dimensional timelike linear subspace of T_qM_0, then the behavior of the sectional curvature on the linear subspace L will be the same as that on the tangent space of some three-dimensional space-time.

Let (M,g) be a three-dimensional space-time. In Beem and Parker,[10] the planes in T_pM are represented as points of $G_2(3) = RP^2$ and the sectional curvature is concretely represented as a ratio of quadratics on RP^2. Note that the notation used in Beem and Parker[10] is for a three-dimensional space-time with x_3 timelike. In the present paper we use the x_0 direction as timelike and hence our present formulas differ slightly in appearance. Using homogeneous coordinates $[y_1:y_2:y_3]$ on RP^2, one obtains

$$K=\frac{A(y_1)^2+By_1y_2+C(y_2)^2+Dy_1y_3+Ey_2y_3+F(y_3)^2}{(y_3)^2-(y_1)^2-(y_2)^2}.$$

With our present notation, $A = R_{0202}$, $B = -2R_{0102}$, $C = R_{0101}$, $D = 2R_{2102}$, $E = 2R_{1201}$, and $F = R_{2121}$.

Letting $y_3 = 0$ be the line at infinity, one obtains

$$K=\frac{Ax^2+Bxy+Cy^2+Dx+Ey+F}{1-x^2-y^2}, \tag{3.1}$$

where $x = y_1/y_3$, $y = y_2/y_3$, and each point (x,y) represents a plane in T_pM. Let

$$Q_1(x,y)=Ax^2+Bxy+Cy^2+Dx+Ey+F$$

and

$$Q_2(x,y)=1-x^2-y^2.$$

The set of (x,y) that satisfies $Q_1(x,y) = 0$ is the homaloidal locus H and the set of (x,y) that satisfies $Q_2(x,y) = 0$ is the null locus N. Since the space-time is three-dimensional, a given null vector of the space-time determines either a complete pole or a fully indeterminate direction. The points in the (x,y) model that are common to both the null locus and the homaloidal locus represent null planes tangent to the null cone of the space-time at a fully indeterminate null direction. Note that if the homaloidal locus does not contain all of the null locus, then it follows easily that the two conics have at most four points in common. This shows that for a three-dimensional space-time each point has either all null directions nondestructive or at most four null directions nondestructive. In fact, we will prove the stronger result that for a point of three-dimensional space-time either all null directions are nondestructive or else at most two null directions are nondestructive.

In this (x,y) model, the points $x^2 + y^2 < 1$ lying inside the unit circle correspond to spacelike planes, the unit circle corresponds to null planes, and the points $x^2 + y^2 > 1$ lying outside the unit circle correspond to timelike planes. The collection of all planes containing a fixed vector v in T_pM corresponds to a line. If v is timelike, the corresponding line in the xy-plane misses the unit circle. The collection of planes in T_pM that contain the null vector $E_0 + E_1$ corresponds to the line $x= -1$ in the (x,y) model. In particular, the null plane in T_pM with basis $\{E_0 + E_1, E_2\}$ corresponds to the point $(-1,0)$. This null plane is indeterminate when $P_1(E_0 + E_1,E_2) = 0$. This yields $R_{2020} + 2R_{2120} + R_{2121} = 0$, which can also be written as $A-D+F=0$.

Lemma 3.1: The sectional curvature of all planes of T_pM containing the null vector $w = E_0 + E_1$ is bounded iff (1) $A-D+F=0$ and (2) $E-B=0$. Furthermore, if the sectional curvature of planes containing w is bounded, then all nondegenerate planes containing w have the same constant curvature $K = -C = -R_{1010}$.

Proof: The sectional curvature of these planes is given by Eq. (3.1) evaluated along the line $x= -1$. Assume first that (1) and (2) hold. Equation (3.1) shows that K is identically equal to the desired $-C$ along the line $x= -1$ and is thus also bounded. Assume now that the sectional curvature is bounded. Then $P_1(w,E_2) = 0$ implies that Eq. (1) must hold. Using (1) and $x= -1$, Eq. (3.1) takes the form

$$K=(-By+Cy^2+Ey)/-y^2. \tag{3.2}$$

Since this must be bounded for all y, one obtains the required equation (2) and $K= -C$ by letting y approach 0. ■

Proposition 3.2: The null vector $w = E_0 + E_1$ defines a nondestructive direction iff (1) $A-D+F=0$ and (2) $E-B=0$.

Proof: Assume w is nondestructive. Proposition 2.5 and Lemma 3.1 imply that (1) and (2) hold. In order to show that (1) and (2) imply w is nondestructive, it suffices to show that the tidal acceleration matrix B converges to a matrix with finite values as one takes a one parameter family of boosts in the x_0x_1 plane corresponding to observers going arbitrarily close to c in the x_1 direction. We omit the details of this calculation since they will be done for the four-dimensional case in the next section. ■

Lemma 3.1 and Proposition 3.2 yield the following result.

Theorem 3.3: Let M be a three-dimensional space-time. The following are equivalent: (a) the null vector w is nondestructive; (b) the set of values of the sectional curvature of all planes containing w is bounded; (c) the set of values of the sectional curvature of all planes containing w is constant.

We now consider the implications for the homaloidal locus of the above equations (1) and (2). Set $D=A+F$. The homaloidal locus $Q_1(x,y) = 0$ becomes

$$Ax^2+Bxy+Cy^2+(A+F)x+Ey+F=0. \tag{3.3}$$

The implicit function theorem yields that if $A\neq F$, then this locus is a manifold near $(-1,0)$. Differentiating Eq. (3.3) with respect to y and solving for dx/dy one obtains

$$\frac{dx}{dy}=\frac{-Bx-2Cy-E}{2Ax+By+A+F},$$

which yields $(dx/dy)=0$ at $(-1,0)$ if $B=E$. Thus, given $A\neq F$ and Eq. (1), the homaloidal locus is tangent to the null locus at $(-1,0)$ if Eq. (2) holds. Assume now that both (1) and (2) hold. In this case the multiplicity of the intersection of homaloidal locus and the null locus is two provided that the conics do not coincide. If $A=F$ and $B=E$, then Eq. (3.3) becomes

$$Ax^2+Bxy+Cy^2+2Ax+By+A=0,$$

which can be rewritten as

$$A(x+1)^2+By(x+1)+Cy^2=0. \tag{3.4}$$

If $C=0$, then this locus may be the entire (x,y) plane, two lines one of which is given by $x+1=0$, or the line $x+1=0$ counted twice. Notice that one again finds that if the homaloidal locus does not contain the null locus, then the multiplicity of intersection of the two loci at $(-1,0)$, is at least two.

If $C \neq 0$, then the quadratic formula may be used on Eq. (3.4) to obtain

$$y = [(x+1)(-B \pm (B^2 - 4AC)^{1/2})]/2C. \tag{3.5}$$

It follows that the homaloidal locus may be the single point $(-1,0)$, a line through $(-1,0)$ of slope $-B/2C$ counted twice, or two lines through $(-1,0)$, neither of which is tangent to the null locus at $(-1,0)$. Note that in all of the cases considered, the homaloidal locus must either be degenerate or tangent to the null locus at $(-1,0)$. Of course, any spatial direction in $T_p M$ may be rotated to point in the positive x_1 direction. This means that the geometric relation of the homaloidal locus to the null locus must be similar to one of the cases described above at all nondestructive null directions. Since none of the cases include the two conics intersecting transversally with multiplicity one, we obtain the following result.

Proposition 3.4: Let p be a fixed point of a three-dimensional space-time (M,g). Then exactly one of the following is true at p.

(1) All null directions at p are nondestructive and the space-time has constant sectional curvature at the point p.

(2) There are exactly two nondestructive null directions at p.

(3) There is exactly one nondestructive null direction at p.

(4) There are no nondestructive null directions at p.

IV. BOOSTS AND NONDESTRUCTIVE DIRECTIONS IN DIMENSION FOUR AND HIGHER

Let (M,g) be four-dimensional with local coordinates $x_0,...,x_3$ centered at a point p and assume that the natural basis is orthonormal. Denoting this frame by $E_0,...,E_3$ at the point p, one may ask if the null vector $w = E_0 + E_1$ is nondestructive. Here E_1 is the first space direction for the observer (i.e., the x_1 direction in our notation). Essentially, the question is whether there can exist (for this observer) small objects going close to speed c in spatial directions close to the x_1 direction such that these objects feel only bounded tidal accelerations. We begin by considering objects which go in *exactly* the x_1 direction with speed arbitrarily close to c. This requirement corresponds to the timelike vectors $X(t)$ of Definition 2.3 being associated with pure boosts.

Three necessary conditions for $w = E_0 + E_1$ to be nondestructive are found by using Propositions 2.4 and 2.5 to obtain $P_1(w,E_2) = 0, P_1(w,E_3) = 0,$ and $P_1(w,E_2 + E_3) = 0$. These three equations yield

$$R_{2020} + 2R_{2120} + R_{2121} = 0, \tag{4.1a}$$

$$R_{3030} + 2R_{3130} + R_{3131} = 0, \tag{4.1b}$$

$$R_{2030} + R_{2031} + R_{2130} + R_{2131} = 0. \tag{4.1c}$$

It is easy to verify that these equations are both necessary and sufficient for this w to be fully indeterminate.

Let $y_0,...,y_3$ be new local coordinates derived from a boost in the x_0x_1-plane. Let $R_{\alpha\beta\delta\sigma}$ denote the components of the curvature at p in the original x coordinates and let $\underline{R}_{\alpha\beta\delta\sigma}$ denote the curvature components at p in the y coordinates.

Define the y coordinates by $y_0 = x_0 \cosh\theta - x_1 \sinh\theta$, $y_1 = -x_0 \sinh\theta + x_1 \cosh\theta$, $y_2 = x_2$, and $y_3 = x_3$. Using $\mathrm{ch}(\theta)$ for $\cosh\theta$ and $\mathrm{sh}(\theta)$ for $\sinh\theta$, one obtains

$$\begin{aligned}
\underline{R}_{1010} &= R_{1010},\\
\underline{R}_{2020} &= \mathrm{ch}^2(\theta)R_{2020} + 2\,\mathrm{ch}(\theta)\mathrm{sh}(\theta)R_{2021}\\
&\quad + \mathrm{sh}^2(\theta)R_{2121},\\
\underline{R}_{1020} &= \mathrm{ch}(\theta)R_{1020} + \mathrm{sh}(\theta)R_{1021},\\
\underline{R}_{1030} &= \mathrm{ch}(\theta)R_{1030} + \mathrm{sh}(\theta)R_{1031},\\
\underline{R}_{2030} &= \mathrm{ch}^2(\theta)R_{2030} + \mathrm{ch}(\theta)\mathrm{sh}(\theta)R_{2031}\\
&\quad + \mathrm{ch}(\theta)\mathrm{sh}(\theta)R_{2130} + \mathrm{sh}^2(\theta)R_{2131},\\
\underline{R}_{3030} &= \mathrm{ch}^2(\theta)R_{3030} + 2\,\mathrm{ch}(\theta)\mathrm{sh}(\theta)R_{3031}\\
&\quad + \mathrm{sh}^2(\theta)R_{3131}.
\end{aligned} \tag{4.2}$$

Letting $X(\theta) = (\cosh\theta)E_0 + (\sinh\theta)\mathrm{E}_1$, the direction of $X(\theta)$ converges to the direction of w as $\theta \to +\infty$. Furthermore, for fixed θ the mapping given by the above boost takes $X(\theta)$ into the unit vector in the y_0 direction. In order for the sectional curvature of planes containing $X(\theta)$ to be uniformly bounded as $\theta \to +\infty$, it is sufficient that the elements of $\underline{R}_{i0j0}$ converge to finite limits as $\theta \to +\infty$. Assuming these limits exist, one may obtain the following equations from $\underline{R}_{1020}$ and $\underline{R}_{1030}$:

$$R_{1020} + R_{1021} = 0, \tag{4.3a}$$

$$R_{1030} + R_{1031} = 0. \tag{4.3b}$$

Lemma 4.1: If all nondegenerate planes containing $w = E_0 + E_1$ have the same sectional curvature, then all of the equations of (4.1) and (4.3) are satisfied.

Proof: Propositions 2.4 and 2.5 yield the three equations of (4.1). The equations of (4.3) follow from Lemma 3.1, applied to the three-dimensional slices corresponding to $x_3 = 0$ and $x_2 = 0$. ■

It follows easily that the equations of (4.1) and (4.3) must be satisfied if $E_0 + E_1$ is nondestructive. Furthermore, the equations of (4.1) and (4.3) together imply that the $\underline{R}_{i0j0}$ matrix converges to a limiting matrix as $\theta \to +\infty$. We have thus proven that the five equations given in (4.1) and (4.3) are necessary and sufficient for $E_0 + E_1$ to be nondestructive. Using Lemma 3.1, one finds these equations imply that all nondegenerate planes containing w have constant sectional curvature $-R_{1010}$.

Proposition 4.2: Let p be a fixed point of the four-dimensional space-time (M,g) and let the curvature tensor be given with respect to an orthonormal basis E_α at p. If $w = E_0 + E_1$, then the following are equivalent:

(a) the null vector w is nondestructive;

(b) there is a number L such that the absolute value of the sectional curvature of every nondegenerate plane containing w is less than L;

(c) every nondegenerate plane containing the null vector w has sectional curvature given by the constant $-R_{1010}$;

(d) the above matrix $\underline{R}_{i0j0}$ converges to a limit as $\theta \to +\infty$;

(e) Eqs. (4.1a), (4.1b), (4.1c), (4.3a), and (4.3b) hold at p.

Physically, the above conditions are necessary and sufficient for very small objects going along a geodesic in the positive x_1-direction to have bounded tidal accelerations at p no matter how close to the speed of light they are traveling. Corresponding conditions may be obtained for objects going in the opposite spatial direction by similar methods.

An interesting observation is the following "directional paradox." The required conditions for being able to go close to c in the positive x_1 direction and to have bounded tidal accelerations are different from the requirements that must be satisfied to be able to go in the negative x_1 direction and to have bounded tidal accelerations. There is an "apparent" lack of symmetry. Of course, the reason for the lack of symmetry is that one must consider the world vectors associated with moving objects and not just the spatial components of these vectors. The world vectors for objects going fast in the respective positive and negative x_1 directions are far from being either parallel or antiparallel.

In order to get a physical interpretation of Eqs. (4.1)–(4.3), let $v=\beta=\tanh\theta$, $(1-v^2)^{-1/2}=\gamma=\cosh\theta$ and let $\theta\to+\infty$. The classical special relativistic magnification of energy associated with the moving mass is $\gamma=\cosh\theta$.

Remark 4.3: Let $w = E_0 + E_1$.

(a) If w is either a complete pole or partially indeterminate, then at least one of the equations of (4.1) fails and from (4.2), it follows that the "high-speed" magnification of tidal acceleration is roughly proportional to γ^2, which is the square of the magnification of energy.

(b) If w is fully indeterminate, but destructive, then the equations of (4.1) hold and it is one of the equations of (4.3) that fails. In this case, the magnification of tidal acceleration is roughly proportional to γ which is the same as that of the energy.

(c) If w is nondestructive, then one has magnification of energy but no unbounded increase in tidal accelerations. Thus, one does not get a corresponding magnification of tidal acceleration.

We obtain the following sufficient condition for a null direction to be nondestructive. We say that two regular varieties intersect *tangentially* at a point iff their tangent hyperplanes coincide at that point.

Theorem 4.4: Let p be a fixed point of a space-time M of dimension at least three. Every tangential intersection of the null and homaloidal loci in $G_2(T_pM)$ determines a nondestructive null direction in T_pM.

Proof: First, compare the list of cases for nondestructive null directions in dimension three given in Sec. III to the complete list of cases[10] for $H\cap N$ to verify the result for dimension three. Then observe that every three-dimensional slice of a tangential intersection is a tangential intersection. ■

Note that the results of Sec. III show that even for three-dimensional space-times, there are nondestructive null directions that correspond to nontangential intersections of $H\cap N$. Thus a tangential intersection fails to be a necessary requirement for the corresponding null vector to be nondestructive.

V. THE GENERIC CONDITION AND PRINCIPAL NULL DIRECTIONS

In this section we relate our results to the generic condition[24] and principal null directions[26] in four-dimensional space-times. An inextendible (null) geodesic γ: $(a,b)\to M$ with tangent vector W is said to satisfy the generic condition if there is some parameter value t such that $W^cW^dW_{[a}R_{b]cd[e}W_{f]}\neq 0$ at the point $\gamma(t)$. It will be satisfied in the Ricci flat case if the null geodesic γ contains some point where the Weyl tensor C_{abcd} is not trivial and W does not lie in one of the so-called principal null directions (there are at most four such). Fix a point p in M and let W be a null vector at p. We will say that W is *nongeneric* if $W^cW^dW_{[a}R_{b]cd[e}W_{f]} = 0$.

Starting with the null vector W, extend to a pseudo-orthonormal basis W, N, E_2, E_3 of T_pM. With our sign convention, this yields $g(W,N)=1$ and $g(W,W)=g(N,N)=g(W,E_i)=G(N,E_i)=0$ for $i=2,3$. Let K_{abcd} denote the components of the curvature tensor in this basis. Clearly, the contravariant components W are given by $W^1 = W^2 = W^3 = 0$ and $W^0 = 1$. The covariant components of W are given by $W_0 = W_2 = W_4 = 0$ and $W_1 = 1$. The nongeneric condition for W is then $W^cW^dW_{[a}K_{b]cd[e}W_{f]} = 0$. In these coordinates, this equation simplifies to

$$W_{[a}K_{b]00[e}W_{f]}=0. \tag{5.1}$$

The indices a, b, e, f can each take on the values from 0 to 3. A close inspection of Eq. (5.1) together with standard curvature identities yields the following result.[22]

Lemma 5.1: The null vector W is nongeneric iff $K_{b00e} = 0$ for all $2\leqslant b, e\leqslant 3$, where K_{abcd} are the components of the curvature tensor calculated in the pseudo-orthonormal frame W, N, E_2, E_3.

In order to relate the result of Lemma 5.1 to the three equations of (4.1), let E_0, E_1, E_2, E_3 be an orthonormal frame at the fixed point p and assume $W = E_0 + E_1$, $N = (E_0 - E_1)/2$. Let R_{abcd} represent the components of the curvature in the orthonormal frame.

Proposition 5.2: The null vector $W = E_0 + E_1$ is nongeneric iff Eqs. (4.1a), (4.1b), and (4.1c) all hold at p.

Proof: The equation $K_{2002} = 0$ is equivalent to the equation

$$g(E_2, R(W,E_2)W)$$
$$=g(E_2, R(E_0+E_1,E_2)(E_0+E_1))=0.$$

Using the multilinearity properties of the metric g and curvature R as well as curvature identities, one finds that $K_{2002} = 0$ is equivalent to Eq. (4.1a). Similarly, one finds that $K_{3003} = 0$ and $K_{2003} = 0$ are equivalent to Eqs. (4.1b) and (4.1c), respectively. ■

This proposition yields the following two corollaries.

Corollary 5.3: A null direction at a fixed point p fails to satisfy the generic condition iff it is a fully indeterminate direction.

Corollary 5.4: If W determines a nondestructive null direction, then the generic condition fails for W.

Assume now that (M,g) is Ricci flat. In this case, the curvature tensor R_{abcd} is equal to the Weyl tensor C_{abcd}. We now consider a null tetrad k, l, m, $\overline{m}$ defined by $k = (E_0 + E_1)/\sqrt{2}$, $l = (E_0 - E_1)/\sqrt{2}$, $m = (E_2 - iE_3)/\sqrt{2}$, and $\overline{m} = (E_2 - iE_3)/\sqrt{2}$. Note that $k^0 = l^0 = 1/\sqrt{2}$, $k^1 = -l^1 = 1/\sqrt{2}$, $m^2 = 1/\sqrt{2}$, $m^3 = -i/\sqrt{2}$. Define two complex coefficients Ψ_0, Ψ_1 by

$$\Psi_0 = C_{abcd}k^am^bk^cm^d$$

and

$$\Psi_1 = C_{abcd}\,k^al^bk^cm^d.$$

The null direction determined by k is a *principal null direction*[26] iff $\Psi_0 = 0$. For (M,g) Ricci flat, this holds if the three equations (4.1a), (4.1b), and (4.1c) are satisfied substituting C_{abcd} for R_{abcd}. The null direction determined by k is a *repeated null direction* iff $\Psi_1 = 0$ and $\Psi_0 = 0$. Using the above k, l, m, and curvature identities, the real part of $\Psi_1 = 0$ yields

$$C_{1002} + C_{1012} = 0. \qquad (5.2)$$

The imaginary part of $\Psi_1 = 0$ yields

$$C_{0103} + C_{0113} = 0. \qquad (5.3)$$

Note that for Ricci flat space-times, Eqs. (5.2) and (5.3) are equivalent to Eqs. (4.3a) and (4.3b), respectively. We have thus established the following theorem.

Theorem 5.5: Let (M,g) be a four-dimensional Lorentzian manifold and assume that (M,g) is Ricci flat. Let p be a point of M and let k be a null vector at p.

(a) The null vector k is fully indeterminate iff it lies in a principal null direction.

(b) The null vector k is nondestructive iff it is a repeated principal null direction.

(c) If (M,g) is not flat, there are at most four fully indeterminate directions at p.

(d) If (M,g) is not flat, there are at most two nondestructive directions at p.

Note that if (M,g) is Ricci flat with exactly two nondestructive directions at each point [i.e., case (d) of Theorem 5.5 holds], then (M,g) is of Petrov type D.

VI. SCHWARZSCHILD SPACE-TIME

In this section we shall give a short discussion of how the results of this paper relate to Schwarzschild space-time. Using "static" coordinates, the exterior Schwarzschild metric is given for signature $(-,+,+,+)$ by

$$ds^2 = -(1-2M/r)dt^2 + (1-2M/r)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2).$$

Using the standard orthonormal basis

$$E_0 = \left(1 - \frac{2M}{r}\right)^{-1/2}\frac{\partial}{\partial t}, \quad E_1 = \left(1 - \frac{2M}{r}\right)^{1/2}\frac{\partial}{\partial r},$$

$$E_2 = r^{-1}\frac{\partial}{\partial \theta}, \quad E_3 = (r\sin\theta)^{-1}\frac{\partial}{\partial \phi},$$

the nonzero components of the curvature tensor all may be obtained from the following equations using symmetry properties of the curvature tensor[23,27]:

$$R_{0101} = -2M/r^3, \quad R_{0202} = R_{0303} = M/r^3,$$

$$R_{2323} = 2M/r^3, \quad R_{1212} = R_{1313} = -M/r^3. \qquad (6.1)$$

Assume $\gamma' = E_0$ and that there is Jacobi field of length $|J|$. It follows easily,[23] that for $J = |J|E_1$ there is a stretching acceleration of magnitude $(2|J|M)/r^3$. For $J = |J|E_2$ and for $J = |J|E_3$, one obtains compressions of magnitude $(|J|M)/r^3$. On the other hand, consider a freely falling observer moving at time $t=0$ in the ϕ direction. Assume that in the above orthonormal frame the moving observer has speed $v = \tanh A$. Thus $(1 - v^2)^{-1/2} = \cosh A$. The world velocity vector of the moving observer becomes $\gamma'(0) = \cosh AE_0 + \sinh A\,E_3$. Let the Jacobi field at $t=0$ be $J = |J|E_1$ in order to measure the tidal acceleration in the radial direction, which in this case is orthogonal to the direction of motion. Using

$$\begin{aligned} J'' &= -R(J,\gamma')\gamma' \\ &= -R(|J|E_1, \cosh A\,E_0 + \sinh A\,E_3) \\ &\quad \times (\cosh A\,E_0 + \sinh A\,E_3), \end{aligned}$$

one finds that

$$J'' = |J|(2M\cosh^2 A + M\sinh^2 A)r^{-3}E_1. \qquad (6.2)$$

Consequently, it is clear that as $A \to +\infty$ (i.e., v approaches c) the tidal acceleration becomes unbounded in the r direction.

Consider a steel ball of radius b that traverses a nonradial geodesic in this model. This ball will have a closest approach to the central mass M. Let r_0 denote the value of r at the closest approach. Equation (6.2) shows that there is some speed (measured by static observers) such that the ball will be destroyed if it is going this speed or faster when $r = r_0$. Clearly, the speed corresponding to destruction depends on the radius b, the tensile strength of the ball and the value of r_0. Notice that no matter how large the value of r_0, there will be some speed yielding certain destruction in this model. It should be mentioned that the tidal accelerations build as the r coordinate of this geodesic approaches its smallest value r_0 and hence this effect does not correspond to tidal accelerations that just act for short time periods. Note in this regard that the classical special relativistic time dilation of the moving particle is given by $(\cosh A)^{-1}$ and that the acceleration J'' in Eq. (6.2) has factors corresponding to $\cosh^2 A$ and $\sinh^2 A$. Thus for the situation described by Eq. (6.2), the "high-speed" magnification of tidal acceleration roughly goes as the square of the increase in energy as the square of the reciprocal of the time dilation. This is a consequence of $E_0 + E_3$ being a complete pole and not just a reflection of the nature of Schwarzschild space-time.

The radial directions (both inwards and outwards) are nondestructive in the Schwarzschild space-time. It is well known[23] that for a fixed point of this space-time, the tidal accelerations are independent of speed when the direction of motion is radial. This means that for a steel ball moving along a radial geodesic, the tidal accelerations calculated at the centroid are the same as those for a similar steel ball at rest in the "static" frame. Tidal forces and radiation for a radially falling body in Schwarzschild space-time have been studied by Mashhoon.[28]

To view the sectional curvature of Schwarzschild space-time in terms of RP^5, use the above orthonormal basis $E_0,...,E_3$. Relative to this basis the metric tensor is $g=\mathrm{diag}(-1,1,1,1)$. Thus, $\Lambda^2 g=\mathrm{diag}(-1,-1,-1,1,1,1)$. At any point $(t_0,r_0,\theta_0,\phi_0)$,

$$Q_1=M(r_0)^{-3}[-2p_{01}{}^2+p_{02}{}^2+p_{03}{}^2 -p_{12}{}^2-p_{13}{}^2+2p_{23}{}^2],$$

$$Q_2=-p_{01}{}^2-p_{02}{}^2-p_{03}{}^2+p_{12}{}^2+p_{13}{}^2+p_{23}{}^2$$

in Plücker coordinates. We recall that these are homogeneous coordinates in RP^5 in which $G_2(4)$ is given by $p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12} = 0$. Now the variety $N = \{p\,|\,Q_2(p) = 0\}$ and similarly $H = \{p\,|\,Q_1(p) = 0\}$.

It follows that

$$H=\{p\,|\,p_{02}{}^2+p_{03}{}^2-p_{12}{}^2-p_{13}{}^2=2(p_{01}{}^2-p_{23}{}^2)\},$$

$$N=\{p\,|\,p_{02}{}^2+p_{03}{}^2-p_{12}{}^2-p_{13}{}^2=-(p_{01}{}^2-p_{23}{}^2)\}.$$

Thus at $H\cap N, p_{01}=\pm p_{23}$.

In order that this intersection be tangential, we must have

$$\nabla Q_1=[-2p_{01}:p_{02}:p_{03}:-p_{12}:-p_{13}:2p_{23}],$$

$$\nabla Q_2=[-p_{01}:-p_{02}:-p_{03}:p_{12}:p_{13}:p_{23}]$$

parallel. Now this happens iff either $p_{01} = p_{23} = 0$ or $p_{02} = p_{03} = p_{12} = p_{13} = 0$. Substituting the latter into the equation for $G_2(4)$, we obtain $p_{01} = p_{23} = 0$ in $H\cap N$, which is not possible with homogeneous coordinates. Therefore we must have $p_{01} = p_{23} = 0$ and at least one of the other four nonzero.

Using the symmetries of the equations, we may as well assume that $p_{13}\neq 0$. Letting $x = p_{02}/p_{13}$, $y = p_{03}/p_{13}$, and $z = p_{12}/p_{13}$, we obtain

$$H\cap N=\{(x,y,z)\,|\,x^2+y^2-z^2=1\},$$

$$G_2(4)=\{(x,y,z)\,|\,x=yz\},$$

in this affine R^3 in the copy of RP^3 given by $p_{01} = p_{23} = 0$ in RP^5.

Now, solve $x=yz$ successively for each variable and substitute into $x^2 + y^2 - z^2 = 1$ to obtain the equations for the tangential part of $H\cap N$ in $G_2(4)$:

$$y=\pm 1, \quad x=\pm z.$$

Therefore, we have two pencils of planes,

$$p_{03}=\pm p_{13}, p_{01}=p_{02}=p_{12}=p_{23}=0,$$

each containing a common null vector: $(1, \pm 1, 0, 0)$, respectively. These are precisely the inward and outward radial directions at each space-time point $(t_0,r_0,\theta_0,\phi_0)$. One may easily verify that there are no other nondestructive null directions.

VII. CONCLUSIONS

In this paper tidal accelerations for freely falling objects of nonzero volume and nonzero mass are considered. One assumes the centroid of the object traverses a timelike geodesic and studies the tidal accelerations using Jacobi fields. It makes sense to fix a point of the geodesic and consider the tidal accelerations associated with Jacobi fields having given values at that point.

We find that, generically, high speeds magnify tidal accelerations in an unbounded fashion. If one investigates tidal accelerations by boosting towards a fixed null direction, then generically the magnification of tidal accelerations due to high speed goes as the square of the magnification of energy. Since one can never know the true physical metric tensor exactly, this suggests that at high speeds the objects under consideration must be destroyed. On the other hand, the spaces of constant (sectional) curvature show that for at least some models these unbounded tidal accelerations need not occur. Furthermore, for some space-times that do not have constant sectional curvature (e.g., Schwarzschild space-time) there are certain directions that are tidally nondestructive. A null direction is shown to be nondestructive iff there is some constant K_0 such that all nondegenerate planes containing this null direction have sectional curvature K_0. It should be remembered though that, in addition to being not physically able to know the metric tensor exactly, one cannot physically determine spatial directions exactly, either.

A natural question is how the results of this paper are related to waves in general relativity. Classically, one has the waves traversing null geodesics. This has the important implication that they cannot have unit speed parametrizations, but must have affine parametrizations that are independent of arc length. Thus sectional curvature does not have the same interpretation for zero rest mass objects traversing null geodesics as for nonzero mass objects traversing timelike geodesics. Mashhoon[29] has considered waves in general relativity and the difficulties with assuming that they traverse null geodesics. The basic problem is that if null geodesics are converging or diverging, then how do waves maintain their integrity? They should tend to spread or contract in the presence of gravitational fields and this tendency should be larger for longer wave lengths and higher gravitational fields (i.e., space-times with larger curvature).

The relation of sectional curvature and wave surfaces has been investigated by Hall and Rendall[14] and by Hall.[30] Using the results of Hall,[30] one may obtain an alternative derivation of parts (a) and (b) of Theorem 5.5.

ACKNOWLEDGMENTS

The authors wish to thank Malcolm MacCallum for some valuable suggestions and Graham Hall for some important advice and conversations. One of the authors (J.B.) was supported in part by NSF grant DMS-8803511.

[1] D. G. Retzloff, B. DeFacio, and P. Dennis, J. Math. Phys. **21**, 96 (1982).

[2] D. G. Retzloff, B. DeFacio, and P. Dennis, J. Math. Phys. **21**, 105 (1982).

[3] J. A. Thorpe, J. Math. Phys. **10**, 1 (1969).

[4] R. S. Kulkarni, Comment. Math. Helv. **54**, 173 (1979).

[5] M. Dajczer and K. Nomizu, Math. Ann. **247**, 279 (1980).

[6] S. G. Harris, Indiana Univ. Math. J. **31**, 289 (1982).

[7] K. Nomizu, Proc. Am. Math. Soc. **89**, 473 (1983).

[8] S. G. Harris, Gen. Relativ. Gravit. **17**, 493 (1985).

[9] L. J. Koch-Sen, J. Math. Phys. **26**, 407 (1985).

[10] J. K. Beem and P. E. Parker, Comment. Math. Helv. **59**, 319 (1984).

[11] G. S. Hall, Gen. Relativ. Gravit. **16**, 495 (1984).

[12] G. S. Hall, Gen. Relativ. Gravit. **16**, 79 (1984).

[13] W. J. Cormack and G. S. Hall, Int. J. Theor. Phys. **18**, 279 (1979).

[14] G. S. Hall and A. D. Rendall, Gen. Relativ. Gravit. **19**, 771 (1987).

[15] A. D. Rendall, J. Math. Phys. **29**, 1569 (1988).

[16] R. S. Kulkarni, Ann. Math. **91**, 311 (1970).

[17] R. S. Kulkarni, Int. J. Math. Math. Sci. **1**, 137 (1978).

[18] B. Ruh, Math. Z . **189**, 371 (1985).

[19] S. T. Yau, Ann. Math. **100**, 121 (1974).

[20] H. Quevedo, "Determination of the metric from the curvature" (to be published).

[21] J. A. Wolf, *Spaces of Constant Curvature* (Publish or Perish, Boston, 1974), 3rd ed.

[22] J. K. Beem and P. E. Ehrlich, "Global Lorentzian geometry," *Pure and Applied Mathematics*, Vol. 67 (Dekker, New York, 1981).

[23] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[24] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime* (Cambridge U. P., Cambridge, 1973).

[25] J. K. Beem and P. E. Parker, Pac. J. Math. **116**, 11 (1985).

[26] D. Kramer, H. Stephani, E. Hertl, M. MacCallum, and E. Schmutzer, *Exact Solutions of Einstein's Field Equations* (Cambridge U. P., Cambridge, 1980).

[27] B. A. O'Neill, "Semi-Riemannian geometry," *Pure and Applied Mathematics*, Vol. 102 (Academic, New York, 1983).

[28] B. Mashhoon, Astrophys. J. **216**, 591 (1977).

[29] B. Mashhoon, Phys. Lett. A **122**, 299 (1987).

[30] G. S. Hall, Z. Naturforsch. A **33**, 559 (1978).

Block-diagonalization in second quantization

A. Tarantelli and L. S. Cederbaum
Theoretische Chemie, Physikalisch-Chemisches Institut, Universität Heidelberg, Im Neuenheimer Feld 253, D-6900 Heidelberg, West Germany

(Received 5 July 1989; accepted for publication 6 December 1989)

The unitary matrix that brings a Hermitian matrix **H** into block-diagonal form can be uniquely determined under very simple and transparent conditions. In this work the block-diagonalization problem is investigated in the framework of the second quantization formalism. Starting with an operator $\hat{H}$ which in any n-particle Fock space has a well-defined matrix representation an attempt was made to answer the question whether the transformation matrices **T** which can be separately given in the various n-particle spaces can be considered as different matrix representations of the same operator $\hat{T}$. Interestingly, the very important result was reached that the block-diagonalization operator $\hat{T}$ exists and is unique. As a particular example, attention was concentrated on the case of an operator $\hat{H}$ given by a one-particle operator. In this case the block-diagonalization operator can be constructed and given in explicit form. This approach is applied to the theory of Green's function where the block-diagonalization of the Hamiltonian has interesting consequences that are illustrated in some details.

I. INTRODUCTION

A Hermitian matrix **H** can be block-diagonalized, i.e., transformed into a block-diagonal matrix by a unitary transformation

$$\tilde{\mathbf{H}} = \mathbf{T}^{\dagger}\mathbf{H}\mathbf{T}. \tag{1.1}$$

The block-diagonal matrix $\tilde{\mathbf{H}}$ consists of square matrices (blocks) along its diagonal and is zero elsewhere.

There are several physical problems that involve the block-diagonalization of Hermitian matrices. Some of these are discussed in Ref. 1 and include examples from quasidegenerate perturbation theory[2,3] and many-body Green's functions.[4-6] The block structure of the resulting block-diagonal matrix, i.e., the dimensions and characterization of the diagonal blocks (see also Sec. II), is usually determined by the problem under investigation and by the physical situation at hand.

In general, there are infinitely many different unitary transformation **T** that block-diagonalize **H** for a given block structure. All of them can be cast into the form[1]

$$\mathbf{T} = \mathbf{S}\mathbf{F}, \tag{1.2}$$

where **S** is the eigenvector matrix of **H** and **F** is any unitary block-diagonal matrix with the same block structure as $\tilde{\mathbf{H}}$. In Ref. 1 it has been investigated under which elementary conditions the block-diagonalization becomes unique. By an elementary condition we mean a condition that is simple and convincing so that it may be considered a "must" in realistic applications. Two different conditions have been found that lead to the <u>*same*</u> unique **F** and thus **T**. It is this unique transformation matrix **T** that plays the central role in the present paper. Whenever unambiguous we shall use in the following the term "block-diagonalization" as synonymous to block-diagonalization with the unique **T**.

In many applications **H** will be the matrix obtained by representing the Hamiltonian $\hat{H}$ of the systems under investigation in a basis $\{|\phi_q\rangle\}$. In these and in other cases Eq. (1.1) uniquely gives rise to the operator relation

$$\hat{\tilde{H}} = \hat{T}^{\dagger}\hat{H}\hat{T} \tag{1.3}$$

once a basis $\{\phi_q\}$ is defined. The block-diagonalization problem addressed in Eq. (1.1) applies to a system with a fixed number n of particles. If n is varied another matrix equation arises (**H** changes with n). On the other hand, it is of interest to simultaneously study the same or another system with different number of particles. This goal can be achieved if we start from Eq. (1.3) and work within the formalism of second quantization.[4-7] This powerful quantum mechanical formalism allows us to write operators (observables) in a form that is independent of the number of particles contained in the system. In the present work we investigate an operator $\hat{T}$ that block-diagonalizes $\hat{H}$ for all numbers n. With this terminology we understand that the operator $\hat{T}$ gives rise in any n-particle space to a matrix **T** that block-diagonalizes the matrix representation of the operator $\hat{H}$ in the same n-particle space representation. The existence of such a general operator $\hat{T}$ will give the block-diagonalization a more general meaning and may lead to a wider field of application.

To illustrate our theory we will then focus our attention to the simple problem where the matrix **H** that has to be block-diagonalized is the matrix representation of a so called one-particle operator $\hat{H}$ describing, for instance, particles in an external field. In this case the "block-diagonalization operator" $\hat{T}$ can be written down in explicit form and its existence and uniqueness can be directly proven.

We apply the block-diagonalization procedure to the theory of Green's functions.[4-7] We have shown elsewhere[8] that the block-diagonalization of the Hamiltonian matrix in a suitable basis of configuration functions assumes in the case of a one-particle operator $\hat{H}$ a very interesting and simple physical interpretation. We will investigate this

0022-2488/90/040828-10$03.00

problem further and derive an explicit form for the transformation matrix **T**. The matrix elements of **T** can be obtained by means of a recursive procedure that determines the transformation matrix uniquely.

II. PRELIMINARIES AND DEFINITIONS

In order to define an operator in second quantization it is first necessary to introduce a set of creation and annihilation operators $a_\alpha^\dagger$ and a_α, respectively, related to some suitable basis of one-particle states $\{|\varphi_\alpha\rangle\}$. The set $\{|\varphi_\alpha\rangle\}$ is an orthonormal complete set of states given by $|\varphi_\alpha\rangle = a_\alpha^\dagger|\text{vac}\rangle$, where $|\text{vac}\rangle$ represents the vacuum configuration. An operator $\hat{H}$, as for example, the Hamiltonian of a physical system, is usually described in this formalism as a sum of two terms:

$$\hat{H}=\hat{H}_0+\hat{H}_I, \tag{2.1}$$

where $\hat{H}_0$ is a one-particle operator diagonal in the one-particle basis $\{|\varphi_\alpha\rangle\}$

$$\hat{H}_0=\sum_\alpha \epsilon_\alpha a_\alpha^\dagger a_\alpha. \tag{2.2}$$

The second term $\hat{H}_I$ is in general composed of a nondiagonal one-particle and two-particle operators. For applications to electronic systems, $\hat{H}_0$ is often naturally chosen as the Hartree–Fock operator, and the set $\{|\varphi_\alpha\rangle\}$ is composed of the one-particle states deriving from a self-consistent energy calculation on the ground state of the system.

The eigenstates of $\hat{H}_0$ in any n-particle Fock space are described by single determinants $|\phi_q^n\rangle$ build up from n one-particle states of the set $\{|\varphi_\alpha\rangle\}$:

$$|\phi_q^n\rangle=a_{q_n}^\dagger a_{q_{n-1}}^\dagger\cdots a_{q_1}^\dagger|\text{vac}\rangle. \tag{2.3}$$

The set $\{|\phi_q^n\rangle\}$ is an orthonormal complete set of states in the n-particle space. The functions ϕ_q^n will be referred to as the n-particle configuration functions.

Now we introduce a "reference space" that is a particular N-particle space and denote by a "reference function" a particular one-determinant function ϕ_0^N belonging to this space:

$$|\phi_0^N\rangle=a_{0_N}^\dagger a_{0_{N-1}}^\dagger\cdots a_{0_1}^\dagger|\text{vac}\rangle. \tag{2.4}$$

The state $|\phi_0^N\rangle$ could be, for example, the N-particle Hartree–Fock ground state of the system. Accordingly, we divide the set of the one-particle states $\{|\varphi_\alpha\rangle\}$ into the two subsets $\{|\varphi_\alpha\rangle\}_\in$ and $\{|\varphi_\alpha\rangle\}_\notin$. The subset $\{|\varphi_\alpha\rangle\}_\in$ contains the N one-particle states that are needed to construct $|\phi_0^N\rangle$. These one-particle states and the corresponding creation and annihilation operators will be referred to as "occupied" states and operators, respectively. The subset $\{|\varphi_\alpha\rangle\}_\notin$ is the complementary set to $\{|\varphi_\alpha\rangle\}_\in$ and contains the "unoccupied" one-particle states. The greek letters are used throughout this work to label the one-particle states of the set $\{|\varphi_\alpha\rangle\}$ and the related creation and annihilation operators. If necessary, we will specify explicitly whether we are referring to occupied or unoccupied one-particle states and whenever unambiguous use for the former the subscripts $i,j,k,l,\ldots$ and for the latter the subscripts $u,v,w,\ldots$.

In any n-particle space the eigenstates $\{|\phi_q^n\rangle\}$ of the operator $\hat{H}_0$ can be now classified according to the number of unoccupied creation operators necessary to construct the eigenstate. Hence, in any n-particle space the set $\{|\phi_q^n\rangle\}$ can be subdivided as follows:

$$\{|\phi_q^n\rangle\}=\{|\phi_q^n\rangle\}_{(0)}\cup\{|\phi_q^n\rangle\}_{(1)}\cup\cdots\cup\{|\phi_q^n\rangle\}_{(n)}. \tag{2.5}$$

The subset $\{|\phi_q^n\rangle\}_{(r)}$ contains the n-particle configuration states which can be represented by a single determinant in which r one-particle states are unoccupied (i.e., $s=n-r$ are occupied). We will refer to the subset $\{|\phi_q^n\rangle\}_{(r)}$ as to the "configuration class r" of the n-particle space. The configuration states of this class can also be classified in the hole (h)–particle (p) notation with respect to the reference state $|\phi_0^N\rangle$ as $(N-s)h-rp$ configurations. For example, in the $(N-1)$-particle space the class $r=0,1,2,\ldots$ contain the $1h,2h-1p,3h-2p,\ldots$ configuration states. This $h-p$ notation is less useful in the present context. In our notation a configuration state is thus characterized by the two parameters n and r, where n specifies the n-particle space and r the class to which the state belongs. Since we do not always need to work with a specific configuration state, but only to know its configuration class, we will denote a configuration state of the class $\{|\phi_q^n\rangle\}_{(r)}$ simply by $|n,r\rangle$. In each n-particle space we have $n+1$ configuration classes r, with r running from 0 to n. It is important to remark that in the N-particle space, which is our reference space, the class 0 is composed of the N-particle reference state $|\phi_0^N\rangle$ [Eq. (2.4)] only. The state $|\phi_q^N\rangle$ is indeed the only configuration state that builds up a class by itself, being the only state that can be constructed with the N occupied one-particle states only.

The simplest example of this classification is given by the set of one-particle configuration states $\{|\phi_q^1\rangle\}$. This set splits into the two possible classes 0 and 1 that obviously coincide with the subsets $\{|\varphi_\alpha\rangle\}_\in$ and $\{|\varphi_\alpha\rangle\}_\notin$ of the one-particle states, respectively.

The general classification introduced above will be used throughout this work. The number N that characterizes our reference space and is used for the definition of the configuration classes is in principle completely arbitrary. However, the choice of a particular N-particle space as reference space may arise naturally in the investigation of the physical problems at hand. For instance, in studying processes that change the number of particles in the target system, it seems reasonable to choose N to be the number of particles in the initial system.

For Hamiltonian operators $\hat{H}$ one usually works with a matrix representation **H** in the basis of the configuration functions, i.e., the eigenfunctions of $\hat{H}_0$. Throughout this work we make use of the configuration classes that give to the matrix **H** a well-defined block structure consistent in all n-particle space. In any n-particle space the matrix representation **H** of $\hat{H}$ can be viewed as a $(n+1)\times(n+1)$ block matrix composed of $(n+1)$ diagonal blocks and of coupling blocks among the configuration classes. Each row and column of blocks is labeled with the index r that runs from 0 to n and is spanned by the corresponding configuration class $\{|\phi_q^n\rangle\}_{(r)}$. The corresponding eigenvector matrix **S** of **H** that appears in Eq. (1.2) can be considered a block matrix as well with the same block structure. In the next sections we will always refer to this block structure and we will be interested in the block-diagonalization of the matrix **H**, i.e., in decoupling the various configuration classes so that $\widetilde{\mathbf{H}}$ becomes block-diagonal. We would like to stress that the block structure used here is only one which can be consistently used in second quantization in all n-particle Fock spaces once annihilation and creation operators for occupied and unoccupied one-particle states are introduced.

The configuration states $\{|\phi_q^n\rangle\}$ and the eigenvector matrix **S** define an "eigenstate operator" $\hat{S}$. The eigenstate operator $\hat{S}$ is that operator which in any n-particle space gives rise to the matrix **S** when represented in the basis $\{|\phi_q^n\rangle\}$. The operator $\hat{S}$ can be viewed as the operator transforming the configuration states to the exact eigenstates. As an interesting example we discuss the case where the operator $\hat{H}$ is the sum of the diagonal term $\hat{H}_0$ [Eq. (2.2)] and a nondiagonal term $\hat{H}_I$ of the form

$$\hat{H}_I \equiv \hat{W} = \sum_{\alpha,\beta} W_{\alpha\beta} a_\alpha^\dagger a_\beta. \tag{2.6}$$

Since the operator $\hat{H}$ is a one-particle operator, its eigenstates can be described by single determinants $\{|\Psi_q^n\rangle\}$ in a suitable one-particle basis set $\{|\tilde{\varphi}_\alpha\rangle\}$. The eigenstate operator $\hat{S}$ is therefore the operator that transforms the one-particle states of the set $\{|\varphi_\alpha\rangle\}$ into the one-particle states of the set $\{|\tilde{\varphi}_\alpha\rangle\}$. The operator $\hat{S}$ can be explicitly written down[5,7] in second quantization form and reads:

$$\hat{S} = e^{i\hat{s}}, \tag{2.7a}$$

$$\hat{s} = \sum_{\alpha,\beta} s_{\alpha\beta} a_\alpha^\dagger a_\beta, \quad \hat{s} = \hat{s}^\dagger. \tag{2.7b}$$

Here $\hat{s}$ is a Hermitian one-particle operator. $\hat{S}$ is unitary and, by definition, it transforms the configuration functions of the set $\{|\phi_q^n\rangle\}$ into the exact eigenfunctions of $\hat{H}$ of the set $\{|\Psi_q^n\rangle\}$ in *any* n-particle space. For example, the application of the eigenstate operator $\hat{S}$ to the N-particle reference function $|\phi_0^N\rangle$ [Eq. (2.4)] yields

$$\hat{S}|\phi_0^N\rangle, = |\Psi_0^N\rangle, \quad |\Psi_0^N\rangle = \prod_{i=1}^N \tilde{a}_{0_i}^\dagger |\text{vac}\rangle. \tag{2.8}$$

The relation between the creation and annihilation operators related to the sets $\{|\varphi_\alpha\rangle\}$ and $\{|\tilde{\varphi}_\alpha\rangle\}$ is given by

$$\begin{aligned} a_\alpha^\dagger &= \hat{S}^\dagger \tilde{a}_\alpha^\dagger \hat{S} \\ &= e^{-i\hat{s}} \tilde{a}_\alpha^\dagger e^{i\hat{s}} \\ &= \sum_\beta \tilde{a}_\beta^\dagger \langle\varphi_\beta|\hat{S}^\dagger|\varphi_\alpha\rangle \\ &= \sum_\beta \tilde{a}_\beta^\dagger \langle\tilde{\varphi}_\beta|\varphi_\alpha\rangle. \end{aligned} \tag{2.9}$$

In general H_I contains also non-one-particle terms. We assume that in this case the operator $\hat{S}$ can be again represented in second quantization by an exponential operator of the form given in Eq. (2.7a), but the operator $\hat{s}$ is no longer given by a one-particle operator as in Eq. (2.7b). Here $\hat{s}$ must be now an infinite sum of many-particle terms.

III. EXISTENCE OF A BLOCK-DIAGONALIZATION OPERATOR

In this section we want to demonstrate that the block-diagonalization operator $\hat{T}$ exists. More precisely, for any n-particle space, the matrix **T** that is the matrix representation of $\hat{T}$ in the basis $\{|\phi_q^n\rangle\}$ block-diagonalizes the matrix representation **H** of $\hat{H}$. The same $\hat{T}$ should, of course, apply to all n. Each of the transformation matrices **T** should fulfill the conditions of Ref. 1 (see also the Introduction). Before starting with the discussion it is useful to briefly review the principal points introduced in Ref. 1. There, two theorems concerned with the uniqueness of the block-diagonalization transformation **T** are proven and discussed. Both theorems can be formulated once the block structure is introduced. The first theorem can be expressed as the theorem of "minimal action of the transformation **T**" and states that **T** follows uniquely from the condition that its Euclidean norm is as near as possible to the norm of the unit matrix, i.e., that $\|\mathbf{T} - \mathbf{1}\| = \min$. The second theorem is introduced in a more general context and asserts that the uniqueness of **T** follows if **T** is "fully determined" by the eigenvector matrix **S** of **H**. With "fully determined" it is meant that a prescription g exists such that the block-diagonal matrix **F** [which together with **S** gives rise to **T**, see Eq. (1.2)] satisfies the condition $\mathbf{F} = g(\mathbf{S})$. Each diagonal block of the matrix **F** depends only on the corresponding diagonal block of the eigenvector matrix **S** and hence

$$\mathbf{F} = g(\mathbf{S}_{\text{BD}}). \tag{3.1}$$

$\mathbf{S}_{\text{BD}}$ denotes the block-diagonal part of the eigenvector matrix **S**. According to our definition of classes, **S** is a matrix with a well-defined block structure and the matrix $\mathbf{S}_{\text{BD}}$ is obtained from **S** by retaining its diagonal blocks and putting to zero the off-diagonal blocks. In this and in the next section we make use of the above cited second and first theorem of Ref. 1, respectively. The relevant formulas will be restated whenever necessary.

In order to prove the existence of an operator $\hat{T}$ which underlies the matrix **T** of Eq. (1.1) we need to transfer the relation expressed by Eq. (3.1) to operators. To be specific, we denote by $\mathbf{A}^{(n)}$ the matrix representation of the operator $\hat{A}$ in the n-particle space. As discussed in the preceding section, we dispose of a general eigenstate operator $\hat{S}$, the matrix representation of which is the eigenvector matrix $\mathbf{S}^{(n)}$ in any n-particle space. In order to transfer Eq. (3.1) to operators we now have to investigate whether it is possible to define an operator $\hat{S}_{\text{BD}}$ which in any n-particle space gives rise to a matrix $\mathbf{S}_{\text{BD}}^{(n)}$. Interestingly, the operator $\hat{S}_{\mathrm{BD}}$ can be obtained by means of projection operators $\hat{\mathscr{P}}_{nr}$ acting on the eigenstate operator $\hat{S}$. The simplest way to proceed is to define the operator $\hat{S}_{\mathrm{BD}}$ as follows:

$$\hat{S}_{\mathrm{BD}}=\sum_{n}\sum_{r=0}^{n}\hat{\mathscr{P}}_{nr}\hat{S}\hat{\mathscr{P}}_{nr}, \tag{3.2}$$

and then determine a suitable form for the projection operator $\hat{\mathscr{P}}_{nr}$. The projection operator $\hat{\mathscr{P}}_{nr}$ can be defined as follows:

$$\hat{\mathscr{P}}_{nr}=\hat{P}_{r}\hat{Q}_{n-r}, \tag{3.3}$$

with

$$\hat{P}_{r}=\delta\left(r\hat{1}-\sum_{u}a_{u}^{\dagger}a_{u}\right), \tag{3.4a}$$

$$\hat{Q}_{s}=\delta\left(s\hat{1}-\sum_{i}a_{i}^{\dagger}a_{i}\right). \tag{3.4b}$$

We remember that the indices u and i refer to unoccupied and occupied one-particle states, respectively. In Eqs.(3.4) the operator $\Sigma_{u}a_{u}^{\dagger}a_{u}(\Sigma_{i}a_{i}^{\dagger}a_{i})$ is the *number operator* which counts the one-particle states belonging to $\{|\varphi_{\alpha}\rangle\}_{\in}$ (belonging to $\{|\varphi_{\alpha}\rangle\}_{\notin}$). Here r is the number of unoccupied one-particle states necessary to construct a configuration function of the class r, i.e., the number which characterizes the configuration class r, and $s=(n-r)$ denotes the number of occupied one-particle states appearing in the configuration functions of the class r. Here $\hat{1}$ is the identity operator. The symbol δ indicates that the operators $\hat{P}_{r}$ and $\hat{Q}_{n-r}$ can be identified with the identity operator when the expression in parentheses vanishes and are zero elsewhere. Clearly the operator $\hat{\mathscr{P}}_{nr}$ [Eq. (3.4)] is the identity operator when the indices r and $n-r$ of the operators $\hat{P}_{r}$ and $\hat{Q}_{n-r}$ of Eqs. (3.4) determine the class r of the n-particle space. In fact, identifying with $|n,r'\rangle$ a configuration state of the n-particle space belonging to the class r', the application of the operators $\hat{P}_{r}$ and Q_{s} on this configuration function yields

$$P_{r}|n,r'\rangle=\delta_{r,r'}|n,r'\rangle, \tag{3.5a}$$

$$Q_{s}|n,r'\rangle=\delta_{s,n-r'}|n,r'\rangle. \tag{3.5b}$$

Hence, according to Eq. (3.2), $\hat{S}_{\mathrm{BD}}$ is that operator which has as matrix representation the block-diagonal matrix $\mathbf{S}_{\mathrm{BD}}^{(n)}$. It is therefore possible to define a matrix $\mathbf{F}^{(n)}$ and the relation (3.1) is valid in all n-particle spaces. The same relation (3.1) leads thus straightforwardly to the corresponding operator relation

$$\hat{F}=g(\hat{S}_{\mathrm{BD}}), \tag{3.6}$$

which asserts the existence and uniqueness of the operator $\hat{F}$.

IV. CONSTRUCTION OF THE BLOCK-DIAGONALIZATION OPERATOR

In this section we want to investigate explicitly the problem of block-diagonalizing a matrix $\mathbf{H}$ which is the representation of a one-particle operator $\hat{H}$ composed of a diagonal part $\hat{H}_{0}$ [Eq. (2.2)] and of a nondiagonal part $\hat{W}$ [Eq. (2.6)]. The eigenvector matrix $\mathbf{S}$ of $\mathbf{H}$ is the matrix representation of an exponential operator $\hat{S}$ [Eqs. (2.7)] which brings the eigenfunctions of $\hat{H}_{0}$ of the set $\{|\phi_{q}^{n}\rangle\}$ into the exact eigenfunctions $\{|\Psi_{q}^{n}\rangle\}$ of $\hat{H}$ in any n-particle space. The Hamiltonian $\hat{H}$ can be represented in the basis of the configuration functions $\{|\phi_{q}^{n}\rangle\}$. With the classification introduced in the preceding chapters for the configuration functions, the matrix representation $\mathbf{H}^{(n)}$ of $\hat{H}$ can be viewed as a block-tridiagonal matrix. By a block-tridiagonal matrix we understand a matrix which has nonvanishing blocks only along the diagonal and the first neighboring diagonals.

Now we want to find an explicit form for the block-diagonalization operator $\hat{T}$. For our purpose we will make use of the Theorem 1 of Ref. 1 according to which the transformation matrix $\mathbf{T}^{(n)}$ derives uniquely from the condition that the Euclidean norm of $(\mathbf{T}^{(n)}-\mathbf{1})$ takes on its minimum or, equivalently, that the trace of $(\mathbf{T}^{(n)}+\mathbf{T}^{(n)\dagger})$, i.e., $\mathrm{Tr}(\mathbf{T}^{(m)}+\mathbf{T}^{(n)\dagger})$ takes on its maximum. Since the matrix $\mathbf{S}_{\mathrm{BD}}^{(n)}\mathbf{F}^{(n)}$ is a block-diagonal matrix where the different classes of configurations are decoupled, we can consider each block separately and write

$$\mathrm{Tr}(\mathbf{S}_{r}^{(n)}\mathbf{F}_{r}^{(n)})+\mathrm{Tr}(\mathbf{S}_{r}^{\dagger(n)}\mathbf{F}_{r}^{\dagger(n)})=\mathrm{Maximum}, \tag{4.1}$$

where with $\mathbf{S}_{r}^{(n)}$ and $\mathbf{F}_{r}^{(n)}$ we indicate the diagonal blocks of the matrices $\mathbf{S}^{(n)}$ and $\mathbf{F}^{(n)}$, respectively.

We begin with the analysis in the one-particle space. This is the simplest n-particle space and the configuration functions ϕ_{q}^{1}'s can be subdivided into two classes only (class 0 and 1; see Sec. II). Since the matrix $\mathbf{F}$ must be a unitary block-diagonal matrix in the one-particle space, we assume that the corresponding operator $\hat{F}$ can be written in second quantization form as follows:

$$\hat{F}=\exp(i(\hat{f}_{h}+\hat{f}_{p})), \tag{4.2a}$$

$$\hat{f}=\sum_{i<j}f_{ij}a_{i}^{\dagger}a_{j},\quad \hat{f}_{h}=\hat{f}_{h}^{\dagger}, \tag{4.2b}$$

$$\hat{f}_{p}=\sum_{u<v}f_{uv}a_{u}^{\dagger}a_{v},\quad \hat{f}_{p}=\hat{f}_{p}^{\dagger}. \tag{4.2c}$$

The operator $\hat{F}$ represented in the one-particle space clearly gives rise to a unitary block-diagonal matrix $\mathbf{F}^{(1)}$, where the blocks spanned by the classes 0 and 1 decouple from each other.

The first term of the lhs of Eq. (4.1) can be explicitly written in the one-particle space in the class $r=0$ as (an analogous result holds for the class $r=1$)

$$\mathrm{Tr}(\mathbf{S}_{0}^{(1)}\mathbf{F}_{0}^{(1)})=\sum_{i}(\mathbf{S}_{0}^{(1)}\mathbf{F}_{0}^{(1)})_{i,i} \tag{4.3}$$

We decompose now the eigenvector matrix $\mathbf{S}_{0}^{(1)}$ according to

$$\mathbf{S}_{0}^{(1)}=\mathbf{UDV}^{\dagger},\quad \mathbf{UU}^{\dagger}=\mathbf{1},\quad \mathbf{VV}^{\dagger}=\mathbf{1}, \tag{4.4}$$

where $\mathbf{D}$ is a diagonal matrix with positive definite diagonal elements D_i. This decomposition is unique up to phase factors.[9,10] With the aid of Eq. (4.4) we can rewrite Eq. (4.3) as

$$\mathrm{Tr}(\mathbf{S}_0^{(1)}\mathbf{F}_0^{(1)})=\mathrm{Tr}(\mathbf{DX}), \tag{4.5}$$

where $\mathbf{X}$ is a unitary matrix defined as

$$\mathbf{X}=\mathbf{V}^\dagger\mathbf{F}_0^{(1)}\mathbf{U}. \tag{4.6}$$

Since $\mathbf{X}$ is a unitary matrix, Eq. (4.5) takes on its maximum when $\mathbf{X}$ is the unit matrix. From Eq. (4.6) it follows that $\mathbf{F}_0^{(1)} = \mathbf{VU}^\dagger$. This result fulfills Eq. (3.1), see also Ref. 1, and determines the number f_{ij} in Eq. (4.2b). The analogous result for the class $r=1$ determines the numbers f_{uv} in Eq. (4.2c). In this way we have determined the operator $\hat{F}$. We will show now that the same operator F leads to matrices $\mathbf{F}^{(n)}$ which satisfy Eq. (4.1). Consequently, the expression given in Eq. (4.2) is a suitable form for describing the operator $\hat{F}$ and we dispose of an explicit formula for the block-diagonalization operator in second quantization.

Before starting with the proof we are reminded that since $\hat{H}$ is a one-particle operator, its eigenfunctions can be described, as well as those of $\hat{H}_0$, by one-determinant functions. Therefore, the matrix elements of $\hat{S}$ [Eqs. (2.7)], $\hat{F}$ [Eqs. (4.2)] and of any one-particle operator possess a very simple and useful form. Consider, for example, one element of the eigenvector matrix $\mathbf{S}^{(n)}$:

$$\begin{aligned}\mathbf{S}_{qq'}^{(n)}&=\langle\phi_q^n|\Psi_{q'}^n\rangle\\&=\langle\mathrm{vac}|a_{q_1}a_{q_2}\cdots a_{q_n}\tilde{a}^\dagger_{q'_n}\tilde{a}^\dagger_{q'_{(n-1)}}\cdots\tilde{a}^\dagger_{q'_1}|\mathrm{vac}\rangle.\end{aligned} \tag{4.7}$$

Here the creation operators $\tilde{a}^\dagger_\alpha$ refer to the exact one-particle states of the set $\{|\tilde{\varphi}_\alpha\rangle\}$ (the one-particle basis set in which the Hamiltonian $\hat{H}$ is diagonal). Equation (4.7) describes the overlap of two determinants and can be therefore rewritten as

$$\begin{aligned}\mathbf{S}_{qq'}^{(n)}&=\det|\langle\varphi_{q_1}|\tilde{\varphi}_{q'_1}\rangle\langle\varphi_{q_2}|\tilde{\varphi}_{q'_2}\rangle\cdots\langle\varphi_{q_n}|\tilde{\varphi}_{q'_n}\rangle|\\&=\det|\langle\varphi_{q_1}|\hat{S}|\varphi_{q'_1}\rangle\langle\varphi_{q_2}|\hat{S}|\varphi_{q'_2}\rangle\cdots\langle\varphi_{q_n}|\hat{S}|\varphi_{q'_n}\rangle|.\end{aligned} \tag{4.8}$$

With this expression we indicate the determinant of the matrix, the elements of which are the overlap of the unperturbed and exact one-particle states used to construct $|\phi_q^n\rangle$ and $|\Psi_{q'}^n\rangle$, respectively.

Making use of this property we can demonstrate that in any n-particle space the relation expressed in Eq. (4.1) is satisfied for the blocks spanned by all the configuration classes. We show here as an explicit example the case of the class 0 in the two-particle space. The treatment of the classes 0 and n of any n-particle space is completely analogous. The proof for the other classes (1 to $n-1$) is more lengthy and is not reported here. In our example we will study only the first term of Eq. (4.1) since the complex conjugate term gives rise to the same result.

A configuration function belonging to the class 0 in the two-particle space is a single determinant built up from two one-particle states of the set $\{|\varphi_\alpha\rangle\}_\epsilon$:

$$|\phi_q^2\rangle=a_j^\dagger a_i^\dagger|\mathrm{vac}\rangle=|\varphi_i\varphi_j\rangle,\quad i<j. \tag{4.9}$$

Hence, the first term of Eq. (4.1) can be explicitly written as follows:

$$\mathrm{Tr}(\mathbf{S}_0^{(2)}\mathbf{F}_0'^{(2)})=\sum_{i<j}(\mathbf{S}_0^{(2)}\mathbf{F}_0'^{(2)})_{ij,ij}. \tag{4.10}$$

We introduced here a matrix $\mathbf{F}_0'^{(2)}$ instead of $\mathbf{F}_0^{(2)}$ and we suppose that the matrix $\mathbf{F}_0'^{(2)}$ is the matrix representation of a one-particle $\hat{F}'$ which can be different from the operator $\hat{F}$ in that the numbers f_{ij} and f_{uv} can differ from those determined by $\mathbf{F}^{(1)}$. We will then show that we obtain necessarily $\hat{F}'=\hat{F}$ if Eq. (4.1) is fulfilled. Writing Eq. (4.10) as an expectation value of the operators $\hat{S}$ and $\hat{F}'$ on the configuration functions of class 0 [see Eq. (4.9)] we obtain

$$\mathrm{Tr}(\mathbf{S}_0^{(2)}\mathbf{F}_0'^{(2)})=\sum_{i<j}\sum_{k<1}\langle\varphi_i\varphi_j|\hat{S}|\varphi_k\varphi_l\rangle\langle\varphi_k\varphi_l|\hat{F}'|\varphi_i\varphi_j\rangle. \tag{4.11}$$

Since $\mathbf{F}'$ is block diagonal by definition, the configuration states $|\varphi_k\varphi_l\rangle$ also belong to the class 0. Equation (4.11) can be rewritten to give

$$\mathrm{Tr}(\mathbf{S}_0^{(2)}\mathbf{F}_0'^{(2)})=\tfrac{1}{2}\{[\mathrm{Tr}(\mathbf{T}_0'^{(1)})]^2-\mathrm{Tr}(\mathbf{T}_0'^{(1)})^2\}, \tag{4.12}$$

where $\mathbf{T}_0'^{(1)}$ is the upper left block of the matrix $\mathbf{T}'^{(1)} = \mathbf{S}^{(1)}\mathbf{F}'^{(1)}$. Using Eq. (4.4) to decompose the eigenvector matrix $\mathbf{S}_0^{(1)}$ and the notation $\mathbf{X}' = \mathbf{V}^\dagger\mathbf{F}_0'^{(1)}\mathbf{U}$ yields:

$$\mathrm{Tr}(\mathbf{S}_0^{(2)}\mathbf{F}_0'^{(2)})=\tfrac{1}{2}\{[\mathrm{Tr}(\mathbf{DX}')]^2-\mathrm{Tr}(\mathbf{DX}')^2\}. \tag{4.13}$$

The rhs of Eq. (4.10) assumes the following explicit form:

$$\frac{1}{2}\sum_{i,j}D_iD_j(X'_{ii}X'_{jj}-X'_{ij}X'_{ji}). \tag{4.14}$$

The factor $(X'_{ii}X'_{jj} - X'_{ij}X'_{ji})$ can be viewed as one of the principal minors of rank 2 of the unitary matrix $\mathbf{X}'$. The rhs of Eq. (4.13) is thus a linear combination of all principal minors of rank 2 of the matrix $\mathbf{X}'$ weighted with positive coefficients. If we now require that the lhs of Eq. (4.13) takes on its maximum, it follows that $\mathbf{X}'$ must be the unit matrix (since $\mathbf{X}'$ is a unitary matrix, the individual principal minors are largest when $\mathbf{X}'=\mathbf{1}$). Hence, by comparison with Eq. (4.6) it follows that $\mathbf{X}'=\mathbf{X}$ and $\mathbf{F}_0'^{(1)} = \mathbf{VU}^\dagger$, and remembering that $\mathbf{F}_0^{(1)} = \mathbf{VU}^\dagger$ [1] we finally obtain that $\hat{F}'=\hat{F}$.

The general proof for the class 0 of a general n-particle space can be carried out in an analogous way. One always obtains

$$\mathrm{Tr}(\mathbf{S}_0^{(n)}\mathbf{F}_0'^{(n)})=\frac{1}{n!}\sum_{i_k}D_{i_1}D_{i_2}\cdots D_{i_n}\mathbf{X}'^{(n)}_{i_1i_2\cdots i_n,i_1i_2\cdots i_n}. \tag{4.15}$$

Here the sum runs over all the occupied one-particle states and $\mathbf{X}'^{(n)}_{i_1i_2\cdots i_n,i_1i_2\cdots i_n}$ is one of the principal minors of rank n of the unitary matrix $\mathbf{X}'$. $\mathrm{Tr}(\mathbf{T}_0'^{(n)})$ could thus be expressed by the trace of $\mathbf{T}_0'^{(1)}$ or powers of it. Equation

(4.15) represents again a linear combination of all principal minors of rank n (taken with positive coefficients) and takes on its maximum when $\mathbf{X}'=\mathbf{X}=\mathbf{1}$.

A simple operator $\hat{F}$ has been introduced and the block-diagonalization of the matrix $\mathbf{H}^{(1)}$ into two blocks allows to fix the number f_{ij} and f_{uv} in Eq. (4.2) and thus the operator $\hat{F}$. We have then shown that the same operator $\hat{F}$ applies to all n-particle spaces: It gives rise to the block-diagonal matrix $\mathbf{F}^{(n)}$, which together with the eigenvector matrix $\mathbf{S}^{(n)}$ block-diagonalizes the Hamiltonian $\mathbf{H}^{(n)}$.It is worth noting that the operator $\hat{T}=\hat{S}\hat{F}$ transforms $\hat{H}$ into two decoupled parts, one related to the occupied and the other to the unoccupied space of creation and annihilation operators. Explicitly this can be written as

$$\hat{\tilde{H}}=\hat{T}^{\dagger}\hat{H}\,\hat{T}=\sum_{i<j}\tilde{h}_{ij}a_i^{\dagger}a_j+\sum_{u<v}\tilde{h}_{uv}a_u^{\dagger}a_v. \qquad (4.16)$$

In this expression the matrix elements $\tilde{h}_{ij}$ and $\tilde{h}_{uv}$ are elements of the block-diagonal matrix $\tilde{\mathbf{H}}^{(1)}=\mathbf{T}^{(1)\dagger}\mathbf{H}^{(1)}\mathbf{T}^{(1)}$. The elements $\tilde{h}_{ij}$ and $\tilde{h}_{uv}$ belong to the two different blocks along the diagonal of $\tilde{\mathbf{H}}^{(1)}$.

V. APPLICATION: GREEN'S FUNCTIONS

Block-diagonalizing the Hamiltonian and decoupling the configuration classes from each other is very useful in the theory of the Green's functions.[4-7]. We continue our study in the second quantization formalism and define a Green's function in a very general form which contains as special cases the various Green's functions known in the literature. For this purpose we introduce a vector of operators $\hat{\mathbf{O}}$ the elements of which are all possible products of creation and annihilation operators

$$\hat{\mathbf{O}}=(a_\alpha,a_\alpha a_\beta,\ldots,a_\alpha^{\dagger},a_\alpha^{\dagger}a_\beta,a_\alpha^{\dagger}a_\beta a_\gamma,\ldots a_\alpha^{\dagger}a_\beta^{\dagger}\ldots). \qquad (5.1)$$

In this expression it is understood that the single vector elements stand for arrays of all possible values of the indices $\alpha,\beta,\ldots$ which run over all the one-particle labels. By means of this operator we define the general Green's function matrix as follows:

$$\mathbf{G}(\omega)=\langle\mathbf{Y}^{\dagger}|\hat{\Omega}^{-1}|\mathbf{Y}\rangle, \qquad (5.2a)$$

where

$$|\mathbf{Y}\rangle=\hat{\mathbf{O}}\,|\Psi_0^N\rangle, \qquad (5.2b)$$

$$\hat{\Omega}=(\omega-E_0^N+\hat{H}). \qquad (5.2c)$$

In Eqs. (5.2) $|\Psi_0^N\rangle$ is the N-particle ground state of the Hamiltonian $\hat{H}$ and E_0^N is its energy. It is easy to see that the elements of $\hat{\mathbf{O}}$ that produce a different change in the number of particles do not couple to each other and therefore the elements of the vector $\hat{\mathbf{O}}$ containing a number n_a of annihilation operators and a number n_c of creation operators such that the quantity $(n_a - n_c)$ is different can be considered separately. As an example we write here the part $\hat{\mathbf{O}}'$ of the vector $\hat{\mathbf{O}}$ composed of simple operators so that $n_a - n_c=1$

$$\hat{\mathbf{O}}'=(a_\alpha,a_\alpha^{\dagger}a_\beta a_\gamma,a_\alpha^{\dagger}a_\beta^{\dagger}a_\gamma a_\delta a_\epsilon,\ldots). \qquad (5.3)$$

The common one-particle Green's function[4-7] is obtained from the first element $\{a_\alpha\}$ of $\hat{\mathbf{O}}'$.

If we require to block-diagonalize $\hat{H}$ in the Green's function matrix, we have to introduce in Eq. (5.2a) the identity operator in the form of $\hat{T}\,\hat{T}^{\dagger}$ and obtain

$$\mathbf{G}(\omega)=\langle\tilde{\mathbf{Y}}^{\dagger}|\hat{\tilde{\Omega}}^{-1}|\tilde{\mathbf{Y}}\rangle, \qquad (5.4a)$$

where

$$|\tilde{\mathbf{Y}}\rangle=\hat{T}^{\dagger}\hat{\mathbf{O}}\,|\Psi_0^N\rangle, \qquad (5.4b)$$

$$\hat{\tilde{\Omega}}=(\omega-E_0^N+\hat{\tilde{H}}). \qquad (5.4c)$$

The operators $\hat{\tilde{H}}$ and thus $\hat{\tilde{\Omega}}$ in Eq. (5.4c) are block-diagonal, i.e., the various configuration classes are decoupled in all n-particle representations.

We investigate now the consequences of the block-diagonalization of the one-particle Hamiltonian $\hat{H}$ in the Green's function $\mathbf{G}$. As mentioned before, the different n-particle Fock spaces are by definition decoupled from each other since the Hamiltonian is a particle number conserving operator. Therefore we may restrict our analysis to the part $\hat{\mathbf{O}}'$ [Eq. (5.3)] of $\hat{\mathbf{O}}$, which couples the N-particle space to the $(N-1)$-particle space and describes the removal of one particle from the exact N-particle ground state $|\Psi_0^N\rangle$. We find it, however, important to stress that all the following considerations can very easily be transferred to other Fock spaces, i.e., to other Green's functions, using trivial modifications only.

Inserting into Eq. (5.2a) a complete set of configuration states, it is easy to see that, if the operators $\hat{O}_\alpha$ and $\hat{O}_\beta$ are identified with the elements of the operator $\hat{\mathbf{O}}'$ the only nonvanishing contributions come from the set $\{|\phi_q^{N-1}\rangle\}$ and Eq. (5.2a) takes on the explicit form

$$\begin{aligned}G_{\alpha\beta}(\omega)=\sum_{q,q'}&\langle\Psi_0^N|\,\hat{O}_\alpha^{\prime\dagger}|\phi_q^{N-1}\rangle\\&\times\langle\phi_q^{N-1}|(\omega-E_0^N+\hat{H})^{-1}|\phi_{q'}^{N-1}\rangle\\&\times\langle\phi_{q'}^{N-1}|\,\hat{O}'_\beta|\Psi_0^N\rangle,\end{aligned} \qquad (5.5a)$$

$$\begin{aligned}\mathbf{G}(\omega)&=\mathbf{Y}^{\dagger}[(\omega-E_0^N)\mathbf{1}+\mathbf{H}]^{-1}\mathbf{Y}\\&=\tilde{\mathbf{Y}}^{\dagger}[(\omega-E_0^N)\mathbf{1}+\tilde{\mathbf{H}}]^{-1}\tilde{\mathbf{Y}}.\end{aligned} \qquad (5.5b)$$

The configuration states $\{|\phi_q^{N-1}\rangle\}$ can be divided, according to our general definition of classes, into N classes r, with r running from 0 to $N-1$. There is a correspondence between the configuration classes of the $(N-1)$-particle space and the blocks of the operator $\hat{\mathbf{O}}'$. By selecting suitably the indices of the simple operators in the expression (5.3) the operator $\hat{\mathbf{O}}'$ can be ulteriorly subdivided into a "physical part" $\hat{\mathbf{O}}_p$ which has a one-to-one correspondence with the configuration classes and an "unphysical part." The physical part contains the operators $a_i,a_u^{\dagger}a_ja_j,\ldots$, i.e., where all creation operators are related to unoccupied and the annihilation operators to occupied one-particle states. The other combinations constitute the unphysical part. As will become clear below, the consideration of the physical part only is sufficient to derive the most important consequences of the application of the block-diagonalization operator $\hat{T}$ to the Green's function.

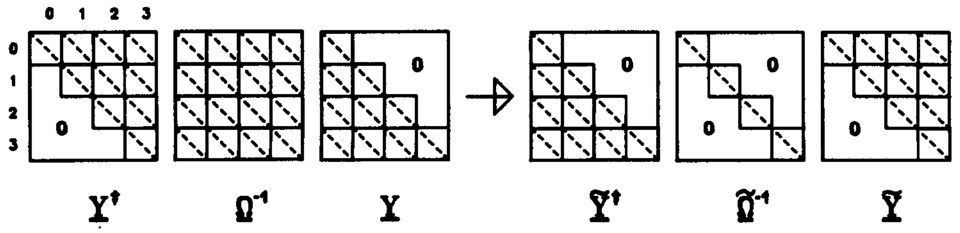


FIG. 1. Action of the operator $\hat{T}$ on the Green's function for the specific case of the three-particle Fock space. In this space there are four configuration classes r, with r running from 0 to 3. The inverse matrix $\tilde{\mathbf{\Omega}}^{-1}$ becomes a block-diagonal matrix and the matrix of residues changes from a lower to an upper triangular block form.

Taking into account in Eqs. (5.2) the operator $\hat{\mathbf{O}}_p$ only, the matrix representation of the Green's function becomes a square matrix, where the blocks can be unambiguously labeled with the indices r of the configuration classes. The matrix of the residues has now a defined block structure and the single blocks can be written as

$$\mathbf{Y}_{rr'} = \langle N-1,r|\,\hat{O}_{r'}\,|\Psi_0^N\rangle. \tag{5.6}$$

Interestingly, the residues matrix $\mathbf{Y}$ is a <u>lower</u> triangular block matrix, i.e., it has nonvanishing contributions only in the diagonal blocks and in the blocks below the diagonal. Moreover, we can show that the block-diagonalization of the Hamiltonian $\hat{H}$ in the expression of the Green's function also has an important consequence: the transformed residues matrix $\tilde{\mathbf{Y}}$ [Eq. (5.4b)] becomes an <u>upper</u> triangular block matrix (see below). As a direct consequence, the spaces spanned by the different configuration classes r couple only with the classes $r' < r$. This fact has an important and easy physical interpretation: Consider for example the usual one-particle Green's function, which is described by the residues $|Y\rangle = a_\alpha|\Psi_0^N\rangle$. If we perform in Eq. (5.2) the transformation defined by Eq. (2.9) of the creation and annihilation operators, thus going from the unperturbed ones $\{a_\alpha\}$ to the "exact" ones $\{\tilde{a}_\alpha\}$, it is easy to see that the one-particle Green's function expressed in the latter operators has occupied indices only. Consequently the one-particle Green's function has only N poles and they are all characterizeable by the occupied one-particle states. It is thus reasonable to search for a representation of the one-particle Green's function in the space spanned only by the configuration class 0 of the $(N-1)$-particle space. By similar arguments, the propagator specified by $|Y\rangle = a_\alpha^\dagger a_\beta a_\gamma|\Psi_0^N\rangle$ can be transformed into a matrix with indices which correspond only to the configuration classes 0 and 1 of the $(N-1)$-particle space. Hence, by using the transformation $\hat{T}$ in the Green's function we simultaneously cast all propagators into their respective smallest configuration spaces possible. This implies a block-diagonal Hamiltonian and an upper triangular matrix of residues. The action of the block-diagonalization operator $\hat{T}$ on the Green's function $\mathbf{G}$ is graphically represented in Fig. 1.

In order to show that the transformed residues matrix $\tilde{\mathbf{Y}}$ is an upper triangular block matrix we recall first that the operator $\hat{T}$ satisfies the following important relation[8]:

$$\hat{T}|\phi_0^N\rangle = |\Psi_0^N\rangle. \tag{5.7}$$

Consider now the block $\tilde{\mathbf{Y}}_{rr'}$ of $\tilde{\mathbf{Y}}$:

$$\tilde{\mathbf{Y}}_{rr'} = \langle N-1,r|\,\hat{T}^\dagger\hat{O}_{r'}\hat{T}\,|\phi_0^N\rangle. \tag{5.8}$$

Since $\hat{T}$ is by definition a one-particle operator, the expression $\hat{T}^\dagger\hat{O}_{r'}\hat{T}$ represents a unitary transformation of the operators a_α and $a_\alpha^\dagger$ appearing in $\hat{O}_{r'}$. The operator $\overline{O}_{r'}$ which derives from this transformation is thus composed of the same number of creation and annihilation operators as $\hat{O}_{r'}$. The operators that appear in $\overline{O}_{r'}$ can always be written as linear combination of the operators a_α and $a_\alpha^\dagger$. Therefore, the block $\tilde{\mathbf{Y}}_{rr'}$ is zero if $\overline{O}_{r'}$ contains more simple operators than necessary to construct a state $|N-1,r\rangle$ of the class r by applying creation and annihilation operators to $|\phi_0^N\rangle$. It follows that the lower triangular blocks of the matrix $\tilde{\mathbf{Y}}$ vanish identically and our statement is proven.

VI. CONNECTION TO THE GRAM–SCHMIDT ORTHOGONALIZATION PROCEDURE AND THE EXPLICIT MATRIX REPRESENTATION OF $\hat{\mathbf{T}}$

The fact that the product $\mathbf{T}^\dagger\mathbf{Y}$ gives rise to an upper triangular block matrix reminds of the classical QR decomposition (or Gram–Schmidt orthogonalization procedure) of a square matrix.[10] As is well known, the QR procedure is a decomposition of a square real matrix $\mathbf{M}$ into a product of a unitary matrix $\mathbf{Q}$ and an upper triangular matrix $\mathbf{R}$ according to

$$\mathbf{M} = \mathbf{QR}. \tag{6.1}$$

This decomposition is unique up to phase factors. In order to generalize Eq. (6.1) to block matrices we have simply to identify the matrices with square matrices having a well determined block structure in which $\mathbf{R}$ is an upper triangular block matrix. In this case the decomposition expressed by Eq. (6.1) is no longer unique. Two different QR decompositions can differ at most by a unitary block-diagonal matrix. To prove our statement, consider another QR decomposition for the block matrix $\mathbf{M}$. With an anologous notation as used in Eq. (6.1) we can write:

$$\mathbf{M} = \mathbf{Q}'\mathbf{R}'. \tag{6.2}$$

From Eqs. (6.1) and (6.2) it follows that

$$\mathbf{R}' = \mathbf{Q}'^\dagger\mathbf{QR}. \tag{6.3}$$

Assuming that the diagonal blocks of the matrices involved in Eq. (6.3) are not singular and taking into account that both $\mathbf{R}$ and $\mathbf{R}'$ are upper triangular block matrices, it necessarily follows that the unitary matrix $\mathbf{Q}'\mathbf{Q}$ is a block-diagonal matrix.

Identifying $\mathbf{Y}$ with the matrix to be QR-decomposed we find that, because of $\mathbf{Y} = \widehat{T}\widetilde{\mathbf{Y}}$, we may consider $\mathbf{Q} = \mathbf{T}$ and $\mathbf{R}' = \widetilde{\mathbf{Y}}$, in accordance with the results of the preceding section. It turns out that the matrix $\mathbf{T}$ and the matrix $\mathbf{Q}$ (which is obtained by simply orthogonalizing the columns of blocks of the matrix $\mathbf{Y}$ by means of the Gram–Schmidt procedure) may differ at most by a unitary block-diagonal matrix. The block-diagonalizing transformation matrix $\mathbf{T}$ is just one of the possible unitary matrices that follow from the *QR*-decomposition of $\mathbf{Y}$. We shall see below how $\mathbf{T}$ is obtained from $\mathbf{Y}$.

Now we attempt to find explicit expressions for the matrix representation of the operator $\widehat{T}$ which block-diagonalizes the Hamiltonian $\widehat{H}$. Such expressions are very helpful in performing practical calculations. In any n-particle space the matrix $\mathbf{T}$ reads[1,8]:

$$\mathbf{T} = \mathbf{U}(\mathbf{U}^\dagger\mathbf{U})^{-1/2}, \tag{6.4}$$

where

$$\mathbf{U} = \mathbf{S}\mathbf{S}_{\mathrm{BD}}^{-1}. \tag{6.5}$$

We already found out elsewhere[8] that the first column of blocks of the matrix $\mathbf{U}$ coincides with the first column of the matrix $\mathbf{Y}$, up to a factor $\langle\phi_0^N|\Psi_0^N\rangle^{-1}$. The calculation of the remaining part of $\mathbf{U}$ can be performed by subsequently computing one column of blocks after the other. The line of the general calculation and the details of the explicit derivation of the second column as an example are reported in the Appendix. Here we quote the results that we have obtained for the first three columns:

$$\mathbf{U}_{r0} = (\langle\phi_0^N|\Psi_0^N\rangle)^{-1}\mathbf{Y}_{r0}, \tag{6.6a}$$

$$\mathbf{U}_{r1} = (\langle\phi_0^N|\Psi_0^N\rangle)^{-1}\{\mathbf{Y}_{r1} - P_{1(r)}\mathbf{Y}_{r0}\mathbf{U}_{01}\}, \tag{6.6b}$$

$$\begin{aligned}\mathbf{U}_{r2} = (\langle\phi_0^N|\Psi_0^N\rangle)^{-1}\{&\mathbf{Y}_{r2} + \tfrac{1}{2}P_{1(r)}P_{2(r)}\mathbf{Y}_{r0}\mathbf{U}_{02}\\ &- P_{2(r)}\mathbf{Y}_{r1}\mathbf{U}_{12}\}.\end{aligned} \tag{6.6c}$$

In these formulas the coefficients $P_{i(r)}$ are given by $P_{i(r)} = i - r$. The general form for the block $\mathbf{U}_{rr'}$ of $\mathbf{U}$ is also given in the Appendix. The expressions presented in Eqs. (6.6) still contain some unknown block elements of the matrix $\mathbf{U}$ itself. These terms are the blocks $\mathbf{U}_{rr'}$ with $r < r'$ and can be obtained by using the necessary condition that the matrix $(\mathbf{U}^\dagger\mathbf{U})$ is a block-diagonal matrix. In other words, we require, according to the definition, that the columns of the matrix $\mathbf{U}$ are orthogonal to each other. As an example, we determine here $\mathbf{U}_{01}$, the knowledge of which completes the computation of the second column of blocks of $\mathbf{U}$. In order to obtain the term $\mathbf{U}_{01}$ we use the condition

$$\sum_r (\mathbf{U}_{r0})^\dagger\mathbf{U}_{r1} = \mathbf{0}, \tag{6.7}$$

and the result reads:

$$\mathbf{U}_{01} = (\boldsymbol{\rho}_1)^{-1}\sum_r (\mathbf{Y}_{r0})^\dagger\mathbf{Y}_{r1}, \tag{6.8a}$$

where

$$\boldsymbol{\rho}_1 = \sum_r (\mathbf{Y}_{r0})^\dagger P_{1(r)}\mathbf{Y}_{r0}. \tag{6.8b}$$

If the Green's function in question is specified, $\boldsymbol{\rho}_1$ can be given explicitly. For instance, dealing with the one-particle Green's function it takes on the following appearance:

$$(\boldsymbol{\rho}_1)_{i,j} = \langle\Psi_0^N|a_i^\dagger\Big(\sum_u a_u^\dagger a_u\Big)a_j|\Psi_0^N\rangle. \tag{6.8c}$$

Once $\mathbf{U}$ is calculated, the transformation matrix $\mathbf{T}$ is obtained by multiplying it with the unitarization factor $(\mathbf{U}^\dagger\mathbf{U})^{-1/2}$, which is a block-diagonal matrix. This completes the computation of an explicit form for the block-diagonalization matrix $\mathbf{T}$.

These formulas have been obtained under the assumption that the Hamiltonian $\widehat{H}$ is a one-particle operator and thus the corresponding N-particle ground state is described by a single determinant. However, we anticipate that these formulas are also meaningful in the more general case of an Hamiltonian where the interaction part contains also two-particle terms, by simply identifying $|\Psi_0^N\rangle$ with the exact N-particle ground state of the system (see next section and Ref. 8). We should mention here that $|\Psi_0^N\rangle$ is the only *a priori* unknown quantity in the expression of $\mathbf{Y}$ and hence of $\mathbf{T}$ [see Eqs. (5.5) and (6.6)].

VII. DISCUSSION AND CONCLUSIONS

The unitary matrix that brings a Hermitian matrix $\mathbf{H}$ into block-diagonal form can be uniquely determined under very simple and transparent conditions. In this work we investigated the block-diagonalization problem in the framework of the second quantization formalism. We started with an operator $\widehat{H}$ that in any n-particle Fock space has a well-defined matrix representation and attempted to answer the question whether the transformation matrices $\mathbf{T}$ that can be separately given in the various N-particle spaces can be considered as different matrix representations of the same operator $\widehat{T}$. Interestingly, we reached the very important result that the block-diagonalization operator $\widehat{T}$ exists and is unique. As a particular example we concentrated our attention to the case of an operator $\widehat{H}$ given by a one-particle operator. In this case the block-diagonalization operator can be constructed and given in explicit form.

As an important application we approached the Green's function theory. The Green's function could be defined in a very general way. This general form contains the information about all propagators related to the same physical process. As a specific example we analyzed the various propagators that describe the process of removal of one particle from the exact ground state of the system. As is well known, the propagators that are suitable to describe this process connect two Fock spaces, which can be characterized as N- and $(N-1)$-particle spaces. We point out, however, that the theory can be straightforwardly generalized to any other Green's function and to any other

Fock space by simple modifications. This is possible because we have at our disposal a block-diagonalization operator $\hat{T}$ which produces the same effects and consequences in any n-particle space.

If the Hamiltonian that appears in the Green's function is taken to be a one-particle operator describing, for instance, particles in an external field, the block-diagonalization of the Hamiltonian matrix allows for an exact reformulation of the propagators which has a very interesting interpretation. From an algebraic point of view the block-diagonalization of the Hamiltonian matrix has as a direct consequence that the physical part of the residues matrix becomes an upper triangular block matrix. The physical interpretation of this transformation can be summarized as follows: each propagator can be exactly reformulated as a matrix in that configuration space which has the same dimension and the same characteristics as the matrix which one would obtain in the space spanned by the exact eigenstates of the Hamiltonian. This space is clearly the smallest space necessary to have an exact representation of the propagators. This property of the propagators can be explicitly proven by taking into account that the Hamiltonian of the system is a one-particle operator and thus its eigenstates are described by single Slater determinants in a suitable basis of one-particle states. Since we can perform the transformation by means of an operator $\hat{T}$ which is uniquely defined, we have an instrument which allows for an exact and simultaneous reformulation of all propagators related to the same process. Moreover, the matrix elements of the transformation matrix **T** can be explicitly constructed by means of a procedure which is strictly related to the well known Gram–Schmidt orthogonalization procedure. We have demonstrated that the results that one obtains by simply orthogonalizing the columns of blocks of the residues matrix to each other and by our transformation may differ at most by a block-diagonal unitary matrix which can be given.

Finally, we would like to briefly discuss the importance of the block-diagonalization operator $\hat{T}$ in the more general case of a Hamiltonian that contains, in addition to the one-particle term, also two-particle interactions. The explicit construction of the matrix elements of **T** shows that all formulas (which are obtained by taking into account that the exact eigenstates of the one-particle operator $\hat{H}$ are Slater determinants) are expressed as a function of the exact N-particle ground state $|\Psi_0^N\rangle$ only. Clearly, if the Hamiltonian contains also two-body terms, $|\Psi_0^N\rangle$ is no longer a determinant and in general is not known exactly. In this case the block-diagonalization operator becomes very complicated. However, the transformation matrix obtained here for the one-particle Hamiltonian is still useful also in the general case if we simply identify $|\Psi_0^N\rangle$ with the exact N-particle ground state. Although with this procedure the transformation **T** does not block-diagonalize exactly the Hamiltonian matrix **H**, the transformed Hamiltonian assumes a very interesting structure.[8] It can be proven that this structure is very useful in computing propagators consistently in perturbation theory.

ACKNOWLEDGMENTS

The authors wish to thank Dr. H.-D. Meyer and Dr. J. Schirmer for helpful discussions.

The Deutsche Forschungsgemeinschaft is gratefully acknowledged for financial support.

APPENDIX

In the following we will carry out the explicit calculation of the second column of blocks of the matrix **U** [Eq. (6.3)]. The procedure of calculation is completely general and can be applied to any Green's function. In particular the results that can be obtained for the matrix **U** are independent of the choice of the n-particle space. The simplest way to derive the matrix **U** is to consider the one-particle Green's function and the related propagators of higher order, i.e., to work in the $(N-1)$-particle configuration space. As mentioned in the text and explained in detail in Ref. 8, the first column of the matrix **U** reads:

$$\mathbf{U}_{r0}=\mathbf{S}_{r0}^{(N-1)}(\mathbf{S}_{00}^{(N-1)})^{-1}=(\langle\phi_0^N|\Psi_0^N\rangle)^{-1}\mathbf{Y}_{r0}. \quad \text{(A1)}$$

Here with $\mathbf{S}_{rr'}^{(N-1)}$ we denote the block rr' of the eigenvector matrix $\mathbf{S}^{(N-1)}$ of $\mathbf{H}^{(N-1)}$. In the following we drop the superscript $(N-1)$ whenever unnecessary.

The second block of columns of the matrix **U** reads, by definition [see Eq. (6.3)],

$$\mathbf{U}_{r1}=\mathbf{S}_{r1}(\mathbf{S}_{11})^{-1}, \quad \text{(A2)}$$

and we attempt now to find an explicit form of this quantity in terms of the blocks of **Y**. For this purpose we analyze the expression

$$\mathbf{Y}_{r1}\mathbf{S}_{11}=\frac{1}{2}\sum_{u}\sum_{i,j}\langle N-1,r|a_u^\dagger a_i a_j|\Psi_0^N\rangle\langle\phi_0^N|a_j^\dagger a_1^\dagger a_u|\Psi_1^{N-1}\rangle. \quad \text{(A3)}$$

The states $|\Psi_1^{N-1}\rangle$ are eigenstates of $\mathbf{H}^{(N-1)}$ and correspond to the configuration class $r=1$.

To start the discussion, we transform Eq. (A3) into a sum of two terms obtained by considering the sum over the unoccupied index u as a sum over all one-particle indices α minus a sum over the occupied ones:

$$\begin{aligned}\mathbf{Y}_{r1}\mathbf{S}_{11}=&\frac{1}{2}\sum_{\alpha}\sum_{i,j}\langle N-1,r|a_\alpha^\dagger a_i a_j|\Psi_0^N\rangle\langle\phi_0^N|a_j^\dagger a_i^\dagger a_\alpha|\Psi_1^{N-1}\rangle\\ &-\frac{1}{2}\sum_{k,i,j}\langle N-1,r|a_k^\dagger a_i a_j|\Psi_0^N\rangle\\ &\times\langle\phi_0^N|a_j^\dagger a_i^\dagger a_k|\Psi_1^{N-1}\rangle. \end{aligned} \quad \text{(A4)}$$

The two terms of Eq. (A4) have now to be handled in a different way. The second term is indeed quite easy to transform into a more suitable form since the operators $a_j^\dagger$, $a_i^\dagger$, and a_k have all occupied indices and therefore act on the reference state $\langle\phi_0^N|$ in a well-known way. Using the anticommutation properties of these operators and bearing in mind that a creation operator with occupied index gives a vanishing contribution when acting on the reference state, we obtain for the second term of Eq. (A4):

$$-s\sum_i \langle N-1,r|a_i|\Psi_0^N\rangle\langle\phi_0^N|a_i^\dagger|\Psi_1^{N-1}\rangle, \tag{A5}$$

where s is the number of occupied one-particle states in $|N-1,r\rangle$, i.e., $s=N-1-r$.

By transforming the operators $a_\alpha^\dagger a_\beta a_j$ into the operators $\tilde{a}_\alpha^\dagger\tilde{a}_\beta\tilde{a}_\gamma$ which act on $|\Psi_0^N\rangle$ [see Eq. (2.9)], we obtain for the first term of Eq. (A4)

$$\frac{1}{2}\sum_{k,i,j}\langle N-1,r|\,\tilde{a}_k^\dagger\tilde{a}_i\tilde{a}_j|\Psi_0^N\rangle\langle\phi_0^N|\,\tilde{a}_j^\dagger\tilde{a}_i^\dagger\tilde{a}_k|\Psi_1^{N-1}\rangle$$

$$+\frac{1}{2}\sum_u\sum_{i,j}\langle N-1,r|\,\tilde{a}_u^\dagger\tilde{a}_i\tilde{a}_j|\Psi_0^N\rangle$$

$$\times\langle\phi_0^N|\,\tilde{a}_j^\dagger\tilde{a}_i^\dagger\tilde{a}_u|\Psi_1^{N-1}\rangle, \tag{A6}$$

where we have taken into account that a creation operator $\tilde{a}_i$ with occupied index gives zero by acting on $|\Psi_0^N\rangle$. The second term of Eq. (A6) can be straightforwardly rewritten to give $\mathbf{S}_{r1}\langle\phi_0^N|\Psi_0^N\rangle$, while the first term gives rise to

$$\tilde{s}\sum_i\langle N-1,r|a_i|\Psi_0^N\rangle\langle\phi_0^N|a_i^\dagger|\Psi_1^{N-1}\rangle. \tag{A7}$$

To obtain this result we made use of the anticommutation relations of the "exact" operators and finally, using Eq. (2.9) we transformed the remaining "exact" operator $\tilde{a}_i$ into the operator a_i. In Eq. (A7) $\tilde{s}$ is the number of "exact" occupied one-particle states in $|\Psi_1^{N-1}\rangle$, i.e., $\tilde{s}=N-2$.

Collecting all above results we have

$$\mathbf{Y}_{r1}\mathbf{S}_{11}=P_{1(r)}\mathbf{Y}_{r0}\mathbf{S}_{01}+\mathbf{S}_{r1}\langle\phi_0^N|\Psi_0^N\rangle. \tag{A8}$$

From this equation, recalling Eq. (A2) it is easy to derive the final form, Eq. (6.6b).

The successive columns of the matrix $\mathbf{U}$ are derived using the same procedure. For example, the third column is calculated starting with the analysis of the expression $\mathbf{Y}_{r2}\mathbf{S}_{22}$ and in general the r'th column is obtained by evaluating $\mathbf{Y}_{rr'}\mathbf{S}_{r'r'}$. The expressions which we derive by the transformations and the anticommutation rules of the operators are composed of more terms and are more complex since the number of operators which have to be handled is larger for larger r'. The general formula for the block $\mathbf{U}_{rr'}$ of $\mathbf{U}$ reads:

$$\mathbf{U}_{rr'}=\langle\phi_0^N|\Psi_0^N\rangle^{-1}\left\{\mathbf{Y}_{rr'}+\sum_{k=0}^{r'-1}c_{r,r',k}\mathbf{Y}_{rk}\mathbf{U}_{kr'}\right\}, \tag{A9a}$$

where

$$c_{r,r',k}=\frac{(-1)^{r'-k}}{(r'-k)!}\prod_{l=k+1}^{r'}(l-r). \tag{A9b}$$

[1] L. S. Cederbaum, J. Schirmer, and H.-D. Meyer, J. Phys. A **22**, 2427 (1989).

[2] B. H. Brandow, Rev. Mod. Phys. **39**, 771 (1967).

[3] V. Kvasnicka, Adv. Chem. Phys. **36**, 345 (1977).

[4] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).

[5] D. J. Thouless, *The Quantum Mechanics of Many-Body Systems* (Academic, New York, 1972).

[6] N. H. March, W. H. Joung, and S. Sampanthar, *The Many Body Problem in Quantum Mechanics* (Cambridge U. P., Cambridge, 1967).

[7] P. Jorgensen and J. Simon, *Second Quantization-Based Methods in Quantum Chemistry* (Academic, New York, 1981).

[8] A. Tarantelli and L. S. Cederbaum, Phys. Rev. A **39**, 1639 (1989).

[9] R. Zurmühl and S. Falk, *Matrizen und Ihre Anwendungen* (Springer, Berlin, 1984), Teil. 1.

[10] B. N. Parlett, *The Symmetric Eigenvalue Problem* (Prentice-Hall, Englewood Cliffs, NJ, 1980).

Path integral solution for a particle confined in a region

L. Chetouani
Departement de Physique Théorique, Institut de Physique, Université de Constantine, Constantine, Algeria

A. Chouchaoui
Laboratoire de Physique Théorique, Institut de Physique, U.S.T.H.B., Bab-Ezzouar, Alger, Algeria

T. F. Hammann
Laboratoire de Mathématiques, Physique Mathématique et Informatique, Faculté des Sciences et Techniques, Université de Haute Alsace, 4, rue des Frères Lumière, 68093 Mulhouse Cédex, France

(Received 24 August 1989; accepted for publication 22 November 1989)

The propagator relative to a particle constrained to move in a finite region of space is calculated in the framework of path integrals. This region of the three-dimensional space is delimited through a sector of opening angle α, and also through the action of two attractive harmonic potentials, one being central and located in the $0xy$ plan, and the other directed along the z axis, with respective pulsations ω and ω_0. It is shown that for $\alpha = \pi/2$ and π the propagator is the sum of propagators evaluated on classical paths. The important case of the edge ($\alpha = 2\pi$) is considered.

I. INTRODUCTION

In this paper, we present a calculation of the propagator relative to a particle confined in a region of the three-dimensional space, delimited through the action of an external force. This study is done in the framework of path integrals.

The external potential that acts upon the particle is made up of an angular part V_S and a harmonic part $V_{\rm HO}$. The first one, V_S, maintains the particle in a sector of opening angle α, situated in the plan $0xy$,

$$V_S(x,y) = V_S(\phi) = \begin{cases} 0, & \text{if } 0<\phi<\alpha, \\ \infty, & \text{elsewhere.} \end{cases}$$

The second potential is a sum of harmonic oscillators,

$$V_{\rm HO}(x,y,z) = \tfrac{1}{2}m\omega^2(x^2+y^2) + \tfrac{1}{2}m\omega_0^2 z^2,$$

of which the pulsations ω and ω_0 are of arbitrary value. For confined particles, the exact solution of the Schrödinger equation exists only for certain particular regions. For the region under consideration the case $\alpha = 2\pi$ has recently been solved exactly in the Schrödinger formalism[1] and through the image method.[2]

In the framework of the path integral formalism, only the pure sector ($V_{\rm HO} = 0$) has been treated.[3]

Our aim is to find the propagator for a particle moving inside the compound potential $V_S + V_{\rm HO}$, utilizing integrals that are calculated either directly or through the image method (Sec. II). For particular values of the opening angle α of the sector, $\alpha = \pi/2$ and $\alpha = \pi$, it is shown that the propagator collapses, as it should, into an algebraic sum on classical paths (Sec. III). Finally, the really important case $\alpha = 2\pi$ is analyzed. This case has been studied in the absence of harmonic forces, by Sommerfeld in the framework of optical diffraction through an edge.[4]

II. PROPAGATOR

In the canonical version the propagator in Cartesian coordinates is written

$$K(\mathbf{r}_f,\mathbf{r}_i;T) = \int \mathscr{D}x\mathscr{D}p_x\mathscr{D}y\mathscr{D}p_y\mathscr{D}z\mathscr{D}p_z \times \exp\left[\frac{i}{\hbar}\int_0^T dt(p_x\dot{x} + p_y\dot{y} + p_z\dot{z} - \frac{p_x^2+p_y^2+p_z^2}{2m} - V_S(x,y) - V_{\rm HO}(x,y,z)\right]. \tag{1}$$

The motion along the z axis being independent of the others, the propagator (1) is factorized in a product of two propagators:

$$K(\mathbf{r}_f,\mathbf{r}_i;T) = K_{xy}(x_f,y_f,x_i,y_i;T)K_z(z_f,z_i;T), \tag{2}$$

where

$$K_z(z_f,z_i;T) = \int \mathscr{D}z\mathscr{D}p_z \exp\left\{\frac{i}{\hbar}\int_0^T dt\left[p_z\dot{z} - \frac{p_z^2}{2m} - \frac{1}{2}m\omega_0^2 z^2\right]\right\} = \left[\frac{m\omega_0}{2i\pi\hbar\sin(\omega_0 T)}\right]^{1/2} \exp\left\{\frac{i}{\hbar}\frac{m\omega_0}{2\sin(\omega_0 T)} \times [(z_f^2+z_i^2)\cos(\omega_0 T) - 2z_f z_i]\right\}, \tag{3}$$

is the known propagator relative to the harmonic oscillator along the z axis,[5] and

$$K_{xy}(x_f,y_f,x_i,y_i;T) = \int \mathscr{D}x\mathscr{D}p_x\mathscr{D}y\mathscr{D}p_y \exp\left\{\frac{i}{\hbar}\int_0^T dt\,[p_x\dot{x} + p_y\dot{y} - ((p_x^2+p_y^2)/2m) - V_S(x,y) - \tfrac{1}{2}m\omega^2(x^2+y^2)]\right\}, \tag{4}$$

is the propagator describing the motion of the particle in the $0xy$ plan.

A. Calculation of K_{xy}

It is advantageous to change the coordinate system, and to go over to polar coordinates:

$$x = \rho \cos \phi,$$

$$y = \rho \sin \phi, \quad 0 \leqslant \rho < \infty \text{ and } 0 \leqslant \phi \leqslant 2\pi.$$

The propagator (4) has then the following form[6]:

$$K_{xy}(\rho_f,\phi_f,\rho_i,\phi_i;T)$$

$$= [\rho_f \rho_i]^{-1/2} \int \mathscr{D}\rho \mathscr{D}p_\rho \mathscr{D}\phi \mathscr{D}p_\phi \times \exp\left\{\frac{i}{\hbar}\int_0^T dt \left[p_\rho \dot{\rho} + p_\phi \dot{\phi} - \frac{p_0^2}{2m} - \frac{p_\phi^2 - \hbar^2/4}{2m\rho^2} - V_S(\phi) - \frac{1}{2} m\omega^2 \rho^2\right]\right\}, \quad (5)$$

the quantum correction $\hbar_2/8m\rho^2$ being exclusively the result of the coordinate change.

On the other hand, the motion of the particle in the plan Oxy is restricted to the interior of the sector ($V_S = \infty$ on the outside) and thus

$$K_{xy} = 0 \quad \text{if the points } (x_i, y_i) = (\rho_i, \phi_i)$$

$$\text{or } (x_f, y_f) = (\rho_f, \phi_f)$$

do not belong to the sector.

Then, if (x_i, y_i) and (x_f, y_f) belong to the sector, or if $0 < \phi_f, \phi_i < \alpha$, and $0 \leqslant \rho_f, \rho_i < \infty$, the propagator to be calculated becomes

$$K_{xy}(\rho_f,\phi_f,\rho_i,\phi_i;T) = (\rho_f \rho_i)^{-1/2} \int \mathscr{D}\rho \mathscr{D}p_\rho \mathscr{D}\phi \mathscr{D}p_\phi \times \exp\left\{\frac{i}{\hbar}\int_0^T dt \left[p_\rho \dot{\rho} + p_\phi \dot{\phi} - \frac{p_\rho^2}{2m} - \frac{p_\phi^2 - \hbar^2/4}{2m\rho^2} - \frac{1}{2} m\omega^2 \rho^2\right]\right\}. \quad (6)$$

Let us increase the domain of variation of ϕ and therefore the opening of the sector, via the following canonical transformation:

$$0 < \phi < \alpha \to 0 < \theta < \pi,$$
$$\theta = (\pi/\alpha)\phi; \quad p_\theta = (\alpha/\pi)p_\phi, \quad (7)$$

and if we take into consideration the transformation of the measure

$$\mathscr{D}\phi \mathscr{D}p_\phi = \lim_{N\to\infty} \prod_{j=1}^{N-1} d\phi_j \prod_{j=1}^{N} \frac{dp_{\phi_j}}{2\pi\hbar} = \frac{\pi}{\alpha} \mathscr{D}\theta \mathscr{D}p_\theta,$$

then (6) becomes

$$K_{xy} = \frac{\pi}{\alpha}(\rho_f \rho_i)^{-1/2} \int \mathscr{D}\rho \mathscr{D}p_\rho \mathscr{D}\theta \mathscr{D}p_\theta \times \exp\left\{\frac{i}{\hbar}\int_0^T dt \left[p_\rho \dot{\rho} + p_\theta \dot{\theta} - \frac{p_\rho^2}{2m} - \frac{(p_\theta \pi/\alpha)^2 - \hbar^2/4}{2m\rho^2} - \frac{1}{2} m\omega^2 \rho^2\right]\right\}. \quad (8)$$

Let us now utilize the following result obtained for the rigid rotator[6,7]:

$$\int \mathscr{D}\theta \mathscr{D}p_\theta \times \exp\left\{\frac{i}{\hbar}\int_0^T dt \left[p_\theta \dot{\theta} - \frac{p_\theta^2}{2ma^2} - \frac{\hbar^2(n^2 - 1/4)}{2ma^2 \sin^2\theta}\right]\right\}$$

$$= \sum_{l=0}^{\infty} \exp\left[-\frac{i}{\hbar}\frac{(l+n+1/2)^2 \hbar^2 T}{2ma^2}\right] \times \left(l + n + \frac{1}{2}\right)\frac{(l+2n)!}{l!}$$

$$(\sin\theta_f \sin\theta_i)^{1/2} P_{l+n}^{-n}(\cos\theta_f) P_{l+n}^{-n}(\cos\theta_i). \quad (9)$$

This formula is simplified when setting $n = 1/2$ and utilizing the relation[8]

$$P_{\nu-1/2}^{-1/2}(\cos\theta) = [2/(\pi \sin\theta)]^{1/2}(\sin(\nu\theta)/\nu),$$

with $\nu = l + 1$

In this case Eq. (9) becomes

$$\int \mathscr{D}\theta \mathscr{D}p_\theta \exp\left[\frac{i}{\hbar}\int_0^T dt \left(p_\theta \dot{\theta} - \frac{p_\theta^2}{2ma^2}\right)\right] = \frac{2}{\pi}\sum_{l=1}^{\infty} \exp\left[\frac{-i}{\hbar}\frac{\hbar^2 l^2 T}{2ma^2}\right]\sin(l\theta_f)\sin(l\theta_i). \quad (10)$$

Now Eq. (10) can also be obtained through the image method: it suffices to set $a = 1$, $L = \pi$ in the calculus of Ref. 9.

It is easy to verify with the help of the Poisson summation formula

$$\sum_{l=-\infty}^{+\infty} f(l) = \sum_{m=-\infty}^{+\infty} \int_{-\infty}^{+\infty} d\phi f(\phi) e^{2i\pi m\phi}$$

that Eq. (10) is the algebraic sum of the propagators relative to all possible classical paths.

Let us first perform the integration in (8) on $\mathscr{D}\theta \mathscr{D}p_\theta$, using the expression (10) (setting $a = \alpha/\pi$), and let us go back to the old variable ϕ [$\phi = (\alpha/\pi)\theta$]:

$$K_{xy} = \frac{2}{\alpha}\sum_{l=1}^{\infty} \sin\left(\frac{l\pi}{\alpha}\phi_f\right)\sin\left(\frac{l\pi}{\alpha}\phi_i\right) \times \left\{[\rho_f \rho_i]^{-1/2}\int \mathscr{D}\rho \mathscr{D}p_\rho \exp\left[\frac{i}{\hbar}\int_0^T dt \left[p_\rho \dot{\rho} - \frac{p_\rho^2}{2m} - \frac{\hbar^2}{2m\rho^2}\left(\left(\frac{l\pi}{\alpha}\right)^2 - \frac{1}{4}\right) - \frac{1}{2}m\omega^2\rho^2\right]\right]\right\}. \quad (11)$$

The variable p_ϕ has thus been quantized ($p_\phi \to \hbar l\pi/\alpha$) as it should.

Let us then perform the integration on $\mathscr{D}\rho \mathscr{D}p_\rho$. The path integral of the harmonic oscillator, supplemented by a centrifugal barrier, has been calculated a long time ago[6] through direct integration. This very integral has recently also been obtained via the image method.[10]

Thus, finally k_{xy} can be evaluated either through direct integration or through utilization of the image method.

Its expression reads

$$K_{xy}(\rho_f,\phi_f,\rho_i,\phi_i;T)$$

$$= \frac{2m\omega}{i\hbar\alpha \sin(\omega T)} \exp\left[\frac{im\omega}{2\hbar}(\rho_f^2 + \rho_i^2)\cot(\omega T)\right] + \sum_{l=1}^{\infty} \sin\left(\frac{l\pi}{\alpha}\phi_f\right)\sin\left(\frac{l\pi}{\alpha}\phi_i\right) I_{l\pi/\alpha}\left(\frac{m\omega\rho_f\rho_i}{i\hbar \sin(\omega T)}\right), \quad (12)$$

where I_ν is the modified Bessel function.

The product of the propagator (12) with the propagator (3) gives the explicit expression of the propagator defined in (1). It is then easy to deduce from it the spectrum and the corresponding wave functions.

B. Spectrum and wave functions

When setting $x=(m\omega/\hbar)\rho_f^2$, $y=(m\omega/\hbar)\rho_i^2$, $z=e^{-2i\omega T}$ ($|z|<1$), in the formula[11]

$$\frac{(xyz)^{-\nu/2}}{1-z}\exp\left[-z\frac{x+y}{1-z}\right]I_\nu\left(\frac{2(xyz)^{1/2}}{1-z}\right)$$
$$=\sum_{n=0}^{\infty}n!\frac{L_n^\nu(x)L_n^\nu(y)}{\Gamma(n+\nu+1)}z^n,$$

it is possible to separate the variables ρ_f, ρ_i, and T in (12). This leads to

$$K_{xy}=\frac{4m\omega}{\alpha\hbar}\exp\left[-\frac{m\omega}{2\hbar}(\rho_f^2+\rho_i^2)\right]\sum_{l=1}^{\infty}\sum_{n=0}^{\infty}\sin\left(\frac{l\pi}{\alpha}\phi_f\right)$$
$$\times\sin\left(\frac{l\pi}{\alpha}\phi_i\right)f_{nl}(\rho_f)f_{nl}(\rho_i)$$
$$\times\exp\left[-i\omega\left(2n+1+\frac{l\pi}{\alpha}\right)T\right], \qquad (13)$$

where

$$f_{nl}(\rho)=\left[\frac{2m\omega}{\hbar}\frac{n!}{\Gamma(n+1+l\pi/\alpha)}\right]^{1/2}$$
$$\times\left(\frac{m\omega}{\hbar}\rho^2\right)^{l\pi/2\alpha}\exp\left[\frac{-m\omega}{2\hbar}\rho^2\right]L_n^{l\pi/\alpha}\left(\frac{m\omega}{\hbar}\rho^2\right),$$

the L_n^ν being the Laguerre polynomials.

In the expression (3), the separation of the variables z_f, z_i, and T is also possible, and has been done a long time ago[5]:

$$K_z=\sum_{m=0}^{\infty}\Phi_m(z_f)\Phi_m(z_i)\exp\left[-i\omega_0\left(m+\frac{1}{2}\right)T\right], \qquad (14)$$

where

$$\Phi_m(z)=\left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4}\frac{1}{[2^n n!]^{1/2}}$$
$$\times\exp\left[-\frac{m\omega_0}{2\hbar}z^2\right]H_n\left(z\sqrt{\frac{m\omega_0}{\hbar}}\right),$$

the H_n being the Hermite polynomials.

The product of the propagators (13) and (14) leads then to the spectrum

$$E_{mnl}=\hbar\omega_0(m+\tfrac{1}{2})+\hbar\omega(2n+1+l\pi/\alpha),$$

as well as to the corresponding wave functions

$$\psi_{mnl}(\rho,\phi,z)=(2/\alpha)^{1/2}\sin(l\pi\phi/\alpha)f_{nl}(\rho)\Phi_m(z),$$

with $(m,n)=0,1,2,\ldots,\infty$, and $l=1,2,3,\ldots,\infty$.

III. PARTICULAR CASES

Let us now examine particular opening angles of the sector. Let us show that for $\alpha=\pi/2$, π, 2π, the propagator (1) [which is a product of the series (12) and the propagator (3)] can bé put into a compact form.

A. First case: $\alpha=\pi/2$

Thanks to the formula[12]

$$\cos[z\sin\phi]=J_0(z)+2\sum_{k=1}^{\infty}J_{2k}(z)\cos(2k\phi),$$

where ϕ is replaced by $\pi/2-\theta$, it is easy to show that the propagator (1) takes the following form:

$$K_{\pi/2}(\mathbf{r}_f,\mathbf{r}_i;T)=F(T)\{\exp[(i/\hbar)S_{cl}(A,D;T)]$$
$$+\exp[(i/\hbar)S_{cl}(I_3,D;T)]$$
$$-\exp[(i/\hbar)S_{cl}(I_1,D;T)]$$
$$-\exp[(i/\hbar)S_{cl}(I_2,D;T)]\}, \qquad (15)$$

where D designates the starting point, $\mathbf{r}_i=(x_i,y_i,z_i)$, of the particle, and A designates its end point $\mathbf{r}_f=(x_f,y_f,z_f)$ at final time. S_c is the action, evaluated along the classical path, of the particle submitted to the potential $V_{\mathrm{HO}}(xyz)$:

$$S_{cl}(A,D;T)$$
$$=(m\omega/2\sin(\omega T))[(x_f^2+y_f^2+x_i^2+y_i^2)\cos(\omega T)$$
$$-2x_fx_i-2y_fy_i]+m\omega_0/(2\sin(\omega_0T))$$
$$\times[(z_f^2+z_i^2)\cos(\omega_0T)-2z_fz_i]$$

and

$$F(T)=\frac{m\omega}{2i\pi\hbar\sin(\omega T)}\left[\frac{m\omega_0}{2i\pi\hbar\sin(\omega_0T)}\right]^{1/2}$$
$$=\left[\mathrm{Det}\left[\frac{i}{2\pi\hbar}\frac{\partial^2S_{cl}}{\partial\mathbf{r}_f\partial\mathbf{r}_i}\right]\right]^{1/2},$$

is the usual fluctuation factor.

The points I_1, I_2, and I_3 are the images of the point A with respect to the reflecting, vertical, and horizontal walls. Their coordinates are $(-x_f,y_f,z_f)$, $(x_f,-y_f,z_f)$, and $(-x_f,-y_f,z_f)$, respectively.

Thus for $\alpha=\pi/2$, the propagator (15) collapses into an algebraic sum on the four classical paths.

B. Second case: $\alpha=\pi$

Thanks to the Fourier expansion[13]

$$\exp[u\cos\phi]=\sum_{l=-\infty}^{+\infty}e^{il\phi}I_l(u),$$

it is easy to show that that propagator collapses into a sum on two classical paths:

$$K_\pi(\mathbf{r}_f,\mathbf{r}_i;T)=F(T)\{\exp[(i/\hbar)S_{cl}(A,D;T)]$$
$$-\exp[(i/\hbar)S_{cl}(I_2,D;T)]\}, \qquad (16)$$

I_2 being the image of D with respect to the horizontal wall.

C. Third case: $\alpha=2\pi$

Thanks to the formula[14]

$$\left[\frac{i}{\pi}\right]^{1/2}\exp[iz\cos(2\phi)]\int_{-\infty}^{\sqrt{2z}\cos\phi}\exp(-it^2)dt$$
$$=\frac{1}{2}J_0(z)+\sum_{k=1}^{\infty}\exp\left[\frac{i\pi k}{4}\right]J_{k/2}(z)\cos(k\phi),$$

one can show that in this case:

$$
\begin{aligned}
&K_{2\pi}(\mathbf{r}_f,\mathbf{r}_i;T)\\
&\quad=\frac{1}{2}F(T)\Big\{\exp\Big[\frac{i}{\hbar}S_{cl}(A,D;T)\Big]\\
&\qquad\times\Big(1+\mathrm{erf}\Big[\Big(\frac{2m\omega\rho_f\rho_i}{i\hbar\sin(\omega T)}\Big)^{1/2}\\
&\qquad\times\cos\Big(\frac{\phi_f-\phi_i}{2}\Big)\Big]\Big]\Big)-\exp\Big[\frac{i}{\hbar}S_{cl}(I_2,D;T)\Big]\\
&\qquad\times\Big\{1+\mathrm{erf}\Big[\Big(\frac{2m\omega\rho_f\rho_i}{i\hbar\sin(\omega T)}\Big)^{1/2}\cos\Big(\frac{\phi_f+\phi_i}{2}\Big)\Big]\Big]\Big\}\Big\},
\end{aligned}
\tag{17}
$$

where

$$\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z\exp(-u^2)du$$

is the standard error function.

One can see that in this case the propagator (17) is decomposed into two parts: one associated with the two classical paths (incidental and reflexive) and the other corresponding to the edge diffraction, expressed through the terms containing the error functions. Although the error functions are superpositions of Gaussians, it is not possible, in this case, to have for the propagator a sum on classical paths.[15]

IV. CONCLUSION

We have calculated exactly in the Feynman approach the propagator (1) of a particle confined in a region, and we have made obvious the fact that one can obtain Eq. (12) and thus (1) either through a direct integration calculus or through the image method. We have shown that for the particular cases of opening angles $\alpha=\pi/2$ and π, the propagator takes quite a remarkable form. It is possible to show in a general fashion that for opening angles α of the sector, such as the ratio π/α is an integer, the propagator collapses into a sum of propagators evaluated on classical paths. For the important case $\alpha=2\pi$ the expression (17)—the main result of this paper—was already obtained in Ref. 1 (but one has to replace ϕ by $\pi+\theta$), in the Schrödinger formalism.

Finally, by suppressing the propagator K_z and by making ω tend towards 0, one comes back to the results of the pure sector.[3]

ACKNOWLEDGMENTS

The authors would like to express their appreciation to Dr. Jean Richert, Centre de Recherches Nucléaires, Strasbourg, France, and to Professor Robert Lutz, Faculté des Sciences, Université de Haute-Alsace, Mulhouse, France, for a critical reading of the manuscript and for valuable suggestions.

[1] F. W. Wiegel and P. W. Van Andel, J. Phys. A: Math. Gen. **20**, 627 (1987).

[2] R. E. Crandall, J. Phys. A **16**, 513 (1983).

[3] L. Chetouani, L. Guechi, and T. F. Hammann, Nuovo Cimento B **101**, 547 (1988).

[4] A. J. W. Sommerfeld, *Optics* (Academic, New York, 1954).

[5] R. P. Feynman and H. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

[6] D. Peak and A. Inomata, J. Math. Phys. **10**, 1422 (1969); W. Langguth and A. Inomata, *ibid.* **20**, 499 (1979).

[7] K. Pak and I. Sokmen, Phys. Lett. A **103**, 298 (1984).

[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1965), p. 1008, Eq. (8.754.3).

[9] W. Janke and H. Kleinert, Lett. Nuovo Cimento **25**, 297 (1979).

[10] I. H. Duru, Phys. Lett. A **112**, 421 (1985).

[11] Reference 8, p. 1038, Eq. (8.976.1).

[12] Reference 8, p. 974, Eq. (8.514.4).

[13] Reference 8, p. 973, Eq. (8.511.4).

[14] Reference 8, p. 973, Eq. (8.511.5).

[15] L. S. Schulman, Phys. Rev. Lett. **49**, 599 (1982).

Criteria for the Kato smoothness with respect to a dispersive N-body Schrödinger operator

Jan Derezinski
Division of Mathematical Methods in Physics, Warsaw University, Hoza 74, 00-682 Warsaw, Poland

(Received 18 October 1988; accepted for publication 6 September 1989)

The operators of the form $H = \omega(D) + \Sigma v_a(\pi^a x)$, which are a natural generalization of N-body Schrödinger operators, are studied. Certain criteria are presented that allow one to verify if a given pseudodifferential operator is Kato smooth with respect to H.

I. INTRODUCTION

Let $\mathscr{H}$ be a Hilbert space, Δ a subset of $\mathbb{R}$, H a self-adjoint operator, and C a bounded operator. Let $E_\Delta(H)$ denote the spectral projection of H onto Δ. We say that C is H-smooth (or Kato smooth with respect to H) on Δ if and only if, for any $\psi \in \operatorname{Ran} E_\Delta(H)$,

$$\int_{-\infty}^{\infty} \|Ce^{iHt}\psi\|^2\, dt < \infty .$$

This concept has been introduced by Kato.[1,2] The H-smooth operators appear very often in the theory of Schrödinger operators, especially when we want to prove the absence of the singular continuous spectrum and the asymptotic completeness of the wave operators.[1,3–11] The theory of H-smooth operators is especially interesting in the context of N-body Schrödinger operators. It can be shown, for instance, that if Δ is a compact interval that avoids the thresholds and eigenvalues of a fairly general N-body Schrödinger operator H, then $(|x|+1)^{-1/2-\epsilon}$ is H-smooth on Δ. This was proved in the three-body case in Refs. 12 and 13 and in the N-body case in Ref. 14 (see, also, Refs. 15 and 16). There also exists an interesting different proof of this fact.[17]

The importance of H-smooth operators is especially evident in the remarkable paper[11] by Sigal and Soffer devoted to the proof of the asymptotic completeness of the short range N-body scattering. The proof that the operators from a sufficiently rich family are Kato smooth with respect to an N-body Schrödinger operator was the crucial step of the proof contained in Ref. 11 (see, also, Ref. 18).

Regular N-body Schrödinger operators are self-adjoint operators on $L^2(\mathbb{R}^{3N})$ of the form

$$H = -\sum_{i=1}^{N} \frac{1}{2m_i}\Delta_i + \sum_{i,j=1}^{N} v_{ij}(x_i - x_j) \qquad (1.1)$$

(see, e.g., Refs. 11, 14, and 19–21). By an appropriate choice of the coordinates in the configuration space the kinetic energy operator can be made equal to minus the Laplacian. The fact that the potentials v_{ij} depend on the differences of the positions of the particles is often not essential in the mathematical analysis of such operators. Thus following Refs. 22–24 instead of (1.1) one can study operators on $L^2(X)$ of the form

$$H = -\Delta + \sum_{a\in\mathscr{A}} v_a(\pi^a x),$$

where X is a Euclidean space isomorphic to $\mathbb{R}^n$, $\{X_a : a\in\mathscr{A}\}$ is a family of its subspaces, $X^a = X/X_a$, the π^a are the canonical surjections from X onto X^a, and the v_a are the real functions on X^a.

One can go one step further. Let us forget about the scalar product in X; let K be the space dual to X, ω a real function on K and $D = (1/i)\nabla$. By a dispersive N-body Schrödinger operator we will mean a self-adjoint operator on $L^2(X)$ of the form

$$H = \omega(D) + \sum_{a\in\mathscr{A}} v_a(\pi^a x) . \qquad (1.2)$$

Such operators have a lot in common with regular N-body Schrödinger operators [for which $\omega(k)$ can be made equal to k^2]. They seem to be of a significant physical interest, e.g., in the physics of interacting relativistic particles or in solid state physics.

This paper is devoted to the study of H-smooth operators, where H is of the type (1.2). We are partly motivated by the possibility of applying our results in the scattering theory of such operators. Besides, we think that H-smooth operators are interesting for their own sake providing some nontrivial information about the properties of the evolution e^{iHt}. Loosely speaking, the H-smoothness of C on Δ means that the particles with the energy in Δ "spend a finite amount of time in B^*B."

We will assume in this paper that the potentials v_a decay in some sense at least as $|\pi^a x|^{-\mu}$, where $\mu > 0$. We will assume, moreover, that for some fixed open interval Δ_0 and any $\epsilon > 0$ the operator $(|x|+1)^{-1/2-\epsilon}$ is H-smooth on Δ_0. It turns out that this property can be proved for a certain class of dispersive N-body Schrödinger operators and an appropriately chosen Δ_0 by using the so-called Mourre estimate (see Ref. 25 and, also, Refs. 12–16 and 24). As we said earlier, this property is well known in the case $\omega(k) = k^2$, if $\overline{\Delta}_0$ does not contain thresholds or bound states.

Throughout this paper we will try to find conditions on the symbol w of a pseudodifferential operator $w(x,D)$ that will guarantee the H-smoothness of $w(x,D)$ on Δ if $\overline{\Delta} \subset \Delta_0$. It turns out that this property depends to a great extent on the phase space support of w. In particular, it is natural to introduce the following definition. Let Ω be a subset of the phase space $X \times K$ and $\Delta \subset \mathbb{R}$. We say that $\Omega \in \mathscr{N}\mathscr{P}_\Delta$ if and only if $w \in S^{-1/2}$ and supp $w \subset \Omega$ imply the H-smoothness of $w(x,D)$ on Δ. (Here S^m denotes the set

$$\{w \in C^\infty(X\times K) \colon |\partial_x^\alpha \partial_k^\beta w| \leqslant c_{\alpha\beta}(|x|+1)^{m-|\alpha|}\}.)$$

0022-2488/90/040842-09$03.00

The letters $\mathscr{N}$ and $\mathscr{P}$ in $\mathscr{N}\mathscr{P}_\Delta$ stand for "no propagation." The family $\mathscr{N}\mathscr{P}_\Delta$ is a variation of the concept of the propagation set used in Ref. 11. A very similar concept was used also in Ref. 18.

Now let us say a few words on how the paper is organized. The first three sections give basic definitions. In Sec. IV, we study a certain subset of the phase space that we call the energy shell. Roughly speaking, we show that the complement of the energy shell belongs to $\mathscr{N}\mathscr{P}_\Delta$ if $\overline{\Delta}\subset\Delta_0$. This fact is easy to prove and was essentially known before.[26] Its proof is based on the so-called geometric method (see, e.g., Refs. 14, 15, 24, and 27–30).

Section V introduces a more powerful technique for finding H-smooth operators that goes back to Putnam and Kato.[2,3,9,11] The main idea of this method is the following fact: if B is bounded and $i[H,B]$ is positive, then the square root of $i[H,B]$ is H-smooth. We use this method jointly with the geometric method and the calculus of pseudodifferential operators to obtain an interesting criterion for finding sets belonging to $\mathscr{N}\mathscr{P}_\Delta$. Roughly speaking this criterion says that if $\Omega\subset X\times K$, $u\in S^0$, $\overline{\Delta}\subset\Delta_0$ and the Poisson bracket of ω and u is non-negative on the energy shell and strictly positive on Ω, then $\Omega\in\mathscr{N}\mathscr{P}_\Delta$.

In Sec. VI we try to formulate a more constructive criterion for verifying if a given Ω belongs to $\mathscr{N}\mathscr{P}_\Delta$. This criterion, we think, is based on an intuitively appealing analysis of the geometry of the phase space. Roughly speaking, we try to understand better the asymptotic properties of the evolution e^{iHt} by thinking in terms of classical physics.

Unfortunately, the family of H-smooth operators that we can get by using the methods of this paper is rather limited. In particular, in the case $\omega(k) = k^2$, Sigal and Soffer had to find a different family of H-smooth operators in order to prove the asymptotic completeness of the N-body short range scattering (see the propagation theorem in Refs. 11 and 18). Let us point out that the techniques that we use in our paper are somewhat different than those used in Refs. 11 and 18. Looseley speaking, in this paper we avoid studying what happens if the particles interact and we concentrate on the free motion. The interaction enters our considerations chiefly through the confinement of the particles to the energy shell and the possible absence of conservation of the momentum in some regions of configuration space.

The new ingredient that makes it possible to obtain a richer family of H-smooth operators in the case $\omega(k) = k^2$ in Refs. 11 and 18 is the exploitation of the properties of the operator $\gamma=\frac{1}{2}(D\cdot(x/|x|)+(x/|x|)\cdot D)$ and the application of the Mourre estimate. Unfortunately, there seems to be no natural generalization of the operator γ in the case of a general dispersive $\omega(k)$. Nevertheless, the family of the H-smooth operators that one can obtain by using the methods of this paper, while not so spectacular as the one described in Refs. 11 and 18, is also quite nontrivial and interesting.

II. GEOMETRY OF THE PHASE SPACE

Throughout this paper X will denote a vector space isomorphic to $\mathbb{R}^n$; $\{X_a : a\in\mathscr{A}\}$ will be a certain finite family of its subspaces. To be consistent with the notation used in the literature we will write $a_1\subset a_2$ whenever $X_{a_1}\supset X_{a_2}$ and $a_1\cup a_2=a_3$ whenever $X_{a_1}\cap X_{a_2}=X_{a_3}$ (see, e.g., Refs. 19–23). Here $X_{a_{\min}}$ will denote X and $X_{a_{\max}}=\{0\}$. We will assume the following properties of $\mathscr{A}$:

(1) $a_{\min}\in\mathscr{A}$,

(2) $a_{\max}\in\mathscr{A}$,

(3) if $a_1,a_2\in\mathscr{A}$, then $a_1\cup a_2\in\mathscr{A}$.

We will denote X/X_a by X^a. The spaces dual to X_a, X, and X^a will be denoted by K_a, K, and K^a. We will fix a scalar product in X that will enable us to identify X^a with a certain subspace of X and K_a with a certain subspace of K. The symbol π_a will denote the projections of K and X onto K_a and X_a; π^a will be the projections of K and X onto K^a and X^a.

If $y\in X$, then let us define the unitary operator U_y on $L^2(X)$ such that $(U_y\varphi)(x)=\varphi(x-y)$. An operator $B\in B(L^2(X))$ will be called a-fibered if and only if $U_yB=BU_y$, for $y\in X_a$.[9,14] Such operators can be decomposed as

$$B=\int_{K_a}^{\oplus} dk_a\, B(k_a)\,,$$

where $k_a\mapsto B(k_a)$ is a function from $L^\infty(K_a,B(L^2(X^a)))$. Here, $|x|$ will denote the Euclidean norm of x; $B(y,r)$ will denote the ball of center y and radius r.

We fix a certain fixed positive C^∞ function $X\ni x\mapsto\langle x\rangle\in\mathbb{R}$ such that, for $|x|>1$, we have $\langle x\rangle=|x|$.

We also define

$$S^m_\delta(X\times K_a)$$
$$=\{u\in C^\infty(X\times K_a):\ |\partial^\alpha_x\partial^\beta_{k_a}u|\leqslant c_{\alpha\beta}\langle x\rangle^{m-|\alpha|\delta}\}\,.$$

We will usually write S^m_δ instead of $S^m_\delta(X\times K)$ and $S^m_\delta(X)$ instead of $S^m_\delta(X\times\{0\})$.

Let $0\leqslant\epsilon$, $0\leqslant\delta\leqslant1$, and $\mathscr{U}\subset X$. Define

$$\mathscr{U}^{\epsilon,\delta}=\{x\in X:\ \operatorname{dist}(x,\mathscr{U})\leqslant\epsilon|x|^\delta\}\,.$$

If $0\leqslant\kappa$ and $Q\subset K_a$, we define

$$Q^\kappa=\{k_a\in K_a:\ \operatorname{dist}(k_a,Q)\leqslant\kappa\}\,.$$

If $\Omega\subset X\times K$, we define

$$\Omega^{\epsilon,\delta,\kappa}=\overline{\bigcup_{(x,k)\in\Omega}\{x\}^{\epsilon,\delta}\times\{k\}^\kappa}\,.$$

A subset Ω of $X\times K$ is called conical if and only if $(x,k)\in X\times K$ and $\lambda\in\mathbb{R}$ implies $(\lambda x,k)\in\Omega$. It is called a-fibered if and only if $(x,k)\in\Omega$ and $\pi_ak=\pi_ak'$ implies $(x,k')\in\Omega$.

We define $Z_a=X_a\setminus\cup_{b\not\subset a}X_b$ and $Z_a^{(\epsilon,\delta)}=X_a^{\epsilon,\delta}\setminus\cup_{b\not\subset a}X_b^{\epsilon,\delta}$. Note that $X=\cup_{a\in\mathscr{A}}Z_a^{(\epsilon,\delta)}$.

In all the above symbols we will usually drop δ if $\delta=1$.

It is easy to see that there exists $\epsilon_0>0$ such that if $0\leqslant\epsilon\leqslant\epsilon_0$, $a\not\subset b$, and $b\not\subset a$, then $X_a^\epsilon\cap X_b^\epsilon=\{0\}$. For such ϵ, the family $\{Z_a^{(\epsilon)}: a\in\mathscr{A}\}$ is a partition of X into disjoint sets.

The rest of this section is devoted to the construction of a family of special partitions of unity on X. Similar partitions of unity (usually with $\delta=1$) are a typical ingredient of the geometric method.[11,23,27–30]

Proposition 2.1: Let $0<\delta<1$ and $0<\epsilon_1<\epsilon_2$; or $\delta=1$

and $0<\epsilon_1<\epsilon_2<\epsilon_0$. Then there exists a family of functions j_a such that

(a) $j_a \in S_\delta^0(X)$;

(b) $0 \leqslant j_a \leqslant 1$;

(c) $\operatorname{supp} j_a \subset (X_a^{\epsilon_2,\delta} \setminus \cup_{b \not\subset a} X_b^{\epsilon_1,\delta}) \cup B(0,R)$, for some R;

(d) $\sum_{a \in \mathscr{A}} j_a = 1$.

The family $\{ j_a : a \in \mathscr{A} \}$ described in Proposition 2.1 will be called a δ partition of unity. In the proof of the above proposition we will need the following property of the sets $A^{\epsilon,\delta}$; this property is easy to show and we omit its proof.

Lemma 2.2: Let $0<\epsilon_1<\alpha<\epsilon_2$, $0 \leqslant \delta \leqslant 1$, and $A \subset X$. Then there exits $0<\beta$ such that

$$(A^{\alpha,\delta})^{\beta,\delta} \subset A^{\epsilon_2,\delta} \cup B(0,1)$$

and

$$(X \setminus A^{\alpha,\delta})^{\beta,\delta} \subset (X \setminus A^{\epsilon_1,\delta}) \cup B(0,1) .$$

Now let us choose $f \in C_0^\infty(X)$ such that $f \geqslant 0$, $\int f(x)dx = 1$, and $\operatorname{supp} f \subset B(0,1)$. Define

$$f^{(\epsilon,\delta)}(x,x') = [\langle x \rangle^\delta \epsilon]^{-n} f(x'/\epsilon\langle x \rangle^\delta) ,$$

where $n = \dim X$. We easily show the following lemma.

Lemma 2.3: Let g be a bounded measurable function on X. Define

$$g^{\epsilon,\delta}(x) = \int g(x-x') f^{(\epsilon,\delta)}(x,x') dx' .$$

Then

(a) $1^{\epsilon,\delta} = 1$,

(b) $\operatorname{supp} g^{\epsilon,\delta} \subset (\operatorname{supp} g)^{\epsilon,\delta} \cup B(0,2)$,

(c) $g^{\epsilon,\delta} \in S_\delta^0(X)$.

Now we can construct our partition of unity.

Proof of Proposition 2.1: We will restrict ourselves to the case $0 \leqslant \delta < 1$. Note that if $a \neq b$, then $Z_a^{(\alpha,\delta)} \cap Z_b^{(\alpha,\delta)}$ is a bounded set. Let $\epsilon_1 < \alpha < \epsilon_2$ and let χ_a be the characteristic function of $Z_a^{(\alpha,\delta)}$. Take $\beta > 0$ such that

$$(Z_a^{(\alpha,\delta)})^{\beta,\delta} \subset X_a^{\epsilon_2,\delta} \cup B(0,1)$$

and

$$(X \setminus Z_a^{(\alpha,\beta)})^{\beta,\delta} \subset (X \setminus X_a^{\epsilon_1,\delta}) \cup B(0,1) .$$

Then the $\chi_a^{\beta,\delta}$, constructed as in Lemma 2.3, satisfy (a)–(c). They also satisfy (d) outside a certain bounded set. Thus it is enough to change $\chi_a^{\epsilon,\delta}$ appropriately within a bounded set to get j_a satisfying the required conditions. Q.E.D.

III. THE HAMILTONIAN

In this section we state precisely the assumptions on N-body dispersive Schrödinger Hamiltonians that we will use in our paper.

Hypothesis A: Let ω be a real valued C^∞ function on K such that

(a) $\omega \geqslant 0$,

(b) $\lim_{|k| \to \infty} \omega(k) = \infty$,

(c) $|\partial_k^\alpha \omega(k)| \leqslant c_\alpha (\omega(k)+1)$, for all multi-indices α,

(d) there exist c and N such that $\omega(k) \leqslant c\langle k-k' \rangle^N \omega(k')$.

Note that hypothesis A is satisfied if $\omega(k) = k^2$.

The unboundedness of ω will cause some technical problems. To deal with them it will be helpful to introduce the spaces

$$S_\delta^m(\omega+1) = \{u \in C^\infty(X \times K): \\ |\partial_x^\alpha \partial_k^\beta u| \leqslant c_{\alpha\beta} \langle x \rangle^{m-|\alpha|\delta} (\omega(k)+1)\} .$$

Note that $\omega \in S_0^0(\omega+1)$. The properties of pseudodifferential operators with symbols from $S_\delta^m(\omega+1)$ are the subject of the Appendix.

Next we describe the assumptions that we will impose on the potentials.

Hypothesis B: Suppose that $1 \geqslant \mu > 0$ and, for any $a \in \mathscr{A}$, we have a real function v_a on X^a such that

(a) $(\omega(D)+1)^{-1} v_a(\pi^a x) \langle \pi^a x \rangle^\mu$ is bounded,

(b) $(\omega(D)+1)^{-1} \nabla v_a(\pi^a x) \langle \pi^a x \rangle^{1+\mu}$ is bounded,

(c) $\lim_{\lambda \to \infty} \|(\omega(D)+\lambda)^{-1} v_a(\pi^a x)\| = 0$.

Note that Hypothesis B(c) guarantees that $v_a(\pi^a x)$ is a relatively bounded perturbation of $\omega(D)$ with an arbitrarily small bound.

Let

$$V(x) = \sum_{a \in \mathscr{A}} v_a(\pi^a x) .$$

We define H to be the self-adjoint operator such that $\mathscr{D}(H) = \mathscr{D}(\omega(D))$ and $H = \omega(D) + V$.

Unfortunately, unlike Hypotheses A and B, our third assumption on the Hamiltonian is implicit.

Hypothesis C: We fix a certain open bounded interval Δ_0 and we assume that for any $\epsilon > 0$ the operator $(|x|+1)^{-1/2-\epsilon}$ is H-smooth on Δ_0.

For a discussion of hypothesis C we refer the reader to the Introduction.

We also need to define

$$V_a(x) = \sum_{b \subset a} v_b(\pi^b x)$$

and $I_a = V - V_a$. Then the so-called "cluster Hamiltonians" H_a are defined as self-adjoint operators such that $\mathscr{D}(H_a) = \mathscr{D}(\omega(D))$ and $H_a = \omega(D) + V_a$. Note that $H_{a_{\min}} = \omega(D)$ and $H_{a_{\max}} = H$. We will often write H_0 instead of $\omega(D)$.

Note that cluster Hamiltonians H_a are a-fibered and we can write

$$H_a = \int_{K_a}^{\oplus} dk_a \, H_a(k_a) \, .$$

Let $\Delta \subset \mathbb{R}$. We define

$$ES_a(\Delta) = \{k_a \in K_a\colon \ \Delta \cap \sigma(H_a(k_a)) \neq 0\},$$

where $\sigma(B)$ denotes the spectrum of B. It is easy to show that for any bounded Δ the set $ES_a(\Delta)$ is bounded.

An important role will be played in our paper by the following subset of $X \times K$:

$$\mathscr{E}\mathscr{S}(\Delta) = \bigcup_{a \in \mathscr{A}} X_a \times \pi_a^{-1}(ES_a(\Delta)) \, .$$

Here $\mathscr{E}\mathscr{S}(\Delta)$ will be called the energy shell for the energy range Δ.

IV. PROPAGATION OUTSIDE THE ENERGY SHELL

It is intuitively clear that quantum particles should tend to "live" in a neighborhood of the energy shell. This intuition can be expressed by the following theorem.

Theorem 4.1: Let $\kappa > 0$, $m < -\frac{1}{2} + \mu$, and $\overline{\Delta} \subset \Delta_0$. Let $u \in S^m$ and $u = 0$ on $\mathscr{E}\mathscr{S}_{\Delta_0}^{0,\kappa}$. Then $u(x,D)$ and $u(x,D)^*$ are H-smooth on Δ.

The above theorem immediately implies the following corollary.

Corollary 4.2: Let $\kappa > 0$ and $\overline{\Delta} \subset \Delta_0$. Then $X \times K \setminus \mathscr{E} S_{\Delta_0}^{0,\kappa} \in \mathscr{N}\mathscr{P}_{\Delta}$.

The following lemma will be needed in the proof of Theorem 4.1.

Lemma 4.3: Suppose that $1 \geqslant \delta > 0$, $z \in S^m$, and j_a is an element of a δ partition of unity constructed in Proposition 2.1. Then

$$j_a(x)(z(x,k) - z(\pi_a x,k)) \in S_\delta^{m-1+\delta} \tag{4.1}$$

and

$$j_a(x) z(\pi_a x,k) \in S_\delta^m \, . \tag{4.2}$$

Proof: First note that

$$\partial_x^\alpha \partial_k^\beta [j_a(x)(z(x,k) - z(\pi_a x,k))] \tag{4.3}$$

is a linear combination of terms

$$\partial_x^{\alpha_1} j_a(x) \partial_x^{\alpha_2} \partial_k^\beta (z(x,k) - z(\pi_a x,k)) \, , \tag{4.4}$$

where $\alpha_1 + \alpha_2 = \alpha$. Now (4.4) equals

$$\partial_x^{\alpha_1} j_a(x) \int_0^1 (\pi^a x) \cdot \nabla_x \, \partial_x^{\alpha_2} \partial_k^\beta z(\pi_a x + \tau \pi^a x, k) d\tau \, . \tag{4.5}$$

The absolute value of (4.5) is less than or equal to

$$c_{\alpha\beta} \langle x \rangle^{-|\alpha_1|\delta} |\pi^a x| \int_0^1 d\tau \langle \pi_a x + \tau \pi^a x \rangle^{m - (|\alpha_a| + 1)}$$
$$\leqslant c_{\alpha\beta} \langle x \rangle^{-(|\alpha_1| - 1)\delta + m - (|\alpha_2| + 1)} \, .$$

(We used the fact that on the support of j_a we have $|\pi^a x| \leqslant \epsilon \langle x \rangle^\delta$ and $\langle \pi_a x + \tau \pi^a x \rangle \leqslant c \langle x \rangle$.)

Consequently, the absolute value of (4.3) is less than or equal to $c'_{\alpha\beta} \langle x \rangle^{m - 1 + \delta - |\alpha|\delta}$. This implies (4.1). Now, (4.2) follows from (4.1) and the fact that $j_a(x) z(x,k) \in S_\delta^m$.
Q.E.D.

The next facts needed in the proof of Theorem 4.1 can be easily shown by using the methods of Sec. 5 of Ref. 18 and by the calculus of pseudodifferential operators contained in the Appendix. (The methods of Sec. 5 of Ref. 18 belong to the standard folklore of Schrödinger operators; they are based especially on the techniques of Ref. 11.)

Lemma 4.4: (a) Let $u \in S_0^m$. Then $\langle x \rangle^\lambda u(x,D) \langle x \rangle^{-m-\lambda}$ is bounded for any $\lambda \in \mathbb{R}$.

(b) Let $F \in C_0^\infty(\mathbb{R})$. Then $\langle x \rangle^\lambda F(H)(H_0 + 1) \langle x \rangle^{-\lambda}$ is bounded for any λ.

(c) Suppose that $F \in C_0^\infty(\mathbb{R})$, $a \in \mathscr{A}$, and j_a is an element of a δ partition of unity. Then

$$\langle x \rangle^\lambda (F(H) - F(H_a)) j_a(x) \langle x \rangle^{\mu\delta - \lambda}$$

is bounded for any λ.

Now we are ready for the proof of Theorem 4.1.

Proof of Theorem 4.1: Clearly if $u \in S^m$, then

$$u(x,D) - u(x,D)^* = B \langle x \rangle^{m-1} \, ,$$

for some bounded operator B. Thus it is enough to show that $u(x,D)^*$ is H-smooth on Δ.

Choose $\rho_a \in C_0^\infty(K_a)$ such that $\rho_a = 1$ on $ES_a(\Delta_0)$ and $\operatorname{supp} \rho_a \subset ES_a(\Delta_0)^\kappa$. Let $0 < \delta < 1$ and let j_a be a δ partition of unity. Choose $F \in C_0^\infty(\mathbb{R})$ such that $\operatorname{supp} F \subset \Delta_0$ and $F = 1$ on Δ. Let us prove the following fact.

Lemma 4.5: Let ρ_a, j_a, and F be as above and let $u \in S_\delta^m$. Then, for any λ,

$$\langle x \rangle^\lambda F(H) \Big[u(x,D) - \sum_{a \in \mathscr{A}} j_a(x) u(\pi_a x,D) \rho_a(D_a) \Big]$$
$$\times \langle x \rangle^{-m + \min(\delta\mu, 1 - \delta) - \lambda}$$

is bounded.

Proof: Write

$$F(H) \Big[u(x,D) - \sum_{a \in \mathscr{A}} j_a(x) u(\pi_a x,D) \rho_a(D_a) \Big]$$
$$= \sum_{a \in \mathscr{A}} \{ [F(H) - F(H_a)] j_a(x) u(x,D) [1 - \rho_a(D_a)]$$
$$+ F(H_a) [\rho_a(D_a), j_a(x) u(x,D)]$$
$$+ F(H) j_a(x) [u(x,D) - u(\pi_a x,D)] \rho_a(D_a) \}$$
$$= C_1 + C_2 + C_3 \, .$$

Now, by Lemmas 4.3 and 4.4 and Proposition A1, the following operators are bounded for any λ:

$$\langle x \rangle^\lambda C_1 \langle x \rangle^{-m + \mu\delta - \lambda} \, ,$$
$$\langle x \rangle^\lambda C_2 \langle x \rangle^{-m + \delta - \lambda} \, ,$$

and

$$\langle x \rangle^\lambda C_3 \langle x \rangle^{-m + 1 - \delta - \lambda} \, .$$

This immediately implies our statement. Q.E.D.

Now we continue with the proof of Theorem 4.1. We set

$$B = \langle x \rangle^{-m + \min(\delta\mu, 1 - \delta)} F(H) u(x,D) \, .$$

Clearly $j_a(x) u(\pi_a x,D) \rho_a(D_a) = 0$, for any $a \in \mathscr{A}$. Thus Lemma 4.5 implies the boundedness of B. But

$$u(x,D)^* E_\Delta(H) = B^* \langle x \rangle^{m - \min(\delta\mu, 1 - \delta)} E_\Delta(H) \, .$$

Thus the theorem follows from hypothesis C. Q.E.D.

V. THEOREM ON POSITIVE POISSON BRACKETS

It is well known that if we find a bounded operator B such that the commutator $i[H,B]$ is positive then the square root of $i[H,B]$ is H-smooth. This fact is known as the Putnam–Kato theorem.[2,3,9] It is easy to show a slightly more general version of this fact.

Lemma 5.1: Let B, C, C_i, and C'_i, for $i = 1,...,k$, belong to $B(\mathscr{H})$; let H be a self-adjoint operator on $\mathscr{H}$; and let $\Delta \subset \mathbb{R}$. Suppose that C_i and C'_i are H-smooth on Δ and

$$E_\Delta(H)i[H,B]E_\Delta(H)$$
$$\geqslant E_\Delta(H)C^*CE_\Delta(H) + \sum_{i=1}^{k} E_\Delta(H)C_i^*C'_iE_\Delta(H). \tag{5.1}$$

Then C is H-smooth on Δ.

In this section we use Lemma 5.1 to obtain an interesting criterion for the local H-smoothness of pseudodifferential operators. This criterion is the central result of our paper. Its advantage over Lemma 5.1 lies in the fact that it is expressed in terms of functions on the phase space $X \times K$, instead of operators on $L^2(X)$. The positivity of $E_\Delta(H)i \times [H,B]E_\Delta(H)$ is replaced in this criterion by the positivity of the Poisson bracket of the kinetic energy $\omega(k)$ and a certain function on the energy shell.

Theorem 5.2: Suppose that $\overline{\Delta} \subset \Delta_0$ and $u \in S^0$. Assume, moreover, the following three conditions.

(i) Suppose that $\epsilon > 0$ and Q_a are bounded subsets of K_a. Let

$$u(x,k) = \sum_{a \in \mathscr{A}} u_a(x,\pi_a k),$$
$$u_a \in S^0(X \times K_a),$$

and

$$\operatorname{supp} u_a \subset \Big(X \setminus \bigcup_{b \not\subset a} X_b^\epsilon\Big) \times Q_a.$$

(ii) Suppose that $\kappa > 0$ and w, w_i, $w'_i \in S^{-1/2}$, for $i = 1,...,k$. Assume, moreover, that

$$\{\omega,u\} - |w|^2 - \sum_{i=1}^{k} \overline{w}_i w'_i \geqslant 0,$$

on $\mathscr{E}\mathscr{S}(\Delta_0)^{0,\kappa}$.

(iii) Let $w_i(x,D)$ and $w'_i(x,D)$ be H-smooth on Δ.

Then $w(x,D)$ is H-smooth on Δ.

The following corollary is an immediate consequence of Theorem 5.2.

Corollary 5.3: Suppose that ϵ, $\kappa > 0$, $\overline{\Delta} \subset \Delta_0$, $(y,p) \in \mathscr{E}\mathscr{S}(\Delta_0)$, and the Q_a are bounded subsets of K_a. Suppose that u is a function on $X \times K$ and the u_a are functions on $X \times K_a$ such that

$$u(x,k) = \sum_{a \in \mathscr{A}} u_a(x,\pi_a k)$$

and

$$\operatorname{supp} u_a \subset \Big(X \setminus \bigcup_{b \not\subset a} X_b^\epsilon\Big) \times Q_a.$$

We assume that u and the u_a's are differentiable for $|x| \neq 0$ and homogeneous of degree zero with respect to x. Finally assume that

$$\{\omega,u\} \geqslant 0, \quad \text{on} \quad \mathscr{E}\mathscr{S}(\Delta_0)^{0,\kappa},$$

and

$$\{\omega,u\}(y,p) > 0. \tag{5.2}$$

Then there exists an open conical set Ω containing (y,p) such that $\Omega \in \mathscr{N}\mathscr{P}_\Delta$.

Theorem 5.2 will follow from Lemma 5.1 and the following propositions.

Proposition 5.4: Suppose that $a \in \mathscr{A}$, u_a satisfies condition (i) of Theorem 5.2 and Δ is a bounded subset of $\mathbb{R}$. Then there exists a bounded operator B such that

$$E_\Delta(H)i[V,u_a(x,D_a)]E_\Delta(H)$$
$$= E_\Delta(H)\langle x\rangle^{-(1+\mu)/2}B\langle x\rangle^{-(1+\mu)/2}E_\Delta(H). \tag{5.3}$$

Proposition 5.5: Suppose that condition (ii) of Theorem 5.2 holds. Let $0 < \mu' < \mu$. then there exists c such that

$$E_\Delta(H)i[H_0,\tfrac{1}{2}(u(x,D) + u(x,D)^*)]E_\Delta(H)$$
$$\geqslant E_\Delta(H)\Big[-c\langle x\rangle^{-(1+\mu')} + w(x,D)^*w(x,D)$$
$$+ \sum_{i=1}^{k} w_i(x,D)^*w'_i(x,D)\Big]E_\Delta(H).$$

Proof of Theorem 5.2 given Propositions 5.4 and 5.5: We set $B = \frac{1}{2}[u(x,D) + u(x,D)^*]$, $C = w(x,D)$, $C_i = w_i(x,D)$, $C'_i = w'_i(x,D)$, for $i = 1,...,k$, $C_{k+1} = \langle x\rangle^{-(1+\mu')/2}$, and $C'_{k+1} = c_1\langle x\rangle^{-(1+\mu')/2}$. Then we apply Lemma 5.1. By Propositions 5.4 and 5.5, we can choose c_1 such that (5.1) is satisfied. Hypothesis A implies that C_{k+1} and C'_{k+1} are H-smooth on Δ. Thus the theorem follows by Lemma 5.1. Q.E.D.

Proposition 5.4 will follow from the following lemma.

Lemma 5.6.

$$\langle x\rangle^\lambda [V,u_a(x,D_a)](H_0 + 1)^{-1}\langle x\rangle^{1+\mu-\lambda}$$

is bounded for any $\lambda \in \mathbb{R}$.

Proof: Set

$$\hat{u}_a(x,z_a) = (2\pi)^{-\dim K_a} \int dk_a u_a(x,k_a)e^{-ik_az_a}.$$

Clearly

$$u_a(x,D_a) = \int \hat{u}_a(x,z_a)e^{iD_az_a}\,dz_a,$$

any, for any N,

$$|\hat{u}_a(x,z_a)| \leqslant c_N\langle z_a\rangle^{-N}.$$

Now

$$[V,\hat{u}_a(x,D_a)] = \sum_{b \not\subset a} [v_b(\pi^b x),u_a(x,D_a)].$$

Moreover,

$$\langle x\rangle^\lambda [u_b(\pi^b x),u_a(x,D_a)](H_0 + 1)^{-1}\langle x\rangle^{1+\mu-\lambda}$$
$$= \langle x\rangle^\lambda \int dz_a \int_0^1 d\tau\, \hat{u}_a(x,z_a)z_a[v_b(\pi^b x + \tau z_a),iD_a]$$
$$\times e^{iz_a\cdot D_a}(H_0 + 1)^{-1}\langle x\rangle^{1+\mu-\lambda}. \tag{5.4}$$

The norm of (5.4) is bounded by

$$\int dz_a \int_0^1 d\tau \| \langle x \rangle^{1+\mu} \langle \pi^b x \rangle^{-1-\mu} \hat{u}_a(x,z_a) z_a \langle \tau z_a \rangle^{1+\mu} \|$$
$$\times \| \langle \pi^b x \rangle^{1+\mu} \langle \pi^b x + \tau z_a \rangle^{-1-\mu} \langle \tau z_a \rangle^{-1-\mu} \|$$
$$\times \| \langle x \rangle^{\lambda} \langle \pi^b x + \tau z_a \rangle^{1+\mu} \nabla v_b (\pi^b x + \tau z_a)$$
$$\times (H_0+1)^{-1} \langle x \rangle^{-\lambda} \| \, \| \langle x \rangle^{\lambda} e^{iz_a D_a} \langle x \rangle^{-\lambda} \|. \qquad (5.5)$$

Since $b \not\subset a$, on supp $\hat{u}_a$ we have $\langle \pi^b x \rangle \geqslant \epsilon \langle x \rangle$. Thus the first norm in (5.5) is bounded by $c_N \langle z_a \rangle^{-N}$, for any N. The second and third norms in (5.5) are bounded uniformly in z_a. Moreover,

$$\| \langle x \rangle^{\lambda} e^{iz_a \cdot D_a} \langle x \rangle^{-\lambda} \| \leqslant c_\lambda \langle z_a \rangle^{|\lambda|}.$$

Thus the integral in (5.4) is absolutely convergent. This ends the proof of the lemma. Q.E.D.

Proof of Proposition 5.4: Take $F \in C_0^\infty(\mathbb{R})$ such that $F=1$ on Δ. Put

$$B = \langle x \rangle^{(1+\mu/2)} i[V, u_a(x,D_a)] F(H) \langle x \rangle^{(1+\mu/2)}.$$

The identity

$$B = \langle x \rangle^{(1+\mu/2)} i[V, u_a(x,D_a)](H_0+1)^{-1}$$
$$\times \langle x \rangle^{(1+\mu/2)} \langle x \rangle^{-(1+\mu/2)} (H_0+1) F(H) \langle x \rangle^{(1+\mu/2)}$$

and an application of Lemmas 5.6 and 4.4 (b) show that B is bounded. Clearly, B satisfies (5.3). Q.E.D.

Proof of Proposition 5.5: Let F, δ, ρ_a, and j_a be as in the proof of Theorem 4.1. Set

$$z(x,k) = \{\omega, u\}(x,k) - |w|^2(x,k) - \sum_{i=1}^{k} \overline{w}_i w_i'(x,k).$$

Clearly $z \in S^{-1}(\omega+1)$. Propositions A1 and A2 imply that

$$i[\omega(D), u(x,D)] - w(x,D)^* w(x,D)$$
$$- \sum_{i=1}^{k} w_i(x,D)^* w_i'(x,D) = z(x,D) + r(x,D),$$

where $r \in S^{-2}(\omega+1)$. By Proposition A3,

$$\langle x \rangle r(x,D)(H_0+1)^{-1} \langle x \rangle \qquad (5.6)$$

is bounded. Lemma 4.5 shows that

$$\langle x \rangle^{[1+\min(\delta\mu, 1-\delta)/2]} F(H)$$
$$\times \left[z(x,D) - \sum_{a \in \mathscr{A}} j_a(x) z(\pi_a x, D) \rho_a(D_a) \right]$$
$$\times (H_0+1)^{-1} \langle x \rangle^{[1+\min(\delta\mu, 1-\delta)/2]} \qquad (5.7)$$

is bounded. Furthermore, since

$$\sum_{a \in \mathscr{A}} j_a(x) z(\pi_a x, k) \rho_a(\pi_a k)$$

is non-negative and belongs to $S_\delta^{-1}(\omega+1)$, we can write, by proposition A4,

$$(H_0+1)^{-1/2} \left[\sum_{a \in \mathscr{A}} j_a(x) z(\pi_a x, D) \rho_a(D_a) + hc \right]$$
$$\times (H_0+1)^{-1/2} \geqslant -c \langle x \rangle^{-1-\delta}. \qquad (5.8)$$

Now our proposition follows from (5.8) and the boundedness of (5.6) and (5.7) by an argument similar to that contained in the proof of Proposition 5.4. Q.E.D.

VI. CLASSICALLY ALLOWED TRAJECTORIES

Suppose that $(y,p) \in \mathscr{E}\mathscr{S}(\Delta_0)$. Does there exist an easy method to determine if a conical neighborhood of (y,p) belongs $\mathscr{N}\mathscr{P}_\Delta$ for $\overline{\Delta} \subset \Delta_0$? This section is devoted to presenting a method that in many cases can serve this purpose.

This method is based on quite suggestive intuition taken from classical physics. First let us give a rather loose description of the intuitive picture hidden here.

Imagine that the quantum evolution e^{iHt} has some common features with the motion of classical particles in the configuration space X. We suppose that they may scatter against the planes X_a, or actually against their neighborhoods X_a^ϵ. More precisely, we assume that the component of the momentum parallel to X_a is conserved while the particles travel through X_a^ϵ but the transversal component may change (even in a discontinuous fashion). Moreover, we suppose that the velocity of particles with momentum k equals approximately $\nabla\omega(k)$ and their motion is confined to $\mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon,k}$. We will show that if all the above described trajectories leave a given conical a-fibered neighborhood of (y,p) within the phase space and do not return to it then a certain (maybe smaller) conical neighborhood of (y,p) belongs to $\mathscr{N}\mathscr{P}_\Delta$.

Now we want to make our intuition precise. Let ϵ, κ, α, β, $T \geqslant 0$. If $\kappa \in K$, we define

$$Vel^{\alpha\beta}(k) = \{w \in X\text{: there exists } k' \in K \text{ such}$$
$$\text{that } |k-k'| \leqslant \alpha$$
$$\text{and } |w - \nabla\omega(k')| \leqslant \beta |\nabla\omega(k')|\}.$$

We say that

$$[0,T] \ni t \to (x,(t),k(t)) \in \mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon,k}$$

is an α,β-classically allowed trajectory (abbreviated as an α, β-CAT) if and only if the following two conditions are true.

(i) $t \to x(t)$ is continuous and differentiable a.e.; moreover if $(dx/dt)(t)$ exists, then $(dx/dt)(t) \in Vel^{\alpha\beta}(k(t))$.

(ii) Let $[t_1,t_2] \subset [0,T]$ and $a \in \mathscr{A}$. Suppose that if $t \in [t_1,t_2]$, then $x(t) \in X_a^\epsilon$. Then $\pi_a k(t)$ is constant for $t \in [t_1,t_2]$.

Now let $\Omega \subset \mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon k}$. We denote by $\Gamma^{\alpha,\beta}(\Omega)$. The set of all $(x,k) \in \mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon k}$ such that there exist $T \geqslant 0$ and an α,β-CAT $[0,T] \ni t \to (x(t),k(t)) \in \mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon,k}$, with $(x(0),k(0)) \in \Omega$ and $(x(T),k(T)) = (x,k)$.

If $k_a \in K_a$, then we define $\widetilde{Vel}_a^{\alpha\beta}(k_a)$ to be equal to the convex cone spanned by $\cup_{\pi_a k' = k_a} Vel^{\alpha\beta}(k')$. It is easy to show the following properties of the operation $\Gamma^{\alpha\beta}$.

Lemma 6.1: Let $\Omega \subset \mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon,\kappa}$. Then we have the following.

(a) $\Gamma^{\alpha\beta}(\Gamma^{\alpha\beta}(\Omega)) = \Gamma^{\alpha\beta}(\Omega)$.

(b) If Ω is conical then $\Gamma^{\alpha\beta}(\Omega)$ is conical.

(c) $\Gamma^{\alpha\beta}(\Omega) \cap X_a^\epsilon \times K$ is a fibered.

(d) Suppose that $\kappa \in K$, $w \in \widetilde{Vel}_a^{\alpha\beta}(\pi_a\kappa)$, and, for $0 \leqslant t \leqslant 1$, we have $x + wt \in X_a^\epsilon$. Then $(x,k) \in \Gamma^{\alpha\beta}(\Omega)$ implies $(x+w,k) \in \Gamma^{\alpha\beta}(\Omega)$.

The main result of this section is the following theorem.

Theorem 6.2: Suppose that α, β, $\kappa > 0$, $\epsilon_0 \geqslant \epsilon > 0$, $a \in \mathscr{A}$, and $\overline{\Delta} \subset \Delta_0$. Let $(y,p) \in Z_a \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$ and let Θ be an open conical a-fibered set such that $(y,p) \in \Theta$

$\subset Z_a^{(\epsilon)} \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$. Assume, moreover, the following two conditions.

(i) There exists $q \in K$ such that $q \cdot y = 0$, $q \cdot \nabla\omega(p) > 0$, and, if $w \in \widetilde{Vel}^{2\alpha\beta}(\pi_a p)$, then $q \cdot w \geqslant 0$.

(ii) If $[0,T] \ni t \to (x(t),k(t))$ is an α,β-CAT such that $(x(0),k(0)) \in \Theta$ and $(x(T),k(T)) \in \Theta$, then $(x(t),k(t)) \in Z_a^{(\epsilon)} \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$, for all $t \in [0,T]$.

Then there exists an open conical set Ω in $X \times K$ such that $(y,p) \in \Omega$ and $\Omega \in \mathscr{N}\mathscr{P}_\Delta$.

Let us say a few words about the meaning of this theorem. Define

$$M = \{(x,k) \in X \times K: x \cdot q = 0\}$$

and

$$M_+ = \{(x,k) \in X \times K: x \cdot q \geqslant 0\}.$$

Clearly condition (i) implies that all the α,β-CAT's starting in a conical a-fibered neighborhood of (y,p) inside M move initially within M_+. Having left $Z_a^{(\epsilon)} \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$ those trajectories may scatter a number of times. Condition (ii) guarantees that they never "come back" into a vicinity of (y,p) after leaving $Z_a^{(\epsilon)} \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$.

The following lemma is based on the above intuition.

Lemma 6.3: Suppose that the hypotheses of Theorem 6.2 are true. Then there exist an open conical a-fibered set $\tilde{\Theta} \subset Z_a^{(\epsilon)} \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$ containing (y,p) and a set $\Psi \subset \mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon,\kappa}$ such that $\Gamma^{\alpha\beta}(\Psi) = \Psi$ and $\tilde{\Theta} \cap M_+$.

Proof: Set

$$\Delta_{\gamma,\rho} = \left\{(x,k) \in M: \left|\frac{x}{|x|} - \frac{y}{|y|}\right| < \rho, |\pi_a \kappa - \pi_a p| < \gamma\right\}$$

and

$$\Theta_{\gamma,\rho,\sigma} = \{(x + t\nabla\omega(p),k): (x,k) \in \Delta_{\gamma,\rho}, |t| < \sigma\}.$$

We set $\tilde{\Theta} = \Theta_{\gamma,\rho,\sigma}$ and $\Psi = \Gamma^{\alpha\beta}(\Delta_{\gamma,\rho})$ for sufficiently small γ, ρ, σ, to be determined later. We may assume that $\gamma \leqslant \alpha$ and $\tilde{\Theta} \subset \Theta$. Clearly $\tilde{\Theta} \cap M_+ \subset \Psi$.

Now suppose that $(x,k) \in \Psi \cap \tilde{\Theta}$. Then there exists an α, β-CAT $[0,T] \ni t \mapsto (x(t),k(t))$ such that $(x(0),k(0)) \in \Delta_{\gamma,\rho}$ and $(x(T), k(T)) = (x,k)$. Condition (ii) implies that $(x(t),k(t)) \in Z_a^{(\epsilon)} \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$, for all $t \in [0,T]$. Consequently, $x(T) - x(0) \in \widetilde{Vel}_a^{\alpha,\beta}(\pi_a k)$. But if α is small enough, then $\widetilde{Vel}_a^{\alpha,\beta}(\pi_a k) \subset \widetilde{Vel}_a^{2\alpha,\beta}(\pi_a p)$. Condition (i) implies now that $(x(T)-x(0)) \cdot q \geqslant 0$. Thus $x(T) \in M_+$. Consequently $\Psi \cap \tilde{\Theta} \subset M_+$, which ends the proof of the lemma. Q.E.D.

The proof of Theorem 6.2 is based on Corollary 5.3. The characteristic function of Ψ would be a good candidate for the function u from this corollary except that it is discontinuous. Thus our strategy is to approximate this characteristic function by an S^0 function.

Proof of Theorem 6.2: Let χ be the characteristic function of Ψ and χ_a be the characteristic functions of $\Psi \cap Z_a^{(\epsilon)} \times K$. Note that $\Gamma^{\alpha,\beta}(\Psi) = \Psi$ implies that the distributional derivative of χ has the following property: if $(x,k) \in \mathrm{Int}\, X_a^\epsilon \times \pi_a^{-1} ES_a(\Delta_0)^\kappa$ and $w \in \widetilde{Vel}_a^{\alpha,\beta}(\pi_a k)$, then

$$w \cdot \nabla_x \chi(x,k) \geqslant 0. \tag{6.1}$$

Now let $\rho, \gamma > 0$. Let S denote the unit sphere in X and ds the invariant measure on S. Choose $f \in C_0^\infty(\mathbb{R})$ such that $f \geqslant 0$, $f > 0$ on a neighborhood of 0 and $\mathrm{supp}\, f \subset [-1,1]$. Define $f_\rho \in C^\infty(S \times S)$ such that

$$f_\rho(s,s') = c_\rho f(|s,s'|^2/\rho^2),$$

where c_ρ is defined by the condition

$$\int_S f_\rho(s,s')ds' = 1.$$

Next choose $g_\gamma \in C_0^\infty(K)$ such that $g_\gamma \geqslant 0$, $\int g_\gamma(k)dk = 1$, and $\mathrm{supp}\, g_\gamma \subset B(0,\gamma)$. We put

$$\chi^{\rho,\gamma}(x,k) = \int_{S \times K} \chi(s',k') f_\rho\left(\frac{x}{|x|}, s'\right) g_\gamma(k - k')ds'\, dk'$$

and

$$\chi_a^{\rho,\gamma}(x,k) = \int_{S \times K} \chi_a(s',k') f_\rho\left(\frac{x}{|x|}, s'\right) g_\gamma(k - k')ds'\, dk'.$$

Lemma 6.4: The $\chi_a^{\rho,\gamma}$ are non-negative bounded functions differentiable for $|x| \neq 0$ and homogeneous of degree zero with respect to x. Moreover,

(a) $\sum_{a \in \mathscr{A}} \chi_a^{\rho,\gamma} = \chi^{\rho,\gamma}$;

(b) $\chi_a^{\rho,\gamma}(x,k)$, if $\pi_a k = \pi_a k'$;

(c) if $\rho \leqslant \epsilon$, then

$$\mathrm{supp}\, \chi_a^{\rho,\gamma} \subset \left(X_a^{\epsilon+\rho} \setminus \bigcup_{b \not\subset a} X_b^{\epsilon-\rho}\right) \times \pi_a^{-1} ES_a(\Delta_0)^{\kappa+\gamma};$$

and (d) if ρ is small enough and $\gamma \leqslant \min(\alpha,\kappa)$, then $\nabla\omega(k) \cdot \nabla_x \chi^{\rho,\gamma}(x,k) \geqslant 0$, on $\mathscr{E}\mathscr{S}(\Delta_0)^{\epsilon-\rho,\kappa-\gamma}$.

Proof: All the statements of the above lemma are straightforward except for (d). To prove (d), we need the following definition. Let $s \in S$, $z \in X$, and $s \cdot z = 0$. Let $\mathbb{R} \ni t \mapsto O_t^{s,z}$ denote the one-parameter group of rotations of X such that $(d/dt)\, O_t^{s,z}(s)_{t=0} = z$ and the subspace orthogonal to s and z is left invariant. Clearly,

$$f_\rho(O_t^{s,z}(s_1), O_t^{s,z}(s_2)) = f_\rho(s_1,s_2);$$

moreover, the measure ds is invariant with respect to $O_t^{s,z}$. Consequently,

$$\chi^{\rho,\gamma}(O_t^{s,z}(s),k) = \int ds'\, dk'\, \chi(O_t^{s,z}(s'),k') f_\rho(s,s') g_\gamma(k - k'). \tag{6.2}$$

Now if $s \in S$ and $k \in K$, define

$$z(s,k) = \nabla\omega(k) - s(s \cdot \nabla\omega(k)).$$

Fix $(x,k) \in X \cdot K$ and set $s = x/|x|$. Then by the homogeneity of χ and by (6.2), we can write

$$\begin{aligned} &\nabla\omega(k) \cdot \nabla_x \chi^{\rho,\gamma}(x,k) \\ &\quad = (1/|x|) z(s,k) \cdot \nabla_x \chi^{\rho,\gamma}(s,k) \\ &\quad = \frac{1}{|x|} \frac{d}{dt} \chi^{\rho,\gamma}(O_t^{s,z(s,k)}(s),k)|_{t=0} \\ &\quad = \frac{1}{|x|} \int ds'\, dk' \left[\frac{d}{dt} O_t^{s,z(s,k)}(s')|_{t=0} + s'(s' \cdot \nabla\omega(k))\right] \\ &\qquad \times \nabla_x \chi(s',k') f_\rho(s,s') g_\gamma(k - k'). \end{aligned} \tag{6.3}$$

We can choose $\rho > 0$ such that if $s_1, s_2 \in S$, $z \in X$, and $|s_1 - s_2| < \rho$, then

$$\left| z - \frac{d}{dt} O_t^{s_1, z}(s_2) \big|_{t=0} \right| \leqslant (\beta - 2\rho) |z|. \tag{6.4}$$

From now on, let us assume that $(x,k) \in \mathscr{E}\mathscr{S}\,(\Delta_0)^{\epsilon - \rho, \kappa - \gamma}$. Let $(s', k') \in S \times K$ be such that $f_\rho(s,s') g_\gamma(k - k') \neq 0$ (recall that $s = x/|x|$). Then

$$|s - s'| < \rho \tag{6.5}$$

and

$$|k - k'| < \gamma \leqslant \min(\alpha, \kappa). \tag{6.6}$$

Consequently

$$(s', k') \in \bigcup_{a \in \mathscr{A}} \operatorname{Int} \chi_a^\epsilon \times \pi_a^{-1} ES_a(\Delta_0)^\kappa.$$

Formulas (6.4) and (6.5) imply the following inequality:

$$\begin{aligned} &\left| \nabla\omega(\kappa) - \frac{d}{dt} O_t^{s, z(s,k)}(s') \Big|_{t=0} - s'(s' \cdot \nabla\omega(\kappa)) \right| \\ &\quad \leqslant |s(s \cdot \nabla\omega(k)) - s'(s' \cdot \nabla\omega(k))| \\ &\qquad + (\beta - 2\rho) |\nabla\omega(k) - s(s \cdot \nabla\omega(k))| \\ &\quad \leqslant \beta |\nabla\omega(k)|. \end{aligned} \tag{6.7}$$

Thus the expression in the square brackets in (6.3) belongs to $Vel^{\alpha\beta}(k')$ on the support of $f(s,s')g(k - k')$. Consequently, by (6.1), expression (6.3) is non-negative. This ends the proof of our lemma. Q.E.D.

Now we continue with the proof of Theorem 6.2. We set $u_a = \chi_a^{\rho,\gamma}$ and $u = \chi^{\rho,\gamma}$. Clearly, $\{\omega, u\} \geqslant 0$ on $\mathscr{E}\mathscr{S}(\Delta_0)^{\kappa - \beta, \epsilon - \rho}$. In order to finish the proof of the theorem it remains to check (5.2).

Let Ω_0 be an open conical a-fibered set containing (y,p) and ρ_1, $\gamma_1 > 0$ be such that $\Omega_0^{\rho_1, \gamma_1} \subset \widetilde{\Theta}$. We can assume additionally that the numbers ρ, γ from the previous lemma satisfy $0 < \rho < \rho_1$ and $0 < \gamma < \gamma_1$. Let χ_{M_+} be the characteristic function of M_+. Then, by Lemma 6.3, we have $\chi = \chi_{M_+}$ on $\Omega_0^{\pi_1, \gamma_1}$. Thus if $(x,k) \in \Omega_0$, then

$$\chi^{\rho,\gamma}(x,k) = \int_S \chi_{M_+}(s) f_\rho\left(\frac{x}{|x|}, s\right) ds.$$

Clearly

$$\nabla\omega(p) \cdot \nabla\chi_{M_+} = \nabla\omega(p) \cdot q \delta_M, \tag{6.8}$$

where δ_M is a translation invariant measure concentrated on M. Next we mimic the arguments contained in the proof of Lemma 6.4 using additionally (6.8) and the positivity of $\nabla\omega(p) \cdot q$. For sufficiently small ρ, we obtain

$$\nabla\omega(p) \cdot \nabla_x \chi^{\rho,\gamma}(y,p) > 0.$$

Now our theorem follows from Corollary 5.3. Q.E.D.

Let us say a few words on the advantages and disadvantages of Theorem 6.2. The theorem can be nontrivial only if the cone $\widetilde{Vel}_a^{\alpha\beta}(\pi_a p)$ is not equal to X. Only then will there exist a vector q such that condition (i) can be satisfied. This always happens if $\nabla\omega(p) \neq 0$, $a = a_{\min}$, and α, β are small enough. Of course, in this case it remains to verify condition (ii). Thus the theorem seems quite interesting and nontrivial in the case $y \in Z_{a_{\min}}$.

Unfortunately, in the case $a \neq a_{\min}$ it often happens that $\widetilde{Vel}_a^{\alpha,\beta}(k_a) = X$, for any $\alpha, \beta > 0$. This takes place, in particular, if $\omega(k) = k^2$. Thus if $\omega(k) = k^2$ and $a \neq a_{\min}$, then Theorem 6.2 is worthless. Note, however, that in the case $\widetilde{Vel}_a^{\alpha,0}(k_a)$ is not equal to X for $k_a \neq 0$ and small enough α (it is equal to

$$\{x \in X: \pi_a x = \lambda x_a', \lambda > 0, |x_a' - 2k_a| < \alpha\}).$$

In fact, we think that it may be possible to prove a modified version of Theorem 6.2 with $\beta = 0$ (and maybe $\epsilon = 0$), which would be more interesting.

Still we think that the general intuition Theorem 6.2 is based on is the right one. Moreover, there are a lot of examples where this theorem is nontrivial if $y \not\subset Z_{a_{\min}}$. For instance, if $\omega(k) = (k^2 + 1)^m$, with $m > 1$, then $\widetilde{Vel}_a^{\alpha,\beta}(k_a)$ is not equal to X, for small enough α, β and nonzero k_a. Thus, in this case, condition (i) of our theorem can be easily satisfied—one needs to check condition (ii), which involves a more detailed investigation of the geometry of the phase space.

ACKNOWLEDGMENT

I am very grateful to I. Sigal for helpful conversations and for bringing my attention to the subject of dispersive N-body Schrödinger operators.

APPENDIX: PSEUDODIFFERENTIAL OPERATORS

In this appendix we describe the properties of the pseudodifferential operators with symbols in $S_\delta^m(\eta)$ (see the definition below).

Most of the time, the pseudodifferential operators that we use in this paper have the symbols in S_δ^m. Their properties are well known.[31,32] Occasionally, however, we need to use slightly more general classes, namely $S_\delta^m(\omega + 1)$. The calculus of the pseudodifferential operators with symbols in very general classes that include $S_\delta^m(\eta)$ is studied in Ref. 31. The results that we present below are essentially contained in Ref. 31. and therefore we omit their proofs.

We say that a strictly positive function of K is slowly varying if and only if there exist c and N such that $\eta(k) \leqslant c(k - k')^N \eta(k')$, for every $k, k' \in K$. Clearly if η is slowly varying, then so is η^m, for any $m \in \mathbb{R}$. Moreover, $\omega + 1$ is slowly varying on K. Now if η is slowly varying on K, then we define

$$\begin{aligned} S_\delta^m(\eta) = \{ &u \in C^\infty(X \times K) : |\partial_x^\alpha \partial_k^\beta u| \\ &\leqslant c_{\alpha\beta} \langle x \rangle^{m - |\alpha|\delta} \eta(k) \}. \end{aligned}$$

If $u \in S_\delta^m(\eta)$ and $\varphi \in \mathscr{S}$ (the space of Schwartz test functions), then we define $u(x,D)\varphi$ by the formula

$$u(x,D)\varphi(x) = \frac{1}{(2\pi)^{\dim X}} \int u(x,k) e^{ixk} \hat{\varphi}(k) dk.$$

It is easy to show that $u(x,D)$ is a bounded operator on $\mathscr{S}$. The next two propositions describe the properties of the superposition and of the adjoints of pseudodifferential operators with symbols in $S_\delta^m(\eta)$.

Proposition A1: Let η_1 and η_2 be slowly varying. Let $u \in S_\delta^{m}(\eta_1)$, $v \in S_\delta^{m_2}(\eta_2)$, and $l = 0,1,\ldots$. Define

$$w_j(x,k) = (1/j!)(iD_k \cdot D_y)^j u(x,k) v(y,p)_{\substack{x=y \\ k=p}} .$$

Then $w_j \in S_\delta^{m_1+m_2-j\delta}(\eta_1 \cdot \eta_2)$ and there exists $r_l \in S_\delta^{m_1+m_2-l\delta}$ $\times (\eta_1 \cdot \eta_2)$ such that

$$u(x,D)v(x,D) = w_0(x,D) + \cdots + w_{l-1}(x,D) + r_l(x,D).$$

In all the propositions below η is a slowly varying function.

Proposition A2: Let $u \in S_\delta^m(\eta)$ and $l = 0.1,\ldots$. Define $w_j(x,k) = (1/j!)(iD_x D_k)^j \bar{u}(x,k)$. Then $w_j \in S_\delta^{m-j\delta}(\eta)$ and there exists $r_l \in S_\delta^{m-l\delta}(\eta)$ such that

$$u(x,D)^* = w_0(x,D) + \cdots + w_{l-1}(x,D) + r_l(x,D).$$

The following proposition is a consequence of the Calderon–Vaillancourt theorem.[31–33]

Proposition A3: Let $u \in S_0^0(\eta)$. Then $u(x,D)\eta(D)^{-1}$ and $\eta(D)^{-1}u(x,D)$ extend to bounded operators on $L^2(X)$.

Finally we state an easy consequence of the sharp Gårding inequality.[31,32,34]

Proposition A4: Let $u \in S_\delta^m(\eta)$ and Re $u \geqslant 0$. Then there exists c such that

$$\eta(D)^{-1/2}(u(x,D) + u(x,D)^*)\eta(D)^{-1/2} \geqslant -c\langle x\rangle^{m-\delta}.$$

[1] T. Kato, "Wave operators and similarity for some non-selfadjoint operators," Math. Ann. **162**, 258 (1966).

[2] T. Kato, "Smooth operators and commutators," Studia Math. **31**, 535 (1968).

[3] C. R. Putnam, *Commutational Properties of Hilbert Space Operators and Related Topics* (Springer, Berlin, 1967).

[4] R. Lavine, "Absolute continuity of Hamiltonian operators with repulsive potentials," Proc. Am. Math. Soc. **22**, 55 (1969).

[5] R. Lavine, "Commutators and scattering theory, I. Repulsive interactions," Commun. Math. Phys. **20**, 301 (1971).

[6] R. Lavine, "Commutators and scattering theory, II. A class of one-body problems," Indiana Univ. Math. J. **21**, 643 (1972).

[7] R. Lavine, "Completeness of the wave operators in the repulsive N-body problem," J. Math. Phys. **14**, 376 (1973).

[8] M. Arai, "Absolute continuity of Hamiltonian operators with repulsive potentials," Publ. Res. Inst. Math. Sci. **7**, 621 (1971/72).

[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics IV* (Academic, New York, 1978).

[10] R. Iorio and M. O'Carroll, "Asymptotic completeness for multiparticle Schrödinger Hamiltonians with weak potentials," Commun. Math. Phys. **27**, 137 (1972).

[11] I. Sigal and A. Soffer, "N-particle scattering problem: Asymptotic completeness for short range systems," Ann. Math. **125**, 35 (1987).

[12] E. Mourre, "Absence of singular continuous spectrum for certain selfadjoint operators," Commun. Math. Phys. **78**, 391 (1981).

[13] E. Mourre, "Operateurs conjugués et proprietes de propagations," Commun. Math. Phys. **91**, 279 (1983).

[14] P. Perry, I. M. Sigal, and B. Simon, "Spectral analysis of N-body Schrödinger operators," Ann. Math. **114**, 519 (1981).

[15] H. L. Cycon, R. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators with Application to Quantum Mechanics and Global Geometry* (Springer, Berlin, 1987).

[16] D. R. Yafaev, "Remarks on the spectral theory for Schrödinger operators of multiparticle type," Zap. Nauch. Sem. LOMI **133**, 277 (1984).

[17] I. M. Sigal and A. Soffer (private communication).

[18] J. Dereziński, "A new proof of the propagation theorem for N-body quantum systems," Commun. Math. Phys. **122**, 203 (1989).

[19] M. Reed and B. Simon, *Method of Modern Mathematical Physics III* (Academic, New York, 1979).

[20] I. M. Sigal, *Scattering Theory for Many Body Quantum Mechanical Systems, Lecture Notes in Mathematics*, Vol. 1 1011, Springer, Berlin, (1983).

[21] G. A. Hagedorn, "Asymptotic completeness for two, three, and four particle Schrödinger operators," Trans. Am. Math. Soc. **258**, 1 (1980).

[22] S. Agmon, *Lectures on the Exponential Decay of Solutions of Second Order Elliptic Equations* (Princeton U.P., Princeton, NJ, 1982).

[23] R. Froese and I. Herbst, "A new proof of the Mourre estimate," Duke Math. J. **49**, 1075 (1982).

[24] R. Froese and I. Herbst, "Exponential bounds and absence of positive eigenvalues for N-body Schrödinger operators," Commun. Math. Phys. **87**, 429 (1982).

[25] J. Dereziński, "The Mourre estimate for dispersive N-body Schrödinger operators," to appear in Trans. Am. Math. Soc.

[26] I. M. Sigal (private communication).

[27] V. Enss, "A note of Hunziker's theorem," Commun. Math. Phys. **52**, 233 (1977).

[28] P. Deift and B. Simon, "A time dependent approach to the completeness of multiparticle quantum scattering," Commun. Pure Appl. Math. **30**, 573 (1977).

[29] B. Simon, "Geometric methods in multiparticle quantum systems," Commun. Math. Phys. **55**, 259 (1977).

[30] I. M. Sigal, "Geometric methods in quantum many body problem. Nonexistence of very negative ions," Commun. Math. Phys. **85**, 309 (1982).

[31] L. Hörmander, *The Analysis of Linear Partial Differential Operators III* (Springer, Berlin, 1985).

[32] M. Taylor, *Pseudodifferential Operators* (Princeton U.P., Princeton, NJ, 1981).

[33] A. P. Calderon and R. Vaillancourt, "On the boundedness pseudodifferential operators," J. Math. Soc. Jpn. **23**, 374 (1972).

[34] L. Hörmander, "Pseudo-differential operators and non-elliptic boundary problems," Ann. Math. **83**, 129 (1966).

The probability operator in quantum theory and quantization of a system of free fields

Sujit Basu[a] and V. V. Kuryshkin
Department of Theoretical Physics, Peoples' Friendship University, Moscow V-302, USSR

(Received 22 February 1989; accepted for publication 22 November 1989)

A theoretical framework for quantization, defined by the normalized positive-definite probability operator establishing dynamical correspondence between classical and quantum Poisson brackets, is presented. The resulting quantum theory, unlike the conventional one, admits consistent probabilistic interpretation. It is shown that, in the nonrelativistic case, quantization based on the probability operator leads to the theory known as "quantum mechanics with a non-negative quantum distribution function." A generalization of the proposed framework to the case of the relativistic theory of fields is attempted. Four auxiliary problems of constructing probability operators of one-dimensional field oscillators in Bose and Fermi algebras are formulated and solved. On the basis of these solutions it is concluded that spinor fields are not quantizable in the Bose algebra with the help of the probability operator (the analog of Pauli's theorem). An equation for the probability operator of a system of free fields is derived from the principles of dynamical correspondence and translational invariance. The physical meaning of the operators corresponding to classical field amplitudes, such as annihilation and creation operators of field quanta with definite energy-momentum, is shown to emerge as a consequence of this equation. It is shown that the quantization of a system consisting only of tensor fields or only of spinor fields in the formalism of the probability operator leads to difficulties. It is shown further that these difficulties can be removed by considering quantization of a system containing both tensor and spinor fields. As an illustration, quantization of a system consisting of a massive vector field and a massive spinor field is considered and it is found that a noncontradictory quantization requires the mass of the vector particle to be less than that of the spinor particle. The probability operator thus acts as a mechanism of selection of the fields to be quantized already at the level of free fields.

I. INTRODUCTION

It is quite well known that quantization of a physical system viewed as a transition from the known classical description to the corresponding quantum one requires the solution of the following four basic problems: (i) determination of a linear space $\mathscr{L}$ of the vector states $|\psi\rangle$ of the system, (ii) choice of an algebra $\mathscr{A}$ of operators, linear in $\mathscr{L}$, (iii) determination of the operators (belonging to $\mathscr{A}$) of physical quantities characterizing the system, and (iv) postulation of the evolution operator of the system.

The first two of these problems are solved by defining a set of generators satisfying specific commutation relations. Thus, in a nonrelativistic quantum description of a system with N degrees of freedom, one takes as generators N pairs of Hermitian operators satisfying standard Heisenberg commutation relations. In the case of relativistic fields modern quantum theory postulates the existence of only two kinds of algebras: Bose–Einstein and Fermi–Dirac. The selection of a specific one from these two is decided by Pauli's spin-statistics theorem connecting the transformation properties of a field with commutation relations of the generators.

The third problem, i.e., the problem of assigning a linear operator $\hat{A}$ to each physical quantity A [represented by a phase-space function $A(q,p,t)$ in the nonrelativistic classical theory of a finite-dimensional system], generally known as the problem of the correspondence rule, is yet to achieve a complete and undisputed final solution, as noted by many authors,[1–6] notwithstanding the fact that a wide variety of correspondence rules have been proposed[7–15] since the birth of quantum mechanics.

Specifically, Neumann's rule,[7] lying at the root of conventional quantum mechanics, is nonunique and any attempt to eliminate nonuniqueness leads to inner contradiction (all quantum operators commute with one another[1]). The possibility of a noncontradictory and unique formulation at the expense of weakening one of Neumann's requirements was studied in Ref. 4, resulting in a wide class of unique rules, the so-called non-Neumann rules. However, there does not seem to be any guiding physical principle for selecting a particular rule from this wide class.

Dirac's correspondence[8] between classical and quantum Poisson brackets has also been studied fairly thoroughly[2,3,5,6,16,17] and the nonuniqueness of this rule has been shown explicitly. Here, again, attempts to liquidate nonuniqueness inevitably lead to contradictions.[2,3]

The majority of the unique rules proposed (e.g., Weyl,[9] Born–Jordan,[10] Rivier,[11] standard,[12] normal,[12] generalized correspondence rule of Cohen[14]) do not guarantee positive-definiteness of the quantum average values of non-negative physical quantities such as dispersion,[2,3,15] thus presenting

[a] Present address: Meteorology and Oceanography Division, Space Applications Centre, Ahmedabad 380 053, India.

0022-2488/90/040851-12$03.00

serious interpretational difficulty. The antinormal rule of Kano,[13] free from this drawback, belongs[18] to the class of unique rules proposed by one of the present authors.[15] However, quantum theory based on such correspondence differs significantly[19–23] from conventional quantum mechanics, although it contains several interesting results.

One can also approach the problem of the correspondence rule from a seemingly different point of view, that of quantum distribution functions (QDF's). The essence of the problem of QDF's lies in an attempt to assign a function $F_\Psi(q,p,t)$ to each normalized state $|\psi\rangle$ with the condition

$$\int F_\Psi(q,p,t)dq\,dp = 1, \tag{1.1}$$

and the distributive property

$$\langle A\rangle_\psi = \langle\psi|\hat{A}\,|\psi\rangle = \int A(q,p,t)F_\psi(q,p,t)dq\,dp \tag{1.2}$$

simultaneously for all physical quantities A. If, in addition, F_ψ is real and non-negative, then it can be interpreted as a joint probability density of coordinates and momenta. However, attempts to introduce such a non-negative QDF in quantum mechanics met with failures. Thus the QDF's proposed and studied by different authors[24–29] turned out to be either complex or real, but sign variable. Non-negative functions constructed by Bopp[30] and Kano[31] also cannot be interpreted strictly as QDF's since they do not yield "correct" (from the standpoint of conventional quantum mechanics) marginal distributions. Later investigations showed[12,14,15] that the possibility of introducing a QDF is uniquely connected with the correspondence rule used for constructing quantum operators from their classical counterparts. In particular, no QDF (even a sign-variable or a complex one) exists[32] in conventional quantum mechanics based on Neumann's rule.

Further investigations[15,21] showed that it is possible to alter quantum mechanics so as to introduce non-negative QDF's. Such an alteration was achieved by means of the correspondence rule formulated in Ref. 15. The resulting theory, named "quantum mechanics with a non-negative QDF"[19–23] is closed and self-consistent. Besides, it contains the essential parts of conventional quantum mechanics and classical statistical theory as particular limiting cases.[19] As the name implies, in this theory there exists a non-negative QDF for each state and as such the theory has a built-in statistical interpretation. Here we want to note that the non-negative QDF's recently considered in the literature[33–36] are just special cases of the class of non-negative QDF's contained in this theory. In fact, from the point of view of the correspondence rules, this class is the only admissible one, as has been shown quite recently by one of the current authors.[37]

Before coming to the purpose of the present paper, we present briefly some concrete results of this theory. Energy levels of a harmonic oscillator,[21] and a hydrogenlike atom[20,22] have been calculated in its framework. In both cases they are shifted in comparison with the results of conventional quantum mechanics. In the case of an oscillator all levels are shifted equally and thus such a shift is nonobservable experimentally. The shift of the energy levels of the hydrogenlike atom is analogous to the Lamb shift. The magnetic moment of a hydrogenlike atom in the state $|nlm\rangle$, contrary to the results of conventional quantum mechanics, depends[38] on all the quantum numbers n,l,m. If the shift of the S levels is identified with the experimentally observed value of the Lamb shift then the magnetic moment is increased by 10^{-6} (in the Bohr magneton), constituting a thousandth part of the anomalous magnetic moment of the electron. Such a result may, in principle, be checked experimentally by measuring magnetic moments of hydrogenlike atoms.

However, the above-mentioned differences of the results from those of conventional quantum mechanics, strictly speaking, are at best of a qualitative character, since they were obtained in a nonrelativistic framework. Strict quantitative comparison apparently requires proper relativistic generalization of the theory, inclusion of spin, etc. It seems to us that a proper generalization is possible only in the case of relativistic field theory, where the classical theory already contains such significantly relativistic concepts as spin (see, however, the recent interesting approach of such a generalization to the case of relativistic quantum mechanics in Ref. 39).

Thus the main purpose of our paper is a mathematically consistent generalization of the formalism of "quantum mechanics with a non-negative QDF" to the case of relativistic fields. The principal mathematical tool of our investigation will be the probability operator.[40] It is well known that a linear unique correspondence rule can be conveniently formulated with the help of a universal basis operator (see, for details, Refs. 23 and 40–44), parametrically depending on coordinates, momenta, and time and simultaneously defining the form and properties of a QDF. This operator has been given various names [the mapping operator,[42] the representation operator,[44] the (quasi)probability operator[40]]. We adopt the last terminology and drop the prefix "quasi," since in our case this operator is positive definite (for details, see Ref. 40). Not only does such an operator explicitly demonstrate the connection of the correspondence rule to the existence of QDF's, but it also provides us with a method of relating the correspondence rule to the fourth aspect of quantization—the temporal evolution of the quantized system. Mutual connection of the third and fourth aspects of quantization along with suitable assumptions leads to an equation[45] for the probability operator, the solution of which (with proper normalization) completely determines the quantization procedure, provided the first two problems (determination of state space and algebra) are solved beforehand. Thus the formalism of a probability operator deals with all the aspects of quantization, dealt with separately earlier, in a unified manner.

The structure of our paper is as follows. Sections II and III are basically of an introductory nature. In Sec. II we introduce formally the concept of the probability operator, and show its role as a universal basis operator in the quantization of finite-dimensional systems. We also present the derivation of the equation obeyed by such an operator following from the principle of dynamical correspondence.[45] In

Sec. III we investigate the problem of the probability operator in nonrelativistic quantum mechanics and show how the property of positive-definiteness and a reasonable demand on the correspondence rule leads to the previously mentioned "quantum mechanics with a non-negative QDF." The remaining sections (IV–VII) are devoted to the generalization to the case of relativistic field theory. Bearing in mind the well-known fact that free classical fields (and we consider only such cases) can be represented as a system of noninteracting one-dimensional oscillators, in Sec. IV we devote our attention to the quantization of such oscillators, which are of two types: those with a positive contribution to the field energy and those with a negative contribution (arising in the case of classical spinor fields). The postulation of the existence of only two types of algebra (Bose–Einstein and Fermi–Dirac) leads us to the consideration of four typical problems concerning one-dimensional field oscillators (two types of oscillators × two types of algebra). Sections V and VI contain the general aspects of quantization of fields on the basis of the probability operator and the derivation of the equation for such an operator from general principles of dynamical correspondence and translational invariance. In Sec. VII we study the quantization of free fields and the consequences of such a quantization. Specifically it is shown that the difficulties of infinite vacuum energy and infinite vacuum charge can be liquidated by quantizing a system comprising both tensor and spinor fields, although such a difficulty persists in the quantization of isolated tensor and spinor fields. At the end of this section we summarize the main conclusions.

II. NONSTANDARD QUANTIZATION BASED ON THE PROBABILITY OPERATOR

The transition from classical theory to the corresponding quantum theory assumes, as one of its necessary procedures, the mapping of a set of physical quantities $\{A\}$, represented in the classical theory by the functions $\{A(q,p,t)\}$ of generalized coordinates $q=(q_1,\ldots,q_N)$, momenta $p=(p_1,\ldots,p_N)$, and time t onto the set of quantum operators $\{\hat{A}\}$ representing the same quantities in quantum theory. Also, in accordance with the basic postulates of quantum theory, the set $\{\hat{A}\}$ belongs to some algebra $\mathscr{A}$ of linear operators acting in a linear complex space $\mathscr{L}$ of state vectors $|\psi\rangle$.

From among the numerous proposed mappings $\hat{A}=O(A(q,p,t))$, known also as correspondence rules, we adopt a linear mapping[23,40] satisfactory from the interpretational point of view and formulated as

$$\hat{A}=O(A(q,p,t))=\int A(q,p,t)\hat{F}(q,p,t)dq\,dp, \qquad (2.1)$$

where the operator $\hat{F}(q,p,t)\in\mathscr{A}$, termed as the probability operator, satisfies the conditions of normalization and positive-definiteness:

$$\int \hat{F}(q,p,t)dq\,dp=\hat{1}, \qquad (2.2a)$$

$$F_\psi(q,p,t)=\langle\psi|\hat{F}(q,p,t)|\psi\rangle\geqslant 0, \quad \forall|\psi\rangle\in\mathscr{L}. \qquad (2.2b)$$

In the quantization defined by (2.1) the quantum average values $\{\langle A\rangle\}$ of the whole set of physical quantities in any normalized state $|\psi\rangle$ evidently can be written by means of formula (1.2) of Sec. I and the function $F_\psi(q,p,t)$, being the quantum average (2.2b) of the operator $\hat{F}$ in the state $|\psi\rangle$, by virtue of its normalization (1.1) [following from (2.2a)] and positive-definiteness, can be considered as the joint probability density of coordinates and momenta of the physical system in the state $|\psi\rangle$.

For establishing the concrete form of the probability operator it is only natural to extend the correspondence rule (2.1) so as to include the evolutionary aspect of the quantization procedure.[45] Such an extension is based on the following reasoning.

In classical theory the time evolutions of coordinates $\langle q\rangle_{\rm cl}=q(t)$ and $\langle p\rangle_{\rm cl}=p(t)$ are determined by the following Hamiltonian equations:

$$d_tq(t)=\partial_pH(q(t),p(t),t), \quad d_tp(t)=-\partial_qH(q(t),p(t),t), \qquad (2.3a)$$

where $H(q,p,t)$ is the classical Hamiltonian of the system. Hence the physical quantities and their time derivatives are determined as

$$\begin{aligned}&\langle A\rangle_{\rm cl}=A(q,p,t)|_{q(t),p(t)},\\ &d_t\langle A\rangle_{\rm cl}=(\partial_tA+\{H,A\})(q,p,t)|_{q(t),p(t)},\end{aligned} \qquad (2.3b)$$

where { , } denotes the classical Poisson bracket.

In quantum theory with the correspondence rule (2.1), the time evolution of the state $|\psi\rangle$ is determined by the Schrödinger equation

$$i\hbar\,\partial_t|\psi(t)\rangle=\hat{H}|\psi(t)\rangle, \quad \hat{H}=\int H(q,p,t)\hat{F}(q,p,t)dq\,dp. \qquad (2.4a)$$

Thus the average value $\langle A\rangle$ and its time derivative are given by

$$\langle A\rangle_Q=\langle\psi|\hat{A}|\psi\rangle, \quad d_t\langle A\rangle_Q=\langle\psi|\partial_t\hat{A}+(i/\hbar)[\hat{H},\hat{A}]|\psi\rangle, \qquad (2.4b)$$

[,] being the commutator.

Suppose that the correspondence (2.1) can be extended to the evolutionary part of the relations (2.3b) and (2.4b). We can write

$$\int(\partial_tA+\{H,A\})\hat{F}\,dq\,dp=\partial_t\hat{A}+\frac{i}{\hbar}[\hat{H},\hat{A}]. \qquad (2.5)$$

Thus the quantization scheme based on the probability operator is reduced to a solution of the set of integrodifferential equations (2.5) (the equation for each $A\in\{A\}$), with the conditions (2.2) and the subsequent determination of the quantum operators via the recipe (2.1).

While searching for the probability operator, we can, in principle, demand the fulfillment of (2.5) not only for the quantities $A\in\{A\}$, but for all possible phase-space functions. Then, because of the arbitrariness of the function $A(q,p,t)$, from (2.5) we obtain the equation for the probability operator,

$$\begin{aligned}&\partial_t\hat{F}(q,p,t)+\{H(q,p,t),\hat{F}(q,p,t)\}\\ &\quad=\frac{i}{\hbar}\int H(q',p',t)[\hat{F}(q,p,t),\hat{F}(q',p',t)]dq'\,dp'.\end{aligned} \qquad (2.6)$$

The solution of (2.6) must satisfy conditions (2.2).

The concrete solution of (2.6) [or, of the set of equations (2.5), if (2.6) does not possess a solution with the required properties] depends not only on the predetermined algebra $\mathscr{A}$ but also on the classical Hamiltonian $H(q,p,t)$. Besides, in the theory obtained subsequently, to any state $|\psi\rangle$ there corresponds a distribution function $F_\psi(q,p,t) \geqslant 0$. Both these observations show that the proposed quantization is not a conventional or standard one.

III. THE PROBABILITY OPERATOR IN NONRELATIVISTIC QUANTUM MECHANICS

To obtain an explicit form of the operator $\hat{F}$ in terms of the generators of the quantum algebra, we start with the following basic assumptions, also valid in the conventional version of nonrelativistic quantum mechanics.

(i) Generators of the algebra to which $\hat{F}$ (and subsequently any quantum operator $\hat{A}$) belongs are N (denoting the number of degrees of freedom) pairs of Hermitian operators $\hat{q}_j = O(q_j)$, $\hat{p}_j = O(p_j)$, with the commutation relations

$$[\hat{q}_j,\hat{q}_k] = [\hat{p}_j,\hat{p}_k] = 0, \quad [\hat{q}_j,\hat{p}_k] = i\hbar\delta_{jk}\hat{1}, \tag{3.1a}$$

where $j,k = 1,\ldots,N$ and $\hat{1}$ denotes the identity operator of the algebra. The algebra defined by (3.1a) is the standard Heisenberg one and is isomorphic to the Bose algebra.

(ii) Between the pairs of variables (q_j,p_j) and the operators $(\hat{q}_j,\hat{p}_j)$ there is a one-to-one correspondence

$$q_j \rightleftarrows \hat{q}_j, \quad p_j \rightleftarrows \hat{p}_j, \quad j = 1,\ldots,N, \tag{3.1b}$$

understood in the sense that $\hat{A}$ explicitly involves the operator $\hat{q}_j$ (operator $\hat{p}_j$), when and only when the corresponding classical function $A(q,p,t)$ explicitly depends on the coordinate q_j (momentum p_j).

By virtue of the commutation relations (3.1a), one can always express any operator in some particular ordered form. We take the probability operator in such an ordered form, defined by the following operator Fourier integral[42]:

$$\hat{F}(q,p,t) = \int f(q,p,\xi,\eta,t)\exp\left\{\frac{i}{\hbar}(\eta\hat{q} + \xi\hat{p})\right\} d\xi\, d\eta, \tag{3.2}$$

where $\eta\hat{q}$ and $\xi\hat{p}$ are the usual scalar products of the N-vectors.

From (3.2) it is evident that the requirements (3.1b), in the sense explained above, can be satisfied only if the function $f(q,p,\xi,\eta,t)$ is of the form

$$f(q,p,\xi,\eta,t) = f_1(\xi,\eta,t)\exp\left\{-\frac{i}{\hbar}\sum_{j=1}^{N}(\eta_j q_j + \xi_j p_j)\right\}.$$

Thus we have the following general expression for the probability operator:

$$\hat{F}(q,p,t) = (2\pi\hbar)^{-2N}\int u(\xi,\eta,t) \times\exp\left\{\frac{i}{\hbar}[\eta(\hat{q}-q) + \xi(\hat{p}-p)]\right\} d\xi\, d\eta. \tag{3.3}$$

The factor $(2\pi\hbar)^{-2N}$ has been introduced for the sake of future convenience. Normalization, Hermiticity, and positive-definiteness of $\hat{F}$ put the following restrictions on the kernel:

$$u(0,0,t) = 1, \quad u^*(-\xi,-\eta,t) = u(\xi,\eta,t), \tag{3.4a}$$

$$\int u(\xi,\eta,t)\langle\psi|\exp\left\{\frac{i}{\hbar}[\eta(\hat{q}-q) + \xi(\hat{p}-p)]\right\}|\psi\rangle d\xi\, d\eta \geqslant 0, \quad \forall|\psi\rangle. \tag{3.4b}$$

Thus the assumptions (3.1) determine the probability operator in the form (3.3) with the kernel $u(\xi,\eta,t)$ having the properties (3.4). An interesting thing to be noted in this connection is that the form (3.3) of the probability operator automatically satisfies the reasonable requirements of invariance of phase-space description under a Galilean transformation (for details, see the paper by Ruggeri[43]).

In order to restrict further the choice of $\hat{F}$ in the correspondence (2.1) we may impose other specific demands. One of such demands may be, for example, Dirac's correspondence between classical and quantum Poisson brackets:

$$O(\{A(q,p,t),B(q,p,t)\}) = (i/\hbar)[O(A),O(B)]. \tag{3.5}$$

Assuming global validity of Dirac's principle (A and B arbitrary) it is easy to obtain

$$\{\hat{F}(q,p,t),\delta(q-q')\delta(p-p')\} = (i/\hbar)[\hat{F}(q,p,t),\hat{F}(q',p',t)], \tag{3.6}$$

from which we observe that the operators $\hat{F}$ at two different phase-space points commute. Thus the global principle of Dirac reduces the quantum theory based on rule (2.1) to a classical-like theory with commuting observables.

Since the correspondence rule (2.1) is simply a consequence of the linearity of the mapping $A \to \hat{A}$, we conclude that a global Dirac principle is, in fact, incompatible with any quantum theory with a linear mapping. However, we note from Eq. (2.5) of Sec. II that, for time-independent A and $\hat{F}$, it reduces to the form (3.5) with the Hamiltonian H in place of B. Thus (2.5) may be considered as a weakened version of Dirac's principle and may, in principle, be satisfied for suitable choices of $\hat{F}$ and the set $\{A\}$ of physical quantities.

As regards Eq. (2.6), the question of solving it with a predetermined algebra can be investigated only for specific physical systems, since it explicitly involves the classical Hamiltonian. Substitution of the operator $\hat{F}$ in the form (3.3), in (2.6) evidently leads to an equation for the function $u(\xi,\eta,t)$, different, in general, for different systems. The solution of such an equation must also satisfy conditions (3.4), irrespective of the system considered. Hence, without going into the details of the equation for u and the process of solving it for specific systems, we will write down the general structure of the function u, valid for any physical system, simply from the conditions (3.4). It has been shown[37] that these conditions can be satisfied if and only if the function $u(\xi,\eta,t)$ has the structure (the result was obtained for the one-dimensional case, but the generalization to N dimensions is trivial)

$$u(\xi,\eta,t) = \int \sum_k \varphi_k^*\Big(x - \frac{1}{2}\xi,t\Big)\varphi_k\Big(x + \frac{1}{2}\xi,t\Big) \times \exp\Big(\frac{i}{\hbar}\eta x\Big)\,dx, \tag{3.7a}$$

where the φ_k are a set of square-integrable functions of the configuration space, satisfying the single condition

$$\int \sum_k |\varphi_k(q,t)|^2\,dq = 1. \tag{3.7b}$$

The non-negative QDF, corresponding to this structure, is given by[37]

$$F_\psi(q,p,t) = (2\pi\hbar)^{-N}\sum_k \bigg|\int \psi(x,t)\varphi_k^*(q-x,t) \times \exp\Big(-\frac{ipx}{\hbar}\Big)\,dx\bigg|^2 \geqslant 0. \tag{3.8}$$

The function (3.8) [or rather a set of such functions for various possible choices of the set $\{\varphi_k\}$, compatible with (3.7b)] coincides with the one proposed by Kuryshkin[15] long ago. Actually the function (3.8) was obtained as a direct consequence of the correspondence rule proposed by him. As stated in Sec. I, the theory with such a correspondence rule was given the quite natural sounding name of "quantum mechanics with a non-negative QDF." What has been achieved by us in this section is thus a reformulation of this theory in terms of the probability operator. We hope to show in the following sections that such a reformulation is an essential and vital step toward extending the scope of this theory to the case of relativistic fields.

IV. PROBABILITY OPERATORS OF ONE-DIMENSIONAL FIELD OSCILLATORS

It will be shown in later sections that a generalization of the quantization scheme developed so far in this paper to the case of a field without interaction requires the field to be represented as a system of noninteracting one-dimensional oscillators. It is widely known[46,47] that such a representation is possible. The probability operator of such a system of oscillators will be obtained from those of the one-dimensional ones comprising the system in a constructive manner. Hence, with a view towards future applications, we consider in this section the problem of constructing the probability operators of such one-dimensional "field oscillators," which may be of two types: those having positive and those having negative contributions, respectively, to the classical field energy. Besides, in conformity with standard practice, we postulate the existence of only two types of algebra—Bose–Einstein and Fermi–Dirac. It thus follows that we have to investigate four individual problems (two types of oscillators × two types of algebra). We will deal with all four cases in this section.

Case 1: We consider the case of an ordinary one-dimensional harmonic oscillator of unit mass with the classical Hamiltonian

$$H(q,p) = p^2/2 + \omega^2 q^2/2, \tag{4.1}$$

and assume that the generators $\hat{q},\hat{p}$ satisfy Heisenberg commutation relations (3.1a) with $N = 1$. For application in the theory of fields we have to seek a representation of the probability operator in terms of the annihilation and creation operators of the oscillator. For this we introduce a pair of complex conjugate variables z, z^* and the familiar annihilation and creation operators (hereafter we adopt the convention $\hbar = 1$)

$$\begin{aligned} z &= (2\omega)^{-1/2}(\omega q + ip), \quad \hat{a} = (2\omega)^{-1/2}(\omega\hat{q} + i\hat{p}),\\ z^* &= (2\omega)^{-1/2}(\omega q - ip), \quad \hat{a}^+ = (2\omega)^{-1/2}(\omega\hat{q} - i\hat{p}). \end{aligned} \tag{4.2}$$

The operators $\hat{a}$ and a^+ satisfy the commutation relation

$$[\hat{a},\hat{a}^+] = \hat{1}, \tag{4.3}$$

defining the Bose algebra.

Physical quantities $A(q,p,t)$ relating to the oscillator (4.1) can now be expressed as functions of the variables z, z^*, and t. We retain the same notation for them although the functional forms will, in general, be quite different. Now introducing the probability operator $\hat{F}(z,z^*,t)$ such that

$$\hat{A} = O(A(z,z^*,t)) = \int A(z,z^*,t)\hat{F}(z,z^*,t)\,d^2z, \tag{4.4}$$

where $d^2z = d(\operatorname{Re} z)d(\operatorname{Im} z)$, and comparing with (2.1) (for $N = 1$), it is readily seen that

$$\hat{F}(z,z^*,t) = 2\hat{F}(q,p,t)\big|_{q=q(z,z^*),p=p(z,z^*)}. \tag{4.5}$$

Conditions (2.2) and Eq. (2.6) are now transformed as

$$\begin{aligned} &\int \hat{F}(z,z^*,t)\,d^2z = \hat{1},\\ &F_\psi(z,z^*,t) = \langle\psi|\hat{F}(z,z^*,t)|\psi\rangle \geqslant 0, \quad \forall|\psi\rangle, \end{aligned} \tag{4.6}$$

$$-i\frac{\partial\hat{F}}{\partial t} + \omega\Big(z^*\frac{\partial\hat{F}}{\partial z^*} - z\frac{\partial\hat{F}}{\partial z}\Big) = \omega\int z'z'^*[\hat{F},\hat{F}']\,d^2z', \tag{4.7}$$

where $\hat{F}' = \hat{F}(z',z'^*,t)$. In obtaining (4.7), we have used the explicit form of the Hamiltonian in new variables:

$$H(z,z^*) = \omega z z^*. \tag{4.8}$$

We will further take $\partial_t\hat{F} = 0$ to guarantee that operators of time-independent quantities such as Hamiltonians do not have explicit time dependence. Thus the first term in the lhs of (4.7) drops out.

Time-independent solutions of (4.7) with the properties (4.6) can be found from the analogous solution in the equivalent (q,p) representation obtained earlier,[48] simply by applying the transformation (4.5). This has been done. The properties of the operator $\hat{F}(z,z^*)$ thus found and the related quantization has been studied in great detail.[49,50] We simply mention the solution here:

$$\hat{F}^{+-}(z,z^*) = \pi^{-2}\int \Omega(\alpha,\alpha^*)\hat{D}(\alpha)\exp(\alpha^* z - \alpha z^*)\,d^2\alpha, \tag{4.9a}$$

where

$$\hat{D}(\alpha) = \exp(\alpha\hat{a}^+ - \alpha^*\hat{a}) \tag{4.9b}$$

is the unitary displacement operator[42,51] and the kernel

$$\Omega(\alpha,\alpha^*) = \sum_{n=0}^{\infty} |c_n|^2 \exp\left(-\frac{1}{2}\alpha\alpha^*\right) L_n(\alpha\alpha^*),$$
$$\sum_{n=0}^{\infty} |c_n|^2 = 1, \tag{4.9c}$$

is the analog of the kernel u of Sec. III. In formula (4.9c), L_n denotes the Laguerre polynomial and the $|c_n|^2$ are a set of non-negative coefficients, arbitrary except for satisfying the second relation in (4.9c). The notation $\hat{F}^{+-}$ has the following meaning. The plus sign signifies the fact that the probability operator relates to an oscillator with positive-definite energy (4.8) and the minus sign is related to the type of algebra (Bose, in this case). The corresponding operator for the same oscillator but in the Fermi algebra (case 2 of our study) thus will be denoted by $\hat{F}^{++}$.

We also mention that, with the help of operator (4.9) and the correspondence (4.4), one can construct the operators (for details, see Refs. 49 and 50)

$$O(z) = \hat{a}, \quad O(z^*) = \hat{a}^+, \tag{4.10}$$
$$\hat{H} = O(H(z,z^*)) = \omega\hat{a}^+\hat{a} + \omega c^{+-}, \tag{4.11a}$$

where

$$c^{+-} = \frac{1}{2} + \sum_n n|c_n|^2 > 0. \tag{4.11b}$$

By exploiting the properties of the displacement operator,[42,51] one can recast the probability operator in the equivalent form[49]

$$\hat{F}^{+-}(z,z^*) = \hat{D}(z)\hat{F}^{+-}(0,0)\hat{D}^+(z), \tag{4.12a}$$
$$\hat{F}^{+-}(0,0) = \frac{1}{\pi}\sum_n |c_n|^2 |n\rangle\langle n|, \tag{4.12b}$$

where $|n\rangle$ are the usual normalized eigenstates of the number operator $\hat{a}^+\hat{a}$. From representation (4.12) the positive-definiteness of $\hat{F}^{+-}$ is quite obvious. It can also be checked easily that the operator (4.12) is normalized. It thus remains to be shown, for self-consistency of this paper, that $\hat{F}^{+-}$ satisfies Eq. (4.7) as well. With this aim in mind we first express the displacement operator (4.9b) in its normally and antinormally ordered forms

$$\hat{D}(\alpha) = e^{-(1/2)\alpha\alpha^*} e^{\alpha\hat{a}^+} e^{-\alpha^*\hat{a}} = e^{(1/2)\alpha\alpha^*} \mathrm{e}^{-\alpha^*\hat{a}} e^{\alpha\hat{a}^+} \tag{4.13}$$

by the use of the Baker–Hausdorff identity.[52] Differentiating these expressions separately with respect to α and α^* (with the convention $\partial\alpha/\partial\alpha^* = \partial\alpha^*/\partial\alpha = 0$ to be observed throughout) we obtain

$$\frac{\partial\hat{D}(\alpha)}{\partial\alpha} = \left(-\frac{1}{2}\alpha^* + \hat{a}^+\right)\hat{D}(\alpha) = \hat{D}(\alpha)\left(\hat{a}^+ + \frac{1}{2}\alpha^*\right), \tag{4.14a}$$

$$\frac{\partial\hat{D}(\alpha)}{\partial\alpha^*} = -\hat{D}(\alpha)\left(\hat{a} + \frac{1}{2}\alpha\right) = -\left(\hat{a} - \frac{1}{2}\alpha\right)\hat{D}(\alpha). \tag{4.14b}$$

From the relations (4.14) and their adjoints, it follows that

$$\begin{aligned} &[\hat{a},\hat{D}(\alpha)] = \alpha\hat{D}(\alpha), \quad [\hat{a}^+,\hat{D}(\alpha)] = \alpha^*\hat{D}(\alpha), \\ &[\hat{a},\hat{D}^+(\alpha)] = -\alpha\hat{D}^+(\alpha), \\ &[\hat{a}^+,\hat{D}^+(\alpha)] = -\alpha^*\hat{D}^+(\alpha). \end{aligned} \tag{4.15}$$

We now observe that the rhs of (4.7) is nothing but the commutator of $\hat{F}(z,z^*)$ with $\hat{H}$. Using the expressions (4.9) and (4.11), respectively, for these operators we obtain the commutator as

$$[\hat{F}^{+-}(z,z^*),\hat{H}] = \omega\left(\hat{a}^+ \frac{\partial\hat{F}^{+-}}{\partial z^*} - \hat{a}\frac{\partial\hat{F}^{+-}}{\partial z}\right). \tag{4.16}$$

To evaluate the derivatives in the right-hand side we use the equivalent expression (4.12) for the operator $\hat{F}^{+-}$, the expressions (4.14) with their adjoints. After straightforward but somewhat lengthy calculations we have the result

$$\begin{aligned} [\hat{F}^{+-}(z,z^*),\hat{H}] &= \omega\left(z^* \frac{\partial\hat{F}^{+-}}{\partial z^*} - z\frac{\partial\hat{F}^{+-}}{\partial z}\right) \\ &\quad + \omega\hat{D}(z)[\hat{F}^{+-}(0,0),\hat{a}^+\hat{a}]\hat{D}^+(z). \end{aligned}$$

The commutator occurring in the right-hand side is evidently zero because of (4.12b) and the fact that $|n\rangle$ are eigenstates of $\hat{a}^+\hat{a}$. With the observation $\partial_t\hat{F}^{+-} = 0$ we find that our proof is complete.

Case 2: Let us now come to the second problem—the quantization of the oscillator (4.8) in the Fermi algebra. The generators of the algebra satisfy

$$[\hat{a},\hat{a}]_+ = [\hat{a}^+,\hat{a}^+]_+ = 0, \quad [\hat{a},\hat{a}^+]_+ = \hat{1}, \tag{4.17}$$

where $[\ ,\ ]_+$ denotes an anticommutator. The required operator $\hat{F}^{++}(z,z^*)$ must satisfy the conditions (4.6) and the equation

$$\omega\left(z^* \frac{\partial\hat{F}}{\partial z^*} - z\frac{\partial\hat{F}}{\partial z}\right) = [\hat{F}(z,z^*),\hat{H}], \tag{4.18a}$$

where

$$\hat{H} = \int H(z,z^*)\hat{F}(z,z^*)d^2z. \tag{4.18b}$$

We also demand that in such a quantization

$$O(z) = \hat{a}, \quad O(z^*) = \hat{a}^+ \tag{4.19}$$

and

$$\hat{H} = \omega(\hat{a}^+\hat{a} + c^{++}), \tag{4.20}$$

where c^{++} is a finite real constant. These demands are consistent with the physical meaning of the generators and the Hamiltonian operator.

The required solution, as can be checked by direct substitution, is the operator

$$F^{++}(z,z^*) = f_1 + f_2\hat{a}^+\hat{a} + z^*f_3\hat{a} + zf_3^*\hat{a}^+, \tag{4.21a}$$

where the functions f_i $(i = 1,2,3)$ are all functions of the argument zz^* and satisfy

$$f_1^* = f_1 \geqslant 0, \quad \int f_1\, d^2z = 1 \tag{4.21b}$$

$$f_2^* = f_2, \quad \int f_2\, d^2z = 0, \quad \int |z|^2 f_2\, d^2z = 1, \tag{4.21c}$$

$$\int |z|^2 f_3\, d^2z = 1, \quad f_1(f_1 + f_2) \geqslant |zf_3|^2. \tag{4.21d}$$

The constant c^{++} in (4.20) can now be written as

$$c^{++} = \int |z|^2 f_1(|z|^2)d^2z > 0. \tag{4.22}$$

The solution (4.21) is the most general solution of (4.18) consistent with the requirements (4.19) and (4.20). Of course, the restrictions (4.21b)–(4.21d) are not sufficient for a unique determination of the functions f_i. Many such choices are possible, and are discussed in detail elsewhere.[53] Just to illustrate that the set $\{f_i\}$ with the properties (4.21b)–(4.21d) is nonempty we mention here the one-parameter family

$$f_1 = \frac{1}{\pi\sigma} e^{-|z|^2/\sigma}, \quad f_2 = \frac{1}{\pi\sigma^3} (|z|^2 - \sigma) e^{-|z|^2/\sigma},$$
$$f_3 = \frac{1}{\pi\sigma^2} e^{-|z|^2/\sigma},$$

where σ is a real non-negative parameter $\geqslant 1$. It can be checked by direct calculation that all conditions are satisfied.

Case 3: The third problem to be investigated is the quantization of a "negative-energy oscillator" (NEO), a system with "classical" Hamiltonian

$$H(z,z^*) = -\omega z z^* \tag{4.23}$$

in the Bose algebra (4.3). As we will see later such a system occurs in the study of spinor fields. The required operator $\hat{F}^{--}$ must satisfy the conditions (4.6). As regards the equation for the operator one has to make a choice between (4.18) and the following:

$$\frac{\partial H}{\partial z}\frac{\partial \hat{F}}{\partial z^*} - \frac{\partial H}{\partial z^*}\frac{\partial \hat{F}}{\partial z} = [\hat{F}(z,z^*), \hat{H}], \tag{4.24a}$$

$$\hat{H} = \int H(z,z^*)\hat{F}(z,z^*)d^2z. \tag{4.24b}$$

For an oscillator with positive energy, Eqs. (4.24), reflecting the principle of dynamical correspondence in terms of the variables z,z^*, simply reduce to Eqs. (4.18). But for the NEO (4.23) they differ [meaning Eqs. (4.18a) and (4.24a)] by the sign of the left-hand side.

However, the question of making a choice is quite irrelevant in the present case, since it will be shown presently that irrespective of such a choice, no probability operator $\hat{F}^{--}(z,z^*)$ exists such that the Hamiltonian possesses the required physical meaning and thus can be written as

$$\hat{H} = \omega(\hat{a}^+\hat{a} + c^{--}), \tag{4.25}$$

where c^{--} as usual is a finite real constant. To show this, let us assume that such an operator $\hat{F}^{--}$ has been found. Since the energy (4.23) is negative and $\hat{F}^{--}$ is positive-definite, by (4.24b) it is obvious that the average value of $\hat{H}$ in any arbitrary state $|\psi\rangle$ must be negative. But in the eigenstate $|n\rangle$, we obtain

$$\langle H\rangle_n = n\omega + \omega c^{--}, \quad n = 0,1,2,\ldots.$$

Since n may be arbitrarily large and c^{--} is finite, the above average must become positive at sufficiently large n. We thus arrive at a contradiction. Hence for a NEO there does not exist any operator $\hat{F}^{--}$ in the Bose algebra compatible with the physical meaning of the Hamiltonian operator.

Case 4: We now consider the fourth and last problem of this section, the quantization of the NEO (4.23) in the Fermi algebra (4.17). As usual, the required operator $\hat{F}^{-+}$ must satisfy (4.6) and the Hamiltonian must be of the form

$$\hat{H} = \omega(\hat{a}^+\hat{a} + c^{-+}), \quad c^{-+} = (c^{-+})^*. \tag{4.26}$$

Assuming that the operator $\hat{F}^{-+}$ satisfies (4.24) we can write down the solution in the form of the operator (4.21a) (see case 2) by interchanging z and z^*, with the corresponding restrictions on the functions f_i. However, the equalities $O(z) = \hat{a}$, $O(z^*) = \hat{a}^+$ cannot now be satisfied for any choice of the set $\{f_i\}$. These violations, as will be shown later, come into conflict with translational invariance. Hence we discard the solution as physically inadmissible. However, this is the most general solution of (4.24) consistent with (4.26). The inevitable conclusion is that Eq. (4.24) itself is physically inadmissible for a NEO and has to be discarded. To save the situation, we now turn to Eq. (4.18), which produces a solution, compatible with all requirements, as

$$\hat{F}^{-+}(z,z^*) = \tilde{f}_1 + \tilde{f}_2\hat{a}^+\hat{a} + z^*\tilde{f}_3\hat{a} + z\tilde{f}_3^*\hat{a}^+, \tag{4.27a}$$

where the functions $\tilde{f}_i$ are again functions of the argument zz^* and satisfy the conditions

$$\tilde{f}_1^* = \tilde{f}_1 \geqslant 0, \quad \int \tilde{f}_1\, d^2z = 1, \tag{4.27b}$$

$$\tilde{f}_2^* = \tilde{f}_2, \quad \int \tilde{f}_2\, d^2z = 0, \quad \int |z|^2\tilde{f}_2\, d^2z = -1, \tag{4.27c}$$

$$\int |z|^2\tilde{f}_3\, d^2z = 1, \quad \tilde{f}_1(\tilde{f}_1 + \tilde{f}_2) \geqslant |z\tilde{f}_3|^2. \tag{4.27d}$$

The constant c^{-+} in the expression (4.26) for $\hat{H}$ is now determined by the function $\tilde{f}_1$ as

$$c^{-+} = -\int |z|^2\tilde{f}_1\, d^2z < 0. \tag{4.28}$$

Again it can be shown that there is sufficient freedom in the choice of the set $\{\tilde{f}_i\}$. It is easy to show that the set $\{\tilde{f}_i\}$ is nonempty. Suppose we have found a solution for case 2. As mentioned earlier, this is always possible. From this solution it is possible to generate a solution for the NEO simply by the transformation $\tilde{f}_1 = f_1 + f_2$, $\tilde{f}_2 = -f_2$, $\tilde{f}_3 = f_3$, since under such a transformation the conditions (4.27b)–(4.27d) for the functions $\tilde{f}_i$ transform to the conditions (4.21b)–(4.21d) for the functions f_i.

Before concluding this section we want to make the following comments: in order to obtain a physically meaningful solution for the probability operator of the NEO, (i) we had to discard the Bose algebra and (ii) we had to discard Eq. (4.24) and decide in favor of Eq. (4.18), which amounts to a redefinition of the classical Poisson bracket.

We thus postulate the equation for the probability operator of an oscillator (either a positive-energy or a negative-energy one) as

$$\varepsilon\left(\frac{\partial H}{\partial z}\frac{\partial \hat{F}}{\partial z^*} - \frac{\partial H}{\partial z^*}\frac{\partial \hat{F}}{\partial z}\right) = \omega\left(z^*\frac{\partial \hat{F}}{\partial z^*} - z\frac{\partial \hat{F}}{\partial z}\right) = [\hat{F}, \hat{H}], \tag{4.29a}$$

$$H(z,z^*) = \varepsilon\omega z z^*, \quad \hat{H} = \int H(z,z^*)\hat{F}(z,z^*)d^2z, \tag{4.29b}$$

where $\varepsilon = +1$ for a positive-energy oscillator and -1 for a negative-energy one. This means that the Poisson bracket in the lhs of (4.29a) must take into account the signature of classical energy. The necessity of such a redefinition can be clear only after a consideration of the quantization of fields to which we now proceed.

V. GENERAL ASPECTS OF QUANTIZATION OF FIELDS ON THE BASIS OF THE PROBABILITY OPERATOR

In the theory of classical fields, all physical quantities are constructions (functions or functionals) from the field function $u(x)=(u_1(x),...,u_N(x))$ and its derivatives, written symbolically as $A(u(x))$. Here as usual $x=(x^0,\mathbf{x})$ denotes the coordinates of Minkowski space-time with the metric tensor $g^{mn}=\mathrm{diag}(1,-1,-1,-1)$. Hence the correspondence rule (2.1) with the probability operator can be generalized to the case of fields as

$$\hat{A}=O(A(u(x)))=\int A(u(x))\hat{F}[u(x)]du(x), \qquad (5.1)$$

where $\hat{F}[u(x)]$ is an operator-valued functional (probability operator) of the field and the integral is taken over all admissible field configurations (solutions of the corresponding classical Lagrange–Euler equations[47]). The conditions (2.2) and Eq. (2.6) will be simply generalized to

$$\int \hat{F}[u(x)]du(x)=\hat{1}, \qquad (5.2a)$$

$$\overset{*}{\Phi}\hat{F}[u(x)]\Phi=F_\Phi[u(x)]\geqslant 0, \qquad (5.2b)$$

$$\partial_{x^0}\hat{F}[u(x)]+\{P^0(u(x)),\hat{F}[u(x)]\}$$
$$=i\int P^0(u'(x))[\hat{F}[u(x)],\hat{F}[u'(x)]]du'(x). \qquad (5.3)$$

Hereafter $\hbar=c=1$, Φ is the arbitrary state vector, $P^0(u(x))$ is the energy of the classical field of the configuration $u(x)$, and $\{\,,\}$ in the lhs of (5.3) denotes the Poisson bracket, to be clarified later. The non-negative functional F_Φ, normalized by virtue of (5.2a), can be interpreted as the probability density of the field configuration $u(x)$ in the state Φ. Specification of the meaning of the functional integrals (domain of integration and measure) appearing in (5.1), (5.2a), and (5.3) obviously requires knowledge of the classical field equations.

In the present paper we restrict ourselves to the consideration of the linear local relativistic theory of a system of fields without interactions, when the Lagrange–Euler equation for each field component is linear and thus in its discrete momentum representation each field component admits[47] Lorenz-invariant decomposition into a sum of positive- and negative-frequency parts in the following manner:

$$u_{sj}(x)=\sum_{s,\mathbf{k},\nu_s}\{v_{sj\mathbf{k}}^{\nu_s+}z_{s\mathbf{k}\nu_s}^{+}e^{ik_sx}+v_{sj\mathbf{k}}^{\nu_s-}z_{s\mathbf{k}\nu_s}^{-}e^{-ik_sx}\}, \qquad (5.4a)$$

$$u_{sj}^{*}(x)=\sum_{s,\mathbf{k},\nu_s}\{\overset{*}{v}{}_{sj\mathbf{k}}^{\nu_s+}\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{+}e^{ik_sx}+\overset{*}{v}{}_{sj\mathbf{k}}^{\nu_s-}\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{-}e^{-ik_sx}\}. \qquad (5.4b)$$

Here u_{sj} is the jth component of the field s, $k_s=(k_s^0,\mathbf{k})$, $k_s^0=\sqrt{\mathbf{k}^2+m_s^2}$, m_s is the mass of the quantum of field s, $k_sx=k_s^0x^0-\mathbf{k}\cdot\mathbf{x}$, and the index ν_s takes care of spin, polarization, etc. The quantities v, being solutions of the corresponding field equations in the momentum representation and reflecting the transformation properties of the fields, can always be suitably normalized to express the classical energy-momentum four-vector P^n through the amplitudes z in the following manner:

$$P^n(u(x))=P^n(\{z_{s\mathbf{k}\nu_s}^{\pm}\},\{\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{\pm}\})$$
$$=\sum_{s,\mathbf{k},\nu_s}k_s^n(\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{+}z_{s\mathbf{k}\nu_s}^{-}+\varepsilon_s\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{-}z_{s\mathbf{k}\nu_s}^{+}), \qquad (5.5a)$$

where $\varepsilon_s=+1$ for the tensor field and -1 for the spinor field. The quantities $v^\pm$ and the amplitudes $z^\pm$ obey the following rules of complex conjugation[47]:

$$(v_{sj\mathbf{k}}^{\nu_s\pm})^*=\overset{*}{v}{}_{sj\mathbf{k}}^{\nu_s\mp},\quad (z_{s\mathbf{k}\nu_s}^{\pm})^*=\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{\mp}. \qquad (5.5b)$$

The relations (5.4) and (5.5) show that any admissible field configuration satisfying free field equations is uniquely represented by the sets of independent amplitudes $\{z_{s\mathbf{k}\nu_s}^{-}\}$ and $\{\overset{*}{z}{}_{s\mathbf{k}\nu_s}^{-}\}$. Thus the operator-valued functional $\hat{F}$ can now be considered as an operator-valued function of the infinite sets of variables $\{z^\pm\}$, $\{\overset{*}{z}{}^\pm\}$:

$$\hat{F}[u(x)]=\hat{F}(\{z_B^\pm\},\{\overset{*}{z}{}_B^\pm\}). \qquad (5.6)$$

Here we have introduced the collective index $B=s\mathbf{k}\nu_s$. The correspondence (5.1) can now be rewritten as

$$\hat{A}=O(A)=\int A(x,\{z_B^\pm\},\{\overset{*}{z}{}_B^\pm\})\hat{F}(\{z_B^\pm\},\{\overset{*}{z}{}_B^\pm\})$$
$$\times\prod_B d^2z_B^-\,d^2\overset{*}{z}{}_B^-, \qquad (5.7)$$

where the integration is over the domains of definition (complex planes) of the independent variables z_B^- and $\overset{*}{z}{}_B^-$.

The corresponding restatement of (5.3) in terms of $z_B^\pm$ and $\overset{*}{z}{}_B^\pm$ requires specification of the Poisson-bracket symbol $\{\,,\}$. For this we consider the different arguments of $\hat{F}$ and the terms in (5.3) corresponding to them separately.

(1) The time variable x^0. As in the case of nonrelativistic theory, we take $\hat{F}$ to be independent of x^0. Thus $\partial_{x^0}\hat{F}=0$.

(2) The independent variable z^- (for simplicity we drop the index B). The corresponding contribution $k^0|z^-|^2$ to the field energy can be reduced to that of a harmonic oscillator by the transformations

$$z^-=(2k^0)^{-1/2}(\omega q+ip),\quad \overset{*}{z}{}^+=(2k^0)^{-1/2}(\omega q-ip),$$
$$P^0(z^-,\overset{*}{z}{}^+)=p^2/2+\omega^2q^2/2=H(q,p),$$
$$d^2z^-=(2k^0)^{-1}\omega\,dq\,dp,$$
$$\hat{F}(...,z^-,\overset{*}{z}{}^-,...)=(2k^0/\omega)\hat{F}(...,q,p,...).$$

Using now Eq. (2.6) of Sec. II, we obtain

$$\omega\left(\overset{*}{z}{}^+\frac{\partial\hat{F}}{\partial\overset{*}{z}{}^+}-z^-\frac{\partial\hat{F}}{\partial z^-}\right)$$
$$=\int k^0\overset{*}{z}{}'^+z'^-[\hat{F},\hat{F}']\cdots d^2z'^-\cdots. \qquad (5.8)$$

(3) The independent variable $\overset{*}{z}{}^-$, with the contribution $\varepsilon|\overset{*}{z}{}^-|^2$ ($\varepsilon=\pm1$) to the field energy. Similar to the previous case, we obtain

$$\overset{*}{\omega}(\varepsilon)\left(z^+\frac{\partial\hat{F}}{\partial z^+}-\overset{*}{z}{}^-\frac{\partial\hat{F}}{\partial\overset{*}{z}{}^-}\right)$$
$$=\int\varepsilon k^0z'^+\overset{*}{z}{}'^-[\hat{F},\hat{F}']\cdots d^2\overset{*}{z}{}'^-\cdots. \qquad (5.9)$$

Actually, only the equation for $\varepsilon = +1$ was obtained by variable transformation to a harmonic oscillator and thus $\overset{*}{\omega}(1) > 0$. For $\varepsilon = -1$, there is no such transformation and the corresponding equation has been written by pure analogy (see the comments at the end of Sec. IV). Thus $\overset{*}{\omega}(-1)$ is merely a real constant (following from the Hermiticity of $\widehat{F}$), but not necessarily positive.

Note that Eqs. (5.8) and (5.9) contain arbitrary sets of real constants $\{\omega_{s\mathbf{k}\nu_s}\},\{\overset{*}{\omega}_{s\mathbf{k}\nu_s}(\varepsilon_s)\}$. For a unique determination of these constants and thus for a unique specification of the Poisson-bracket concept, we have to resort to some general physical principle. We show in Sec. VI that the requirement of translational invariance of the theory provides us with such a principle.

VI. TRANSLATIONAL INVARIANCE AND THE EXPLICIT FORM OF THE EQUATION FOR THE PROBABILITY OPERATOR OF A SYSTEM OF FREE FIELDS

The well known condition of compatibility of the transformational properties of the field operators $\hat{u}_{sj}(x)$ and the state vector Φ with respect to translation[47] is

$$i\frac{\partial \hat{u}_{sj}(x)}{\partial x_n} = [\hat{u}_{sj}(x),\widehat{P}^n]. \tag{6.1}$$

Substituting in this, for $n=0$, the field operators $\hat{u}_{sj}(x)$ and the energy operator $\widehat{P}^0$, constructed by the general recipe (5.7) from the classical expressions (5.4a) and (5.5a), respectively, we obtain

$$\begin{aligned}\pm k^0_B &\int z^{\mp}_B \widehat{F}(\{z^{\pm}_B\},\{\overset{*}{z}{}^{\pm}_B\}) \prod_B d^2 z^-_B\, d^2\overset{*}{z}{}^-_B\\ &= \int z^{\mp}_B P^0(\{z^{\pm}_{B'}\},\{\overset{*}{z}{}^{\pm}_{B'}\})\\ &\quad\times[\widehat{F}(\{z^{\pm}_B\},\{\overset{*}{z}_B\}),\widehat{F}(\{z^{\pm}_{B'}\},\{\overset{*}{z}{}^{\pm}_{B'}\})]\\ &\quad\times\prod_B d^2 z^-_B\, d^2\overset{*}{z}{}^-_B \prod_{B'} d^2 z^-_{B'}\, d^2\overset{*}{z}{}^-_{B'}.\end{aligned}$$

Transforming the rhs by means of the relations (5.8) and (5.9) of the previous section, we can write

$$\begin{aligned}\text{rhs} = \int z^{\mp}_B \sum_c &\Big\{\omega_c\Big(\overset{*}{z}{}^+_c \frac{\partial}{\partial\overset{*}{z}{}^+_c} - z^-_c\frac{\partial}{\partial z^-_c}\Big)\\ &+ \overset{*}{\omega}_c(\varepsilon_c)\Big(z^+_c\frac{\partial}{\partial z^+_c} - \overset{*}{z}{}^-_c\frac{\partial}{\partial \overset{*}{z}{}^-_c}\Big)\Big\}\\ &\times \widehat{F}(\{z^{\pm}_c\},\{\overset{*}{z}{}^{\pm}_c\})\prod_c d^2 z^-_c\, d^2\overset{*}{z}{}^-_c.\end{aligned} \tag{6.2}$$

The above expression can be integrated by using the following integral identities:

$$\begin{aligned}&\int zz^*\,\partial_{z^*}\widehat{F}\,d^2z = \frac{1}{2}\int z^2\,\partial_z\widehat{F}\,d^2z = -\int z\widehat{F}\,d^2z,\\ &\int zz^*\,\partial_z\widehat{F}\,d^2z = \frac{1}{2}\int (z^*)^2\,\partial_{z^*}\widehat{F}\,d^2z = -\int z^*\widehat{F}\,d^2z,\\ &\int z^*\,\partial_{z^*}\widehat{F}\,d^2z = \int z\,\partial_z\widehat{F}\,d^2z = -\int \widehat{F}\,d^2z,\end{aligned} \tag{6.3}$$

where $\widehat{F}=\widehat{F}(\ldots,z,\overset{*}{z},\ldots)$ and $\widehat{F}\to 0$ as $|z|\to\infty$. The identities (6.3) can be readily checked by changing to the polar coordinates $z = r\exp(i\theta)$, $d^2z = r\,dr\,d\theta$. Now considering the expression (6.2) for different independent z^-_B and $\overset{*}{z}{}^-_B$, integrating and equating to the lhs, we obtain the unique result

$$\omega_{s\mathbf{k}\nu_s} = k^0_s,\quad \overset{*}{\omega}_{s\mathbf{k}\nu_s}(\varepsilon_s) = k^0_s,$$

both for $\varepsilon_s = +1$ and $\varepsilon_s = -1$.

Finally, summing up the relations (5.8) and (5.9), written separately for each independent variable, we obtain the equation for the probability operator of a system of free fields:

$$\begin{aligned}\sum_B k^0_B&\Big\{\Big(\overset{*}{z}{}^+_B\frac{\partial}{\partial\overset{*}{z}{}^+_B} - z^-_B\frac{\partial}{\partial z^-_B}\Big)\\ &+\Big(z^+_B\frac{\partial}{\partial z^+_B} - \overset{*}{z}{}^-_B\frac{\partial}{\partial\overset{*}{z}{}^-_B}\Big)\Big\}\widehat{F}(\{z^{\pm}_B\},\{\overset{*}{z}{}^{\pm}_B\})\\ &= \int [\widehat{F}(\{z^{\pm}_B\},\{\overset{*}{z}{}^{\pm}_B\}),\widehat{F}(\{z^{\pm}_{B'}\},\{z^{*\pm}_{B'}\})]\\ &\qquad\times P^0(\{z^{\pm}_{B'}\},\{\overset{*}{z}{}^{\pm}_{B'}\})\prod_{B'} d^2z^-_{B'}\, d^2\overset{*}{z}{}^-_{B'},\end{aligned} \tag{6.4}$$

where $k^0_B = k^0_s$, B, and B' are collective indices.

Comparison of Eq. (6.4) with the same equation written in the form (5.3) of the previous section, taking into account the relations (5.5a) and (5.7) and the fact that $\partial_{x^0}\widehat{F}=0$, leads to an explicit definition of the classical Poisson bracket which automatically takes into account the signatures ε_s appearing in the classical expression (5.5a) for the field energy (see the comment at the end of Sec. IV).

Thus the problem of quantizing a system of free fields in the formalism of the probability operator is reduced to that of finding a solution of Eq. (6.4), restricted further by the conditions

$$\int \widehat{F}(\{z^{\pm}_B\},\{\overset{*}{z}{}^{\pm}_B\})\prod_B d^2z^-_B\, d^2\overset{*}{z}{}^-_B = \hat{1}, \tag{6.5a}$$

$$\overset{*}{\Phi}\widehat{F}(\{z^{\pm}_B\},\{\overset{*}{z}{}^{\pm}_B\})\Phi = F_{\Phi}(\{z^{\pm}_B\},\{\overset{*}{z}{}^{\pm}_B\})\geqslant 0. \tag{6.5b}$$

It is to be noted that the quantization scheme considered here depends both on the structure of the system of fields to be quantized [classical energy P^0 explicitly enters Eq. (6.4)] and the predetermined algebra of operators. It is thus remarkable that irrespective of these dependences, for the operators $O(z^{\pm}_{s\mathbf{k}\nu_s})$, $O(\overset{*}{z}{}^{\pm}_{s\mathbf{k}\nu_s})$, and $\widehat{P}^0$ determined by the rule (5.7) we have the relations

$$\begin{aligned}\mp k^0_s O(z^{\pm}_{s\mathbf{k}\nu_s}) &= [O(z^{\pm}_{s\mathbf{k}\nu_s}),\widehat{P}^0]\\ \mp k^0_s O(\overset{*}{z}{}^{\pm}_{s\mathbf{k}\nu_s}) &= [O(\overset{*}{z}{}^{\pm}_{s\mathbf{k}\nu_s}),\widehat{P}^0]\end{aligned} \tag{6.6}$$

as direct consequences of Eq. (6.4). This can be readily seen by multiplying both sides of (6.4) by $z^-_{s\mathbf{k}\nu_s}$, $z^+_{s\mathbf{k}\nu_s}$, $\overset{*}{z}{}^-_{s\mathbf{k}\nu_s}$, and $\overset{*}{z}{}^+_{s\mathbf{k}\nu_s}$ (separately each time) and integrating over all variables with the use of identities (6.3). These relations lead to the physical meaning of the operators $O(z^-_{s\mathbf{k}\nu_s})$, $O(\overset{*}{z}{}^-_{s\mathbf{k}\nu_s})$ as the annihilation operators and $O(\overset{*}{z}{}^+_{s\mathbf{k}\nu_s})$, $O(z^+_{s\mathbf{k}\nu_s})$ as the creation operators of the quantum of energy $k^0_s = \sqrt{\mathbf{k}^2 + m^2_s}$ of the fields s. Relations similar to (6.6) with $\mathbf{k}$ and $\mathbf{P}$ replacing k^0_s and P^0 can also be easily established, since along with Eq. (5.3), with $\partial_{x^0} = 0$, the probability operator also satisfies the equations

$$\{\mathbf{P},\hat{F}\} = i[\hat{F}[u(x)],\hat{\mathbf{P}}]. \tag{6.7}$$

This can be seen by starting from (6.1), for $n = 1,2,3$, and following the same procedures. Hence we can write

$$\begin{aligned} (O(z^{\pm}_{s\mathbf{k}\nu_s}))^{+} &= O(\overset{*}{z}{}^{\mp}_{s\mathbf{k}\nu_s}), \\ [\hat{P}^{n},O(z^{\pm}_{s\mathbf{k}\nu_s})] &= \pm k^{n}_{s}O(z^{\pm}_{s\mathbf{k}\nu_s}), \end{aligned} \tag{6.8}$$

leading to the complete identification of the operators $O(z^{-}_{s\mathbf{k}\nu_s})$, $O(\overset{*}{z}{}^{-}_{s\mathbf{k}\nu_s})$ as annihilation operators and of $O(\overset{*}{z}{}^{+}_{s\mathbf{k}\nu_s})$, $O(z^{+}_{s\mathbf{k}\nu_s})$ as creation operators of the quantum of energy-momentum $k^{n}_{s} = (\sqrt{\mathbf{k}^2 + m_s^2},\mathbf{k})$, whereby m_s can now strictly be identified as the mass of the quantum of field s.

VII. QUANTIZATION OF A SYSTEM OF FREE FIELDS IN THE FORMALISM OF THE PROBABILITY OPERATOR

For the sake of convenience we introduce the "partial" operators $\hat{F}^{1}_{s\mathbf{k}\nu_s}$ and $\hat{F}^{2}_{s\mathbf{k}\nu_s}$, connected with the probability operator by the integral relations

$$\hat{F}^{1}_{B}(z^{-}_{B},\overset{*}{z}{}^{+}_{B}) = \int \hat{F}(\{z^{\pm}_{B'}\},\{\overset{*}{z}{}^{\pm}_{B'}\}) \times \prod_{B'\neq B} d^2 z^{-}_{B'} \prod_{B'} d^2 \overset{*}{z}{}^{-}_{B'}, \tag{7.1a}$$

$$\hat{F}^{2}_{B}(\overset{*}{z}{}^{-}_{B},z^{+}_{B}) = \int \hat{F}(\{z^{\pm}_{B'}\},\{\overset{*}{z}{}^{\pm}_{B'}\}) \times \prod_{B'} d^2 z^{-}_{B'} \prod_{B'\neq B} d^2 \overset{*}{z}{}^{-}_{B'}. \tag{7.1b}$$

It immediately follows that the energy-momentum operator can be written as

$$\hat{P}^{n} = \sum_{s,\mathbf{k},\nu_s} k^{n}_{s}(\hat{N}^{1}_{s\mathbf{k}\nu_s} + \hat{N}^{2}_{s\mathbf{k}\nu_s}), \tag{7.2a}$$

$$\begin{aligned} \hat{N}^{1}_{B} &= \int |z|^2 \hat{F}^{1}_{B}(z,z^*)d^2 z, \\ \hat{N}^{2}_{B} &= \varepsilon_B \int |z|^2 \hat{F}^{2}_{B}(z,z^*)d^2 z. \end{aligned} \tag{7.2b}$$

Carrying out the integrations over all but one variable in the relations (6.4) and (6.5), we further obtain

$$k^{0}_{B}\left(z^* \frac{\partial}{\partial z^*} - z\frac{\partial}{\partial z}\right)\hat{F}^{i}_{B}(z,z^*) = [\hat{F}^{i}_{B}(z,z^*),\hat{P}^{0}], \tag{7.3a}$$

$$\int \hat{F}^{i}_{B}(z,z^*)d^2 z = \hat{1}, \tag{7.3b}$$

$$\overset{*}{\Phi}\hat{F}^{i}_{B}(z,z^*)\Phi = F^{i}_{B}(z,z^*) \geqslant 0. \tag{7.3c}$$

Let us now go over to the construction of the probability operator of a system of free fields as the solution of Eq. (6.4) with the properties (6.5). We will make the following assumptions (by analogy with the standard theory of quantized fields[47]).

(a) The algebra $\mathscr{A}$, to which the probability operator and consequently operators of all physical quantities belong, is the standard[47] product of Bose and Fermi algebras.

(b) The generators of $\mathscr{A}$ are the creation operators $\hat{a}^{+}_{B}$, $\hat{\overset{*}{a}}{}^{+}_{B}$ and annihilation operators $\hat{a}^{-}_{B}$, $\hat{\overset{*}{a}}{}^{-}_{B}$ of the quanta of the field, corresponding by rule (5.7) to the classical amplitudes z^{+}_{B}, $\overset{*}{z}{}^{+}_{B}$, z^{-}_{B}, $\overset{*}{z}{}^{-}_{B}$, respectively.

(c) The operators $\hat{N}^{i}_{B}$ in the expression for the energy-momentum operator, determined by the relations (7.2b) and (7.1), coincide with (apart from additive c numbers) the number operators of the corresponding quanta:

$$\hat{N}^{1}_{B} = \hat{a}^{+}_{B}\hat{\overset{*}{a}}{}^{-}_{B} + c^{1}_{B}, \quad \hat{N}^{2}_{B} = \hat{\overset{*}{a}}{}^{+}_{B}\hat{a}^{-}_{B} + c^{2}_{B}, \tag{7.4}$$

where the c^{i}_{B} are finite c-number constants.

Assumptions (a)–(c), regarding the nature of the algebra, the physical meaning of the generators, and the physical meaning of the energy-momentum operator (7.2a), allow us to solve the problem formulated above in a constructive manner.

Thus from (a) and (c) it follows that the operators $\hat{N}^{i}_{B}$ commute with any operator of the subalgebra $\mathscr{A}_{jB} \in \mathscr{A}$, not containing the generators entering $\hat{N}^{i}_{B}$. Hence

$$[\hat{F}^{i}_{B}(z,z^*),\hat{P}^{0}] = k^{0}_{B}[\hat{F}^{i}_{B}(z,z^*),\hat{N}^{i}_{B}], \tag{7.5}$$

and from (7.2a), (7.4), and (7.5), we obtain the equation for $\hat{F}^{i}_{B}$:

$$\left(z^* \frac{\partial}{\partial z^*} - z\frac{\partial}{\partial z}\right)\hat{F}(z,z^*) = [\hat{F}(z,z^*),\hat{a}^{+}\hat{a}^{-}]. \tag{7.6a}$$

Here and from now on indices are dropped whenever possible without invoking confusion.

Conditions on $\hat{F}(z,z^*)$ in (7.6a), following from (7.2b), (7.3b), and (7.3c) and the assumptions (b) and (c) are

$$\int \hat{F}(z,z^*)d^2 z = \hat{1}, \quad \overset{*}{\Phi}\hat{F}\Phi \geqslant 0, \tag{7.6b}$$

$$\int z\hat{F}(z,z^*)d^2 z = \hat{a}^{-}, \quad \int z^*\hat{F}(z,z^*)d^2 z = \hat{a}^{+}, \tag{7.6c}$$

$$\pm \int |z|^2 \hat{F}(z,z^*)d^2 z = \hat{a}^{+}\hat{a} + c, \quad |c| < \infty. \tag{7.6d}$$

Here the plus in (7.6d) relates to the operator $\hat{F}^{1}_{B}$ and to the operator $\hat{F}^{2}_{B}$ for $\varepsilon_B = +1$ while the minus relates to the operator $\hat{F}^{2}_{B}$ for $\varepsilon_B = -1$.

Besides, according to assumption (a), the operators $\hat{a}^{-}$ and $\hat{a}^{+}$ satisfy one of the following two types of commutation relations:

$$[\hat{a}^{-},\hat{a}^{+}]_{\pm} = \hat{1}, \quad [\hat{a}^{-},\hat{a}^{-}]_{\pm} = 0. \tag{7.7}$$

Thus the problem of finding all the operators $\hat{F}^{i}_{B}$ is reduced to solving four typical problems (7.6) [two types of conditions (7.6d) × two types of algebra (7.7)]. All of these four one-dimensional problems were studied in Sec. IV. We obtained the result that the operators $\hat{F}^{+-}$, $\hat{F}^{++}$, and $\hat{F}^{-+}$ exist while $\hat{F}^{--}$ does not exist. Thus we can conclude that quantization of the spinor field, in the formalism of probability operator, cannot be carried out in the Bose algebra. This is the analog of the well-known Pauli theorem in the formalism of the probability operator.

We now construct the probability operator of a system of free fields [solution of (6.4) with properties (6.5)] as a symmetrized product of the operators $\hat{F}^{+-}$, $\hat{F}^{++}$, $\hat{F}^{-+}$ of the one-dimensional problems, with the following observations.

(i) To each tensor degree of freedom, in the probability operator there corresponds an operator $\widehat{F}^{+-}$, to each spinor degree of freedom with $\varepsilon_s = +1$ there corresponds an operator $\widehat{F}^{++}$ and to each spinor degree of freedom with $\varepsilon_s = -1$ an operator $\widehat{F}^{-+}$. The operators enter with corresponding indices.

(ii) The probability operator $\widehat{F}$, as a symmetrized product of $\widehat{F}^{+-}$, $\widehat{F}^{++}$, and $\widehat{F}^{-+}$, automatically satifies Eq. (6.4), normalization (6.5a), and relations (7.1).

(iii) The condition of positive-definiteness (6.5b) is trivially satisfied for the product of the operators $\widehat{F}^{+-}$ (generators commute), but imposes additional limitations on the parameter functions $(f_i, \tilde{f}_i)$ for the symmetrized product of the operators $\widehat{F}^{++}$ and $\widehat{F}^{-+}$ (the corresponding generators of the Fermi algebra anticommute). For details, see Ref. 53.

However, it is to be emphasized that the explicit mathematical expression for the probability operator $\widehat{F}$ of a system of free fields is only of academic interest, the most important fact being its existence. In practice one only needs to know the operators $\widehat{F}^{+-}$, $\widehat{F}^{++}$, and $\widehat{F}^{-+}$ and their different paired symmetrized products. This is explained by the fact that all the physical quantities for such a system are either linear in amplitudes $z^{\pm}_{s\mathbf{k}\nu_s}$, $\overset{*}{z}{}^{\pm}_{s\mathbf{k}\nu_s}$ (e.g., field components and their positive- and negative-frequency parts), or quadratic in them [energy-momentum four-vector, charge

$$Q(u(x)) = \sum_{s,\mathbf{k},\nu_s} (|z^{-}_{s\mathbf{k}\nu_s}|^2 - \varepsilon_s |\overset{*}{z}{}^{-}_{s\mathbf{k}\nu_s}|^2), \tag{7.8}$$

and others], or bilinear (Lagrangian, tensor of energy-momentum, current, etc.). To find the corresponding operators it is sufficient to know the "partial" operators (7.1) and the integral of $\widehat{F}$ over all variables except two. These latter ones are, because of the established structure of $\widehat{F}$, given by a typical symmetrized product

$$\widehat{F}^{i_1 i_2}_{B_1 B_2}(z_1,z_1^*,z_2,z_2^*) = \frac{1}{2}\,[\widehat{F}^{i_1}_{B_1}(z_1,z_1^*),\widehat{F}^{i_2}_{B_2}(z_2,z_2^*)]_+, \tag{7.9}$$

where each $\widehat{F}^{i}_{B}(z,z^*)$ is one of the operators $\widehat{F}^{+-}$, $\widehat{F}^{++}$, or $\widehat{F}^{-+}$.

We will now study the simple consequences of quantization based on the probability operator. By the general rule (5.7) of Sec. V we find the operators of energy and charge from the corresponding classical expressions as

$$\widehat{P}^0 = \sum_B k^0_B (\hat{a}^+_B \hat{\overset{*}{a}}{}^-_B + \hat{\overset{*}{a}}{}^+_B \hat{a}^-_B) + P^0_{(0)}, \tag{7.10}$$

$$\widehat{Q} = \sum_B (\hat{a}^+_B \hat{\overset{*}{a}}{}^-_B - \hat{\overset{*}{a}}{}^+_B \hat{a}^-_B) + Q_{(0)}, \quad\bullet \tag{7.11}$$

where the "vacuum" contributions $P^0_{(0)}, Q_{(0)}$ can be written as

$$\begin{gathered} P^0_{(0)} = \sum_B (\mathbf{k}^2 + m_B^2)^{1/2} p_{(0)}(B,\varepsilon_B), \\ Q_{(0)} = \sum_B q_{(0)}(B,\varepsilon_B). \end{gathered} \tag{7.12}$$

The quantities $p_{(0)}(B,\varepsilon_B)$, $q_{(0)}(B,\varepsilon_B)$ are the constants c^{+-}, c^{++}, c^{-+} of Sec. IV and are expressible through the quantization parameters $[\{c_n\},\{f_i\},\{\tilde{f}_i\}]_{s\mathbf{k}\nu_s}$ of the individual operators $\widehat{F}^{+-}$, $\widehat{F}^{++}$, and $\widehat{F}^{-+}$. Nullification of $P^0_{(0)}$ and $Q_{(0)}$, if possible, is to be attained only through the choices of the quantization parameters. Such formal nullification attained in the standard theory of the quantized fields[47] by writing the operators in a normal form is not applicable in our case, since such a procedure is incompatible with the rule (5.7) of construction of operators and the properties (6.5) of the probability operator. Below we consider concrete examples.

System of tensor fields: Here all $p_{(0)}(B,1) = c_B^{+-} > 0$ [see formula (4.11b) of Sec. IV], and thus $P^0_{(0)} = \infty$. Hence a system comprising only tensor fields is not quantizable in the formalism of the probability operator.

System of spinor fields: Here all $q_{(0)}(B,1) = c_B^{++} > 0$ [see formula (4.22) of Sec. IV], and all $q_{(0)}(B,-1) = -c_B^{-+} > 0$ [see formula (4.28) of Sec. IV]. As a result, $Q_{(0)} = \infty$. Thus, again, a system consisting only of spinor fields is not quantizable.

System of tensor and spinor fields: Here it is possible, in principle, to attain simultaneously both the equalities $P^0_{(0)} = 0$, $Q_{(0)} = 0$ by a proper choice of quantization parameters, not, however, for any arbitrary system configuration. We illustrate this by the example of the quantization of a system consisting of a complex vector field $V^n(x)$ and a Dirac spinor field $\Psi(x)$. Denoting the polarization index of the vector field ν_V by α and the spin index ν_S of the spinor field by σ, we can write the conditions of zero vacuum energy and zero vacuum charge as

$$\sqrt{\mathbf{k}^2 + m_V^2}\sum_\alpha p^V_{(0)\mathbf{k}\alpha} + \sqrt{\mathbf{k}^2 + m_S^2}\sum_\sigma p^S_{(0)\mathbf{k}\sigma} = 0, \tag{7.13}$$

$$\sum_\alpha q^V_{(0)\mathbf{k}\alpha} + \sum_\sigma q^S_{(0)\mathbf{k}\sigma} = 0. \tag{7.14}$$

Expressing the quantities $p_{(0)}$ and $q_{(0)}$ through the parameters $\{c_n\}$, $\{f_i\}$, and $\{\tilde{f}_i\}$ of the quantization we obtain

$$\begin{aligned} &\sqrt{\mathbf{k}^2 + m_V^2}\sum_\alpha (2 + \mu_1(\mathbf{k},\alpha) + \mu_2(\mathbf{k},\alpha)) \\ &\quad + \sqrt{\mathbf{k}^2 + m_S^2}\sum_\sigma \left[\int |z|^2 (f_{1\mathbf{k}\sigma} - \tilde{f}_{1\mathbf{k}\sigma})\, d^2 z\right] = 0, \end{aligned}$$

$$\begin{aligned} &\sum_\alpha (\mu_1(\mathbf{k},\alpha) - \mu_2(\mathbf{k},\alpha)) \\ &\quad + \sum_{|\sigma|} \left[\int |z|^2 (f_{1\mathbf{k}\sigma} + \tilde{f}_{1\mathbf{k}\sigma})\, d^2 z\right] = 0, \end{aligned}$$

where

$$\mu_1(\mathbf{k},\alpha) = \sum_{n=0}^{\infty} n|c^1_{n\mathbf{k}\alpha}|^2, \quad \mu_2(\mathbf{k},\alpha) = \sum_{n=0}^{\infty} n|c^2_{n\mathbf{k}\alpha}|^2.$$

From the above two equations it is easy to derive

$$\begin{aligned} &2\sum_\sigma \int |z|^2 f_{1\mathbf{k}\sigma}\, d^2 z \\ &\quad = \Bigg[\sum_\alpha \mu_2(\mathbf{k},\alpha) - \mu_1(\mathbf{k},\alpha) \\ &\qquad - \sqrt{\frac{\mathbf{k}^2 + m_V^2}{\mathbf{k}^2 + m_S^2}}\,(2 + \mu_1(\mathbf{k},\alpha) + \mu_2(\mathbf{k},\alpha))\Bigg]. \end{aligned} \tag{7.15}$$

This last relation cannot be satisfied if $m_V \geqslant m_S$, by any choices of the non-negative parameters μ_1, μ_2, since in that case the rhs is negative while the lhs is positive. Thus for a noncontradictory quantization (in the sense of null vacuum energy and null vacuum charge) m_V must be less than m_S. Such a limitation can be lifted by adding another tensor field to the system.

Thus in this paper we have presented a general theoretical framework for the quantization of physical systems on the basis of the probability operator. It has been shown that the resulting quantum theory, contrary to the conventional one, has a consistent probabilistic interpretation. In the nonrelativistic case, such quantization leads to the theory, previously developed under the name of "quantum mechanics with a non-negative QDF." We have shown further that a logical and mathematically consistent generalization of this framework to the case of relativistic theory of fields is possible, at least in the case of free fields. Such a scheme of quantization, based on the probability operator, puts forward definite limitations not only on the parameters of quantization, but also on the structure of the field to be quantized (only a system comprising tensor and spinor fields simultaneously is quantizable with definite restrictions on the masses of quanta), already at the level of free fields. The probability operator thus can be said to act as a selection criterion. A distinct advantage of this quantization scheme is that the average values of relevant physical quantities (and matrix elements between two states) can be calculated in a classical manner (by phase-space integration in nonrelativistic theory and by functional integration in the case of fields). To calculate the average value (or matrix elements) of any quantity it is sufficient to know only the average (or matrix elements) of the single probability operator. This might be of great advantage in calculating matrix elements of the scattering matrix in the theory of fields. Thus, in our opinion, it might be interesting and promising to study further implications of such quantization by extending the framework proposed in this paper to include the case of interacting fields. Needless to say, there remain many obstacles, primarily of a mathematical nature, to be overcome before such an extension is possible, which thus remains a study for the future.

ACKNOWLEDGMENTS

The authors gratefully acknowledge stimulating and helpful discussions with Professor Ya. P. Terletsky and several other colleagues of the Department of Theoretical Physics of the Patrice Lumumba Peoples' Friendship University.

[1] G. Temple, Nature **135**, 951 (1935).
[2] H. J. Groenwold, Physica **12**, 405 (1946).
[3] J. R. Shewell, Am. J. Phys. **27**, 16 (1959).
[4] V. V. Kuryshkin, I. A. Lyabis, and Yu. I. Zaparovanny, Ann. Fond. L. de Broglie **3**, 45 (1978).
[5] P. R. Chernoff, Hadronic J. **4**, 879 (1981).
[6] A. J. Kalnay, Hadronic J. **4**, 1127 (1981).
[7] J. von Neumann, Nachr. Akad. Wiss. Göttingen Math. Phys. **Kl**, 245 (1927); *Mathematical Foundations of Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1955).
[8] P. A. M. Dirac, Proc. R. Soc. London Ser. A **110**, 561 (1926); *The Principles of Quantum Mechanics* (Oxford U.P., Oxford, 1958).
[9] H. Weyl, Z. Phys. **46**, 1 (1927); *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1950).
[10] M. Born and P. Jordan, Z. Phys. **34**, 858 (1925).
[11] D. C. Rivier, Phys. Rev. **83**, 862 (1951).
[12] C. L. Mehta, J. Math. Phys. **5**, 677 (1964).
[13] Y. Kano, J. Phys. Soc. (Jpn.) **19**, 1558 (1964).
[14] L. Cohen, J. Math. Phys. **7**, 781 (1966).
[15] V. V. Kuryshkin, Izv. Vusov Fiz. **11**, 102 (1971).
[16] P. B. Guest, Rep. Math. Phys. **6**, 99 (1974).
[17] L. Castellani, Nuovo Cimento A **48**, 359 (1978).
[18] S. Basu, Ind. J. Phys. A **57**, 67 (1983).
[19] V. V. Kuryshkin, Ann. Inst. H. Poincaré **17**, 81 (1972).
[20] V. V. Kuryshkin, Int. J. Theor. Phys. **7**, 451 (1973).
[21] V. V. Kuryshkin, C. R. Acad. Sci. (Paris) B **274**, 1163 (1972).
[22] V. V. Kuryshkin and Yu. I. Zaparovanny, C. R. Acad. Sci. (Paris) B **277**, 17 (1974).
[23] V. V. Kuryshkin, in *The Uncertainty Principle and Foundations of Quantum Mechanics* (Wiley, London, 1977), p. 61.
[24] E. Wigner, Phys. Rev. **40**, 749 (1932).
[25] Ya. P. Terletsky, Zh. Eksp. Teor. Fiz. **7**, 1290 (1937).
[26] D. I. Blokhintzev, J. Phys. (Moscow) **2**, 71 (1940).
[27] J. E. Moyal, Proc. Camb. Philos. Soc. **45**, 99 (1949).
[28] H. Margenau and R. N. Hill, Prog. Theor. Phys. **26**, 722 (1961).
[29] T. S. Shankara, Prog. Theor. Phys. **37**, 1335 (1967).
[30] F. Bopp, Ann. Inst. H. Poincaré **15**, 81 (1956).
[31] Y. Kano, J. Math. Phys. **6**, 1913 (1965).
[32] L. Cohen, Phil. Sci. **33**, 317 (1966).
[33] R. F. O'Connell and E. P. Wigner, Phys. Lett. A **85**, 121 (1981).
[34] E. Prugovečki, Phys. Rev. Lett. **49**, 1065 (1982).
[35] P. Bertrand, J. P. Doremus, B. Izrar, V. T. Nguyen, and M. R. Feix, Phys. Lett. A **94**, 415 (1983).
[36] K. Wódkiewicz, Phys. Rev. Lett. **52**, 1064 (1984).
[37] S. Basu, Phys. Lett. A **114**, 303 (1986).
[38] N. V. Sidorkov, Izv. Vusov Fiz. **4**, 107 (1983).
[39] P. R. Holland, A. Kyprianidis, and J. P. Vigier, Physica A **139**, 619 (1986).
[40] Yu. I. Zaparovanny, V. V. Kuryshkin, and I. A. Lyabis, Sov. J. Phys. **21**, 336 (1978).
[41] S. P. Misra and T. S. Shankara, J. Math. Phys. **9**, 299 (1968).
[42] G. S. Agarwal and E. Wolf, Phys. Rev. D **2**, 2161 (1970).
[43] G. J. Ruggeri, Prog. Theor. Phys. **46**, 1703 (1971).
[44] M. D. Srinivas and E. Wolf, Phys. Rev. D **11**, 1477 (1975).
[45] V. V. Kuryshkin, I. A. Lyabis, and Yu. I. Zaparovanny, Ann. Ford. L. de Broglie **5**, 105 (1980).
[46] A. Messiah, *Quantum Mechanics* (North-Holland, Amsterdam, 1962), Vol. 2.
[47] N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Interscience, New York, 1959).
[48] S. Basu, Sov. J. Phys. **25**, 956 (1982).
[49] S. Basu and I. A. Lyabis, Proc. Ind. Nat. Sci. Acad. A **49**, 509 (1983).
[50] S. Basu and I. A. Lyabis, Ann. Fond. L. de Broglie **8**, 271 (1983).
[51] K. E. Cahill and R. J. Glauber, Phys. Rev. **177**, 1857 (1969).
[52] A. Messiah, *Quantum Mechanics* (North-Holland, Amsterdam, 1961), Vol. 1.
[53] S. Basu, Ph.D. thesis, Patrice Lumumba Peoples' Friendship University, Moscow, 1984 (unpublished).

The causal geometry of twistor space

Robert J. Low
Mathematics Department, Coventry Polytechnic, Priory Street, Coventry CV1 5FB, England

(Received 18 August 1989; accepted for publication 15 November 1989)

The problem of classifying the nature of the vector connecting a pair of points in Minkowski space is examined within the twistor theoretic framework. Two approaches are considered, one algebraic and the other geometric. The latter of the two is studied in some detail, providing some insight into the relation between the causal structure of Minkowski space and the geometry of projective twistor space.

I. INTRODUCTION

The question of whether a pair of points in Minkowski space are separated by a timelike interval or a spacelike (or null) one is of fundamental importance; its answer determines whether fields at one point can be affected by data at the other. In Minkowski space with the standard coordinates the answer is simple to find, for two points x and y with position vectors x^a and y^a are connected by a timelike vector if $\|x^a - y^a\|^2$ is positive, and so on. The point whose t coordinate is greater is the one to the future of the other.

However, if one takes the point of view that twistor space is fundamental and space-time is a derived structure, then the problem becomes rather murkier. Hitherto, the study of causal relations via twistor theory has received little attention. Presumably, this is at least partly because from the twistor point of view it is complexified compactified Minkowski space that arises naturally, and in this case the notion of a causal structure is not naturally well defined, although that of a conformal one is. Below, in Sec. II, we will see how the nature of the causal separation of a point from the origin of Minkowski space may be ascertained using twistor algebra. Although this algebraic approach may be extended to arbitrary pairs of points, it also suggests a geometric approach to the problem that extends more naturally, and that is examined in Sec. III. In Sec. IV we consider briefly the points at infinity in compactified Minkowski space, and finally in Sec. V the point of view is applied to the consideration of worldlines in M.

The terminology of Hawking and Ellis[1] will be used regarding causality theory; for an introduction to twistor theory, see the texts of Huggett and Tod[2] or Penrose and Rindler.[3] We will follow these references for "twistorial" notation and terminology, but for convenience a list of appropriate notation is appended:

M is Minkowski space,

$M^\#$ is compactified Minkowski space,

$\tilde{M}^\#$ is compactified but not identified Minkowski space,

I is the null cone at infinity in $M^\#$,

CM is complexified Minkowski space,

$CM^\#$ is complexified compactified Minkowski space,

CI is the null cone at infinity in $CM^\#$,

T is twistor space, i.e., $\mathbb{C}^4\setminus\{0\}$, with coordinates $(\omega^A, \pi_{A'})$ and the Hermitian form Φ defined by $\Phi(\omega^A, \pi_{A'}) = \omega^A\bar{\pi}_A + \bar{\omega}^{A'}\pi_{A'}$.

PT is projective twistor space, with homogeneous coordinates $Z^\alpha = (\omega^A, \pi_{A'})$,

PN is projective real twistor space, given by $\Phi(Z^\alpha) = 0$,

If $\pi_{A'}$ is a spinor, then p_a is the corresponding null vector, so that $\pi_{A'}\bar{\pi}_A = p_a$ (using the abstract index convention).

The null geodesic in M through x with cotangent p_a is given in PN by $(ix^{AA'}\pi_{A'}, \pi_{A'})$, where $x^{AA'}$ are the spinorial coordinates of $x \in M$.

I, the twistor line corresponding to I is given by $\{(\omega^A, 0)\colon \omega^A \in \mathbb{C}^2\{0\}\}$,

PN^{I} and PT^{I} are $PN\setminus\mathrm{I}$ and $PT\setminus\mathrm{I}$, respectively.

Finally, we will use the convention that if $x \in CM^\#$, then $X \subset PT$ is its sky, and if $\gamma \in PT$, then $\Gamma \subset CM^\#$ is the corresponding α plane; furthermore, if $\gamma \in PN$, then $\Gamma \subset M^\#$ is the corresponding null geodesic. One exception to this convention will be the use of L_0 to represent the sky of the origin of Minkowski space.

II. TWISTOR ALGEBRA AND CAUSAL RELATIONS

In order to study the problem of classifying the causal nature of the separation of two points, it is convenient to begin with the problem of deciding whether a point in M is separated from the origin of M by a timelike, null, or spacelike interval, and, if the first, whether it is to the past or future of the origin. So consider the point $x \in M$, x not the origin, with position vector x^a. We want some way of attacking the problem using twistors, so recall that a nonprojective twistor $(\omega^A, \pi_{A'})$ is a future-pointing null geodesic with the associated tangent spinor $\pi_{A'}$ such that $p_a = \pi_{A'}\bar{\pi}_A$ is the tangent vector to it. The projective twistor forgets the actual spinor $\pi_{A'}$ remembering it—and hence p_a—only up to scale.

But now an elementary result from the geometry of Minkowski space tells us that

$x \in I^+(0)$ iff $x^a p_a > 0$
for all future pointing (fp) null p_a,

$x \in J^+(0)$ iff $x^a p_a \geqslant 0$ for all fp null p_a,

$x \in J^+(0) \setminus I^+(0)$ iff $x \in J^+(0)$ and $x^a p_a = 0$
for some p_a unique up to scale,

0022-2488/90/040863-05$03.00

$x\in M\setminus J(0)$ iff $x^a p_a$ takes on positive and negative values as $\mathrm{p_a}$ varies over fp null vectors,

with similar statements for I^- and J^-, but with $>$ replaced by $<$, etc.

Now recall that if $Z^\alpha = (ix^{AA'}\pi_{A'}, \pi_{A'})$, then $Z^\alpha \bar{Z}_\alpha = -2\,\mathrm{Im}(x^a p_a)$, and when $x\in M$, this is automatically zero. However, the expression does contain $x^a p_a$, which motivates the following idea. Given $Z^\alpha = (\omega^A, \pi_{A'})$, define the twistor $W^\alpha(Z^\alpha) = (-i\omega^A, \pi_{A'})$; then we obtain $W^\alpha(Z^\alpha)\bar{W}_\alpha(Z^\alpha) = 2\,\mathrm{Re}(x^a p_a)$, and the separation of a point from the origin can be classified using this.

One possible means of proceeding is to define a new inner product Φ on T by $\Phi_0(Z^\alpha) = \Phi(W^\alpha(Z^\alpha))$. Then in just the same way as Φ splits up T into T^+, T^-, and N, we can use Φ_0 to split it up as follows:

$$T^f = \{Z^\alpha\in T\colon\ \Phi_0(Z^\alpha) > 0\},$$
$$T^s = \{Z^\alpha\in T\colon\ \Phi_0(Z^\alpha) = 0\},$$
$$T^p = \{Z^\alpha\in T\colon\ \Phi_0(Z^\alpha) < 0\},$$

and since the sign of $\Phi_0(Z^\alpha)$ (although not the value) projects down to PT, we obtain PT^f, PT^s, and PT^p in just the same way as PT^+, etc. Again, just as before, PT^s is the common boundary of the two other regions, so that $PT^f\cup PT^s = \overline{PT}^f$, and so on. Finally, intersecting with PN defines the regions PN^f, PN^s, and PN^p. This gives the following classification.

Proposition 2.1: If $x\in M$ is not the origin, then

$x\in I^+(0)$ iff $X\subset PN^f$,

$x\in J^+(0)$ iff $X\subset \overline{PN}^f$,

$x\in M\setminus J(0)$ iff X intersects PN^f and PN^p,

and similarly for I^- and J^-. □

One now faces the natural question of how the families PN^f, PN^s, and PN^p are mirrored in Minkowski space. In fact, the null geodesics of M are split up into three classes by this partition, the class a given geodesic lies in depending on its relationship with $N(0)$, the null cone of the origin.

Proposition 2.2: PN^f consists of those null geodesics in M which are spacelike separated from 0 for all sufficiently large negative values of t, eventually enter $I^+(0)$, and remain there thereafter. PN^s consists of those null geodesics whose intersection with $I(0)$ is empty (together with the null cone at infinity in $M^\#$, and PN^p consists of the null geodesics in $I^-(0)$ for t sufficiently large which eventually leave $I(0)$ and never enter again.

Proof: Any null geodesic Γ in M is described by a pair (x^a, p^a) where $x^a\in\{(x,y,z,0)\in M\}$, and is parametrically given by $\Gamma = \{x^a + tp^a\colon t\in\mathbb{R}\}$. The result follows immediately from the facts that $\|x^a + tp^a\|^2 = \|x^a\|^2 + 2tx^a p_a$, and that $\gamma\in PN^f$ if and only if $x^a p_a \geqslant 0$, etc. □

An alternative point of view is that PN^s is characterized by the fact that $Z^\alpha\in PN$ lies in PN^s if for $W^\alpha\in L_0$ we have $Z^\alpha \bar{W}_\alpha \neq 0$ except for a single $W^\alpha\in L_0$, and for that this one, $Z^{[\alpha}W^{\beta]}$ intersects I. This point of view then allows us to build the rest of PT^s by allowing Z^α to lie in PT, and finally we can extend the construction to include points of M other than the origin in exactly the same way. This has the advantage that it is easy to extend to points other than the origin, whereas the construction of the form corresponding to Φ is rather awkward.

On the other hand, the mapping $Z^\alpha\to W^\alpha(Z^\alpha)$ given by $(\omega^A, \pi_{A'})\to(-i\omega^A, \pi_{A'})$ can be regarded as a rotation—in a certain sense—and this approach also motivates a natural extension to points other than the origin.

III. TWISTOR GEOMETRY AND CAUSAL RELATIONS

So consider the mapping acting on PT by $(\omega^A, \pi_{A'})\to(i\omega^A, \pi_{A'})$ where $(\omega^A, \pi_{A'})$ are the usual homogeneous coordinates on PT. This is an operation of order four, and can be written as $(\omega^A, \pi_{A'})\to(e^{i\pi/2}\omega^A, \pi_{A'})$, prompting the consideration of $(\omega^A, \pi_{A'})\to(e^{i\theta}\omega^A, \pi_{A'})$ for $\theta\in[0, 2\pi)$. In other words, there is an action of U(1) on PT containing this mapping. The question is, how does one specify this particular representation of U(1) amongst all the representations of U(1) with PT for a representation space?

If $\rho(\theta)\colon PT\to PT$ is given by $(\omega^A, \pi_{A'})\to(e^{i\theta}\omega^A, \pi_{A'})$ then we can observe that $\rho(\theta)$ fixes both L_0 and I pointwise; moreover, regarding the action on T given by this action on the homogeneous coordinates, we observe that I is fixed pointwise in T as well, even although L_0 is rotated about inside itself in such a way as to preserve all the complex lines in L_0 through the origin. In this sense, $\rho(\theta)$ is a rotation in PT induced by a rotation of PT which fixes I pointwise and L_0 projectively. In fact, it is (almost) the unique such representation of U(1) on PT which is free and preserves L_0 and I in the way described just above.

Let

$$\rho(\theta) = \begin{bmatrix} A(\theta) & B(\theta) \\ C(\theta) & D(\theta) \end{bmatrix}.$$

Since $\rho(\theta)$ preserves L_0 and **I** up to proportionality, $B = C = 0$; since I is fixed pointwise in T, $D(\theta) = \mathbf{I}$; and since L_0 is fixed in PT, $A(\theta) = \alpha(\theta)\mathbf{I}$. Finally, by freeness, we have $\alpha(\theta) = e^{\pm i\theta}$. The choice of $+i\theta$ or $-i\theta$ is just the choice of which timelike direction to call future, and which to call past, and we make the choice of $+i\theta$.

Considering the action of $\rho(\theta)$ on PT, we note that if z^a lies in the future tube, i.e., $z^a = x^a - iy^a$ with y^a timelike and future pointing, then $\rho(\pi/2)(Z)$ is the sky of a point whose real part is timelike and future pointing. Equivalently, $\rho(\pi/2)\colon PT^+\to PT^f$, and also maps PN to PT^s, and PT^- to PT^p.

Now, let $y\in M$. Then there are only two representations of U(1) on PT fixing Y and I in the same way as $\rho(\theta)$ fixes L_0 and I, and only one of these is compatible with $\rho(\theta)$. Call this representation $\rho_Y(\theta)$. It is now clear that under the action of $\rho_Y(\theta)$, PT^+ is sent to the set of twistors $PT_Y{}^f$ such that if $X\subset PT_Y{}^f$ then $\mathrm{Re}(x)\in I^+(y)$ for if $x = \xi^a - i\eta^a$ with η^a timelike and future pointing, then the action of $\rho_Y(\pi/2)$ gives

$$x^a \to x^a - y^a \to \eta^a + i(\xi^a - y^a) \to \eta^a + y^a + i(\xi^a - y^a),$$

and since η^a is timelike and future pointing, it follows that $y^a + \eta^a$ lies inside $\mathbf{I}^+(y)$.

We are now almost in a position in which we can start to consider the relationship between the geometry of twistor space and the causal geometry of Minkowski space in a little more detail. First, though, we make one more observation. If $y \in M$, then $PT_Y{}^f$ consists of those twistors $(iz^{AA'}\pi_{A'}, \pi_{A'})$ such that $\mathrm{Re}(z^a - y^a)p_a > 0$ for all future pointing null p^a. We thus obtain

Lemma 3.1: Let $x, y \in M$. Then

$$x < y \quad \text{iff } PT_Y{}^f \subseteq PT_X{}^f \quad \text{iff } PT_Y{}^p \supseteq PT_X{}^p.$$

□

And since $x \boldsymbol{<} y$ iff $x < y$ and $X \cap Y = \varnothing$, twistorial proofs of such statements as

$$x < y \quad \text{and} \quad y < z \quad \text{implies that } x < z,$$

etc. become straightforward. For example, we have the following.

Proposition 3.1: Let x, y, and z be points of M such that $x < y$ and $y \boldsymbol{<} z$. Then $x \boldsymbol{<} z$.

Proof: The twistorial future of Y lies inside that of X, and Z lies inside the twistorial future of Y, therefore it also lies inside that of X. □

Lemma 3.2: Let $x, y \in M$. Then one of the following situations must hold:

(1) X lies in the twistorial future of Y,

(2) all but one point of X lies in the twistorial future of Y, and that point lies on Y,

(3) X intersects the twistorial future and the twistorial past of Y, and also intersects $PN_Y{}^s$ in a one parameter family,

(4) a situation analogous to (1) or (2) but with past replacing future.

Proof: This follows immediately from the corresponding analysis of the way that skies in PT lie relative to PN.[4] □

The set $PN_X{}^f$ will be called the twistorial future of X, and its closure, the twistorial causal future. Twistorial pasts are defined similarly. Note that all this is only defined here for skies corresponding to points of real Minkowski space. The points at infinity form the topic of the next section.

IV. POINTS AT INFINITY

Any sky in PN^{I} is unambiguously associated with a twistorial future and past set, obtained as described above. However, the skies of points on I, and the points i^+, i^-, and i^0 in compactified (but not identified) Minkowski space now appear naturally, although in the usual way of considering twistor space it is compactified and identified Minkowski space that occurs.

For if we consider the sky of some point at infinity, there is no unique way to assign a twistorial future or past to this sky. The rotation prescription breaks down and no longer specifies the twistorial future as it did for the sky of a point in M. This can be seen quite clearly from the fact that a point on the light cone at infinity in compactified identified Minkowski space corresponds to a pair of points in $\widetilde{M}^{\#}$, and these points have distinct future and past sets. More severely, in the case of the point i, the axes of rotation degenerate to a single sky, which clearly leaves the rotation undefined.

We avoid this problem by considering the route by which the point at infinity is approached. Regarding these points as the endpoints in compactified Minkowski space of curves with no endpoint in Minkowski space, we can associate twistorial futures in the following way.

Let z be a point in $\widetilde{M}^{\#}$, and let c be a curve in M with endpoint z. Define the twistorial future of Z approached in this way to be the limit of the future sets of the points on c, and similarly the twistorial past.

The net effect of this is that to each sky in PN that intersects I, we will obtain two possible twistorial future and past sets: for one, the twistorial past set will be empty (corresponding to the point on I^-), and for the other the twistorial future set will be empty (the point on I^+). Also, for I itself we will obtain three possibilities, namely an empty future set and a past set consisting of all of PN^{I}, corresponding to i^+, an empty past set and a future set consisting of all of PN^{I}, corresponding to i^-, and both the past and future sets being empty, corresponding to i^0.

This enables us to rebuild from PN the whole causal structure of compactified, but not identified, Minkowski space up from these subsets of twistor space, by using the characterization of causal ordering obtained in the previous section.

V. WORLDLINES

There are three classes of curves in M from the physical point of view, namely timelike curves, null curves, and others. These others may have spacelike sections or always be causal, but are in this latter case timelike at some points and null at others—in either case can they correspond to the worldlines of material particles. The timelike curves correspond to massive particles, and the null ones to zero rest mass particles. One can also classify those candidates for worldlines as being geodesics or not—i.e., corresponding to free particles or particles acted upon by some force. Here we will only be concerned with the case of curves that are everywhere timelike or everywhere null, as a single particle cannot change from massive to massless or vice versa. For the sake of brevity, all such curves will be understood to be future pointing unless the contrary is explicitly stated.

Now, a curve in M gives us a projective ruled surface (see Semple and Kneebone[5] for definitions of most of the projective geometry terms used here) in PN, which is ruled by the skies of the points on the curve in M. A causal curve is characterized by the fact that the corresponding ruled surface is such that each generator lies to the twistorial causal future of all previous ones, and a timelike curve by the fact that each generator lies in the twistorial future of all previous ones. Although the distinction between causal and timelike curves can be made in twistorial terms, that between (non-null) geodesic and nongeodesic worldlines cannot, for the simple reason the structure of twistor space depends only on the conformal structure of M, and different choices of conformal factor correspond to different choice on timelike geodesic.

A null geodesic is characterized by the fact that there is a common point on each of the generators of the ruled surface—so in this case it is a pencil, whose vertex is the point of PN corresponding to the null geodesic. A null curve is infinitesimally a null geodesic, with the consequence that infinitesimally separated generators of the corresponding rules surface will intersect. In other words, a null curve in M corresponds to a developable in PN. It is also clear that any developable corresponds to a null curve.

To summarize the above, then, curves in M correspond to ruled surfaces in PN. A pencil gives a null geodesic, a developable gives a null curve, and a scroll (i.e., any ruled surface which is not a developable[6]) that moves into the twistorial future corresponds to a timelike curve. (In fact, since two points are timelike separated in M if and only if there is a null curve that is not a geodesic joining them, we see that two points are timelike separated if and only if there is a nonsingular developable in PN^1 whose boundary consists of their skies.) All other ruled surfaces either change character by being a scroll at some times and a developable others, or are acausal, and in neither case do they correspond to physically meaningful curves in M. In fact, it is easy to see that scrolls with self-intersections (other than pencils) correspond to curves with spacelike sections, since they must correspond to curves in M with null-separated points.

We can also use these ideas to give intrinsically twistorial proofs of results to do with the causal geometry of Minkowski space.

Proposition 5.1: Let c be a future and past endless timelike curve in M, which approaches i^- to the past and i^+ to the future, and let x be a point not on c. Then the null cone of x intersects c in precisely two points.

Proof: Regard c as a map from $\mathbf{R} \rightarrow M$, with $c(t)$ approaching $i^{\pm}$ as t approaches $\pm\infty$. Then $C(t)$ starts off at I and ends up at I, and for all values of t in $(-\infty, \infty)$ is travelling into the twistorial future. Initially, $C(t)$ lies in the twistorial past of X, and eventually it ends up in its twistorial future. We must therefore consider the process by which it gets from one to the other.

As t increases, it takes on some value, say t_0, at which some element of $C(t_0)$ first intersects $PN_X{}^s$—this intersection must be a singleton, and hence lies in X. Also, eventually a value t_1 is reached after which no $C(t)$ intersects $PN_X{}^s$, and $C(t_1)$ intersects it in a singleton, which again must lie in X. In between, $C(t)$ always intersects $PN_X{}^s$ in a one-parameter family, for if it did not, then there must be some value of t between t_0 and t_1 for which $C(t)$ either lies in the twistorial future or the twistorial past of X. It cannot lie in the twistorial past, since $C(t)$ is always moving into the twistorial future; likewise, it cannot lie in the twistorial future, or else we must have $t > t_0$.

It then follows that for $t < t_0$, $C(t)$ lies in the twistorial past of X, for $t = t_0$, $C(t)$ intersects X, for $t_0 < t < t_1$ $C(t)$ intersects $PN_X{}^s$ in a one-parameter family, for $t = t_1$ $C(t)$ intersects X, and for $t > t_1$ $C(t)$ lies in the twistorial future of X.

Translated into Minkowski space terms, this means that the timelike curve c starts off in the past of x, crosses its light cone, spends some time spacelike separated from it, then crosses the light cone one more time, after which it remains in the future of x. □

Proposition 5.2: Let c be a causal curve in M such that x and z lie on c, but $x \notin \mathbf{I}^-(z)$. Then the curve is a segment of a null geodesic between x and z.

Proof: Let y be a point of c between x and z. Then the pairs x and y, and y and z must be null separated, else $z \notin \mathbf{I}^+(x)$. Now consider the skies X, Y, Z. We know that each pair must intersect. Now, Y lies in the twistorial future of X except for the point at which they intersect. Furthermore, Z lies in that of Y, except for the point of intersection. But the twistorial future of Y contains no points of X, and since Z intersects X, we see that the points of intersection must coincide, hence x, y, and z lie on a common null geodesic. This holds for every point y between x and z on c, and so the result follows. □

VI. CONCLUSIONS

We have observed that the causal structure of Minkowski space has a reasonably straightforward interpretation in terms of twistor geometry, from (at least) two different points of view. The above notions allow us to give a twistor interpretation of compactified but not identified Minkowski space, and to discuss the causal structure of M—and hence any portion of a space-time conformal to a subset of M—in twistorial terms.

However, as the linear structure of PT was involved in this in an essential way, it seems unlikely that the techniques will extend to curved space-time in any straightforward way. The study of the causal structure of a general space-time in terms of its space of null geodesics is a much more difficult problem, although partial answers can be found, for example if one restricts attention to causally simple regions, or considers only a restricted class of space-times.[7]

ACKNOWLEDGMENTS

I am much obliged to Professor R. Penrose of the Mathematical Institute, Oxford, and Dr. T. N. Bailey, then of the Mathematical Institute, Oxford, for helpful discussions.

Part of this work was done with the aid of an SERC Research Studentship.

[1] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge U. P., Cambridge, 1973).

[2] S. A. Hugget and K. P. Tod, *An Introduction to Twistor Theory* (Cambridge U. P., Cambridge, 1985).

[3] R. Penrose and W. Rindler, *Spinors and Space-time* (Cambridge U. P., Cambridge, 1986), Vol. 2.

[4] R. R. Moore, M.S. thesis, Oxford University, 1978.

[5] J. G. Semple and G. T. Kneebone, *Algebraic Projective Geometry* (Clarendon, Oxford, 1952).

[6] J. G. Semple and L. Roth, *Introduction to Algebraic Geometry* (Clarendon, Oxford, 1949).

[7] R. J. Low, Ph.D. thesis, Oxford University, 1988.

An example of a nontrivial causally simple space-time having interesting consequences for boundary constructions

Petra Rübe
Technische Universität Berlin, Fachbereich 3—Mathematik, Sekr. MA 8-3, Strasse des 17 Juni 136, D-1000 Berlin 12, West Germany

(Received 5 July 1989; accepted for publication 30 November 1989)

An example is given of a causally simple space-time that may serve as a counterexample for various purposes, such as showing that for a general causally simple space-time the chronological common past of a terminal indecomposable future set need not be an indecomposable set.

I. INTRODUCTION

The following "multipurpose counterexample" is one result of a more thorough investigation into some technical aspects of different kinds of causal boundary constructions for general relativistic space-times.

One of the most convincing candidates for such constructions is the well-known procedure indicated by Budic and Sachs.[1] However, it only yields reasonable results for space-times that are at least causally continuous. Thus an important part of our investigation had been restricted to causally continuous, causally simple, and globally hyperbolic space-times.

In this context one often has to do with propositions that are trivially true for globally hyperbolic space-times and "trivially false" for space-times that are only causally continuous, but not causally simple (in that case there is most often rather a primitive counterexample obtained from some low-dimensional Minkowski space in the usual cutting-and-gluing way). However, it is not always easy to decide if the proposition in question holds for arbitrary causally simple space-times. (Although our experience has shown that in most such cases there is either an easy proof or not a quite so easy-to-find counterexample, still this example remains to be found—and there *might* once be an exception to our empiric "rule.")

Thus the problem that served as a starting point for this part of our investigation was the following: Given a terminal indecomposable future set F in a space-time M, under which conditions it is necessarily true that the chronological common past $\downarrow F$ of F is also an *indecomposable* set, if it is not empty? The analog for *proper* indecomposable sets is generally true for reflecting space-times [Ref. 2, p. 290, Prop. 1.3]. The proposition is also trivially true for *terminal* sets in *globally hyperbolic* space-times, since in that case necessarily $\downarrow F=\phi$. There is, however, an example of a causally continuous (but not causally simple) open submanifold of three-dimensional Minkowski space that contains a terminal indecomposable future set F, such that $\downarrow F$ is neither indecomposable nor empty [Ref. 3, p. 49 f., (1.3.9)].

Knowing that the above proposition were also true in the causally simple case would provide us with quite a strong practical tool for examining causal boundaries of such spaces. In fact, it can be shown that the proposition holds for a class of causally simply space-times that includes open submanifolds of Minkowski space [Ref. 3, p. 51 ff., (1.3.11)]. Our example shows, however, that the proposition fails for causally simple space-times in general.

II. DEFINITIONS AND NOTATIONS

We use the term *space-time* in the widest sense possible in this context, defining it simply as a connected, time-oriented Lorentz manifold of dimension $\geqslant 2$. Orientability may also be demanded, but will not be needed. Our example will, of course, be orientable.

Here, $\mathbb{R}_1^3$ shall denote three-dimensional Minkowski space, i.e., $\mathbb{R}_1^3$ is $\mathbb{R}^3$ with the standard Minkowski metric $dx\otimes dx+dy\otimes dy-dt\otimes dt$ (x, y, and t are Euclidean coordinate functions of $\mathbb{R}^3$).

As usual, I^+, I^-, J^+, J^- shall denote chronological (resp. causal) futures (resp. pasts) in a given space-time.

Consider, on a time-oriented manifold M, the set of all Lorentzian metrics that are compatible with the time orientation of M (i.e., the vector field that defines the time orientation shall be timelike with respect to all of these Lorentz metrics). On this set an ordering relation $\leqslant$ may be defined by $g_1\leqslant g_2$ if the causal relations induced by g_1 are contained in those induced by g_2.

A subset of space-time is called *indecomposable* if it is either the past or the future of some causal curve in M. (See Ref. 4, p. 547 ff. for equivalent descriptions.) An indecomposable set is called a *terminal* indecomposable set if it is neither the past nor the future of a single point.

The *chronological common* past $\downarrow U$ of an open subset U of a space-time M is defined as

$$\downarrow U:\ =I^-\Big(\bigcap_{x\in U} I^-(x)\Big).$$

The *chronological common future* $\downarrow U$ is defined dually (cf. Ref. 1, p. 1303).

A space-time M is called *causally continuous* if M is distinguishing and for all $x\in M$ $I^-(x) = \downarrow I^+(x)$ and $I^+(x) = \uparrow I^-(x)$. There are many equivalent descriptions of this important causality condition (cf. e.g., Ref. 2).

Here, M is called *globally hyperbolic* if M is strongly causal and for all $x, y\in M$ $J^+(x)\cap J^-(y)$ is a compact set. There are as well useful equivalent definitions, but they have not been needed for this work.

M is called *causally simple* if M is distinguishing and for all $x\in M$ $J^+(x)$ and $J^-(y)$ are closed sets. Unfortunately, no equivalent characterization is known to us that would illuminate the significance of this "intermediate" causality condition from some other aspect (e.g.,

0022-2488/90/040868-03$03.00

time-functions). In particular, there seems to be no "elegant" way of checking causal simplicity of nontrivial examples.

It is well known that global hyperbolicity implies causal simplicity which, in turn, implies causal continuity, but that no two of these conditions are equivalent.

III. THE EXAMPLE

We now give a precise description of our "prototype." For a detailed proof that it does indeed possess the desired properties see Ref. 3, p. 55 ff., (1.3.12).

Here, M, considered as a point set, is roughly speaking the first quadrant of an open full cylinder in three-dimensional Minkowski space, the axis of the cylinder coinciding with the time axis of $\mathbb{R}_1^3$ and the upper part of the plane parts of the (topological) boundary of the cylinder being rounded off for technical reasons.

More precisely choose some $a\in]2,\infty[$ and define

$$M:=\{(x,y,z)\in\mathbb{R}_1^3 \,|\, x>0,\ y>0,\ x^2+y^2<1,$$
$$-1<t<\min\{a\sqrt{x},\ a\sqrt{y}\}\}.$$

[The "queer" upper bound for t is necessary for M to serve as a counterexample in (2) of Sec. IV; it is not necessary for the applications (1) and (3): In these cases $-1<t<1$ would also do.]

We now define on M two Lorentzian metrics $\hat{g}$ and $\hat{\hat{g}}$ in the following way: $\hat{g}$ is the restriction to M of the standard Minkowski metric of $\mathbb{R}_1^3$ and $\hat{\hat{g}}$ is the pullback of the standard Minkowski metric of $\mathbb{R}_1^3$ via

$$f\colon\ M\to\mathbb{R}_1^3,\quad (x,y,t)\longmapsto(\sqrt{x^2+y^2},\ \arctan(y/x),t).$$

The causal structure of $(M,\hat{g})$ is well known, and as $f{:}(M,\hat{\hat{g}})\to f(M)\subsetneq\mathbb{R}_1^3$ is an isometry, it is also easy to calculate the causal structure of $(M,\hat{\hat{g}})$. In fact, the causal cones of $(M,\hat{\hat{g}})$ are contained in the corresponding causal cones of $(M,\hat{g})$, the decisive point for our purpose being that the converse is not true.

The idea now is to provide M with a C^∞-Lorentz metric g that coincides with $\hat{g}$ in the "upper part" U of M and with $\hat{\hat{g}}$ in the "lower part" W, these two areas being separated by an arbitrarily small "area of transition" V.

More precisely choose $\epsilon\in]0,\frac{1}{4}[$ arbitrarily small and set

$$U:=\{(x,y,z)\in M\,|\,t\geqslant 0\},$$

$$V:=\{(x,y,z)\in M\,|\,-\epsilon\cdot x\cdot y\cdot(1-\sqrt{x^2+y^2})<t<0\},$$

$$W:=\{(x,y,z)\in M\,|\,t\leqslant-\epsilon\cdot x\cdot y\cdot(1-\sqrt{x^2+y^2})\}.$$

It is easy to indicate explicitly a C^∞ function $\eta\colon\ M\to[0,1]$ with $\eta|_U=0$ and $\eta|_W=1$. $g:=\eta\cdot\hat{\hat{g}}+(1-\eta)\cdot\hat{g}$ is then a Lorentz metric on M with $g|_U=\hat{g}$ and $g|_W=\hat{\hat{g}}$. (M,g) is the desired example.

It is easy to verify that $F:=\{\ (x,y,z)\epsilon M\,|\,t>0,\ x^2+y^2<t^2\}$ is a terminal indecomposable set on (M,g) (since it is the chronological future of the causal curve $\gamma{:}]0,b[\to M,\ \tau\mapsto(\tau,\tau,2\tau)$, with $b\in\mathbb{R}^+$ sufficiently small) and that $\downarrow F=\{(x,y,z)\in M\,|\,t<0,x^2+y^2<t^2\}$ is *not* an indecomposable set of (M,g).

The proof that (M,g) is, in fact, causally simple is rather lengthy, since the only criterion available is the one given in the definition (see Sec. II) and one has to check explicitly that $p\in\overline{J^\pm(q)}$ implies $p\in J^\pm(q)$ for all combinations of $p/q\in U/V/W$, respectively, using that $\hat{g}$ (resp. $\hat{\hat{g}}$) is an upper (resp. lower) bound for g (with respect to the ordering relation indicated in Sec. II) as well as the (evident) causal simplicity of $(U,\hat{g}|_U)$ and $(W,\hat{\hat{g}}|_W)$.

It has to be noted, however, that causal simplicity gets lost as soon as we modify M by "widening up" the angle at the t axis, defining for M something like

$$M:=\{(\rho\cdot\cos(\vartheta),\rho\cdot\sin(\vartheta),t)\,|\,\rho\in]0,1[,\vartheta\in]0,\Theta[,$$
$$t\in]-1,\varphi(\rho,\vartheta)[\},$$

with $\Theta>\pi/2$ and φ some appropriate real-valued function with $\varphi(\rho,\vartheta)>\rho$.

Such modifications preserve causal simplicity if $\Theta\in]0,\pi/2]$ and φ such that M remains a convex set.

IV. APPLICATIONS

In our work on causal boundaries the example just described has been used on three occasions.

(1) The first application concerning chronological common pasts of indecomposable terminal future sets has already been described in the Introduction and in Sec. III of this paper.

The other applications are of a more technical nature and concern comparison between different types of causal boundary constructions.

(2) (M,g) as constructed in Sec. III illustrates the fact that *even for causally simple space-times* the Geroch–Kronheimer–Penrose (GKP) identifying relation R_H [Ref. 4, p. 563] need not be contained in the Budic–Sachs (BS) "hull-pair" identifying relation [Ref. 1, p. 1303]. (For globally hyperbolic space-times these two relations are, once again, trivially identical.)

Indeed, taking $P:=\mathrm{I}^-(\gamma)$ for $\gamma{:}]0,1[\to\mathrm{M},\gamma{:}]0,1[\to\mathrm{M}$, $\tau\mapsto(\tau\cdot\cos(\vartheta_0),\ \tau\cdot\sin(\vartheta_0),\ -\tau)$ ($\vartheta_0\in]0,\pi/2[$ arbitrary, but fixed) and F as in Sec. III one sees that the elements of the "intermediate space" $M^\#$ [Ref. 4, p. 563] represented by P and F have to be identified in the GKP construction, although P and F are terminal sets which do *not* form a "hull-pair" [Ref. 3, p. 91 ff., (2.3.8)].

(3) Finally, for comparing different boundary constructions, one can consider them from the viewpoint of "natural mappings" existing between them: Given two prescriptions for boundary constructions A and B, under which circumstances can we be sure that for *any* space-time M there is a "natural mapping" from the "A-completion" to the "B-completion" of M? And if there is one, is it necessarily continuous?

Unfortunately, the answers to these questions are mostly negative. In general, small variations in the constructions lead to completely incompatible results. Thus, except for the globally hyperbolic case, there need not be any "natural mapping" between the standard GKP and BS constructions [Ref. 3, p. 110], and although in the globally hyperbolic case the existing "natural mapping" is (trivially) bijective, neither direction has to be continuous [Ref. 3, pp. 111–116].

One can, however, modify the GKP-construction maintaining the original procedure, but starting from a finer topology of $M^{\#}$ [Ref. 3, pp. 102–107]. There is always a natural mapping from this modified GKP completion to the BS completion, but the *example described above shows that it need not be continuous.* In fact, only a minor and, in every other respect harmless, modification in the definition of the Budic–Sachs "extended causality relations" [Ref. 3, p. 122 (3.3.4)] is needed to repair this defect [Ref. 3, p. 124 (3.3.8) and p. 128 (3.3.11)].

[1] R. Budic and R. K. Sachs, "Causal boundaries for general relativistic space-times," J. Math. Phys. **15**, 1302 (1974).

[2] S. W. Hawking and R. K. Sachs, "Causally continuous spacetimes," Commun. Math. Phys. **35**, 287 (1974).

[3] P. Rübe, "Kausale Randkonstruktionen für Raum-Zeiten der Allgemeinen Relativitätstheorie," Ph.D. thesis, Technische Univ., Berlin, 1988.

[4] R. Geroch, E. H. Kronheimer, and R. Penrose, "Ideal points in spacetime," Proc. R. Soc. London Ser. A **327**, 545 (1972).

Initial value problem for colliding gravitational plane waves. III

Isidore Hauser
Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13675
Frederick J. Ernst
Department of Mathematics and Computer Science and Department of Physics, Clarkson University, Potsdam, New York 13676

(Received 2 September 1989; accepted for publication 29 November 1989)

The development of a homogeneous Hilbert problem (HHP) approach to the initial value problem (IVP) for colliding gravitational plane waves with noncollinear polarization that began in two earlier papers [I. Hauser and F. J. Ernst, J. Math. Phys. **30**, 872 (1989) and **30**, 2322 (1989)] is continued. After formulating the HHP, the description of how one can apply it to generate a new family of solutions of the colliding wave problem that generalizes a three-parameter family constructed by Ernst, García, and Hauser [J. Math. Phys. **29**, 681 (1988)] using a double-Harrison transformation is given. Then the proof that the solution of the new HHP indeed solves the IVP that is posed is presented. A matrix Fredholm equation of the second kind that is equivalent to the HHP is also deduced. This will be used in a sequel to complete the proof of existence of solutions of the HHP and the proof that certain assumed differentiability hypotheses are in fact valid.

I. INTRODUCTION

In the first paper[1] of our series, we presented a new Abel transform method of solution of the initial value problem (IVP) for colliding gravitational plane waves, valid when the polarizations of the incident plane waves are collinear. Subsequently, in the second paper[2] of our series, we demonstrated that the Abel transform solution could be derived anew using either of two forms of the Hilbert problem and we indicated how one of these two Hilbert problems could be generalized to a matrix homogeneous Hilbert problem (HHP) when the polarizations are noncollinear.

We should like to emphasize that the new HHP is rather different from the matrix HHP that was employed by Ernst, García, and Hauser[3] (EGH) to derive a three-parameter generalization of the Ferrari–Ibañez–Bruni family of colliding wave solutions.[4] The earlier HHP was, aside from some relatively minor details, a direct translation of the HHP used in connection with the Geroch group of transformations of one stationary axisymmetric space-time into another.[5] On the other hand, the formulation of the new HHP was motivated by a desire to find an effective method of solving the IVP for colliding plane-fronted gravitational waves with noncollinear polarizations.

In this paper we shall present the details concerning the new matrix HHP and demonstrate one way that we have succeeded in employing it to construct a new family of colliding wave solutions with noncollinear polarizations. We shall also present a matrix Fredholm equation of the second kind which is equivalent to our new HHP and is especially useful in connection with various proofs.

In a separate paper authored with Li, the new family of solutions that we have obtained will be described in detail and in the next paper of the present series the Fredholm equation will be applied to establish existence of a solution of the new HHP, and prove other essential theorems.

It will be assumed that the reader is familiar with the general features of colliding gravitational wave solutions, as exemplified by the famous Nutku–Halil solution[6] and the more recent Chandrasekhar–Xanthopoulos solution.[7] We shall continue to describe such colliding wave solutions in terms of a chart (x^1,x^2,u,v) that was described in detail in Ref. 1 (Sec. II), where we also described the four regions into which the chart was divided by the null hypersurfaces $u=0$ and $v=0$, which separate the incident plane-wave regions II and III from the Minkowski space region I and the region IV in which the waves interact. See Fig. 1. We continue to employ the line element

$$ds^2=\hat{\rho}\hat{F}^{-1}|\hat{E}\,dx^1+i\,dx^2|^2-(2/\sqrt{\hat{\rho}})e^{2\hat{\Gamma}}\,du\,dv\,,\qquad(1.1)$$

where $\hat{F}:=\operatorname{Re}\hat{E}>0$, $\hat{\rho}>0$, and where the real fields $\hat{\rho}$, $\hat{\Gamma}$, and the complex Ernst potential $\hat{E}$ depend at most upon the null coordinates u,v. At the outset we shall select the Killing vectors so that at the two-surface of collision $u=v=0$ we have $\hat{\rho}(0,0)=1$, $\hat{E}(0,0)=1$ and we shall scale $du\,dv$ so that $\hat{\Gamma}(0,0)=0$. The caret symbol will be suppressed whenever we desire to consider the restriction of a function to the interaction region IV of the gravitational waves.

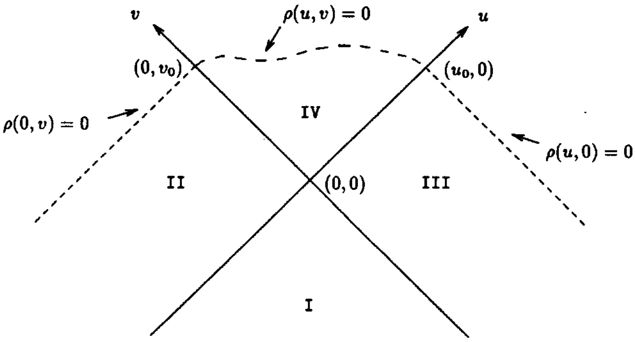


FIG. 1. Region I–IV for sample choices of $r(u)$ and $s(v)$ such that $r(u_0)=1$ and $s(v_0)=-1$.

0022-2488/90/040871-11$03.00

In region II the various fields depend only upon v, while in region III they depend only upon u: In region IV they depend upon both u and v. In particular, in region IV, we express ρ and its conjugate field z in the form

$$\rho(u,v) = \tfrac{1}{2}(s(v) - r(u)),$$
$$z(u,v) = \tfrac{1}{2}(s(v) + r(u)), \qquad (1.2)$$

where

$$r(u) := 1 - 2\rho(u,0), \quad s(v) := 2\rho(0,v) - 1. \qquad (1.3)$$

With our choice of $\rho(0,0)$, it follows that

$$r(0) = -1, \quad s(0) = 1. \qquad (1.4)$$

The formal definition of region IV is

$$\mathrm{IV} := \{(u,v) \in R^2 : 0 \leqslant u < u_0, 0 \leqslant v < v_0, r(u) < s(v)\}, \qquad (1.5)$$

where in this paper we shall assume that

$$r(u_0) = 1, \quad s(v_0) = -1, \qquad (1.6)$$

in other words, that

$$\rho(u_0,0) = \rho(0,v_0) = 0.$$

A slightly more general situation was considered in Ref. 1 (Sec. II).

As we demonstrated in Ref. 1 (Sec. II), the imposition of the vacuum field equations at the $u = 0$ and $v = 0$ hypersurfaces requires that

$$\dot r(0) = 0; \quad \dot s(0) = 0, \qquad (1.7)$$

while the "wave front conditions" require that

$$\dot r(u) > 0, \quad \text{for all } 0 < u < u_0, \qquad (1.8)$$
$$\dot s(v) < 0, \quad \text{for all } 0 < v < v_0. \qquad (1.9)$$

In our formulation of the IVP for colliding gravitational plane waves one regards $r(u)$ and $s(v)$ to be prescribed $\mathbf{C}^2$-differentiable functions and $E_3(u) := E(u,0)$ and $E_2(v) := E(0,v)$ to be prescribed $\mathbf{C}^1$-differentiable functions, from which $E(u,v)$ and $\Gamma(u,v)$ are to be determined throughout region IV. It is understood, of course, that $r(u)$ and $s(v)$ satisfy conditions (1.4) and (1.7)–(1.9) and that u_0 and v_0 are defined by Eqs. (1.6).

After $E(u,v)$ is determined, the field $\Gamma(u,v)$ is evaluated using Eqs. (2.23), (2.31), (2.40), and (2.41) in Ref. 1. Not every choice of $r(u)$, $s(v)$, $E_3(u)$, and $E_2(v)$ that satisfies Eqs. (1.4) and (1.7)–(1.9) is admissible. The reader should be familiar with the "colliding wave conditions" [Ref. 1, Eqs. (2.36)–(2.38)].

In the earlier papers we employed r and s to designate functions of u and v, respectively, and designate alternative null coordinates. We shall continue to depend upon context to distinguish one meaning of r, s from the other. The mapping $(u,v) \to (r(u),s(v))$ is one-to-one, bicontinuous, and maps IV as defined by Eq. (1.5) onto the set

$$D_{\mathrm{IV}} := \{(r,s) : -1 \leqslant r < 1, -1 < s \leqslant 1, r < s\}. \qquad (1.10)$$

The inverse of this mapping was denoted by Σ in Ref. 2.

Note: We shall often have occasion to express fields alternatively in terms of u, v or r, s. Wherever possible, we shall, beginning in this paper, use lightface type to signify the field expressed as a function of u,v and the corresponding boldface type to signify the field expressed as a function of r,s as, for example, in

$$\boldsymbol{\rho}(r(u),s(v)) := \rho(u,v), \quad \mathbf{z}(r(u),s(v)) := z(u,v).$$

In Sec. II of Ref. 2 we introduced a structure, also denoted by D_{IV}, which was not a $\mathbf{C}^2$ manifold, but to which were applicable many of the concepts and results of $\mathbf{C}^2$ manifold theory. In particular, we defined the duality operator $*$, the exterior differentiation operator d on zero- and one-forms, and the exterior differentiation operators $d^2 := d \wedge d$ and $d{*}d := d \wedge ({*}d)$ on zero-forms.

We also introduced the useful function

$$\chi(r,s,\tau) := \chi_3(r,\tau)\chi_2(s,\tau), \qquad (1.11)$$

where

$$\chi_3(r,\tau) := \left(\frac{\tau+1}{\tau-r}\right)^{1/2}, \quad \chi_2(s,\tau) := \left(\frac{\tau-1}{\tau-s}\right)^{1/2}; \qquad (1.12)$$

$\chi_3(-1,\tau) = \chi_2(1,\tau) = 1$; and where for fixed $r \neq -1$ and $s \neq 1$, we employ those holomorphic branches of $\chi_3(r,\tau)$ and $\chi_2(s,\tau)$ that have the cuts $[-1,r]$ and $[s,1]$, respectively, on the real axis of the complex τ plane and which satisfy $\chi_3(r,\infty) = \chi_2(s,\infty) = 1$. The domain of the function χ is

$$D := \{(r,s,\tau) : (r,s) \in D_{\mathrm{IV}}, \tau \in D_{(r,s)}\}, \qquad (1.13)$$

where

$$D_{(r,s)} := C - ([-1,r] \cup [s,1]) \qquad (1.14)$$

is the complex plane minus two cuts on the real axis. Finally, for any given $\tau \in C - \{-1,1\}$ we introduced the τ section of D, the set

$$D_\tau := \{(r,s) \in D_{\mathrm{IV}} : \tau \in D_{(r,s)}\}. \qquad (1.15)$$

The field χ and the various sets that we have recalled here are just as useful in connection with the noncollinear case as they were in connection with the collinear case treated in Ref. 2. The relation [Ref. 2, Eq. (2.27)]

$$(\tau - \mathbf{z} + \boldsymbol{\rho}{*})d\chi = d\mathbf{z}\,\chi \qquad (1.16)$$

will also be useful.

II. THE POTENTIALS H AND $A(\tau)$

In region IV the line element (1.1) may be expressed in the form

$$ds^2 = \rho \sum_{a,b=1}^{2} S_{ab}\, dx^a\, dx^b - \frac{2}{\sqrt{\rho}}\, e^{2\Gamma}\, du\, dv, \qquad (2.1)$$

where S denotes that 2×2 matrix function whose domain is IV and whose values are given by

$$S(u,v) := F(u,v)^{-1} \begin{pmatrix} E(u,v)E(u,v)^* & \omega(u,v) \\ \omega(u,v) & 1 \end{pmatrix}, \qquad (2.2)$$

where $F := \operatorname{Re} E$ and $\omega := \operatorname{Im} E$ In particular, we note that $\det S = 1$, $S(0,0) = I$. The corresponding matrix function $\mathbf{S}$ has domain D_{IV} and values given by

$$\mathbf{S}(r(u),s(v)) := S(u,v),$$

where

$$\det \mathbf{S} = 1, \quad \mathbf{S}(-1,1) = I.$$

Aside from the equation that determines Γ, the relevant vacuum field equations are

$$d*d\rho=0\,, \tag{2.3}$$

$$d(\rho\mathbf{S}\Omega*d\mathbf{S})=0\,, \tag{2.4}$$

where

$$\Omega:=\begin{pmatrix}0 & i\\ -i & 0\end{pmatrix}. \tag{2.5}$$

A. The Kinnersley H potential

A 2×2 matrix generalization H of the Ernst potential[8] $\mathscr{E}$ was introduced by Kinnersley[9] in connection with the stationary axisymmetric field problem. Because of the field equation (2.4), it is clear that one can always introduce a 2×2 matrix function $\mathbf{H}$, with domain IV and values determined by

$$d\mathbf{H}=d(-\rho\mathbf{S}+\mathbf{z}\Omega)-\rho\mathbf{S}\Omega*d\mathbf{S}\,, \tag{2.6}$$

where $\rho=\frac{1}{2}(s-r)$, $\mathbf{z}=\frac{1}{2}(s+r)$. Moreover, integration constants can be selected so that

$$\begin{aligned}&\operatorname{Re}\mathbf{H}=-\rho\mathbf{S},\quad \mathbf{H}-\mathbf{H}^T=2\mathbf{z}\Omega,\\ &\mathbf{H}(-1,1)=-I.\end{aligned} \tag{2.7}$$

B. Examples for the collinear case

In the collinear case considered in Ref. 2 one has $\mathbf{S}=e^{-2\sigma_3\psi}$, where σ_3 is the usual Pauli matrix and ψ is a real scalar field.[10] Equation (2.6) and conditions (2.7) immediately yield

$$\mathbf{H}^{\text{coll}}=-\rho e^{-2\psi\sigma_3}+\mathbf{z}\Omega+2i\delta\sigma_1\,,$$

where $\delta(r,s)$ is that integral of $d\delta=\rho*d\psi$ that satisfies $\delta(-1,1)=0$.

In particular, for the Kasner metric of index n, where $\psi=-(n/2)\ln\rho$, we have

$$\mathbf{H}^{\text{K}}=-\rho^{I+n\sigma_3}+\mathbf{z}\Omega-in\mathbf{z}\sigma_1\,.$$

C. The Yang–Mills type potential $A(\tau)$

Using Eq. (2.6) it is easy to show that $\mathbf{H}(r,s)$ satisfies the "self-duality relation"

$$\mathbf{S}\Omega*d\mathbf{H}=d\mathbf{H}\,, \tag{2.8}$$

which we derived by another method in an earlier paper,[5] where we also showed that for any given value of the complex parameter τ, the field

$$\mathbf{A}(\tau):=\tfrac{1}{2}(\tau-\mathbf{z}+\rho*)^{-1}\,d\mathbf{H}\,\Omega \tag{2.9}$$

is a Yang–Mills type of potential, i.e., it satisfies

$$d\mathbf{A}-\mathbf{A}\mathbf{A}=0\,. \tag{2.10}$$

We prefer to think of $\mathbf{A}(\tau)$ as one-parameter family of "connection" one-forms. It should be noted that the concept of $\mathbf{A}(\tau)$ is implicit in a paper[11] of Kinnersley and Chitre (KC).

III. THE SPECTRAL POTENTIAL $P(\tau)$

For the purposes of defining precisely what we mean by the P potential it is convenient to introduce the temporary notation $\mathbf{P}(\tau)$ as follows.

Definitions: For any given τ in $C-\{-1,1\}$, let $\mathbf{P}(\tau)$ be that 2×2 matrix function whose domain is D_τ [Eq. (1.15)] and which satisfies

$$d\mathbf{P}(\tau)=\mathbf{A}(\tau)\mathbf{P}(\tau)\,, \tag{3.1}$$

$$\mathbf{P}(\tau)(-1,1)=I\,. \tag{3.2}$$

Moreover, let $\mathbf{P}$ denote that function whose domain is D [Eq. (1.13)] and which has the values $\mathbf{P}(r,s,\tau):=\mathbf{P}(\tau)(r,s)$.

Note: Later we shall use $\mathbf{P}(\tau)$ as an abbreviation for $\mathbf{P}(r,s,\tau)$. This same dual usage was introduced in Ref. 2 (Sec. II C) in connection with scalar fields such as $\Psi(\tau)$.

The spectral potential $\mathbf{P}(\tau)$ is closely related to the spectral potential $F(t)$ of Ref. 11. The latter function, with $t=1/2\tau$ and expressed in terms of r, s, also satisfies

$$d\mathbf{F}(t)=\mathbf{A}(\tau)\mathbf{F}(t)\,.$$

However, in place of the stringent condition $\mathbf{P}(-1,1,\tau)=I$, KC imposed only conditions equivalent to the following: (i) that $\mathbf{F}(t)$ be a holomorphic function of t in a neighborhood of $t=0$; (ii) that $\mathbf{F}(0)=\Omega$; (iii) that $\dot{\mathbf{F}}(0)=\mathbf{H}$; (iv) that $\det\mathbf{F}(t)=-1/\lambda(t)$, where $\lambda(t):=\sqrt{(1-2t\mathbf{z})^2-(2t\rho)^2}$; and (v) that

$$\mathbf{F}^\dagger(t)[\Omega-t\Omega(\mathbf{H}+\mathbf{H}^\dagger)\Omega]\mathbf{F}(t)=\Omega\,.$$

The possibility of imposing further conditions on $F(t)$ so as to maximize its domain of holomorphy in the t plane was first considered by us in Ref. 5. If $F(t)$, in any gauge, is expressed as a function of (r,s) over the domain D_{IV}, then one may compute $\mathbf{P}(r,s,\tau)$ using the formula

$$\mathbf{P}(r,s,\tau)=\mathbf{F}(r,s,1/2\tau)\mathbf{F}(-1,1,1/2\tau)^{-1}\,. \tag{3.3}$$

A. Examples for the collinear case

For the Kasner case, where $H=H^{\text{K}}$, the P potential is easily evaluated. In the case of the isotropic Kasner metric ($n=0$) one obtains the especially simple result

$$\mathbf{P}_0^{\text{K}}(\tau)=\tfrac{1}{2}(I+\Omega)\chi_3(\tau)+\tfrac{1}{2}(I-\Omega)\chi_2(\tau)\,. \tag{3.4}$$

From this result we can easily generate the P potential for any other collinear case, including all the other Kasner metrics.

Definitions: In the collinear case, where $\mathbf{S}=\exp(-2\sigma_3\psi)$, for any given τ in $C-\{-1,1\}$, let $\Psi(\tau)$ be that function whose domain is D_τ [Eq. (1.15)] and which satisfies[2]

$$d\Psi(\tau)=\chi(\tau)(\tau-\mathbf{z}-\rho*)d\psi/(\tau^2-1)\,, \tag{3.5}$$

$$\Psi(\tau)(-1,1)=0\,. \tag{3.6}$$

Moreover, let Ψ denote that function whose domain is D [Eq. (1.13)] and which has the values $\Psi(r,s,\tau):=\Psi(\tau)(r,s)$.

Note: Later we shall use $\Psi(\tau)$ simply as an abbreviation for $\Psi(r,s,\tau)$.

The P potential for the collinear case is then given by

$$\mathbf{P}^{\text{coll}}(\tau)=e^{-\psi\sigma_3}\mathbf{P}_0^{\text{K}}(\tau)e^{(\tau\sigma_3+i\sigma_1)\Psi(\tau)}\,. \tag{3.7}$$

Specializing to the Kasner case, in which $\psi=-(n/2)\ln\rho$, one readily identifies

$$\Psi^{\text{K}}(\tau)=-\frac{n}{2\mu_0(\tau)}\ln\left[\frac{\rho[\tau+\mu_0(\tau)]}{\tau-\mathbf{z}+\mu(\tau)}\right], \tag{3.8}$$

where

$$\boldsymbol{\mu}(\tau) := [(\tau - r)(\tau - s)]^{1/2}, \tag{3.9}$$

$$\boldsymbol{\mu}_0(\tau) := [\tau^2 - 1]^{1/2}. \tag{3.10}$$

We employ that branch of $\boldsymbol{\Psi}^{\mathrm{K}}(\tau)$ for which $\boldsymbol{\Psi}^{\mathrm{K}}(\infty) = 0$. Note that for given $r,s,\boldsymbol{\Psi}^{\mathrm{K}}(\tau)$ is a holomorphic function of τ on $C - ([-1,r] \cup [s,1])$.

B. An example for the noncollinear case

Reference 3 employed a double-Harrison transformation[12] on Kasner metrics in order to construct a three-parameter family of colliding wave solutions with noncollinear polarizations. The spectral potential for the resulting solutions was given by EGH's Theorem 1, together with Eqs. (58)-(61). In EGH (Sec. V) it was further stated that the solution for seed solutions other than the Kasner metrics was still given by Eqs. (58)-(61) provided that their Theorem 1 was replaced by the more general Theorem 2.

Using our Eq. (3.3) we can infer from the EGH F potential a corresponding P potential of the form

$$\mathbf{P}^{\mathrm{EGH}}(\tau) = e^{\mathbf{B}^J \eta(\tau)} \sigma_3 \mathbf{P}^{\mathrm{seed}}(\tau) \sigma_3 e^{-\mathbf{B}_1^J \eta(\tau)}, \tag{3.11}$$

where J is a real 2×2 constant matrix and $\mathbf{B}^J$ is a 2×2 matrix function whose domain is D_{IV} and which is given by

$$\mathbf{B}^J := [\mathbf{N}^J + (\mathbf{N}^J)^T]\Omega / [\mathbf{N}^J - (\mathbf{N}^J)^T]\Omega, \tag{3.12}$$

$$\mathbf{N}^J := \mathbf{G}(-1)\tfrac{1}{2}(I + J)\Omega \mathbf{G}(1)^T, \tag{3.13}$$

$$\mathbf{G}(\tau) := [\boldsymbol{\lambda}(\tfrac{1}{2}\tau)]^m \sigma_3 \mathbf{F}^{\mathrm{seed}}(\tfrac{1}{2}\tau)\sigma_3. \tag{3.14}$$

Moreover,

$$\eta(\tau) := \tfrac{1}{2}\ln[(\tau + 1)/(\tau - 1)], \tag{3.15}$$

$$J^2 = I, \quad \mathrm{tr}\, J = 0, \tag{3.16}$$

$$\mathbf{B}_1^J := \mathbf{B}^J(-1,1), \tag{3.17}$$

and, although

$$\begin{aligned} \det[\mathbf{F}(1/2\tau)] &= -[\boldsymbol{\lambda}(1/2\tau)]^{-1} \\ &= -\tau[(\tau - r)(\tau - s)]^{-1/2} \to \infty \end{aligned} \tag{3.18}$$

when $\tau = 1$ and $s \to 1$ and when $\tau = -1$ and $r \to -1$, it is assumed that a real number $m > \tfrac{1}{2}$ exists such that $\mathbf{G}(\tau)$ satisfies the following conditions at $\tau = \pm 1$.

(i) The function $\mathbf{G}(r,s,\pm 1)$ has a continuous extension to all of the boundary points $(r,1)$ and $(-1,s)$ of D_{IV} such that $G(u,v,\pm 1)$ has continuous first derivatives and a continuous mixed second derivative with respect to u and v at the points $(u,0)$ and $(0,v)$ of IV.

(ii) The matrix $\mathbf{N}^J$ is not symmetric at any point in D_{IV}.

In particular, F^{K} satisfies conditions (i) and (ii) with $m = 1$ and it is possible that one can prove that these conditions hold for all members of the gauge described in EGH. For $F = F^{\mathrm{K}}$ one finds, in the notation of EGH,

$$2T^*(n,\nu,\nu')B^J = \begin{pmatrix} (x-y)T(n,\nu',\nu) - (x+y)T^*(n,\nu',\nu) & -i\rho^{1+n}[T^*(n,\nu,\nu') + T^*(n+2,\nu,\nu')] \\ i\rho^{1-n}[T^*(n,\nu,\nu') + T^*(n-2,\nu,\nu')] & -(x-y)T(n,\nu',\nu) + (x+y)T^*(n,\nu',\nu) \end{pmatrix} \tag{3.19}$$

C. Properties of the spectral potential P

Using standard theorems on ordinary differential equations,[13] one can show that the solution of Eq. (3.1) with the initial value (3.2) has certain properties. First, the function $P(u,v,\tau)$ and the derivatives P_u, P_v, and P_{uu} exist and are continuous functions of (u,v,τ) over the domain

$$\{(u,v,\tau) : (u,v) \in \mathrm{IV}, \tau \in D_{(r(u),s(v))}\}.$$

Moreover, for any given (r,s) in D_{IV}, $\mathbf{P}(\tau)$ is a holomorphic function of τ throughout $D_{(r,s)}$.

Additional properties can be derived using the same methods that we employed in Ref. 5, namely, by proving that the differentials of certain quantities vanish and then using the relations $\mathbf{P}(-1,1,\tau) = I$ and $\mathbf{H}(-1,1) = -I$. In this way one can, for example, show that $\mathbf{P}(\infty) = I$. In proving

$$\mathbf{P}^\dagger(\tau)[\tau\Omega - \tfrac{1}{2}\Omega(\mathbf{H} + \mathbf{H}^\dagger)\Omega]\mathbf{P}(\tau) = \tau\Omega + I$$

one uses the relations[5]

$$d\mathbf{H}^\dagger\, \Omega\, d\mathbf{H} = d\mathbf{H}^\dagger\, \Omega {*d}\mathbf{H} = 0.$$

To see that

$$\det \mathbf{P}(\tau) = \chi(\tau)$$

one uses $\mathbf{z} = \tfrac{1}{2}\,\mathrm{tr}(\mathbf{H}\Omega)$ and

$$\begin{aligned} \frac{d(\det \mathbf{P})}{\det \mathbf{P}} &= \mathrm{tr}(d\mathbf{P}(\tau)\mathbf{P}(\tau)^{-1}) \\ &= (\tau - \mathbf{z} + \boldsymbol{\rho}{*})^{-1}\, d\mathbf{z} = \frac{d\chi(\tau)}{\chi(\tau)}. \end{aligned}$$

IV. THE HHP ADAPTED TO (P_3,P_2)

In this section we shall formulate the matrix HHP that is the main subject of this paper. We employ the subscripts 3 and 2 to designate initial values specified at $v = 0$, $0 \leqslant u < u_0$ and $u = 0$, $0 \leqslant v < v_0$, respectively. The notation is meant to remind one that the former initial values derive from the incident plane wave in region III, while the latter initial values derive from the incident plane wave in region II. In fact, we already used this notation when in Sec. I we described the initial data as $r(u)$, $s(v)$, $E_3(u)$, and $E_2(v)$. As regards $r(u)$, $s(v)$, $E_3(u)$, and $E_2(v)$, it is sufficient for all purposes in this section and in the later sections to assume that these initial data functions are $\mathbf{C}^1$ and satisfy conditions (1.4), (1.6), (1.8), (1.9), and $E_3(0) = E_2(0) = 1$.

A. The initial spectral potential pair (P_3,P_2)

The restrictions of $P(u,v,\tau)$ to the null hypersurfaces $u = 0$ and $v = 0$ define the initial spectral potential pair

$$P_3(u,\tau) := P(u,0,\tau), \quad P_2(v,\tau) := P(0,v,\tau). \tag{4.1}$$

The construction of $\mathbf{P}_j$ $(j = 3,2)$ from the initial data is not difficult in principle, although in practice it may not always be possible to write the expressions in closed form. The ordinary differential equations that determine $\mathbf{P}_j$ $(j = 3,2)$ are

$$d\mathbf{P}_3 = \mathbf{A}_3\mathbf{P}_3, \tag{4.2}$$

$$d\mathbf{P}_2 = \mathbf{A}_2\mathbf{P}_2, \tag{4.3}$$

where $\mathbf{P}_3(-1,\tau) = \mathbf{P}_2(1,\tau) = I$. Here

$$\mathbf{A}_3(r,\tau) := \mathbf{A}(r,1,\tau) = \frac{d\mathbf{H}_3\,\Omega}{2(\tau - r)}$$

and

$$\mathbf{A}_2(s,\tau) := \mathbf{A}(-1,s,\tau) = \frac{d\mathbf{H}_2\,\Omega}{2(\tau - s)}$$

can be expressed in terms of the respective initial values of S and thus in terms of the prescribed initial data using Eq. (2.6).

The following properties are easily established.

(i) For any given r such that $-1 \leqslant r < 1$, $\mathbf{P}_3(r,\tau)$ is a holomorphic function of τ throughout $C - [-1,r]$ (in the sense that it can be holomorphically extended to include the point $\tau = 1$). Likewise, for any given s such that $-1 < s \leqslant 1$, $\mathbf{P}_2(s,\tau)$ is a holomorphic function of τ throughout $C - [s,1]$.

(ii) The function $\mathbf{Q}_2(\tau) := \mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1}$ is holomorphic on $C - [s,1]$ and the function $\mathbf{Q}_3(\tau) := \mathbf{P}(\tau)\mathbf{P}_2(\tau)^{-1}$ is holomorphic on $C - [-1,r]$.

Property (i) is implied by standard theorems on ordinary differential equations.[13] To prove (ii) integrate (3.1) along alternative paths in D_τ consisting of straight line segments:

$$(-1,1) \to (r,1) \to (r,s)$$

and

$$(-1,1) \to (-1,s) \to (r,s)\,.$$

Details will be left to the reader.

B. Admissible (P_3,P_2)

Rather than refer back to the actual initial data, it is often convenient to think of $\mathbf{P}_j$ ($j = 3,2$) as if they constituted the initial data. Of course, these "data" cannot be prescribed arbitrarily.

Definition: The ordered pair $(\mathbf{P}_3,\mathbf{P}_2)$ of 2×2 matrix functions, with the respective domains

$$\operatorname{dom} \mathbf{P}_3 = \{(r,\tau): -1 \leqslant r < 1, \tau \in C - [-1,r]\}\,, \qquad (4.4)$$

$$\operatorname{dom} \mathbf{P}_2 = \{(s,\tau): -1 < s \leqslant 1, \tau \in C - [s,1]\} \qquad (4.5)$$

will be called *admissible* if the following conditions hold.

(i) For any given $\tau \neq -1$, $d\,\mathbf{P}_3(\tau)$ exists and is continuous (in the sense defined in Ref. 2) over dom $\mathbf{P}_3(\tau)$, i.e.,

$\{r: -1 \leqslant r < \tau$ if τ is real and $-1 < \tau < 1$,

and $-1 \leqslant r < 1$ for all other $\tau\}$.

For any given $\tau \neq 1$, $d\,\mathbf{P}_2(\tau)$ exists and is continuous over dom $\mathbf{P}_2(\tau)$, i.e.,

$\{s: \tau < s \leqslant 1$ if τ is real and $-1 < \tau < 1$,

and $-1 < s \leqslant 1$ for all other $\tau\}$.

(ii) There exist initial data functions $r(u)$, $s(v)$, $E_3(u)$, and $E_2(v)$, with the properties specified for those functions, such that Eqs. (4.2) and (4.3) are satisfied.

C. Properties of (H_3,H_2) and (P_3,P_2)

The self-duality relations for the H potential imply that

$$\tfrac{1}{2}(\mathbf{H}_3 + \mathbf{H}_3^\dagger)\Omega(\mathbf{H}_3)_r = r(\mathbf{H}_3)_r\,, \qquad (4.6)$$

$$\tfrac{1}{2}(\mathbf{H}_2 + \mathbf{H}_2^\dagger)\Omega(\mathbf{H}_2)_s = s(\mathbf{H}_2)_s\,. \qquad (4.7)$$

On the other hand, it is simple to see that

$$\det \mathbf{P}_j = \chi_j\,. \qquad (4.8)$$

For any given r such that $-1 \leqslant r < 1$, $\mathbf{P}_3(r,\tau)$ is a holomorphic function of τ throughout $C - [-1,r]$. Similarly, for any given s such that $-1 < s \leqslant 1$, $\mathbf{P}_2(s,\tau)$ is a holomorphic function of τ throughout $C - [s,1]$. In a neighborhood of $\tau = \infty$,

$$\mathbf{P}_j(\tau) = I + (I + \mathbf{H}_j)\Omega/2\tau + O(\tau^{-2})\,. \qquad (4.9)$$

Moreover, the quadratic relation

$$\mathbf{P}_j^\dagger(\tau)[\tau\Omega - \Omega\tfrac{1}{2}(\mathbf{H}_j + \mathbf{H}_j^\dagger)\Omega]\mathbf{P}_j(\tau) = \tau\Omega + I \qquad (4.10)$$

holds.

D. Statement of the HHP adapted to (P_3,P_2)

We may now formulate our new HHP as follows.

Definition: Let $(\mathbf{P}_3,\mathbf{P}_2)$ denote any given admissible initial spectral potential pair. Then *the HHP adapted to* $(\mathbf{P}_3,\mathbf{P}_2)$ is the search for a 2×2 matrix function $\mathbf{P}$ with domain D such that for any given (r,s) in D_{IV}, $\mathbf{P}(\tau) = \mathbf{P}(r,s,\tau)$ satisfies the following three conditions in the τ plane.

(i) The function $\mathbf{Q}_2(\tau) := \mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1}$ is holomorphic on $C - [s,1]$ (meaning that it has a holomorphic extension to the domain $C - [s,1]$).

(ii) The function $\mathbf{Q}_3(\tau) := \mathbf{P}(\tau)\mathbf{P}_2(\tau)^{-1}$ is holomorphic on $C - [-1,r]$.

(iii) The value at infinity is $\mathbf{P}(\infty) = I$.

Note: For any given (r,s) in D_{IV}, $\mathbf{P}(r,s,\tau)$ is holomorphic throughout $C - ([-1,r] \cup [s,1])$.

Definition: The 2×2 matrix field $\mathbf{H}$ is defined by

$$(I + \mathbf{H})\Omega := \{2\tau[\mathbf{P}(\tau) - I]\}_{\tau=\infty}\,, \qquad (4.11)$$

or, equivalently, in a neighborhood of $\tau = \infty$ by

$$\mathbf{P}(\tau) = I + (I + \mathbf{H})\Omega/2\tau + O(\tau^{-1})\,. \qquad (4.12)$$

Definition: The 2×2 matrix field S is defined by

$$\mathbf{S} := -\rho^{-1}\operatorname{Re}\mathbf{H}\,. \qquad (4.13)$$

We stress that although we have used the symbol $\mathbf{P}(r,s,\tau)$ for the solution of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$ and the suggestive symbols $\mathbf{H}(r,s)$ and $\mathbf{S}(r,s)$ as well, it remains to be shown that the $\mathbf{P}(r,s,\tau)$ so defined satisfies the differential equation (3.1) and has all the other properties of a bona fide P potential and that $\mathbf{H}(r,s)$ and $\mathbf{S}(r,s)$ have the properties expected for Kinnersley's H potential and the metric matrix S. In fact, that is the crux of the problem being addressed in this paper.

V. A NEW METHOD OF GENERATING COLLIDING WAVE SOLUTIONS

Before proceeding to the proof that the solution of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$ actually solves the IVP, we shall consider one possible way to employ the HHP to generate new families of colliding wave solutions. Although we are still novices when it comes to exploiting the new HHP, its main use may turn out to be connected with proving theorems not easily proved by other techniques.

Here we shall consider any given seed metric, the spectral potential of which will be denoted by $\mathbf{P}^{\text{seed}}(\tau)$. The P potential $\mathbf{P}^{\text{EGH}}(\tau)$, which results from the application of a double-Harrison transformation,[12] has already been cited in Eq. (3.11). Imagine carrying out upon the same seed metric *two* such transformations, differing only in the choice of the real constant matrix J. From the result of the first transformation let $\mathbf{P}_3(\tau)$ be evaluated and from the result of the second transformation let $\mathbf{P}_2(\tau)$ be evaluated. In this way one easily obtains the obviously admissible pair

$$\mathbf{P}_3(\tau) = e^{\mathbf{B}_3^J\eta(\tau)}\sigma_3\mathbf{P}_3^{\text{seed}}(\tau)\sigma_3 e^{-\mathbf{B}_1^J\eta(\tau)}, \tag{5.1}$$

$$\mathbf{P}_2(\tau) = e^{\mathbf{B}_2^K\eta(\tau)}\sigma_3\mathbf{P}_2^{\text{seed}}(\tau)\sigma_3 e^{-\mathbf{B}_1^K\eta(\tau)}. \tag{5.2}$$

Here J and K are the selected constant matrices, each of which is trace-free and has a square equal to the unit matrix. Our objective is to construct the solution of the HHP adapted to this $(\mathbf{P}_3,\mathbf{P}_2)$ since from the solution of the HHP a larger family of solutions than EGH obtained using a double-Harrison transformation on the same seed metric can be constructed.

The explicit expression for $\mathbf{B}^J$ was given in the equations following Eq. (3.11). Using the expressions for $\mathbf{B}^J$ and $\mathbf{B}^K$, one can easily establish the useful relations

$$[I+\mathbf{B}_3^J][I-\mathbf{B}_3^K] = 0, \tag{5.3}$$

$$[I-\mathbf{B}_2^J][I+\mathbf{B}_2^K] = 0. \tag{5.4}$$

In particular, relations (5.3) and (5.4) tell us that

$$\mathbf{B}_1^J = \mathbf{B}_1^K = :\mathbf{B}_1. \tag{5.5}$$

We consider it amazing that the solution of the HHP adapted to this $(\mathbf{P}_3,\mathbf{P}_2)$ can be expressed in a form similar to Eq. (3.11), namely,

$$\mathbf{P}(\tau) = e^{\mathbf{B}^{JK}\eta(\tau)}\sigma_3\mathbf{P}^{\text{seed}}(\tau)\sigma_3 e^{-\mathbf{B}_1\eta(\tau)}, \tag{5.6}$$

where $[\mathbf{B}^{JK}]^2 = I$. However, if we try a $\mathbf{P}(\tau)$ of this form, it follows, for example, that

$$\mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1} = e^{\mathbf{B}^{JK}\eta(\tau)}\sigma_3\{\mathbf{P}^{\text{seed}}(\tau)\mathbf{P}_3^{\text{seed}}(\tau)^{-1}\}\sigma_3 \mathrm{e}^{-\mathbf{B}_3^J\eta(\tau)}, \tag{5.7}$$

where $\mathbf{P}^{\text{seed}}(\tau)\mathbf{P}_3^{\text{seed}}(\tau)^{-1}$ is known to be holomorphic on $C-[s,1]$. The remaining $\eta(\tau)$-dependent factors can only give rise to singularities at $\tau = \pm 1$. In fact, if we reexpress Eq. (5.7) in the form

$$\begin{aligned}\mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1} = {}&\{\tfrac{1}{2}(I+\mathbf{B}^{JK})e^{\eta(\tau)} + \tfrac{1}{2}(I-\mathbf{B}^{JK})e^{-\eta(\tau)}\}\\ &\times\sigma_3\{\mathbf{P}^{\text{seed}}(\tau)\mathbf{P}_3^{\text{seed}}(\tau)^{-1}\}\sigma_3\\ &\times\{\tfrac{1}{2}(I-\mathbf{B}_1)e^{\eta(\tau)} + \tfrac{1}{2}(I+\mathbf{B}_1)e^{-\eta(\tau)}\},\end{aligned} \tag{5.8}$$

where

$$e^{2\eta(\tau)} = (\tau+1)/(\tau-1),$$

then it becomes obvious that there can be at most a simple pole at $\tau = 1$ and a simple pole at $\tau = -1$. The object is to select $\mathbf{B}^{JK}$ so that the pole at $\tau = -1$ is eliminated from $\mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1}$ and the pole at $\tau = 1$ is eliminated from $\mathbf{P}(\tau)\mathbf{P}_2(\tau)^{-1}$. The choice that is found to accomplish the task is

$$\mathbf{B}^{JK} = \frac{[I+\mathbf{B}^J][I+\mathbf{B}^K] - [I-\mathbf{B}^K][I-\mathbf{B}^J]}{[I+\mathbf{B}^J][I+\mathbf{B}^K] + [I-\mathbf{B}^K][I-\mathbf{B}^J]}. \tag{5.9}$$

It is left to the reader to check this result and verify that $[\mathbf{B}^{JK}]^2 = I$ given that $[\mathbf{B}^J]^2 = [\mathbf{B}^K]^2 = I$.

The H potential of the new space-time with P potential given by Eqs. (5.6) and (5.9) is easily computed using Eq. (4.12). We find that

$$I+\mathbf{H} = -\sigma_3\{I+\mathbf{H}^{\text{seed}}\}\sigma_3 + 2[\mathbf{B}^{JK} - \mathbf{B}_1]\Omega. \tag{5.10}$$

As noted by EGH, the double-Harrison transformation with

$$\eta(\tau) = \tfrac{1}{2}\ln[(\tau+1)/(\tau-1)]$$

does *not* preserve the colliding wave conditions.[14] To generate a colliding wave solution one must apply the transformation to a solution such as a Kasner metric, which is not itself a colliding wave solution. The same observation applies, of course, to our new method of generating a solution. In a separate paper authored with Li, the family of space-times that is obtained when our approach is applied to Kasner seed metrics will be considered.

VI. PROOF THAT THE HHP SOLVES THE IVP

We desire to establish that the solution of the HHP exists, is unique, and has all the properties that one requires of a spectral potential.[15] The proof of the existence of a solution will be postponed to a future paper of this series. Tentatively assuming that a solution exists, we can establish uniqueness and tentatively assuming certain differentiability premises, we can show that the solution exhibits all the properties that one requires of a bona fide spectral potential.

Let us begin with some theorems which require no special assumptions other than the existence of a solution $\mathbf{P}(r,s,\tau)$ of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$.

Theorem 1:

$$\mathbf{P}(r,1,\tau) = \mathbf{P}_3(r,\tau), \tag{6.1}$$

$$\mathbf{P}(-1,s,\tau) = \mathbf{P}_2(s,\tau). \tag{6.2}$$

To prove (6.1) set $s = 1$ in conditions (i) and (ii) of the HHP in Sec. IV D and then use the fact that $\mathbf{P}_3(-1,\tau) = \mathbf{P}_2(1,\tau) = I$ to prove that $\mathbf{P}(r,1,\tau)\mathbf{P}_3(r,\tau)^{-1}$ is holomorphic on $C-\{1\}$ and $\mathbf{P}(r,1,\tau)$ is holomorphic on $C-[-1,r]$. However, this can only be true if $\mathbf{P}(r,1,\tau)\mathbf{P}_3(r,\tau)^{-1}$ is holomorphic on C. Therefore, by Liouville's theorem, Eq. (4.9), and condition (iii) of the HHP in Sec. IV D,

$$\mathbf{P}(r,1,\tau)\mathbf{P}_3(r,\tau)^{-1} = \mathbf{P}(r,1,\infty)\mathbf{P}_3(r,\infty)^{-1} = I.$$

Equation (6.2) is proved in a similar fashion.

Theorem 2:

$$\mathbf{P}(-1,1,\tau) = I. \tag{6.3}$$

Proof: Use Eq. (6.1) and the fact that $\mathbf{P}_3(-1,\tau) = I$.

Theorem 3:

$$\det\mathbf{P}(\tau) = \chi(\tau). \tag{6.4}$$

Proof: From conditions (i)–(iii) of the HHP in Sec. IV D and Eq. (4.8) it follows that $[\det\mathbf{P}(\tau)]\chi_3(\tau)^{-1}$ is holomorphic on $C-[s,1]$, $[\det\mathbf{P}(\tau)]\chi_2(\tau)^{-1}$ is holomorphic on $C-[-1,r]$, and $\det\mathbf{P}(\infty) = 1$. However, since

$\chi(\tau) = \chi_3(\tau)\chi_2(\tau)$, we conclude that $[\det \mathbf{P}(\tau)]\chi(\tau)^{-1}$ is holomorphic on $C-[s,1]$, as well as on $C-[-1,r]$; therefore, it is holomorphic on C. Again, by Liouville's theorem we have $[\det \mathbf{P}(\tau)]\chi(\tau)^{-1} = 1$.

Corollary: $\mathbf{P}(r,s,\tau)^{-1}$ exists.

Theorem 4: The solution of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$ is unique.

Proof: Let $\mathbf{P}$ and $\mathbf{P}'$ denote any solutions of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$. Then from conditions (i) and (ii) of the HHP in Sec. IV D it follows that $\mathbf{P}'(\tau)\mathbf{P}(\tau)^{-1}$ is holomorphic on $C-[-1,r]$, as well as on $C-[s,1]$, whereupon it is holomorphic on C. Using Liouville's theorem and condition (iii) of the HHP in Sec. IV D it follows that $\mathbf{P}'(\tau)\mathbf{P}(\tau)^{-1} = I$.

Theorem 5:

$$\operatorname{dom}\mathbf{H} = D_{\mathrm{IV}}, \quad \mathbf{H}(-1,1) = -I. \tag{6.5}$$

Proof: Theorem 5 follows immediately from the definition of $\mathbf{H}$, together with the definition of the HHP and Theorem 2.

Theorem 6:

$$\mathbf{H}(r,1) = \mathbf{H}_3(r)\,, \quad \mathbf{H}(-1,s) = \mathbf{H}_2(s)\,. \tag{6.6}$$

Proof: Theorem 6 follows immediately from the definition of $\mathbf{H}$, Eq. (4.9), and Theorem 1.

Theorem 7:

$$\mathbf{P}^\dagger(\tau)[\tau\Omega - \tfrac{1}{2}\Omega(\mathbf{H}+\mathbf{H}^\dagger)\Omega]\mathbf{P}(\tau) = \tau\Omega + I\,. \tag{6.7}$$

Proof: Conditions (i) and (ii) of the HHP in Sec. IV D and Eq. (4.10) permit one to conclude that

$$\begin{aligned}
&[\mathbf{P}^\dagger(\tau)]^{-1}(\tau\Omega + I)\mathbf{P}(\tau)^{-1}\\
&\quad = \{\mathbf{P}_j(\tau)\mathbf{P}(\tau)^{-1}\}^\dagger\\
&\qquad \times[\tau\Omega - \tfrac{1}{2}\Omega(\mathbf{H}_j + \mathbf{H}_j^\dagger)\Omega]\{\mathbf{P}_j(\tau)\mathbf{P}(\tau)^{-1}\}\\
&\quad = [\mathbf{Q}^\dagger_{5-j}(\tau)]^{-1}[\tau\Omega - \tfrac{1}{2}\Omega(\mathbf{H}_j + \mathbf{H}_j^\dagger)\Omega]\\
&\qquad \times\mathbf{Q}_{5-j}(\tau)^{-1}\,.
\end{aligned}$$

is holomorphic on $C-[s,1]-\{\infty\}$, is holomorphic on $C-[-1,r]-\{\infty\}$, and has a simple pole at $\tau=\infty$. Therefore,

$$[\mathbf{P}^\dagger(\tau)]^{-1}(\tau\Omega + I)\mathbf{P}(\tau)^{-1} = \tau A + B\,, \tag{6.8}$$

where A,B are τ-independent 2×2 matrices. However, from Eq. (4.12) it follows that

$$\mathbf{P}(\tau)^{-1} = I - (I+\mathbf{H})\Omega/2\tau + O(\tau^{-2})\,.$$

Inserting this into Eq. (6.8), we readily identify

$$A = \Omega\,, \quad B = -\tfrac{1}{2}\Omega(\mathbf{H}+\mathbf{H}^\dagger)\Omega$$

and Theorem 7 follows.

Corollary:

$$\det[\tau\Omega - \tfrac{1}{2}\Omega(\mathbf{H}+\mathbf{H}^\dagger)\Omega] = -\mu(\tau)^2\,. \tag{6.9}$$

The solution $P(\tau)$ of the HHP adapted to (P_3,P_2) must be shown to satisfy the differential equation (3.1) and have the other properties of a bona fide P potential. We should like to point out a relevant result concerning any complex-valued function $f(u,v,\tau)$ whose domain $\mathscr{D}$ is the same as that of P, Q_3, or Q_2 and which, for fixed (u,v) in IV, is (like P, Q_3, and Q_2) a holomorphic function of τ throughout:

$$\mathscr{D}_{(u,v)} := \{\tau\in C : (u,v,\tau)\in\mathscr{D}\}\,.$$

A result that follows from a well-known theorem is that $df(u,v,\tau)$ and $f_{uv}(u,v,\tau)$ are also holomorphic functions of τ throughout $\mathscr{D}(u,v)$ if we grant that f_u, f_v, and f_{uv} exist and are continuous functions of (u,v,τ) throughout $\mathscr{D}$. Moreover, if one considers the expansion

$$f(u,v,\tau) = \sum_{n=0}^{\infty}\frac{f^{(n)}(u,v)}{(2\tau)^n}$$

in a neighborhood of $\tau=\infty$, then $f^{(n)}, f_u^{(n)}, f_v^{(n)}$, and $f_{uv}^{(n)}$ exist and are continuous throughout IV and

$$df(\tau) = \sum_{n=0}^{\infty}\frac{df^{(n)}}{(2\tau)^n}\,, \quad f_{uv}(\tau) = \sum_{n=0}^{\infty}\frac{f_{uv}^{(n)}}{(2\tau)^n}\,.$$

We shall use this result frequently in the sequel. However, to do so we shall have to adopt as a working hypothesis a premise which will permit us to show that various functions that we encounter do exist and are continuous in the appropriate domains.

Premise: We shall assume as a working hypothesis the premise that

$$Q_j(\tau),\quad [Q_j]_u(\tau),\quad [Q_j]_v(\tau),\quad [Q_j]_{uv}(\tau) \tag{6.10}$$

exist and are continuous functions of (u,v,τ) over the respective domains

$$\{(u,v,\tau):(u,v)\in\mathrm{IV}, \tau\in C-[-1,r(u)]\}, \quad \text{if } j=3\,, \tag{6.11}$$

$$\{(u,v,\tau):(u,v)\in\mathrm{IV}, \tau\in C-[s(v),1]\}, \quad \text{if } j=2\,. \tag{6.12}$$

We shall establish the validity of this premise in a future paper of our series on the IVP.

Theorem 8: $P(\tau)$, $dP(\tau)$, and $P_{uv}(\tau)$ exist and are continuous functions of (u,v,τ) throughout

$$\{(u,v,\tau):(u,v)\in\mathrm{IV}, \tau\in D_{(r(u),s(v))}\}\,.$$

Proof: Theorem 8 is an immediate consequence of the assumption that $(\mathbf{P}_3,\mathbf{P}_2)$ is admissible, our new premise, and the definitions of Q_j $(j=3,2)$.

Corollary: H, dH, and H_{uv} exist and are continuous functions of (u,v) throughout IV.

Theorem 9:

$$P_u(\tau) = [H_u\Omega/2(\tau - r(u))]P(\tau)\,, \tag{6.13}$$

$$P_v(\tau) = [H_u\Omega/2(\tau - s(v))]P(\tau)\,. \tag{6.14}$$

Proof: Note that by the corollary to Theorem 3, $P(\tau)$ has an inverse. From Eqs. (4.1) and (4.3) and the definitions of Q_j $(j=3,2)$ it follows that

$$\begin{aligned}
P_u(\tau)P(\tau)^{-1} &= [Q_2]_u(\tau)Q_2(\tau)^{-1}\\
&\quad + Q_2(\tau)[[H_3]_u\Omega/2(\tau - r)]Q_2(\tau)^{-1}\,,\\
&= [Q_3]_u(\tau)Q_3(\tau)^{-1}\,.
\end{aligned}$$

However, for fixed (u,v) in IV, conditions (i) and (ii) of the HHP in Sec. IV D tell us that $Q_2(\tau)$ is holomorphic on $C-[-1,r]$ and $Q_3(\tau)$ is holomorphic on $C-[s,1]$. Therefore, bearing in mind the comments at the beginning of this section, $P_u(\tau)P(\tau)^{-1}$ must be holomorphic on C except for a simple pole at $\tau=r$. Therefore, we can express it in the form

$$P_u(\tau)P(\tau)^{-1} = A + B/2(\tau - r)\,,$$

where A and B are τ-independent matrices. Next, in the above equation, one expands $P(\tau)$ and $P_u(\tau)$ in a neighborhood of $\tau=\infty$ and concludes that $A=0$ and $B=H_u\Omega$, thus establishing Eq. (6.13). Equation (6.14) is established similarly.

Corollary:

$$d\mathbf{P}(\tau)=\mathbf{A}(\tau)\mathbf{P}(\tau)\,, \tag{6.15}$$

where

$$\mathbf{A}(\tau){:}=\tfrac{1}{2}(\tau-\mathbf{z}+\boldsymbol{\rho}*)^{-1}\,d\mathbf{H}\,\Omega \tag{6.16}$$

and

$$\mathbf{z}{:}=\tfrac{1}{2}(s+r)\,,\quad \boldsymbol{\rho}{:}=\tfrac{1}{2}(s-r)\,.$$

Theorem 10:

$$\tfrac{1}{2}(H+H^\dagger)\Omega H_u=r(u)H_u\,, \tag{6.17}$$

$$\tfrac{1}{2}(H+H^\dagger)\Omega H_v=s(v)H_v\,. \tag{6.18}$$

Proof: As in the proof of Theorem 9, we begin here with the observation that

$$P_u(\tau)P(\tau)^{-1}=[Q_2]_u(\tau)Q_2(\tau)^{-1}+P(\tau)[P_3(\tau)]^{-1}$$
$$\times[[H_3]_u\Omega/2(\tau-r)]Q_2(\tau)^{-1}\,.$$

Here we apply the operator $[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega]$ to both sides and use Theorem 7 to show that

$$[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega]P_u(\tau)P(\tau)^{-1}$$
$$=[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega][Q_2]_u(\tau)Q_2(\tau)^{-1}$$
$$+[Q_2^\dagger(\tau)]^{-1}[P_3^\dagger(\tau)]^{-1}(\tau\Omega+I)$$
$$\times[P_3(\tau)]^{-1}[[H_3]_u\Omega/2(\tau-r)]Q_2(\tau)^{-1}\,.$$

Now, using Eq. (4.10), we may express the above equation in the form

$$[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega]P_u(\tau)P(\tau)^{-1}$$
$$=[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega][Q_2]_u(\tau)Q_2(\tau)^{-1}$$
$$+[Q_2^\dagger(\tau)]^{-1}[\tau\Omega-\tfrac{1}{2}\Omega(H_3+H_3^\dagger)\Omega]$$
$$\times[[H_3]_u\Omega/2(\tau-r)]Q_2(\tau)^{-1}\,.$$

Finally, using Eq. (4.6), we conclude that

$$[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega]P_u(\tau)P(\tau)^{-1}$$
$$=[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega][Q_2]_u(\tau)Q_2(\tau)^{-1}$$
$$+\tfrac{1}{2}[Q_2^\dagger(\tau)]^{-1}\Omega[H_3]_u\Omega[Q_2(\tau)]^{-1}\,.$$

However, by condition (i) of the HHP in Sec. IV D, $Q_2(\tau)$ is holomorphic on $C-[s(v),1]$. Therefore, $[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega]P_u(\tau)P(\tau)^{-1}$ is holomorphic on $C-[s(v),1]$, except perhaps for a simple pole at $\tau=\infty$. On the other hand, we infer from Theorem 10 that

$$[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega]P_u(\tau)P(\tau)^{-1}$$
$$=[\tau\Omega-\tfrac{1}{2}\Omega(H+H^\dagger)\Omega][H_u\Omega/2(\tau-r(u))]\,,$$

which is holomorphic throughout C except perhaps for a simple pole at $\tau=r(u)$. Consequently, the above expression must be independent of τ, from which Eq. (6.17) follows. Equation (6.18) is proved similarly.

Corollary:

$$\tfrac{1}{2}(\mathbf{H}+\mathbf{H}^\dagger)\Omega\,d\mathbf{H}=(\mathbf{z}-\boldsymbol{\rho}*)d\mathbf{H}\,. \tag{6.19}$$

Theorem 11:

$$\tfrac{1}{2}(\mathbf{H}-\mathbf{H}^T)=\mathbf{z}\Omega\,. \tag{6.20}$$

Proof: By Eq. (1.16) and Theorem 3, it follows that

$$(\tau-\mathbf{z}+\boldsymbol{\rho}*)^{-1}\,d\mathbf{z}=\frac{d\,[\det\mathbf{P}(\tau)]}{\det\mathbf{P}(\tau)}=\mathrm{tr}(d\,\mathbf{P}(\tau)\mathbf{P}(\tau)^{-1})\,.$$

However, from the corollary to Theorem 9, one may conclude that the rhs equals

$$\tfrac{1}{2}(\tau-\mathbf{z}+\boldsymbol{\rho}*)^{-1}\,d[\mathrm{tr}(\mathbf{H}\Omega)]$$

and hence that

$$\mathbf{z}=\tfrac{1}{2}\,\mathrm{tr}(\mathbf{H}\Omega)+\text{const.}$$

The constant may be shown to vanish by considering the point of collision $r=-1$, $s=1$, where $\mathbf{z}=0$ and $\mathbf{H}(-1,1)=-I$. The above equation is (for two-dimensional matrices) equivalent to Eq. (6.20).

Corollary:

$$\mathbf{S}{:}=-\boldsymbol{\rho}^{-1}\,\mathrm{Re}\,\mathbf{H}\text{ is symmetric.} \tag{6.21}$$

Theorem 12:

$$\det\mathbf{S}=1\,. \tag{6.22}$$

Proof: From Theorem 11 and its corollary, we infer that

$$\tfrac{1}{2}(\mathbf{H}+\mathbf{H}^\dagger)=-\boldsymbol{\rho}\mathbf{S}+\mathbf{z}\Omega\,. \tag{6.23}$$

Substituting result (6.23) into Eq. (6.9), we obtain the relation

$$\det[(\tau-\mathbf{z})\Omega+\boldsymbol{\rho}\Omega\mathbf{S}\Omega]=-\boldsymbol{\mu}(\tau)^2=\boldsymbol{\rho}^2-(\tau-\mathbf{z})^2\,.$$

Setting $\tau=\mathbf{z}$ in the above equation, we obtain Eq. (6.22).

Theorem 13:

$$d(\boldsymbol{\rho}\mathbf{S}\Omega{*}d\,\mathbf{S})=0\,, \tag{6.24}$$

Proof: Equation (6.19) can be expressed in the alternative form

$$\mathbf{S}\Omega\,d\mathbf{H}={*}d\mathbf{H}\,.$$

Taking the imaginary part of the above equation, one can show that $\boldsymbol{\rho}\mathbf{S}\Omega{*}d\,\mathbf{S}$ is a closed one-form.

It should be recalled that Eq. (6.24) is equivalent to an Ernst equation for the complex field

$$\mathbf{E}{:}=(1+i\mathbf{S}_{12})/\mathbf{S}_{22}\,.$$

Moreover, from Eq. (6.6) it can be seen that $\mathbf{E}(r,s)$ satisfies the requisite initial conditions

$$\mathbf{E}(r,1)=\mathbf{E}_3(r),\quad \mathbf{E}(-1,s)=\mathbf{E}_2(s)\,.$$

Except for the proof of existence and of the working hypothesis concerning $Q_j\,(j=3,2)$, this concludes our proof that the HHP adapted to admissible $(\mathbf{P}_3,\mathbf{P}_2)$ solves the IVP we posed.

VII. A FREDHOLM INTEGRAL EQUATION

We begin with some definitions that will be useful in this section.

Definition: For any given (r,s) in D_{IV}, $\mathscr{C}_{(r,s)}$ will denote the set of all (Γ_3,Γ_2) such that Γ_3 and Γ_2 are piecewise smooth, simple, positively oriented, and nonintersecting closed contours in the finite complex plane and such that

$$\begin{aligned}&[-1,r]\subset\Gamma_3^+\,,\quad [s,1]\subset\Gamma_2^+\,,\\&[-1,r]\subset\Gamma_2^-\,,\quad [s,1]\subset\Gamma_3^-\,.\end{aligned} \tag{7.1}$$

Here, as in earlier papers by the present authors, the superscript plus denotes that portion of the complex plane that lies in the interior of the contour and the superscript minus denotes that portion of the complex plane that lies outside the contour.

A. Derivation of the Fredholm equation

Let (r,s) be any point in D_{IV} and (Γ_3,Γ_2) be any member of $\mathscr{C}_{(r,s)}$. Suppose that $\mathbf{P}(\tau)$ is the solution of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$. Then since $\mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1}$ is holomorphic in $C-[s,1]$ and $\mathbf{P}(\tau)\mathbf{P}_2(\tau)^{-1}$ is holomorphic in $C-[-1,r]$, for any $\tau\in\Gamma_3^-$ one has

$$\frac{1}{2\pi i}\int_{\Gamma_3} d\sigma\frac{\mathbf{P}(\sigma)\mathbf{P}_3(\sigma)^{-1}}{\sigma-\tau}=0 \tag{7.2}$$

and for any $\tau\in\Gamma_2^-$, one has

$$\frac{1}{2\pi i}\int_{\Gamma_2} d\sigma\frac{\mathbf{P}(\sigma)\mathbf{P}_2(\sigma)^{-1}}{\sigma-\tau}=0\,. \tag{7.3}$$

Consider now any $\tau\in C-(\Gamma_3^+\cup\Gamma_3\cup\Gamma_2^+\cup\Gamma_2)$. Then

$$\frac{1}{2\pi i}\int_{\Gamma_3} d\sigma\frac{\mathbf{P}(\sigma)\{\mathbf{P}_3(\sigma)^{-1}\mathbf{P}_3(\tau)-I\}}{\sigma-\tau}+\frac{1}{2\pi i}\int_{\Gamma_2} d\sigma\frac{\mathbf{P}(\sigma)\{\mathbf{P}_2(\sigma)^{-1}\mathbf{P}_2(\tau)-I\}}{\sigma-\tau}+\frac{1}{2\pi i}\int_{\Gamma_3\cup\Gamma_2} d\sigma\frac{\mathbf{P}(\sigma)}{\sigma-\tau}=0\,,$$

which implies in turn that

$$\mathbf{P}(\tau)=I+\frac{1}{2\pi i}\int_{\Gamma_3} d\sigma\frac{\mathbf{P}(\sigma)\{\mathbf{P}_3(\sigma)^{-1}\mathbf{P}_3(\tau)-I\}}{\sigma-\tau}+\frac{1}{2\pi i}\int_{\Gamma_2} d\sigma\frac{\mathbf{P}(\sigma)\{\mathbf{P}_2(\sigma)^{-1}\mathbf{P}_2(\tau)-I\}}{\sigma-\tau}, \tag{7.4}$$

where we have used the fact that $\mathbf{P}(\infty)=I$ to establish the existence of the integrals on the rhs and evaluate one integral. Since the "kernels" in the above integrands are holomorphic functions of τ over $C-[-1,r]$ and $C-[s,1]$, respectively, it follows that Eq. (7.4) holds for all $\tau\in D_{(r,s)}$. In particular, Eq. (7.4) holds for all τ on $\Gamma_3\cup\Gamma_2$.

B. Equivalence to the HHP

Having derived the Fredholm equation, we now show that it is equivalent to the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$.

Theorem 14: Assume that a solution of the Fredholm equation (7.4) exists and let $\mathbf{P}(\tau)$ be the holomorphic extension (which is obviously unique) of the solution to the domain $D_{(r,s)}$. Then $\mathbf{P}(\tau)$ exists and is obtained from Eq. (7.4) simply by letting $\tau\in D_{(r,s)}$ on the rhs. Moreover, $\mathbf{P}(\tau)$ is the solution of the Hilbert problem adapted to $(\mathbf{P}_3,\mathbf{P}_2)$.

Proof: Upon letting $\tau\in D_{(r,s)}$ in Eq. (7.4), we obtain the holomorphic extension of the solution. Clearly, $\mathbf{P}(\infty)=I$.

If one defines

$$\mathbf{M}_j(\tau):=\frac{1}{2\pi i}\int_{\Gamma_j} d\sigma\frac{\mathbf{P}(\sigma)\mathbf{P}_j(\sigma)^{-1}\mathbf{P}_j(\tau)}{\sigma-\tau} \tag{7.5}$$

for $\tau\in C-(\Gamma_3\cup\Gamma_3^+\cup\Gamma_2\cup\Gamma_2^+)$, then from Eq. (7.4) it follows that

$$\mathbf{M}_3(\tau)=-\mathbf{M}_2(\tau)\,. \tag{7.6}$$

However, one can see from definition (7.5) that $\mathbf{M}_3(\tau)$ and $\mathbf{M}(\tau)$ have unique holomorphic extensions such that $\mathbf{M}_3(\tau)$ is holomorphic on $C-[-1,r]$ and $\mathbf{M}_3(\infty)=0$ and $\mathbf{M}_2(\tau)$ is holomorphic on $C-[s,1]$ and $\mathbf{M}_2(\infty)=0$. Therefore, from Eq. (7.6) it follows that

$$\mathbf{M}_3(\tau)=0\,,\quad \mathbf{M}_2(\tau)=0\,,$$

i.e., for $\tau\in\Gamma_j^-$,

$$\frac{1}{2\pi i}\int_{\Gamma_j} d\sigma\frac{\mathbf{P}(\sigma)\mathbf{P}_j(\sigma)^{-1}}{\sigma-\tau}=0\,.$$

Hence, for $\tau\in\Gamma_j^+$,

$$\mathbf{P}(\tau)\mathbf{P}_j(\tau)^{-1}=\frac{1}{2\pi i}\int_{\Gamma_j} d\sigma\frac{\mathbf{P}(\sigma)\mathbf{P}_j(\sigma)^{-1}}{\sigma-\tau}\,.$$

Therefore, $\mathbf{P}(\tau)\mathbf{P}_3(\tau)^{-1}$ is holomorphic on $[-1,r]$ and $\mathbf{P}(\tau)\mathbf{P}_2(\tau)^{-1}$ is holomorphic on $[s,1]$. Because $\mathbf{P}(\tau)$ is holomorphic on $D_{(r,s)}$ it follows that the conditions of the HHP adapted to $(\mathbf{P}_3,\mathbf{P}_2)$ are satisfied. Q.E.D.

VIII. STANDARD FORM OF THE FREDHOLM EQUATION

Equation (7.4) can be reexpressed in the standard form of a Fredholm equation of the second kind:

$$\mathbf{P}(\tau)-\frac{1}{2\pi i}\int_{\Gamma_3\cup\Gamma_2} d\sigma\,\mathbf{P}(\sigma)\mathbf{K}(\sigma,\tau)=I\,. \tag{8.1}$$

As usual, we have suppressed the dependence of the solution $\mathbf{P}$ and the kernel $\mathbf{K}$ upon r and s. Moreover, the kernel depends upon one's choice of (Γ_3,Γ_2) in $\mathscr{C}_{(r,s)}$.

Definitions: The symbol $\mathbf{K}_3$ will denote that function whose domain is

$$\operatorname{dom}\mathbf{K}_3:=\{(r,\sigma,\tau):-1\leqslant r<1,\sigma\in C-[-1,r],\tau\in C-[-1,r]\} \tag{8.2}$$

and whose values are given by

$$\mathbf{K}_3(r,\sigma,\tau):=[\mathbf{P}_3(r,\sigma)^{-1}\mathbf{P}_3(r,\tau)-I]/(\sigma-\tau)\,. \tag{8.3}$$

Similarly, the symbol $\mathbf{K}_2$ will denote that function whose domain is

$$\operatorname{dom}\mathbf{K}_2:=\{(s,\sigma,\tau):-1\leqslant s<1,\sigma\in C-[s,1],\tau\in C-[s,1]\} \tag{8.4}$$

and whose values are given by

$$\mathbf{K}_2(s,\sigma,\tau):=[\mathbf{P}_2(s,\sigma)^{-1}\mathbf{P}_2(s,\tau)-I]/(\sigma-\tau)\,. \tag{8.5}$$

Equations (8.2)–(8.5) imply that for fixed (r,s) in D_{IV}, $\mathbf{K}_3(r,\sigma,\tau)$ is a holomorphic function of (σ,τ) over $(C-[-1,r])^2$ and $\mathbf{K}_2(s,\sigma,\tau)$ is a holomorphic function of (σ,τ) over $(C-[s,1])^2$. Moreover,

$$\mathbf{K}_3(-1,\sigma,\tau)=\mathbf{K}_2(1,\sigma,\tau)=0\,. \tag{8.6}$$

The symbol $\mathbf{K}$ will denote that function whose domain is

$$\operatorname{dom}\mathbf{K}:=\{(r,s,\Gamma_3,\Gamma_2,\sigma,\tau):(r,s)\in D_{\mathrm{IV}},\ (\Gamma_3,\Gamma_2\in\mathscr{C}_{(r,s)},\ \sigma\in\Gamma_3\cup\Gamma_2,\tau\in D_{(r,s)}\} \tag{8.7}$$

and whose values are given by

$$\mathbf{K}(r,s,\Gamma_3,\Gamma_2,\sigma,\tau):=\begin{cases}\mathbf{K}_3(r,\sigma,\tau)\,, & \text{when } \sigma\in\Gamma_3\,,\\ \mathbf{K}_2(s,\sigma,\tau)\,, & \text{when } \sigma\in\Gamma_2\,.\end{cases} \tag{8.8}$$

Definition: Let $\mathbf{F}$ be any function whose domain is D, i.e., the same as dom $\mathbf{P}$, and whose values are $m\times 2$ matrices with complex entries. Then $\mathbf{F}_{(r,s)}$ will denote that function whose domain is $D_{(r,s)}$ and whose values are defined by $\mathbf{F}_{(r,s)}(\tau):=\mathbf{F}(r,s,\tau)$.

Theorem 15: Suppose that $\mathbf{F}_{(r,s)}$ is holomorphic for given (r,s) in D_{IV}. Then the integral

$$(\mathbf{F}\cdot\mathbf{K})(r,s,\tau):$$
$$=\frac{1}{2\pi i}\int_{\Gamma_3\cup\Gamma_2} d\sigma\, \mathbf{F}(r,s,\sigma)\mathbf{K}(r,s,\Gamma_3,\Gamma_2,\sigma,\tau) \qquad (8.9)$$

has the same value for all (Γ_3,Γ_2) in $\mathscr{C}_{(r,s)}$. In other words, the value of the integral is independent of the choice of (Γ_3,Γ_2).

Proof: The validity of Theorem 15 follows easily from the holomorphy statements following Eq. (8.5).

Let (r,s) denote any point in D_{IV} and (Γ_3,Γ_2) denote any member of $\mathscr{C}_{(r,s)}$. The Fredholm equation is, in full explicitness,

$$\mathbf{P}(r,s,\tau)-\frac{1}{2\pi i}\int_{\Gamma_3\cup\Gamma_2} d\sigma\, \mathbf{P}(r,s,\sigma)\mathbf{K}(r,s,\Gamma_3,\Gamma_2,\sigma,\tau)=I, \qquad (8.10)$$

where $\mathbf{P}(r,s,\sigma)$ is a Lebesgue integrable 2×2 matrix function of σ on $\Gamma_3\cup\Gamma_2$. Since the kernel is a continuous function of (σ,τ) on the compact set $(\Gamma_3\cup\Gamma_2)^2$ it follows that any solution $\mathbf{P}(r,s,\tau)$ is a continuous function of τ on $\Gamma_3\cup\Gamma_2$. Therefore, without loss of generality, we can restrict ourselves to solutions $\mathbf{P}(r,s,\tau)$ which are continuous functions of τ on $\Gamma_3\cup\Gamma_2$. Moreover, from Theorem 15, $\mathbf{P}_{(r,s)}$ is holomorphically extendable to $D_{(r,s)}$; this extension is obtained by simply letting τ in the kernel of Eq. (8.10) range over $D_{(r,s)}$.

Henceforth, $\mathbf{P}_{(r,s)}$ will be understood to have $D_{(r,s)}$ as its domain and to be holomorphic over that domain.

Granting the existence of the solution $\mathbf{P}_{(r,s)}$ for each (r,s) in D_{IV}, the Fredholm equation is expressible in the form

$$\mathbf{P}-\mathbf{P}\cdot\mathbf{K}=\mathscr{I}, \qquad (8.11)$$

where $\mathbf{P}$ has domain D, $\mathscr{I}$ has domain $D_{\text{IV}}\times C$, and $\mathscr{I}(r,s,\tau)=I$ for all (r,s,τ). Furthermore, the solution $\mathbf{P}_{(r,s)}$ is independent of the choice of (Γ_3,Γ_2) and $\mathbf{P}(r,s,\infty)=I$.

IX. PERSPECTIVES

In a future paper of this series we shall present proofs based upon the Fredholm equation of the existence of solutions of the HHP adapted to (P_3,P_2) and the validity of the differentiability–continuity premises made in this paper. In addition, we shall show that the assumption of certain differentiability–continuity and analyticity properties for the initial value data permits one to infer corresponding differentiability–continuity and analyticity properties, respectively, for the solution P of the HHP. The solution for $\mathbf{P}$ in terms of the resolvent kernel will be discussed in some detail.

ACKNOWLEDGMENT

This research was supported in part by Grant No. PHY-8908231 from the National Science Foundation.

APPENDIX A: CAUCHY INTEGRAL EQUATION

Using methods very similar to those that were employed in Sec. VII in the derivation of the Fredholm equation and the proof that it is equivalent to the HHP, one can also derive the following Cauchy equation and show that it too is equivalent to the HHP adapted to (P_3,P_2):

$$\mathbf{P}(\tau)=I+\frac{1}{2\pi i}\int_{\Gamma_3} d\sigma\frac{\mathbf{P}(\sigma)\{\mathbf{P}_3(\sigma)^{-1}-I\}}{\sigma-\tau}$$
$$+\frac{1}{2\pi i}\int_{\Gamma_2} d\sigma\frac{\mathbf{P}(\sigma)\{\mathbf{P}_2(\sigma)^{-1}-I\}}{\sigma-\tau}$$

for any (Γ_3,Γ_2) in $\mathscr{C}_{(r,s)}$ and any $\tau\in C-(\Gamma_3^+\cup\Gamma_3\cup\Gamma_2^+\cup\Gamma_2)$.

We find in general that the Fredholm equation is more useful in connection with proving various theorems such as those that will be contained in the sequel to this series.

APPENDIX B: COLLINEAR CASE

The first general solution of the IVP for colliding gravitational plane waves with collinear polarizations was obtained by Szekeres[16] by using the Green's function method of Riemann. New forms of the same solution were obtained in Refs. 1 and 2 using different methods. Here we shall show the connection between the HHP of this paper and the method that was used in Ref. 2.

For the collinear case, Eqs. (1.12) and (3.4)–(3.7) yield, upon setting $s=1$ for $j=3$ and $r=-1$ for $j=2$,

$$\mathbf{P}_j(\tau)=e^{-\psi_j\sigma_3}\mathbf{P}^{\text{K}}_{0j}(\tau)e^{(\tau\sigma_3+i\sigma_1)\Psi_j(\tau)}, \qquad (\text{B}1)$$

where $\boldsymbol{\Psi}(\tau)$ is expressed in terms of the initial data function ψ_j by a simple integral which is given by Eqs. (1.12) and (2.23) of Ref. 2. Equations (3.4), (3.7), and (B1) further yield

$$\mathbf{P}(\tau)\mathbf{P}_j(\tau)^{-1}$$
$$=e^{-\psi\sigma_3}$$
$$\times\exp\{\mathbf{q}(\tau)\sigma_3[\boldsymbol{\phi}(\tau)-\chi_{j'}(\tau)\boldsymbol{\phi}_j(\tau)]\}$$
$$\times\mathbf{P}^{\text{K}}_{0j'}(\tau)e^{\psi_j\sigma_3}, \qquad (\text{B}2)$$

where

$$j':=\begin{cases}2, & \text{if } j=3,\\ 3, & \text{if } j=2,\end{cases}$$
$$\mathbf{q}(\tau):=\tfrac{1}{2}(I+\sigma_2)(\tau-r)+\tfrac{1}{2}(I-\sigma_2)(\tau-s), \qquad (\text{B}3)$$
$$\boldsymbol{\phi}(\tau):=-\chi(\tau)\boldsymbol{\Psi}(\tau),$$

and

$$\boldsymbol{\phi}_j(\tau)=-\chi_j(\tau)\boldsymbol{\Psi}_j(\tau).$$

Hence, the HHP adapted to the $(\mathbf{P}_3,\mathbf{P}_2)$ of Eq. (B1) is solved if one finds $\boldsymbol{\phi}(\tau)$ and ψ such that (i) $\boldsymbol{\phi}(\tau)-\chi_2(\tau)\boldsymbol{\phi}_3(\tau)$ is holomorphic on $C-[s,1]$, (ii) $\boldsymbol{\phi}(\tau)-\chi_3(\tau)\boldsymbol{\phi}_2(\tau)$ is holomorphic on $C-[-1,r]$, and (iii) $\boldsymbol{\phi}(\infty)=0$ and $\psi:=[-\tau\boldsymbol{\phi}(\tau)]_{\tau=\infty}$.
The above *Hilbert problem adapted to* $(\boldsymbol{\phi}_3,\boldsymbol{\phi}_2)$ (as it was called in Ref. 2) was solved explicitly in Ref. 2 and the solution is given by Eqs. (1.12) and (1.21) of that paper.

[1] I. Hauser and F. J. Ernst, J. Math. Phys. **30**, 872 (1989).
[2] I. Hauser and F. J. Ernst, J. Math. Phys. **30**, 2322 (1989).
[3] F. J. Ernst, A. Gracía-Díaz, and I. Hauser, J. Math. Phys. **29**, 681 (1988). See, also, the two earlier papers of that series: J. Math. Phys. **28**, 2155, 2951 (1987).

[4] V. Ferrari, J. Ibañez, and M. Bruni, Phys. Rev. D **36**, 1053 (1987).

[5] I. Hauser and F. J. Ernst, J. Math. Phys. **21**, 1126 (1980).

[6] Y. Nutku and M. Halil, Phys. Rev. Lett. **39**, 1379 (1977).

[7] S. Chandrasekhar and B. C. Xanthopoulos, Proc. R. Soc. London Ser. A **408**, 175 (1986).

[8] F. J. Ernst, Phys. Rev. **167**, 1175 (1968).

[9] W. Kinnersley, J. Math. Phys. **18**, 1529 (1977). See, also, W. Kinnersley and D. M. Chitre, J. Math. Phys. **18**, 1538 (1977).

[10] In Ref. 2, we let $\gamma = \psi$. We shall no longer use the symbol γ for that purpose.

[11] W. Kinnersley and D. M. Chitre, J. Math. Phys. **19**, 1926 (1978).

[12] B. K. Harrison, Phys. Rev. Lett. **41**. 1197 (1978).

[13] S. Lefschetz, *Differential Equations: Geometric Theory* (Dover, New York, 1977), 2nd ed.

[14] If one wishes to apply a double-Harrison-like transformation to a given colliding wave solution to generate a new one, one can accomplish this by moving the singularities of $\eta(\tau)$ away from $\tau = \pm 1$.

[15] We can prove all of this in a brief way by using the known theorems on the existence and uniqueness of the solution of the Cauchy problem for hyperbolic differential equations of the second order such as the Ernst equation for $E(u,v)$. However, we shall instead construct our proofs by applying holomorphic function theory to our HHP (and to the equivalent Fredholm equations) since that better fits the needs of those who want to master the HHP and its uses.

[16] P. Szekeres, J. Math. Phys. **13**, 286 (1972).

Complete integrability of two-dimensional gravity with dynamical torsion

M. O. Katanaev
Steklov Mathematical Institute, Vavilov Street, 42, GSP-1, 117966, Moscow, USSR

(Received 2 August 1989; accepted for publication 1 November 1989)

The most general Lagrangian for two-dimensional gravity with dynamical torsion is considered. A general solution of the system of nonlinear equations of motion is found. Also found is a global solution of the Cauchy problem.

I. INTRODUCTION

The two-dimensional theory of gravity has attracted a growing interest at present.[1–8] One of the reasons for studying this model is pedagogical because it provides a deeper insight into the four-dimensional gravity and its quantization. Another reason arises from close connection with string models[9] in which a two-dimensional metric on a string world sheet acquires dynamics at the quantum level.[10]

One usually adopts a constant curvature equation for a metric as an equation of motion for two-dimensional gravity.[1–8] In conformal gauge it results in the integrable Liouville equation that presents instructive problems at the quantum level.[11–14] As far as the bosonic string model is concerned, the additional Liouville mode arises after quantization in the number of dimensions not equal to the critical one.[10] This mode can already be added at the classical level and provides an interesting modification of the standard bosonic string action.[15–17]

A constant curvature equation for two-dimensional gravity cannot be obtained from an action principle for an invariant Lagrangian without auxiliary fields. This happens because in two dimensions the Hilbert–Einstein Lagrangian equals a total derivative and yields no equation of motion.

The problem of introducing a dynamic for two-dimensional metric from a purely geometric viewpoint has been solved in the framework of Riemann–Cartan geometry when metric and torsion are considered as independent dynamical variables.[18] The model is called two-dimensional gravity with dynamical torsion.

Metric and torsion[19] are fundamental and independent geometrical concepts. If matter fields are coupled minimally to metric and torsion, then the *canonical* energy-momentum tensor of matter is the source for metric and the *canonical* spin tensor is the source for torsion (see, for example, Ref. 20). This difference offers a different physical interpretation of forces provided by metric and torsion. It is natural to assume that metric describes gravity interaction between masses like in general relativity whereas torsion describes a new, probably short-range interaction between spins.

In the present paper it is proved that two-dimensional gravity with dynamical torsion is a new completely integrable model. That is a general solution of the equations of motion following from the Lagrangian

$$L = -\sqrt{-g}\,(\tfrac{1}{4}\gamma R^{2}_{\alpha\beta\gamma\delta} + \tfrac{1}{4}\beta T^{2}_{\alpha\beta\gamma} + \lambda), \quad g = \det g_{\alpha\beta}, \tag{1}$$

which contains a curvature squared term, a torsion squared term, and a cosmological constant is found. This Lagrangian is the most general Lagrangian yielding second-order equations of motion for zweibein and Lorentz connection. It also found a global solution of the Cauchy problem.

The Lagrangian L was proposed in the context of the bosonic string theory to overcome difficulties with tachyon and critical dimension of a space-time.[18] The model was called a string with dynamical geometry because metric and torsion, which define the geometry of a string world sheet, became independent dynamical variables. Functional integration indicated that addition of the Lagrangian L to the bosonic string Lagrangian modifies the theory in such a way that the notion of critical dimension disappears.[18] This fact is based on the violation of the conformal symmetry by L. The problem of tachyon is solved at the classical level in Ref. 21.

In two-dimensional space-time, the dynamical torsion theory was discussed in Ref. 22 where the Cauchy problem was anlayzed and the equations of motion were integrated in the stationary limit. This result is generalized in the present paper. It is proved that two-dimensional gravity with dynamical torsion is a new integrable model. Theorems 1 and 2 of this paper yield a general solution of the equations of motion when torsion is nonzero, the solution being expressed by the new type of special functions. Theorem 3 yields a zero torsion solution. In the letter case scalar curvature must be constant and the equations of motion are reduced to the Liouville equation. Thus the constant curvature two-dimensional gravity is the zero torsion sector of gravity theory with dynamical torsion.

The plan of the paper is as follows. In Sec. II the Riemann–Cartan geometry is briefly reviewed and the Lagrangian is discussed. In Sec. III the equations of motion are solved. Section IV contains examples of stationary and homogeneous space-times with nonzero torsion and the linear approximation. In Sec. V a global solution of the Cauchy problem is found. Section VI concludes the paper.

II. THE LAGRANGIAN

Let M be a two-dimensional manifold[19] with local coordinates ζ^{α}, $\alpha = 0, 1$. Geometry on M is defined by a metric tensor $g_{\alpha\beta}(\zeta)$ which is symmetric in its indices and by a linear connection $\Gamma_{\alpha\beta}{}^{\gamma}(\zeta)$. No symmetry under permutation of the indices of Γ is assumed. In Riemann–Cartan geometry, one postulates that the linear connection is metric compatible, that is the covariant derivative of the metric equals to zero:

$$\nabla_{\alpha} g_{\beta\gamma} = \partial_{\alpha} g_{\beta\gamma} - \Gamma_{\alpha\beta}{}^{\delta} g_{\delta\gamma} - \Gamma_{\alpha\gamma}{}^{\delta} g_{\beta\delta} = 0. \tag{2}$$

This equation quarantees that raising and lowering of tensor

0022-2488/90/040882-10$03.00

indices commutes with covariant differentiation. Equation (2) is algebraic for metrical connection Γ and can be solved. Its general solution has the following form:

$$\begin{aligned}\Gamma_{\alpha\beta\gamma} &= \Gamma_{\alpha\beta}{}^{\delta} g_{\delta\gamma} \\ &= \tfrac{1}{2}(\partial_\alpha g_{\beta\gamma} + \partial_\beta g_{\alpha\gamma} - \partial_\gamma g_{\alpha\beta}) \\ &\quad + \tfrac{1}{2}(T_{\alpha\beta\gamma} + T_{\gamma\alpha\beta} - T_{\beta\gamma\alpha}),\end{aligned}$$

where

$$T_{\alpha\beta}{}^{\gamma} = -T_{\beta\alpha}{}^{\gamma} = T_{\alpha\beta\delta} g^{\delta\gamma} = T_{\alpha\beta}{}^{\gamma} T_{\beta\alpha}{}^{\gamma}$$

is a torsion tensor which is antisymmetric in the first pair of indices.

General relativity is based on Riemannian geometry where the equation for torsion to be zero is postulated. Corresponding metrical connection is called Christoffel's symbols and is defined by the metric only

$$\widetilde{\Gamma}_{\alpha\beta\gamma} = \tfrac{1}{2}(\partial_\alpha g_{\beta\gamma} + \partial_\beta g_{\alpha\gamma} - \partial_\gamma g_{\alpha\beta}).$$

Metric and torsion form a complete set of independent geometrical notions in Riemann–Cartan geometry and define metrical connection uniquely. Curvature tensor is defined in terms of metrical connection as usual

$$R_{\alpha\beta\gamma}{}^{\delta} = \partial_\alpha \Gamma_{\beta\gamma}{}^{\delta} - \Gamma_{\alpha\gamma}{}^{\epsilon} \Gamma_{\beta\epsilon}{}^{\delta} - (\alpha \leftrightarrow \beta).$$

In the following sections, an equivalent realization of Riemann–Cartan geometry will be used. Let metric have the signature $(+\ -)$ then one can introduce zweibein $e_\alpha{}^a$, $a = 0, 1$, and Lorentz connection $\omega_\alpha{}^{ab} = -\omega_\alpha{}^{ba}$ by the following formulas:

$$e_\alpha{}^a e_\beta{}^b \eta_{ab} = g_{\alpha\beta}, \quad \eta_{ab} = \mathrm{diag}(+\ -), \tag{3}$$

$$\nabla_\alpha e_\beta{}^a = \partial_\alpha e_\beta{}^a - \Gamma_{\alpha\beta}{}^{\gamma} e_\gamma{}^a - \omega_\alpha{}^{ab} e_{\beta b} = 0. \tag{4}$$

Equation (3) defines zweibein up to a local Lorentz rotation and Eq. (4) uniquely defines Lorentz connection in terms of zweibein and metrical connection. Note that the number of components of Lorentz connection equals that of torsion. Zweibein has one component more than metric due to the symmetry under local Lorentz rotation.

Curvature and torsion have the following form in terms of zweibein and Lorentz connection:

$$\begin{aligned}R_{\alpha\beta}{}^{ab} &= \partial_\alpha \omega_\beta{}^{ab} - \omega_\alpha{}^{ac} \omega_{\beta c}{}^{b} - (\alpha \leftrightarrow \beta), \\ T_{\alpha\beta}{}^{a} &= \partial_\alpha e_\beta{}^a - \omega_\alpha{}^{ab} e_{\beta b} - (\alpha \leftrightarrow \beta).\end{aligned}$$

The transformation of Greek indices into Latin ones and vice versa is carried out by the zweibein.

Two-dimensional gravity with dynamical torsion is described by Lagrangian (1), where γ and β are coupling constants and λ is cosmological constant. This Lagrangian is invariant under general coordinate transformations and local Lorentz rotation. The presence of dimension-full constants γ and λ breaks the conformal (or Weyl) invariance. Up to a total divergence L is the unique invariant Lagrangian (among polynomials in curvature and torsion) which yields the second-order equations of motion for zweibein and Lorentz connection. Moreover, there are no pseudoinvariants that can be added to this Lagrangian. The curvature-squared term includes a kinetic term for the Lorentz connection, whereas the torsion-squared term includes a kinetic term for the zweibein and a mass term for the Lorentz connection.

The uniqueness of L can be easily proved if one notices that in two dimensions curvature tensor is completely defined by the scalar curvature $R = R_{ab}{}^{ab}$,

where ϵ_{ab} is the antisymmetric tensor, $\epsilon_{ab} = -\epsilon_{ba}$, $\epsilon_{01} = 1$, and torsion tensor is completely defined by the pseudovector $T^{*c} = T_{ab}{}^{c} \epsilon^{ab}$,

$$T_{ab}{}^{c} = -\tfrac{1}{2}\epsilon_{ab} T^{*c}.$$

III. SOLUTION OF THE EQUATIONS OF MOTION

In the present section we write down the equations of motion for zweibein and Lorentz connection then fix the conformal gauge and prove three theorems that give a general solution for the equations of motion.

To simplify calculations, let us parametrize Lorentz connection by pseudovector field B_α

$$\omega_\alpha{}^{ab} = B_\alpha \epsilon^{ab}. \tag{5}$$

This is always possible in two dimensions. Then curvature takes the following form:

$$R_{\alpha\beta}{}^{ab} = F_{\alpha\beta} \epsilon^{ab}, \quad F_{\alpha\beta} = \partial_\alpha B_\beta - \partial_\beta B_\alpha.$$

Varying the Lagrangian (1) with respect to B_α and $e_\alpha{}^a$ one gets the following equations of motion:

$$2\gamma \nabla_\beta F^{\alpha\beta} - \beta T^{\alpha}{}_{ab} \epsilon^{ab} = 0, \tag{6}$$

$$\begin{aligned}&\beta \nabla_\beta T^{\beta\alpha}{}_{a} + \beta T^{abc} T_{abc} - (\beta/4) T_{bcd} T^{bcd} e^{\alpha}{}_{a} \\ &\quad - 2\gamma F^{\alpha\beta} F_{a\beta} + (\gamma/2) F_{\beta\gamma} F^{\beta\gamma} e^{\alpha}{}_{a} - \lambda e^{\alpha}{}_{a} = 0.\end{aligned} \tag{7}$$

Here, ∇_β means the covariant derivative with Lorentz connection when it acts on tensors with Latin indices and metrical connection without torsion (Christoffell's symbols) when it acts on tensors with Greek indices. The difference between Greek and Latin indices arises after integration by parts because of the identity $\partial_\alpha \sqrt{-g} = \sqrt{-g}\,\widetilde{\Gamma}_{\alpha\beta}{}^{\beta}$.

Equations (6) and (7) can be completely integrated in the conformal gauge

$$e_\alpha{}^a = e^{\varphi} \delta^a_\alpha, \tag{8}$$

where $\varphi(\zeta)$ is a scalar field. The conformal gauge (8) can always be fixed in two dimensions by means of general coordinate and local Lorentz transformations that are parametrized by three independent functions. Then, Eqs. (6) and (7) take the following form:

$$2\gamma \partial_\beta (\epsilon^{-2\varphi} F^{\alpha\beta}) - \beta(\epsilon^{\alpha\beta} \partial_\beta \varphi - B^{\alpha}) = 0, \tag{9}$$

$$\begin{aligned}&\beta(\delta^\alpha_a \Box\varphi - \partial_a\, \partial^\alpha \varphi + \partial_a \varphi\, \partial^\alpha \varphi - \tfrac{1}{2}\delta^\alpha_a \partial_\beta \varphi\, \partial^\beta \varphi - \epsilon_a{}^{\alpha} \partial_\beta B^{\beta} \\ &\quad + \epsilon_a{}^{\beta} \partial_\beta B^{\alpha} - B_a B^{\alpha} + \tfrac{1}{2}\delta^\alpha_a B_\beta B^{\beta} + \epsilon^{\alpha\beta} \partial_\beta \varphi B_a \\ &\quad - \epsilon^{\alpha\beta} \partial_a \varphi B_\beta + \delta^\alpha_a \epsilon^{\beta\gamma} \partial_\beta \varphi B_\gamma) - 2\gamma e^{-2\varphi} F^{\alpha\beta} F_{a\beta} \\ &\quad + (\gamma/2) e^{-2\varphi} \delta^\alpha_a F^{\beta\gamma} F_{\beta\gamma} - \lambda e^{2\varphi} \delta^\alpha_a = 0,\end{aligned} \tag{10}$$

where $\Box = \partial_\beta \partial^\beta$. Here, raising and lowering of all indices are carried out by means of the Minkowskian metric, and the distinction between Greek and Latin indices disappears.

Let us denote by $C^n(D)$ a space of functions that have derivatives up to the nth order in some region D of their arguments. Then the following theorem holds.

Theorem 1: Let $B^\alpha \in C^2(D)$ and $\chi \in C^3(D)$ in some two-dimensional region of coordinates $\{\zeta^\alpha\}$. Then Eq. (9) has the following general solution in D:

$$B^\alpha = \epsilon^{\alpha\beta}\partial_\beta\chi, \tag{11}$$

where the scalar field χ satisfies the equation

$$2\gamma\Box\chi + \beta(\varphi - \chi)e^{2\varphi} = 0. \tag{12}$$

Proof: Let us introduce vector field E_α dual to B^α, $B^\alpha = \epsilon^{\alpha\beta}E_\beta$. Then Eq. (9) after multiplication by $\epsilon_{\alpha\beta}$ takes the following form:

$$-2\gamma\partial_\beta(e^{-2\varphi}\partial_\gamma E^\gamma) - \beta\partial_\beta\varphi + \beta E_\beta = 0. \tag{13}$$

Equation (13) shows that vector field E_β is a gradient of a scalar field that we denote by χ', $E_\beta = \partial_\beta\chi'$. Substitution of this expression into Eq. (13) yields the equation

$$-2\gamma\partial_\beta(e^{-2\varphi}\Box\chi') - \beta\partial_\beta(\varphi - \chi') = 0,$$

where the first integral has the form

$$-2\gamma e^{-2\varphi}\Box\chi' - \beta(\varphi - \chi') = c, \quad c = \text{const}.$$

After shifting the scalar field $\chi' = \chi + c/\beta$ one gets Eqs. (11) and (12).

The inverse statement is "if β_α is expressed by Eq. (11) in terms of the scalar field satisfying Eq. (12), then it satisfies Eq. (9)." This inverse statement can be verified by straightforward calculations.

The assumption $B^\alpha \in C^2(D)$ and $\chi \in C^3(D)$ are needed for Eqs. (9), (11), and (12) to be defined on D. □

Having found a general solution of Eq. (9), one can write down equations of motion (9) and (10) in terms of two scalar fields φ and $f = \varphi - \chi$. It is also convenient to rescale coordinates

$$\zeta^0 = \sqrt{(2\gamma/\beta)}\,\tau, \quad \zeta^1 = \sqrt{(2\gamma/\beta)}\,\sigma$$

and introduce the following notations: $\dot\varphi = \partial\varphi/\partial\tau$, $\varphi' = \partial\varphi/\partial\sigma$. Then equations of motion can be written as two dynamical equations and two constraints.

Proposition 1: Equations of motion (6) and (7) in conformal gauge are equivalent to the following system of equations:

$$\Box f + (f^2 - \Lambda)e^{2\varphi} = 0, \tag{14}$$

$$\Box\varphi + (f^2 + f - \Lambda)e^{2\varphi} = 0, \tag{15}$$

$$2f'' + \dot f^2 + f'^2 - 2\dot\varphi\dot f - 2\varphi' f' + (f^2 - \Lambda)e^{2\varphi} = 0, \tag{16}$$

$$\dot f' + \dot f f' - \dot f\varphi' - f'\dot\varphi = 0, \tag{17}$$

where $\Lambda = 4\lambda\gamma/\beta^2$.

The proof is based on Theorem 1 and simple, but lengthy, algebraic manipulations which are omitted here.

The form (14)–(17) of the equations of motion breaks down the manifest Lorentz covariance but is useful in analyzing the Cauchy problem.[22] Let us calculate the number of independent functions in the complete set of initial data. Suppose that φ and $\dot\varphi$ are defined at the initial time $\tau = 0$. Then from constraints (16) and (17) one can find initial data for f and $\dot f$ up to some integrating constants. It means that f is not an independent dynamical variable. The opposite conclusion is valid also. One can choose f as an independent dynamical variable and find initial data for φ using constraints (16) and (17). Next it must be proved that if f and φ satisfy the dynamical equations (14) and (15) then the constraints (16) and (17) are satisfied during the evolution provided they are satisfied at the initial time. This is done in Ref. 21 using the Hamiltonian approach where it is demonstrated that constraints (16) and (17) are of the first class.

A general solution of equations of motion (14)–(17) can be found in light cone coordinates

$$\xi = \sigma - \tau, \quad \eta = \sigma + \tau.$$

Let us denote derivatives by means of the indices. For example, $\varphi_\xi = \partial\varphi/\partial\xi$.

Proposition 2: In the light cone coordinates, equations of motion (14)–(17) take the following form:

$$-4f_{\xi\eta} + (f^2 - \Lambda)e^{2\varphi} = 0, \tag{18}$$

$$-4\varphi_{\xi\eta} + (f^2 + f - \Lambda)e^{2\varphi} = 0, \tag{19}$$

$$f_{\eta\eta} + f_\eta^2 - 2\varphi_\eta f_\eta = 0, \tag{20}$$

$$f_{\xi\xi} + f_\xi^2 - 2\varphi_\xi f_\xi = 0. \tag{21}$$

Proof: The proof is based on simple algebraic manipulations that are only sketched here. Equations (18) and (19) follow from (14) and (15). Equation (20) is the following linear combination (14)–(16)–2(17). Equation (21) follows from (17) after the use of (20). □

In the proof of Theorem 2 we will use the following simple lemmas showing that Eqs. (18)–(21) are not independent from each other.

Lemma 1: Equation (20) is one of the first integrals of Eqs. (18) and (19).

Proof: Let us differentiate (20) with respect to ξ,

$$f_{\eta\eta\xi} + 2f_\eta f_{\eta\xi} - 2\varphi_{\eta\xi}f_\eta - 2\varphi_\eta f_{\eta\xi} = 0,$$

and use Eqs. (18) and (19) for excluding the mixed derivatives $f_{\eta\xi}$ and $\varphi_{\eta\xi}$. Then one gets an identity. This procedure can be reversed. That is the following linear combination of equations:

$$\tfrac{1}{4}(18)_\eta + \tfrac{1}{2}f_\eta(18) - \tfrac{1}{2}f_\eta(19) - \tfrac{1}{2}\varphi_\eta(18)$$

after integration over ξ yields the equation

$$f_{\eta\eta} + f_\eta^2 - 2\varphi_\eta f_\eta = H(\eta),$$

where $H(\eta)$ is an arbitrary function. Hence, Eq. (20) is one of the first integrals of Eqs. (18) and (19) which corresponds to $H = 0$. □

Lemma 2: Equation (21) is one of the first integrals of Eqs. (18) and (19).

Proof: Proof repeats the one of Lemma 1 with the change of coordinates $\xi \leftrightarrow \eta$. □

As follows from the proved lemmas, Eq. (19) is the consequence of (18), (20) or (28), (21). But Eq. (18) does not follow from (19), (20) or (19), (21) because after integration over η or ξ one gest an arbitrary function.

In what follows, derivatives of one argument functions are denoted by primes. One easily distinguishes them from σ derivatives from the context.

Theorem 2: Let $f \in C^3(D)$, $\varphi \in C^2(D)$, and $f_\eta \neq 0$ or $f_\xi \neq 0$ in some two-dimensional region D of coordinates (ξ,η). Then general solution of equations of motion (18)–(21) in D has the following form:

$$f=\theta(F\pm G), \tag{22}$$

$$e^{2\varphi}=|\theta'|F'G'e^{\theta}, \tag{23}$$

where θ is the one-argument function defined by the following ordinary differential equation:

$$4|\theta'| = \pm\,[(\theta^2-2\theta+2-\Lambda)e^{\theta}+A\,], \quad A=\text{const}, \tag{24}$$

$F(\xi)\in C^2(O_1)$ and $G(\eta)\in C^2(O_2)$ are arbitrary functions defined on the intervals O_1 and O_2, $D\subset O_1\times O_2$, and satisfying the inequalities $F'>0$, $G'>0$. Primes denote derivatives by the arguments. One must choose either upper signs in formulas (22) and (24) or the lower signs.

Proof: The assumptions $f_\eta\neq 0$ and $f_\xi\neq 0$ are equivalent to each other as the obvious consequence of Eq. (18).

Due to the previous lemmas, Eq. (19) can be omitted as the consequence of (18), (20) or (18), (21).

Equations (20) and (21) can be integrated after dividing them by f_η *and* f_ξ:

$$\ln|f_\eta| + f - 2\varphi + \widetilde{F}(\xi) = 0, \tag{25}$$

$$\ln|f_\xi| + f - 2\varphi + \widetilde{G}(\eta) = 0, \tag{26}$$

where $\widetilde{F}(\xi)$ and $\widetilde{G}(\eta)$ are arbitrary functions of their arguments. Let us introduce two monotonic functions $F(\xi)$ and $G(\eta)$ defined by the following equations:

$$F'=\frac{dF}{d\xi}=e^{\widetilde{F}}>0, \quad G'=\frac{dG}{d\eta}=e^{\widetilde{G}}>0.$$

Then the difference between (25) and (26) is equivalent to the following equation:

$$|f_\eta|/G' = |f_\xi|/F'. \tag{27}$$

It is always possible to introduce new coordinates $(\xi,\eta)\rightarrow(F+G,F-G)$, because the Jacobian of this transformation is positive. Due to the moduli sign in Eq. (27) there are two cases. The first is $f_\eta f_\xi>0$. Then f does not depend on the coordinate $F-G$ because

$$\begin{aligned} f_{F-G} &= f_\eta\,\frac{\partial\eta}{\partial(F-G)} + f_\xi\,\frac{\partial\xi}{\partial(F-G)} \\ &= -\frac{f_\eta}{G'}+\frac{f_\xi}{F'}=0, \end{aligned}$$

as follows from Eq. (27). In the second case, $f_\eta f_\xi<0$, f is a function of one variable also, $f=(F-G)$, because $f_{F+G}=0$ due to Eq. (27). The fact that f is a function of one argument only is a crucial consequence of Eqs. (20) and (21). In what follows we denote this function by θ.

Let us consider the case $f=\theta(F+G)$. Then Eqs. (18), (20), and (21) are equivalent to the following two equations

$$-4\theta''F'G' + (\theta^2-\Lambda)e^{2\varphi}=0, \tag{28}$$

$$\ln|\theta'G'| + \theta - 2\varphi + \ln F' = 0, \tag{29}$$

which follows from (18) and (25), (26). Equation (29) can be solved with respect to φ,

$$e^{2\varphi}=|\theta'|F'G'e^{\theta}. \tag{30}$$

After substitution of this solution into Eq. (28) it can be integrated to give

$$4|\theta'| - (\theta^2-2\theta+2-\Lambda)e^{\theta} - A = 0, \quad A=\text{const}. \tag{31}$$

In the second case $f=\theta(F-G)$, and the system of Eqs. (18), (25), and (26) takes the form

$$4\theta''F'G' + (\theta^2-\Lambda)e^{2\varphi}=0, \tag{32}$$

$$\ln|\theta'G'| + \theta - 2\varphi + \ln F' = 0, \tag{33}$$

which differs from Eqs. (28) and (29) only by the sign in the first equation. Equations (32) and (33) can be integrated as in the previous case to give (30) and the equation for θ

$$4|\theta'| + (\theta^2-2\theta+2-\Lambda)e^{\theta} + A = 0, \tag{34}$$

differing from Eq. (31) only by the signs. Formulas (30), (31), and (34) yield a general solution of Eqs. (18)–(21) when $f_\eta\neq 0$ or $f_\xi\neq 0$.

The assumptions φ, F, $G\in C^2(D)$ are needed because Eqs. (18)–(21) are of the second order. The assumption $f\in C^3(D)$ is needed because in deriving Eq. (19) from (18) and (20) or (21) one must differentiate (18), and in order to satisfy the requirement $\chi\in C^3(D)$ of Theorem 1. □

The solution from Theorem 2 will be briefly discussed in the next sections. Theorem 2 does not give a general solution because of the assumption $f_\eta\neq 0$ or $f_\xi\neq 0$. The following theorem covers this gap.

Theorem 3: Let $f,\varphi\in C^2(D)$ and $f_\eta=0$ or $f_\xi=0$ in some two-dimensional region D of coordinates (ξ,η). Then the system of Eqs. (18)–(21) has a solution only for $\Lambda\geqslant 0$.

For positive Λ a general solution takes the form

$$f=\pm\sqrt{\Lambda}, \tag{35}$$

$$e^{2\varphi}=4F'G'/\sqrt{\Lambda}(F\pm G)^2, \tag{36}$$

where $F(\xi)\in C^2(O_1)$ and $G(\eta)\in C^2(O_2)$ are arbitrary functions defined on the intervals O_1 and O_2, $D\subset O_1\times O_2$, and satisfying the inequality $F'G'>0$. One must choose either upper signs in formulas (35) and (36) or the lower signs.

For zero Λ a general solution is as follows:

$$f=0, \quad \varphi=F+G. \tag{37}$$

Proof: The assumptions $f_\eta=0$ and $f_\xi=0$ are equivalent to each other as the obvious consequence of Eq. (18).

When $f_\eta=0$, Eq. (18) has the solution (35) only for $\Lambda\geqslant 0$. For constant f Eqs. (20) and (21) are satisfied and Eq. (19) transforms into the Liouville equation

$$-4\varphi_{\xi\eta} \pm \sqrt{\Lambda}e^{2\varphi}=0. \tag{38}$$

General solution of Liouville equation has a well-known form (36).

When $\Lambda=0$, then $f=0$ and Eq. (19) transforms into the d'Alembert equation which has a general solution (37).

The assumptions of $f,\varphi\in C^2(D)$ are obvious. □

From a geometric view point, Theorem 2 describes space-times with nonzero curvature and torsion, whereas Theorem 3 corresponds to a space-time with zero torsion and constant curvature. To prove this statement one has to write down scalar curvature and torsion squared, which is geometric invariant in terms of θ. Straightforward calculations yield the result

$$R=-(\beta/\gamma)\theta, \quad T_{abc}T^{abc}=-4(\beta/\gamma)|\theta'|e^{-\theta}.$$

Now one can easily verify that Theorem 3 describes space-times with constant curvature and zero torsion,

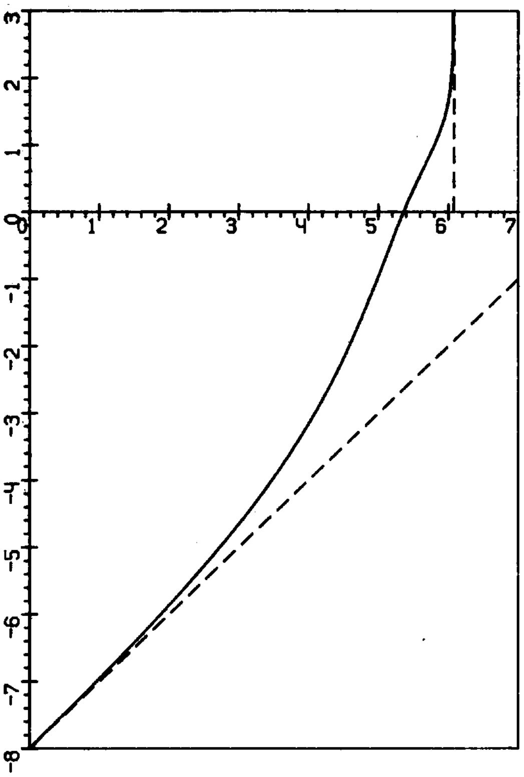


FIG. 1. Qualitative behavior of solutions of the equation $\theta' = (\theta^2 - 2\theta + 2 - \Lambda)e^\theta + A$ for different choices of the parameters Λ and A. The numerical solutions are plotted for the values of Λ and A written in square brackets. (This is the same for Figs. 2–9.) $\Lambda < 0$, $A \geqslant 0$ or $\Lambda = 1$, $A > 0$ or $\Lambda > 1$, $A > 2e^{\sqrt{\Lambda}}(\sqrt{\Lambda - 1})$. [$\Lambda = 0.5$, $A = 1.0$.]

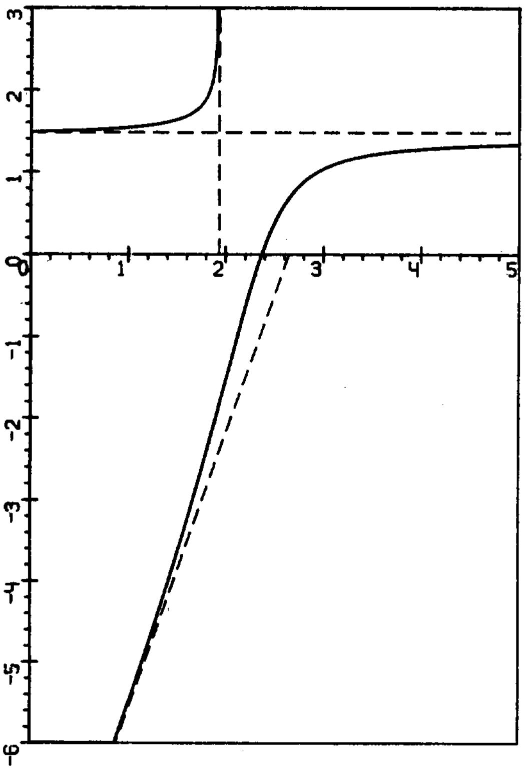


FIG. 2. $\Lambda \geqslant 0$, $A = 2e^{\sqrt{\Lambda}}(\sqrt{\Lambda} - 1)$. [$\Lambda = 2.0$.]

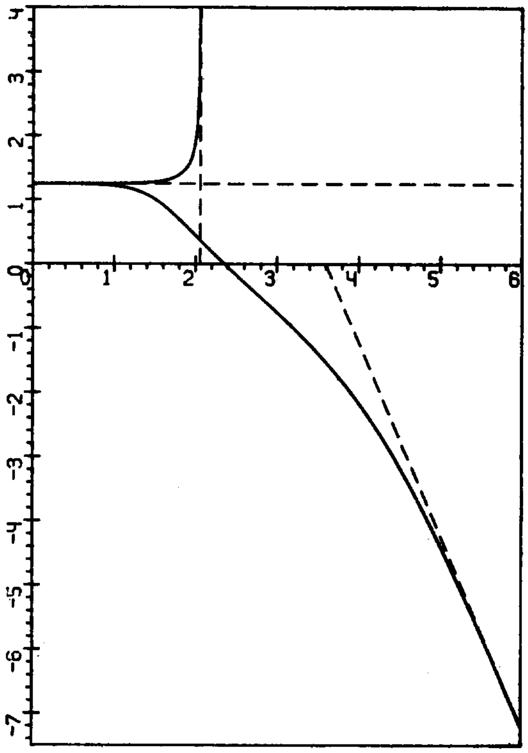


FIG. 3. $\Lambda > 0$, $A < -2e^{-\sqrt{\Lambda}}(\sqrt{\Lambda} + 1)$ or $\Lambda \leqslant 0$, $A < -2(2 - \Lambda)$. [$\Lambda = 0.2$, $A = -3.0$.]

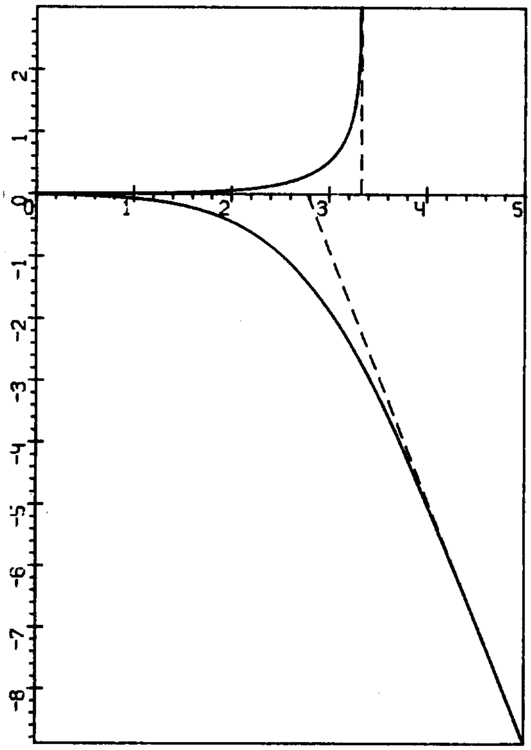


FIG. 4. $\Lambda \leqslant 0$, $A = \Lambda - 2$. [$\Lambda = -2.0$.]

$$R = \mp 2\sqrt{\lambda/\gamma}, \quad T_{abc} = 0.$$

The fact that zero torsion and constant curvature space-times satisfy the equations of motion can be traced back to Eqs. (9) and (10), and in this form has been proved in Ref. 22.

IV. SPACE-TIMES WITH NONTRIVIAL TORSION

General solution of equations of motion for two-dimensional gravity with dynamical torsion found in the previous section consists of two sectors. The first describing space-times with nonzero torsion and curvature is new whereas the second describing zero torsion and constant curvature space-times is already known. In the present section, we discuss only the first sector which is covered by Theorem 2.

In fact, Theorem 2 formulated as a *local* one yields a *global* general solution defined on the whole coordinate plane with little exception. This happens because new special function θ defined by Eq. (24) is smooth on the whole real axis or has one singular point. When $A = 0$, θ is expressible in terms of the integral logarithm functions. In a general case, Figs. 1–9 show numerical solutions of the equation

$$\theta' = (\theta^2 - 2\theta + 2 - \Lambda)e^{\theta} + A,$$

corresponding to the upper sign in Eq. (24) written without moduli sign and the factor 4 which is absorbed by rescaling the argument. There are nine qualitatively different cases for different choices of Λ and A. Each case contains up to four branches and every one of them can be shifted independently along the x axis. All branches are smooth on the whole coor-

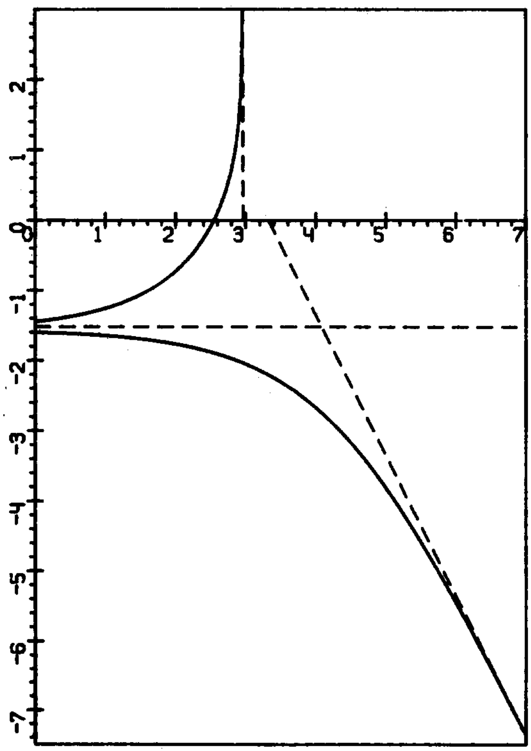


FIG. 5. $0<\Lambda<1$, $2e^{\sqrt{\Lambda}}(\sqrt{\Lambda}-1)<A<0$ or $\Lambda<0$, $\Lambda-2<A<0$. [$\Lambda = -2.0, A = -2.0$.]

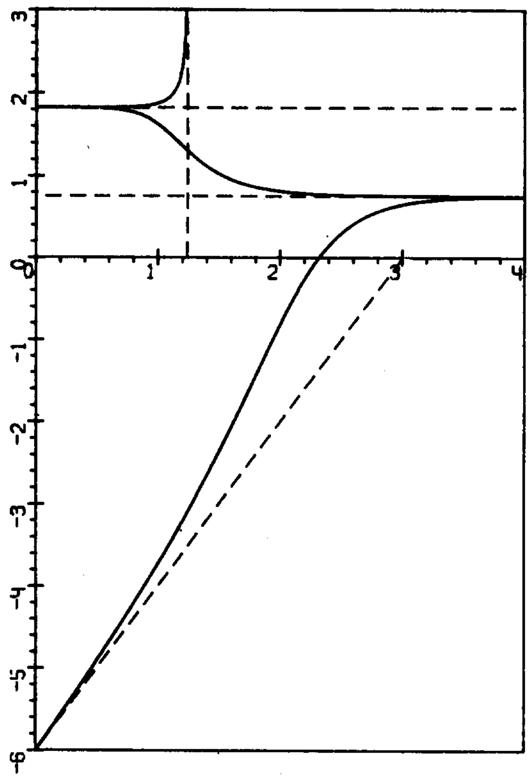


FIG. 6. $\Lambda>1, 0<A<2e^{\sqrt{\Lambda}}(\sqrt{\Lambda}-1)$. [$\Lambda = 2.0, A = 2.0$.]

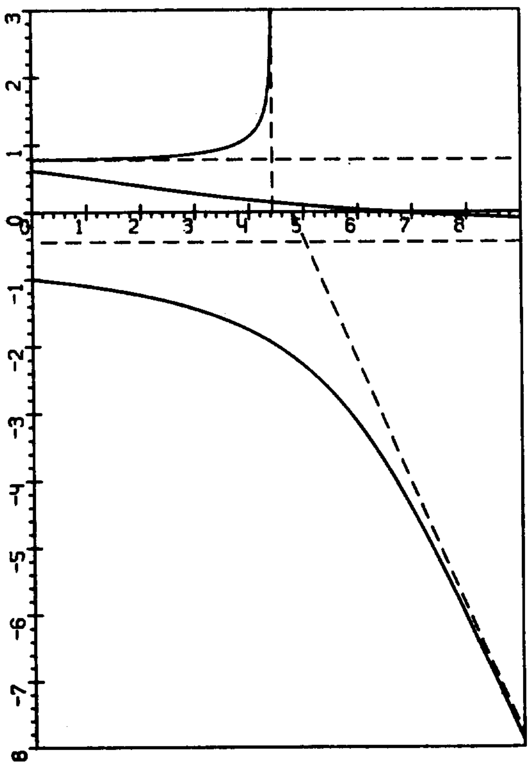


FIG. 7. $\Lambda>0, A = -2e^{-\sqrt{\Lambda}}(\sqrt{\Lambda}+1)$. [$\Lambda = 0.2$.]

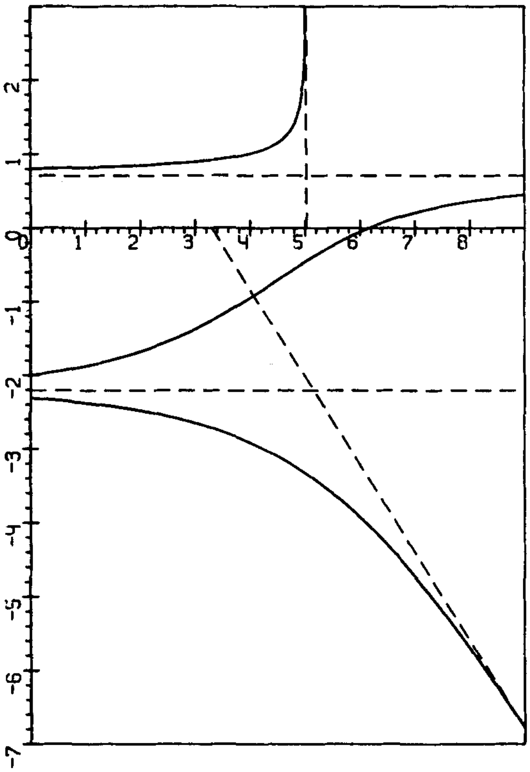


FIG. 8. $0<\Lambda<1, A=-2e^{\sqrt{\Lambda}}(1-\sqrt{\Lambda})$. [$\Lambda=0.5$.]

dinate line except the uppermost branches defined only on a half-line and having singularity at a point, say, x_1. It means that the function θ is smooth on the whole coordinate plane $(\xi,\eta)\in R^2$ except for the upper branches that are defined on the half-plane and have singularity at the boundary $F(\xi)+G(\eta)=x_1$. Thus the solution (22)–(24) is globally defined except for the upper branches.

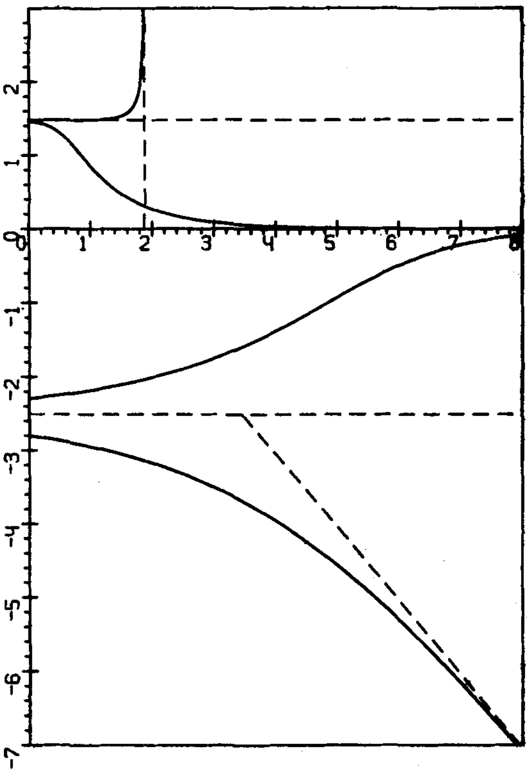


FIG. 9. $\Lambda\geqslant 1$, $-2e^{-\sqrt{\Lambda}}(\sqrt{\Lambda}+1)<A<0$ or $0<\Lambda<1$, $-2e^{-\sqrt{\Lambda}}(\sqrt{\Lambda}+1)<A<2e^{\sqrt{\Lambda}}(\sqrt{\Lambda}-1)$. [$\Lambda=1.0, A=-1.0$.]

Another important note should be made about the choice of signs in Eqs. (22) and (24). The upper signs correspond to spacelike argument $F+G$ whereas the lower choice corresponds to timelike argument $F-G$. To prove this statement let us consider tangent vectors to the line $F-G=\text{const}$. At any point of the coordinate plane the following inequality holds:

$$\left|\frac{d\sigma}{d\tau}\right| = \left| -\frac{(F-G)_\tau}{(F-G)_\sigma}\right| = \left|\frac{F'+G'}{F'-G'}\right| > 1.$$

This means that coordinate line $F+G$ is spacelike. Similar arguments prove that coordinate line $F-G$ is timelike.

To get a solution of the equations of motion (18)–(21) one must distinguish spacelike and timelike arguments. For a spacelike argument one must choose only those branches shown in Figs. 1–9 that have positive derivative because the right-hand side of Eq. (24) must be positive and double them by parity transformation, $x\to -x$. For a timelike argument one must choose the remaining branches with negative derivative and double them by time reversing, $x\to -x$, x being timelike.

At this point we consider stationary and homogeneous space-times as two simplest examples. They correspond to the following choice

$$F(\xi)=\xi, \quad G(\eta)=\eta.$$

Then for the spacelike argument all geometrical quantities will depend only from σ, and the branches on Figs. 1–9 with positive derivative show the dependence from the space coordinate σ. For the timelike argument, the branches with negative derivative show the dependence from the time coordinate τ.

There does exist stationary spherically symmetric solutions. For one-dimensional space, spherical symmetry reduces to the reflection symmetry. Therefore every one of the uppermost branches in Figs. 1–9 describes one-half of the spherically symmetric solution. One of them, with singularity at the origin, is shown in Fig. 10. It is tempting to call the singularity a two-dimensional black hole but this is not true. To get a physical interpretation of the singularity one has to analyze trajectories of pointlike particles. They can be found explicitly.[23] The analysis shows that singularity has a repulsive character and no particle can penetrate through it.

Homogeneous space-times have no singularity. Their evolution will go on forever and will be approaching a space of constant curvature and zero torsion or will be tending to infinity. In a general case, the final state is not described by Theorem 3 because the asymptotics can differ from $f=\pm\sqrt{\Lambda}$.

At the end of this section we consider the simplest example of linear approximation in order to understand the kinematics of space-time with nontrivial torsion. Let Minkowskian space-time with zero torsion be the zero-order

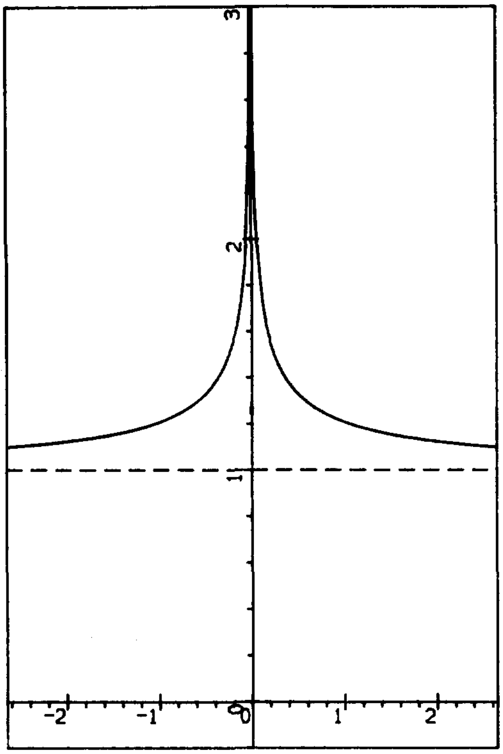


FIG. 10. The spherically symmetric solution of the equation $|\theta'| = (\theta^2 - 2\theta + 2 - \Lambda)e^{\theta} + A$ for $\Lambda = 1.0, A = 0.0$ with singularity at the origin.

approximation (vacuum). This solution exists only for zero cosmological constant. Then at the first-order approximation one gets the following equations of motion:

$$\Box\varphi = 0, \quad \Box f = 0, \quad f'' = 0, \quad \dot{f}' = 0$$

as the consequence of Proposition 1. Thus f does not describe a dynamical mode, while φ describes a massless excitation in Minkowskian space-time. The situation coincides with that of constant curvature two-dimensional gravity with zero cosmological constant because at the linear approximation near Minkowskian space-time torsion equals zero.

Torsion plays a dynamical role at the higher-order approximations or at the linear approximation near a more general type of vacuum solution. In the latter case, a stationary vacuum solution has no translational symmetry, and the theory includes all troubles inherent for quantization of constant curvature gravity[11–14] which we will not discuss here.

V. THE CAUCHY PROBLEM

Using the general solution found in Sec. III we formulate the Cauchy problem for space-times with nontrivial torsion and find a global solution for smooth initial data in the present section. The Cauchy problem for the Liouville equation describing space-times with zero torsion was considered in Ref. 24.

To get a unique smooth solution of the Cauchy problem the initial data must specify the corresponding branch of a general solution from Theorem 2. To this end one must get rid of the moduli sign in Eq. (24) and make a choice of the signs in the right-hand side of it. Moreover, one must choose the branch if there is still more than one. Before formulating the theorem let us discuss how the initial data for f and A make the unique choice of the branches.

Let f and $\dot{f}$ be defined at the initial time $\tau = 0$ on the whole axis $-\infty < \sigma < +\infty$,

$$f(\sigma,0) = f_0(\sigma), \quad \dot{f}(\sigma,0) = f_1(\sigma), \tag{39}$$

where f_0 and f_1 are sufficiently smooth functions. The sign in the right-hand side of Eq. (24) is defined by the sign of the product $f_\xi f_\eta$ as was shown in the proof of Theorem 2. It can be calculated at the initial time

$$f_\xi(\sigma,0) = \tfrac{1}{2}(f_0' - f_1), \quad f_\eta(\sigma,0) = \tfrac{1}{2}(f_0' + f_1). \tag{40}$$

Thus to get a global smooth solution of the Cauchy problem the functions f_0 and f_1 must satisfy one of the following inequalities:

$$(f_0' - f_1)(f_0' + f_1) > 0 \quad [f = \theta(F + G)],$$

$$(f_0' - f_1)(f_0' + f_1) < 0 \quad [f = \theta(F - G)].$$

To get rid of the moduli sign in Eq. (24) the sign of the derivative θ' should be known. It is also fixed by the initial data. The derivatives f_ξ and f_η are as follows:

$$f_\xi = \theta' f', \quad f_\eta = \pm\theta' G', \tag{41}$$

where ± correspond to different arguments of the solution $\theta(F \pm G)$. Because $F' > 0$ and $G' > 0$ then the sign of θ' is fixed by the signs of f_ξ and f_η.

Thus the initial data (39) should be divided into the four classes:

$$f_0' - f_1 > 0, \quad f_0' + f_1 > 0 \quad [f = \theta(F + G), \quad \theta' > 0], \tag{42}$$

$$f_0' - f_1 < 0, \quad f_0' + f_1 < 0 \quad [f = \theta(F + G), \quad \theta' < 0], \tag{43}$$

$$f_0' - f_1 > 0, \quad f_0' + f_1 < 0 \quad [f = \theta(F - G), \quad \theta' > 0], \tag{44}$$

$$f_0' - f_1 < 0, \quad f_0' + f_1 > 0 \quad [f = \theta(F - G), \quad \theta' < 0]. \tag{45}$$

For each class of the initial data, Eq. (24) can be written without a moduli sign and with a definite sign in the right-hand side of it.

To get a global smooth solution of the Cauchy problem, the value of A cannot be chosen arbitrarily too. This happens because for every branch shown in Figs. 1–9 the right-hand side of Eq. (24) must be positive at any moment and thus at the initial time. Therefore, in the cases (42) and (43) the following inequality must be satisfied:

$$(f_0^2 - 2f_0 + 2 - \Lambda)e^{f_0} + A > 0, \tag{46}$$

whereas in the cases (44) and (45) the inequality must be opposite

$$(f_0^2 - 2f_0 + 2 - \Lambda)e^{f_0} + A < 0 \tag{47}$$

for all σ.

Theorem 4: If the initial data (39) $f_0 \in C^3(-\infty < \sigma < +\infty)$, $f_1 \in C^2(-\infty < \sigma < +\infty)$ satisfy one of the inequalities (42)–(45) and the constant A satisfies the corresponding inequality (46) or (47) then the solution of the Cauchy problem for the equations of motion (14)–(17) is unique and takes the following form:

$$f=\theta(x),$$

where θ is the branch of solution of Eq. (24) containing f_0 in its range and having positive derivative for (42), (44) and negative derivative for (43), (45). The argument is defined by the initial data through the formula

$$x=\tfrac{1}{2}\theta^{-1}(f_0(\sigma-\tau))+\tfrac{1}{2}\theta^{-1}(f_0(\sigma+\tau))$$
$$+\frac{1}{2}\int_{\sigma-\tau}^{\sigma+\tau}dt\,f_1(t)\theta'^{-1}(f_0(t)). \tag{48}$$

The function φ is defined by the equation

$$e^{2\varphi}=\frac{1}{4}|\theta'|\left|\left(\frac{\partial x}{\partial\sigma}\right)^2-\left(\frac{\partial x}{\partial\tau}\right)^2\right|e^{\theta}. \tag{49}$$

Solution of the Cauchy problem is smooth and is defined on the whole half-plane $\tau>0$ or on the strip between the lines $\tau=0$ and $x(\sigma,\tau)=x_1$ where x_1 is the singular point for the upper branches of Eq. (24) with positive derivative.

Proof: If one of the inequalities (42)–(45), and the constant A satisfies the corresponding inequality (46) or (47), are satisfied then there always exists a branch θ that contains f_0 in its range and it is unique up to a translation along x axis as was discussed at the beginning of the present section.

Let us express the argument through the initial data when they satisfy the inequalities (42) and (46). Then Eq. (24) takes the following form:

$$4\theta'=(\theta^2-2\theta+2-\Lambda)e^{\theta}+A. \tag{50}$$

Using Eqs. (40) and (41) one easily finds equations defining the argument through the initial data

$$F'=2\{(f_0'-f_1)/[(f_0^2-2f_0+2-\Lambda)e^{f_0}+A]\}, \tag{51}$$

$$G'=2\{(f_0'+f_1)/[(f_0^2-2f_0+2-\Lambda)e^{f_0}+A]\}. \tag{52}$$

The argument of the solution $x=F+G$ takes the following form:

$$x=2\int_{\sigma_0}^{\xi}d\sigma\frac{f_0'}{(f_0^2-2f_0+2-\Lambda)e^{f_0}+A}$$
$$+2\int_{\sigma_0}^{\eta}d\sigma\frac{f_0'}{(f_0^2-2f_0+2-\Lambda)+e^{f_0}+A}$$
$$+2\int_{\xi}^{\eta}d\sigma\frac{f_1}{(f_0^2-2f_0+2-\Lambda)e^{f_0}+A}, \tag{53}$$

where the constant of integration σ_0 can be evaluated at the initial time

$$\theta(x(\sigma,\tau=0))=f_0(\sigma).$$

At this point one sees that the arbitrariness in shifting the branch along the x axis is compensated by the choice of σ_0.

To transform the argument (53) into the form (48) let us rewrite Eq. (50)

$$4d\theta/[(\theta^2-2\theta+2-\Lambda)e^{\theta}+A]=dt. \tag{54}$$

The first two integrals entering (53) can be rewritten using the following formula:

$$\int_{f_0(\sigma_0)}^{f_0(\xi)}\frac{d\theta}{(\theta^2-2\theta+2-\Lambda)e^{\theta}+A}$$
$$=\tfrac{1}{4}\theta^{-1}(f_0(\xi))-\tfrac{1}{4}\theta^{-1}(f_0(\sigma_0)), \tag{55}$$

where θ^{-1} is the inverse of the solution of (54), which always exists because $\theta'\neq 0$. The remaining integral in (53) can be written using the notion of the inverse function too. To this end let us differentiate (55) by $f_0(\xi)$. Then

$$[(f_0^2-2f_0+2-\Lambda)e^{f_0}+A]^{-1}=\tfrac{1}{4}\theta'^{-1}(f_0(\xi)).$$

Now one sees that Eqs. (53) and (48) are equivalent.

The only difference in the proof of Eq. (48) for another choice of the initial data is different signs in Eqs. (50)–(52) but the final result remains unchanged.

Let us discuss the region where the solution of the Cauchy problem is defined. When the initial data satisfy inequalities (46) and (47), and the solution is described by the uppermost branch shown in Figs. 1–9, then the solution of the Cauchy problem is defined on the strip between the line $\tau=0$ and the line $x(\sigma,\tau)=x_1$ where x_1 is the singular point for the branch. To prove this statement one has to note that during the evolution of a space-time the argument x of the solution is increasing, and its value at the initial time at any point σ cannot exceed or be equal to x_1 because the initial data are smooth and belong to the same branch. The increasing of the argument follows from the inequality

$$\frac{dx}{d\tau}(\xi=\text{const},\eta)=G'>0,$$

and the fact that coordinate lines $\xi=\text{const}$ are future directed.

For all other choices of the initial data satisfying inequalities (42)–(45) and (46) and (47) the solution of the Cauchy problem is defined on the whole half-plane $\tau>0$ because corresponding branches have no singularity or the singular point lies in the past.

The expression (49) for φ follows from (23) and expression for the argument $x=F\pm G$.

Conditions $f_0\in C^3$ and $f_1\in C^2$ are needed because the solution f must have derivatives up to the third order as follows from Theorem 2. □

It is instructive to check directly that the solution $f=\theta(x)$ where the argument is defined by Eq. (48) does satisfy the initial data (39). To this end let us calculate x and $\dot{x}$ at the initial time

$$x(\sigma,0)=\theta^{-1}(f_0(\sigma)),$$
$$\dot{x}(\sigma,0)=f_1(\sigma)\theta'^{-1}(f_0(\sigma)).$$

Now one sees that Theorem 4 yields the correct answer

$$\theta(x(\sigma,0))=f_0(\sigma),$$
$$\dot{\theta}(x(\sigma,0))=\theta'\dot{x}$$
$$=\theta'(\theta^{-1}(f_0(\sigma)))f_1(\sigma)\theta'^{-1}(f_0(\sigma))=f_1(\sigma).$$

The last equality follows after differentiation of the definition $f_0=\theta(\theta^{-1}(f_0))$ with respect to f_0.

Thus we have found all smooth globally defined (except for the singularity in the future) solutions of the Cauchy problem. At first sight the inequalities imposed on the initial data restrict the freedom of will. One may ask what will be if the initial data are smooth but do not satisfy inequalities (42)–(45) or (46) and (47). In this case the upper half-plane will have wedges where the solution of the Cauchy problem will be multivalued or will not be defined by the initial data at all. These solutions can be regarded as smooth solutions defined not on the upper half-plane but on the manifold with more complicated topology.

VI. CONCLUSION

In the present paper a general solution of the equations of motion for two-dimensional gravity with dynamical torsion is found. There are two sectors in the theory. The first describes two-dimensional space-times with nontrivial torsion and curvature. The second describes space-times with zero torsion. It turns out that in the second case curvature must be constant, and equations of motion reduce to the Liouville equation. In this way two-dimensional gravity with dynamical torsion solves the long standing problem of construction of purely geometric Lagrangian yielding the Liouville equation. Now it appears as one of the sectors of two-dimensional gravity with dynamical torsion corresponding to zero torsion.

Having found a general solution of the equations of motion we have found a global solution of the Cauchy problem for space-times with nontrivial torsion. Initial data can be chosen in such a way that singularity will appear after finite time of evolution or there exist globally defined smooth solutions. This situation reminds us of that for the Liouville equation.[24]

Recently, a stable instantonlike solution for the theory has been found by Akdeniz, Kizilersü, and Rizaoglu[25] (see also Ref. 26).

Two-dimensional gravity with dynamical torsion[18] was introduced in the context of the string theory and seems to solve the problem of critical dimension and existence of tachyon in the bosonic string theory. Due to its purely geometric origin, the model perhaps will find a variety of applications in mathematics and physics.

ACKNOWLEDGMENTS

The author would like to thank I. V. Volovich for valuable discussions and A. K. Pogrebkov for the suggestion simplifying the proof of Theorem 2.

[1] R. Jackiw, in *Quantum Theory of Gravity*, edited by S. Christensen (Hilger, Bristol, 1984), pp. 403–420; Nucl. Phys. B **252**, 343 (1985).

[2] C. Teitelboim, Phys. Lett. B **126**, 41 (1983); in *Quantum Theory of Gravity*, edited by S. Christensen (Hilger, Bristol, 1984), pp. 327–344; T. Banks and L. Susskind, Int. J. Theor. Phys. **23**, 475 (1984).

[3] N. Sanchez, Nucl. Phys. B **266**, 487 (1986); Proc. NATO Adv. Study Inst. Cargese, 15–31 July 1986 (Pub. New York, 1987), pp. 371–383.

[4] J. D. Brown, M. Henneaux, and C. Teitelboim, Phys. Rev. D **33**, 319 (1986).

[5] R. Balbinot and R. Floreanini, Phys. Lett. B **151**, 401 (1985); R. Floreanini, Ann. Phys. **167**, 317 (1986).

[6] T. Fukuyama and K. Kamimura. Phys. Rev. D **35**, 3768 (1987), Phys. Lett. B **200**, 75 (1988).

[7] M. Martellini, Ann. Phys. **167**, 437 (1986).

[8] K. Li, Phys. Rev. D **34**, 2292 (1986).

[9] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge U. P., Cambridge, 1987).

[10] A. M. Polyakov, Phys. Lett. B **103**, 207 (1981).

[11] E. D'Hoker and R. Jackiw, Phys. Rev. D **26**, 3517 (1982); Phys. Rev. Lett. **50**, 1719 (1983).

[12] E. Braaten, T. Curtright, and C. Thorn, Ann. Phys. **147**, 365 (1983).

[13] J.-L. Gervais and A. Neveu, Nucl. Phys. B **224**, 329 (1983).

[14] H. J. Otto and G. Weigt, Z. Phys. C **31**, 219 (1986).

[15] R. Marnelius, Nucl. Phys. B **211**, 14 (1983).

[16] L. Johansson and R. Marnelius, Phys. Rev. D **32**, 1445 (1985); Nucl. Phys. B **254**, 201 (1985).

[17] S. Hwang and R. Marnelius, Nucl. Phys. B **271**, 369 (1986).

[18] M. O. Katanaev and I. V. Volovich, Phys. Lett. B **175**, 413 (1986); Pis'ma JETP **43**, 212 (1986) (in Russian).

[19] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963) Vol. 1 (1963); Vol. 2 (1969).

[20] F. W. Hehl, Found. Phys. **15**, 451 (1985).

[21] M. O. Katanaev, Teor. Mat. Fiz. **80**, 239 (1989) (in Russian).

[22] M. O. Katanaev and I. V. Volovich, "Two-dimensional gravity with dynamical torsion and strings," preprint CERN-TH. 4771/87 (1987).

[23] M. O. Katanaev and I. V. Volovich, to be published in Ann. Phys.

[24] G. P. Jorjadze, A. K. Pogrebkov, and M. K. Polivanov, Dokl. Akad. Nauk USSR **243**, 318 (1978) (in Russian); A. K. Pogrebkon, Dokl. Akad. Nauk USSR **244**, 873 (1979) (in Russian).

[25] K. G. Akdeniz, A. Kizilersü, and E. Rizaoglu, Phys. Lett. B **215**, 81 (1988).

[26] K. G. Akdeniz, C. Dane, and M. Hortacsu, Phys. Rev. D **37**, 3074 (1988).

Renormalized equations for linear transport in stochastic media

F. Malvagi and G. C. Pomraning
School of Engineering and Applied Science, University of California, Los Angeles, Los Angeles, California 90024-1597

(Received 1 August 1989; accepted for publication 6 December 1989)

The master equation is used to obtain a model describing the ensemble-averaged intensity corresponding to linear particle transport in randomly mixed immiscible fluids. An asymptotic limit corresponding to small amounts of opaque fluids admixed with large amounts of transparent fluids is employed to reduce the complexity of the description. In the limit of a single transparent fluid, a renormalized transport equation is obtained, involving an effective source and effective interaction coefficients that account, in a simple way, for the statistical nature of the problem in this asymptotic limit.

I. INTRODUCTION

In recent years there has been considerable interest in the problem of describing linear particle transport in a stochastic medium consisting of two randomly mixed immiscible fluids.[1–16] The goal in this work has been to develop a relatively simple and accurate description for the ensemble-averaged solution of the stochastic transport problem. The generic linear transport equation treated in these papers, and which we will consider here, is written

$$\frac{1}{v}\frac{\partial\psi}{\partial t}+\mathbf{\Omega}\cdot\nabla\psi+\sigma\psi=\frac{\sigma_s}{4\pi}\phi+S, \tag{1}$$

where

$$\phi=\int_{4\pi}d\mathbf{\Omega}\,\psi(\mathbf{\Omega})\,. \tag{2}$$

In writing Eqs. (1) and (2) we have used the notation of neutron transport theory, but our considerations are equally applicable in any linear transport setting. The dependent variable in Eq. (1) is the angular flux $\psi(\mathbf{r},\mathbf{\Omega},t)$, with $\mathbf{r}$, $\mathbf{\Omega}$, and t denoting the spatial, angular (neutron flight direction), and time variables, respectively. The quantity $\phi(\mathbf{r},t)$ is the scalar flux, v is the neutron speed, $\sigma(\mathbf{r},t)$ is the macroscopic total cross section, $\sigma_s(\mathbf{r},t)$ is the macroscopic scattering cross section, and $S(\mathbf{r},\mathbf{\Omega},t)$ is the external (nonscattering) source of neutrons. We have assumed isotropic and coherent (no energy exchange) scattering in Eq. (1), but this simplification is not necessary for the essentials of the considerations to follow. Thus Eq. (1) is a monoenergetic equation, and there is no need to display the energy variable which is simply a parameter.

The master equation approach described by van Kampen[17] has been used[2,4,9,11,16] to obtain, in the case of a binary mixture, a set of two coupled transport equations in order to describe (in the case of Markov mixing statistics) the ensemble-averaged angular flux. This approach is easily extended to the case of M randomly mixed immiscible fluids. The ensemble-averaged flux, $\langle\psi\rangle$, is given by

$$\langle\psi\rangle=\sum_{i=1}^{M}p_i\psi_i\,, \tag{3}$$

where the ψ_i satisfy the coupled transport equations

$$\left[\frac{1}{v}\frac{\partial}{\partial t}+\mathbf{\Omega}\cdot\nabla+\sigma_i\right]p_i\psi_i =\frac{\sigma_{si}}{4\pi}p_i\phi_i+p_iS_i +\sum_{j\neq i}^{M}\frac{p_j\psi_j}{\lambda_{ji}}-p_i\psi_i\sum_{j\neq i}^{M}\frac{1}{\lambda_{ij}},\quad 1\leqslant i\leqslant M. \tag{4}$$

Here, $\psi_i(\mathbf{r},\mathbf{\Omega},t)$ is the ensemble-averaged flux given that the space-time point $\mathbf{r},t$ is in fluid i, $\phi_i(\mathbf{r},t)$ is the integral of ψ_i over all solid angle, and $p_i(\mathbf{r},t)$ is the probability of the space-time point $\mathbf{r},t$ being in fluid i. The $\lambda_{ij}(\mathbf{r},\mathbf{\Omega},t)$ are the Markov transition probabilities describing the transition from fluid i to fluid j. They are defined by the equation

$$\mathrm{Prob}(i\to j)=ds/\lambda_{ij}\,,\quad i\neq j\,, \tag{5}$$

where $\mathrm{Prob}(i\to j)$ is the probability of the fluid mixture being in fluid j at point $s+ds$, given that it is in fluid i at point s. The probabilities p_i in Eqs. (3) and (4) are related to the λ_{ij} according to

$$\frac{dp_i}{ds}=\sum_{j\neq i}^{M}\frac{p_j}{\lambda_{ji}}-p_i\sum_{j\neq i}^{M}\frac{1}{\lambda_{ij}}\,. \tag{6}$$

The quantities σ_i, σ_{si}, and S_i in Eq. (4) are the cross sections and external source associated with the ith fluid, and are taken as deterministic functions of their arguments. The statistical nature of the problem enters through the statistics of the fluid mixing, i.e., through the knowledge as to what fluid is present in the mixture at space point $\mathbf{r}$ and time t. That is, σ, σ_s, and S in Eq. (1) are M-state discrete random variables.

This master equation description of linear transport in an M-component Markovian mixture is known to be exact in the absence of time dependence and the scattering interaction in the underlying transport problem.[2,7,8,17] In the presence of time dependence and scattering, it is a heuristic model[4,9,11,13,16] that appears to be, at least for a binary mixture ($M=2$), qualitatively as well as semiquantitatively accurate.[13] A model of this same generic form has also been proposed for mixing statistics more general than Markovian in the case of a binary mixture.[4] We also remark that a question has recently been raised[14] concerning the physical realizability of Markov statistics in three-dimensional geometry for

0022-2488/90/040892-09$03.00

other than layered slabs. In two-dimensional geometry, the physical realizability of Markov statistics for a binary mixture has been demonstrated.[18] For the purpose of this paper, we accept Eqs. (3) and (4) as a reasonable model of linear transport in an M-component stochastic mixture, and investigate a certain asymptotic limit of this description.

The limit we consider is that corresponding to N of the M fluids being "thin (transparent)" and present in large amounts, and the remaining $M-N$ fluids being "thick (opaque)" and present in small amounts. By "thick" we mean that the cross sections σ_i and σ_{si} as well as the source S_i are large compared to the corresponding quantities for the "thin" fluids. If we let subscripts n and k denote thin and thick, respectively, we can quantify this characterization of the two types of fluids by writing

$$\sigma_n = O(1); \quad \sigma_{sn} = O(1); \quad S_n = O(1), \tag{7}$$

$$\sigma_k = O(1/\epsilon); \quad \sigma_{sk} = O(1/\epsilon); \quad S_k = O(1/\epsilon), \tag{8}$$

where ϵ is a formal smallness parameter. To introduce the presumption concerning the presence of small amounts of the thick fluids, we scale the Markov transition probabilities as

$$\lambda_{nn} = O(1); \quad \lambda_{nk} = O(1); \quad \lambda_{kk} = O(1); \quad \lambda_{kn} = O(\epsilon). \tag{9}$$

Physically, Eq. (9) states that the predominant transitions are from thick to thin fluids. Introducing the scaling given by Eq. (9) into Eq. (6), we deduce that p_i, the probability of being in the ith fluid, scales as

$$p_n = O(1); \quad p_k = O(\epsilon). \tag{10}$$

Hence we see that scaling the transition probabilities according to Eq. (9) yields the result that the thick fluids are present in small amounts.

In the remainder of this paper we show that, in lowest order in ϵ, these scalings reduce the M coupled differential equations given by Eq. (4) to N such equations, with the remaining $M-N$ equations being algebraic. Thus in this asymptotic limit, the complexity of the statistical transport description is reduced, roughly speaking, in proportion to the number of thick fluids. In particular, for a single thin fluid ($N=1$), we obtain a single renormalized transport equation involving an effective source and effective cross sections that account for the statistical nature of the problem in a very simple way. In our development, it is not necessary to specify which of the fluids is thin; only the number N of thin materials must be specified. This leads to a relatively robust reduced description in that the interchange of the designation of two fluids, one as thick and the other as thin, does not affect the result. As part of our development, we also give the necessary initial and boundary layer analyses needed to obtain the correct initial and boundary conditions on the reduced set of differential equations.

We close this introduction by noting that this asymptotic limit has already been developed in the case of a binary mixture ($M=2$).[15] In this case, N is necessarily one since there can be only one thin material and one thick material. The treatment of a binary mixture is particularly simple since one can "symmetrize" the equations for ψ_1 and ψ_2 through the change of variables[9,16]

$$\langle\psi\rangle = p_1\psi_1 + p_2\psi_2, \quad \theta = (p_1p_2)^{1/2}(\psi_2 - \psi_1). \tag{11}$$

Introducing the thick and thin scalings given by Eqs. (7)–(10) into the coupled equations for $\langle\psi\rangle$ and θ leads, in a very simple way, to a renormalized equation for $\langle\psi\rangle$ whose effective cross sections and source are invariant under the interchange of indices 1 and 2. This implies, as we noted earlier, that one need not identify which fluid is thin. When more than two fluids are present in the mixture ($M>2$), there seems to be no useful symmetrization analog of Eq. (11) and, as we shall see, it is much more involved to obtain a reduced set of equations which is symmetric upon the interchange of fluid indices. Further, the present treatment has an advantage over the earlier treatment[15] even for $M=2$ in that in any order of reduction one preserves the exact deep-in exponential decay rate for the time-independent, source-free, purely absorbing problem with spatially independent cross sections σ_i and transition probabilities λ_{ij}. This is the case, for any choice of N, no matter how far the actual physical problem is removed from the asymptotic limit that led to the reduced equations. For $M=2$, these two treatments agree, as they must, if the smallness parameter ϵ is, in reality, a vanishingly small number. That is, the two treatments are asymptotically equivalent, but the present treatment is more robust in that it yields certain exact results, as just noted, away from the asymptotic limit. We also note that the earlier binary mixture paper[15] did not consider the initial and boundary layer analyses needed to obtain the proper initial and boundary conditions for the reduced set of equations.

II. THE REDUCED EQUATIONS

Of the M fluids constituting the mixture, we assume that N of these are thin and $M-N$ are thick, as discussed in the last section. If we identify by fluid index which of the fluids are thin and which are thick, the use of the scalings given by Eqs. (7) through (10) leads to a very simple analysis. Specifically, if we introduce these scalings into Eq. (4) and seek an asymptotic solution according to

$$\psi_i = \sum_{n=0} \epsilon^n \psi_i^{(n)}, \tag{12}$$

we find, upon equating coefficients of ϵ^n, that the $\psi_i^{(0)}$ satisfy the same equations as given by Eq. (4), but with the space and time derivatives deleted from the thick fluid equations. Thus we have, with an error of $O(\epsilon)$, that the ψ_i satisfy a reduced set of equations consisting of N differential equations and $M-N$ algebraic equations. However, this very simple result has the major drawback that this reduced set of equations is not symmetric in the fluid indices. It is necessary to specify which fluid indices correspond to the thin fluids and which correspond to the thick fluids. Accordingly, we consider an alternate treatment in which it is only necessary to specify N, the number of thin fluids. The final result will be symmetric in the fluid indices, which means that it is unnecessary to identify any given fluid component as either thick or thin. In this sense, the treatment about to be given is much more robust than the simple treatment just sketched. We shall also see that it preserves, for all N and independent of how close the physical problem is to the asymptotic limit, the deep-in decay length as discussed at the end of the last section. The simple treatment just sketched does not have this property.

We rewrite Eq. (4) in matrix notation as

$$\frac{d\boldsymbol{\Psi}}{ds} + \boldsymbol{\Sigma}\boldsymbol{\Psi} = \frac{1}{4\pi}\boldsymbol{\Sigma}_s\boldsymbol{\Phi} + \mathbf{Q}, \tag{13}$$

where d/ds is the convective derivative in Eq. (4), $\boldsymbol{\Psi}$ is an M-component column vector with components $p_i\psi_i$, $\mathbf{Q}$ is a similar vector with components p_iS_i, $\boldsymbol{\Phi}$ is the integral of $\boldsymbol{\Psi}$ over all solid angle, $\boldsymbol{\Sigma}_s$ is an $M \times M$ diagonal matrix with diagonal elements σ_{si}, and $\boldsymbol{\Sigma}$ is an $M \times M$ matrix with elements given by

$$\Sigma_{ii} = \sigma_i + \sum_{j\neq i}^{M}\frac{1}{\lambda_{ij}}, \quad \Sigma_{ij} = -\frac{1}{\lambda_{ji}}, \quad i\neq j. \tag{14}$$

Prior to introducing the scalings given by Eqs. (7)–(10) into Eq. (13), we diagonalize the matrix $\boldsymbol{\Sigma}$. We introduce a new dependent variable $\boldsymbol{\chi}$ according to

$$\boldsymbol{\Psi} = \mathsf{M}\boldsymbol{\chi}, \tag{15}$$

where M is the modal matrix corresponding to $\boldsymbol{\Sigma}$, i.e., the columns of M are the eigenvectors of $\boldsymbol{\Sigma}$. Substituting Eq. (15) into Eq. (13) and left-multiplying the result by M^{-1} gives

$$\mathsf{M}^{-1}\frac{d(\mathsf{M}\boldsymbol{\chi})}{ds} + \boldsymbol{\Lambda}\boldsymbol{\chi} = \frac{1}{4\pi}\mathsf{M}^{-1}\boldsymbol{\Sigma}_s\mathsf{M}\boldsymbol{\eta} + \mathsf{M}^{-1}\mathbf{Q}. \tag{16}$$

Here, $\boldsymbol{\Lambda}$ is the diagonal matrix whose elements are the eigenvalues of $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{\Lambda} = \mathsf{M}^{-1}\boldsymbol{\Sigma}\mathsf{M}, \tag{17}$$

and

$$\boldsymbol{\eta} = \mathsf{M}^{-1}\boldsymbol{\Phi} = \int_{4\pi} d\boldsymbol{\Omega}\,\boldsymbol{\chi}(\boldsymbol{\Omega}). \tag{18}$$

In obtaining Eq. (16) we have assumed isotropic statistics, i.e., the Markov transition probabilities, λ_{ij}, are independent of direction $\boldsymbol{\Omega}$. This implies that the modal matrix M is independent of $\boldsymbol{\Omega}$ and hence passes through the angular integration which relates $\boldsymbol{\eta}$ to $\boldsymbol{\chi}$ according to Eq. (18). The more general case of anisotropic statistics, $\lambda_{ij} = \lambda_{ij}(\boldsymbol{\Omega})$, can also be treated, with some additional algebra, but for simplicity we treat only the case of isotropic statistics in this paper. The eigenvalues of $\boldsymbol{\Sigma}$, which we denote by ν satisfy

$$\boldsymbol{\Sigma}\mathbf{x} = \nu\mathbf{x}, \tag{19}$$

and the eigenvectors $\mathbf{x}$ are the columns of M. Thus the eigenvalues follow from

$$D(\nu) = \det[\boldsymbol{\Sigma} - \nu\,\mathsf{I}] = \sum_{n=0}^{M}(-1)^{M-n}c_n\nu^n = 0, \tag{20}$$

where I is the identity matrix and the coefficient c_n is the sum of all principal minors of $\boldsymbol{\Sigma}$ of order $M-n$, with $c_M = 1$. In particular, c_0 is the determinant of $\boldsymbol{\Sigma}$. If we assume N thin fluids, it is clear from the scalings given by Eqs. (7)–(10) that the matrix $\boldsymbol{\Sigma}$ has N columns of $O(1)$, and $M-N$ columns of $O(1/\epsilon)$. Then the scaling of the c_n is given by

$$c_n = \begin{cases} O(\epsilon^{-(M-N)}), & 0\leqslant n\leqslant N, \\ O(\epsilon^{-(M-n)}), & N+1\leqslant n\leqslant M. \end{cases} \tag{21}$$

From Eqs. (20) and (21), it is easily seen that N eigenvalues will be $O(1)$ and $M-N$ will be $O(1/\epsilon)$. That is, using the scaling given by Eq. (21) in Eq. (20) gives, if $\nu = O(1)$,

$$D(\nu) \sim \sum_{n=0}^{N}(-1)^{M-n}c_n\nu^n = 0, \quad \nu = O(1), \tag{22}$$

where all terms in the sum in Eq. (22) are $O(\epsilon^{-(M-N)})$. On the other hand, if we assume $\nu = O(1/\epsilon)$, we find

$$D(\nu) \sim \sum_{n=N}^{M}(-1)^{M-n}c_n\nu^n = 0, \quad \nu = O(1/\epsilon), \tag{23}$$

where all terms in the sum in Eq. (23) are $O(\epsilon^{-M})$. Equation (22) has N roots and Eq. (23) has $M-N$ roots. Thus all M eigenvalues are recovered with this scaling. It is important to note that we did not have to identify a given fluid as either thick or thin in these considerations; we only needed to specify N, the number of thin fluids, to obtain N eigenvalues of $O(1)$ and $M-N$ eigenvalues of $O(1/\epsilon)$. For concreteness, we order the eigenvalues such that

$$\nu_1 \leqslant \nu_2 \leqslant \cdots \leqslant \nu_M. \tag{24}$$

Thus the first N columns of the modal matrix M are the eigenvectors of $\boldsymbol{\Sigma}$ corresponding to $O(1)$ eigenvalues, and the remaining $M-N$ columns are the eigenvectors corresponding to $O(1/\epsilon)$ eigenvalues. In constructing the modal matrix, we conceptually normalize each eigenvector to length $O(1)$ so that the introduction of M into the analysis does not introduce any artificial scaling into the problem.

To proceed, we introduce the diagonal matrices L and U with elements

$$L_{ii} = \begin{cases} 0, & 1\leqslant i\leqslant N, \\ 1, & N+1\leqslant i\leqslant M, \end{cases} \tag{25}$$

$$U_{ii} = \begin{cases} 1, & 1\leqslant i\leqslant N, \\ 0, & N+1\leqslant i\leqslant M. \end{cases} \tag{26}$$

It is clear that $\mathsf{L}+\mathsf{U}$ is simply the identity matrix I. Thus the diagonal eigenvalue matrix $\boldsymbol{\Lambda}$ given by Eq. (17) scales as

$$\boldsymbol{\Lambda} \to \boldsymbol{\Lambda}\mathsf{U} + (1/\epsilon)\boldsymbol{\Lambda}\mathsf{L}. \tag{27}$$

It is easily seen from the scalings given by Eqs. (7) through (10) that all elements of the vector $\mathbf{Q}$ have the same scaling, which is $O(1)$. Since M has been constructed to be $O(1)$, M^{-1} is also $O(1)$ and thus

$$\mathsf{M}^{-1}\mathbf{Q} = O(1). \tag{28}$$

To obtain the scaling on the term involving $\boldsymbol{\Sigma}_s$ in Eq. (16), we introduce the matrix c defined by

$$\boldsymbol{\Sigma}_s = \mathsf{c}\boldsymbol{\Sigma}; \quad \mathsf{c}^{-1} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_s^{-1}. \tag{29}$$

Since $\boldsymbol{\Sigma}_s$ is diagonal, its inverse is simply computed and using the scalings given by Eqs. (7)–(10) one easily verifies that $\mathsf{c}^{-1} = O(1)$, which implies that $\mathsf{c} = O(1)$ since $|\mathsf{c}|\neq 0$. Thus $(\mathsf{M}^{-1}\mathsf{c}\mathsf{M}) = O(1)$ and we conclude that

$$\mathsf{M}^{-1}\boldsymbol{\Sigma}_s\mathsf{M} = \mathsf{M}^{-1}\mathsf{c}\boldsymbol{\Sigma}\mathsf{M} = (\mathsf{M}^{-1}\mathsf{c}\mathsf{M})\boldsymbol{\Lambda} \tag{30}$$

is a matrix with the first N columns of $O(1)$ and the remaining $M-N$ columns of $O(1/\epsilon)$, i.e.,

$$\mathsf{M}^{-1}\boldsymbol{\Sigma}_s\mathsf{M} \to \mathsf{M}^{-1}\boldsymbol{\Sigma}_s\mathsf{M}\mathsf{U} + (1/\epsilon)\mathsf{M}^{-1}\boldsymbol{\Sigma}_s\mathsf{M}\mathsf{L}. \tag{31}$$

Finally, we need to consider the scaling of the derivative term in Eq. (16). If we assume that the present considerations involve scalings away from initial and boundary layers (i.e., we seek an interior solution), we take d/ds to be $O(1)$. We also assume that $d\mathsf{M}/ds$ is $O(1)$. That is, we assume that the σ_i and λ_{ij} are sufficiently slowly varying in time and space such that differentiating M does not introduce terms larger than $O(1)$. Combining all of these considerations, we find that in the interior (away from initial and boundary layers) Eq. (16) scales as

$$\mathsf{M}^{-1}\frac{d(\mathsf{M}\chi)}{ds}+\Lambda\mathsf{U}\chi+\frac{1}{\epsilon}\Lambda\mathsf{L}\chi$$
$$=(1/4\pi)\left[\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{U}\eta+(1/\epsilon)\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{L}\eta\right]+\mathsf{M}^{-1}\mathbf{Q}\,. \tag{32}$$

Having introduced the scaling, we can easily return to the physical variable Ψ [see Eq. (15)]. We find the equivalent scaled equation

$$\frac{d\Psi}{ds}+(\Sigma\mathsf{M}\mathsf{U}\mathsf{M}^{-1})\Psi+\frac{1}{\epsilon}(\Sigma\mathsf{M}\mathsf{L}\mathsf{M}^{-1})\Psi$$
$$=(1/4\pi)\left[(\Sigma_s\mathsf{M}\mathsf{U}\mathsf{M}^{-1})\Phi+(1/\epsilon)(\Sigma_s\mathsf{M}\mathsf{L}\mathsf{M}^{-1})\Phi\right]+\mathbf{Q}\,. \tag{33}$$

We emphasize that the scaled equations given by Eqs. (32) and (33) depend only upon N, the number of fluid components identified as thin. We did not need to specify, by identifying the fluid indices i, which of the fluids are thin and which are thick.

In seeking an asymptotic solution to these scaled equations, it is algebraically easier to use Eq. (32) rather than Eq. (33). Accordingly, we seek a solution to Eq. (32) as a power series in ϵ, i.e.,

$$\chi=\sum_{n=0}\epsilon^n\chi^{(n)}\,. \tag{34}$$

Using Eq. (34) in Eq. (32) and equating coefficients of ϵ^n, we obtain as the first two equations

$$\Lambda\mathsf{L}\chi^{(0)}=(1/4\pi)\,\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{L}\eta^{(0)}\,, \tag{35}$$

$$\mathsf{M}^{-1}\frac{d(\mathsf{M}\chi^{(0)})}{ds}+\Lambda\mathsf{U}\chi^{(0)}+\Lambda\mathsf{L}\chi^{(1)}$$
$$=(1/4\pi)\left[\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{U}\eta^{(0)}+\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{L}\eta^{(1)}\right]+\mathsf{M}^{-1}\mathbf{Q}. \tag{36}$$

Integration of Eq. (35) over all solid angle gives, assuming isotropic statistics,

$$(\Lambda-\mathsf{M}^{-1}\Sigma_s\mathsf{M})\mathsf{L}\eta^{(0)}=0\,. \tag{37}$$

Since the matrix multiplying $\mathsf{L}\eta^{(0)}$ in Eq. (37) is nonsingular, we have

$$\mathsf{L}\eta^{(0)}=0\,. \tag{38}$$

Using Eq. (38) in Eq. (35) then gives, since Λ is nonsingular,

$$\mathsf{L}\chi^{(0)}=0\,. \tag{39}$$

Now, left-multiplying Eq. (36) by L gives

$$\mathsf{L}\mathsf{M}^{-1}\frac{d(\mathsf{M}\chi^{(0)})}{ds}+\Lambda\mathsf{L}\chi^{(1)}$$
$$=\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}+(1/4\pi)(\mathsf{L}\mathsf{M}^{-1}\Sigma_s\mathsf{M})(\mathsf{U}\eta^{(0)}+\mathsf{L}\eta^{(1)})\,. \tag{40}$$

The derivative term in Eq. (40) is of $O(\epsilon)$ and can legitimately be neglected; it actually belongs in the next higher order (in ϵ) equation. To see this, we use the identity

$$\chi^{(0)}=(\mathsf{L}+\mathsf{U})\chi^{(0)}\,, \tag{41}$$

and use of Eq. (39) in Eq. (41) yields

$$\chi^{(0)}=\mathsf{U}\chi^{(0)}\,. \tag{42}$$

Then the derivative term in Eq. (40) can be written

$$\mathsf{L}\mathsf{M}^{-1}\frac{d(\mathsf{M}\chi^{(0)})}{ds}=\mathsf{L}\mathsf{M}^{-1}\left(\frac{d\mathsf{M}}{ds}\right)\mathsf{U}\chi^{(0)}\,. \tag{43}$$

If one examines the eigenvalue problem given by Eq. (19) under the assumption of N thin fluids, one deduces that each $O(1/\epsilon)$ eigenvalue has a corresponding eigenvector with all elements $O(1)$, and each $O(1)$ eigenvalue has a corresponding eigenvector with N elements $O(1)$ and the remainder $O(\epsilon)$. The ordering of the $O(\epsilon)$ elements depends upon which fluids are identified as thin. However, no matter what identification is made, the product $\mathsf{M}^{-1}(d\mathsf{M}/ds)$ always scales the same; namely, all elements of this product are $O(1)$ except for those in a submatrix in the lower left consisting of N columns and $M-N$ rows. All elements in this submatrix scale as $O(\epsilon)$. Then it is easily verified that the right-hand side of Eq. (43) is $O(\epsilon)$, assuming, as we have before, that the derivative operating on M does not increase the magnitude of any of the matrix elements in an ϵ scaling sense. An integration of Eq. (40) over all solid angle then gives, neglecting the derivative term,

$$(\Lambda-\mathsf{L}\mathsf{M}^{-1}\Sigma_s\mathsf{M})\mathsf{L}\eta^{(1)}=\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}_0+\mathsf{L}\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{U}\eta^{(0)}\,, \tag{44}$$

where we have defined

$$\mathbf{Q}_0=\int_{4\pi}d\Omega\,\mathbf{Q}(\Omega)\,. \tag{45}$$

The matrix H, defined as

$$\mathsf{H}=\Lambda-\mathsf{L}\mathsf{M}^{-1}\Sigma_s\mathsf{M}\,, \tag{46}$$

is nonsingular, and thus Eq. (44) yields

$$\mathsf{L}\eta^{(1)}=\mathsf{H}^{-1}\mathsf{L}(\mathsf{M}^{-1}\mathbf{Q}_0+\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{U}\eta^{(0)})\,. \tag{47}$$

We use this result as follows. We return to Eq. (36) and left-multiply by U to obtain

$$\mathsf{U}\mathsf{M}^{-1}\frac{d(\mathsf{M}\chi^{(0)})}{ds}+\Lambda\mathsf{U}\chi^{(0)}$$
$$=\mathsf{U}\mathsf{M}^{-1}\mathbf{Q}+(1/4\pi)\mathsf{U}\mathsf{M}^{-1}\Sigma_s\mathsf{M}(\mathsf{U}\eta^{(0)}+\mathsf{L}\eta^{(1)})\,. \tag{48}$$

Substituting Eq. (47) for $\mathsf{L}\eta^{(1)}$ into Eq. (48) yields the significant result

$$\mathsf{U}\mathsf{M}^{-1}\frac{d(\mathsf{M}\chi^{(0)})}{ds}+\Lambda\mathsf{U}\chi^{(0)}$$
$$=\mathsf{U}\mathsf{M}^{-1}\mathbf{Q}+(1/4\pi)\mathsf{U}\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{H}^{-1}\mathsf{L}(\mathsf{M}^{-1}\mathbf{Q}_0$$
$$+\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{U}\eta^{(0)})+(1/4\pi)\mathsf{U}\mathsf{M}^{-1}\Sigma_s\mathsf{M}\mathsf{U}\eta^{(0)}\,. \tag{49}$$

To obtain our final result, we use Eq. (42) in the derivative term in Eq. (49), and recognize that

$$\chi=\chi^{(0)}+O(\epsilon)\,. \tag{50}$$

Then we find, from Eqs. (39) and (49), the reduced set of equations (in the χ variable) corresponding to N thin fluids. These equations are

$$\mathsf{L}\chi = 0 + O(\epsilon) \tag{51}$$

and

$$\mathsf{U}\mathsf{M}^{-1}\frac{d(\mathsf{M}U\chi)}{ds} + \Lambda\mathsf{U}\chi = \mathsf{U}\mathsf{M}^{-1}\mathbf{Q} + (1/4\pi)\mathsf{U}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}_0 + (1/4\pi)\mathsf{U}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\Lambda\mathsf{U}\eta + O(\epsilon)\,, \tag{52}$$

where we have used the identity

$$\mathsf{I} + \mathsf{M}\mathsf{H}^{-1}\mathsf{L}\mathsf{M}^{-1}\mathbf{\Sigma}_s = \mathsf{M}\mathsf{H}^{-1}\mathsf{M}^{-1}\mathbf{\Sigma}\,. \tag{53}$$

Alternately, we can express this result in terms of the original variables in the problem by using Eqs. (15) and (17). We find

$$\mathsf{L}\mathsf{M}^{-1}\mathbf{\Psi} = 0 + O(\epsilon)\,, \tag{54}$$

and, if we again use Eq. (42) in the derivative term in Eq. (52),

$$\mathsf{U}\mathsf{M}^{-1}\frac{d\mathbf{\Psi}}{ds} + \mathsf{U}\mathsf{M}^{-1}\mathbf{\Sigma}\mathbf{\Psi} = \mathsf{U}\mathsf{M}^{-1}\mathbf{Q} + (1/4\pi)\mathsf{U}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}_0 + (1/4\pi)\mathsf{U}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\Lambda\mathsf{U}\mathsf{M}^{-1}\mathbf{\Phi} + O(\epsilon)\,, \tag{55}$$

with the matrix H given by

$$\mathsf{H} = (\mathsf{M}^{-1}\mathbf{\Sigma} - \mathsf{L}\mathsf{M}^{-1}\mathbf{\Sigma}_s)\mathsf{M}\,. \tag{56}$$

Equations (54) and (55), or equivalently Eqs. (51) and (52), represent the reduced set of equations corresponding to N thin fluids. Equation (54) represents $M - N$ algebraic equations and Eq. (55) represents N differential equations. The ensemble-averaged flux, $\langle\psi\rangle$, is given in terms of the solution to these equations by the simple expression

$$\langle\psi\rangle = \mathbf{P}^T\mathbf{\Psi}\,, \tag{57}$$

where $\mathbf{P}$ is an M-component column vector with all components one, and the superscript T means transpose, i.e., $\mathbf{P}^T$ is an M-component row vector with all components one.

We emphasize two items concerning our final result. First, these equations are symmetric with respect to the fluid indices. To use these equations, one needs only specify N, the number of thin fluids; one does not need to identify which fluids are thin and which are thick. Aside from physical considerations concerning how many fluids can legitimately be characterized as thick and thin, the choice of N is dictated by the amount of reduction one wishes to make in the full set of equations given by Eq. (4). The smaller the choice of N, the less differential equations need to be solved to obtain $\langle\psi\rangle$. Of course, if the physical problem is far from the asymptotic limit being considered, the accuracy of the result deteriorates as N becomes smaller. Second, recalling that Λ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}$, it is clearly seen from Eq. (52) that the Nth order reduction maintains exactly the N largest characteristic decay lengths of the full problem given by Eq. (4). In particular, for $N = 1$ the dominant (largest) decay length is preserved, which means that even in this lowest order approximation one obtains the proper deep-in exponential decay rate for a time-independent, source-free, purely absorbing problem with the σ_i and λ_{ij} spatially independent. We consider explicitly the renormalized transport equation associated with $N = 1$ later on in this paper. Prior to this, however, we give the necessary initial and boundary layer analyses to obtain the initial and boundary conditions on Eq. (55).

III. INITIAL AND BOUNDARY LAYER ANALYSES

In this section we obtain, via initial and boundary layer analyses, the initial and boundary conditions that apply to Eq. (55). Considering first the initial conditions, we write the initial conditions on the full set of equations given by Eq. (4), or equivalently in matrix notation by Eq. (13), as

$$\mathbf{\Psi}(\mathbf{r},\mathbf{\Omega},0) = \gamma(\mathbf{r},\mathbf{\Omega})\,, \tag{58}$$

where γ is the prescribed known initial data. In the initial layer (t near 0), one expects a rapid variation of $\mathbf{\Psi}$ with time. To account for this, we introduce a scaled time $\tau = t/\epsilon$, and then the convective derivative is written

$$\frac{d}{ds} = \frac{1}{\epsilon v}\frac{\partial}{\partial\tau} + \mathbf{\Omega}\cdot\mathbf{\nabla}\,. \tag{59}$$

We now assume that in the initial layer $\partial/\partial\tau = O(1)$, and we further assume that the spatial variation of γ is $O(1)$ so that $\mathbf{\Omega}\cdot\mathbf{\nabla}$ in Eq. (59) is $O(1)$. Introducing Eq. (59) into Eq. (32), the scaled transport equation for χ in the initial layer (away from the boundary layer) is given by

$$\mathsf{M}^{-1}\left(\frac{1}{\epsilon v}\frac{\partial}{\partial\tau} + \mathbf{\Omega}\cdot\mathbf{\nabla}\right)\mathsf{M}\chi_i + \Lambda\mathsf{U}\chi_i + \frac{1}{\epsilon}\Lambda\mathsf{L}\chi_i = \frac{1}{4\pi}\left(\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{U}\eta_i + \frac{1}{\epsilon}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{L}\eta_i\right) + \mathsf{M}^{-1}\mathbf{Q}\,. \tag{60}$$

Here we have subscripted both χ and η with an "i" to emphasize that Eq. (60) holds in the initial layer.

To lowest order in ϵ, it is clear from Eq. (60) that we have, with an error of $O(\epsilon)$,

$$\frac{1}{v}\mathsf{M}^{-1}\frac{\partial(\mathsf{M}\chi_i)}{\partial\tau} + \Lambda\mathsf{L}\chi_i = \frac{1}{4\pi}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{L}\eta_i\,. \tag{61}$$

We can remove M from under the $\partial/\partial\tau$ operator in Eq. (61) since $\partial\mathsf{M}/\partial t$ has consistently been assumed to be $O(1)$, and hence $\partial\mathsf{M}/\partial\tau$ is $O(\epsilon)$. That is, in the initial layer analysis, we properly treat all properties of the ith fluid, namely σ_i, σ_{si}, and the λ_{ij}, as time independent, equal to their values at $t = 0$. Then left-multiplication of Eq. (61) by U gives

$$\frac{1}{v}\frac{\partial(\mathsf{U}\chi_i)}{\partial\tau} = \frac{1}{4\pi}\mathsf{U}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{L}\eta_i\,. \tag{62}$$

From Eqs. (15) and (58), we deduce that the initial condition on Eq. (62) is

$$\mathsf{U}\chi_i(0) = \mathsf{U}\mathsf{M}^{-1}\gamma\,. \tag{63}$$

To deal with the right-hand side of Eq. (62), we return to Eq. (61) and left-multiply by L. This yields a closed set of equations for $\mathsf{L}\chi_i$ given by

$$\frac{1}{v}\frac{\partial(\mathsf{L}\chi_i)}{\partial\tau} + \Lambda\mathsf{L}\chi_i = \frac{1}{4\pi}\mathsf{L}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{L}\eta_i\,, \tag{64}$$

and integration over all solid angle gives

$$\frac{1}{v}\frac{\partial(\mathsf{L}\eta_i)}{\partial\tau} + \mathsf{H}(\mathsf{L}\eta_i) = 0\,, \tag{65}$$

where the matrix H is given by Eq. (46). The initial condition on Eq. (65) again follows from Eqs. (15) and (58) as

$$\mathsf{L}\eta_i(0) = \mathsf{LM}^{-1}\gamma_0\,, \tag{66}$$

where

$$\gamma_0 = \int_{4\pi} d\mathbf{\Omega}\,\gamma(\mathbf{\Omega})\,. \tag{67}$$

Since the matrix H is nonsingular, we can apply H^{-1} to Eq. (65). Using the result for $\mathsf{L}\eta_i$ in Eq. (62) gives

$$\frac{\partial}{\partial\tau}(\mathsf{U}\chi_i) = -\frac{1}{4\pi}\frac{\partial}{\partial\tau}(\mathsf{UM}^{-1}\mathbf{\Sigma}_s\mathsf{MH}^{-1}\mathsf{L}\eta_i)\,. \tag{68}$$

Integration of Eq. (68) over $0\leqslant\tau<\infty$, using Eq. (63) for $\mathsf{U}\chi_i(0)$, and recognizing that $\mathsf{L}\eta_i$ vanishes as τ increases without bound [see Eqs. (65) and (66)], we find that $\mathsf{U}\chi_i(\infty)$ is given by

$$\mathsf{U}\chi_i(\infty) = \mathsf{UM}^{-1}\gamma + \frac{1}{4\pi}\mathsf{UM}^{-1}\mathbf{\Sigma}_s\mathsf{MH}^{-1}\mathsf{L}\eta_i(0)\,. \tag{69}$$

This large time initial layer result must match with the small time interior solution, and this asymptotic matching requirement gives the initial conditions on the reduced set of (interior) equations derived in the last section. These conditions are, using Eq. (66) for $\mathsf{L}\eta_i(0)$,

$$\mathsf{U}\chi(0) = \mathsf{UM}^{-1}\gamma + (1/4\pi)\mathsf{UM}^{-1}\mathbf{\Sigma}_s\mathsf{MH}^{-1}\mathsf{LM}^{-1}\gamma_0\,. \tag{70}$$

Finally, using Eq. (15) which relates χ to $\mathbf{\Psi}$, we find the proper initial conditions on the reduced set of equations given by Eq. (55). Recalling that our analysis here is to lowest order in ϵ, we have

$$\begin{aligned}\mathsf{UM}^{-1}&\mathbf{\Psi}(\mathbf{r},\mathbf{\Omega},0)\\ &= \mathsf{UM}^{-1}\gamma(\mathbf{r},\mathbf{\Omega}) + \frac{1}{4\pi}\mathsf{UM}^{-1}\mathbf{\Sigma}_s\mathsf{MH}^{-1}\mathsf{LM}^{-1}\gamma_0(\mathbf{r})\\ &\quad + O(\epsilon)\,,\end{aligned} \tag{71}$$

where all elements of the various matrices in Eq. (71) are to be evaluated at $t=0$.

We note several items concerning Eq. (71). First, these initial conditions contain an error of $O(\epsilon)$, which is consistent with the error in the differential equations to which they apply [see Eq. (55)]. Second, these initial conditions apply at each spatial point **r**; the variable **r** is simply a parameter in Eq. (71). Third, the original initial data $\gamma(\mathbf{r},\mathbf{\Omega})$ is explicitly contained in Eq. (71), but so is $\gamma_0(\mathbf{r})$, the angular integral of $\gamma(\mathbf{r},\mathbf{\Omega})$. The appearance of γ_0 is to be expected since $\mathbf{Q}_0$, the angular integral of the external source **Q**, is contained in the reduced equations [see Eq. (55)], and an initial condition is equivalent to a delta function in time source at $t=0$. We see, however, that the relatively complex terms involving γ_0 in Eq. (71) and $\mathbf{Q}_0$ in Eq. (55) both vanish for a purely absorbing ($\sigma_{si}=0$) problem. Last, we note that Eq. (71) was derived by an asymptotic matching of $\mathsf{U}\chi_i(\infty)$ to $\mathsf{U}\chi(0)$. We should also ensure that $\mathsf{L}\chi_i(\infty)$ matches to $\mathsf{L}\chi(0)$. This is easily seen to be the case. Equation (65) predicts that $\mathsf{L}\eta_i$ decays exponentially with τ, and from Eq. (64) we then deduce that $\mathsf{L}\chi_i$ also decays exponentially with τ, approaching zero as τ increases without bound. This indeed matches with the small time interior solution for $\mathsf{L}\chi$ which is identically zero [see Eq. (51)].

We now take up the boundary conditions on Eq. (55). We take the boundary conditions on the full set of equations to be

$$\mathbf{\Psi}(\mathbf{r}_s,\mathbf{\Omega},t) = \mathbf{\Gamma}(\mathbf{r}_s,\mathbf{\Omega},t)\,,\quad \mathbf{n}\cdot\mathbf{\Omega}>0\,, \tag{72}$$

where $\mathbf{\Gamma}$ describes the known incoming angular flux, and **n** is a local unit inward normal vector at a surface point $\mathbf{r}_s$. In the boundary layer (**r** near $\mathbf{r}_s$) we expect a rapid spatial variation of $\mathbf{\Psi}$ in a direction normal to the surface. If we introduce a local inward-pointing normal coordinate z, with $z=0$ corresponding to $\mathbf{r}=\mathbf{r}_s$, and further introduce a scaled variable $\tau = z/\epsilon$, the convective derivative is written

$$\frac{d}{ds} = \frac{1}{v}\frac{\partial}{\partial t} + \frac{\mu}{\epsilon}\frac{\partial}{\partial\tau} + (\mathbf{\Omega}\cdot\mathbf{\nabla})_p\,, \tag{73}$$

where $\mu = \mathbf{n}\cdot\mathbf{\Omega}$ and $(\mathbf{\Omega}\cdot\mathbf{\nabla})_p$ denotes the streaming operator in the plane perpendicular to z. We assume that in the boundary layer $\partial/\partial\tau = O(1)$. We further assume that the temporal and spatial variations of $\mathbf{\Gamma}$ along the surface as well as the local radius of curvature of the surface are $O(1)$. Then the time derivative and the perpendicular spatial gradient in Eq. (73) are $O(1)$. Thus if we introduce Eq. (73) into Eq. (32), the scaled transport equation in the boundary layer (away from the initial layer) is given by

$$\begin{aligned}\mathsf{M}^{-1}&\left[\frac{1}{v}\frac{\partial}{\partial t} + \frac{\mu}{\epsilon}\frac{\partial}{\partial\tau} + (\mathbf{\Omega}\cdot\mathbf{\nabla})_p\right]\mathsf{M}\chi_b\\ &+ \mathsf{AU}\chi_b + \frac{1}{\epsilon}\mathsf{AL}\chi_b\\ &= \mathsf{M}^{-1}\mathbf{Q} + \frac{1}{4\pi}\left(\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{MU}\eta_b + \frac{1}{\epsilon}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{ML}\eta_b\right).\end{aligned} \tag{74}$$

Here we have subscripted both χ and η with a "b" to emphasize that Eq. (74) holds in the boundary layer.

To lowest order in ϵ, Eq. (74) yields the equation, with an error of $O(\epsilon)$,

$$\mu\mathsf{M}^{-1}\frac{\partial(\mathsf{M}\chi_b)}{\partial\tau} + \mathsf{AL}\chi_b = \frac{1}{4\pi}\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{ML}\eta_b\,. \tag{75}$$

From this point on, the analysis closely parallels the initial layer analysis and we omit the algebraic details. We obtain as the boundary conditions on Eq. (55)

$$\begin{aligned}(\mathbf{n}\cdot\mathbf{\Omega})&\mathsf{UM}^{-1}\mathbf{\Psi}(\mathbf{r}_s,\mathbf{\Omega},t)\\ &= (\mathbf{n}\cdot\mathbf{\Omega})\mathsf{UM}^{-1}\mathbf{\Gamma}(\mathbf{r}_s,\mathbf{\Omega},t)\\ &\quad + (1/4\pi)\mathsf{UM}^{-1}\mathbf{\Sigma}_s\mathsf{MH}^{-1}\mathsf{LF}(\mathbf{r}_s,t)\\ &\quad + O(\epsilon)\,,\quad \mathbf{n}\cdot\mathbf{\Omega}>0\,.\end{aligned} \tag{76}$$

The various matrices in Eq. (76) are to be evaluated at the surface point $\mathbf{r}_s$ in question. Here the vector $\mathsf{LF}(\mathbf{r},t)$, which physically is a vector of particle currents, is defined by an angular integration according to

$$\mathsf{LF} = \int_{4\pi} d\mathbf{\Omega}\,\mu\mathsf{L}\chi_b\,. \tag{77}$$

The vector $\mathsf{L}\chi(\tau,\mathbf{\Omega},t)$ satisfies the canonical half-space "multigroup" albedo equation given by

$$\mu\frac{\partial(\mathsf{L}\chi_b)}{\partial\tau} + \mathsf{AL}\chi_b = \frac{1}{4\pi}\mathsf{LM}^{-1}\mathbf{\Sigma}_s\mathsf{ML}\eta_b\,,\quad 0\leqslant\tau<\infty\,, \tag{78}$$

with boundary conditions

$$\mathsf{L}\chi_b(0,\mathbf{\Omega},t) = \mathsf{L}\mathsf{M}^{-1}\mathbf{\Gamma}(\mathbf{r}_s,\mathbf{\Omega},t)\,, \quad \mathbf{n}\cdot\mathbf{\Omega} > 0\,, \tag{79}$$

$$\mathsf{L}\chi_b(\infty,\mathbf{\Omega},t) < \infty\,. \tag{80}$$

In the initial layer analysis, the equation analogous to Eq. (78), namely Eq. (64), could be solved explicitly and simply [see Eqs. (65)–(67)] for the required $\tau = 0$ term, $\mathsf{L}\boldsymbol{\eta}_i(0)$, in Eq. (69). Such simplicity is not the case here. In principle, but with algebraic complexity, Eqs. (78)–(80) can be solved via a Wiener–Hopf analysis, a singular eigenfunction technique,[19] or by employing invariance methods.[20] Then $\mathbf{F}(\mathbf{r}_s,t)$, needed in the boundary conditions given by Eq. (76), follows from Eq. (77) evaluated at $\mathbf{r} = \mathbf{r}_s$. For one thick fluid ($M - N = 1$), Eqs. (78)–(80) become scalar, and the solution of these equations is well known and relatively simple.[19,20] The surface quantity needed in Eq. (76), $\mathsf{L}\mathbf{F}(\mathbf{r}_s,t)$, is in this case a simple angular integral involving the incoming flux $\mathbf{\Gamma}$ and Chandrasekhar's well-known H function.[20] For two thick materials ($M - N = 2$), the necessary analysis of Eqs. (78)–(80) has been given by Siewert.[21] To our knowledge, the solution of Eqs. (78)–(80) for $M - N > 2$ is unavailable in the literature, but in principle can be found by one of the three methods alluded to above.

Finally, we make a few remarks concerning the boundary conditions given by Eq. (76) in the same vein as those we made for the initial conditions given by Eq. (71). These are: (1) The boundary conditions and the differential equations to which they apply are consistent in that both are in error by $O(\epsilon)$; (2) the time variable is simply a parameter in these boundary conditions; (3) these boundary conditions contain an isotropic component, namely $\mathsf{L}\mathbf{F}(\mathbf{r}_s,t)$, as is expected because of the close correspondence of an incident flux and a surface external source; and (4) asymptotic matching of the boundary layer and interior solutions for $\mathsf{U}\chi$ was used to derive the boundary conditions, and it is easily verified that the boundary layer and interior solutions for $\mathsf{L}\chi$ also match as required since both are zero in the matching region.

IV. A RENORMALIZED TRANSPORT EQUATION

In this section we consider a special case of our analysis corresponding to one thin fluid ($N = 1$). Then the reduced set of equations becomes scalar, and we can write our results as a renormalized transport equation. That is, we obtain an equation for $\langle\psi\rangle$, the ensemble-averaged flux, of the form given by Eq. (1), but with effective cross sections and an effective external source which account for the statistical nature of the problem in this asymptotic limit.

In this case, the matrix U has only one nonzero element, and for any column vector $\mathbf{v}$ and square matrix A we have the identities

$$\mathsf{U}\mathbf{v} = \mathbf{J}v_1\,; \quad \mathsf{U}\mathsf{A} = \mathsf{B}\,, \tag{81}$$

where $\mathbf{J}$ is an M-component column vector with all zero components except for the first component which is one, v_1 is the first component of $\mathbf{v}$, and all rows of the matrix B are zero except for the first row, which coincides with the first row of A. From the second equality in Eq. (81) we see that Eq. (52) contains only one nonzero equation. This can be isolated by left-multiplying Eq. (52) by $\mathbf{J}^T$. Making use of the first equality in Eq. (81) and $\mathbf{J}^T\mathsf{U} = \mathbf{J}^T$, we find

$$\begin{aligned}\mathbf{J}^T\mathsf{M}^{-1}\frac{d(\mathsf{M}\mathbf{J}\chi_1)}{ds} + \nu_1\chi_1 &= \mathbf{J}^T\mathsf{M}^{-1}\mathbf{Q} + (1/4\pi)\mathbf{J}^T\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}_0 \\ &\quad + (1/4\pi)(\mathbf{J}^T\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathbf{J})\nu_1\eta_1 + O(\epsilon)\,.\end{aligned} \tag{82}$$

To relate χ_1 to $\langle\psi\rangle$, the ensemble-averaged flux, we use

$$\langle\psi\rangle = \mathbf{P}^T\mathbf{\Psi} = \mathbf{P}^T\mathsf{M}\chi\,, \tag{83}$$

where $\mathbf{P}$ is a column vector with all components one. We rewrite Eq. (83) as

$$\langle\psi\rangle = \mathbf{P}^T\mathsf{M}(\mathsf{L} + \mathsf{U})\chi\,, \tag{84}$$

and using Eqs. (51) and (81) we deduce

$$\langle\psi\rangle = (\mathbf{P}^T\mathsf{M}\mathbf{J})\chi_1 + O(\epsilon)\,. \tag{85}$$

Using Eq. (85) for χ_1 in Eq. (82) gives

$$\begin{aligned}(\mathbf{P}^T\mathsf{M}\mathbf{J})\mathbf{J}^T\mathsf{M}^{-1}\frac{d}{ds}\left[\frac{\mathsf{M}\mathbf{J}}{(\mathbf{P}^T\mathsf{M}\mathbf{J})}\langle\psi\rangle\right] + \nu_1\langle\psi\rangle &= (\mathbf{P}^T\mathsf{M}\mathbf{J})(\mathbf{J}^T\mathsf{M}^{-1}\mathbf{Q}) \\ &\quad + (1/4\pi)(\mathbf{P}^T\mathsf{M}\mathbf{J})(\mathbf{J}^T\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}_0) \\ &\quad + (1/4\pi)(\mathbf{J}^T\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathbf{J})\nu_1\langle\phi\rangle + O(\epsilon)\,.\end{aligned} \tag{86}$$

Now, by examining the modal matrix M in the case of one thin fluid ($N = 1$), one can deduce

$$\frac{\mathsf{M}\mathbf{J}}{(\mathbf{P}^T\mathsf{M}\mathbf{J})} = \mathbf{c} + O(\epsilon)\,, \tag{87}$$

where $\mathbf{c}$ is a vector with only one nonzero component. This component, whose position is determined by which fluid is identified as thin, can be set to one by normalizing, to $O(\epsilon)$, the eigenvector corresponding to the smallest eigenvalue of $\mathbf{\Sigma}$ to unit length. Then the term given by the left-hand side of Eq. (87) can be taken outside the derivative in Eq. (86), and we arrive at the renormalized transport equation

$$\frac{d\langle\psi\rangle}{ds} + \sigma_{\text{eff}}\langle\psi\rangle = \frac{\sigma_{s,\text{eff}}}{4\pi}\langle\phi\rangle + S_{\text{eff}} + O(\epsilon)\,, \tag{88}$$

where

$$\sigma_{\text{eff}} = \nu_1\,, \tag{89}$$

$$\sigma_{s,\text{eff}} = (\mathbf{J}^T\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathbf{J})\nu_1\,, \tag{90}$$

$$\begin{aligned}S_{\text{eff}} &= (\mathbf{P}^T\mathsf{M}\mathbf{J})(\mathbf{J}^T\mathsf{M}^{-1}\mathbf{Q}) \\ &\quad + (1/4\pi)(\mathbf{P}^T\mathsf{M}\mathbf{J})(\mathbf{J}^T\mathsf{M}^{-1}\mathbf{\Sigma}_s\mathsf{M}\mathsf{H}^{-1}\mathsf{L}\mathsf{M}^{-1}\mathbf{Q}_0)\,.\end{aligned} \tag{91}$$

We emphasize that ν_1 in Eqs. (89) and (90) is the smallest eigenvalue of the matrix $\mathbf{\Sigma}$, and hence this renormalized equation possesses the exact deep-in characteristic decay length in the absence of sources (scattering and external) of the full transport description given by Eq. (4).

The expressions for $\sigma_{s,\text{eff}}$ and S_{eff} can be written in a somewhat more explicit form by considering the eigenvalue problem corresponding to the transpose of the matrix $\mathbf{\Sigma}$. We denote these eigenvectors by $\mathbf{y}$; the eigenvalues will again be ν, the eigenvalues defined by Eq. (19). Thus we have the eigenvalue problem

$$\mathbf{\Sigma}^T\mathbf{y} = \nu\mathbf{y}\,. \tag{92}$$

If the eigenvectors $\mathbf{x}$ and $\mathbf{y}$ are normalized such that

$$\mathbf{y}^T\mathbf{x} = 1\,, \tag{93}$$

it is well known that the matrix M^*, the modal matrix corresponding to $\boldsymbol{\Sigma}^T$ (the columns of M^* are the eigenvectors $\mathbf{y}$), is related to the modal matrix M by

$$\mathsf{M}^{-1} = (\mathsf{M}^*)^T. \tag{94}$$

Then Eqs. (90) and (91) can be rewritten as

$$\sigma_{s,\text{eff}} = \frac{(\mathbf{y}_1^T\boldsymbol{\Sigma}_s\mathsf{MH}^{-1}\mathsf{M}^{-1}\mathbf{x}_1)\nu_1}{(\mathbf{y}_1^T\mathbf{x}_1)}, \tag{95}$$

$$S_{\text{eff}} = \frac{(\mathbf{P}^T\mathbf{x}_1)(\mathbf{y}_1^T\mathbf{Q})}{(\mathbf{y}_1^T\mathbf{x}_1)} + \frac{1}{4\pi}\left[\frac{(\mathbf{P}^T\mathbf{x}_1)(\mathbf{y}_1^T\boldsymbol{\Sigma}_s\mathsf{MH}^{-1}\mathsf{LM}^{-1}\mathbf{Q}_0)}{(\mathbf{y}_1^T\mathbf{x}_1)}\right], \tag{96}$$

where $\mathbf{x}_1$ and $\mathbf{y}_1$ are the eigenvectors of Eqs. (19) and (92) corresponding to the smallest eigenvalue ν_1.

The initial and boundary conditions for the renormalized transport equation given by Eq. (88) follow by similar manipulations of Eqs. (71) and (76) as

$$\langle\psi(\mathbf{r},\boldsymbol{\Omega},0)\rangle = \frac{(\mathbf{P}^T\mathbf{x}_1)[\mathbf{y}_1^T\boldsymbol{\gamma}(\mathbf{r},\boldsymbol{\Omega})]}{(\mathbf{y}_1^T\mathbf{x}_1)} + \frac{1}{4\pi}\left[\frac{(\mathbf{P}^T\mathbf{x}_1)[\mathbf{y}_1^T\boldsymbol{\Sigma}_s\mathsf{MH}^{-1}\mathsf{LM}^{-1}\boldsymbol{\gamma}_0(\mathbf{r})]}{(\mathbf{y}_1^T\mathbf{x}_1)}\right] + O(\epsilon)\,, \tag{97}$$

and

$$(\mathbf{n}\cdot\boldsymbol{\Omega})\langle\psi(\mathbf{r}_s,\boldsymbol{\Omega},t)\rangle = (\mathbf{n}\cdot\boldsymbol{\Omega})\left[\frac{(\mathbf{P}^T\mathbf{x}_1)[\mathbf{y}_1^T\boldsymbol{\Gamma}(\mathbf{r}_s,\boldsymbol{\Omega},t)]}{(\mathbf{y}_1^T\mathbf{x}_1)}\right] + \frac{1}{4\pi}\left[\frac{(\mathbf{P}^T\mathbf{x}_1)[\mathbf{y}_1^T\boldsymbol{\Sigma}_s\mathsf{MH}^{-1}\mathsf{L}\mathbf{F}(\mathbf{r}_s,t)]}{(\mathbf{y}_1^T\mathbf{x}_1)}\right] + O(\epsilon)\,,\quad \mathbf{n}\cdot\boldsymbol{\Omega}>0\,. \tag{98}$$

Unfortunately, in this asymptotic limit of a single thin fluid ($N=1$) which leads, as we have seen, to a scalar transport description, one must solve a matrix canonical transport problem defined by Eqs. (78) through (80) to obtain the vector $\mathbf{F}$ needed in Eq. (98) to specify the boundary conditions. This matrix transport problem is, on one hand, complex in that it involves $M-1$ coupled transport equations. On the other hand, it is simple in the sense that it is a time-independent, source-free halfspace problem whose coefficients (σ_i, σ_{si}, and λ_{si}) are spatially independent.

To make contact with earlier work,[15] we consider the renormalized transport equation just derived in the simplest case of a binary mixture ($M=2$). In this case we have only two Markov transition probabilities and we simplify the notation somewhat by setting $\lambda_{12}=\lambda_1$ and $\lambda_{21}=\lambda_2$. Then the eigenvalues satisfy [see Eq. (20)]

$$\nu^2 - \left(\sigma_1+\sigma_2+\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right)\nu + \left(\sigma_1\sigma_2+\frac{\sigma_2}{\lambda_1}+\frac{\sigma_1}{\lambda_2}\right) = 0\,. \tag{99}$$

Since σ_{eff} is the smallest eigenvalue [see Eq. (89)], we have

$$\sigma_{\text{eff}} = \frac{1}{2}\left(\sigma_1+\sigma_2+\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right) - \frac{1}{2}\left[\left(\sigma_1+\sigma_2+\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right)^2 - 4\left(\sigma_1\sigma_2+\frac{\sigma_2}{\lambda_1}+\frac{\sigma_1}{\lambda_2}\right)\right]^{1/2}. \tag{100}$$

An asymptotically equivalent expression for σ_{eff} can be found by neglecting the ν^2 term in Eq. (99) when searching for the $O(1)$ eigenvalue since this term is $O(\epsilon)$ compared to the other terms in Eq. (99). This gives

$$\sigma_{\text{eff}} = \frac{(\sigma_1\sigma_2+\sigma_2/\lambda_1+\sigma_1/\lambda_2)}{(\sigma_1+\sigma_2+1/\lambda_1+1/\lambda_2)}. \tag{101}$$

Neither of these two expressions for σ_{eff} agree with the result reported earlier[15] for a binary mixture. This earlier result is (for homogeneous statistics)

$$\sigma_{\text{eff}} = \frac{(\sigma_1\sigma_2+\sigma_2/\lambda_1+\sigma_1/\lambda_2)}{[(\lambda_1\sigma_2+\lambda_2\sigma_1)/(\lambda_1+\lambda_2)+1/\lambda_1+1/\lambda_2]}. \tag{102}$$

We see that all three expressions for σ_{eff} are symmetric in the indices, so that one does not need to specify which fluid is thin. However, we also see that all three expressions are different. Nevertheless, they are all asymptotically equivalent as they must be. That is, if we explicitly identify one of the two fluids, say fluid 1, as thin, all three expressions for σ_{eff} yield the common result

$$\sigma_{\text{eff}}(1\ \text{thin}) = \frac{\sigma_1\lambda_1+\sigma_2\lambda_2+\sigma_1\sigma_2\lambda_1\lambda_2}{\lambda_1(1+\sigma_2\lambda_2)} + O(\epsilon)\,. \tag{103}$$

In general, Eq. (100) is to be preferred over Eqs. (101) and (102) in that it is most robust; it gives the exact deep-in decay length, as discussed earlier, no matter how far the physical problem is from the asymptotic limit under consideration.

To compute the effective scattering cross section and external source as given by Eqs. (95) and (96), we need to find the eigenvectors $\mathbf{x}_i$ and $\mathbf{y}_i$, $i=1,2$, according to Eqs. (19) and (92), construct M and M^{-1}, and perform the matrix multiplications indicated in Eqs. (95) and (96). The resulting expressions for $\sigma_{s,\text{eff}}$ and S_{eff} are algebraically complex and will not be given here. We point out, however, that the earlier expressions reported for $\sigma_{s,\text{eff}}$ and S_{eff} (see Ref. 15) differ from the expressions that result from Eqs. (95) and (96). However, these two sets of expressions, both of which are invariant under the interchange of fluid indices, are asymptotically equivalent. If we identify fluid 1 as the thin fluid and define $\sigma_{ai}=\sigma_i-\sigma_{si}$, both sets of expressions predict

$$\sigma_{s,\text{eff}}(1\ \text{thin}) = \frac{\sigma_{s1}\lambda_1(1+\sigma_{a2}\lambda_2)+\sigma_{s2}\lambda_2/(1+\sigma_2\lambda_2)}{\lambda_1(1+\sigma_{a2}\lambda_2)} + O(\epsilon) \tag{104}$$

and

$$S_{\text{eff}}(1\ \text{thin}) = p_1S_1(\boldsymbol{\Omega}) + p_2S_2(\boldsymbol{\Omega})/(1+\sigma_2\lambda_2) + \frac{\sigma_{s2}\lambda_2p_2}{4\pi(1+\sigma_2\lambda_2)(1+\sigma_{a2}\lambda_2)}\int_{4\pi} d\boldsymbol{\Omega}'\,S_2(\boldsymbol{\Omega}') + O(\epsilon)\,. \tag{105}$$

Thus the situation for $\sigma_{s,\text{eff}}$ and S_{eff} is the same as for σ_{eff}. Different asymptotic treatments yield algebraically different results, but these different results are all asymptotically equivalent, differing from each other by $O(\epsilon)$.

ACKNOWLEDGMENTS

The work of the first author (F.M.) was partially supported by the Los Alamos National Laboratory and the work of the second author (G.C.P.) was partially supported by the United States Department of Energy.

[1] C. D. Levermore, G. C. Pomraning, D. L. Sanzo, and J. Wong, J. Math. Phys. **27**, 2526 (1986).

[2] D. Vanderhaegen, J. Quant. Spectros. Radiat. Transfer **36**, 557 (1986).

[3] D. Vanderhaegen, J. Quant. Spectros. Radiat. Transfer **39**, 333 (1988).

[4] C. D. Levermore, G. C. Pomraning, and J. Wong, J. Math. Phys. **29**, 995 (1988).

[5] G. C. Pomraning, J. Quant. Spectros. Radiat. Transfer **40**, 479 (1988).

[6] G. C. Pomraning, Transport Theory Stat. Phys. **17**, 595 (1988).

[7] D. Vanderhaegen and C. Deutsch, J. Stat. Phys. **54**, 331 (1989).

[8] G. C. Pomraning, J. Quant. Spectros. Radiat. Transfer **41**, 103 (1989).

[9] G. C. Pomraning, C. D. Levermore, and J. Wong, *Lecture Notes in Pure and Applied Mathematics*, Vol. 115, edited by P. Nelson, V. Faber, T. Manteuffel, D. Seth, and A. White (Marcel Dekker, New York, 1989), pp. 1–35.

[10] D. C. Sahni, Ann. Nucl. Energy **16**, 397 (1989).

[11] D. C. Sahni, J. Math. Phys. **30**, 1554 (1989).

[12] G. C. Pomraning, J. Quant. Spectros. Radiat. Transfer **42**, 279 (1989).

[13] M. L. Adams, E. W. Larsen, and G. C. Pomraning, J. Quant. Spectros. Radiat. Transfer **42**, 253 (1989).

[14] R. Sanchez, J. Math. Phys. **30**, 2498 (1989).

[15] F. Malvagi, C. D. Levermore, and G. C. Pomraning, Transport. Theory Stat. Phys. **18**, 287 (1989).

[16] C. D. Levermore, "Transport in randomly mixed media with inhomogeneous anisotropic statistics," in preparation.

[17] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).

[18] P. Switzer, Am. Math. Stat. **36**, 1859 (1965).

[19] K. M. Case and P. F. Zweifel, *Linear Transport Theory* (Addison–Wesley, Reading, MA, 1967).

[20] S. Chandrasekhar, *Radiative Transfer* (Dover, New York, 1960).

[21] C. E. Siewert and Y. Ishiguro, J. Nucl. Energy **26**, 251 (1972).

Propagators for relativistic systems with non-Abelian interactions

R. A. Corns
Instituut voor Theoretische Fysica, Universiteit Leuven, B-3030 Leuven, Belgium
T. A. Osborn
Department of Physics, University of Manitoba, Winnipeg MB R3T 2N2, Canada

(Received 9 June 1989; accepted for publication 1 November 1989)

The relativistic evolution of a system of particles in the proper-time Schwinger–DeWitt formalism is investigated. For a class of interactions that can be represented as Fourier transforms of bounded complex matrix-valued measures, a Dyson series representation of the propagator is obtained. This class of interactions is non-Abelian and includes both external electromagnetic and Yang–Mills fields. The study of the relativistic problem is facilitated by embedding the original quantum evolution into a larger class of evolution problems that result if one makes an analytic continuation of the metric tensor $g_{\mu\nu}$. This continuation is chosen so that the extended propagator shares (for all signatures of $g_{\mu\nu}$) the Gaussian decay properties typical of heat kernels. Estimates of the nth-order Dyson iterate kernels are found that ensure the absolute convergence of the perturbation series. In this fashion a number of analytic and smoothness properties of the propagator are determined. In particular, it is demonstrated that the convergent Dyson series representation constructs a fundamental solution of the equations of motion.

I. INTRODUCTION

A Hamiltonian suitable for relativistic quantum evolution of a finite number of particles interacting via non-Abelian fields is

$$H(x,\tau) = (1/2m)g_{\mu\nu}(p^\mu - a^\mu(x,\tau))(p^\nu - a^\nu(x,\tau)) + v(x,\tau)\,. \tag{1.1}$$

Here τ is to be identified with the proper time, $x\in\mathbb{R}^d$ is a generic d-dimensional point that describes the space-time coordinates of the system, and m is a mass parameter. The metric tensor $g_{\mu\nu}$ is assumed to be the x, τ independent diagonal matrix appropriate for special relativity. Summation with respect to the repeated Lorentz indices μ or ν is always implied. Let $x = (x^1,x^2,...,x^d)$ be the contravariant representation of x. Then the differential operator $p^\mu = -i\hbar\,\partial/\partial x_\mu$ denotes the momentum conjugate to $x_\mu = g_{\mu\nu}x^\nu$. For one-body problems, a^μ and v will generally be functions of all the components of x. In an N-body application, the components of a^μ and v associated with a given particle will depend on just the variables in $\mathbb{R}^d$ that describe the space-time location of that particle.

The internal degrees of freedom in this problem are carried by the vector space $\mathbb{C}^s$. For example, if the system has N particles and if the ith particle has spin or isospin n_i, then the dimension of $\mathbb{C}^s$ is

$$s = \sum_{i=1}^{N}(2n_i + 1)\,. \tag{1.2}$$

The associated quantum states $\psi(\tau)$ are elements of the Hilbert space $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$. For each possible x, τ, and μ the values of the external field $a^\mu(x,\tau)$ and interparticle field $v(x,\tau)$ are bounded (noncommuting) linear operators on the space $\mathbb{C}^s$. In many applications[1,2] $a^\mu(x,\tau)$ are Hermitian operators defined by a sum of SU(n) matrices. In this latter case the parts of $a^\mu(x,\tau)$ proportional to the identity I (on $\mathbb{C}^s$) describe the electromagnetic field potentials while the remaining elements of the sum define Yang–Mills interactions.

The equation of motion for the propagator K is the Schwinger–DeWitt proper-time realization[3–5] of the Schrödinger equation,

$$i\hbar\frac{\partial}{\partial\tau}K(x,\tau;y,\tau_0) = H(x,\tau)K(x,\tau;y,\tau_0)\,, \tag{1.3a}$$

together with the delta-function initial condition at the time τ_0,

$$\lim_{\tau\to\tau_0+} K(x,\tau;y,\tau_0) = \delta(x-y)I\,. \tag{1.3b}$$

For a class of analytic vector and scalar fields a^μ and v defined as Fourier transforms of τ-dependent measures, the principal goal of this paper is to obtain explicit solutions of the proper-time Schrödinger initial-value problem (1.3a) and (1.3b). These solutions of (1.3a) will take the form of a convergent kernel-valued Dyson[6] series. This investigation establishes many of the analytic properties of the propagator K in the variables x, y, τ, τ_0 and in the physical parameters of interest—the mass m and Planck's constant $\hbar$.

The subsequent analysis in this paper will interpret (1.3a) as an s-component system of classical partial differential equations (PDE's) and find its fundamental solution $K(x,\tau;y,\tau_0)$ via a pointwise convergent infinite series. However, much of the motivation for our treatment stems from the abstract $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$ analog to the PDE (1.3a). Specifically, let the pair $H(\tau)$ and H_0 denote the closed operators on $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$ defined, respectively, by $H(x,\tau)$ and its field-free ($a^\mu = v = 0$) restriction. As a consequence the perturbing operator $W(\tau)$, connecting $H(\tau)$ to H_0, is formally defined by

$$H(\tau) = H_0 + W(\tau)\,. \tag{1.4a}$$

Acting on sufficiently smooth $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$ functions, $W(\tau)$

0022-2488/90/040901-15$03.00

becomes

$$W(\tau) = -(1/m)a^{\mu}(\cdot,\tau)p_{\mu} - (1/2m)[(p_{\mu}a^{\mu})(\cdot,\tau) - a_{\mu}(\cdot,\tau)a^{\mu}(\cdot,\tau)] + v(\cdot,\tau)\,. \tag{1.4b}$$

The first term on the right side of (1.4b) is a linear differential operator while the remaining factors are a multiplication operator.

The $\mathscr{L}^2(\mathbb{R}^d, \mathbb{C}^s)$-space evolution problem[7] is to determine the solution $\psi(\tau)$ satisfying

$$i\hbar\frac{d}{d\tau}\psi(\tau) = H(\tau)\psi(\tau)\,, \tag{1.5a}$$

$$\psi(\tau_0) = \psi_0\,, \tag{1.5b}$$

where ψ_0 is some initial data function lying in the domain of $H(\tau_0)$.

The evolution operator $U(\tau,\tau_0)$ associated with the problem (1.5a) and (1.5b) is the linear operator mapping ψ_0 to $\psi(\tau)$, i.e.,

$$\psi(\tau) = U(\tau,\tau_0)\psi_0\,. \tag{1.6a}$$

The statement that the operator $U(\tau,\tau_0)$ is an integral operator whose kernel is the propagator K provides the bridge between the operator and the pointwise description of the evolution problem, namely,

$$(U(\tau,\tau_0)\psi_0)(x) = \int K(x,\tau;y,\tau_0)\psi_0(y)dy\,. \tag{1.6b}$$

The familiar Dyson expansion[8–10] for $U(\tau, \tau_0)$ is a consequence of attempting to solve the integral equation equivalent to (1.5a) and (1.5b) by iteration. The resulting formal series for $U(\tau,\tau_0)$ then reads

$$U(\tau,\tau_0) = \sum_{n=0}^{\infty} D_n(\tau,\tau_0)\,, \tag{1.7a}$$

where the nth Dyson iterate is

$$D_n(\tau,\tau_0) = \left(\frac{1}{i\hbar}\right)^n \int_< d\boldsymbol{\tau}_n\, U_0(\tau,\tau_n) \times W(\tau_n)U_0(\tau_n,\tau_{n-1}) \times \cdots \times W(\tau_1)U_0(\tau_1,\tau_0). \tag{1.7b}$$

Here $U_0(\tau,\tau_0)$ is the evolution operator for the free Hamiltonian H_0. The variable $\boldsymbol{\tau}_n$ is $(\tau_1, \tau_2,...,\tau_n)$ and the integral subscript $<$ denotes the n-dimensional time-ordered domain $\tau_0 \leqslant \tau_1 \leqslant \cdots \leqslant \tau_n \leqslant \tau$.

One aspect of our method for investigating the solutions of (1.3a) and (1.3b) is to embed the original problem in a larger family of evolution problems that results if one makes an analytic continuation of the metric tensor g. In this way the original problem may be viewed as the boundary value of this enlarged class of solutions. The complex-valued extended metric tensor and the related scalar products are defined as follows. The diagonal matrix g has eigenvalues $\{e_{\mu}\}$ that are restricted to be either $+1$ or -1. Let σ_+ and σ_- represent the number of positive and negative eigenvalues, respectively. The signature of g is then $\sigma_+ - \sigma_-$. The real parameter $-\infty < \epsilon < \infty$ will identify the extended matrix $g(\epsilon)$ and the value $\epsilon = 0$ will specify the original metric tensor, i.e., $g_{\mu\nu}(0) = g_{\mu\nu}$. The $d \times d$ matrix $g(\epsilon)$ is diagonal and is obtained from g by replacing all the positive eigenvalues (diagonal entries) of g by $(1 - i\epsilon)/(1 + i\epsilon)$ and the negative eigenvalues by $-(1 + i\epsilon)/(1 - i\epsilon)$. A notable advantage of this parametrization of the extended problem is that $g(\epsilon)$ is a unitary transformation satisfying

$$g(\epsilon)^{-1} = g(\epsilon)^{\dagger} = g(-\epsilon)\,, \tag{1.8}$$

where † represents the adjoint on the space $\mathbb{C}^d$.

A convenient notation for the scalar product of two $\mathbb{C}^d$-valued vectors v and w is

$$\langle v,w\rangle = g_{\mu\nu}v^{\mu}w^{\nu}\,. \tag{1.9a}$$

If v and w are real vectors, then (1.9a) is the usual indefinite scalar product associated with g; if $\sigma_- = 1$ and $\sigma_+ = 3$, then (1.9a) is the Lorentz scalar product in Minkowski spacetime. For $g(\epsilon)$ we define a symmetric bilinear product

$$\langle v,w\rangle_{\epsilon} = g_{\mu\nu}(\epsilon)v^{\mu}w^{\nu}\,. \tag{1.9b}$$

This extended scalar product has an ϵ-independent estimate

$$|\langle v,w\rangle_{\epsilon}| \leqslant |v|\,|w|\,, \tag{1.9c}$$

where $|\cdot|$ is the Euclidean distance norm on $\mathbb{C}^d$. Conclusion (1.9c) follows from the fact that $g(\epsilon)$ is unitary. As a reasonable abuse of notation, we shall refer to $g(\epsilon)$ as an extended metric tensor in spite of the fact that for $\epsilon \neq 0$, $g(\epsilon)$ is non-Hermitian.

The ϵ-extended problem is defined when $g_{\mu\nu}$ in the Hamiltonian (1.1) is replaced by $g_{\mu\nu}(\epsilon)$. The corresponding modified Hamiltonian is the differential operator

$$H(x,\tau;\epsilon) = (1/2m)\langle p - a(x,\tau), p - a(x,\tau)\rangle_{\epsilon} + v(x,\tau)\,. \tag{1.10}$$

In order to understand the motivation for introducing the particular extension of the metric tensor described above it suffices to examine the analytic form of the free propagator in the variable ϵ. The free evolution kernel is the solution of

$$i\hbar\frac{\partial}{\partial\tau}K_0(x,\tau;y,\tau_0;\epsilon) = -\frac{\hbar^2}{2m}g_{\mu\nu}(\epsilon)\frac{\partial}{\partial x^{\mu}}\frac{\partial}{\partial x^{\nu}}K_0(x,\tau;y,\tau_0;\epsilon)\,, \tag{1.11}$$

which satisfies the initial condition (1.3b). The explicit solution reads

$$K_0(x,\tau;y,\tau_0;\epsilon) = N[\det g(-\epsilon)]^{1/2} \times \exp\{[im/2\hbar(\tau-\tau_0)]\langle x-y,x-y\rangle_{-\epsilon}\}I \tag{1.12}$$

where the normalization factor is $N = (m/[2\pi i\hbar(\tau-\tau_0)])^{d/2}$. Writing

$$\langle x-y,x-y\rangle_{-\epsilon} = \frac{1-\epsilon^2}{1+\epsilon^2}\langle x-y,x-y\rangle + i\frac{2\epsilon}{1+\epsilon^2}|x-y|^2\,, \tag{1.13}$$

it is seen that for forward evolution $\tau > \tau_0$ with $\epsilon > 0$ the function K_0 has Gaussian decay with respect to $|x-y|$. Of course as $\epsilon \to 0+$, K_0 becomes the Lorentz invariant scalar that defines the interaction-free relativistic propagator. The presence of Gaussian decay in the propagators mimics what happens in the heat equation.[11,12] This decay property means that the kernels K_0 and K define bounded operators on the spaces $\mathscr{L}^p(\mathbb{R}^d, \mathbb{C}^s)$, $1 \leqslant p \leqslant \infty$, whenever $\epsilon > 0$. Knowledge of the continuity properties of the kernels $K(x,\tau;y,\tau_0;\epsilon)$ in the variable ϵ has, in similar circumstances with $\sigma_+ = d$, allowed one to prove[12,13] that the $\epsilon \to 0+$ limit of this function is the correct integral kernel for the evolution operator $U(\tau,\tau_0)$ and, in particular, that the representation (1.6b) is valid.

In framing the mathematical definition of the evolution problem we have sought the widest generality consistent with our method of solution. For example, we allow $a^\mu(x,\tau)$ to be an arbitrary bounded linear operator on $\mathbb{C}^s$ rather than the Hermitian operators they usually are in applications. This serves to illustrate that our method does not require that $H(x, \tau)$ define, via extension on $\mathscr{L}^2(\mathbb{R}^d, \mathbb{C}^s)$, a self-adjoint operator $H(\tau)$. Similarly, metric tensors with any signature are permitted. The differential operator $H(x, \tau)$ is always second order but its type changes as σ_- varies. For $\sigma_- = 0$ (or d), it is elliptic and suitable for nonrelativistic quantum mechanics; if $\sigma_- = 1$ (or $d-1$), it is hyperbolic and describes Minkowski space-time; and finally if $\sigma_- \geqslant 2$ (or $\leqslant d-2$), it is ultrahyperbolic. This latter class of differential operator enters the relativistic theories[14-17] of Horwitz, Piron, and others if there are two or more particles. Our constructive solution is successful for all σ_-. In the cases where $\sigma_\pm$ differ from the Minkowski values, the Lorentz transformation is understood in the generalized sense of a representation of the de Sitter group $\mathrm{SO}(\sigma_+, \sigma_-)$ of transformations of $\mathbb{R}^d$ which leave the product $\langle x, x\rangle$ invariant.

In applications linked to the Klein–Gordon equation the parameter m equals 1 and the physical mass of the system, mc^2, enters as a part of the scalar potential $v(x, \tau)$. Problems in which the constituents have different mass values can easily be accommodated in our formalism by a suitable scaling of the space-time coordinates of the various particles.[18] We keep the value of m as an explicit variable in our representations since the analytic scaling in m provides one method of derivation for the derivative field approximations[19,20] to this system.

A Dyson series analysis similar to the one developed here was used to treat the scalar wave function ($s = 1$) problem for nonrelativistic Hamiltonians.[9] The new results found in this paper extend this constructive series method to include relativistic dynamics as well as non-Abelian interactions. One major difference in emphasis between the approach taken here and that found in Ref. 9 is the use of the abstract evolution theory. In Ref. 9 the theory of linear differential equations in Banach space[7] was used in parallel with the Dyson series method. The present set of results uses only a coordinate space kernel valued Dyson series description. The optimal union, in the relativistic context, of the abstract evolution theory and the pointwise characterization of the propagator given here will be the subject of further study.

In Sec. II the class of Fourier image interactions that serve to define the Hamiltonian $H(x, \tau; \epsilon)$ is described. In Sec. III the formulas that construct the nth Dyson iterate kernel are derived. In Sec. IV the convergence properties of the sum over n of the Dyson kernels is established and it is proved that the resultant kernel is a fundamental solution of the proper-time Schrödinger equation. In Sec. V our results are summarized for wave functions and it is verified that we have constructed a solution of the partial differential equation corresponding to (1.5a) and (1.5b).

II. ANALYTIC FIELDS AND MEASURES

The family of potential fields that define the Hamiltonians (1.1) and (1.10) are taken to be Fourier images of bounded measures. Specifically, $a = (a^1, a^2,\dots,a^d)$ and v are assumed to have the general form

$$a(x,\tau) = \int e^{i\langle x,\alpha\rangle}\, d\gamma(\alpha,\tau)\,, \tag{2.1}$$

$$v(x,\tau) = \int e^{i\langle x,\alpha\rangle}\, d\nu(\alpha,\tau)\,. \tag{2.2}$$

In the integrals (2.1) and (2.2) the quantity α denotes the (wave vector) variable of integration that runs over the domain $\mathbb{R}^d$. The evolution variable τ is assumed to take its values in the fixed time segment $[0, T]$. It is common to refer to a as a vector potential and v as a scalar potential. Of course, a and v determine, respectively, a vector in $\mathbb{C}^d$ and a scalar in $\mathbb{C}$ only after the expectation values with respect to a particular quantum state $\psi \in \mathscr{L}^2(\mathbb{R}^d, \mathbb{C}^s)$ have been computed. Similar potentials have been described in detail in Refs. 9 and 21; thus we give only enough detail to define these potentials fully and indicate some of the new structural features not found in these prior works.

The norm sign $|\cdot|$ is employed with several meanings. In the case where its argument is in $\mathbb{C}$, it denotes the absolute value; if its argument is a d-component multi-index $\phi = (\phi_1, \phi_2,\dots,\phi_d)$, then it is the sum $|\phi| = \phi_1 + \cdots + \phi_d$; if its argument is a matrix $A \in \mathbb{C}^{n\times m}$, then it implies the Euclidean norm

$$|A|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} |A_{ij}|^2\,; \tag{2.3}$$

and finally, if the argument is a measure, then it denotes the corresponding total variation measure. The appropriate meaning will follow from the context. For example, if $n = m = s$, then (2.3) is the norm we systematically use to estimate the values of the various kernels such as K_0 and K. Note that the identity matrix has a norm equal to $\sqrt{s}$.

Let $(\mathbb{C}^{s\times s})^r$ denote the space of r-tuples of $s\times s$ matrices over the field $\mathbb{C}$. In the present circumstances r is d for the vector potential a and it is 1 for the scalar potential v. Let the quantity ρ be an arbitrary element of the Banach space $\mathscr{M}(\mathbb{R}^d, (\mathbb{C}^{s\times s})^r)$ of $(\mathbb{C}^{s\times s})^r$-valued Borel measures on $\mathbb{R}^d$ having finite total variation norm. On $\mathscr{M}(\mathbb{R}^d, (\mathbb{C}^{s\times s})^r)$ the norm of ρ is defined via its associated total variation measure $|\rho|$. Let $\mathscr{B}$ represent the Borel measurable sets on $\mathbb{R}^d$. Label the countable partitions of $e \in \mathscr{B}$ by $\pi = \{e_i\}_{i=1}^{\infty}$. Then

$$|\rho|(e) = \sup_{\pi} \sum_{i=1}^{\infty} |\rho(e_i)|\,. \tag{2.4a}$$

The norm that makes the space $\mathscr{M}(\mathbb{R}^d, (\mathbb{C}^{s\times s})^r)$ into a Banach space is

$$\|\rho\| = |\rho|(\mathbb{R}^d) < \infty \; . \tag{2.4b}$$

We frequently use the polar decomposition of the measure ρ relative to the positive scalar measure $|\rho|$. This decomposition asserts (cf. Ref. 21, Lemma 1; Ref. 22, Theorem 6.12) that there is a Borel measurable function $\hat{\rho}$: $\mathbb{R}^d \to (\mathbb{C}^{s\times s})^r$, with Euclidean norm $|\hat{\rho}(\alpha)| = 1$, for all $\alpha \in \mathbb{R}^d$, such that

$$\int_e d\rho(\alpha) = \int_e \hat{\rho}(\alpha) d|\rho(\alpha)|, \quad e\in\mathscr{B} \; . \tag{2.5}$$

Let $S_k \subset \mathbb{R}^d$ be a closed ball of radius $k > 0$ and centered at the origin. The Banach subspace of $\mathscr{M}(\mathbb{R}^d, (\mathbb{C}^{s\times s})^r)$ consisting of those measures whose support is contained in S_k is denoted by $\mathscr{M}(S_k, (\mathbb{C}^{s\times s})^r)$. Next we consider measures suitable for the transforms (2.1) and (2.2). These are continuous one-variable functions on the time interval $[0, T]$ whose values are measures, namely,

$$\rho(\cdot): \; [0, T] \to \mathscr{M}(S_k, (\mathbb{C}^{s\times s})^r) \; . \tag{2.6}$$

With these definitions in place we state the hypotheses on the fields a and v that will be used throughout the remainder of the paper. Since these potential field hypotheses are always assumed they will not be cited as part of the ensuing lemmas and propositions.

Potential Class (A): Let $k < \infty$. The potentials a and v are said to be in class (A) if a and v are the Fourier images, Eqs. (2.1) and (2.2), of the time-dependent measures $\gamma(\tau)$ and $\nu(\tau)$ satisfying

(1) $\gamma(\tau) \in \mathscr{M}(S_{k/2}, (\mathbb{C}^{s\times s})^d)$, $\tau\in[0, T]$;
(2) $\nu(\tau) \in \mathscr{M}(S_k, \mathbb{C}^{s\times s})$, $\tau\in[0, T]$;
(3) both $\gamma(\cdot)$ and $\nu(\cdot)$ are continuous on $[0, T]$.

A simple but important fact is that the space $\mathscr{M}(S_k, (\mathbb{C}^{s\times s})^d)$ is invariant with respect to multiplication by the extended metric tensor. If $\gamma(\tau)$ is in $\mathscr{M}(S_k, (\mathbb{C}^{s\times s})^d)$, then the d-tuple $\{g_{\mu\nu}(\epsilon)\gamma^\nu(\tau)\}_{\mu=1}^d$ is a measure in $\mathscr{M}(S_k, (\mathbb{C}^{s\times s})^d)$. Further, observe that hypothesis (A) implies that $a(\cdot, \tau)$ and $v(\cdot, \tau)$ are, respectively, $(\mathbb{C}^{s\times s})^d$- and $\mathbb{C}^{s\times s}$-valued analytic functions of x. In earlier related work,[9,12,21] it was always assumed that a and v were either real-valued or Hermitian.

Expanding the Hamiltonian $H(x, \tau; \epsilon)$ leads to cross terms of the type

$$\langle p, a(x,\tau)\rangle_\epsilon = g_{\mu\nu}(\epsilon) p^\mu a^\nu(x,\tau) \; , \tag{2.7}$$

$$\langle a(x,\tau), a(x,\tau)\rangle_\epsilon = g_{\mu\nu}(\epsilon) a^\mu(x,\tau) a^\nu(x,\tau) \; . \tag{2.8}$$

Expressions for the functions (2.7) and (2.8) in terms of the measures $\gamma(\tau)$ and $\nu(\tau)$ are repeatedly used in the subsequent construction of the Dyson kernels. Consider (2.7) first. The definition (1.9b) of $\langle \cdot,\cdot \rangle_\epsilon$, the polar factorization (2.5), and the properties of class (A) imply

$$\langle p, a(x,\tau)\rangle_\epsilon = \int \langle \hbar\alpha, \hat{\gamma}(\alpha,\tau)\rangle_\epsilon e^{i\langle x,\alpha\rangle} \, d|\gamma|(\alpha,\tau) \; . \tag{2.9}$$

Next examine (2.8). The right side of this expression is structurally similar to $v(x, \tau)$ and so it is helpful to know the associated measure occurring in the Fourier description of this function. This is given by the following convolution measure:

$$\langle a(x,\tau), a(x,\tau)\rangle_\epsilon = \int e^{i\langle x,\alpha\rangle} \, d\langle \gamma(\tau) * \gamma(\tau)\rangle_\epsilon \; , \tag{2.10a}$$

where $\langle \gamma(\tau) * \gamma(\tau)\rangle_\epsilon$ is the measure in $\mathscr{M}(S_k, \mathbb{C}^{s\times s})$ defined by

$$\langle \gamma(\tau) * \gamma(\tau)\rangle_\epsilon(e) = \int \chi_e(\alpha + \alpha') \langle \hat{\gamma}(\alpha,\tau), \hat{\gamma}(\alpha',\tau)\rangle_\epsilon \times d(|\gamma(\tau)| \times |\gamma(\tau)|) \; . \tag{2.10b}$$

In this latter integral $|\gamma(\tau)| \times |\gamma(\tau)|$ is the two-dimensional product measure on $S_{k/2} \times S_{k/2}$ having the variable (α, α'). The function χ_e is the characteristic function for the Borel set e. The convolution operation in (2.10b) has the effect of enlarging the domain of support of the resultant measure. Provided that the support of $\gamma(\tau)$ is within $S_{k/2}$, the support of $\langle \gamma(\tau) * \gamma(\tau)\rangle_\epsilon$ lies within S_k. This behavior is our reason for requiring the measure spaces to have the support restrictions stated in conditions (1) and (2) of the class (A). Also, formulas (2.10a) and (2.10b) show that the process of choosing $\epsilon \neq 0$ leads to an ϵ dependence in the measure $\langle \gamma(\tau) * \gamma(\tau)\rangle_\epsilon$ but leaves unchanged the Lorentz invariant phase factor $e^{i\langle x,\alpha\rangle}$.

Finally it is convenient to define a measure $\mu(\tau)$ that constructs the p-independent part of the Hamiltonian $H(x, \tau; \epsilon)$. Upon setting

$$\mu(\tau) = (1/2m)\langle \gamma(\tau) * \gamma(\tau)\rangle_\epsilon + \nu(\tau) \; , \tag{2.11a}$$

it follows that

$$\frac{1}{2m}\langle a(x,\tau), a(x,\tau)\rangle_\epsilon + v(x,\tau) = \int e^{i\langle x,\alpha\rangle} \, d\mu(\alpha,\tau) \; . \tag{2.11b}$$

In the study of the convergence properties of the Dyson series, τ-uniform bounds of the vector and scalar potentials play a key role. For potentials in class (A) property (3) implies that $\|\gamma(\cdot)\|$ and $\|\nu(\cdot)\|$ are real continuous functions on $[0, T]$. Thus one may define the finite bounds

$$\gamma_T = \sup_{\tau\in[0,T]} \|\gamma(\tau)\|, \quad \nu_T = \sup_{\tau\in[0,T]} \|\nu(\tau)\| \; . \tag{2.12}$$

The quantities γ_T and ν_T provide [cf. the representations (2.1) and (2.2)] the uniform pointwise bounds

$$|a(x,\tau)| \leqslant \gamma_T, \quad |v(x,\tau)| \leqslant \nu_T \; , \tag{2.13}$$

for all $(x,\tau) \in \mathbb{R}^d \times [0, T]$. Furthermore, definition (2.11a) means that $\|\mu(\tau)\|$ has the bound

$$\|\mu(\tau)\| \leqslant (\gamma_T^2/2m) + \nu_T \equiv \mu_T \; . \tag{2.14}$$

Observe that γ_T, ν_T, and μ_T are ϵ independent.

The utilization of potentials whose Fourier images are measures for the study of dynamics has a long history. Their first use dates back to Ito's early work[23] on the Feynman path integral. Further extensive use of these potentials in more recent studies of the path integral may be found in Albeverio and Høegh-Krohn[24] as well as Cameron and Storvick.[25] These potentials have played a central role[9,12,21] in the study of the summability of Dyson series in various nonrelativistic problems. Of course, it is to be appreciated that the convergence properties of the Dyson series found in this paper are specific to the class of Fourier image potentials of class (A). For potentials that are not analytic or have unbounded values, the Dyson series is typically[10] an asymptotic rather than a convergent series.

III. DYSON KERNELS

This section contains an explicit computation of the formulas that characterize the integral kernels of the nth Dyson iterate $D_n(\tau, \tau_0; \epsilon)$ that arise in the evolution generated by $H(x, \tau; \epsilon)$. Let $Q = (x, \tau; y, \tau_0)$ denote the pair of final and initial space-time points. The kernel $d_n(Q; \epsilon)$ constructs $D_n(\tau, \tau_0; \epsilon)$ via

$$[D_n(\tau,\tau_0;\epsilon)\psi_0](x) = \int d_n(Q;\epsilon)\psi_0(y)dy. \qquad (3.1)$$

The representations found below completely determine the geometrical and analytical character of $d_n(Q; \epsilon)$. The computation is implemented by obtaining a mixed coordinate–momentum space representation for the operator $D_n(\tau, \tau_0; \epsilon)$ and then Fourier transforming this mixed form to find $d_n(Q; \epsilon)$. This kernel calculation is described informally, although it is not too difficult to recast the analysis in a mathematically rigorous fashion.[9,12,21] We proceed in this way since the subsequent treatment of the Dyson series in Secs. IV and V will require only the formulas for $d_n(Q; \epsilon)$ and is insensitive to the method used to find them. Nevertheless it is important to understand where the representations of $d_n(Q; \epsilon)$ come from and why they are correct.

As a first calculation we determine the mixed coordinate–momentum space kernels of $D_0(\tau, \tau_0; \epsilon)$ and $D_1(\tau, \tau_0; \epsilon)$. The operator $D_0(\tau, \tau_0; \epsilon)$ is, in fact, the free evolution operator

$$U_0(\tau,\tau_0;\epsilon) = \exp\{-i(\tau-\tau_0)H_0(\epsilon)/\hbar\}.$$

The $H_0(\epsilon)$ evolution of a plane wave state is given by the generalized-function identity

$$\begin{aligned}&\exp\{-[i(\tau-\tau_0)/\hbar]H_0(\epsilon)\}\exp(i\langle x,\alpha\rangle)\\ &\quad= \exp\{-[i\hbar(\tau-\tau_0)/2m]\langle\alpha,\alpha\rangle_\epsilon + i\langle x,\alpha\rangle\}. \qquad (3.2)\end{aligned}$$

Note that the phase argument of the exponential on the right-hand side of (3.2) contains both the $g(\epsilon)$ and $g(0)$ scalar products. This mixture of scalar product types is a common feature of the representations of d_n. Let ψ_0 be a suitably smooth $\mathscr{L}^2(\mathbb{R}^d, \mathbb{C}^s)$ test function and suppose $\hat{\psi}_0$ is its Fourier transform, having the same normalization as (2.1), i.e.,

$$\psi_0(x) = \int e^{i\langle x,\alpha\rangle}\hat{\psi}_0(\alpha)d\alpha. \qquad (3.3)$$

Identity (3.2) implies

$$\begin{aligned}[D_0(\tau,\tau_0;\epsilon)\psi_0](x) &\equiv [U_0(\tau,\tau_0;\epsilon)\psi_0](x)\\ &= \int \exp\{-[i\hbar(\tau-\tau_0)/2m]\\ &\qquad\times\langle\alpha_0,\alpha_0\rangle_\epsilon + i\langle x,\alpha_0\rangle\}\hat{\psi}_0(\alpha_0)d\alpha_0. \qquad (3.4a)\end{aligned}$$

The mixed kernel of the zeroth Dyson iterate is that function $\hat{d}_0$ which satisfies

$$[D_0(\tau,\tau_0;\epsilon)\psi_0](x) = \int \hat{d}_0(x,\tau;\alpha_0,\tau_0;\epsilon)\hat{\psi}_0(\alpha_0)d\alpha_0. \qquad (3.4b)$$

Comparing Eqs. (3.4a) and (3.4b) we see that

$$\begin{aligned}&\hat{d}_0(x,\tau;\alpha_0,\tau_0;\epsilon)\\ &\quad= \exp\{-[i\hbar(\tau-\tau_0)/2m]\langle\alpha_0,\alpha_0\rangle_\epsilon + i\langle x,\alpha_0\rangle\}I. \qquad (3.4c)\end{aligned}$$

The first iterate is given by the formula

$$\begin{aligned}&[D_1(\tau,\tau_0;\epsilon)\psi_0](x)\\ &\quad= \frac{1}{i\hbar}\int_{\tau_0}^{\tau} d\tau_1[U_0(\tau,\tau_1;\epsilon)W(\tau_1;\epsilon)\\ &\qquad\times U_0(\tau_1,\tau_0;\epsilon)\psi_0](x). \qquad (3.5)\end{aligned}$$

We proceed by computing the effect of the interaction $W(\tau;\epsilon)$ on the state (3.4a). Equation (2.11) determines the p-independent parts of the operator $W(\tau;\epsilon)$ and Eq. (2.9) provides the second of the two p-dependent terms of W. The first p-dependent term is $-m^{-1}\langle a(x,\tau),p\rangle_\epsilon$ and the effect of this momentum operator on the plane wave state is

$$\langle a(x,\tau),p\rangle_\epsilon e^{i\langle x,\alpha_0\rangle} = \langle a(x,\tau),\hbar\alpha_0\rangle_\epsilon e^{i\langle x,\alpha_0\rangle}. \qquad (3.6)$$

We then combine this with the representation (2.1) for $a(x,\tau)$ and the resulting expression is of the same form as those found in Eqs. (2.9) and (2.11). Upon defining

$$\psi_1(\tau) = W(\tau;\epsilon)U_0(\tau,\tau_0;\epsilon)\psi_0, \qquad (3.7a)$$

one readily finds

$$\begin{aligned}[\psi_1(\tau)](x) &= \int d\alpha_0\Big[d\mu(\alpha_1,\tau) - \frac{\hbar}{m}\Big\langle\alpha_0 + \frac{1}{2}\alpha_1,\hat{\gamma}(\alpha_1,\tau)\Big\rangle_\epsilon d|\gamma|(\alpha_1,\tau)\Big]\\ &\qquad\times\exp\{-[i\hbar(\tau-\tau_0)/2m]\langle\alpha_0,\alpha_0\rangle_\epsilon + i\langle x,\alpha_0+\alpha_1\rangle\}\hat{\psi}_0(\alpha_0). \qquad (3.7b)\end{aligned}$$

For each fixed α_0 and τ the object within the square brackets is a measure in the space $\mathscr{M}(S_k, \mathbb{C}^{s\times s})$.

To obtain $D_1(\tau,\tau_0;\epsilon)\psi_0$, act with $U_0(\tau,\tau_1;\epsilon)$ on $\psi_1(\tau_1)$ and follow that by an integration over the variable τ_1 from τ_0 to τ. If the $d\alpha_0$ integration is done last, the result may be written as a mixed coordinate–momentum space kernel for D_1. Specifically, letting $\hat{Q} = (x,\tau;\alpha_0,\tau_0)$, one has

$$[D_1(\tau,\tau_0;\epsilon)\psi_0](x) = \int \hat{d}_1(\hat{Q};\epsilon)\hat{\psi}_0(\alpha_0)d\alpha_0, \qquad (3.8a)$$

where

$$\hat{d}_1(\hat{Q};\epsilon)=\frac{1}{i\hbar}\int_{\tau_0}^{\tau}d\tau_1\int\left[d\mu(\alpha_1,\tau_1)-\frac{\hbar}{m}\left\langle\alpha_0+\frac{1}{2}\alpha_1,\hat{\gamma}(\alpha_1,\tau_1)\right\rangle_\epsilon d|\gamma|(\alpha_1,\tau_1)\right]$$
$$\times\exp\{-[i\hbar(\tau-\tau_1)/2m]\langle\alpha_0+\alpha_1,\alpha_0+\alpha_1\rangle_\epsilon-[i\hbar(\tau_1-\tau_0)/2m]\langle\alpha_0,\alpha_0\rangle_\epsilon+i\langle x,\alpha_0+\alpha_1\rangle\}. \tag{3.8b}$$

In obtaining (3.8b) the operator $U_0(\tau,\tau_1;\epsilon)$ was evaluated by again employing (3.2) with $\alpha\to\alpha_0+\alpha_1$ and $\tau-\tau_0\to\tau-\tau_1$. For $\epsilon>0$, the exponential function [cf. (1.13)] is a decaying Gaussian as $|\alpha_0|\to\infty$. This means that the integration in (3.8a) is absolutely convergent for a wide class of test functions $\hat{\psi}_0$.

The nth mixed Dyson kernel is found by repeating the above process n times. The result of this calculation is the multiple integral

$$\hat{d}_n(\hat{Q};\epsilon)=\frac{1}{(i\hbar)^n}\int_< d\boldsymbol{\tau}_n\int\left[d\mu(\alpha_n,\tau_n)-\frac{\hbar}{m}\left\langle\sum_{j=0}^{n-1}\alpha_j+\frac{1}{2}\alpha_n,\hat{\gamma}(\alpha_n,\tau_n)\right\rangle_\epsilon d|\gamma|(\alpha_n,\tau_n)\right]$$
$$\times\cdots\times\left[d\mu(\alpha_1,\tau_1)-\frac{\hbar}{m}\left\langle\alpha_0+\frac{1}{2}\alpha_1,\hat{\gamma}(\alpha_1,\tau_1)\right\rangle_\epsilon d|\gamma|(\alpha_1,\tau_1)\right]$$
$$\times\exp\left(-\frac{i\hbar}{2m}\sum_{i,j=0}^{n}(\tau-\tau_{i\vee j})\langle\alpha_i,\alpha_j\rangle_\epsilon+i\left\langle x,\sum_{j=0}^{n}\alpha_j\right\rangle\right). \tag{3.9}$$

The notation for the indices i,j is $i\vee j=\max(i,j)$. The quadratic momentum factor in the exponential in (3.9) is a consequence of the identity

$$\sum_{i=0}^{n}(\tau_{i+1}-\tau_i)\left\langle\sum_{j=0}^{i}\alpha_j,\sum_{k=0}^{i}\alpha_k\right\rangle_\epsilon=\sum_{i,j=0}^{n}(\tau-\tau_{i\vee j})\langle\alpha_i,\alpha_j\rangle_\epsilon, \tag{3.10}$$

valid for all times $\tau_0<\tau_1<\cdots<\tau_n<\tau_{n+1}=\tau$. The order of the $\mathbb{C}^{s\times s}$-valued measures in (3.9) is important since these operators on $\mathbb{C}^s$ do not generally commute.

The coordinate space kernel $d_n(Q;\epsilon)$ is obtained as a Fourier transform of $\hat{d}_n(\hat{Q};\epsilon)$, namely,

$$d_n(Q;\epsilon)=\frac{1}{(2\pi)^d}\int\hat{d}_n(x,\tau;\alpha_0,\tau_0;\epsilon)e^{-i\langle y,\alpha_0\rangle}\,d\alpha_0. \tag{3.11}$$

Although it is an exercise of some complexity, the $d\alpha_0$ integration can be done exactly. We sketch this calculation and devise a notation to express the result in a convenient form.

Proceed by extracting all the α_0 dependence from the exponential in (3.9). Let

$$f_1(Q;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n;\epsilon)=\exp\left(-\frac{i\hbar}{2m}\sum_{i,j=1}^{n}(\tau-\tau_{i\vee j})\langle\alpha_i,\alpha_j\rangle_\epsilon+i\left\langle x,\sum_{j=1}^{n}\alpha_j\right\rangle\right), \tag{3.12}$$

and note that partial derivative ∂_{y_μ} $(\mu=1,\dots,d)$, given by

$$\partial_{y_\mu}=\frac{\partial}{\partial y_\mu}=\begin{cases}\dfrac{\partial}{\partial y^\mu}, & \text{if } g_{\mu\mu}=1,\\[2ex] -\dfrac{\partial}{\partial y^\mu}, & \text{if } g_{\mu\mu}=-1,\end{cases} \tag{3.13}$$

acts on the plane wave state as

$$(-i\partial_{y_\mu})^l e^{-i\langle x-y,\alpha_0\rangle}=(\alpha_0^\mu)^l e^{-i\langle x-y,\alpha_0\rangle},\quad l=1,2,\dots. \tag{3.14}$$

Using (3.12)–(3.14) we can formally write

$$d_n(Q;\epsilon)=\frac{1}{(i\hbar)^n}\int_< d\boldsymbol{\tau}_n\int\left[d\mu(\alpha_n,\tau_n)-\frac{\hbar}{m}\left\langle i\,\partial_y+\alpha_1+\cdots+\alpha_{n-1}+\frac{1}{2}\alpha_n,\hat{\gamma}(\alpha_n,\tau_n)\right\rangle_\epsilon d|\gamma|(\alpha_n,\tau_n)\right]$$
$$\times\cdots\times\left[d\mu(\alpha_1,\tau_1)-\frac{\hbar}{m}\left\langle i\,\partial_y+\frac{1}{2}\alpha_1,\hat{\gamma}(\alpha_1,\tau_1)\right\rangle_\epsilon d|\gamma|(\alpha_1,\tau_1)\right]f_1(Q;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n;\epsilon)J(Q;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n;\epsilon), \tag{3.15a}$$

where

$$J(Q;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n;\epsilon)=\frac{1}{(2\pi)^d}\int d\alpha_0\exp\left(-\frac{i\hbar(\tau-\tau_0)}{2m}\langle\alpha_0,\alpha_0\rangle_\epsilon-i\frac{\hbar}{m}\sum_{j=1}^{n}(\tau-\tau_j)\langle\alpha_0,\alpha_j\rangle_\epsilon+i\langle x-y,\alpha_0\rangle\right). \tag{3.15b}$$

The integral J is in the form of a Fresnel integral with a linear (in α_0) shifted phase. As such it can be computed in closed form. The result is

$$J(Q;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n;\epsilon)=N[\det g(-\epsilon)]^{1/2}\exp\left(\frac{im}{2\hbar(\tau-\tau_0)}\langle x-y,x-y\rangle_{-\epsilon}-i\left\langle x-y,\sum_{j=1}^{n}\frac{\tau-\tau_j}{\tau-\tau_0}\alpha_j\right\rangle\right)$$
$$\times\exp\left(\frac{i\hbar}{2m}\sum_{i,j=1}^{n}\frac{(\tau-\tau_i)(\tau-\tau_j)}{\tau-\tau_0}\langle\alpha_i,\alpha_j\rangle_\epsilon\right). \tag{3.16}$$

Here N is the normalization constant found in the free propagator. The integral (3.15b) is absolutely convergent if $\epsilon > 0$ and (3.16) shows the function it defines has a well-defined limit as $\epsilon \rightarrow 0$, which in turn is a continuous function of Q, $\boldsymbol{\alpha}_n$, and $\boldsymbol{\tau}_n$.

A very useful form of $d_n(Q;\epsilon)$ emerges when the exponentials in J are commuted through the $i\,\partial_y$ derivatives in (3.15b). This commutation process changes the structure of the integrand in (3.15a) and introduces a number of new terms. We introduce appropriate notations to describe this situation. Denote by $\vdots\ \vdots$ an ordering operation. If A_i, B_j, C_k, etc. are a series of operators, then $\vdots\ \vdots$ is defined as the product of these operators with increasing index value as read from right to left—i.e., if $j>k>i$, then

$$\vdots\, A_i B_j C_k \,\vdots = B_j C_k A_i. \tag{3.17}$$

In our applications the indices i, j, k are attached to the time labels τ_i, τ_j, τ_k, etc. Since the integral (3.15a) orders the time by $\tau_n \geqslant \tau_{n-1} \geqslant \cdots \geqslant \tau_1 \geqslant \tau_0$, the effect of $\vdots\ \vdots$ on its arguments will be to ensure the correct ordering of the noncommuting operator-valued measures. An additional useful convention is to let $[r]$ be the greatest integer less than or equal to r. The following lemma provides the required identities needed to carry out the commutation process.

Lemma 1: Let $\eta_1, \eta_2, \ldots, \eta_r$ be a set of r $(\mathbb{C}^{s\times s})^d$ matrices. If $z\in\mathbb{C}$ and $x\in\mathbb{C}^d$, then the formula

$$\langle \eta_r, \partial_x\rangle_\epsilon \cdots \langle \eta_1, \partial_x\rangle_\epsilon e^{z\langle x,x\rangle_{-\epsilon}} = e^{z\langle x,x\rangle_{-\epsilon}} \sum_{l=0}^{[r/2]} (2z)^{r-l} \sum_{\mathbf{q}_{r,l}} \vdots\, \langle x, \eta_{q_1}\rangle \cdots \langle x, \eta_{q_{r-2l}}\rangle \langle \eta_{q_{r-2l+1}}, \eta_{q_{r-2l+2}}\rangle_\epsilon \cdots \langle \eta_{q_{r-1}}, \eta_{q_r}\rangle_\epsilon \,\vdots \tag{3.18}$$

is valid. The summation convention for $\mathbf{q}_{r,l}$ is the following. For a given l and r, $\mathbf{q}_{r,l}$ represents a two-stage selection from $\{\eta_j\}_{j=1}^r$ into particular subsets. First choose $r-2l$ elements from $\{\eta_j\}_{j=1}^r$ and label these with the subscripts $q_1, q_2, \ldots, q_{r-2l}$. Next out of the remaining $2l$ η's form l pairs and use the subscripts $\{q_{r-2l+1}, q_{r-2l+2}\}, \ldots, \{q_{r-1}, q_r\}$ to label these. The summation over $\mathbf{q}_{r,l}$ involves all distinct choices of this type. There are $r![2^l(r-2l)!l!]^{-1}$ terms in this sum.

The nonrelativistic version of this algebraic lemma is found in Ref. 9. The combinatorial aspects of the nonrelativistic and relativistic formula are identical. A verification that the ϵ indices in (3.18) are correct can be seen from the formulas

$$\langle \eta, \partial_x\rangle_\epsilon e^{z\langle x,x\rangle_{-\epsilon}} = 2z\langle \eta, x\rangle e^{z\langle x,x\rangle_{-\epsilon}}, \tag{3.19a}$$

$$\langle \eta_i, \partial_x\rangle_\epsilon \langle \eta_j, x\rangle = \langle \eta_i, \eta_j\rangle_\epsilon. \tag{3.19b}$$

In order to complete the calculation of d_n one needs to expand the product of measures in (3.15a). We devise a notation for these expanded measures. Let n be the order of the Dyson kernel and for each r, $0<r<n$, define $\mathbf{j}_r$ to be a set of integers $\mathbf{j}_r = \{j_1 j_2, \ldots, j_r\}$ if $r>0$ and the empty set if $r=0$. Let $J_{n,r}$ denote the set of all subsets of $\{1,\ldots,n\}$ having r elements. If $r=0$, then $J_{n,0} = \{\mathbf{j}_0\} = \{\varnothing\}$. There are $\binom{n}{r}$ elements $\mathbf{j}_r$ in $J_{n,r}$. To each $\mathbf{j}_r \in J_{n,r}$ we associate a product measure. First define

$$d\rho_l(\alpha_l,\tau_l) = d\mu(\alpha_l,\tau_l) - \frac{\hbar}{m}\left\langle \frac{1}{2}\alpha_l + \sum_{j=1}^{l-1} \alpha_j, \hat{\gamma}(\alpha_l,\tau_l)\right\rangle_\epsilon d\,|\gamma|(\alpha_l,\tau_l), \tag{3.20a}$$

together with the associated polar factorization

$$d\rho_l(\alpha_l,\tau_l) = \hat{\rho}_l(\alpha_l,\tau_l)\, d\,|\rho_l|(\alpha_l,\tau_l). \tag{3.20b}$$

One constructs a positive real-valued measure on $(\mathbf{R}_1^d \times \cdots \times \mathbf{R}_n^d, \mathscr{B} \times \cdots \times \mathscr{B})$ via

$$d\Lambda^n(\mathbf{j}_r;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n) = d\,|\rho_1|(\alpha_1,\tau_1) \times \cdots \times d\,|\gamma|(\alpha_{j_1},\tau_{j_1}) \times \cdots \times d\,|\gamma|(\alpha_{j_r},\tau_{j_r}) \times \cdots \times d\,|\rho_n|(\alpha_n,\tau_n). \tag{3.20c}$$

Formula (3.20c) is to be understood in the following sense. If $r=0$, the measure involves only the products of $d\,|\rho_i|$, for $i=1,\ldots,n$. On the other hand, if $r>0$ and $\mathbf{j}_r = \{j_1,\ldots,j_r\}$, then the j_ith term of the product for the $r=0$ case has the measure $d\,|\rho_{j_i}|(\alpha_{j_i},\tau_{j_i})$ replaced by the measure $d\,|\gamma|(\alpha_{j_i},\tau_{j_i})$, for each $i = 1$–r.

The notation above means that the product of measures and differential operators in (3.15a) may be restated as

$$\begin{aligned} &\vdots \prod_{l=1}^n \left[d\mu(\alpha_l,\tau_l) - \frac{\hbar}{m}\left\langle i\,\partial_y + \alpha_1 + \cdots + \alpha_{l-1} + \frac{1}{2}\alpha_l, \hat{\gamma}(\alpha_l,\tau_l)\right\rangle_\epsilon d\,|\gamma|(\alpha_l,\tau_l)\right] \vdots \\ &\quad = \vdots \prod_{l=1}^n \left[d\rho_l(\alpha_l,\tau_l) - \frac{i\hbar}{m}\langle \partial_y, \hat{\gamma}(\alpha_l,\tau_l)\rangle_\epsilon d\,|\gamma|(\alpha_l,\tau_l)\right] \vdots \\ &\quad = \sum_{r=0}^n \sum_{\mathbf{j}_r\in J_{n,r}} d\Lambda^n(\mathbf{j}_r;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n) \vdots \left[\prod_{l\notin\mathbf{j}_r} \hat{\rho}_l(\alpha_l,\tau_l) \prod_{l\in\mathbf{j}_r}\right] \left(\frac{-i\hbar}{m}\right) \langle \partial_y, \hat{\gamma}(\alpha_l,\tau_l)\rangle_\epsilon \vdots\,. \end{aligned} \tag{3.21}$$

Consider next the behavior of the exponential phase arguments in (3.15a) and (3.16). The $\boldsymbol{\alpha}_n$-quadratic portion of (3.16) is independent of the y variables and so commutes with the operator ∂_y in the integrand of (3.15a). Adding the $\boldsymbol{\alpha}_n$-quadratic phases of f_1 and J gives (up to the multiplicative constant $-i\hbar/2m$)

$$\sum_{j,k=1}^n \left[(\tau - \tau_{j\vee k}) - \frac{(\tau-\tau_k)(\tau-\tau_j)}{\tau-\tau_0} \right] \langle \alpha_j, \alpha_k\rangle_\epsilon = \sum_{j,k=1}^n G_{jk}\langle \alpha_j, \alpha_k\rangle_\epsilon, \tag{3.22a}$$

where

$$G_{jk} \equiv G(\tau_j,\tau_k;\tau,\tau_0) = (\tau-\tau_0)\mathscr{G}(\xi_j,\xi_k). \tag{3.22b}$$

Here $\mathscr{G}$ is the one-dimensional fixed end point Green's function for the unit interval. Specifically, if $\xi_> = \max(\xi,\xi')$ and $\xi_< = \min(\xi,\xi')$, then

$$\mathscr{G}(\xi,\xi') = \xi_<(1-\xi_>). \tag{3.22c}$$

The argument ξ_j in (3.22b) is the fractional elapsed time,

$$\xi_j = (\tau_j - \tau_0)/(\tau - \tau_0). \tag{3.22d}$$

The addition of the $\boldsymbol{\alpha}_n$-linear phase parts of f_1 and J also leads to a simple result. The $\boldsymbol{\alpha}_n$-linear phase of J is y dependent and moving these factors through the ∂_y operators changes the integrand (3.15a). However, these modifications of the integrand leave the exponential phase of J unchanged. Summing these parts of the phases in f_1 and J gives

$$i\left\langle x, \sum_{j=1}^{n} \alpha_j\right\rangle - i\left\langle x-y, \sum_{j=1}^{n} \frac{\tau-\tau_j}{\tau-\tau_0}\alpha_j\right\rangle = i\sum_{j=1}^{n}\langle w(\xi_j;Q),\alpha_j\rangle. \tag{3.23}$$

The function $w(\cdot;Q)$ is a linear path from the initial point y to the final point x,

$$w(\xi_j;Q) = y - \xi_j(x-y). \tag{3.24}$$

In the flat Minkowski-type manifold on which this problem is set, w is a geodesic connecting the end points of Q. Note that the Green's function $G(\tau_j,\tau_k;\tau,\tau_0)$ and the path $w(\xi_j;Q)$ associated with it are independent of the parameter ϵ and the signature of $g_{\mu\nu}$.

Putting together the above identities yields the final representation for $d_n(Q;\epsilon)$. This result can be written in the form of the free evolution kernel times a multiple integral,

$$d_n(Q;\epsilon) = K_0(Q;\epsilon)\tilde{d}_n(Q;\epsilon). \tag{3.25}$$

In order to write $\tilde{d}_n$ in a compact form, set

$$\mathscr{P}_n(\mathbf{j}_r) = \sum_{l=0}^{[r/2]}\left(\frac{im}{\hbar(\tau-\tau_0)}\right)^{r-l}\sum_{\mathbf{q}_{r,l}}\Phi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n)\Psi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n), \tag{3.26a}$$

where Φ_l and Ψ_l are the functions

$$\Phi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n) = \prod_{i=0}^{l-1}\langle\hat{\gamma}(\alpha_{q_{r-2i}},\tau_{q_{r-2i}}),\hat{\gamma}(\alpha_{q_{r-2i-1}},\tau_{q_{r-2i-1}})\rangle_\epsilon, \tag{3.26b}$$

$$\Psi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n) = \prod_{i=1}^{r-2l}\langle\hat{\gamma}(\alpha_{q_i},\tau_{q_i}),y-X_n\rangle. \tag{3.26c}$$

Here the complex vector X_n is given in terms of its components by

$$X_n^\mu = x^\mu - \frac{\hbar}{m}g^{\mu\nu}g(\epsilon)_{\nu\beta}\sum_{j=1}^{n}(\tau-\tau_j)\alpha_j^\beta. \tag{3.26d}$$

In both (3.26b) and (3.26c) it is unnecessary to specify the order of the noncommuting terms in the product because the index ordering operation $\vdots\ \vdots$ will properly account for the placements of the matrices concerned. The summation convention for $\mathbf{q}_{r,l}$ is that of Lemma 1, with the correspondence $\eta_i \leftrightarrow \hat{\gamma}(\alpha_{j_i},\tau_{j_i})$.

In applying Lemma 1 in order to compute the integrand of $d_n(Q;\epsilon)$ and hence $\tilde{d}_n(Q;\epsilon)$, one first takes the $x-y$ quadratic and linear terms in the phase of J and completes the square. After this revision of the phase of J has taken place, one can apply Lemma 1 to compute the partial derivatives appearing in (3.15a). Assembling all the above steps together, one obtains, for $n \geqslant 1$,

$$\tilde{d}_n(Q;\epsilon) = \frac{1}{(i\hbar)^n}\sum_{r=0}^{n}\left(\frac{-i\hbar}{m}\right)^r\sum_{\mathbf{j}_r\in J_{n,r}}\int_< d\boldsymbol{\tau}_n\int d\Lambda^n(\mathbf{j}_r;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)$$
$$\times\exp\left(\frac{-i\hbar}{2m}\sum_{j,k=1}^{n}G_{jk}\langle\alpha_j,\alpha_k\rangle_\epsilon + i\sum_{j=1}^{n}\langle w(\xi_j;Q),\alpha_j\rangle\right)\vdots\,\mathscr{P}_n(\mathbf{j}_r)\prod_{\substack{k=1\\k\notin\mathbf{j}_r}}^{n}\hat{\rho}_k(\alpha_k,\tau_k)\vdots\,. \tag{3.27}$$

If $n=0$, define $\tilde{d}_0(Q;\epsilon) = I$. The function $\tilde{d}_n(Q;\epsilon)$ will be referred to as the reduced Dyson kernel.

The explicit representation (3.25)–(3.27) for $d_n(Q;\epsilon)$ is one of the principal results of this paper. Note that the multiple integral (3.27) is well defined for all ϵ because its integrand is a jointly continuous function of all its variables and the variables of integration range over a compact set. The characterization (3.25)–(3.27) is the relativistic and non-Abelian generalization of formula (6.20) in Ref. 9. Although this formula is both intricate and long, it will prove easy to investigate analytically and easy to find suitable bound estimates that guarantee the summability of the Dyson series. After the substitution of the original defining formulas for $\Lambda^n(\mathbf{j}_r;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)$, $\mathscr{P}_n(\mathbf{j}_r)$, etc., the integral representation (3.27) will manifestly display all the analytic dependence on ϵ, $\hbar$, m, and Q.

IV. SUMMABILITY ANALYSIS

In the previous section the fact that the space-time manifolds of special relativity are flat was used in order to introduce a Fourier analysis of the integral kernel of the nth Dyson iterate, $D_n(\tau,\tau_0;\epsilon)$. This formal calculation provided an explicit expression for the nth iterate kernel, $d_n(Q;\epsilon)$. The objective of the present section is twofold. First, bounds of $d_n(Q;\epsilon)$ are found that suffice to ensure that the summation over n is well defined in a pointwise sense. The second half of the section will verify that this sum is the fundamental solution of the proper-time Schrödinger equation (1.3a) with Hamiltonian $H(x,\tau;\epsilon)$.

Perhaps the most basic convergence problem here is found in the sum of the reduced kernels

$$F(Q;\epsilon)=\sum_{n=0}^{\infty}\tilde{d}_n(Q;\epsilon). \tag{4.1}$$

We now outline a four-step method for bounding $\tilde{d}_n(Q;\epsilon)$. Forward evolution $\tau\geqslant\tau_0$ is always assumed.

(a) Consider the exponential function appearing in (3.27). The second term in the phase factor is

$$i\sum_{j=1}^{n}\langle w(\xi_j;Q),\alpha_j\rangle,$$

which is pure imaginary and hence will not change the modulus of the exponential function. In order to understand the α-quadratic portion of the exponential it is helpful to employ the real and imaginary decomposition of the extended Lorentz scalar,

$$\begin{aligned}\langle v,w\rangle_\epsilon=&[(1-\epsilon^2)/(1+\epsilon^2)]\langle v,w\rangle\\&-[2i\epsilon/(1+\epsilon^2)]v\cdot w.\end{aligned} \tag{4.2a}$$

Here $v\cdot w$ denotes the Euclidean scalar product $\Sigma_{\mu=1}^{d}v^\mu w^\mu$. Decomposition (4.2a) allows one to write

$$\begin{aligned}&\sum_{j,l=0}^{n}G_{jl}\langle\alpha_j,\alpha_l\rangle_\epsilon\\&\quad=\frac{1-\epsilon^2}{1+\epsilon^2}\sum_{j,l=0}^{n}G_{jl}\langle\alpha_j,\alpha_l\rangle-\frac{2i\epsilon}{1+\epsilon^2}\sum_{j,l=0}^{n}G_{jl}\alpha_j\cdot\alpha_l.\end{aligned} \tag{4.2b}$$

The imaginary valued term on the right has the form previously encountered in the study of the nonrelativistic evolution problem [Ref. 12, Eq. (2.37)]. For $\tau_0\leqslant\tau_1\leqslant\cdots\leqslant\tau_n\leqslant\tau$, the identity

$$\begin{aligned}&\sum_{j,l=0}^{n}G_{jl}\alpha_j\cdot\alpha_l\\&\quad=\sum_{l=1}^{n}\left(\frac{1}{\tau-\tau_l}-\frac{1}{\tau-\tau_{l-1}}\right)\left|\sum_{j=l}^{n}(\tau-\tau_j)\alpha_j\right|^2\geqslant0\end{aligned} \tag{4.2c}$$

shows that the modulus of the exponential in (3.27) is bounded by 1.

(b) Next examine the measure volume associated with the momentum and time integrations. Formulas (3.20a) and (3.20c) together with estimates (2.12) and (2.14) and the restriction that $|\alpha_j|\leqslant k$ imply that

$$\|\Lambda^n(\mathbf{j}_r;\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)\|\leqslant(\mu_T+m^{-1}\hbar nk\gamma_T)^{n-r}\gamma_T^r. \tag{4.3a}$$

Inequality (4.3a) is applicable for all values of n, r, $\mathbf{j}_r$, $\boldsymbol{\alpha}_n$, $\boldsymbol{\tau}_n$, and ϵ that enter in the integral (3.27). The time-ordered integration gives one the well known result

$$\int_< d\boldsymbol{\tau}_n=\frac{(\tau-\tau_0)^n}{n!}. \tag{4.3b}$$

(c) The most elaborate structure in (3.27) is given by the integrand function containing $\mathscr{P}_n(\mathbf{j}_r)$. The norm identity $|\hat{\rho}_l(\alpha,\tau')|=1$ $(\tau'\in[0,T])$ means that

$$\left|:\mathscr{P}_n(\mathbf{j}_r)\prod_{\substack{k=1\\k\notin\mathbf{j}_r}}^{n}\hat{\rho}_k(\alpha_k,\tau_k):\right|\leqslant|\mathscr{P}_n(\mathbf{j}_r)|. \tag{4.4a}$$

To estimate the functions Ψ_l and Φ_l that occur in $\mathscr{P}_n(\mathbf{j}_r)$, we use the unitarity of $g_{\mu\nu}(\epsilon)$ and the identity $|\hat{\gamma}(\alpha,\tau')|=1$. These ensure that $|\Phi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n)|\leqslant1$ and that the factors entering in $\Psi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n)$ have the bound

$$|\langle\hat{\gamma}(\alpha_{q_i},\tau_{q_i}),y-X_n\rangle|\leqslant|x-y|+\hbar nk(\tau-\tau_0)/m. \tag{4.4b}$$

Since the sum over the allowed $\mathbf{q}_{r,l}$ in (3.26a) has $r![2^l(r-2l)!l!]^{-1}$ terms and our estimates of Ψ_l and Φ_l depend upon r and l but not on the value of $\mathbf{q}_{r,l}$ one sees that

$$\begin{aligned}&\sum_{\mathbf{q}_{r,l}}|\Psi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n)\Phi_l(\mathbf{q}_{r,l},\boldsymbol{\alpha}_n)|\\&\quad\leqslant\frac{r!}{2^l(r-2l)!l!}\left[|x-y|+\frac{\hbar nk(\tau-\tau_0)}{m}\right]^{r-2l}.\end{aligned} \tag{4.4c}$$

Rather than bound $|\mathscr{P}_n(\mathbf{j}_r)|$, it is more appropriate to bound $(\tau-\tau_0)^n|\mathscr{P}_n(\mathbf{j}_r)|$. The factor $(\tau-\tau_0)^n$ comes from the time scale factor in (4.3b). Since $r-2l\geqslant0$ and $r\leqslant n$, it is seen that every occurrence of $(\tau-\tau_0)$ in $(\tau-\tau_0)^n|\mathscr{P}_n(\mathbf{j}_r)|$ has a non-negative power. Let $\delta\tau$ be a time displacement bound, i.e., $(\tau-\tau_0)\leqslant\delta\tau\leqslant T$. Then one has

$$\begin{aligned}&(\tau-\tau_0)^n|\mathscr{P}_n(\mathbf{j}_r)|\\&\quad\leqslant(\delta\tau)^n\sum_{l=0}^{[r/2]}\left(\frac{m}{\hbar\delta\tau}\right)^{r-l}\frac{r!}{2^l(r-2l)!l!}\\&\qquad\times(|x-y|+\hbar nk\delta\tau/m)^{r-2l}.\end{aligned} \tag{4.4d}$$

Replacing the coefficient $r!/[(r-2l)!l!]$ by the larger value $n^l\binom{r}{l}$ and then extending the sum over l from $[r/2]$ to r we obtain

$$\begin{aligned}&(\hbar/m)^r(\tau-\tau_0)^n|\mathscr{P}_n(\mathbf{j}_r)|\\&\quad\leqslant(\delta\tau)^{n-r}[|x-y|+\hbar nk\delta\tau/m+1/2k]^r.\end{aligned} \tag{4.4e}$$

(d) The final step is to carry out the summation over $\mathbf{j}_r$ and r. The bounds occurring in (4.3a), (4.3b), and (4.4e) are independent of the value of $\mathbf{j}_r$. There are $\binom{n}{r}$ terms in the sum over subsets $\mathbf{j}_r\in J_{n,r}$. Thus, at this stage, the bound for $|\tilde{d}_n(Q;\epsilon)|$ reads

$$|\tilde{d}_n(Q;\epsilon)| \leqslant \frac{1}{n!\hbar^n} \sum_{r=0}^{n} \binom{n}{r} \left(\mu_T\delta\tau + \frac{\hbar nk\gamma_T\delta\tau}{m}\right)^{n-r} \times \left[|x-y|\gamma_T + \frac{\hbar nk\gamma_T\delta\tau}{m} + \frac{\gamma_T}{2k}\right]^r = \frac{1}{n!\hbar^n}\left[|x-y|\gamma_T + \mu_T\delta\tau + \frac{\gamma_T}{2k} + \frac{2\hbar nk\gamma_T\delta\tau}{m}\right]^n. \tag{4.5a}$$

The n dependence of this last bound can be further simplified if we use $n^n/n! < e^n$ and $(1+n^{-1}a)n \leqslant e^a$, $a \geqslant 0$, to obtain

$$|\tilde{d}_n(Q;\epsilon)| \leqslant \left(\frac{2ek\gamma_T\delta\tau}{m}\right)^n \times \exp\left[\frac{m}{2\hbar k\delta\tau}\left(|x-y| + \frac{1}{2k} + \frac{\mu_T\delta\tau}{\gamma_T}\right)\right]. \tag{4.5b}$$

Estimate (4.5b) is independent of $\tau \in [\tau_0, \tau_0 + \delta\tau]$ and $\epsilon \geqslant 0$. The quantities μ_T, γ_T, and k are all finite constants associated with the values or the supports of the potentials in class (A).

If one defines a variable

$$\Theta = 2ek\gamma_T\delta\tau/m, \tag{4.6}$$

it follows from (4.5b) that the series (4.1) is absolutely convergent whenever $\Theta < 1$. Estimates similar to (4.5b) apply to the x, y, and τ derivatives of $\tilde{d}_n(Q;\epsilon)$. One can verify that the x or y partial derivatives may be interchanged with the integrals in (3.27). One then investigates the multiple integral (3.27) with the integrand modified by the partial derivatives. Note that all the x, y dependence is contained in the phase factor

$$\sum_{j=1}^{n} \langle w(\xi_j;Q),\alpha_j\rangle$$

and in the function $\Psi_l(\mathbf{q}_{r,l},\alpha_n)$. In a like fashion the first partial derivative with respect to τ or τ_0 may be calculated. Recall that $\gamma(\tau)$ and $\nu(\tau)$ are continuous in τ though not necessarily differentiable in τ. Here, the present study of the Dyson series differs from the former nonrelativistic approach,[9,19] wherein it assumed that these measures were continuously differentiable in τ. There it was necessary to use these stronger assumptions on the potentials in order to utilize the evolution theory in Banach space.[7] In expression (3.27) each measure $d\gamma(\tau_i)$ and $d\nu(\tau_i)$ appears within the corresponding time integral over τ_i. The only τ dependence to appear in (3.27) is in the upper limit of the τ_n domain of integration and in the manifest τ dependence found in $\mathscr{P}_n(\mathbf{j}_r)$ and ξ_j. After employing the Leibnitz rule for differentiating parametric integrals, the estimating method (a)–(d) is again applicable.

Prior to summarizing these results in lemma form, it is helpful to introduce some additional notation. Let $\mathbb{I}^d_+$ be the set of d-tuples of non-negative integers. The multi-index $\phi \in \mathbb{I}^d_+$ will specify a partial derivative with respect to the variable x, while $\theta \in \mathbb{I}^d_+$ will be a y partial derivative, e.g.,

$$\partial_x^\phi = \left(\frac{\partial}{\partial x_1}\right)^{\phi_1} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\phi_d}. \tag{4.7}$$

The triangular time domain obeying $\tau_0 < \tau$ is defined to be the set

$$\Delta_T^0 = \{(\tau_0,\tau) \in [0,T]^2 : \tau_0 < \tau\}$$

and the closure of Δ_T^0 will be denoted by Δ_T. Finally if $\delta\tau$ is a time displacement bound, the striplike region it defines is denoted by

$$\Delta_T^0(\delta\tau) = \{(\tau_0,\tau) \in \Delta_T^0 : \tau - \tau_0 < \delta\tau\},$$

with closure $\Delta_T(\delta\tau)$. The convergence condition $\Theta < 1$ is fulfilled if the time pair $(\tau_0,\tau) \in \Delta_T(\delta\tau)$, where $\delta\tau < m/2ek\gamma_T$. From (4.5b) we extract the estimating function

$$b(x,\delta\tau) = \exp\{m|x|/2\hbar k\delta\tau\}. \tag{4.8}$$

The reduced kernel $\tilde{d}_n(Q;\epsilon)$ is characterized by the following.

Lemma 2: Let $\tilde{d}_n : \Delta_T \times \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty) \to \mathbb{C}^{s\times s}$ be the function defined by the integral (3.27). For all $n \geqslant 0$, $\tilde{d}_n$ has partial derivatives to arbitrary order in x,y that are jointly continuous in the domain $\Delta_T \times \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty)$. Further, $\tilde{d}_n$ has a first-order partial derivative in τ continuous in the domain $\Delta_T^0 \times \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty)$. If $0 \leqslant \delta\tau < T - \tau_0$ is an arbitrary time displacement bound, then, throughout the domain $\Delta_T(\delta\tau) \times \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty)$, $\tilde{d}_n$ and its derivatives satisfy the estimates

$$|\tilde{d}_n(Q;\epsilon)| \leqslant \Theta^n c_1 b(x-y,\delta\tau), \tag{4.9a}$$

$$|\partial_x^\phi \partial_y^\theta \tilde{d}_n(Q;\epsilon)| \leqslant n^{|\phi|+|\theta|}\Theta^n c_2 b(x-y,\delta\tau), \tag{4.9b}$$

$$\left|\frac{\partial}{\partial\tau}\tilde{d}_n(Q;\epsilon)\right| \leqslant \frac{n^2\Theta^n}{\tau-\tau_0}(c_3 + c_4|x-y|)b(x-y,\delta\tau), \tag{4.9c}$$

where the positive constants c_i $(i = 1,2,3,4)$ are independent of n, Q, and ϵ.

Proof (sketch)*:* The discussion prior to the lemma shows how the inequalities (4.9a)–(4.9c) are obtained. The claim that $\tilde{d}_n(Q;\epsilon)$, $\partial_x^\phi\partial_y^\theta\tilde{d}_n(Q;\epsilon)$, and $(\partial/\partial\tau)\tilde{d}_n(Q;\epsilon)$ are continuous in an appropriate domain of (Q,ϵ) follows from a straightforward application of the dominated convergence theorem. □

It is worthwhile to contrast the approaches used to bound the Dyson series in the relativistic and nonrelativistic problems. In the nonrelativistic case, the convergence properties of (4.1) for analytical potentials are known in substantial detail[9] and the associated derivative field asymptotics[19] (valid as $m \to \infty$) have been fully worked out. The method (a)–(d) is a copy of the nonrelativistic analysis with the indefinite Lorentz scalar $\langle u,w\rangle$ replacing the Euclidean inner product $u\cdot w$. Because of the inequality (1.9c) the indefinite character of $\langle u,w\rangle$ has no adverse impact on the estimating procedure (a)–(d). The combinatorial and indexing problems for the series in the relativistic and nonrelativistic problems are identical. Arguably the most critical ingredient in the estimating procedure is the bound of the exponential in step (a) implied by the inequality (4.2c). The definition of the extended metric tensor $g(\epsilon)$ was devised so as to ensure the validity of (4.2c). If the measures $\gamma(\tau)$ and $\nu(\tau)$ were continuously differentiable on $[0,T]$, then the singular factor $(\tau-\tau_0)^{-1}$ can be deleted from the estimate (4.9c).

The function $F(Q;\epsilon)$ is central in our analysis of the propagator $K(Q;\epsilon)$. The analytic and smoothness properties of $F(Q;\epsilon)$ are described as follows.

Proposition 1: Let $0<\delta\tau<m/2ek\gamma_T$ be a time displacement bound.

(1) For each $(Q,\epsilon)\in\Delta_T(\delta\tau)\times\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)$, the sum (4.1) is absolutely convergent and provides a pointwise definition of the function

$$F\colon \Delta_T(\delta\tau)\times\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)\to\mathbb{C}^{s\times s}.$$

(2) The function F is continuous. For all $(Q,\epsilon)\in\Delta_T(\delta\tau)\times\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)$, F has the bound

$$|F(Q;\epsilon)|\leqslant[c_1/(1-\Theta)]b(x-y,\delta\tau). \tag{4.10}$$

(3) The function F has partial derivatives to arbitrary order in x,y that are jointly continuous on the domain $\Delta_T(\delta\tau)\times\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)$.

(4) The function F has a first-order partial derivative in τ that is jointly continuous on the domain $\Delta_T^0(\delta\tau)\times\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)$.

Proof: The inequality $\delta\tau<m/2ek\gamma_T$ guarantees that $\Theta<1$ and this together with estimate (4.9a) shows that the series (4.1) is absolutely convergent and that F obeys the bound (4.10). Let Ω be any compact subset of $\mathbb{R}^d\times\mathbb{R}^d$. The inequality (4.9a) implies the series (4.1) is uniformly convergent on the compact domain $\Delta_T(\delta\tau)\times\Omega\times[0,\epsilon_+]$, where $0<\epsilon_+<\infty$. Each term $\tilde{d}_n$ is uniformly continuous on $\Delta_T(\delta\tau)\times\Omega\times[0,\epsilon_+]$. Thus the sum F is continuous on this compact set. Since Ω and ϵ_+ are arbitrary, it follows that F is continuous throughout $\Delta_T(\delta\tau)\times\mathbb{R}^d\times\mathbb{R}^d\times\ [0,\infty)$. This establishes properties (1) and (2).

Estimates (4.9b) and (4.9c) suffice to show that the sum (4.1) may be differentiated term by term. The continuity properties of $\partial_x^\phi\partial_y^\theta F$ and $\partial F/\partial\tau$ result from an application of the argument used to prove the continuity of F. □

In view of the product identity (3.25) relating $\tilde{d}_n$ to d_n, the convergence of (4.1) implies the convergence of

$$K(Q;\epsilon)=\sum_{n=0}^{\infty}d_n(Q;\epsilon). \tag{4.11}$$

In particular, K admits the product representation

$$K(Q;\epsilon)=K_0(Q;\epsilon)F(Q;\epsilon). \tag{4.12}$$

Although it is not made explicit in Lemma 2 and Proposition 1, the inequality (4.9a) may be used to show that $F(Q;\epsilon)$ is a smooth bounded function as $m^{-1}\to0+$. In sharp contrast to this, $K_0(Q;\epsilon)$ has an essential singularity as $m^{-1}\to0+$, as can be seen from inspection of (1.12). Thus the factorization (4.12) provides a precise characterization of the singular behavior of $K(Q;\epsilon)$ in a neighborhood of $m^{-1}=0$. In the analysis of the nonrelativistic propagator,[9,19] it was possible to let one variable (complex mass) implement the embedding process (which occurs if $\operatorname{Im} m>0$) and play the role of the small parameter in the derivative field approximation. In the relativistic case, $H(x,\tau)$ is no longer an elliptic partial differential operator and for this reason the embedding procedure is carried out separately by the variable ϵ in the metric tensor while the asymptotic scaling is still controlled by $m^{-1}\to0+$.

We now turn to the general problem of verifying that $K(Q;\epsilon)$ is the solution of the ϵ-extended proper-time Schrödinger equation

$$i\hbar\frac{\partial}{\partial\tau}K(Q;\epsilon)=H(x,\tau;\epsilon)K(Q;\epsilon). \tag{4.13}$$

We proceed by finding a recurrence relation that links d_n to d_{n-1}. A summation over n of this recurrence formula then leads to identity (4.13).

In Sec. III, $d_n(Q;\epsilon)$ was defined, via (3.11), to be the Fourier transform of $\hat{d}_n(\hat{Q};\epsilon)$. It is helpful to clarify the precise meaning attached to this transform. Let $\mathscr{S}(\mathbb{R}^d,\mathbb{C}^{s\times s})$ denote the Schwartz space of $\mathbb{C}^{s\times s}$-valued functions of rapid decrease on $\mathbb{R}^d$. Specifically, $\Gamma\in\mathscr{S}(\mathbb{R}^d,\mathbb{C}^{s\times s})$ if

$$\|\Gamma\|_{\theta,\phi}=\sup_{x\in\mathbb{R}^d}|x^\theta\partial_x^\phi\Gamma(x)|<\infty, \tag{4.14}$$

for all $\theta,\phi\in\mathbb{I}_+^d$.

Lemma 3: Let $\hat{d}_n\colon\Delta_T\times\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)\to\mathbb{C}^{s\times s}$ be the function defined by Eqs. (3.4c) and (3.9). For each $n\geqslant0$, $\hat{d}_n$ is a continuous function.

Assume $\epsilon>0$. For each $(x,\tau_0,\tau)\in\mathbb{R}^d\times\Delta_T^0$ the following hold.

(1) For all $\phi\in\mathbb{I}_+^d$,

$$\alpha_0\mapsto\partial_x^\phi\hat{d}_n(x,\tau;\alpha_0,\tau_0;\epsilon)\in\mathscr{S}(\mathbb{R}_{\alpha_0}^d,\mathbb{C}^{s\times s}).$$

(2) The Fourier transform (3.11) is a bicontinuous bijection from $\mathscr{S}(\mathbb{R}_{\alpha_0}^d,\mathbb{C}^{s\times s})$ onto $\mathscr{S}(\mathbb{R}_x^d,\mathbb{C}^{s\times s})$.

(3) For all $\phi\in\mathbb{I}_+^d$,

$$\partial_x^\phi d_n(Q;\epsilon)=\frac{1}{(2\pi)^d}\int\partial_x^\phi\hat{d}_n(\hat{Q};\epsilon)e^{-i\langle y,\alpha_0\rangle}\,d\alpha_0, \tag{4.15a}$$

$$\partial_\tau d_n(Q;\epsilon)=\frac{1}{(2\pi)^d}\int\partial_\tau\hat{d}_n(\hat{Q};\epsilon)e^{-i\langle y,\alpha_0\rangle}\,d\alpha_0. \tag{4.15b}$$

Proof: (sketch): The integral (3.9) is an absolutely convergent integral with a finite domain of integration, $(S_k)^n\times[0,T]^n$. The continuity property of $\hat{d}_n$ in the variable $(\hat{Q};\epsilon)$ follows from an application of the dominated convergence theorem that exploits the continuity properties of the integrand in Eq. (3.9). For $n=0$, the continuity of $\hat{d}_0$ follows from inspection of formula (3.4c).

Consider statements (1) and (2). Initially let $|\phi|=0$. Examine (3.4c) and (3.9). The modulus of the exponential function is controlled by the real part of its phase, viz.,

$$\begin{aligned}\operatorname{Re}&\left\{-\frac{i\hbar}{2m}\sum_{j,l=0}^{n}(\tau-\tau_{j\vee l})\langle\alpha_j,\alpha_l\rangle_\epsilon\right\}\\&=\frac{-\hbar\epsilon}{m(1+\epsilon^2)}\sum_{j=0}^{n}(\tau_{j+1}-\tau_j)|\alpha_0+\cdots+\alpha_j|^2\\&\leqslant\frac{\hbar\epsilon(\tau-\tau_0)}{m(1+\epsilon^2)}(-|\alpha_0|^2+2nk|\alpha_0|).\end{aligned} \tag{4.16a}$$

The equality in (4.16a) is a consequence of (4.2a) and the inequality employs $|\alpha_j|\leqslant k$, $j=1,\ldots,n$. Estimate (4.16a) together with the measure volume bound analogous to (4.3a) and (4.3b) shows that

$$|\hat{d}_n(\hat{Q};\epsilon)| \leqslant \frac{(\tau-\tau_0)^n}{n!}\left[\frac{\mu_T}{\hbar}+\frac{\gamma_T(|\alpha_0|+nk)}{m}\right]^n$$

$$\times \exp\{[\hbar\epsilon(\tau-\tau_0)/m(1+\epsilon^2)]$$

$$\times(-|\alpha_0|^2+2nk\,|\alpha_0|)\}. \qquad (4.16b)$$

Thus if $\epsilon>0$ and $\tau>\tau_0$, $\hat{d}_n(\hat{Q};\epsilon)$ has a Gaussian decay bound in the variable α_0.

Next compute the α_0 partial derivatives of $\hat{d}_n(\hat{Q};\epsilon)$. Once again the analytic properties of the integrand of (3.9) and the compact domain of integration allow one to justify without difficulty (cf. Ref. 26, Appendix B.3; Ref. 19, Proposition 2) that the differential operator $\partial^{\phi'}_{\alpha_0}$ may be passed through the integration. Using estimates like those leading to (4.16b) will yield a bound for $|\partial^{\phi'}_{\alpha_0}\hat{d}_n(\hat{Q};\epsilon)|$ similar to (4.16b)—namely, the same Gaussian decay factor appears but it is now multiplied by a polynomial in $|\alpha_0|$ of order $n+|\phi'|$. This estimate shows that the norms $\|\hat{d}_n(x,\tau;\cdot,\tau_0;\epsilon)\|_{\theta',\phi'}$ are finite for all $\theta',\phi'\in\mathbb{I}^d_+$, provided $\epsilon>0$ and $\tau>\tau_0$. Thus (1) is established if $|\phi|=0$. A parallel argument applies if $|\phi|>0$. Property (2) is a direct consequence of (1) and the Fourier inversion theorem for the space $\mathscr{S}(\mathbb{R}^d,\mathbb{C}^{s\times s})$ (Ref. 27, Theorem IX.1).

(3) Assuming $\epsilon>0$ and $\tau>\tau_0$, a variation of the derivation above shows that $\partial^{\phi}_x\hat{d}_n(\hat{Q};\epsilon)$ and $\partial_\tau\hat{d}_n(\hat{Q};\epsilon)$ satisfy Gaussian decay estimates that suffice to guarantee the interchange of the limits shown in (4.15a) and (4.15b). □

Two special examples of (4.15a) and (4.15b) are important in understanding the recurrence relations for d_n. For the second-order partial differential operator $H_0(\epsilon)$ and the first-order partial differential operator $W(x,\tau;\epsilon)$ we have

$$H_0(\epsilon)d_n(Q;\epsilon)=\frac{1}{(2\pi)^d}\int H_0(\epsilon)\hat{d}_n(\hat{Q};\epsilon)e^{-i\langle y,\alpha_0\rangle}\,d\alpha_0, \qquad (4.17)$$

$$W(x,\tau;\epsilon)d_n(Q;\epsilon)$$

$$=\frac{1}{(2\pi)^d}\int W(x,\tau;\epsilon)\hat{d}_n(\hat{Q};\epsilon)e^{-i\langle y,\alpha_0\rangle}\,d\alpha_0. \qquad (4.18)$$

Identities (4.17) and (4.18) are valid for $\epsilon>0$ and $\tau>\tau_0$. If $\epsilon=0$, then the Fourier transform (3.11) is not absolutely integrable and must be reinterpreted as some type of improper integral.

Lemma 4: Assume that $\epsilon\geqslant 0$. For all $n\geqslant 0$,

(1) the factorization identity (3.25) relating d_n to $\tilde{d}_n$ holds for all $Q\in\Delta^0_T\times\mathbb{R}^d\times\mathbb{R}^d$;

(2) for each $(\tau_0,\tau,x)\in\Delta^0_T\times\mathbb{R}^d$, the functions $\hat{d}_n$ satisfy, for all $\alpha_0\in\mathbb{R}^d$, the recurrence relation

$$i\hbar\frac{\partial}{\partial\tau}\hat{d}_n(\hat{Q};\epsilon)=H_0(\epsilon)\hat{d}_n(\hat{Q};\epsilon)+W(x,\tau;\epsilon)\hat{d}_{n-1}(\hat{Q};\epsilon); \qquad (4.19)$$

and

(3) for all $Q\in\Delta^0_T\times\mathbb{R}^d\times\mathbb{R}^d$, the functions d_n satisfy the recurrence relation

$$i\hbar\frac{\partial}{\partial\tau}d_n(Q;\epsilon)=H_0(\epsilon)d_n(Q;\epsilon)+W(x,\tau;\epsilon)d_{n-1}(Q;\epsilon). \qquad (4.20)$$

The functions d_{-1} and $\hat{d}_{-1}$ are defined to be zero.

Proof: (1) Let $\epsilon>0$. The computation in Sec. III that transforms (3.11) into (3.15a) and (3.15b) only involves the changing of the order of integration with respect to $d\alpha_0$ and the $d\mu\,d\gamma$ integrals. One knows from Lemma 3 that (3.11) is absolutely convergent, so Fubini's theorem justifies this change of integration order. The evaluation of the Gaussian integral (3.15b) is (for positive ϵ) a standard exercise [cf. Ref. 9, Eqs. (6.15) and (6.16)]. To obtain the validity of (3.25) for $\epsilon=0$, take the limit as $\epsilon\to 0+$ and note for fixed Q and $\tau>\tau_0$ that d_n, $\tilde{d}_n$, and K_0 are continuous functions of ϵ on the closed interval $[0,\epsilon_+]$, $\epsilon_+>0$.

(2) For $(\tau_0,\tau)\in\Delta^0_T$ define the n-dimensional time-ordered domain

$$\Delta_n(\tau,\tau_0)=\{\boldsymbol{\tau}_n\in[0,T]^n\colon\ \tau_0\leqslant\tau_1\leqslant\cdots\leqslant\tau_n\leqslant\tau\}. \qquad (4.21)$$

In addition, let f_2 denote the exponential function and $d\Lambda^n_0$ be the $(\alpha_0,\boldsymbol{\alpha}_n)$-dependent product measure appearing in (3.9):

$$f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)$$

$$=\exp\left(-\frac{i\hbar}{2m}\sum_{i,j=0}^{n}(\tau-\tau_{i\vee j})\langle\alpha_i,\alpha_j\rangle_\epsilon+i\left\langle x,\sum_{j=0}^{n}\alpha_j\right\rangle\right), \qquad (4.22a)$$

$$d\Lambda^n_0(\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)=\left[d\mu(\alpha_n,\tau_n)-\frac{\hbar}{m}\left\langle\sum_{j=0}^{n-1}\alpha_j\right.\right.$$

$$\left.\left.+\frac{1}{2}\alpha_n,\hat{\gamma}(\alpha_n,\tau_n)\right\rangle_\epsilon d\,|\gamma|(\alpha_n,\tau_n)\right]$$

$$\times\cdots\times\left[d\mu(\alpha_1,\tau_1)-\frac{\hbar}{m}\left\langle\alpha_0\right.\right.$$

$$\left.\left.+\frac{1}{2}\alpha_1,\hat{\gamma}(\alpha_1,\tau_1)\right\rangle_\epsilon d\,|\gamma|(\alpha_1,\tau_1)\right]. \qquad (4.22b)$$

First write (3.9) as the iterated integral

$$\hat{d}_n(\hat{Q};\epsilon)=\frac{1}{(i\hbar)^n}\int_{\tau_0}^{\tau}d\tau_n\left[\int_{\Delta_{n-1}(\tau_n,\tau_0)}d\boldsymbol{\tau}_{n-1}\right.$$

$$\left.\times\int d\Lambda^n_0(\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)\right]. \qquad (4.23)$$

The Leibnitz rule for differentiating integrals with variable limits is used to evaluate the τ derivative of $\hat{d}_n(\hat{Q};\epsilon)$. The Leibnitz rule is applicable since the integrand in the square bracket defines, for each $(x,\alpha_0,\epsilon)\in\mathbb{R}^d\times\mathbb{R}^d\times[0,\infty)$, a continuously differentiable function of τ on the domain $(\tau_n,\tau)\in[\tau_0,T]^2$. In this way one obtains

$$i\hbar\frac{\partial}{\partial\tau}\hat{d}_n(\hat{Q};\epsilon)=T_1(\hat{Q};\epsilon)+T_2(\hat{Q};\epsilon), \qquad (4.24)$$

where T_1 is the term coming from the differentiation of the upper limit of the $d\tau_n$ integral in (4.23) and T_2 is the term resulting from the τ derivative acting on f_2.

Consider T_1 first. The exponential argument of f_2 has the property

$$\left.\sum_{j,l=0}^{n}(\tau-\tau_{j\vee l})\langle\alpha_j,\alpha_l\rangle_\epsilon\right|_{\tau_n=\tau}=\sum_{j,l=0}^{n-1}(\tau-\tau_{j\vee l})\langle\alpha_j,\alpha_l\rangle_\epsilon \qquad (4.25a)$$

or, equivalently,

$$f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)|_{\tau_n=\tau}=e^{i\langle x,\alpha_n\rangle}f_2(\hat{Q};\boldsymbol{\alpha}_{n-1},\boldsymbol{\tau}_{n-1}). \qquad (4.25b)$$

As a consequence of this and

$$\int \left[d\mu(\alpha_n,\tau) - \frac{\hbar}{m} \left\langle \sum_{j=0}^{n-1} \alpha_j + \frac{1}{2}\alpha_n, \hat{\gamma}(\alpha_n,\tau) \right\rangle_\epsilon \times d\,|\gamma|(\alpha_n,\tau) \right] e^{i\langle x,\alpha_0 + \cdots + \alpha_n\rangle} = W(x,\tau;\epsilon) e^{i\langle x,\alpha_0 + \cdots + \alpha_{n-1}\rangle}, \tag{4.25c}$$

we obtain

$$T_1(\hat{Q};\epsilon) = \frac{1}{(i\hbar)^{n-1}} \int_< d\boldsymbol{\tau}_{n-1}\, W(x,\tau;\epsilon) \times \int d\Lambda_0^{n-1}(\boldsymbol{\alpha}_{n-1},\boldsymbol{\tau}_{n-1}) f_2(\hat{Q};\boldsymbol{\alpha}_{n-1},\boldsymbol{\tau}_{n-1}) = W(x,\tau;\epsilon)\hat{d}_{n-1}(\hat{Q};\epsilon). \tag{4.25d}$$

In evaluating the term T_2, we note that f_2 satisfies

$$i\hbar \frac{\partial}{\partial\tau} f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n) = H_0(\epsilon) f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n). \tag{4.26a}$$

Thus

$$T_2(\hat{Q};\epsilon) = \frac{1}{(i\hbar)^n} \int_< d\boldsymbol{\tau}_n \times \int d\Lambda_0^n(\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n) H_0(\epsilon) f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n) = H_0(\epsilon)\hat{d}_n(\hat{Q};\epsilon). \tag{4.26b}$$

Combining (4.24), (4.25d), and (4.26b) leads to (4.19). Obtaining the final line in both (4.25d) and (4.26b) requires the passing of first- or second-ordered partial derivatives in x through the $\boldsymbol{\alpha}_n$ and $\boldsymbol{\tau}_n$ integrals. Obvious estimates of $|\partial_x^\phi f_2(\hat{Q};\boldsymbol{\alpha}_n,\boldsymbol{\tau}_n)|$ suffice to justify this interchange of limiting orders. The case $n = 1$ follows an argument similar to the above and the case $n = 0$ is a straightforward calculation. The recurrence relation (4.19) holds for all $\tau > \tau_0$ and $\epsilon \geqslant 0$.

Consider the proof of statement (3). Initially assume $\epsilon > 0$. Lemma 3 (1) shows that the right-hand terms of (4.19) are elements of $\mathscr{S}(\mathbb{R}^d_{\alpha_0};\mathbb{C}^{s\times s})$. The pointwise character of equality (4.19) means that this conclusion extends to $\partial_\tau \hat{d}_n(\hat{Q};\epsilon)$. Take the $d\alpha_0$ Fourier transform of all the terms in (4.19) to obtain

$$\int i\hbar \frac{\partial}{\partial\tau} \hat{d}_n(\hat{Q};\epsilon) e^{-i\langle x,\alpha_0\rangle}\, d\alpha_0 = \int H_0(\epsilon)\hat{d}_n(\hat{Q};\epsilon) e^{-i\langle x,\alpha_0\rangle}\, d\alpha_0 + \int W(x,\tau;\epsilon)\hat{d}_{n-1}(\hat{Q};\epsilon) e^{-i\langle x,\alpha_0\rangle}\, d\alpha_0. \tag{4.27}$$

Employing relationships (4.15b), (4.17), and (4.18) gives the recurrence relation (4.20) for $\epsilon > 0$. The integral representation (3.27) of $\tilde{d}_n(Q;\epsilon)$, the explicit analytic form of $K_0(Q;\epsilon)$ and the identity (3.25) show (cf. Lemma 2), for fixed Q with $\tau > \tau_0$, that each function in (4.20) is continuous in ϵ on the closed interval $[0,\epsilon_+]$, $\epsilon_+ > 0$. Thus identity (4.20) may be extended to $\epsilon = 0$ by continuity. □

The final task is to establish that $K(Q;\epsilon)$ is the pointwise solution of (4.13), which, in the limit $\tau \to \tau_0$, satisfies the delta function initial condition (1.3b). The delta function behavior is best characterized in terms of a multidimensional stationary phase expansion. The following lemma adapts general results on oscillatory integrals to the present problem wherein the phase factor is the extended Lorentz scalar $\langle x - y, x - y\rangle_{-\epsilon}$ [cf. (1.12)].

Lemma 5: For each fixed $\epsilon \geqslant 0$, let $\lambda_+ > 0$. Suppose that

$$h(\cdot,\cdot;\epsilon): \mathbb{R}^d_x \times [0,\lambda_+] \to \mathbb{C}^s$$

is a continuous function with the properties (i) there exists a compact set $\Omega \subset \mathbb{R}^d_x$ such that $\operatorname{supp} h(\cdot,\lambda;\epsilon) \subseteq \Omega$, for all $\lambda \in [0,\lambda_+]$; and (ii) for $|\phi| \leqslant d$, the partial derivatives $\partial_x^\phi h(x,\lambda;\epsilon)$ exist and are continuous in the domain $\mathbb{R}^d_x \times [0,\lambda_+]$.

Let $I_\epsilon(\lambda)$ be the oscillatory integral

$$I_\epsilon(\lambda) = [1/(i\pi\lambda)]^{d/2} [\det g(-\epsilon)]^{1/2} \times \int_{\mathbb{R}^d} e^{(i/\lambda)\langle x,x\rangle_{-\epsilon}} h(x,\lambda;\epsilon)\, dx. \tag{4.28}$$

Then

$$\lim_{\lambda \to 0_+} I_\epsilon(\lambda) = h(0,0;\epsilon). \tag{4.29}$$

Proof: If $\epsilon = 0$, the principal term $h(0,0;\epsilon)$ is the same as found in Theorem 2.2 of Fedoriuk's article[28] on the stationary phase method. A slight modification of Fedoriuk's proof and hypothesis gives the result above. □

Below $C_0^n(\mathbb{R}^d,\mathbb{C}^s)$ will denote the compactly supported functions on $\mathbb{R}^d$ having continuous nth-order derivatives.

Proposition 2: Suppose the potentials a and v are in the class (A) and $\delta\tau$ is a time displacement bound with $\delta\tau < m[2ek\gamma_T]^{-1}$. If

$$K: \Delta_T^0(\delta\tau) \times \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty) \to \mathbb{C}^{s\times s}$$

is the function defined by Eq. (4.11), then, for each point $(Q,\epsilon) \in \Delta_T^0(\delta\tau) \times \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty)$, the partial differential equation (4.13) is identically fulfilled.

Furthermore, if $\psi_0 \in C_0^d(\mathbb{R}^d,\mathbb{C}^s)$ and $(\tau_0,\tau) \in \Delta_T^0(\delta\tau)$, then

$$\lim_{\tau \to \tau_0 +} \int_{\mathbb{R}^d} K(Q;\epsilon)\psi_0(y)\, dy = \psi_0(x), \quad x \in \mathbb{R}^d. \tag{4.30}$$

Proof: Let (Q,ϵ) be a fixed point in the domain of K. Observe that the τ derivative of K may be written

$$i\hbar \frac{\partial}{\partial\tau} K(Q;\epsilon) = \sum_{n=0}^\infty i\hbar \frac{\partial}{\partial\tau} d_n(Q;\epsilon) = \sum_{n=0}^\infty [H_0(\epsilon) d_n(Q;\epsilon) + W(x,\tau;\epsilon) d_n(Q;\epsilon)] = H(x,\tau;\epsilon) K(Q;\epsilon). \tag{4.31}$$

The estimate (4.9c) shows the n summation may be interchanged with the τ derivative. The recurrence relation (4.20) gives the second identity in (4.31). Estimate (4.9b) shows that the operators $H_0(\epsilon)$ and $W(x,\tau;\epsilon)$ may be pulled back through the sum over n to obtain the last equality of (4.31).

In order to verify (4.30), use representation (4.12) for $K(Q;\epsilon)$ and set $\lambda = m/2\hbar(\tau - \tau_0)$. Changing the variable of integration with the substitution $y \to y' = x - y$ determines the h of Lemma 5 to be

$$h(y',\lambda;\epsilon) = F(x,\tau_0 + m/2\hbar\lambda;x - y',\tau_0;\epsilon)\psi_0(x - y'). \tag{4.32}$$

Hypothesis (i) of Lemma 5 is obeyed since we may take Ω to be the closure of any open ball containing $x - \operatorname{supp} \psi_0$. The differentiability and continuity requirements are met as a consequence of Proposition 1(3) and the fact $\psi_0 \in C_0^d (\mathbb{R}^d, \mathbb{C}^s)$. In evaluating the right-hand side of (4.29), we obtain the function

$$h(0,0;\epsilon) = F(x,\tau_0;x,\tau_0;\epsilon)\psi_0(x). \tag{4.33}$$

This diagonal value of F is determined by the series in Eq. (4.1). By definition, $\tilde{d}_0(x,\tau;x,\tau_0;\epsilon)$ is the identity matrix in $\mathbb{C}^{s\times s}$ whereas formula (3.27) for $\tilde{d}_n$ $(n\geqslant 1)$ implies that $\tilde{d}_n (x,\tau_0;x,\tau_0;\epsilon)$ vanishes. Thus the value of $F(x,\tau_0;x,\tau_0;\epsilon)$ is the unit matrix I and consequently (4.33) reads $h(0,0;\epsilon) = \psi_0(x)$. This establishes (4.30). □

It is worthwhile to note that other proofs of the recurrence relation (4.20) are possible. One of these alternate derivations, which we have completed, is to compute the needed space and time derivatives of $d_n (Q;\epsilon)$ directly from the representation obtained via (3.25) and (3.27). This approach has the merit that it is not necessary to use the ϵ-continuity properties of $d_n (Q;\epsilon)$ in order to establish the recurrence relation at $\epsilon = 0$. However, the derivatives get entangled with the multiple summations and combinatorics that enter (3.27) and as a result the calculations are of a forbidding length.

V. WAVE FUNCTION EVOLUTION

We conclude the analysis of relativistic evolution with a discussion of the wave function. Let $\psi_0 \in \mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$ be a suitable test function. For each ψ_0 the fundamental solution $K(Q;\epsilon)$, found in Eqs. (4.11) and (4.12), constructs an associated wave function ψ by the formula

$$\psi(x;\tau,\tau_0;\epsilon) = \int K(Q;\epsilon)\psi_0(y)dy. \tag{5.1}$$

Evidently a desirable feature of the test function ψ_0 is that it ensures that the integral (5.1) be absolutely convergent. Given estimate (4.10), this integrability is guaranteed if ψ_0 is an $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$ function of compact support. If $\epsilon > 0$, then the compactness of the support of ψ_0 is unnecessary. However, the summary given below of wave function behavior is stated in a fashion such that the results are equally applicable for all values of $\epsilon \geqslant 0$.

Corollary 1: Assume that a and v are interactions in the class (A) and that $\delta\tau$ is a time displacement bound with $\delta\tau < m/(2ek\gamma_T)$. If $\psi_0 \in C_0^d(\mathbb{R}^d,\mathbb{C}^s)$, then define the continuous function

$$\psi\colon \mathbb{R}^d \times \Delta_T^0(\delta\tau) \times [0,\infty) \to \mathbb{C}^s$$

by formula (5.1).

(1) ψ is a smooth function in the sense that $\partial_\tau\psi$ and $\partial_x^\phi\psi$ $(|\phi| \geqslant 0)$ exist and are continuous.

(2) For each $\epsilon \geqslant 0$, the function ψ is a classical (pointwise) solution of

$$i\hbar \frac{\partial}{\partial\tau}\psi(x;\tau,\tau_0;\epsilon) = H(x,\tau;\epsilon)\psi(x;\tau,\tau_0;\epsilon) \tag{5.2}$$

in the domain $\mathbb{R}^d \times \Delta_T^0(\delta\tau)$. Furthermore, ψ satisfies the initial data condition

$$\lim_{\tau\to\tau_0+} \psi(x;\tau,\tau_0;\epsilon) = \psi_0(x), \quad x\in\mathbb{R}^d. \tag{5.3}$$

Proof: Properties (3) and (4) of Proposition 1, the explicit form of $K_0(Q;\epsilon)$, and Eq. (4.12) suffice to show that the x and τ derivatives may be interchanged with the dy integral in (5.1), i.e.,

$$\partial_x^\phi\psi(x;\tau,\tau_0;\epsilon) = \int \partial_x^\phi K(Q;\epsilon)\psi_0(y)dy, \quad |\phi|\geqslant 0, \tag{5.4a}$$

$$\frac{\partial}{\partial\tau}\psi(x;\tau,\tau_0;\epsilon) = \int \frac{\partial}{\partial\tau}K(Q;\epsilon)\psi_0(y)dy. \tag{5.4b}$$

The continuity properties of ψ, $\partial_x^\phi\psi$, and $\partial_\tau\psi$ are all verified by employing the dominated convergence theorem. Setting

$$L(x,\tau;\epsilon) = i\hbar\frac{\partial}{\partial\tau} - H(x,\tau;\epsilon), \tag{5.5a}$$

one has

$$L(x,\tau;\epsilon)\int K(Q;\epsilon)\psi_0(y)dy = \int L(x,\tau;\epsilon)K(Q;\epsilon)\psi_0(y)dy = 0. \tag{5.5b}$$

The first equality above is justified by the identities (5.4a) and (5.4b). Proposition 2 asserts that $L(x,\tau;\epsilon)K(Q;\epsilon)$ is zero; whence (5.2) follows. Equation (5.3) is a restatement of Eq. (4.30). □

Corollary 1 indicates that the solution of (5.2) as constructed by (5.1) is an element of $C^\infty(\mathbb{R}^d,\mathbb{C}^s)$. This smoothness is a consequence of the analytic nature of the coefficients of the partial differential operator $H(x,\tau;\epsilon)$. It is reasonable to take $C_0^d(\mathbb{R}^d,\mathbb{C}^s)$ as the initial data space as this set of functions is dense in $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$.

The approach used throughout the paper has been to construct the fundamental solution of the equation of motion (4.13) via a pointwise absolutely convergent series. The associated analysis has not used any facts from the Banach space evolution theory, which also is applicable to this problem. Of course a more comprehensive understanding of the relativistic evolution problem will emerge when the pointwise and abstract characterizations are combined. For example, with the estimates given so far for $K(Q;\epsilon)$, it is not evident how one would prove (if $\epsilon = 0$) that ψ in (5.1) is an $\mathscr{L}^2(\mathbb{R}^d,\mathbb{C}^s)$ function. However, this fact is an automatic consequence of the Banach space evolution theory. In the nonrelativistic case $(\sigma_+ = d)$ this unified (pointwise and abstract) treatment[9,13] has been carried out in detail by using an ϵ-embedding technique. In the present treatment the ϵ-continuation process is responsible for establishing the nature of the Fourier equivalence of $\hat{d}_n(\hat{Q};\epsilon)$ and $d_n(Q;\epsilon)$; for making the generalized Fresnel integral (3.15b) well defined; and for providing a simple proof of the recurrence relation (4.20).

ACKNOWLEDGMENTS

The authors have benefited from discussions of the various portions of this material with our colleagues Frank Molzahn and Larry Horwitz.

This work has been supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). R.A.C. gratefully acknowledges a postdoctoral fellowship awarded by NSERC and thanks the Instituut voor Theoretische Fysica at the Universiteit Leuven for its warm hospitality.

[1] C. N. Yang and R. L. Mills, Phys. Rev. **96**, 191 (1954).

[2] S. K. Wong, Nuovo Cimento A **65**, 689 (1970).

[3] J. Schwinger, Phys. Rev. **82**, 664 (1951).

[4] B. S. DeWitt, *Dynamical Theory of Groups and Fields* (Gordon and Breach, New York, 1965).

[5] B. S. DeWitt, Phys. Rep. C **19**, 295 (1975).

[6] F. J. Dyson, Phys. Rev. **75**, 846, 1736 (1949).

[7] S. G. Krein, *Linear Differential Equations in Banach Space* (Am. Math. Soc., Providence, RI, 1971).

[8] M. Reed and B. Simon, *Methods of Modern Mathematical Physics III: Scattering Theory* (Academic, New York, 1979), Sec. XI.15.

[9] T. A. Osborn, L. Papiez, and R. Corns, J. Math. Phys. **28**, 103 (1987).

[10] T. L. Gill and W. W. Zachary, J. Math. Phys. **28**, 1459 (1987).

[11] B. Simon, Bull. Am. Math. Soc. **7**, 447 (1982).

[12] T. A. Osborn and Y. Fujiwara, J. Math. Phys. **24**, 1093 (1983).

[13] R. A. Corns, Ph.D. thesis, University of Manitoba, 1988.

[14] L. P. Horwitz and C. Piron, Helv. Phys. Acta **48**, 316 (1973).

[15] P. Roman and J. P. Leveille, J. Math. Phys. **15**, 1760 (1974).

[16] L. P. Horwitz and A. Soffer, Helv. Phys. Acta **53**, 112 (1980).

[17] R. Arshansky and L. P. Horwitz, Phys. Lett. A **128**, 123 (1988); J. Math. Phys. **30**, 66 (1989).

[18] T. A. Osborn and F. H. Molzahn, Phys. Rev. A **34**, 1669 (1986).

[19] L. Papiez, T. A. Osborn, and F. H. Molzahn, J. Math. Phys. **29**, 642 (1988).

[20] J. A. Zuk, Phys. Rev. D **33**, 3645 (1986); **34**, 1791 (1986).

[21] T. A. Osborn, R. A. Corns, and Y. Fujiwara, J. Math. Phys. **26**, 453 (1985).

[22] W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1974).

[23] K. Ito, in *Proceedings of the Fourth Berkeley Symposium on Mathematics Statistics and Probability* (Univ. of California Press, Berkeley, 1961), Vol. II, p. 227.

[24] S. A. Albeverio and R. J. Høegh-Krohn, *Mathematical Theory of Feynman Path Integrals* (Springer, Berlin, 1976).

[25] R. Cameron and D. Storvick, J. Anal. Math. **38**, 34 (1986).

[26] B. C. Carlson, *Special Functions of Applied Mathematics* (Academic, New York, 1977).

[27] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).

[28] M. V. Fedoriuk, Usp. Mat. Nauk. **26**(3), 67 (1971) [Russ. Math. Surv. **26**(3), 65 (1971)].

Finite energy solutions for (1+1)-dimensional σ models

Bernard Piette[a]
Institut de Physique Theorique, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

Wojciech J. Zakrzewski
Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, England

(Received 30 August 1989; accepted for publication 6 December 1989)

Various properties of finite energy solutions of the $(1+1)$-dimensional sigma models are studied. It is shown that the energy densities of these solutions exhibit some extended lump-like structures that cross each other without interaction, and that the procedure of adding a soliton to a given solution developed by Harnad, Schnider, and Saint-Aubin [Commun. Math. Phys. **92**, 329 (1984); **43**, 33 (1984)] mixes the matrix elements of this solution in a complicated way, but it does not modify its energy momentum tensor density. The one parameter family of conserved currents discovered by Eichenherr, Forger, and Pohlmeyer [Nucl. Phys. B **155**, 381 (1979); **164**, 528 (1980); Commun. Math. Phys. **46**, 207 (1976)] must be considered to distinguish between such solutions. Finally the Uhlenbeck method of construction of Euclidean solutions is modified to make it applicable to the construction of Minkowskian solutions for some classes of sigma models. It is shown that this method does not modify the energy momentum tensor density either.

I. INTRODUCTION

In 1976 Pohlmeyer showed[1] that two-dimensional nonlinear σ models are integrable. Since then they have been the subject of many studies. Different methods have been developed to construct, in more or less explicit form, classical solutions of their equations of motion both in the Euclidean and Minkowski case. The fact that solutions can be constructed explicitly can be seen as a consequence of the integrability of these models.

The notion of integrability is, however, not well defined. It comes from the theory of ordinary differential equations where by Arnold's theorem[2] a system is said to be integrable if it possesses as many conserved quantities as it has its degrees of freedom. Arnold's theorem provides a method of exploiting the existence of these conserved quantities to reduce the process of the construction of solutions of these equations to an algebraic problem. For systems of partial differential equations, which can be thought as describing systems with an infinite number of degrees of freedom, no such theorem exists but, by analogy, such systems are commonly said to be integrable if they possess an infinite number of conservation laws. The problem with such a definition is that, as everybody knows, "infinity plus or minus one is still infinity" and thus, even if a given system has an infinite number of conservation laws, it is very hard to know whether these conserved quantities are sufficient to characterize completely and uniquely a given solution.

We can certainly say that a system is integrable if we know how to construct explicitly all its solutions, or, in practice, large classes of its solutions.

Usually, however, we are not interested in all solutions as we also require that the solutions possess some further properties. For example, in classical field theories, we are usually interested in only those solutions whose energy is finite.

Integrable systems are also considered as models which can be solved by means of the inverse scattering method or equivalently by the so-called Bäcklund transformation method.[3,4] In these methods one associates with the nonlinear system of partial differential equations a pair of linear differential equations dependent on an additional parameter, for which the original nonlinear equations serve as the compatibility condition. This pair of linear equations, called the Lax pair in the literature, is usually the starting point for the construction of an infinite number of conserved quantities as well as the construction of the Bäcklund transformations. The idea of a Bäcklund transformation can be summarized as follows: one starts from a known solution, and tries to construct a new one by adding to it what is called a soliton. When one looks at a new solution, it appears that the transformation has added to the original solution an extended structure which propagates with a given speed, and which preserves its shape. Usually such a transformation is highly nontrivial to perform. A simple example of such a construction is the one developed for the sine-Gordon equation.[5] The Bäcklund transformation or the inverse scattering method are just some of the many techniques for constructing solutions of nonlinear models, but as one can guess, specific models usually have their specific methods for constructing their solutions.

Since Pohlmeyer published his paper,[1] many attempts have been made to construct solutions of both the Minkowskian and Euclidean classical σ models. The class of σ models we are interested in is described by a matrix Q which, for simplicity, we will assume to be unitary, and the expression for the action of these models is given by

[a] Chercheur IISN, Belgium.

0022-2488/90/040916-08$03.00

$$S=\int dx^2\,\frac{1}{4}\mathrm{Tr}[\partial_\mu Q^\dagger \partial_\mu Q], \tag{1.1}$$

where the summation over μ is performed either with the Euclidean or the Minkowskian metric, and where, due to the unitary of Q,

$$Q^\dagger Q=1. \tag{1.2}$$

Further algebraic constraints can be imposed on Q to define other classes of models (see Ref. 4 for a complete list). If, in particular, Q is chosen to be Hermitian, the model is the so-called Grassmannian sigma model.

From the above action it is easy to derive the equations of motion of the model i.e., the Euler–Lagrange equations,

$$\partial_\mu(Q^\dagger\,\partial_\mu Q)=0. \tag{1.3}$$

Restricting ourselves to the Minkowskian case we find that it is convenient to introduce the light-cone coordinates $\xi=x+t$, $\eta=x-t$. Then if we define

$$A_\xi=Q^\dagger\,\partial_\xi Q,\quad A_\eta=Q^\dagger\,\partial_\eta Q, \tag{1.4}$$

the action and the equations of motion of the model can be rewritten as

$$S=-\int dx^2\,\mathrm{Tr}[A_\xi A_\eta], \tag{1.5}$$

$$\partial_\eta A_\xi+\partial_\xi A_\eta=0. \tag{1.6}$$

As the model is integrable we can associate with it a pair of linear equations[3] (called the Lax pair equations)

$$\begin{aligned}\partial_\xi\Psi&=A_\xi\Psi/(1+\lambda),\\ \partial_\eta\Psi&=A_\eta\Psi/(1-\lambda),\end{aligned} \tag{1.7}$$

from which one can derive an infinite number of conservation laws[1,6]. Defining

$$j_\xi=\partial_\xi Q\,Q^\dagger,\quad j_\eta=\partial_\eta Q\,Q^\dagger. \tag{1.8}$$

we see that we have a one parameter family of conserved currents corresponding to each solution[6]

$$\begin{aligned}j_\xi(\omega)&=[(1-\omega)/2(1+\omega)]U(\omega)\partial_\xi Q\,Q^\dagger U(\omega)^{-1},\\ j_\eta(\omega)&=[(1+\omega)/2(1-\omega)]U(\omega)\partial_\eta Q\,Q^\dagger U(\omega)^{-1},\end{aligned} \tag{1.9}$$

where U is a function of ξ, η, and ω, which satisfies

$$\begin{aligned}\partial_\xi U&=[2\omega/(\omega+1)]Uj_\xi,\\ \partial_\eta U&=[2\omega/(\omega-1)]Uj_\eta.\end{aligned} \tag{1.10}$$

We are interested in the construction of the explicit solutions of the Eqs. (1.3). Moreover, we are interested only in those solutions that have finite energy. At this stage we have to make a distinction between the Euclidean and the Minkowskian models. The solutions of the Euclidean models in two dimensions can be considered as static solutions of the Minkowskian models in 2 + 1 dimensions, implying that the two-dimensional action of each such solution is its total energy. On the other hand, the energy densities of the Minkowskian models in 1 + 1 dimensions are not explicit and have to be computed.

Applying the standard methods, one shows that the energy momentum tensor for the Minkowskian models is given by[7,8]

$$\begin{aligned}T^{00}&=2\,\mathrm{Tr}[A_\xi A_\xi+A_\eta A_\eta],\\ T^{10}&=2\,\mathrm{Tr}[A_\xi A_\xi-A_\eta A_\eta].\end{aligned} \tag{1.11}$$

Many classes of solutions of (1.3) are known. One of the aims of this work is to study some of these solutions and in particular those obtained for the Minkowskian models by applying various solution generating techniques and show how to construct finite energy solutions using these techniques.

Before doing so let us briefly summarize the known results from some Euclidean models, as we will use these results as a guide for the Minkowskian models. In the Euclidean case for some classes of sigma models all finite energy solutions can be constructed explicitly, whereas no such results have been proved so far for the Minkowskian models.

The first complete set of solutions for a Euclidean sigma model was found by Borchers and Garber[9] who determined all finite energy solutions of the S^{2n+1} models [which they called the $O(2n+1)$ models]. Their construction has been extended[10,11] to the $\mathbb{C}P^{N-1}$ models and to the construction of some classes of solutions of Grassmannian models. The set of all these solutions can be described with ease. For the $\mathbb{C}P^{N-1}$ models, for example, the set of solutions splits into N subsets, each solution in any of these subsets being completely characterized by a polynomial holomorphic vector. In fact the subsets correspond to the Gramm–Schmidt orthonormalization procedure applied to the holomorphic vector and its first $N-1$ derivatives. Moreover, the models possess a topological charge and any solution can be uniquely characterized by its energy and the topological charge densities.

Recently Uhlenbeck reduced the problem of the construction of all solutions of the $U(N)$ sigma models to that of solving a system of algebraic and first-order partial differential equations. Using her theorem we were then able to construct explicitly all solutions of the $U(3)$ and $U(4)$ sigma models.[12] For these models, like for the $\mathbb{C}P^{N-1}$ models, the energy density is not sufficient to characterize uniquely a given solution. The $U(N)\sigma$ models do not have a topological charge but, by construction, the set of solutions also exhibits a subset-like structure. The Uhlenbeck construction is based on the Lax pair approach to the sigma models. To construct all solutions, one determines first some elementary solutions called by Uhlenbeck the one-uniton solutions[13,14] [which are in fact the self-dual solutions for the Grassmannian models imbedded in $U(N)$] and then adds to them up to $N-2$ additional unitons. The method resembles the procedure of adding solitons to a Minkowskian solution. However, the addition of a uniton to a given solution changes the energy density of this solution. The energy density of the new solution exhibits, in general, additional peaks; their number, position, and shape depend on the properties of the added uniton.[13]

The construction of classical solutions of the Minkowskian sigma models has also received much attention. First Pohlmeyer showed how solutions of the sine–Gordon equation can be related to the solutions of the S^2 sigma model. He also discussed the so-called dual symmetry that can be exploited to construct some further solutions of this model. Later, Eichenherr, and Forger[6] have extended these results to other sigma models.

A next major step in the construction of explicit solutions of the Minkowskian models was achieved by Harnad, Schnider, and Saint-Aubin[4] who, using the work of Zakharov and Mikhailov,[3] developed an algorithm for the addition of any number of solitons to any given solution. Starting from a given solution one has first to solve the Lax pair problem (1.7) for this particular solution. Then the addition of a soliton to this solution can be achieved by purely algebraic operations. This method has then been modified to make it applicable to the construction of Euclidean solutions as well.[16,17]

The problem with all constructions developed for the Minkowskian models is that the construction does not guarantee that the new solutions have finite energy. It is interesting to determine how the energy density of a given solution is modified when one applies the above mentioned procedure and adds a soliton to it. In the next chapter, we will show that the addition of a soliton does not alter the energy momentum density of the original configuration. For this reason we will also argue that the term "addition of a soliton" is not well suited for finite energy solutions as this procedure corresponds merely to a sophisticated internal mixing of the matrix elements of the original solution.

Finally let us recall some properties of the $(1+1)$-dimensional sigma models. As is well known, these models are conformaly invariant, which means that in two dimensions[6,7] the tensor of the energy momentum density for these solutions satisfies the wave equation. The expressions for the energy density of such models are thus of the form

$$T^{00}=f(x+t)+g(x-t), \tag{1.12}$$

where f and g are arbitrary functions. This density thus describes two extended structures moving with the velocity of light in both directions and crossing each other without any apparent interaction (the lumps just add to each other). This does not mean that the construction of solutions is completely trivial, as the nonlinearity of the model makes the simple addition of solutions impossible. The superposition of two solutions can thus result in some internal rearrangement, but this effect does not manifest itself in the energy density. We will discuss this point further in the next section in which we will discuss some explicit examples.

II. ADDITION OF SOLUTIONS TO FINITE ENERGY SOLUTIONS

In the Introduction, we have described some methods that have been developed to construct solutions of the Minkowskian sigma models. What is the energy density of these solutions? To answer this question we will consider here some of the different methods mentioned before.

In his seminal paper[1] Pohlmeyer showed that solutions of the sine–Gordon equation can lead to some solutions of the S^2 model [also called $O(3)$ sigma model]. This is an interesting observation, as the sine–Gordon equation is one of the best known nonlinear equations, and Bäcklund transformations have been developed for this equation. The field of the S^2 model can be described by a three-component real vector q. Assuming that $q_\xi^\dagger q_\xi$ and $q_\eta^\dagger q_\eta$ are both constant, Pohlmeyer proved that the angle α between the vectors $q_\xi = \partial_\xi q$ and $q_\eta = \partial_\eta q$ satisfies the sine–Gordon equation. Unfortunately, as the energy density for the S^2 model is given by

$$E=q_\xi^t q_\xi + q_\eta^t q_\eta, \tag{2.1}$$

we see that the assumptions made imply that the energy density E is constant and thus that the total energy of these solutions is infinite. This means that this embedding of the sine–Gordon model into the S^2 one is not very useful in the context of classical field theories.

In this paper[1] Pohlmeyer showed also how to exploit the so-called dual symmetry of the model to construct further new solutions. Given a solution Q of the sigma model, one computes first the solution U of the Lax pair problem (1.9) for this particular solution and multiplying them together obtains a new solution which now depends on an additional parameter ω. However, as can be easily checked, the gauge fields A_ξ and A_η for the old and the new solutions are exactly the same [Eqs. (2.19) and (2.20) in Ref. 6]. As a consequence the energy momentum tensor and the Lagrangian density are the same for both the new and the old solutions. Thus it would appear that this construction does not really add a soliton to the original solution. At most it corresponds to a sophisticated mixing of the matrix elements of the original solution.

One of the most elegant procedures for the construction of solutions of the sigma models is the method developed by Harnad, Schnider, and Saint-Aubin.[4] This method also corresponds to the addition of solitons to a given solution. In this procedure one starts by constructing a simple solution Q_0 of the model and constructs a matrix χ dependent on a parameter λ, and which satisfies a pair of linear differential equations [(Eq. 2.7) in Ref. 4]. Then, as the authors show, $Q = \chi_{(\lambda=0)} Q_0$ is a new solution of the model. Once the Lax pair problem (1.7) has been solved for the starting solution, the construction of any multisoliton solution reduces to a purely algebraic computation (even though it very quickly becomes too tedious to perform in practice).

The difficult part of this construction is finding a starting solution. In their paper[4] Harnad *et al.* suggest to consider what they call the "vacuum" solutions

$$Q_0=\exp[(i/2)(A\xi+B\eta)], \tag{2.2}$$

where A and B are two constant Hermitian matrices which commute with each other. A few explicit solutions were constructed using this method[18,19] starting from the solution of the form (2.2).

Thus if we choose

$$A=B=\begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \tag{2.3}$$

the vacuum solution is given by

$$Q_0=\begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix}, \tag{2.4}$$

where $x=(1/2)(\xi+\eta)$.

Addition a soliton to this solution is rather tedious, but quite straightforward. Defining[19]

$$\begin{aligned} &\omega=\tfrac{1}{2}[\xi/(1+\lambda)+\eta/(1-\lambda)], \\ &u=2\,\mathrm{Re}(\omega), \quad v=2\,\mathrm{Im}(\omega), \\ &\lambda=\lambda'-i\lambda'', \end{aligned} \tag{2.5}$$

we have

$$\begin{aligned} g_{11}=g_{22}=&\cos(x)-(1/\Lambda^2)[(\lambda'/\lambda'')\sin(x)\sinh(2v) \\ &+\cos(x)\cosh(2v)+(|\lambda|^2-1)[\cos(2u-3x) \\ &-|\lambda|^2\cos(2u-x)]], \\ g_{12}=&-\overline{g}_{21} \\ =&-\sin(x)-(1/\Lambda^2)\Big\{(\lambda'/\lambda'')\cos(x)\sinh(2v) \\ &-\sin(x)\cosh(2v)+(|\lambda|^2-1)^{-1}[\sin(2u-3x) \\ &+|\lambda|^2\sin(2u-x)]+2i\Big[(\lambda'/\lambda'')\sin(u-x) \\ &\times\cosh(v)-\frac{|\lambda|^2+1}{|\lambda|^2-1}\cos(u-x)\sinh(v)\Big]\Big]\Big\}, \end{aligned} \tag{2.6}$$

where

$$\Lambda^2=|\lambda|^2\left[\frac{4\cos^2(u-x)}{(1-|\lambda|^2)^2}+\frac{\cosh^2 v}{\lambda''^2}\right]. \tag{2.7}$$

This solution is the same as the one given in Ref. 19 apart from an overall multiplication by a constant matrix.

When we plot the matrix elements of these two solutions,[19] we see that the construction does indeed appear to add solitons to the original solution. Moreover the solitons propagate with a speed that depends on a given parameter (which can essentially be chosen arbitrarily). The fact that the solitons propagate with arbitrary speed may be surprising as we know that the energy density can only exhibit structures which can propagate with the speed of light. This paradox is resolved by observing that "the solitons" appear only in the graphs of the matrix elements; their contribution cancels in the energy density. To see this we have to compute the energy density of the original solution, and then check how the addition of a soliton modifies this density.

Let us look first at the second problem. When we add a soliton to a solution Q_0, the gauge fields A_ξ and A_η of the new solution are given by[4]

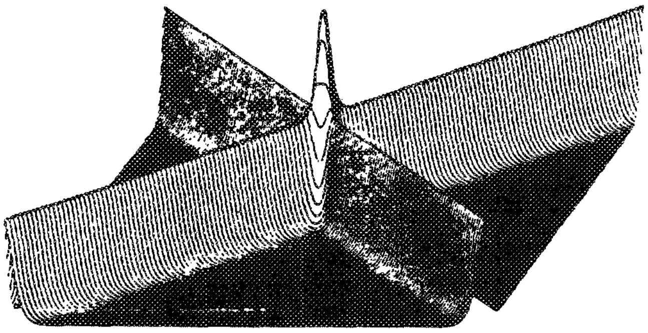


FIG. 1. Energy density (2.11) of the solutions (2.10) and (2.6).

$$A_\xi=\chi(-1)A^0_\xi\chi(-1)^{-1}, \quad A_\eta=\chi(1)A^0_\eta\chi(1)^{-1}, \tag{2.8}$$

where A^0_ξ and A^0_η are the gauge fields (1.4) corresponding to Q_0. From (2.8) it is clear that the energy momentum tensor density (1.11) is invariant under the addition of a soliton. In other words, Q has exactly the same energy density as Q_0. This would suggest that the addition of a soliton to the solution does not appear to add anything. This is, however, not completely true as the energy momentum density does not characterize a given solution completely. To characterize it completely we must also take into account the "internal" degrees of freedom. Before we investigate this question further let us compute the energy density for the solution Q_0. From (1.11) and (2.2) it is easy to show that

$$E=\mathrm{Tr}[AA]+\mathrm{Tr}[BB]. \tag{2.9}$$

Once again we see that the solutions have a constant energy density and thus the total energy if infinite.

Is it possible to construct a finite energy solution for the Minkowskian sigma model? The answer to this question is positive. All we have to do is to recall that the two-dimensional sigma models are conformally invariant. This means that

$$Q_0=\exp\{(i/2)[Af(\xi)+Bg(\eta)]\}, \tag{2.10}$$

where f and g are any real functions and A and B are given by Eq. (2.3), satisfies (1.6) and the energy density for this class of solutions is given by

$$E=(\partial_\xi f)^2\mathrm{Tr}[AA]+(\partial_\eta g)^2\mathrm{Tr}[BB]. \tag{2.11}$$

Finite energy solutions can thus be obtained by choosing f and g appropriately. For example we can choose for f the hyperbolic tangent of any polynomial in ξ, and for g a similar function but dependent on η. The figure (Fig. 1) exhibits the energy density of the solution for which A and B have been chosen as in (2.3) and

$$f=\tanh(\xi), \quad g=\tanh(\eta). \tag{2.12}$$

This solution is given by (2.4) where

$$x=\tfrac{1}{2}[\tanh(\xi)+\tanh(\eta)]. \tag{2.13}$$

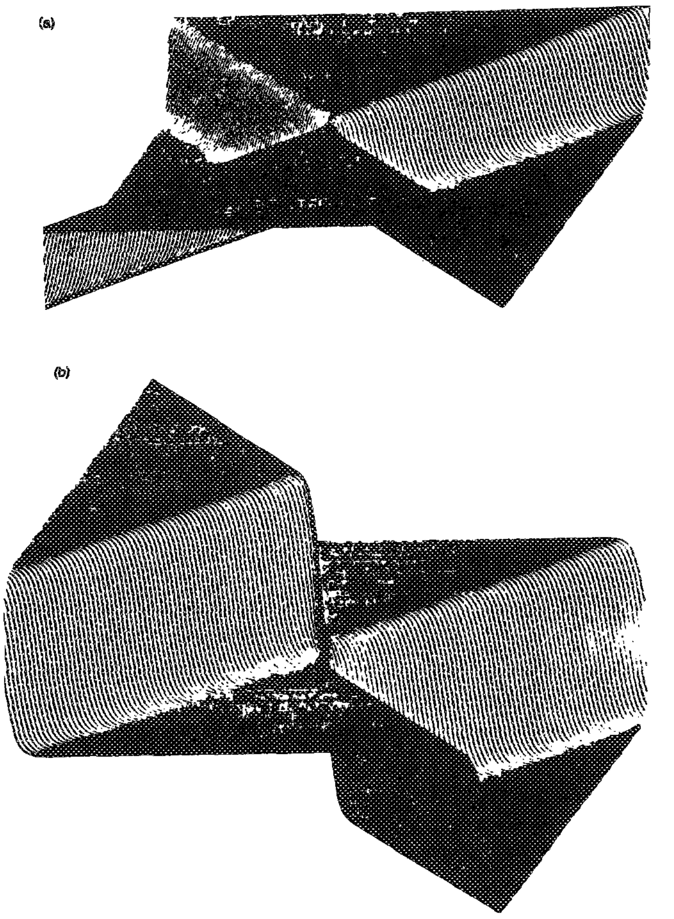


FIG. 2.(a) Finite energy solution (2.10): Q_{11}. (b) Finite energy solution (2.10): Q_{12}.

The matrix elements of this solution are represented in Fig. 2. Looking at the energy density and the graphs of the matrix elements we see that these solutions do indeed represent two solitons propagating in opposite directions with the speed of light.

We can now, using the method of Harnad *et al.*, add a "soliton" to this solution. Such a solution is given by (2.6) in which one has to replace ξ by $\tanh(\xi)$ and η by $\tanh(\eta)$, and x by (2.13). The analytical expression for this solution is much more complicated than (2.10) but we know that they both have the same energy density (2.11). The matrix elements, on the other hand, are slightly different. In Figs. 3 and 4 we show matrix elements for this solution for different values of λ. Once again the matrix elements of this solution exhibit two solitons that propagate in opposite directions with the speed of light. We observe that this solution looks like a λ-dependent mixture of the matrix elements of (2.10).

Let us stress that this is a rather different result from the one we have obtained by adding a soliton to the vacuum solutions (2.2). In that case the addition of a soliton has really added a breathing structure that propagates with a speed that depends on λ.[18,19] We see that, as in the Euclidean case, the requirement that the total energy is finite provides some constraints on the properties of solutions.

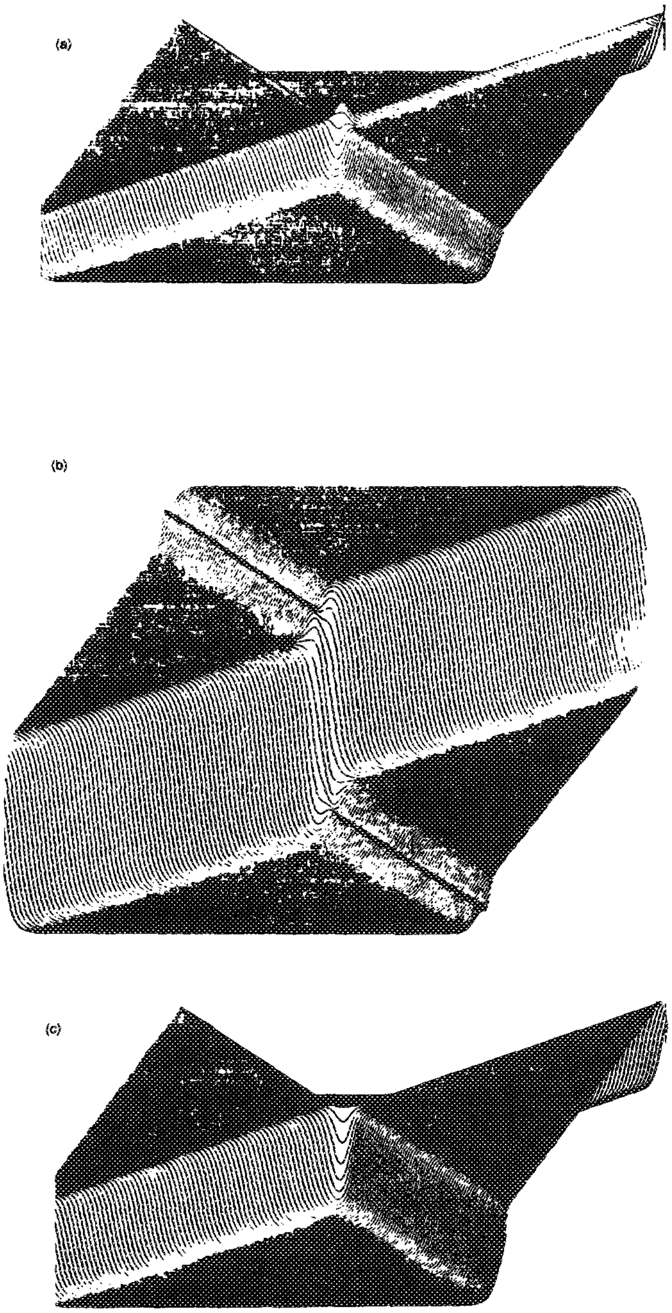


FIG. 3.(a) Finite energy solution (2.6) and (2.12) for $\lambda = -1 + i$: Q_{11}. (b) Finite energy solution (2.6) and (2.12) for $\lambda = -1 + i$:Re Q_{12}. (c) Finite energy solution (2.6) and (2.12) for $\lambda = -1 + i$:Im Q_{12}.

How can we thus characterize all finite energy solutions? The energy density by itself is not sufficient as there exist many different solutions that have exactly the same energy density. Integrable models are usually characterized by an infinite number of conservation laws. Two-dimensional sigma models do have an infinite number of such conserved quantities.[1,6] What we should do is then to compute explicitly the conserved currents (1.9) for the solutions we have analyzed so far and compare the expressions for these currents to see if there is any difference between them. Unfortunately computing these currents for

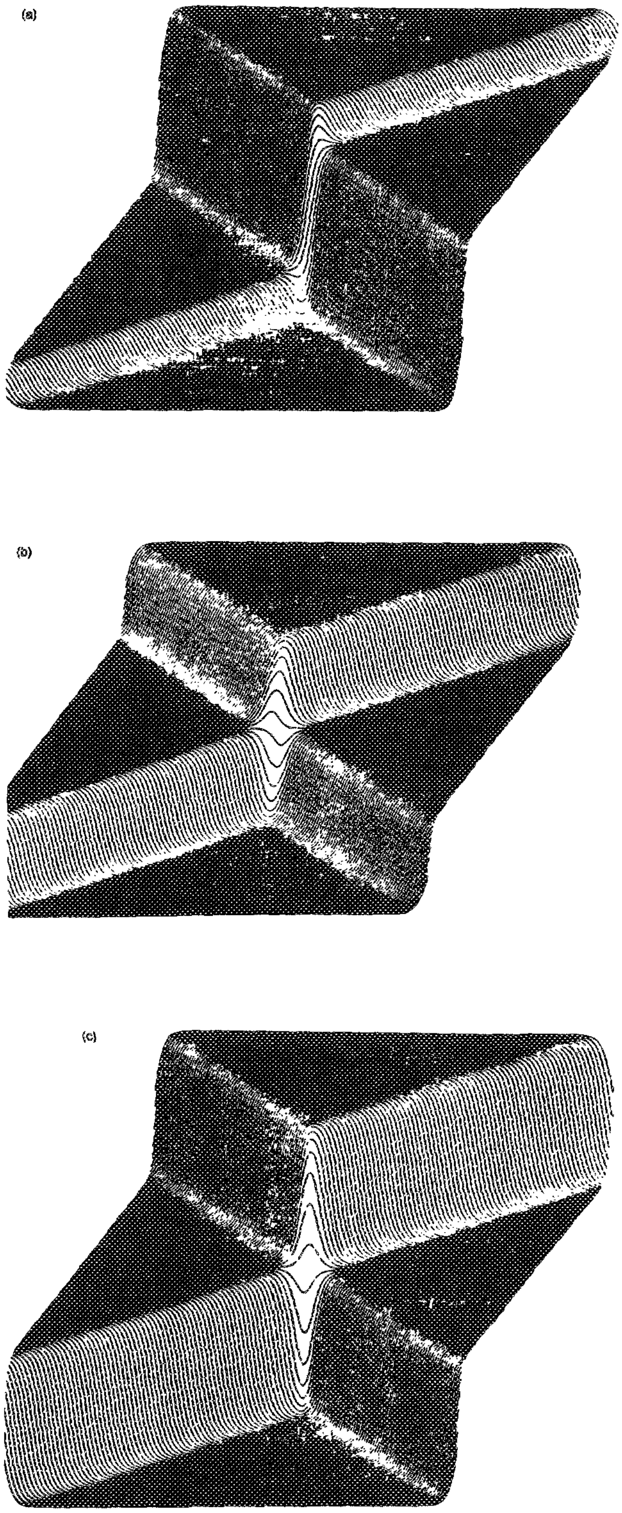


FIG. 4. Finite energy solution (2.6) and (2.12) for $\lambda = -0.8 + 2i{:}Q_{11}$. (b) Finite energy solution (2.6) and (2.12) for $\lambda = -0.8 + 2i{:}\mathrm{Re}Q_{12}$. (c) Finite energy solution (2.6) and (2.12) for $\lambda = -0.8 + i{:}\mathrm{Im}Q_{12}$.

the solutions (2.6), (2.13) is too tedious to be performed in practice, but it seems likely that these currents are modified by the addition of a soliton. Nevertheless, we can consider a class of simpler solutions having the same energy density and compare their currents.

The solutions we can consider are the solutions (2.10) with A and B given by constant Hermitian matrices. The one parameter family of conserved currents (1.9) for this class of solutions is given by

$$j_\xi(\omega) = [(1-\omega)/2(1+\omega)]iA\,\partial_\xi f,$$
$$j_\eta(\omega) = [(1+\omega)/2(1-\omega)]iB\,\partial_\eta g. \tag{2.14}$$

For this particular class of solutions, we observe that there are actually only two independent conserved currents (up to a common factor). However, these two currents give us all the information we need about the matrices A and B and the functions f and g, and determine the solution uniquely up to a constant factor. Thus we see that the currents contain information that is not contained in the energy density.

As the energy density of a solution remains unchanged after the addition of a soliton we can interpret this construction as a modification of the internal degrees of freedom of the solution. Moreover, when we look at the graphs of the matrix elements of our explicit solutions (Figs. 2–4), we see that the shape of the solitons is different after the scattering. This means that solitons interact with each other in such a way that in their interaction they modify only their internal (or local) degrees of freedom.

We have already mentioned that the set of the finite energy solutions of the $\mathbb{C}P^{N-1}$ sigma models splits into different subsets.[11] Does the addition of a soliton to the Minkowskian model give a similar structure to these solutions. The answer to this question is negative. The different subsets of Euclidean solutions are really disconnected, whereas the one soliton solution (2.6) (which is actually a solution of the $\mathbb{C}P^1$ model[19]), which we have discussed before, goes over to the solution (2.4) in the limit of λ going to $\infty + 0i$. Thus we see that this family of one uniton solutions is a continuous deformation of the vacuum solution (2.10).

III. CONSTRUCTION OF SOLUTIONS OF THE HYPERBOLIC COMPLEX SIGMA MODELS

In the previous section we have seen that the so-called procedure of the addition of a soliton to a finite action solution does not really add a soliton to this solution as the energy density of the field configuration remains unchanged during the process. Is there any method that could possibly add a real soliton to any finite energy solution? For the Euclidean model this is the case as the Uhlenbeck construction modifies a solution in a nontrivial way by adding unitons to it. These unitons correspond to real peaks in the energy density. Both the multi-soliton method and the Uhlenbeck method are based on the Lax pair problem (1.7) for the model, but they have been developed, respectively, for the Minkowskian and the Euclidean models. Can one adapt either one of them to the other class of models?

The multi-soliton construction has already been modified to permit the construction of Euclidean solutions.[16,17] When we compare the procedure developed for the construction of the finite energy solutions of the $\mathbb{C}P^{N-1}$

model[11] (the Uhlenbeck approach can be thought of as its further generalization), it is clear that the solitonic method corresponds to a rather trivial modification of the solutions.[16]

It is possible to go the other way and adapt the Uhlenbeck construction to the construction of Minkowskian solutions of some sigma models. The original method developed for the $\mathbb{C}P^{N-1}$ model has already been adapted[20] to the construction of Minkowskian solutions of the so-called hyperbolic complex Grassmannian models. Following this generalization, we will now modify the Uhlenbeck approach to the construction of Minkowskian solutions for the fields valued in hyperbolic complex unitary groups.[20]

The central point of the Uhlenbeck construction depends on the use of complex coordinates in the Euclidean plane. Light-cone coordinates are commonly used for two-dimensional Minkowskian problems, but the fact that they are not related to each other by the complex conjugation makes them less convenient for an Uhlenbeck-like construction. The idea put forward in Ref. 20 was to use another type of coordinates in the Minkowskian plane, namely the hyperbolic complex coordinates defined by

$$z_+ = x + \epsilon t, \quad z_- = x - \epsilon t, \tag{3.1}$$

where by definition ϵ satisfies

$$\epsilon^2 = 1, \quad \overline{\epsilon} = -\epsilon, \tag{3.2}$$

and where $\overline{}$ stands for the hyperbolic complex conjugation. All the algebraic properties of such hyperbolic complex numbers were described in detail in Ref. 20. Here let us just state that these numbers can be manipulated very much like the usual complex numbers. Thus a hyperbolic complex number a can be written as $a = r + \epsilon h$, where r and h are two real numbers representing, respectively, the real and the hyperbolic-imaginary parts of that number. The hyperbolic complex conjugation is defined by $\overline{a} = r - \epsilon h$ and the "norm" of a is given by $|a|^2 = \overline{a}a = r^2h^2$. Note that this "norm" is not positive definite.

As for complex numbers, rather than using the formal constant ϵ, we can use matrices to represent the hyperbolic numbers. Thus we can write

$$a = \begin{pmatrix} r & h \\ h & r \end{pmatrix}. \tag{3.3}$$

The product of two hyperbolic numbers can then be performed by the usual matrix multiplication and the hyperbolic complex conjugation is given by the conjugation by the matrix

$$\eta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{3.4}$$

i.e.,

$$\overline{a} = \eta a \eta = \begin{pmatrix} r & -h \\ -h & r \end{pmatrix}. \tag{3.5}$$

The fields of the sigma model we want to consider can be taken to be given by a square matrix Q whose entries are all hyperbolic complex numbers, and which satisfies $Q^\dagger Q = 1$, where $Q^\dagger$ now stands for $\overline{Q}^{\dagger}$. The action and the equations of motion of the model are still given by (1.1) and (1.3), respectively. Using the variables (3.1), we define

$$A_+ = \tfrac{1}{2}Q^\dagger \partial_{z_+} Q, \quad A_- = \tfrac{1}{2}Q^\dagger \partial_{z_-} Q, \tag{3.6}$$

and rewrite the action and the equations of motion as

$$S = -4 \int dx^2 \, \mathrm{Tr}[A_+ A_-], \tag{3.7}$$

$$\partial_{z_+} A_- + \partial_{z_-} A_+ = 0. \tag{3.8}$$

The Uhlenbeck procedure now corresponds to the following: assume that Q is a solution of (1.3) and that R is a projector that satisfies

$$\begin{aligned} &RA_-(1-R) = 0, \\ &(1-R)(\partial_{z_-} R + A_- R) = 0. \end{aligned} \tag{3.9}$$

Then

$$\overline{Q} = Q(1-2R) \tag{3.10}$$

is a new solution of (1.3). To prove this it is sufficient to observe that $\overline{A}_- = (1/2)\overline{Q}^\dagger \partial_{z_-} \overline{Q} = A_- + \partial_- R$. Using the fact that $A_+ = -A_-^\dagger$ we see that $\overline{A}_-$ and $\overline{A}_+$ also satisfy (3.8).

Having adapted the Uhlenbeck procedure we can now look for some simple solutions and then add unitons to them. However, before we do this let us observe that the energy momentum tensor is given by

$$\begin{aligned} &T^{00} = 2\,\mathrm{Tr}[A_+ A_+ + A_- A_-], \\ &T^{10} = 2\,\mathrm{Tr}[A_+ A_+ - A_- A_-], \end{aligned} \tag{3.11}$$

and is not modified by the addition of a uniton. To see this we observe that from Eqs. (3.9) it follows that $\mathrm{Tr}[\overline{A}_- \overline{A}_-] = \mathrm{Tr}[A_- A_-]$ and so that the energy momentum density of the original solution and of the solution obtained from it by the addition of a uniton are the same. Thus, as far as this property is concerned, the Minkowskian version of the Uhlenbeck procedure is very similar to the methods of Harnad *et al.*

It is easy to construct many nontrivial solutions of this model. For example, all solutions of the Euclidean unitary sigma models constructed in Ref. 14 can be transformed into solutions of our hyperbolic model by replacing the imaginary constant i by the hyperbolic equivalent ϵ. Unfortunately all these solutions were obtained by adding unitons to constant solutions for which the energy momentum density (3.11) is identically zero. Thus all these solutions describe vacuum field configurations.

However, some finite energy solutions can be constructed very easily. Take for example

$$Q = \exp(\epsilon\{A[f(z_+) + f(z_-)] + B[g(z_+) + g(z_-)]\}), \tag{3.12}$$

where A and b are two Hermitian matrices that commute with each other, and f and g are any functions that can be so chosen that the total energy is finite. For example we can take $B = 0$, $f = \tanh$ and

$$A=\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.13}$$

For this very special choice, Q is given by

$$Q=\begin{pmatrix} \mathrm{ch}^2+\epsilon \mathrm{ch}\,\mathrm{sh} & \mathrm{sh}^2+\epsilon \mathrm{sh}\,\mathrm{ch} & 0 \\ \mathrm{sh}^2+\epsilon \mathrm{sh}\,\mathrm{ch} & \mathrm{ch}^2+\epsilon \mathrm{ch}\,\mathrm{sh} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.14}$$

where $\mathrm{ch}=\cosh(f(z_+) + f(z_-))$ and $\mathrm{sh}=\sinh(f(z_+) + f(z_-))$ (see Ref. 20 to see how to compute a function of hyperbolic complex numbers). The use of the hyperbolic constant ϵ might seem odd, but it is easy to express this solution in terms of real matrices by using the matrix representation (3.3) of the hyperbolic complex number. In this case Q becomes a 6×6 real matrix:

$$Q=\begin{bmatrix} \mathrm{ch}^2 & \mathrm{ch}\,\mathrm{sh} & \mathrm{sh}^2 & \mathrm{ch}\,\mathrm{sh} & 0 & 0 \\ \mathrm{ch}\,\mathrm{sh} & \mathrm{ch}^2 & \mathrm{ch}\,\mathrm{sh} & \mathrm{sh}^2 & 0 & 0 \\ \mathrm{sh}^2 & \mathrm{ch}\,\mathrm{sh} & \mathrm{ch}^2 & \mathrm{ch}\,\mathrm{sh} & 0 & 0 \\ \mathrm{ch}\,\mathrm{sh} & \mathrm{sh}^2 & \mathrm{ch}\,\mathrm{sh} & \mathrm{ch}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.15}$$

The energy density (3.11) for this solution is simply given by

$$E=[\cosh(x+t)]^{-2}+[\cosh(x-t)]^{-2}, \tag{3.16}$$

which shows that the total energy of this solution is finite, and that this solution corresponds to two lumps which propagate in opposite directions with the speed of light.

We can now try to add a uniton to this particular solution. First we calculate the gauge field

$$A_-=(\epsilon/2)A\partial_{z_-}f(z_-) \tag{3.17}$$

for this solution, and then solve the Uhlenbeck equations (3.9). It is easy to check that in this particular case a solution of the Uhlenbeck equation is

$$R=vv^\dagger/v^\dagger v, \tag{3.18}$$

where

$$v=\begin{pmatrix} 1 \\ -1 \\ \alpha(z_+) \end{pmatrix}, \tag{3.19}$$

and where α is any hyperbolic-complex holomorphic function. Thus

$$\overline{Q}=Q(1-2R) \tag{3.20}$$

describes a family of one uniton solutions and, from what we have said before, all these solutions have exactly the same energy density (3.16) as Q. Thus $\overline{Q}$ can be considered as the solution (3.14) modified by the addition of a vacuum solution.

IV. CONCLUSIONS

One can construct classes of finite energy solutions of some two-dimensional sigma models by using simple Ansätze. The different methods for the construction of such nontrivial solutions, like the Bäcklund transformation or the Uhlenbeck procedure, either give solutions that have an infinite energy or modify a given solution in a nontrivial way but without modifying its energy momentum tensor density. To characterize completely a given solution one needs, in addition to the energy momentum tensor, also to consider other currents. These currents can be interpreted as providing us with a description of the internal degree of freedom of the solution and they are altered when one modifies the given solution by adding a soliton to it.

We have also looked in some detail at the time evolution of the solitonic structures exhibited by our solutions. Looking at various matrix elements that exhibit these structures we interpret our results as showing that the solitons of our two-dimensional sigma models scatter internally. This is because we find that during the scattering process their internal degrees of freedom do change while their energy density exhibits unperturbed evolution with the velocity of light.

ACKNOWLEDGMENTS

Some part of the work reported here was performed when B. P. visited the University of Durham. He would like to thank the University of Durham of its hospitality and J.-P. Antoine for making his visit possible. He would also like to thank the Fonds National de la Recherche Scientifique, Belgium and the Royal Society, UK for the financial support of his visit organized under the European Science Exchange Programme scheme.

We would also like to thank R. Ward for useful conversations.

[1] K. Pohlmeyer, Commun. Math. Phys. **46**, 207 (1976).

[2] See for example: R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin, New York, 1987).

[3] V. E. Zakharov and A. V. Mikhailov, Zh. Eksp. Teor. Fiz. **74**, 1953 (1978) [Sov. Phys. JETP **47**, 1017 (1979)].

[4] J. Harnad, Y. Saint-Aubin, and S. Shnider, Commun. Math. Phys. **92**, 329 (1984); **93**, 33 (1984).

[5] C. Rogers and W. F. Shadwick, *Bäcklund Transformations and Their Applications* (Academic, New York, 1982).

[6] H. Eichenherr and M. Forger, Nucl. Phys. B **155**, 381 (1979); Nucl. Phys. B **164**, 528 (1980); B **282**, 745(E) (1987).

[7] M. Lüscher and K. Pohlmeyer, Nucl. Phys. B **137**, 46 (1978).

[8] Y. Y. Goldshmidt and E. Witten, Phys. Lett. B **91**, 392 (1980).

[9] H. J. Borchers and W. D. Garber, Commun. Math. Phys. **72**, 77 (1980).

[10] W. J. Zakrzewski, Geom. Phys. **1**, 39 (1984).

[11] A. M. Din and W. J. Zakrzewski, Nucl. Phys. B **174**, 397 (1980); Phys. Lett. B **95**, 419 (1980).

[12] B. Piette and W. J. Zakrzewski, Nucl. Phys. B **300**, 207 (1980).

[13] K. Uhlenbeck, University of Chicago preprint (1985) to be published in J. Differential Geometry.

[14] B. Piett and W. J. Zakrzewski, Durham University preprint DTP-29/87.

[15] B. Piette and W. J. Zakrzewski, Nucl. Phys. B **300**, 223 (1988).

[16] R. Sasaki, J. Math. Phys. **28**, 1786 (1985).

[17] J.-P. Antoine and B. Piette, J. Math. Phys. **29**, 1687 (1988).

[18] Y. Saint-Aubin, Lett. Math. Phys. **6**, 441 (1983).

[19] B. Piette, J. Math. Phys. **29**, 2190 (1988).

[20] D. Lambert and B. Piette, Class. Quantum Grav. **5**, 307 (1988).

Fourier transforms on $\mathscr{A}/\mathscr{G}$ and knot invariants

Shahn Majid
Department of Mathematics and Computer Science, University College of Swansea, Singleton Park, Swansea SA2 8PP, United Kingdom

(Received 15 March 1989; accepted for publication 27 September 1989)

Fourier transformation on $\mathscr{A}/\mathscr{G}$ leads to some elementary insight into Witten's expression for the Jones polynomial.

I. INTRODUCTION

This paper is in response to the interest among knot theorists to gain insight into the complex nature of Witten's recent formulation of the Jones polynomial[1] and related invariants. It is based on some previous unpublished work[2] on "functional Fourier transforms" of quantum gauge field theories. It will be somewhat formal since we do not attempt to define the measure for path integration on the space $\mathscr{A}/\mathscr{G}$ of connections modulo gauge transformations. From the point of view of theoretical physics, Witten has indicated that his approach to knot invariants could be relevant to physics beyond the Planck scale and indeed, in the case of non-Abelian gauge fields the dependence of the knot invariant on the framing appears to come out of conformal field theory.[1] There is a certain amount to be cleared up here even in the U(1) case that we look at in detail.

Fourier transforms may be performed on arbitrary locally compact Abelian groups ω as follows. Let $\hat{\omega} = \mathrm{Hom}(\omega, S^1)$ be the group of continuous characters of ω. These form a group (the Pontryagin dual) by pointwise multiplication and $\hat{\hat{\omega}} \cong \omega$ cf Ref. 3. If $f \in C(\hat{\omega}, \mathbb{C})$ (the C^* algebra of continuous complex-valued functions on $\hat{\omega}$ that vanish at infinity), the Fourier transform of f is defined as

$$\tilde{f}(g) = \int_{\hat{\omega}} d\chi\, \chi(g) f(\chi), \tag{1}$$

where $d\chi$ denotes the left-invariant Haar measure on $\hat{\omega}$. The Fourier transform map is an isomorphism between the C^* algebra $C(\hat{\omega}, \mathbb{C})$ and $C^*(\omega, \mathbb{C})$ (the C^* algebra associated to functions on ω with convolution product).[3]

II. FOURIER TRANSFORM OF CHERN–SIMONS FUNCTIONAL

In this paper we apply such a Fourier transform *formally* to the following groups. Let M be an oriented three-manifold with $H^1(M) = 0$. If M is not compact, we implicitly assume suitable boundary conditions for all fields. Let

$$\omega = \frac{\text{free Abelian group on oriented knots in } M}{\text{erasure of overlaps of opposite orientation}}. \tag{2}$$

We suppose that the knots are piecewise smooth. Let $\mathscr{A}/\mathscr{G}(\mathrm{U}(1))$ denote

$$\frac{\{(L,A); L \in \mathrm{U}(1) \text{ bundles over } M, A \in \text{connections on } L\}}{\text{local gauge transformations}}. \tag{3}$$

This forms a group under pointwise addition of connections *and multiplication of bundle transition functions*. The group $\mathscr{A}/\mathscr{G}(\mathrm{U}(1))$ is essentially the dual of ω [more precisely, $\mathscr{A}/\mathscr{G}(\mathrm{U}(1)) \subseteq \hat{\omega}$, $\omega \subseteq \widehat{\mathscr{A}/\mathscr{G}}(\mathrm{U}(1))$] according to the pairing

$$\langle A, \kappa \rangle = \exp \iota \left(\int_\kappa A \right). \tag{4}$$

We have adopted conventions in which the $U(1)$ connection is written ιA so that A is real in local coordinates. These groups are not naturally locally compact, so there is, in fact, no left-invariant Haar measure. However, the measure $\mathscr{D}[A]$ formally used by physicists on $\mathscr{A}/\mathscr{G}$, essentially

$$\mathscr{D}[A] = \Pi_{x \in M} dA(x) \tag{5}$$

(times a gauge-fixing factor to take care of the fact that we only wish to integrate over a quotient space of $\mathscr{A}$) i.e., essentially a product of Lebesgue measures on the variables $A(x)$, is designed to be formally left invariant for the pointwise addition on $\mathscr{A}/\mathscr{G}(\mathrm{U}(1))$ described.

We are, therefore, formally in a position to Fourier transform interesting functions on $\mathscr{A}/\mathscr{G}(\mathrm{U}(1))$ to functions on knots. Thus we have the following theorem.[2]

Theorem 1: The Fourier transform of the exponential of the Chern–Simons functional $\alpha/2 \int_M A \wedge dA$ is the exponential of the self-linking number of the knot,

$$\begin{aligned}\widetilde{\mathrm{CS}}(\kappa) &= \int \mathscr{D}[A] \exp\left(\iota \int_\kappa A\right) \exp\left(\frac{\iota\alpha}{2} \int_M A \wedge da\right)\\ &= \widetilde{\mathrm{CS}}(\phi) \exp[-(\iota/2\alpha)\mathrm{link}(\kappa,\kappa)],\end{aligned}$$

where ϕ denotes the empty set (the identity in ω).

Before giving our elementary proof of this, we need a lemma on the self-linking number. Note that the linking number $\mathrm{link}(\kappa_1,\kappa_2)$ is defined for disjoint knots κ_1,κ_2. We define the self-linking number in $\mathbb{R}^3$ to be $\lim_{e\to 0}[1/\mathrm{Vol}(B_\epsilon)]\int_{B_\epsilon} d\vec{\epsilon}\, \mathrm{link}(\kappa,\kappa_{\vec{\epsilon}})$, where $\kappa_{\vec{\epsilon}}$ denotes κ displaced uniformly by the vector $\vec{\epsilon}$ and B_ϵ is the ball of radius ϵ. One may expect that link $(\kappa,\kappa_{\vec{\epsilon}})$ is then defined for almost all $\vec{\epsilon}$ so that the integral is well defined. Further, generically, the limit link (κ,κ) exists. For a general manifold we envisage a similar definition using a metric connection. Note that link (κ,κ) is not necessarily an integer and not a diffeomorphism invariant. Of course, we can think of link $(\kappa,\kappa_{\vec{\epsilon}})$ as an invariant of the pair $(\kappa,\vec{\epsilon})$. Where defined, linking number behaves biadditively and self-linking number behaves quadratically with respect to the group structure on ω.

Lemma 2: Let κ be a knot in M. Consider an infinitely thin solenoid wound along the knot such that the magnetic field produced by the solenoid has unit strength (and points

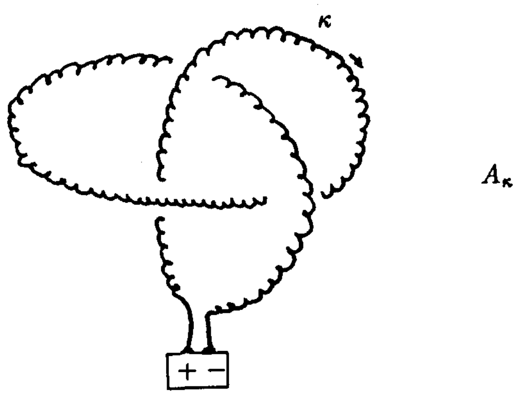


FIG. 1. The vector potential A_κ.

tangentially to the knot). This is depicted in Fig. 1. Let A_κ denote the vector potential for this electromagnetic configuration. [For concreteness, we suppose A_κ is fixed uniquely by the further condition $d*A_\kappa = 0$ (Coulomb gauge).] Then

$$\int_M A_{\kappa_1} \wedge dA_{\kappa_2} = \text{link}(\kappa_1,\kappa_2), \quad \int_M A_\kappa \wedge dA_\kappa = \text{link}(\kappa,\kappa).$$

Proof: Mathematically, the magnetic field $*F_\kappa$ (corresponding to curvature F_κ) is characterized as the (distributional) form on M such that

$$\int_\kappa B = \int_M F_\kappa \wedge B, \quad \forall B \in \wedge^1(M)$$

[cf. the Poincare dual of κ (Ref. 4) except that we do not suppose $dB = 0$]. In particular,

$$\int_M (dF_\kappa)\phi = -\int_M F_\kappa \wedge d\phi = -\int_{\partial\kappa} \phi = 0, \qquad \forall \phi \in \wedge^0(M).$$

Hence, $dF_\kappa = 0$. If $H^1(M) = 0$ then $F_\kappa = dA_\kappa$ and if in addition, M is compact, A_κ is uniquely fixed by $d*A_\kappa = 0$ using Hodge theory. [In fact, one can see, in general, that $[F_\kappa] \in H^2(M,\mathbb{Z})$ so that there exists a U(1) connection A_κ with curvature F_κ.] Then,

$$\int_M dA_{\kappa_1} \wedge A_{\kappa_2} = \int_{\kappa_1} A_{\kappa_2} = \int_{\text{span}\,\kappa_1} dA_{\kappa_2}.$$

Now dA_{κ_2} is the curvature corresponding to a magnetic field along κ_2, with δ-function cross section: The only points in the span of κ_1 that contribute to the integral are those at which κ_2 intersects span κ_1, which contribute ± 1 according to the orientation. □

Proof of Theorem 1: Using this lemma and integration by parts, we have

$$\begin{aligned}\widetilde{\text{CS}}(\kappa) &= \int \mathscr{D}[A]\exp\Big(\iota\int_\kappa A\Big)\exp\Big(\frac{\iota\alpha}{2}\int_M A\wedge dA\Big) = \int \mathscr{D}[A]\exp\Big(\iota\int_M dA_\kappa \wedge A + \frac{\alpha}{2} A\wedge dA\Big)\\ &= \int \mathscr{D}[A]\exp\Big[\frac{\iota\alpha}{2}\int_M\Big(A+\frac{A_\kappa}{\alpha}\Big)\wedge\Big(A+\frac{A_\kappa}{\alpha}\Big)\Big]\exp\Big(-\frac{\iota}{2\alpha}\int_M A_\kappa\wedge dA_\kappa\Big)\\ &= \widetilde{\text{CS}}(\phi)\exp[-(\iota/2\alpha)\text{link}(\kappa,\kappa)],\end{aligned}$$

where in the last line we used formal translation invariance of the measure $\mathscr{D}[A]$ to change variable of integration, $\mathscr{D}[A] = \mathscr{D}[A + A_\kappa/\alpha]$, and Lemma 2. □

The proof given is a generalization of the familiar fact that the Fourier transform of a Gaussian $e^{(\iota\alpha/2)x^TQx}$ is proportional to a Gaussian, $e^{-(\iota/2\alpha)p^TQ^{-1}p}$. Here Q is a symmetric invertible $n\times n$ matrix, $x\in\hat{\omega} = \mathbb{R}^n$, $p\in\omega = \mathbb{R}^n$ and these two groups are dual according to the pairing $\langle x,p\rangle = e^{\iota x^Tp}$.

III. FRAMING DEPENDENCE AND DISCUSSION

In the language of Witten,[1] Theorem 1 says $\langle 0|\exp(\iota\int_\kappa A)|0\rangle = \langle 0|0\rangle\exp - (\iota/2\alpha)\text{link}(\kappa,\kappa)$. If κ_1,κ_2 both have vanishing self-linking number (e.g., if they are planar in $\mathbb{R}^3$), then

$$\text{link}(\kappa_1+\kappa_2,\kappa_1+\kappa_2) = 0 + 2\text{link}(\kappa_1,\kappa_2) + 0,$$

so that

$$\begin{aligned}&\Big\langle 0\Big|\exp\Big(\iota\int_{\kappa_1} A\Big)\exp\Big(\iota\int_{\kappa_2} A\Big)\Big|0\Big\rangle\\ &\quad = \Big\langle 0\Big|\exp\int_{\kappa_1+\kappa_2} A\Big|0\Big\rangle\\ &\quad = \langle 0|0\rangle\exp\Big[-\frac{\iota}{\alpha}\text{link}(\kappa_1,\kappa_2)\Big]\end{aligned}$$

(cf Ref. 1). However, it should be stressed that *this is only true for κ_i of zero self-linking number*: The function on knots constructed by the U(1) quantum field theory is the self-linking number, rather than a link invariant.

The above now suggests an interpretation of the more general quantum field theory of $\mathscr{A}/\mathscr{G}$ with non-Abelian structure group G (connections on an associated vector bundle, associated by a representation of G) as studied by Witten.[1] Now $\kappa\mapsto\text{Tr}\exp(\iota\int_\kappa A)$ is not a character of ω, but because of trace identities in particular representations of particular groups [such as SU(2)], there are relations between $\text{Tr}\exp(\iota\int_{\kappa_1}A)\exp(\iota\int_{\kappa_2}A)$ and $\text{Tr}\exp(\iota\int_{\kappa_1}A)$, $\text{Tr}\exp(\iota\int_{\kappa_2}A)$. Second, $\mathscr{A}/\mathscr{G}$ is not a group. Third, even if we could put some kind of group structure on it, it would not be commutative. Thus ω would not be its dual group: we should work with a suitable noncommutative analog. It is possible that these problems can be solved by putting a Hopf algebra structure on formal linear combinations of points in $\mathscr{A}/\mathscr{G}$ (a point consisting of a connection and a representation). For Hopf algebras there is a dual Hopf algebra[5] and notions of Fourier transform precisely generalizing those for Albelian groups referred to in Sec. I and used in Sec. II.

Finally, the Chern–Simons functional is no longer Gaussian, having a cubic A^3 term, which further complicates the approach of Sec. II. As a result of these complications, one does not obtain a self-linking number, but a generalization. In the case of $G = \mathrm{SU}(2)$ – spin $\frac{1}{2}$, we apparently obtain the Jones invariant.[1] Note however, that as it stands, even the self-linking number is not an actual invariant: rather, one must think of link $(\kappa,\kappa_{\vec{\epsilon}})$ as an invariant of the pair $(\kappa,\vec{\epsilon})$.

As such, we do not need to work with uniform $\vec{\epsilon}$. A pair $(\kappa,\vec{\eta})$ is a *framed knot* if κ is a knot and $\vec{\eta}$ is a nonvanishing nontangential vector field defined on κ. To be concrete, we suppose that at each point on κ, $\vec{\eta}$ has uniform norm $\eta > 0$ and is orthogonal to $\dot{\kappa}$, both with respect to a fixed Riemannian metric on M. Witten then proposes[1] that the quantum field theory in the U(1) case, and its generalizations, can be somehow modified by the chosen $\vec{\eta}$, such as to obtain an invariant of framed knots. Thus in the U(1) case we have

$$\left\langle 0 \middle| \exp\left(\iota \int_\kappa A\right) \middle| 0 \right\rangle_{\vec{\eta}} = \langle 0|0\rangle \exp\left[-\frac{\iota}{2\alpha} \operatorname{link}(\kappa,\kappa_{\vec{\eta}}) \right]. \tag{6}$$

Next it is claimed that the framing dependence of these framed knot inavariants is particularly simple and can be factored out to obtain something that depends only on the knot. We now analyze this issue in the case of $G = \mathrm{U}(1)$ and knots in $M = \mathbb{R}^3$.

Indeed, the self-linking number in $\mathbb{R}^3$ was previously studied in Ref. 6 (in another context), where it is called the *writhing number*. Reference 6 showed that for sufficiently small η

$$\operatorname{link}(\kappa,\kappa_{\vec{\eta}}) = \operatorname{twist}(\kappa,\vec{\eta}) + \operatorname{link}(\kappa,\kappa),$$

where

$$\operatorname{twist}(\kappa,\vec{\eta}) = \int_\kappa dt \, \frac{(\dot{\vec{\eta}} \times \vec{\eta}) \cdot \dot{\kappa}}{\eta^2 |\dot{\kappa}|}$$

measures the twist of $\vec{\eta}$ about κ. Note that the twist is not necessarily an integer and not a diffeomorphism invariant. Thus increasing the twist of $\vec{\eta}$ by 1 changes $\langle 0|\exp(\iota \int_\kappa A)|0\rangle_{\vec{\eta}}/\langle 0|0\rangle$ by a factor of $q = e^{-\iota/2\alpha}$ (cf. Ref. 1 with $\alpha = -k/4\pi$). Unfortunately, any attempt to divide by $q^{\operatorname{twist}(\kappa,\vec{\eta})}$ results only in $\operatorname{link}(\kappa,\kappa)$. This is not an invariant. It is not clear that the SU(2) case would work any better.

Instead of starting with framed knots invariants, I would like to propose the following variation, motivated by Ref. 7. It was argued in Refs. 2 and 6 that the self-linking number in $\mathbb{R}^3$ coincides with the *oriented self-crossing number*, cross $(\kappa,\vec{\epsilon})$, averaged over all projections $\vec{\epsilon}$ onto $\mathbb{R}^2$. Here $\vec{\epsilon}$ is again a uniform vector of norm $\epsilon > 0$. Viewing the knot from a generic direction $\vec{\epsilon}$ we define the oriented self-crossing number as the number of points at which the knot crosses itself, counted ± 1 according to orientation; $\overset{\uparrow}{\underset{|}{\rightarrow}}$ is $+1$. Instead of Eq. (6), we suppose that the quantum field theory is modified according to $\vec{\epsilon}$ such that,

$$\left\langle 0 | \exp\left(\iota \int_K A\right) | 0 \right\rangle_{\vec{\epsilon}} = \langle 0|0\rangle \exp\left[-\frac{\iota}{2\alpha} \operatorname{link}(\kappa,\kappa_{\vec{\epsilon}}) \right] \tag{7}$$

for $\epsilon > 0$ sufficiently small. Again, this depends on $\vec{\epsilon}$. Indeed, $\operatorname{link}(\kappa,\kappa_{\vec{\epsilon}}) = \operatorname{cross}(\kappa,\vec{\epsilon})$ for ϵ sufficiently small. Thus dividing $\langle 0|\exp(\iota \int_\kappa A)|0\rangle_{\vec{\epsilon}}/\langle 0|0\rangle$ through by $q^{\operatorname{cross}(\kappa,\vec{\epsilon})}$ gives a knot invariant, the identity function.

Similarly, in the case of SU(2)-spin $\frac{1}{2}$, one may expect that $\langle 0|\exp(\iota \int_\kappa A)|0\rangle_{\vec{\epsilon}}/\langle 0|0\rangle$ has a similar ϵ dependence. Dividing through by $q^{\operatorname{cross}(\kappa,\vec{\epsilon})}$ one would then obtain a knot invariant, the Jones invariant, cf. the state models of Ref. 7.

Thus the Jones invariant should be regarded as the ratio of the "exponentiated SU(2)-spin$\frac{1}{2}$-self-crossing number in projection $\vec{\epsilon}$" to the exponentiated U(1)-self-crossing number in projection $\vec{\epsilon}$, obtained respectively by "Fourier transforming" on $\mathscr{A}/\mathscr{G}(\mathrm{SU}(2))$ and on $\mathscr{A}/\mathscr{G}(U(1))$, both theories modified by a choice of $\vec{\epsilon}$, which cancels out in the ratio.

Secondly, returning to framed knots, the above analysis also suggests that in the unmodified SU(2)-spin $\frac{1}{2}$ quantum field theory, $\langle 0|\exp(\iota \int_\kappa A)|0\rangle/\langle 0|0\rangle$ gives a well-defined function on knots, the exponentiated SU(2)-spin $\frac{1}{2}$ self-linking number, which can then be turned into a framed knot invariant by multiplying by an SU(2) analog of $q^{\operatorname{twist}(\kappa,\vec{\eta})}$. This is far from the line taken in Ref. 1, but presumably the definition of $\langle 0|\exp(\iota \int_\kappa A)|0\rangle_{\vec{\eta}}/\langle 0|0\rangle$ implicit in Ref. 1 could be reinterpreted as a definition, coming out of conformal field theory, of such an SU(2)-spin $\frac{1}{2}$ twist or $G - \rho$ twist for general group G and representation ρ.

IV. FOURIER TRANSFORM OF YANG–MILLS FUCTIONAL

In the framework of Sec. I, we can consider the Fourier transform of other functions on $\mathscr{A}/\mathscr{G}$. Thus we have the following theorem.[2]

Theorem 3: The Fourier transform of the exponential of the Yang–Mills functional $(\beta/2)\int_M *dA \wedge dA$ is the exponential of the self-inductance of the knot,

$$\begin{aligned}\widetilde{\mathrm{YM}}(\kappa) &= \int \mathscr{D}[A] \exp\left(\iota \int_\kappa A\right) \exp\left(\frac{\iota\beta}{2} \int_M *dA \wedge dA\right) \\ &= \widetilde{\mathrm{YM}}(\phi) \exp[-(\iota/2\beta) \operatorname{ind}(\kappa,\kappa)].\end{aligned}$$

The mutual inductance of two disjoint knots in $\mathbb{R}^3$ is defined by the formula

$$\operatorname{ind}(\kappa_1,\kappa_2) = \frac{1}{4\pi} \int_{\kappa_1} ds \int_{\kappa_2} dt \, \frac{\dot{\kappa}_1(s) \cdot \dot{\kappa}_2(t)}{|\kappa_1(s) - \kappa_2(t)|}. \tag{8}$$

This is analogous to the Gauss formula in $\mathbb{R}^3$ (Ref. 8),

$$\begin{aligned}&\operatorname{link}(\kappa_1,\kappa_2) \\ &\quad = \frac{1}{4\pi} \int_{\kappa_1} ds \int_{\kappa_2} dt \, \frac{(\kappa_1(s) - \kappa_2(t)) \cdot (\dot{\kappa}_1 \times \dot{\kappa}_2)}{|\kappa_1(s) - \kappa_2(t)|^3}.\end{aligned}$$

Physically, $\operatorname{ind}(\kappa_1,\kappa_2)$ is defined as the energy of interaction due the force between knots κ_1 and κ_2 carrying unit electric currents. Like the self-linking number, we define self-inductance as the average inductance between κ and $\kappa_{\vec{\epsilon}}$. Unlike the self-linking number, the self-inductance diverges as $\epsilon \to 0$. This infinity is well known to radio engineers: The self-inductance diverges as the thickness of the wire goes to zero. The infinity can presumably be handled by a renormalization procedure analogous to the renormalization of the Yang–Mills action in quantum field theory.[9] For our purposes we imagine an implicit finite $\epsilon \neq 0$. One could also regulate the infinity by introducing $\iota\epsilon$ in the denominator of Eq.

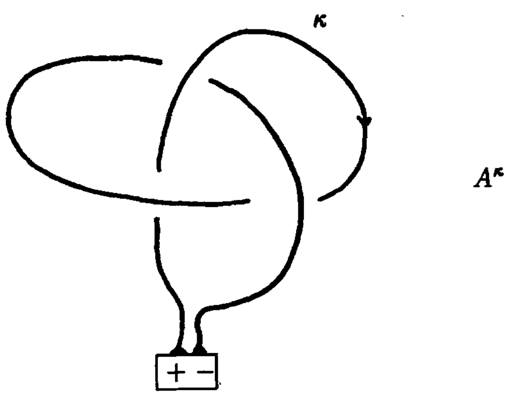


FIG. 2 The vector potential A^κ.

(8), more closely in analogy with the quantum field theory. Note that ind is biadditive.

Lemma 4: Let κ be a knot in M. Consider a current of unit strength flowing around the knot. This is depicted in Fig. 2. Let A^κ denote the potential for this electromagnetic configuration. [We suppose that A^κ is fixed by the condition $d*A^\kappa = 0$ (Coulomb gauge).] Then

$$\int_M *dA^{\kappa_1}\wedge dA^{\kappa_2} = \mathrm{ind}(\kappa_1,\kappa_2),$$

$$\int_M *dA^{\kappa}\wedge dA^{\kappa} = \mathrm{ind}(\kappa,\kappa).$$

Proof: The proof is straightforward. The current in κ_1 is a co-closed one-form, so by Hodge theory [e.g., if M is compact and using $H^1(M) = 0, H^2(M) = 0$] there is a magnetic field corresponding to a potential A^{κ_1} such that $*d*dA^{\kappa_1}$ is a unit vector along κ_1 with a δ-function cross section. Explicit formulas exist for this. The force experienced by κ_2 due to its current is then given by the Lorentz force formula. □

Proof of Theorem 3: We first observe from the proofs of Lemmas 2 and 4 that $*d*dA^\kappa = *dA_\kappa$, so that $*dA^\kappa = A_\kappa$ [plus an exact part, which we assume is forced to be zero since $d*(*dA^\kappa) = 0$.] We also assume integration by parts. Then we have

$$\begin{aligned}\widetilde{\mathrm{YM}}(\kappa) &= \int \mathscr{D}[A]\exp\left(\iota\int_\kappa A\right)\exp\left(\frac{\iota\beta}{2}\int_M *dA\wedge dA\right) = \int \mathscr{D}[A]\exp\left(\iota\int_M dA_\kappa\wedge A + \frac{\beta}{2}*dA\wedge dA\right)\\ &= \int \mathscr{D}[A]\exp\left(\iota\int_M d*dA^\kappa\wedge A + \frac{\beta}{2}*dA\wedge dA\right) = \int \mathscr{D}[A]\exp\left(\iota\int_M *dA^\kappa\wedge dA + \frac{\beta}{2}*dA\wedge dA\right)\\ &= \int \mathscr{D}[A]\exp\left[\frac{\iota\beta}{2}\int_M\left(*dA + \frac{*dA^\kappa}{\beta}\right)\wedge d\left(A + \frac{A^\kappa}{\beta}\right)\right]\exp\left(-\frac{\iota}{2\beta}\int_M *dA^\kappa\wedge dA^\kappa\right)\\ &= \widetilde{\mathrm{YM}}(\phi)\exp\left(-\frac{\iota}{2\beta}\mathrm{ind}(\kappa,\kappa)\right),\end{aligned}$$

where in the last line we used formal translation invariance of the measure $\mathscr{D}[A]$ to change variable of integration, $\mathscr{D}[A] = \mathscr{D}[A + A^\kappa/\beta]$, and Lemma 4. □

Rigorous definitions of the measure $\mathscr{D}[A]$ are known in the $U(1)$ case for trivial bundles, so that Sec. II and IV can presumably be stated rigorously. The measures are then not exactly translation invariant so that the theorems above must be stated with somewhat more care. The non-Abelian cases are expected to present somewhat more difficulty. In any event, it is hoped that the elementary observations of this note help to put Ref. 1 into an interesting context. Other aspects of Ref. 2, related to speculation about high energy physics, are available in Ref. 10.

Some of the results on Chern–Simons and linking-number have also been obtained independently (in another context) in the recent work[11] cited in Ref. 1.

ACKNOWLEDGMENTS

I would like to thank L. Kauffman for encouragement to write this up and for pointing out Ref. 6, as well as for explaining Ref. 7 to me. This was useful for the discussion in Sec. III.

This work was supported by an SERC Research Assistantship.

[1] E. Witten, "Some Geometric Applications of Quantum Field Theory," Proc. ICMP, Swansea (1988).

[2] S. Majid, "Non-commutative $\mathscr{A}/\mathscr{G}$: a New Approach to Quantization of Photons," unpublished (1986).

[3] A. A. Kirillov, *Elements of the Theory of Representations* (Springer, Heidelberg, 1976).

[4] R. Bott and L. Tu, *Differential Forms in Algebraic Topology* (Springer, New York, 1982).

[5] M. E. Sweedler, *Hopf Algebras* (Benjamin, New York, 1969).

[6] J. H. White, Am. J. Math. **91**, 693 (1969); F. B. Fuller, Proc. Natl. Acad. Sci. USA **68**, 815 (1971).

[7] L. Kauffman, "Polynomial Invariants in Knot Theory," preprint (1988).

[8] D. Rolfsen, *Knots and Links* (Publish or Perish, 1976).

[9] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

[10] S. Majid, "Non-commutative Geometric $\mathscr{A}/\mathscr{G}$ and String Field Theory," preprint (1988).

[11] A. M. Polyakov, Mod. Phys. Lett. A **3**, 325 (1988).

Spherically symmetric static SU(2) Einstein–Yang–Mills fields

H. P. Künzle and A. K. M. Masood-ul-Alam
Department of Mathematics, University of Alberta, Edmonton, Alberta T6G 2G1, Canada

(Received 29 August 1989; accepted for publication 22 November 1989)

The discrete family of global solutions of the static spherically symmetric SU(2) Einstein–Yang–Mills equations that were recently numerically obtained by Bartnik and McKinnon [Phys. Rev. Lett. **61**, 141 (1988)] is studied in greater detail, both numerically and analytically. A similar discrete sequence of numerical solutions outside a regular event horizon is shown to exist for every radius of the horizon.

I. INTRODUCTION

Bartnik and McKinnon[1] have recently found numerically a discrete family of globally regular static spherically symmetric solutions of the SU(2) Einstein–Yang–Mills (EYM) equations on $\mathbb{R}^3 \times \mathbb{R}$ for which the potential falls off asymptotically like $1/r$. These are most interesting solutions both physically and mathematically since, in view of Lichnerowicz's theorem,[2] the existence of such nontrivial global solutions of a fundamental field theory comes somewhat as a surprise. (See Ref. 3 for a detailed discussion).

This phenomenon seems to result from the combination of the Yang–Mills with the gravitational equations because there are no pure static Yang–Mills solutions in the three-dimensional space.[4] There also appear to be no simple topological arguments for the existence of a discrete family of solutions as in the case of the monopoles of the Yang–Mills–Higgs equations. In fact, the solutions of this family are characterized by *one* "quantized" mass parameter. Moreover, these mass "eigenvalues" appear to converge to 1 (i.e., the limiting mass in relativistic units is equal to the radius of the "core," in which the mass-energy density is concentrated). They arise simply from the boundary conditions at the center and at infinity together with the considerable nonlinearity of the equations.

The solutions of Bartnik and McKinnon suggest that also black hole solutions of this system might exist. Although a "no hair" theorem was recently extended by Galt'sov and Ershov[5] to the SU(2) Yang–Mills fields in the "essentially Abelian" case, they left open the possibility that solutions which behave like those of Bartnik and McKinnon in the asymptotic region might form regular event horizons. Our numerical calculations strongly suggest that indeed a family of such solutions exists, characterized by a discrete mass parameter for every choice of the radius of the event horizon. This would answer in the negative Yasskin's[6] conjecture that all the black hole solutions of the EYM equations that are asymptotically flat with a nonsingular event horizon and with gauge fields that fall off like $1/r$ at infinity are of the form of his family that generalizes the Reissner–Nordström solutions in a somewhat trivial way to higher-dimensional gauge groups.

Of course, the existence of these classes of solution has not yet been proved rigorously. We do not know at the present time how difficult such a proof is going to be. Also, one might want to investigate the stability within the set of all static solutions or even dynamic stability, a problem that may be even harder. For the time being, we make a more modest attempt to explore at least some properties of this interesting system of equations, mostly under the assumption that solutions exist.

In Sec. II we establish our notation and review the derivation of the basic differential equations in the static spherically symmetric case. We then derive some elementary properties in Sec. III, some of which have already been noted by Bartnik and McKinnon.[1] After a short report in Sec. IV on our numerical results, which qualitatively agree with those in Ref. 1, we derive in Sec. V a number of global estimates on the parameters that describe the solutions. Finally, in Sec. VI we show that to prove existence of a regular event horizon for an asymptotically flat solution amounts again to solving a singular boundary value problem. The numerical evidence supports our belief that there also exists a discrete family of solutions.

II. STATIC SPHERICALLY SYMMETRIC FIELD EQUATIONS

In a general static space-time whose metric we write in the form

$$g = -V^2\, dt \otimes dt + \gamma_{ij}\, dx^i \otimes dx^j,$$

with $\partial_t V = \partial_t \gamma_{ij} = 0$, where γ is a Riemannian metric on a three-manifold Σ, we consider a stationary Yang–Mills connection

$$A = (A^{\mathrm{k}}_0\, dt + A^{\mathrm{k}}_i\, dx^i) E_{\mathrm{k}}$$

with $\partial_t A^{\mathrm{k}}_\alpha = 0$ and write

$$F = {}^4DA = (E^{\mathrm{k}}_i\, dt \wedge dx^i + \tfrac{1}{2} B^{\mathrm{k}}_{ij}\, dx^i \wedge dx^j) E_{\mathrm{k}}.$$

[Here E_{k} denotes a basis vector of the Lie algebra of SU(2) and we may choose $E_{\mathrm{k}} = (-i/2)\sigma_{\mathrm{k}}$ in terms of the Pauli matrices.]

We consider here only magnetic type fields, i.e., we assume that A^{k}_0 and therefore E^{k}_0 vanish. The Yang–Mills field equations then are

$$D^r B^{\mathrm{k}}_{ri} + V^{-1}\, \partial^r V B^{\mathrm{k}}_{ri} = 0, \qquad (1)$$

where D now denotes the three-gauge-covariant derivative, e.g., on Lie Algebra-valued one-forms,

$$D_i \Phi^{\mathrm{k}}_j := \nabla_i \Phi^{\mathrm{k}}_j - A^{\mathrm{l}}_i \epsilon^{\mathrm{k}}{}_{\mathrm{lm}} \Phi^{\mathrm{m}}_j.$$

The mass-energy density is given by

0022-2488/90/040928-08$03.00

$$4\pi\mu = \tfrac{1}{2}\|B\|^2 = \tfrac{1}{4}B^{\mathbf{k}}_{\ ij}B_{\mathbf{k}ij}$$

and the stress tensor

$$4\pi T_{ij} = B^{\mathbf{k}}_{\ ir}B^{r}_{\ \mathbf{k}j} - \tfrac{1}{2}\gamma_{ij}\|B\|^2.$$

For general YM fields over arbitrary manifolds the notion of a symmetry group is not completely straightforward since first an appropriate action of the group on the principal bundle must be defined (see Ref. 7). Such a choice is not difficult, however, in the case of spherical symmetry of an SU(2)-gauge field with trivial principal bundle. There are natural bundle automorphisms by left actions of SU(2) on itself and via its projection onto SO(3) on the base manifold Σ. An invariant connection is then defined by a connection form A satisfying

$$\mathscr{L}_{X_{\mathbf{k}}}A^{\mathbf{l}} = -\epsilon^{\mathbf{l}}_{\ \mathbf{k}\,\mathbf{m}}A^{\mathbf{m}}$$

for the three generators $X_{\mathbf{k}}$ of SO(3) on Σ. On a flat (Σ,γ) the calculation leads to Witten's[8] ansatz, which is easily adapted to curved spherically symmetric (Σ,γ).

If n is the outward pointing unit vector field orthogonal to the orbits of SO(3), $\tau = \frac{1}{6}\epsilon_{ijk}\,dx^i\wedge dx^i\wedge dx^k$ the volume element on Σ, and$\{e_{\mathbf{k}}\}$ an orthonormal frame field on Σ serving as a basis of the Lie algebra of SO(3) with $\{\theta^{\mathbf{k}}\}$ the dual frame field, then we can write $A^{\mathbf{k}} = \theta^{\mathbf{k}}_{\ i}A^{i}_{\ j}\,dx^j$, where the $A^{i}_{\ j}$ are now the components of a tensor field on Σ and have the form

$$A^{i}_{\ j} = X\delta^i_j + (Y-X)n^in_j + Z\epsilon^{i}_{\ jk}n^k \tag{2}$$

for scalar functions X, Y, Z of the curvature radius r only.

This gauge potential (contrary to another one sometimes used, e.g., in Ref. 1) will be globally regular on Σ in most cases. If one requires that it be C^2 at the center it follows easily that

$$X = X_0 + X_1r + O(r^2),$$
$$Y = X_0 + X_1r + O(r^2),$$
$$Z = Z_1r + O(r^2). \tag{3}$$

Gauge transformations $A \mapsto \hat{A} = \mathrm{ad}_{g^{-1}}A + g^{-1}\,dg$ with

$$g\colon \Sigma \to \mathrm{SU}(2)\colon\quad x \mapsto \tilde{g}(r)\theta^{\mathbf{k}}_{\ i}n^i\sigma_{\mathbf{k}} \tag{4}$$

preserve the structure (2) of A. In fact, if we let

$$w := 1 - rZ + irX =: |w|e^{i\gamma} \tag{5}$$

and let $\tilde{g}(r) = \sin(\frac{1}{2}\Lambda(r))$ the effect of (4) is

$$\hat{w} = we^{i\Lambda},\quad \hat{Y} = Y + n^i\,\partial_i\Lambda = Y + \frac{d\Lambda}{dr}.$$

We can thus choose Λ so as to make $\hat{Y}$ zero. The YM equations (1) now also decompose into three terms proportional to δ^i_j, n^in_j, and $\epsilon^{i}_{\ jk}n^k$ and the nn term becomes $d\gamma/dr = 0$ so that the remaining gauge transformation with constant Λ can be used to make w real.

It now turns out that one of the Einstein equations becomes a consequence of the others and the YM equations so that finally the whole system of EYM equations reduces (in suitable units) to[1]

$$S^{-1}S' = 2r^{-1}w'^2, \tag{6}$$
$$m' = Nw'^2 + \tfrac{1}{2}r^{-2}W^2, \tag{7}$$
$$r^2Nw'' + (2m - W^2r^{-1})w' + Ww = 0, \tag{8}$$

where we have put $' := d/dr$, $W := 1 - w^2$, and $S := VN^{-1/2}$ and introduced the Schwarzschild type mass function $m(r)$ so that $N = 1 - 2m/r$, and where we write the three-metric now in the form

$$\gamma = N^{-1}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2).$$

The YM curvature is then given by

$$F^{\mathbf{k}}_{\ ij} = \theta^{\mathbf{k}}_{\ l}[B_T\epsilon^{l}_{\ ij} + (B_L - B_T)n^l\epsilon_{ijr}n^r], \tag{9}$$

where $B_L = -r^{-2}W$ and $B_T = r^{-1}N^{1/2}w'$ are the radial and angular components. In terms of these the mass energy density becomes

$$4\pi\mu = \tfrac{1}{2}B^2_L + B^2_T = m'r^2$$

and the radial and angular pressures become

$$4\pi p_r = B^2_T - \tfrac{1}{2}B^2_L \quad\text{and}\quad 4\pi p_\theta = \tfrac{1}{2}B^2_L,$$

respectively. Then $T^{\alpha}_{\ \alpha} = -\mu + p_r + 2p_\theta = 0$, as it should for a pure gauge field.

From Eqs. (3) and (5) it follows that

$$w = 1 - \beta r^2 + O(r^3),\quad \text{for } r \to 0. \tag{10}$$

Similarly, it is well known that for a spherically symmetric space-time with a regular center

$$N(0) = 1 \quad\text{and}\quad S(0) = S_0 > 0, \tag{11}$$

whence also $m(0) = 0$.

For $r \to \infty$ we require asymptotic flatness so that

$$N \to 1 \quad\text{and}\quad S \to 1. \tag{12}$$

III. ELEMENTARY PROPERTIES AND LOCAL ANALYSIS

Clearly the obvious singularities of the system (6)–(8) are where $r = 0$ or $r \to \infty$ or where $N = 0$. The latter case can physically only occur when a black hole event horizon is being formed and will be examined more closely in Sec. VI.

At $r = 0$ we require w and N to be of the form (10) and (11), respectively, for physical reasons. It is then easy to show that a formal power series in r for w and N can be constructed consistently to arbitrary orders and that both m and w depend only on the choice of β. In fact, it is straightforward to adapt the textbook existence and uniqueness proof to obtain the following lemma.

Lemma 1: There exists a unique solution (m,w) of (7) and (8) with the initial values

$$w(0) = 1,\quad w'(0) = 0,\quad w''(0) = -2\beta,\quad m(0) = 0$$

for small enough r. This solution is analytic in r and β.

Let us now assume that N and w are smooth functions and that $N > 0$ on some interval $I \subset]0,\infty[$. Then Eq. (8) shows that at a critical point of w the second derivative has the same sign as Ww. It follows that the critical points of w are (nondegenerate) maxima (minima) iff $0 < w < 1$ $(0 > w > -1)$. The function w will therefore oscillate in the strip $[-1,1]$ and if it ever crosses ± 1 then $|w|$ will grow monotonically. If, on the other hand, $w'(r_*) = 0$, $w(r_*) = \pm 1$, $0 < r_* < \infty$, and $N > 0$, then the standard uniqueness theorem for ordinary differential equations shows that $w = \pm 1$ and $m = \text{const}$ is the only solution. In summary we have[1] the following lemma.

Lemma 2: Where $N>0$ the function w cannot have a minimum except where $0>w>-1$ and no maximum except where $0<w<1$.

Adopting the physical boundary conditions and requiring regularity for w we get a slightly stronger result.

Lemma 3: If w is C^2, $w(r_0)>1$, $w'(r_0)\geqslant 0$ [or $\leqslant 0$ if $w(r_0)<-1$] and $0<N<1, S>0$ on $I=[r_0,\infty[$ and if the asymptotic conditions (12) hold, then w cannot have a finite limit as $r\to\infty$ (and, in fact, must grow at least linearly with r).

Remark: The condition $w'(r_0)\geqslant 0$ will always be satisfied if $w<1$ before it leaves the strip $[-1,1]$.

Proof: By Lemma 2 we have $w'\geqslant 0$ on I. But one can write (8) [using (6)] in the form

$$(NSw')'=-r^{-2}SWw, \tag{13}$$

so that

$$NSw'=N(r_0)S(r_0)w'(r_0)-\int_{r_0}^{r}ds\,s^{-2}SWw.$$

Since $w>1$ and hence $W<0$ in the integrand, this shows that for large r, at least, $w'(r)$ is thus bounded below by the value of NSw at r_0.

If w grows linearly in r, some components of the YM curvature (9) need not go to zero at ∞, which is not the situation we wish to study. It is more reasonable to assume that w has a finite limit w_∞ for $r\to\infty$. Lemma 3 then shows that $|w_\infty|\leqslant 1$. In fact, we *assume* that from now on

$$w=w_\infty+O\left(\frac{1}{r}\right),\quad w'=O\left(\frac{1}{r^2}\right),\quad w''=O\left(\frac{1}{r^3}\right), \tag{14}$$

then Eq. (8) shows that $w_\infty(1-w_\infty^2)=0$ so that $w_\infty=\pm 1$ or 0. But the following argument of Galt'sov and Ershov[5] shows that $w_\infty=0$ cannot occur: They integrate (13) from the largest critical point r_* of w to ∞ getting

$$\int_{r_*}^{\infty}dr\,r^{-2}SWw=0.$$

Since $|w|<1$, hence $W>0$, w must change sign between r_* and ∞ unless it is identically zero. [The case $w_\infty=\pm 1$ is sometimes[1,5] referred to as the one of vanishing magnetic charge, although this statement only makes sense if the two-form F is integrated over a two-sphere at infinity in an appropriate gauge (cf. Ref. 9).]

In order to start a numerical integration at ∞ we need the following lemma.

Lemma 4: The system (7) and (8) admits a formal asymptotic series in powers of $1/r$,

$$m=m_\infty-\sum_{k=1}^{\infty}m_kr^{-k}, \tag{15}$$

$$w=w_\infty\left[1-\frac{\alpha}{r}+\sum_{k=2}^{\infty}w_kr^{-k}\right], \tag{16}$$

where $w_\infty=\pm 1$ and all coefficients m_k, w_k are determined in terms of m_∞ and α.

The proof is straightforward. Observe that Eqs. (7) and (8) are invariant under the transformation $w\to -w$, so that to every solution (m,w) also $(m,-w)$ is a solution. Therefore m_k and w_k do not depend on w_∞.

We will also assume from now on that the function $m(r)$ has a limit m_∞ as $r\to\infty$.

Next we consider the local behavior of solutions near a finite nonzero value r_0 of r. Not surprisingly, if w and m are assumed to be analytic in r near r_0 the power series are fully determined by the values of w, w', and m at r_0 as long as $N(r_0)\neq 0$, i.e., $2m(r_0)\neq r_0$, since a unique solution in a neighborhood of r_0 with these initial values is guaranteed by standard theorems.

If, however, $N(r_0)=0$ there is only a one-parameter family of regular analytic solutions. More precisely we have the following lemma.

Lemma 5: The system (7) and (8) admits a formal power series solution at $r_0\in]0,\infty[$, where $N(r_0)=0$, provided that $N'(r_0)\neq 0$ and $w(r_0)\neq 0$. All coefficients of the series for m and w are then determined in terms of the value $w_0=w(r_0)$.

We will need this result to construct numerical black hole solutions.

When integrating numerically upwards from $r=0$ with the initial conditions (10) and (11) we see that w develops a singularity as $N\downarrow 0$. It appears that w and m remain finite, but w' blows up. We try therefore solutions of the form

$$w=w_0+t^{\gamma}\tilde{w},\quad m=\tfrac{1}{2}r_0+\tilde{m},$$

where $t=r_0-r$, $0<\gamma<1$,and $\tilde{w}$ and $\tilde{m}$ are regular functions. It turns out that for $\gamma=\frac{1}{2}$ we can construct a formal series consistently. We have the following lemma.

Lemma 6: If

$$m=\tfrac{1}{2}r_0-\sum_{k=2}^{\infty}m_ks^k$$

and

$$w=\sum_{k=0}^{\infty}w_ks^k,$$

where $s=\sqrt{r_0-r}$, all coefficients m_k and w_k are well defined in terms of r_0, w_0, and w_2 provided $r_0^2\neq(1-w_0^2)^2$, i.e., $N'(r_0)\neq 0$.

The proof is again straightforward but lengthy. In particular, we find that

$$w_1^2=r_0\neq 0, \tag{17}$$

showing that w' tends to ∞ at r_0 like $(r_0-r)^{-1/2}$.

Incidentally, if r tends to r_0 from above and $s=\sqrt{r-r_0}$ then Eq. (17) is replaced by $w_1^2=-r_0$, which has no real solution. The function w behaves thus quite differently as it approaches r_0 from below or from above.

We have not proved that the solution of the type given in Lemma 6 is the only possible singular one through a point r_0 with $N(r_0)=0$. The numerical solution satisfying our boundary conditions at $r=0$, however, exhibits exactly this behavior.

Finally, it is interesting to speculate what happens if N has a higher-order zero at a finite value r_0. This is the limiting case when the minimum of N is exactly 0. So far, we have not been able to find any consistent analytic approximation to m and w in the neighborhood of such a point. The numerical experiments also show that this must be quite a horrible singularity.

IV. NUMERICAL RESULTS

We have repeated the numerical calculations reported by Bartnik and McKinnon[1] and find the same qualitative results shown by their graphs. Since our numbers differ somewhat we describe our method in a little more detail.

For this singular two-point boundary value problem we use the method "shooting to a fitting point" as described in Press *et al.*[10] using double precision and an accuracy of $\epsilon \leqslant 10^{-12}$ in the adaptive step size Runge–Kutta method. The integration starts at $r = 0.01$ or less and at $r \geqslant 10^6$, at which points we calculate the initial values of m and w using the asymptotic series to arbitrarily high terms until they converge within ϵ. Starting with some values m_∞, α, and β we iterate until the total difference of m, w, and w' at the fitting point is less than about 10^{-9}. The actual errors in the "eigenvalues" m_k, are, of course, much greater than 10^{-9} since quite a few numerical cancellation errors occur in the summation of the power series as well as the numerical integration. The numbers obtained for the location of the zeros of w are very approximate. They were obtained by polling the function calculating derivatives for changes in the sign of w.

Our results are as shown in Table I.

From the graphs corresponding to Table I (cf. Ref. 1) it is also seen that the amplitude of the oscillations of w is very small where the zeros accumulate near $r = 1$ and gets bigger between the larger roots.

On the basis of these numbers we are tempted to formulate the following conjectures.

There exists an infinite sequence of asymptotically flat globally regular solutions (m_k, w_k) of the EYM equations on $\mathbb{R}^3$, parametrized by the number of zeros of the potential component w.

As $k \to \infty$ the total mass m_∞ of the solutions tends to 1 from below, the parameter β tends to a finite value of about 0.7065, α tends to ∞.

The function $N(r)$ has for all solutions only one minimum near $r = 1$. This minimum value approaches zero as $k \to \infty$.

The zeros of solution w_k accumulate near $r = 1$.

For large k the function w_k approaches the value r_0, where N has its minimum, qualitatively like $\sqrt{r - r_0}$ from the left and somewhat like $(r - r_0)\sin(r - r_0)^{-1}$ from the right.

B_T^2 and B_L^2, and hence the mass-energy density and stresses, fall off rapidly for $r > 1$.

Clearly these numerical solutions pose interesting mathematical problems, although it is not yet known whether they are stable and of physical importance. In the next section we begin such a study by deriving at least some bounds on the parameters m_∞ and β analytically.

TABLE I. Parameters of the solutions $(m_k(r), w_k(r))$ of Eqs. (7) and (8) satisfying the boundary conditions.

k	m_∞	α	β	Zeros of w
1	0.828 6	0.8934	0.453 7	1.55
2	0.971 3	8.864	0.651 7	1.10, 3.70
3	0.995 3	58.93	0.697 0	0.990, 1.67, 14.2
4	0.999 2	366.3	0.704 9	0.969, 1.15, 3.69, 77.2
5	0.999 9	2251	0.706 2	0.966, 1.03, 1.69, 14.1, 460
6	0.999 98	13 820	0.706 4	0.965, 1.004, 1.15, 3.69, 74.9, 2820
7	0.999 996	80 300	0.706 41	0.965, 0.999, 1.03, 1.71, 14.9, 492, 18720

V. SOME GLOBAL ESTIMATES

In this section we show analytically that a regular solution (m,w) satisfying the initial conditions (10) and (11) ceases to exist before w reaches 0.76 if $\beta \geqslant 15.4$ (Theorem 1) and also establish an upper bound for the total mass m_∞ (Theorem 2) for a regular solution of the boundary value problem. In view of the numerical results these values are probably not the best possible.

Theorem 1: If (m,w) is a solution of the system (7) and (8) with the initial conditions (10), $N(0) = 1$ and if $\beta > 15.4$, then $|w'|$ tends to ∞ before w reaches 0.76, and hence no regular solution of the boundary value problem exists.

We prove Theorem 1 using the next three lemmas.

Lemma 7: Suppose on $[0, r_0]$ a solution of the system (7) and (8) with $\beta > 0$ exists and that $1 \geqslant w \geqslant 0$ and $1 \geqslant N > 0$. Then

$$2(1 - w) \geqslant rN|w'| \quad \text{on} \quad [0,r_0]. \tag{18}$$

Proof: We have by straightforward calculation

$$(r^2Nw')' + 2(w'^2 - 1)rNw' + Ww = 0. \tag{19}$$

For a given $r \in [0,r_0]$, we integrate (19) and use

$$\int_0^r w'^3 sN\, ds \leqslant 0$$

and

$$\int_0^r sNw'\, ds \geqslant \int_0^r sw'\, ds = rw - \int_0^r w\, ds$$

to obtain

$$-2rw + \int_0^r w(3 - w^2)ds + r^2Nw' \geqslant 0. \tag{20}$$

But on $]0,r_0]$, w is decreasing (Lemma 2) so that we have $w(3 - w^2) \leqslant 2$. Thus (20) yields (18).

Remark: In particular, (18) implies that if $|w'|$ blows up in $[0,r_0]$, N tends to zero. The converse is also true as can be seen from

$$(N|w|')' + (2w'^2r^{-1} - \tfrac{1}{2}w(1 + w)r)N|w'| \geqslant 0, \tag{21}$$

which is obtained from (7), (8), and (18). Because (21) implies for $s_0 \leqslant r$, that

$$N(r)|w'(r)| \geqslant N(s_0)|w'(s_0)| \times \exp\left(\int_{s_0}^r \left[\frac{1}{2} w(1 + w)s - 2w'^2 s^{-1}\right] ds\right)$$

so that if $N(r)$ tends to zero then $|w'(r)|$ blows up.

Lemma 8: Suppose on $[0,R]$ the solution exists with $w \geqslant w(R) = 0.76$. Then on $[0,R]$ we have

$$r|w'| \geqslant Ww + r^{-2}(1 - w)^3. \tag{22}$$

Proof 1: Since $\beta > 0$ it follows from the series expansion at $r = 0$ that there exists an r^*, with $0 < r^* \leqslant R$, such that, on $[0,r^*]$, (22) holds. Suppose, to the contrary, that, on $]r^*,\hat{r}] \subset [0,R]$,

$$r|w'| < Ww + r^{-2}(1-w)^3. \tag{23}$$

Then using (23) and (7) and (8) we have, on $]r^*,\hat{r}]$,

$$r^2N|w'|' - rN|w'| > W^2r^{-1}|w'| - r^{-2}(1-w)^3. \tag{24}$$

We divide (24) by rN and estimate the right-hand side of the resulting inequality using (18). We thus obtain

$$r|w'|' - |w'| > \tfrac{1}{2}W(1+w)r^{-1}|w'|^2 - \tfrac{1}{2}(1-w)^2r^{-2}|w'|. \tag{25}$$

Integrating (25) on $[r^*,r]$, we get

$$\begin{aligned} r|w'| > {} & r^*|w'(r^*)| + 2(w(r^*) - w) \\ & + \frac{1}{2}\int_{r^*}^{r} W(1+w)|w'|^2s^{-1}\,ds \\ & - \frac{1}{2}\int_{r^*}^{r}(1-w)^2|w'|s^{-2}\,ds. \end{aligned} \tag{26}$$

Now using Hölder's and Cauchy's inequalities we have, for a positive constant C_1,

$$\begin{aligned} \int_{r^*}^{r}(1-w)^2|w'|s^{-2}\,ds \leqslant {} & \frac{1}{2}C_1\int_{r^*}^{r}(1-w)^3s^{-3}\,ds \\ & + \frac{1}{2C_1}\int_{r^*}^{r}(1-w)|w'|^2s^{-1}\,ds. \end{aligned} \tag{27}$$

Since w is decreasing on $[r^*,r]$, the first term on the right-hand side of (27) is less than $\frac{1}{4}C_1(1-w)^3(r^{*-2} - r^{-2})$; whence, from (27) and (26) and the fact that

$$r^*|w'(r^*)| = W(r^*)w(r^*) + r^{*-2}(1 - w(r^*))^3,$$

we get that, for any $r \in [r^*,\hat{r}]$,

$$\begin{aligned} r|w'| > {} & W(r^*)w(r^*) + r^{*-2}(1-w(r^*))^3 + 2(w(r^*) - w) \\ & - \frac{1}{4}C_1C_2(1-w)^3(r^{*-2} - r^{-2}) \\ & + \frac{1}{2}\left((1+w)^2 - \frac{C_2}{C_1}\right)\int_{r^*}^{r}(1-w)|w'|^2s^{-1}\,ds \\ & + \left(C_2 - \frac{1}{2}\right)\int_{r^*}^{r}(1-w)^2|w'|s^{-2}\,ds, \end{aligned} \tag{28}$$

where C_2 is a positive constant. For $r \in [r^*,R]$, now

$$W(r^*)w(r^*) + 2(w(r^*) - w) \geqslant W(r)w(r).$$

Hence, choosing $C_1 = \frac{8}{7}(1 - w(r^*))^3(1-w)^{-3}$ in (28) and $C_2 = \frac{7}{2}$, we get

$$\begin{aligned} r|w'| > {} & Ww + r^{-2}(1-w)^3 \\ & + \tfrac{1}{2}\left[(1+w)^2 - \tfrac{49}{16}(1-w(r^*))^{-3}(1-w)^3\right] \\ & \times \int_{r^*}^{r}(1-w)|w'|^2s^{-1}\,ds. \end{aligned}$$

But, since $(1+w)^2 \geqslant 3.0976$ for $r \in [0,R]$, this equation contradicts (23) in a neighborhood of r^*.

Lemma 9: Let R be as in Lemma 8. Then

$$R^{-1} < 7.536. \tag{29}$$

Proof: Given $\epsilon > 0$, let $\tilde{r} = \tilde{r}(r,\epsilon)$ be the largest value of $s \in [0,r]$ such that

$$W(\tilde{r}) \leqslant \tilde{r}\sqrt{\epsilon + 2m(\tilde{r})\tilde{r}^{-1}}. \tag{30}$$

Then on $[\tilde{r},r]$ (if $\tilde{r} \neq r$), $W(s) \geqslant s\sqrt{\epsilon + 2m(s)/s}$, so that on $[\tilde{r},r]$, $sN' \leqslant -\epsilon$ giving, by straighforward calculations,

$$\epsilon\int_{\tilde{r}}^{r}N^{-1/2}\,ds \leqslant -\int_{\tilde{r}}^{r}2s(N^{1/2})'ds \leqslant 4m(r). \tag{31}$$

On the other hand, for $r \leqslant R$, we have

$$\begin{aligned} w^2(\tilde{r}) - w^2(r) & \leqslant 2w(\tilde{r})\int_{\tilde{r}}^{r}|w'|ds \\ & \leqslant 2w(\tilde{r})\left(\int_{\tilde{r}}^{r}|w'|^3N\,ds\right)^{1/3}\left(\int_{\tilde{r}}^{r}N^{-1/2}\,ds\right)^{2/3}, \end{aligned} \tag{32}$$

where in the last step we have used Hölder's inequality. Now, for $r \leqslant R$, we have

$$\int_{\tilde{r}}^{r}|w'|^3N\,ds \leqslant \sqrt{m(r)r}, \tag{33}$$

which follows because (19) yields

$$2\int_0^r w'^3N\,ds \geqslant \int_0^r\left(w'N - W\frac{w}{s}\right)ds.$$

Thus, by virtue of Hölder's inequality and $w^2 \leqslant 1$, we get

$$2\int_0^r|w'|^3N\,ds \leqslant \left[r\int_0^r w'^2N\,ds\right]^{1/2} + \left[r\int_0^r W^2s^{-2}\,ds\right]^{1/2}.$$

Now, using the formula $\sqrt{x} + \sqrt{y} \leqslant \sqrt{2}\sqrt{x+y}$ and (7), we get (33).

Using (33) and (31) in (32), we have

$$w^2(\tilde{r}) - w^2(r) \leqslant 2(4/\epsilon)^{2/3}(m(r))^{5/6}r^{1/6}w(\tilde{r}). \tag{34}$$

For $\tilde{r} \neq r$, we have, from (30),

$$w(\tilde{r}) = \left[1 - \sqrt{\epsilon\tilde{r} + 2m(\tilde{r})}\sqrt{\tilde{r}}\right]^{1/2}. \tag{35}$$

We note that, for $\tilde{r} = r$, the left-hand side of (34) vanishes. Hence on $[0,r]$, where $r \leqslant R$, (34) and (35) coupled with the triangle inequality imply

$$\begin{aligned} W(r) \leqslant {} & W(\tilde{r}) + 2(4/\epsilon)^{2/3}(m(r))^{5/6}r^{1/6} \\ & \times\left[1 - \sqrt{\epsilon\tilde{r} + 2m(\tilde{r})}\sqrt{\tilde{r}}\right]^{1/2} \end{aligned} \tag{36}$$

Finally using (30) in (36) and the facts that $W(R) = 0.4224$ and $2m(R) \leqslant R$, we get

$$\begin{aligned} 0.4224 \leqslant {} & \sqrt{\epsilon\tilde{r} + 2m(\tilde{r})}\sqrt{\tilde{r}} \\ & + 2\sqrt{2}R\epsilon^{-2/3}\left[1 - \sqrt{\epsilon\tilde{r} + 2m(\tilde{r})}\sqrt{\tilde{r}}\right]^{1/2}. \end{aligned} \tag{37}$$

Now we take $\epsilon = 3.1$ in (37). Then the two cases, namely, $\sqrt{3.1\tilde{r} + 2m(\tilde{r})}\sqrt{\tilde{r}} \leqslant$ or $\geqslant 0.2424$ give, considered separately, in either case $1 < 7.536R$. Hence the lemma follows.

Proof of Theorem 1: The idea is to integrate (22) to bound β in terms of R^{-1} and then to use Lemma 9, in case a regular solution exists so far. In view of the complexities of the inequality (22) we integrate the following two inequalities implied by (22): On $[0,\bar{r}]$ such that $w^2(\bar{r}) = 0.81$, we use

$$r|w'| \geqslant Ww. \tag{38}$$

On $[\bar{r},R]$, we use

$$r|w'| \geqslant (1-w)^3r^{-2} + 10.445. \tag{39}$$

Integrating (38), we get

$$\bar{r}^{-2} \geqslant 8.526\beta. \tag{40}$$

Integrating (39) on $[\overline{r},R]$ and using (40) and (29), we then get $\beta \leqslant 15.4$. The theorem now follows by virtue of the remark after Lemma 7 and the fact that $w' \leqslant 0$ while $w \geqslant 0$, which holds in view of Lemma 2.

Theorem 2: There is no regular solution of the boundary value problem for total mass $m_\infty \geqslant 2.524$.

We first prove the following lemmas.

Lemma 10: Let c_i and c_j $(c_j > c_i)$ be two consecutive critical points of a globally regular solution w of the boundary value problem. Then, for any $\tau > 1$ and $r \in [c_i, c_j,]$ we have

$$r^\tau N(r)|w'(r)| \leqslant \frac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2}(r^\tau - c_i^\tau) - \sigma\int_{c_i}^{r} Wws^{\tau-2}\,ds, \tag{41}$$

where σ is the sign of w' in $]c_i, c_j[$

Proof: We have, for any τ,

$$(r^\tau Nw')' - \tau r^{\tau-1}Nw' + 2w'^3Nr^{\tau-1} + Wwr^{\tau-2} = 0. \tag{42}$$

Since Hölder's and Young's inequalities as well as $N \leqslant 1$ yield

$$\tau\int_{c_i}^{r}|w'|Ns^{\tau-1}\,ds$$

$$\leqslant 2\int_{c_i}^{r}|w'|^3Ns^{\tau-1}\,ds + \frac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2}(r^\tau - c_i^\tau),$$

we get (41) from (42).

We shall use the following lemma to estimate the integral of w'^2N over some suitable interval such that w does not change sign on this interval.

Lemma 11: For any r, a globally regular solution $w(r)$ of the boundary value problem satisfies

$$N|w'| \leqslant 0.93. \tag{43}$$

Proof: From (41), we have

$$N|w'| \leqslant \frac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2} + r^{-\tau}\int_0^r Ws^{\tau-2}\,ds.$$

Using Hölder's inequality,

$$\int_0^r W^2s^{-2}\,ds \leqslant 2m,$$

and $2m < r$, we then have

$$N|w'| \leqslant (\sqrt{2}/3\sqrt{3})\tau^{1/2} + (2\tau - 1)^{-1/2}.$$

Choosing $\tau = 4$ we arrive at (43).

Remark: In an interval $[s_1, s_2]$ such that w does not change sign, (43) gives

$$\int_{s_1}^{s_2} w'^2N\,ds < 0.93|w(s_1) - w(s_2)| \leqslant 0.93, \tag{44}$$

since $|w| < 1$. However, if w changes sign then the following lemma leads to a better estimate than (44).

Lemma 12: Let $\xi_2 \in [c_i, c_j]$ be such that $w(\xi_2) = 0$. Then, for $\tau > 1$,

$$\int_{\xi_2}^{c_j} Nw'^2\,dr \leqslant \frac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2} + \frac{2}{3\sqrt{3}}(\tau - 1)^{-1}\xi_2^{-1} \tag{45}$$

and for any $\overline{s} \in [c_i, \xi_2]$,

$$\int_{\overline{s}}^{\xi_2} Nw'^2\,dr \leqslant \frac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2}|w(\overline{s})| + (\tau - 1)^{-1}\overline{s}^{-1}\left[\tfrac{1}{12} + \tfrac{1}{2}w^2(\overline{s}) - \tfrac{1}{4}w^4(\overline{s})\right]. \tag{46}$$

Proof: Let ξ_1 be such that $c_i \leqslant \xi_1 \leqslant \xi_2$ and $|w(\xi_1)| = 1/\sqrt{3}$ provided such ξ_1 exists. Our aim is to estimate the term involving Ww in (41). We note that

$$\sup_{|w| \in [0,1]} W|w| = 2/3\sqrt{3},$$

and the sup occurs at $|w| = 1/\sqrt{3}$ Thus we have (depending on whether $\mathrm{sgn}(w') = \pm 1$ on $[c_i, c_j]$)

$$\mp\int_{c_i}^{r} Wws^{\tau-2}\,ds$$

$$\leqslant\begin{cases} W|w|(\tau-1)^{-1}r^{\tau-1}, & \text{for} \quad r \in [c_i, \xi_1], \\ (2/3\sqrt{3})(\tau-1)^{-1}r^{\tau-1}, & \text{for} \quad r \in [\xi_1, \xi_2], \\ (2/3\sqrt{3})(\tau-1)^{-1}\xi_2^{\tau-1}, & \text{for} \quad r \in [\xi_2, c_j]. \end{cases} \tag{47}$$

Using the third condition of (47) and (41) we get

$$\int_{\xi_2}^{c_j} Nw'^2\,ds \leqslant \left(\frac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2} + \frac{2}{3\sqrt{3}}(\tau-1)^{-1}\xi_2^{-1}\right)|w(c_j)|,$$

whence (45) follows. On the other hand, for $\overline{s} \in [c_i, \xi_2]$, using the first two of (47) and (41), we get

$$\int_{\overline{s}}^{\xi_2} Nw'^2\,ds \leqslant \begin{cases} \dfrac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2}|w(\overline{s})| + (\tau-1)^{-1}\overline{s}^{-1}\left[\displaystyle\int_{|w(\xi_1)|}^{|w(\overline{s})|} W|w|d\,|w| + \dfrac{2}{3\sqrt{3}}|w(\xi_1)|\right], & \text{if} \quad \xi_1 > \overline{s}, \\ \left(\dfrac{\sqrt{2}}{3\sqrt{3}}\tau^{1/2} + \dfrac{2}{3\sqrt{3}}(\tau-1)^{-1}\overline{s}^{-1}\right)|w(\overline{s})|, & \text{if} \quad \xi_1 \leqslant \overline{s}. \end{cases} \tag{48}$$

Now

$$\int_{1/\sqrt{3}}^{|w(\overline{s})|} W|w|d\,|w| = \frac{1}{2}w^2(\overline{s}) - \frac{1}{4}w^4(\overline{s}) - \frac{5}{36}.$$

Also for $\xi_1 \leqslant \overline{s}$, $|w(\overline{s})| \leqslant 1/\sqrt{3}$ and $(2/3\sqrt{3})|w(\overline{s})| \leqslant \frac{2}{9} + \frac{1}{2}w^2(\overline{s}) - \frac{1}{4}w^4(\overline{s}) - \frac{5}{36}$. Hence (48) yields (46).

Lemma 13: Let $c > 0$ be a critical point of a regular solution w of the boundary value problem. Then,

$$\int_c^\infty Nw'^2\,ds \leqslant \frac{4}{3\sqrt{3}c}. \tag{49}$$

Proof: We take $\tau = 0$ in (42) and then integrate on the interval between two consecutive critical points of w, namely, c_k and c_{k+1}. So, for $r \in [c_k, c_{k+1}]$, we have

$$N(r)|w'(r)| \leqslant \int_{c_k}^{r} W|w|s^{-2}\,ds. \tag{50}$$

Using this and $W|w| \leqslant 2/3\sqrt{3}$ we get

$$\int_{c_k}^{c_{k+1}} Nw'^2\, ds \leqslant \frac{2}{3\sqrt{3}}\left(\frac{1}{c_k} - \frac{1}{c_{k+1}}\right)\int_{c_k}^{c_{k+1}} |w'|ds. \tag{51}$$

Estimating the integral on the right-hand side of (51) by 2 and then summing over all the critical points of c we get (49).

Proof of Theorem 2: For $c \geqslant 2$, we find, using (7) and $m(2) < 1$,

$$m(c) < \min\left[1.25 + \int_2^c Nw'^2\, ds - (2c)^{-1}, \frac{c}{2}\right]. \tag{52}$$

To estimate $\int_2^c Nw'^2\, ds$, we use Lemma 12 and the remark after Lemma 11. We put $\bar{s} = 2$ in (46) and $c = c_j$ in (45). In case there exists $\xi_2 \in [2,c]$ such that $w(\xi_2) = 0$ we put $\tau = 2$ in (46) and $\tau = 2.5$ in (45) to get, respectively,

$$\int_2^{\xi_2} Nw'^2\, ds \leqslant \frac{2}{3\sqrt{3}}|w(2)| + \frac{1}{4}w^2(2) - \frac{1}{8}w^4(2) + \frac{1}{24} \tag{53}$$

and

$$\int_{\xi_2}^c Nw'^2\, ds \leqslant 0.5586. \tag{54}$$

As the right-hand side of (53) is less than 0.5432 for $|w| \leqslant 1$, we have, adding (53) and (54),

$$\int_2^c Nw'^2\, ds < 1.102. \tag{55}$$

In case no ξ_2 exists in $[2,c]$, then (44) applies, so that in either case (55) holds.

But(7) and (49) imply that

$$m_\infty \leqslant m(c) + (4/3\sqrt{3} + \tfrac{1}{2})c^{-1}. \tag{56}$$

Hence from (56), (52), and (55), we get

$$m_\infty < \min(2.352, \tfrac{1}{2}(c + c^{-1})) + (4/3\sqrt{3})c^{-1}. \tag{57}$$

For $c \geqslant 2$, both $c + c^{-1}$ and $c + c^{-1}(1 + 8/3\sqrt{3})$ are increasing functions of c. Hence considering the two cases, namely, $2 \leqslant c \leqslant 4.48$ and $c > 4.48$, separately, we get that in either case (57) implies $m_\infty < 2.524$.

VI. BLACK HOLE SOLUTIONS

Yasskin[6] formulated the conjecture that the only solutions of the EYM equations that are asymptotically flat and stationary with an $1/r$ fall off for the potential and an event horizon with spherical topology were his generalizations of the Kerr–Newman solutions which are essentially Abelian. Very recently Galt'sov and Ershov[5] proved nonexistence of SU(2)-EYM black hole solutions that are non-Abelian and satisfy $\lim_{r\to\infty} w(r) = 0$ in the notation of our Sec. III. They specifically leave open the possibility of black hole solutions when $w_\infty = \pm 1$. We will show here that there are indeed such black holes if a similar kind of singular two-point boundary value problem as in the Bartnik and McKinnon case admits solutions. Numerical calculations indicate that solutions exist.

Kruskal's coordinate construction for a static spherically symmetric space-time consists of replacing the coordinates (t,r) by new coordinates (u,v) such that

$$-V^2\, dt^2 + N^{-1}\, dr^2 = \mathscr{L}(du^2 - dv^2), \tag{58}$$

with $\mathscr{L} \neq 0$ where $V = 0$. We assume that $\mathscr{L}$ depends on u and v only via r. It follows that

$$u = e^{kx}\cosh kt, \quad v = e^{kx}\sinh kt,$$

where k is a constant and x a new radial coordinate determined by

$$\frac{dr}{dx} = VN^{1/2} = NS =: \mathscr{F}. \tag{59}$$

Now, since $S(\infty) = 1$, it follows from (6) that

$$S = \exp\left(\int_r^\infty dr\, r^{-1}w'^2\right)$$

is positive everywhere. The event horizon occurs where the timelike Killing vector field becomes null, i.e., where V or, equivalently, N or $\mathscr{F}$, vanishes. Assume asymptotic flatness again and let r_H denote the largest value of r for which $N(r) = 0$.

If $N'(r_H) > 0$, so that $A := \mathscr{F}'(r_H) > 0$, then we can write

$$\mathscr{F}(r) = A(r - r_H)[1 + A(r - r_H)X']$$

and hence

$$x = A^{-1}\ln(r - r_H) + X(r),$$

for small positive $r - r_H$, where $X(r)$ is a C^1 function of r satisfying $X(r_H) = 0$. It now follows that

$$\mathscr{L} = k^{-2}A(r - r_H)^{(1 - 2k/A)}[1 + A(r - r_H)X']^{-1}Se^{-2X}.$$

In order that $\mathscr{L}(r_H) \neq 0$, one chooses $k = A/2$. We have thus rederived the well-known result: For a regular event horizon to occur at the largest zero r_H of N it is sufficient that $N'(r_H) > 0$ and $S(r_H) < \infty$. [There is no need for $S(r_H) = 1$, as Galt'sov and Ershov[5] seem to require.]

Numerical experiments now show that it is the second condition that is hard to satisfy. As N decreases from 1 when r decreases from ∞, w', and very likely also S, tend to blow up when N reaches zero. However, for a discrete set of values of the parameters m_∞ and α this does not happen. To construct some numerical solutions that represent black hole solutions we can therefore require that w', and hence S, remain finite at r_H. We proceed as follows.

(a) Pick values m_∞ and α and compute m, w, and w' for a very large r from the asymptotic series of Lemma 4.

(b) Pick a value r_H ($< 2m_\infty$) and a value w_H (with $0 < |w_H| < 1$) and calculate m, w, w' at $r_H + \epsilon$ using the power series of Lemma 5 (making sure that

$$N'(r_H) = r_H^{-3}[r_H^2 - (1 - w_H^2)^2] > 0).$$

(c) Shoot to a fitting point by numerical integration from both sides and improve the parameters m_∞, α, and w_H by Newton's method.

For every solution (m,w) thus obtained there will also be the solution $(m, -w)$ for which the angular component B_T of the YM curvature also has the opposite sign.

There remain many questions to be investigated for this black hole case. So far our numerical results indicate that, for every choice of r_H, the following are true.

There exists a discrete (infinite?) sequence of solutions (m_k, w_k) characterized by the number k of zeros of w_k in the interval $]r_H, \infty[$.

The parameters m_∞ for these solutions depend only slightly on k and seem, if $r_H > 1$, to converge to a limit greater than $\frac{1}{2}r_H$. As r_H is chosen larger this limit itself tends towards $\frac{1}{2}r_H$. For $r_H < 1$ the total mass m_∞ also grows with k and may converge to 1 again.

The parameters α grow, apparently without bound, as $k \to \infty$.

The values $|w_H|$ tend to zero as $k \to \infty$, for $r_H > 1$, and to a limiting value greater than zero, if $r_H < 1$ [since $N'(r_H)$ must remain positive].

If $r_H < 1$, then N develops a minimum around $r = 1$ whose value decreases towards zero as k increases.

The mass-energy density is very small for $r > 1$ whether or not $r_H < 1$.

Solutions with large numbers of zeros are very difficult to obtain numerically. In fact, for $w_H = 0$ the power series of Lemma 5 becomes trivial. Such a limiting black hole solution, if it exists, is very difficult to approximate analytically since then N seems to develop a multiple root at the horizon.

ACKNOWLEDGMENTS

We thank Werner Israel for a helpful discussion.

This work was supported in part by NSERC Grant No. A8059.

[1] R. Bartnik and J. McKinnon, "Particlelike solutions of the Einstein–Yang–Mills equations," Phys. Rev. Lett. **61**, 141 (1988).

[2] A. Lichnerowicz, *Théories relativistes de la gravitation et de l'éléctromagnetisme* (Masson, Paris, 1955).

[3] P. Breitenlohner, D. Maison, and G. Gibbons, "Fourdimensional black holes from Kaluza–Klein theories," Commun. Math. Phys. **120**, 295 (1988).

[4] S. Deser, "Absence of static solutions in source-free Yang–Mills theory," Phys. Lett. B **64**, 463 (1976).

[5] D. V. Galt'sov and A. A. Ershov, "Non-abelian baldness of colored black holes," Phys. Lett. A **138**, 160 (1989).

[6] P. B. Yasskin, "Solutions for gravity coupled to massless gauge fields," Phys. Rev. D **12**, 2212 (1975).

[7] J. Harnad, S. Shnider, and J. Tafel, "Group actions on principal bundles and dimensional reduction," Lett. Math. Phys. **4**, 107 (1980).

[8] E. Witten, "Some exact multipseudoparticle solutions in classical Yang–Mills theory," Phys. Rev. Lett. **38**, 121 (1977).

[9] P. T. Chruściel and W. Kondracki, "Some global charges in classical Yang–Mills theory," Phys. Rev. D **36**, 1874 (1987).

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge U.P., Cambridge, 1986).

A class of exact solutions of Yang's K gauge equation for SU(2) gauge fields

Y. Matsuno
Department of Physics, Faculty of Liberal Arts, Yamaguchi University, Yamaguchi 753, Japan

(Received 18 July 1989; accepted for publication 15 November 1989)

Using a simple ansatz, Yang's K gauge equation for SU(2) gauge fields is reduced to a system of nonlinear ordinary differential equations. Exact solutions for the equations are obtained together with corresponding gauge potentials.

I. INTRODUCTION

In search of the SU(2) gauge fields in four-dimensional Euclidean space, Yang derived conditions of self-duality and obtained nonlinear partial differential equations that describe gauge potentials.[1] His equations are divided into the two types according to the choice of the gauge. One is an equation with a Hermitian gauge or the K gauge, and the other is an equation with the R gauge. The latter equation has been studied extensively and various classes of exact solutions have been published.[2–5] The investigation of the former one, on the other hand, has scarcely been done.

The purpose of this paper is to construct exact solutions of Yang's self-dual equation with the K gauge. Since the equation itself is a quite complicated nonlinear partial differential equation for a vector field and hence it seems to be intractable, we shall introduce a simple ansatz that the fields depend only on one variable. Under the assumption, Yang's equation is reduced to a system of nonlinear ordinary differential equations. Exact solutions for the equations are constructed and corresponding gauge potentials are calculated explicitly.

II. EXACT SOLUTIONS

Yang's K gauge equation for a real vector field $\mathbf{v}$ is written in the form[1]

$$\tfrac{1}{2}(1-v^2)\mathbf{v}_{\mu\mu}+2(\mathbf{v}\cdot\mathbf{v}_\mu)\mathbf{v}_\mu-(\mathbf{v}_\mu\cdot\mathbf{v}_\mu)\mathbf{v}$$
$$-2(\mathbf{v}_1\times\mathbf{v}_2-\mathbf{v}_3\times\mathbf{v}_4)=0,\quad (v\equiv|\mathbf{v}|), \tag{2.1}$$

where the subscript μ indicates the differentiation with respect to the Euclidean coordinate x_μ and the repeated greek index μ runs from 1 to 4, i.e., $\mathbf{v}_{\mu\mu}=\Sigma_{\mu=1}^4\partial^2\mathbf{v}/\partial x_\mu^2$, for example.

We shall seek solutions of Eq. (2.1) that depend only on one variable ϕ, where ϕ is a function of x_μ $(\mu=1\sim4)$. Under this situation, the fourth term on the left-hand side of Eq. (2.1) vanishes identically and Eq. (2.1) is reduced to the equation

$$\tfrac{1}{2}(1-v^2)\mathbf{v}'\Delta\phi+\{\tfrac{1}{2}(1-\mathbf{v}^2)\mathbf{v}''+2(\mathbf{v}\cdot\mathbf{v}')\mathbf{v}'-(\mathbf{v}'\cdot\mathbf{v}')\mathbf{v}\}$$
$$\times(\nabla\phi)^2=0. \tag{2.2}$$

Here, the prime appended to $\mathbf{v}$ denotes the differentiation with respect to ϕ, and Δ and ∇ are the Laplace and the gradient operators in four-dimensional Euclidean space, respectively. Furthermore, we may decouple Eq. (2.2) into the following two equations:

$$\tfrac{1}{2}(1-v^2)\mathbf{v}''+2(\mathbf{v}\cdot\mathbf{v}')\mathbf{v}'-(\mathbf{v}'\cdot\mathbf{v}')\mathbf{v}=0, \tag{2.3}$$

$$\Delta\phi=0. \tag{2.4}$$

These are the basic equations that we consider in this paper.

We shall now integrate Eq. (2.3). First, it follows from the vector product of $\mathbf{v}$ and Eq. (2.3) that

$$\tfrac{1}{2}(1-v^2)(\mathbf{v}\times\mathbf{v}')'+2vv'\mathbf{v}\times\mathbf{v}'=0, \tag{2.5}$$

which is readily integrated as

$$\mathbf{v}\times\mathbf{v}'=\mathbf{c}(1-v^2)^2, \tag{2.6}$$

where

$$\mathbf{c}=(c_1,c_2,c_3),\quad c=|\mathbf{c}| \tag{2.7}$$

is a real constant vector. Denoting the components of $\mathbf{v}$ by

$$\mathbf{v}=(F,G,H), \tag{2.8}$$

and introducing the functions f, g, and h through the relations

$$F=(1-v^2)f, \tag{2.9a}$$

$$G=(1-v^2)g, \tag{2.9b}$$

$$H=(1-v^2)h, \tag{2.9c}$$

the vector equation (2.6) is equivalent to the following equations:

$$fg'-f'g=c_1, \tag{2.10a}$$

$$gh'-g'h=c_2, \tag{2.10b}$$

$$hf'-h'f=c_3. \tag{2.10c}$$

From (2.10), one finds that g and h are expressed in terms of f as

$$g=-c_1f\int^\phi\frac{d\phi}{f^2}, \tag{2.11}$$

$$h=-\frac{1}{c_1}\left(c_2-c_1c_3\int^\phi\frac{d\phi}{f^2}\right)f. \tag{2.12}$$

Next, taking the scalar product of Eq. (2.3) and $\mathbf{v}$ gives

$$\tfrac{1}{2}(1-v^2)(vv')'+2(vv')^2-\tfrac{1}{2}(1+v^2)\mathbf{v}'\cdot\mathbf{v}'=0. \tag{2.13}$$

Substituting the relation

$$\mathbf{v}'\cdot\mathbf{v}'=v'^2+c^2(1-v^2)^4/v^2, \tag{2.14}$$

which stems from the square of Eq. (2.6), into (2.13), one obtains the following ordinary differential equation for v:

$$\tfrac{1}{2}(1-v^2)(vv')'+2(vv')^2-\tfrac{1}{2}(1+v^2)$$
$$\times\{v'^2+c^2(1-v^2)^4/v^2\}=0. \tag{2.15}$$

0022-2488/90/040936-03$03.00

If we introduce the variable P by the relation

$$v^2 = 1 - (2/(P+1)), \tag{2.16}$$

then Eq. (2.15) is considerably simplified and it reads in the form

$$(P^2-1)P'' - PP'^2 - 16c^2P = 0. \tag{2.17}$$

This equation is readily integrated to yield the solution

$$P = d\cosh(a\phi + b), \tag{2.18a}$$

with

$$d = \sqrt{1 + 16c^2/a^2}, \tag{2.18b}$$

where a and b are real integration constants. Therefore, we have

$$v^2 = 1 - 2/[d\cosh(a\phi+b)+1]. \tag{2.19}$$

At this stage, the procedure to obtain $\mathbf{v}$ is straightforward. First, substitution of Eq. (2.11) and Eq. (2.12) into the relation $v^2 = F^2 + G^2 + H^2 = (1-v^2)^2 \times (f^2 + g^2 + h^2)$ yields

$$\int^Q \left\{(c_1^2 + c_3^2)Q^2 - \frac{2c_2c_3}{c_1}Q + 1 + \frac{c_2^2}{c_1^2}\right\}^{-1} dQ = \int^\phi \frac{(1-v^2)^2}{v^2}\, d\phi, \tag{2.20}$$

where we have put

$$Q = \int^\phi \frac{d\phi}{f^2} \tag{2.21}$$

for simplicity. Integration of Eq. (2.20) is easily performed by noting (2.19). The result is

$$(c_1^2 + c_3^2)Q = (c_2c_3/c_1) + c\tan[\tan^{-1}\{(1/\sqrt{\mathrm{d}^2-1}) \times \tanh(a\phi+b)\} + \theta], \tag{2.22}$$

where θ is a real constant.

Finally, it follows from (2.9), (2.11), (2.12), (2.21), and (2.22) that one obtains, after some tedious calculations, the explicit expressions for the vector $\mathbf{v} = (F,G,H)$ as follows:

$$F = \pm \alpha R\cosh(a\phi + b - \delta), \tag{2.23a}$$

$$G = \pm \beta R\cosh(a\phi + b + \epsilon), \tag{2.23b}$$

$$H = \pm \gamma R\cosh(a\phi + b - \eta), \tag{2.23c}$$

where

$$R = \{d\cosh(a\phi+b)+1\}^{-1}, \tag{2.24a}$$

$$\alpha = c^{-1}\sqrt{(c_1^2 + c_3^2)(d^2\cos^2\theta - 1)}, \tag{2.24b}$$

$$\beta = c^{-1}\sqrt{\{d^2(c_2c_3\cos\theta + cc_1\sin\theta)^2 - (cc_1)^2 - (c_2c_3)^2\}/(c_1^2+c_3^2)}, \tag{2.24c}$$

$$\gamma = c^{-1}\sqrt{\{d^2(c_1c_2\cos\theta - cc_3\sin\theta)^2 - (c_1c_2)^2 - (cc_3)^2\}/(c_1^2+c_3^2)}, \tag{2.24d}$$

$$\delta = \tanh^{-1}(\tan\theta/\sqrt{d^2-1}), \tag{2.24e}$$

$$\epsilon = \tanh^{-1}\left\{\frac{1}{\sqrt{d^2-1}}\,\frac{cc_1 - c_2c_3\tan\theta}{c_2c_3 + cc_1\tan\theta}\right\}, \tag{2.24f}$$

$$\eta = \tanh^{-1}\left\{\frac{1}{\sqrt{d^2-1}}\,\frac{cc_3 + c_1c_2\tan\theta}{c_1c_2 - cc_3\tan\theta}\right\}. \tag{2.24g}$$

It should be remarked that the arbitrary constants included in (2.23) are c_1, c_2, c_3, a, b, and θ, while c and d are expressed by these constants [see (2.7) and (2.18b)] and hence (2.23) represents a general solution of Eq. (2.3). The solutions of Eq. (2.1) are then determined perfectly by (2.23) and solutions of Eq. (2.4). Although various classes of exact solutions exist for Eq. (2.4), we shall not discuss them here.

III. GAUGE POTENTIALS

The gauge potentials $\mathbf{b}_\mu$ ($\mu = 1\sim4$) in the K gauge are expressed in terms of $\mathbf{v}$ as follows[1]:

$$\mathbf{b}_1 = 2(\mathbf{v}\times\mathbf{v}_1 + \mathbf{v}_2)(1-v^2)^{-1}, \tag{3.1a}$$

$$\mathbf{b}_2 = 2(\mathbf{v}\times\mathbf{v}_2 - \mathbf{v}_1)(1-v^2)^{-1}, \tag{3.1b}$$

$$\mathbf{b}_3 = 2(\mathbf{v}\times\mathbf{v}_3 - \mathbf{v}_4)(1-v^2)^{-1}, \tag{3.1c}$$

$$\mathbf{b}_4 = 2(\mathbf{v}\times\mathbf{v}_4 + \mathbf{v}_3)(1-v^2)^{-1}. \tag{3.1d}$$

These quantities are easily evaluated by using (2.6) and (2.23). The results are expressed in the form

$$\mathbf{b}_1 = R(4\phi_1\mathbf{c} + \phi_2\mathbf{B}), \tag{3.2a}$$

$$\mathbf{b}_2 = R(4\phi_2\mathbf{c} - \phi_1\mathbf{B}), \tag{3.2b}$$

$$\mathbf{b}_3 = R(4\phi_3\mathbf{c} - \phi_4\mathbf{B}), \tag{3.2c}$$

$$\mathbf{b}_4 = R(4\phi_4\mathbf{c} + \phi_3\mathbf{B}), \tag{3.2d}$$

where the components of the vector $\mathbf{B} = (B_1,B_2,B_3)$ are given by

$$B_1 = \pm a\alpha\{\sinh(a\phi + b - \delta) - d\sinh\delta\}, \tag{3.3a}$$

$$B_2 = \pm a\beta\{\sinh(a\phi + b + \epsilon) + d\sinh\epsilon\}, \tag{3.3b}$$

$$B_3 = \pm a\gamma\{\sinh(a\phi + b - \eta) - d\sinh\eta\}, \tag{3.3c}$$

and $\phi_\mu = \partial\phi/\partial x_\mu$. The field strengths $\mathbf{f}_{\mu\nu}$, defined by

$$\mathbf{f}_{\mu\nu} = \frac{\partial\mathbf{b}_\mu}{\partial x_\nu} - \frac{\partial\mathbf{b}_\nu}{\partial x_\mu} - \mathbf{b}_\mu\times\mathbf{b}_\nu, \tag{3.4}$$

are then derived from (3.2), the explicit expressions of which are not written down here. One can observe from (3.2) and (3.3) that the gauge potentials take finite values provided that ϕ_μ ($\mu = 1\sim4$) are finite.

ACKNOWLEDGMENT

The author thanks Professor M. Nishioka for informative discussions about gauge theories.

[1] N. Yang, Phys. Rev. Lett. **38**, 1317 (1977).
[2] S. Takeno, Prog. Theor. Phys. **66**, 1250 (1981).
[3] P. K. Chanda and D. Ray, Phys. Rev. D **31**, 3183 (1985).
[4] B. Leaute and G. Marcilhacy, J. Math. Phys. **28**, 774 (1987).
[5] D. Papadopoulos, Phys. Lett. A **127**, 341 (1988).

Bargmann–Wigner equations in de Sitter space

W. F. Heidenreich
Institut für Theoretische Physik A, TU Clausthal, D 3392 Clausthal-Zellerfeld, West Germany
M. Lorente
Departamento de F ísica, Facultad de Ciencias, Universidad de Oviedo, E 33007 Oviedo, Spain

(Received 8 August 1989; accepted for publication 15 November 1989)

Bargmann–Wigner equations in (3,2)-de Sitter space are found for all spins $s \geqslant 1$ and all masses. The massless fields have gauge freedom; they can be extended to indecomposable representations of the form of Gupta–Bleuler triplets.

I. INTRODUCTION

Although there is no experimental evidence for elementary particles with spin higher than 1, at least particles up to spin 2 are considered necessary by theoreticians for the description of gravity. On the other side there is strong evidence from the various tests of general relativity for space-time being curved. Consequently, higher spin fields in curved space-time have been considered in the literature (see, e.g., Ref. 1).

In flat space, spin is connected to the quantum numbers of the little group of the Poincaré group, i.e., to SU(2) for massive particles and to E(2) for massless ones. We conclude that group-theoretic techniques should be emphasized if possible.

All the physically essential concepts of flat space can be generalized to (3,2)-de Sitter space with constant negative curvature. Its group of motion SO(3,2) is a deformation of the Poincaré group SO(3,1)3⊗T(4) (resp. a covering thereof). In this space, equations for arbitrarily high spin have been discussed by various authors,[2] mostly using tensors or spinor–tensors. A detailed analysis has been given in the cases of massless spin-1 and spin-2 fields.[3,4] As in flat space, a description using field potentials with gauge freedom is possible in these cases. The particles are not described by an irreducible representation of the group of motions, but by an indecomposable one, of the type of a Gupta–Bleuler triplet. Now—in contrast to flat space—there are two inequivalent Gupta–Bleuler triplets, i.e., two inequivalent fields with gauge freedom in de Sitter space. They also differ with respect to a discrete reflection $\mathscr{R}$. This doubling of fields corresponds to the two helicities of flat space. Two formulations for massive fields, which differ in their $\mathscr{R}$ behavior, are also possible, but they are equivalent under SO(3,2).

Here we investigate multispinor fields for all spins $\geqslant 1$, integer or half-integer. Our notation as explained below follows Refs. 5 and 6, which discuss spinor fields with and without gauge freedom. In Sec. III we treat the massive case, in Sec. IV the massless one, emphasizing their gauge freedom. In both cases we first use the second-order Casimir operator as a field equation. Stronger restrictions are obtained by employing the Bargmann–Wigner field equations.[7] They give us two fields of $2s$-multispinors to each spin $s \geqslant 1$. In the massless cases their solution spaces carry inequivalent Gupta–Bleuler triplets. Appendix B gives some explicit states of these triplets.

II. PRELIMINARIES AND NOTATION

We consider fully symmetric multispinor fields

$$\Psi_{\{A_1\cdots A_{2r}\}}(u), \quad r = 1, \tfrac{3}{2}, 2, \ldots, \tag{1}$$

on (3,2)-de Sitter space. The coordinates u_α, $\alpha = 1,2,3,4,6$, satisfy

$$\eta_{\alpha\beta} u^\alpha u^\beta \equiv u_\alpha u^\alpha \equiv u_1^2 + u_2^2 + u_3^2 - u_4^2 - u_6^2 < 0, \quad u = \lambda u, \quad \lambda > 0; \tag{2}$$

the index 5 is omitted, to allow a straightforward extension to conformal space. The projective coordinates of this space of half-rays cover the de Sitter hyperbola only. All statements we make can be extended to the universal covering space, at the price of more involved expressions. The spinor indices A_i take values 1 to 4.

The reflection $\mathscr{R}$ maps $u_\alpha \to -u_\alpha$; fields symmetric (resp. antisymmetric) under $\mathscr{R}$ transform like

$$\Psi(-u) = \pm \Psi(u). \tag{3}$$

The de Sitter Lie algebra so(3,2)$\cong$sp(4,R) acts on these fields as

$$J_{\alpha\beta} = L_{\alpha\beta} + S_{\alpha\beta}, \tag{4}$$

where L is the orbital part acting on the coordinates,

$$L_{\alpha\beta} = -i(u_\alpha \partial_\beta - u_\beta \partial_\alpha), \tag{5}$$

and S is the spin part, acting on the indices,

$$S = \sum_{i=1}^{2r} \overset{(i)}{S}, \tag{6}$$

where $\overset{(i)}{S}$ acts only on the index A_i by

$$(S_{\alpha\beta})_{B_i A_i} = \overset{(i)}{S}_{\alpha\beta} = (1/4i)(\overset{(i)}{\beta}_\alpha \gamma_\beta - \overset{(i)}{\beta}_\beta \gamma_\alpha). \tag{7}$$

The 4×4 matrices β and γ satisfy

$$\beta_\alpha \gamma_\beta + \beta_\beta \gamma_\alpha = 2\eta_{\alpha\beta}, \tag{8}$$

for each i. Where necessary, we use the explicit base

$$\beta_j = \gamma_j = \begin{pmatrix} 0 & i\sigma_j \\ -i\sigma_j & 0 \end{pmatrix}, \quad \text{for} \quad j = 1,2,3,$$

$$\beta_4 = \gamma_4 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \tag{9}$$

0022-2488/90/040939-09$03.00

$$\beta_6 = -\gamma_6 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{10}$$

Here σ_j are the Pauli matrices. Sometimes we will also need

$$\beta_5 = \gamma_5 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -\prod_{\nu=1}^{4} \beta_\nu. \tag{11}$$

The form Eq. (4) of the elements of the Lie algebra means that our fields carry the tensor product of a finite representation described by the spin part and a scalar representation described by the orbital part:

$$\Psi_{\{A_1 \cdots A_{2r}\}}(u) = \psi_{\{A_1 \cdots A_{2r}\}} \otimes \phi(u). \tag{12}$$

The K subalgebra (Lie algebra of the maximal essentially compact subgroup) so(3) ⊕ so(2) of so(3,2) gives us three quantum numbers, which we call energy E, angular momentum l, and its third component l_3. We denote so(3) ⊕ so(2) multiplets by their quantum numbers as (E,l); they form a $(2l+1)$-dimensional space which we call a (generalized) weight. Irreducible lowest weight representations can be labeled uniquely by the quantum numbers of these lowest weights as $D(E_0,s)$.

The representations with $E_0 = 1 + s$ and $D(2,0)$ are called massless, as $D(1+s,s)$ for $s \geqslant \frac{1}{2}$ and $D(1,0) \oplus D(2,0)$ can be extended to the conformal group; those with $E_0 > 1 + s$ [except $D(2,0)$] we will call massive. With a de Sitter radius of 10^{26} m, the electron would have $E_0 \approx 10^{37}$. For

$$E_0 + s = 0, -1, -2, \ldots, \tag{13}$$

the irreducible representations $D(E_0,s)$ are finite [they are the unitary ones of so(5)].

The second-order Casimir operator $\frac{1}{2} J^2 = \frac{1}{2} J_{\alpha\beta} J^{\alpha\beta}$ of so(3,2) has eigenvalues

$$E_0(E_0 - 3) + s(s+1) \tag{14}$$

for the irreducible lowest weight representation $D(E_0,s)$.

The spinor ψ_A carries the four-dimensional representation, which is $D(-\frac{1}{2},\frac{1}{2})$ in our notation, the fully symmetric spinor $\psi_{\{A_1,\ldots,A_{2r}\}}$ carries $D(-r,r)$. The weight diagrams of these finite representations are

$$D(-r,r) = \bigoplus_{l=c}^{r} \bigoplus_{E=-l}^{l} (E,l), \tag{15}$$

with $c = 0$ for integer r and $c = \frac{1}{2}$ for half-integer r (see Fig. 1). Their eigenvalues of the second-order Casimir operator are

$$\tfrac{1}{2} S^2 = 2r(r+2). \tag{16}$$

The spinor ψ_A has an indefinite invariant scalar product, which in our basis of β and γ matrices is given by $\Sigma_A \bar{\psi}_A \psi_A$, where $\bar{\psi} = \psi^\dagger \gamma_4$. Therefore the multispinor has an indefinite invariant scalar product

$$\bar{\psi}\psi = \sum_{A_i} \bar{\psi}_{\{A_1 \cdots A_{2r}\}} \psi_{\{A_1 \cdots A_{2r}\}}, \tag{17}$$

where

$$\bar{\psi}_{\{A_1 \cdots A_{2r}\}} = \sum_{B_i} \psi^\dagger_{\{B_1 \cdots B_{2r}\}} (\gamma_4)_{B_1 A_1} \cdots (\gamma_4)_{B_{2r} A_{2r}}. \tag{18}$$

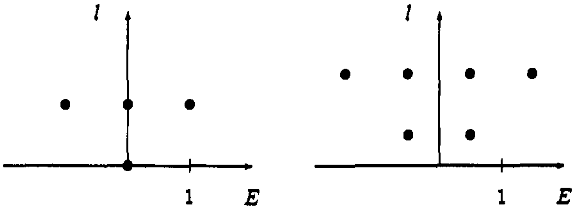


FIG. 1. Weight diagrams of $D(-1,1)$ and $D(-\frac{1}{2},\frac{1}{2})$.

Next we use the Casimir operator to get field equations for the scalar field $\phi(u)$. We denote the lowest energy of the corresponding scalar representation by N. Then $D(N,0)$ has the eigenvalues $N(N-3)$. So we get the second-order field equation

$$\begin{aligned} \tfrac{1}{2} L^2 \phi(u) &\equiv \tfrac{1}{2} L_{\alpha\beta} L^{\alpha\beta} \phi(u) \\ &= [-u^2 \partial^2 + (u\,\partial)(u\,\partial + 3)]\phi(u) \\ &= N(N-3)\phi(u). \end{aligned} \tag{19}$$

We have to fix the behavior of the fields along the half-rays; this can be done by fixing the degree of homogeneity of the fields to

$$(u\,\partial)\phi(u) \equiv (u^\alpha \partial_\alpha)\phi(u) = n\phi(u). \tag{20}$$

On de Sitter space (but not on its spatial infinity), n is arbitrary. We will discuss some choices later.

The solution space of Eq. (19) carries not only $D(N,0)$ but also $D(3-N,0)$. For $N > \frac{5}{2}$ the later representation is not unitarizable. We will not mention it further if it does not give unitarizable representations in the tensor products we consider later. In addition, there are negative energy representations; all our statements hold for them also.

The positive definite invariant scalar product of the field $\phi(u)$ has the form

$$-i \int d^3\bar{u}\, \phi^*(u) \overleftrightarrow{\partial}_t \phi(u), \tag{21}$$

where t is the de Sitter time, and the integration is over a spacelike hypersurface (see Ref. 3).

We conclude that the field equation

$$\begin{aligned} &(\tfrac{1}{2}L^2 - N(N-3))\Psi_{\{A_1 \cdots A_{2r}\}}(u) \\ &\quad = [-u^2 \partial^2 + (u\,\partial)(u\,\partial + 3) - N(N-3)] \\ &\qquad \times \Psi_{\{A_1 \cdots A_{2r}\}}(u) = 0 \end{aligned} \tag{22}$$

of the symmetric multispinor carries on its solution space the tensor product

$$D(-r,r) \otimes D(N,0). \tag{23}$$

This can be reduced by comparing weight diagrams if all terms of the Clebsch–Gordan series are away from reduction points. Reduction points appear if weights in the weight diagram are Weyl-equivalent to the lowest weight, i.e., if considered as lowest weight, they give the same eigenvalues for all Casimir operators. At these points the weight diagram of the irreducible representation is reduced by the weights of the Weyl-equivalent one. In particular, we need the first reduction points (reduction points with maximal energy), which appear at $E_0=\frac{1}{2}$ for the scalar representation, and at $E_0=1+s$, for $s\geqslant 1$. In the latter case the weight diagram of the massless representations $D(1+s,s)$ is reduced by the weights of $D(2+s,s-1)$. If we do not mention otherwise, we will be away from reduction points. Then the weights of the finite representation, shifted by the energy N, become lowest weights in the reduction, i.e., in our case

$$D(-r,r)\otimes D(N,0)=\bigoplus_{l=c}^{r}\ \bigoplus_{E=-l}^{l}\ D(N+E,l),\qquad(24)$$

with $c=0$ for integer r and $c=\frac{1}{2}$ for half-integer r (Ref. 5).

III. FIELD EQUATIONS FOR MASSIVE FIELDS OF ARBITRARY SPIN

We want to find field equations whose solution spaces carry any one of the terms in the Clebsch–Gordan series [Eq. (24)], where we restrict ourselves to $N>1+r$, to stay away from reduction points. For this purpose we can employ again the second-order Casimir operator, this time of the tensor product. With Eq. (4) it takes the form

$$\tfrac{1}{2}J^2=\tfrac{1}{2}L^2+LS+\tfrac{1}{2}S^2.\qquad(25)$$

By fixing the factors in the tensor product [Eq. (23)], we fix the eigenvalues of $\frac{1}{2}L^2$ [see Eq. (19)] and of $\frac{1}{2}S^2$ [see Eq. (16)]. Any term $D(E_0,s)$ in the Clebsch–Gordan series [Eq. (24)] has eigenvalues of $\frac{1}{2}J^2$ given by Eq. (14). If all of these are different—which can be shown for the values of r and N under consideration—then we can obtain field equations for each $D(E_0,s)$ by choosing the eigenvalues of LS appropriately. We get explicitly

$$\begin{aligned}LS\Psi(u)=[&E_0(E_0-3)+s(s+1)\\&-N(N-3)-2r(r+2)]\Psi(u).\end{aligned}\qquad(26)$$

The solution space of this equation and Eq. (22) carries a unitary $D(E_0,s)$; the scalar product is just a combination

$$-i\int d^3\bar{u}\,\overline{\Psi}(u)\overleftrightarrow{\partial}_t\Psi(u)\qquad(27)$$

of the scalar products [Eqs. (17) and (21)].

To each massive representation $D(E_0,s)$, with $s\geqslant 1$, $E_0>1+s$, which we consider here, there are infinitely many values $r\geqslant s$ and N, which can be used to construct field equations for it. But it would be unnecessarily complicated to use fields with more than $2s$ spinor indices, to describe spin-s objects. Even if we restrict ourselves to the cases $r=s$, we still have, in general, $2s+1$ possibilities to choose N for a given E_0 and s. These possibilities are further restricted if we use Bargmann–Wigner equations, which we consider next.

Bargmann–Wigner equations for massive fields: Using Eq. (6) we can decompose the term LS in the second-order Casimir operator into a sum

$$LS=\sum_{i=1}^{2s}\overset{(i)}{LS}.\qquad(28)$$

We put $r=s$ in the sequel. While LS commutes with L^2 and S^2, the various terms $\overset{(i)}{LS}$ do *not* commute among one another, and with S^2. So in general they will not have a common set of eigenfunctions. Only for particular cases may this be the case.

We want to consider the set of $2s$ equations for symmetric spinors,

$$\overset{(i)}{LS}\Psi_{\{A_1\cdots A_i\cdots A_{2s}\}}(u)=m\Psi_{\{A_1\cdots A_{2s}\}},\quad i=1,\ldots,2s.\qquad(29)$$

A straightforward calculation, using Eqs. (7) and (8), yields

$$(\overset{(i)}{LS})\,(\overset{(i)}{LS})=\tfrac{1}{2}L^2-3\overset{(i)}{LS},\quad i=1,\ldots,2s.\qquad(30)$$

So we get, for our field Ψ,

$$\tfrac{1}{2}L^2\Psi=m(m+3)\Psi.\qquad(31)$$

Comparing with Eq. (19) of a scalar field ϕ shows that the solutions of Eq. (29) are a subset of the tensor product

$$D(-s,s)\otimes D(N,0),\qquad(32)$$

with $N=-m$, or $N=m+3$.

In these two cases we get

$$LS\Psi=2sm\Psi=-2sN\Psi\ \ [\text{resp. } LS\Psi=2s(N-3)\Psi].\qquad(33)$$

Comparing with Eq. (26), we find that the first case requires $E_0=N-s$, and the second one requires $E_0=N+s$. (Further solutions $E_0=s-N+3$ [resp. $E_0=-(N+s-3)$] are ignored as they lead to representations that do not lie in the tensor product [Eq. (24)] with $r=s$.)

So a given set of $2s$ equations (29) can only describe $D(E_0,s)$ with

$$\begin{aligned}&E_0=-m-s,\quad\text{if } m<-(2s+1),\\&E_0=m+s+3,\quad\text{if}\quad m>-2.\end{aligned}\qquad(34)$$

The first case uses the term with the smallest lowest energy in the Clebsch–Gordan series [Eq. (24)], the second case uses the term with the largest one (see the circles in Fig. 2).

We can bring Eq. (29) in a more traditional form by using

$$\overset{(i)}{LS}=-(\beta^\alpha u_\alpha\overset{(i)}{\gamma^\beta}\partial_\beta)+(u\,\partial)\equiv-(\beta u\overset{(i)}{\gamma}\partial)+(u\,\partial).\qquad(35)$$

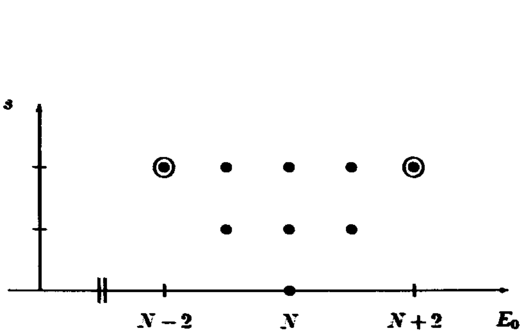


FIG. 2. Lowest weights in the tensor product $D(-2,2)\otimes D(N,0)$ (dots), and the terms used by the Bargmann–Wigner equations (circles).

By choosing the degree of homogeneity $n = -N$, we obtain, in the first case ($E_0 = N - s$),

$$(\beta u \overset{(i)}{\gamma} \partial)\Psi_1 = 0 \quad \text{or} \quad (\overset{(i)}{\gamma}\partial)\Psi_1 = 0, \tag{36}$$

and, in the second case ($E_0 = N + s$),

$$(\beta u \overset{(i)}{\gamma} \partial)\Psi_2 = (-2E_0 + 2s + 3)\Psi_2. \tag{37}$$

We have not yet shown that there are any normalizable solutions of Eq. (29) at all. In the first case, the tensor product (24) can help us again. The lowest state of $D(E_0,s)$ in this case has energy $E = N - s$, i.e., the lowest energy in the tensor product. The state with biggest third component of angular momentum, $l_3 = s$, is simply

$$\Psi_1^0 = \begin{pmatrix}1\\0\\0\\0\end{pmatrix} \times \cdots \times \begin{pmatrix}1\\0\\0\\0\end{pmatrix} u_+^{-N}, \quad \text{with } u_\pm = u_4 \pm iu_6, \tag{38}$$

where we have chosen the degree of homogeneity $n = -N$. It is easy to show that it satisfies Eq. (36) by using the explicit expressions

$$(\gamma\partial) = i\begin{pmatrix} 2\partial_- & \sigma_i\partial^i \\ -\sigma_i\partial^i & -2\partial_+ \end{pmatrix}, \quad \text{with } \partial_\pm = \frac{\partial}{\partial u_\pm}. \tag{39}$$

Because of the invariance of the equations, they must carry a $D(N-s,s)$ on the solution space.

In the second case $E_0 = N + s$, the corresponding lowest weight solutions look more difficult. If we introduce the shorthand notation

$$G_0 = 1, \quad G_q = \sum_{1<i_1<\cdots<i_q}^{2s} (\beta u \overset{(i_1)}{\gamma}_5)\cdots(\beta u \overset{(i_q)}{\gamma}_5), \quad q = 1,2,..., \tag{40}$$

then they are

$$\Psi_2^0 = \sum_{r=0}^{[s]} c_r (u^2)^r G_{2(s-r)} u_+^{-N-2s} \underbrace{\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \cdots \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix}}_{2s \text{ terms}}, \tag{41}$$

where the coefficients are given by the recursion relation

$$c_0 = 1, \quad c_{r+1} = [(2r+1)/(1-2N-2r)]c_r. \tag{42}$$

The solutions transform under $\mathscr{R}$ as

$$\Psi_{1,2}(-u) = (-1)^{E_0 \pm s}\Psi_{1,2}(u). \tag{43}$$

For integer spins, the two fields to a given representation transform the same way; for half-integer spins they differ in their sign.

For spin 1 the two fields that carry the same representation $D(E_0,1)$ are related by the invertible transformation

$$\Psi_2(u) = \left[(\beta u \overset{(1)}{\gamma}_5)(\beta u \overset{(2)}{\gamma}_5) + u^2/(3-2E_0)\right]\Psi_1(u); \tag{44}$$

similar formulas hold for all spins.

It may be esthetically pleasing to give yet another form of the Bargmann–Wigner equations. We obtain with

$$(u\,\partial)\Psi_1' = -(2s+1)\Psi_1' : \quad (\beta u \overset{(1)}{\gamma} \partial)\Psi_1' = (E_0 - s - 1)\Psi_1', \tag{45}$$

in the first case, and with

$$(u\,\partial)\Psi_2' = -2\Psi_2' : \quad (\beta u \overset{(i)}{\gamma} \partial)\Psi_2' = -(E_0 - s - 1)\Psi_2', \tag{46}$$

in the second case. For spin-$\frac{1}{2}$ these equations are discussed in Ref. 5. They look particularly simple in the massless limit, which we consider next.

IV. MASSLESS FIELDS

A. Group-theoretic analysis

Here we want to describe massless fields with spin $s \geqslant 1$ using fully symmetric multispinors with $2s$ indices. So we consider only the biggest possible spins, which can be obtained with a given multispinor. As compared to our discussion of massive fields, the situation is complicated by terms on reduction points in the tensor product Eq. (24), and the appearance of indecomposable representations. Correspondingly we have massless particles with gauge freedom.

The irreducible massless representations $D(1+s,s)$, $s \geqslant 1$, are Weyl-equivalent to

$$D(1+s,s) \overset{w}{\sim} D(2+s,s-1)$$
$$\overset{w}{\sim} D(2-s,s,) \overset{w}{\sim} D(1-s,s-1) \tag{47}$$

(see Fig. 3). The second term is always a unitary representation, the third one an infinite not unitarizable one, and the last one is finite dimensional. All of them are used below.

The massless field can appear as an invariant subspace, or in a quotient space of an indecomposable representation. In such a representation, only the terms in Eq. (47) can be subrepresentations together with the massless one. As was mentioned before, the irreducible massless representations $D(1+s,s)$ have a weight diagram, which is reduced by the weights of the $D(2+s,s-1)$, as compared with the general

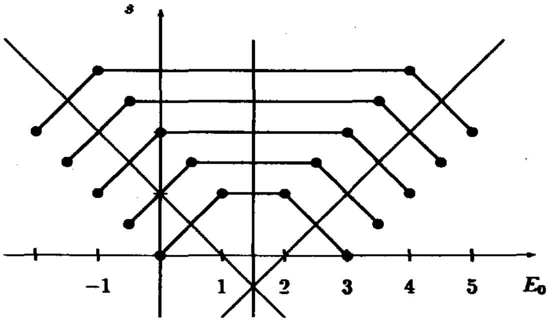


FIG. 3. The irreducible lowest weight representations that may appear in indecomposable representations with massless particles of spin $s<3$ are connected by fat lines.

case. Similarly, the weight diagram of the $D(2-s,s)$ is reduced by the weights of $D(1+s,s)$, and that of $D(1-s,s-1)$ is reduced by the weights of $D(2-s,s)$ (leaving only finitely many).

The solution spaces of the scalar field equations (19) also require special considerations now. They carry a direct sum of the lowest weight representations $D(N',0)$, with $N'=N$ and $N'=3-N$. Here we will need integer values of N. For $N=2$ and $N=1$ we have $D(2,0)$ and $D(1,0)$ in the solution space. These are the two unitary massless representations of the de Sitter Lie algebra. For nonpositive integer N', $D(N',0)$ is finite. A closer look shows that it is the invariant subspace of an indecomposable

$$D(1,1-N')\rightarrow D(N',0), \tag{48}$$

which solves the scalar field equation (19).

Keeping these complications in mind, we can find the values of N for which the tensor product [Eq. (23)] with $r=s\geqslant 1$ can contain massless representations with spin s in the reduction, i.e., we consider

$$\begin{aligned} &D(-s,s)\otimes D(N,0), \quad \text{for } N\notin\{0,-1,-2,...\} \\ &(\text{resp. } D(-s,s)\otimes[D(1,1-N)\rightarrow D(N,0)], \\ &\qquad\qquad \text{for } N\in\{0,-1,-2,...\}). \end{aligned} \tag{49}$$

We find that N must be integer, and

$$(2-2s)\leqslant N\leqslant 1+2s. \tag{50}$$

For the case of spin 1, see Fig. 4.

As in the massive case, we can use the second-order Casimir operator to find field equations for the various summands in the decompositions of the tensor products, which have different Casimir eigenvalues. Again we do not want to discuss all possibilties, but restrict ourselves to Bargmann–Wigner-like equations (29), i.e.,

$$\overset{(i)}{LS}\Psi_{\{A_1,\dots,A_i,\dots,A_{2s}\}}(u)=m\Psi_{\{A_1,\dots A_{2s}\}}(u), \quad i=1,2,\dots,2s. \tag{51}$$

With the same steps as from Eq. (29) to Eq. (34), but accepting all four solutions for the lowest energy N of the scalar representations, we get for the massless cases $E_0=1+s$, $s\geqslant 1$, the values

$$N\in\{1+2s,2,1,2-2s\}. \tag{52}$$

The first and last one (resp. the values $N=1,2$) have the same scalar field equations.

Using the information above, we conjecture that the tensor products [Eqs. (49)] contain in these cases the following Gupta–Bleuler triplets:

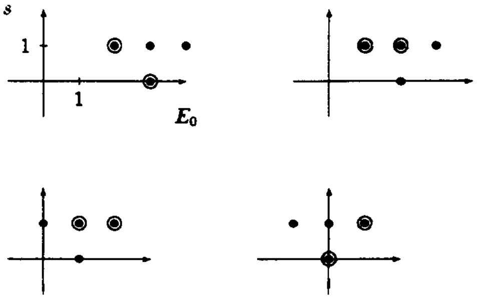


FIG. 4. The dots represent lowest weights in the reduction of $D(-1,1)\otimes D(N,0)$ for $N=3$, 2, 1, and 0. Terms Weyl-equivalent to the photon representation $D(2,1)$ are encircled; lowest weights of invariant subspaces are not shown.

$$D(-s,s,)\otimes D(1+2s,0)=D(2+s,s,-1)\rightarrow D(1+s,s)\rightarrow D(2+s,s-1)\oplus\cdots, \tag{53}$$

$$D(-s,s,)\otimes D(2,0)=D(1+s,s)\rightarrow D(2-s,s)\rightarrow D(1+s,s)\oplus\cdots, \tag{54}$$

$$D(-s,s,)\otimes D(1,0)=D(1+s,s)\rightarrow D(2-s,s)\rightarrow D(1+s,s)\oplus\cdots, \tag{55}$$

$$D(-s,s,)\otimes(D(1,2s-1)\rightarrow D(2-2s,0))=D(2-s,s)\rightarrow\{D(1+s,s)\oplus D(1-s,s-1)\}\rightarrow D(2-s,s)\oplus\cdots. \tag{56}$$

In the first and in the last case, the massless representations appear in the central part of the triplet; this we expect to give a field potential with gauge freedom. In the other two cases, we have the massless representation in an invariant subspace; these we want to use, to describe them irreducibly, as field strengths.

B. Field equations

As in the massive case, we first discuss equations that can be obtained from the second-order Casimir operator, and later use the explicit ground states from Appendix B to find the solution spaces of the Bargmann–Wigner equations.

The eigenvalues of the second-order Casimir operator are $2(s^2-1)$ for the massless representation $D(1+s,s)$; so we expect for it the equation

$$Q\Psi\equiv(\tfrac{1}{2}J^2-2(s^2-1))\Psi=0, \tag{57}$$

which also holds for the invariant subspace of pure gauge states. For the field potential we choose the degree of homogeneity

$$u\,\partial\Psi=-(1+2s)\Psi; \tag{58}$$

then the scalar field equation is

$$\partial^2\Psi=0. \tag{59}$$

For the solutions of this equation, we get

$$Q=-\sum_{i=1}^{2s}(\beta u\overset{(i)}{\gamma}\partial). \tag{60}$$

In all our explicit cases given in Appendix B, the scalar modes are mapped by Q on gauge modes; then the full triplet satisfies only

$$Q^2\Psi=0. \tag{61}$$

1. Spin 1

The description of massless spin-1 particles using a vector field gives in de Sitter space two inequivalent field potentials of electrodynamics,[3] whose pure gauge states carry a $D(1,1)$ [resp. a $D(3,0)$]. The corresponding inequivalent Gupta–Bleuler triplets also appear in our Bargmann–Wigner case here. In Appendix B 1 we give the states with lowest energy of the Gupta–Bleuler triplet of the (3,0)-gauge theory, and the leaks between them. We call it the high triplet. With the help of the explicit states we can find some more simple field equations for various subspaces. The states of the (1,1)-gauge theory look most simple if we choose degree of homogeneity 0 for them, as is done in Appendix B 2; they belong to what we call the low triplet. For our purposes here we want fields with $u\,\partial\Psi = -3\Psi$. To achieve this we can multiply the states of Appendix B 2 by $(u^2)^{-3/2}$, i.e., we use $\Psi(u) = (u^2)^{-3/2}\psi(u)$. This does not affect the action of the raising and lowering operators, and the states satisfy both Eq. (58) and the wave equation [Eq. (59)]. As we will discuss, with this choice of degree the same equations hold for invariant subspaces in both cases, the (1,1)- and the (3,0)-gauge theories.

The pure gauge states satisfy for both triplets

$$\sum_{i=1}^{2} \overset{(i)}{(\beta u\gamma_5)}\Psi_g = 0, \tag{62}$$

while the same operator on the physical ground state gives

$$\sum_{i=1}^{2} \overset{(i)}{(\beta u\gamma_5)}\Psi_p^0 = \frac{\overset{(1)}{(\beta u\gamma_5)} + \overset{(2)}{(\beta u\gamma_5)}}{u_+^3}\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \tag{63}$$

in the case of the (3,0) theory, and

$$\sum_{i=1}^{2} \overset{(i)}{(\beta u\gamma_5)}(u^2)^{-3/2}\psi_p^0 = -\frac{16}{\sqrt{2}}(u^2)^{-1/2}\frac{\overset{(1)}{(\beta u\gamma_5)}\overset{(2)}{(\beta u\gamma_5)} - u^2}{u_+^3}\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \tag{64}$$

in the case of the (1,1) theory. The finite state [Eq. (B7)] of the (1,1) theory also satisfies Eq. (62). So the solutions of Eq. (62) are the invariant subspace of pure gauge modes and the finite state only.

In the vector field description of massless spin-1 particles, the pure gauge states have the form $A_\mu = \partial_\mu\Lambda$, i.e., the vector field is obtained from a scalar field. To find here a corresponding expression, we make the ansatz $\Psi_g(u) = M^\alpha\,\partial_\alpha\Lambda(u)$, where $\Lambda(u)$ is a scalar field, and M^α is a 4×4 matrix; we have to choose it in such a way that the transformation properties of the scalar and the two-spinor Ψ match. This condition requires

$$\Psi_g = \left[\overset{(1)}{(\beta u\gamma_5)}(\beta^\alpha\gamma_1\beta_3) + \overset{(2)}{(\beta u\gamma_5)}(\beta^\alpha\gamma_1\beta_3)^T\right]\partial_\alpha\Lambda(u). \tag{65}$$

The explicit form stems from γ_α being antisymmetric for $\alpha = 1,3$ and symmetric otherwise. It can also be checked by application on the ground states, e.g., $\Lambda(u) = u_+^{-3}$ for the (3,0) theory. The same formula holds for the finite state [Eq. (B7)] with $\Lambda(u) = (i/2)\ln u_+$.

The physical states satisfy the Bargmann–Wigner equations (36):

$$\overset{(i)}{(\gamma\,\partial)}\Psi(u) = 0, \quad i = 1,2; \tag{66}$$

the same must hold in the invariant subspace of pure gauge states, but it does not hold for the finite state [Eq. (B7)] of the (1,1) theory.

The full triplet satisfies, in addition to the wave equation [Eq. (59)], the conditions

$$\sum_{i=1}^{2} \overset{(i)}{(\beta_5\gamma\,\partial)}\Psi = 0, \tag{67}$$

which can be checked by explicit application on the states Eqs. (B4) and (B14). If they are satisfied, the condition Eq. (61) also holds.

All the equations discussed here for spin 1 have solutions belonging to the two inequivalent indefinite Hilbert spaces of the (3,0)- [resp. the (1,1)-] gauge theories. These two solution manifolds differ also in their behavior under the reflection $\mathscr{R}$. If we call the corresponding fields Ψ_1 (resp. Ψ_2), then

$$\Psi_{1,2}(-u) = \mp\Psi_{1,2}(u), \tag{68}$$

i.e., the minus sign holds for the (3,0) theory and the plus sign for the (1,1) theory.

The usual formulation of electrodynamics not only has a vector potential A_μ, which describes photons with gauge freedom, but also a field strength $F_{\mu\nu}$, which carries the photon representation irreducibly. We discuss its analog next.

Field strength: The states on the right-hand side of Eqs. (63) and (64) are ground states of photon representations, which, because of Eq. (62) and the commutator $[J_{\alpha\beta},\beta u\gamma_5] = 0$, do not leak into pure gauge states; they are the ground states of the irreducible invariant subspaces in the second (resp. third) tensor products Eqs. (54) and (55). The corresponding fields

$$\chi_{1,2} = \sum_{i=1}^{2} \overset{(i)}{(\beta u\gamma_5)}\Psi_{1,2} \tag{69}$$

are the field strengths of the electromagnetic two-spinor. The two field strengths are related by

$$\begin{aligned}\chi_1 &= -\frac{1}{2}(u^2)^{-1/2}\sum_{i=1}^{2}\overset{(i)}{(\beta u\gamma_5)}\chi_2,\\ \chi_2 &= \frac{1}{2}(u^2)^{-1/2}\sum_{i=1}^{2}\overset{(i)}{(\beta u\gamma_5)}\chi_1.\end{aligned} \tag{70}$$

They satisfy the Bargmann–Wigner equations

$$(u\,\partial)\chi = -2\chi, \quad \overset{(i)}{(\gamma\,\partial)}\chi = 0, \quad i = 1,2, \tag{71}$$

and they differ in their behavior with respect to the discrete reflection $\mathscr{R}$:

$$\chi_{1,2}(-u) = \pm\chi_{1,2}(u). \tag{72}$$

2. All spins

For spin $\geqslant \frac{3}{2}$, we have explicit states only for the high triplet [Eq. (53)] given in Appendix B3. So the equations discussed below have been checked for these states only, although we conjecture that they also hold for the states of the low triplet in the tensor product [Eq. (56)]. The full triplet satisfies, in addition to the field equations (58) and (59),

$$\sum_{i=1}^{2s} (\beta_5 \overset{(i)}{\gamma} \partial)\Psi = 0. \tag{73}$$

The invariant subspaces of physical and gauge modes satisfy in addition the Bargmann–Wigner equations:

$$(\overset{(i)}{\gamma} \partial)\Psi(u) = 0, \quad i = 1,...,2s. \tag{74}$$

If we use the shorthand notation of Eq. (40), then the pure gauge states satisfy the equation

$$G\Psi_g \equiv \sum_{r=0}^{[s-1/2]} (-u^2)^r G_{2(s-r)-1}\Psi_g = 0. \tag{75}$$

As for spin 1 we conjecture the same for the finite representation of the low triplet.

The $\mathscr{R}$-quantum numbers of the states of the high triplets are given by

$$\Psi_1(-u) = (-1)^{2s+1}\Psi_1(u), \tag{76}$$

while the fields $\Psi_2(u)$ of low triplets have the opposite $\mathscr{R}$-quantum numbers.

Acting with the operator G from Eq. (75) on the lowest physical states, we obtain states with degree of homogeneity (-2). The fields

$$\chi_{1,2} = G\Psi_{1,2} \tag{77}$$

are the field strengths of the spin s fields. They satisfy

$$(u\,\partial)\chi = -2\chi, \quad (\overset{(i)}{\gamma} \partial)\chi = 0, \quad i = 1,...,2s. \tag{78}$$

ACKNOWLEDGMENTS

W. H. thanks the Departmento de Física of the Universidad de Oviedo for their hospitality; M. L. thanks the Arnold Sommerfeld Institute for the hospitality extended to him.

This work was partially supported by the Vicerectorado de Investigación at the Universidad de Oviedo.

APPENDIX A: CARTAN BASIS

In order to calculate explicit states, we use the explicit form of the Cartan basis. Here we give it in terms of the basis $J_{\alpha\beta}$ of the Lie algebra so(3,2). With the energy raising and lowering operators

$$J_j^{\pm} = J_{j6} \pm iJ_{j4}, \quad j = 1,2,3, \tag{A1}$$

we use

$$H_1 = J_{46}, \tag{A2}$$

$$H_2 = J_{12}, \tag{A3}$$

$$E_{\pm\mu} = (1/\sqrt{2})(J_{23} \pm iJ_{31}), \tag{A4}$$

$$E_{\pm\alpha_1} = \pm(i/\sqrt{2})J_3^{\pm}, \tag{A5}$$

$$E_{\pm\alpha_2} = \pm(i/2)(J_1^{\pm} \pm iJ_2^{\pm}), \tag{A6}$$

$$E_{\pm\beta} = \mp(i/2)(J_1^{\pm} \mp iJ_2^{\pm}), \tag{A7}$$

where E_μ is the simple compact root vector, E_β the simple noncompact root vector, and E_{α_1} and E_{α_2} are noncompact root vectors. In a root space with Cartesian basis e_1,e_2 corresponding to H_1,H_2, the roots have the following values: $\mu = (0,1), \beta = (1,-1), \alpha_1 = (1,0)$, and $\alpha_2 = (1,1)$.

APPENDIX B: SOME EXPLICIT STATES

In this appendix we collect some explicit states, which are used in the text, and give the action of ladder operators on them.

1. Spin 1, high triplet

The lowest state in the triplet

$$D(3,0) \to D(2,1) \to D(3,0) \tag{B1}$$

is

$$\Psi_p^0 = \frac{1}{u_+^3}\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix}. \tag{B2}$$

It has quantum numbers $(E = 2, l = 1, l_3 = -1)$. Acting with the raising operator

$$E_\beta E_\mu^2 + E_{\alpha_1}E_\mu + E_{\alpha_2},$$

we get the state

$$\Psi_g^0 = 3i\frac{(\beta u \overset{(1)}{\gamma_5}) - (\beta u \overset{(2)}{\gamma_5})}{2u_+^4} \times \left[\begin{pmatrix}1\\0\\0\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix} - \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \begin{pmatrix}1\\0\\0\\0\end{pmatrix}\right], \tag{B3}$$

with quantum numbers $(E = 3, l = 0)$. The state

$$\Psi_s^0 = \frac{3}{2u_+^3}\left[\begin{pmatrix}0\\0\\1\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix} - \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \times \begin{pmatrix}1\\0\\0\\0\end{pmatrix}\right] + \text{symm} \tag{B4}$$

has the same quantum numbers. It is a relative lowest weight, for which

$$E_{-\alpha_2}\Psi_s^0 = -(3i/2)\Psi_p^0 \tag{B5}$$

holds.

2. Spin 1, low triplet

Next we want to give states in the low spin-1 triplet

$$D(1,1) \to \{D(2,1) \oplus D(0,0)\} \to D(1,1). \tag{B6}$$

The absolute lowest state in the triplet is

$$\psi_f^0=\frac{\overset{(1)}{(\beta u\gamma_5)}-\overset{(2)}{(\beta u\gamma_5)}}{u_+}\times\left[\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}-\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\right], \tag{B7}$$

with quantum numbers ($E=0$, $l=0$). It belongs to the finite representation, which here is the trivial one. Acting with the raising operator E_β on it we get

$$\psi_g^0=\frac{1}{u_+^2}\left[\overset{(1)}{(\beta u\gamma_5)}\overset{(2)}{(\beta u\gamma_5)}+u^2\right]\left[\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\right], \tag{B8}$$

with quantum numbers ($E=1$, $l=1$, $l_3=-1$). A cyclic state ψ_s^0 for the full triplet, with the same quantum numbers, is

$$\begin{aligned}\psi_s^0=&\frac{4u^2}{u_+^2}\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}-\psi_g^0\\&-i\frac{\overset{(1)}{(\beta u\gamma_5)}-\overset{(2)}{(\beta u\gamma_5)}}{u_+}\\&\times\left[\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}-\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\right].\end{aligned} \tag{B9}$$

It can be constructed by taking the limit

$$\psi_s^0=\lim_{\epsilon\to 0}\left[(\psi_g^\epsilon-c\psi_1^\epsilon)/\epsilon\right], \tag{B10}$$

where ψ_g^ϵ is the lowest weight of $D(1+\epsilon,1)$, with $l_3=-1$, i.e., a deformation of ψ_g to a noninteger energy, and ψ_1^ϵ is the state at ($E=1+\epsilon$, $l=1$, $l_3=-1$) which belongs to a representation $D(\epsilon,0)$, i.e., a deformation of the finite one. The constant c is chosen such that $\lim_{\epsilon\to 0}(\psi_g^\epsilon-c\psi_1^\epsilon)=0$.

From this cyclic state we get $\psi_f^0=-E_{-\beta}\psi_s^0$ and also a "physical state" with quantum numbers ($E=2$, $l=1$, $l_3=-1$) by using the raising operators

$$(E_{\alpha_1}+E_\beta E_\mu)\psi_s^0=\psi_p^0. \tag{B11}$$

Explicitly it is

$$\begin{aligned}\psi_p^0=&\frac{-8u^2}{\sqrt{2}}\left[\frac{\overset{(1)}{(\beta u\gamma_5)}+\overset{(2)}{(\beta u\gamma_5)}}{u_+^3}\right]\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\\&+\frac{2i}{\sqrt{2}u_+^2}\left[u^2+\overset{(1)}{(\beta u\gamma_5)}\overset{(2)}{(\beta u\gamma_5)}\right]\\&\times\left[\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\0\\0\\1\end{pmatrix}+\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\right].\end{aligned} \tag{B12}$$

It satisfies $E_{-\alpha_1}\psi_p^0=\frac{1}{2}\psi_g^0$.

The two triplets are isomorphic to the triplets of de Sitter electrodynamics.[3]

3. All spins, high triplet

In the case of the high triplet

$$D(2+s,s-1)\to D(1+s,s)\to D(2+s,s-1), \tag{B13}$$

we also found an expression for the ground states for all spins $\geqslant 1$. A cyclic state with the quantum numbers ($E=s+2$, $l=s-1$, $l_3=-s+1$) is

$$\begin{aligned}\Psi_s^0=\frac{(2s+1)}{2u_+^{2s+1}}\Bigg[&(2s-1)\begin{pmatrix}0\\0\\1\\0\end{pmatrix}\times\underbrace{\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\cdots\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}}_{(2s-1)\text{ terms}}\\&\underbrace{-\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\times\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\cdots\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}-\cdots-\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\cdots\times\begin{pmatrix}0\\1\\0\\0\end{pmatrix}\times\begin{pmatrix}1\\0\\0\\0\end{pmatrix}}_{(2s-1)\text{ terms}}\Bigg]+\text{symm.}\end{aligned} \tag{B14}$$

Using a lowering operator on it,

$$E_{-\alpha_2}\Psi_s^0 = -is(4s^2-1)\Psi_p^0, \tag{B15}$$

we get the lowest "physical" state

$$\Psi_p^0 = \frac{1}{u_+^{2s+1}} \underbrace{\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \cdots \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix}}_{2s \text{ terms}}, \tag{B16}$$

with the quantum numbers $(E=s+1, l=s, l_3=-s)$. Acting on this one with the raising operators

$$[E_\beta E_\mu^2 + (2s-1)E_{\alpha_1}E_\mu + (2s-1)s\, E_{\alpha_2}]\Psi_p^0 = \Psi_g^0, \tag{B17}$$

we get a pure gauge state with the same quantum numbers as those of Ψ_s^0. It is explicitly

$$\begin{aligned}\Psi_g^0 = \frac{(2s+1)i(\overset{(1)}{\beta u}\gamma_5)}{2u_+^{2s+2}} & \left[(2s-1)\begin{pmatrix}1\\0\\0\\0\end{pmatrix} \times \underbrace{\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \cdots \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix}}_{(2s-1) \text{ terms}} \right. \\ & \left. \underbrace{-\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \begin{pmatrix}1\\0\\0\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \cdots \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix} - \cdots - \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \cdots \times \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \times \begin{pmatrix}1\\0\\0\\0\end{pmatrix}}_{(2s-1) \text{ terms}} \right] + \text{symm.}\end{aligned} \tag{B18}$$

[1] W. Rindler and R. Penrose, *Spinors and Spacetime* (Cambridge U.P., Cambridge, 1984).

[2] C. Fronsdal, Phys. Rev. D **20**, 848 (1979); J. Fang and C. Fronsdal, *ibid.* **22**, 136 (1980); M. A. Vasiliev, Fortschr. Phys. **35**, 741 (1987); J. P. Gazeau and M. Hans, J. Math. Phys. **29**, 2533 (1988).

[3] B. Binegar, C. Fronsdal, and W. Heidenreich, Ann. Phys. (NY) **149**, 254 (1983).

[4] C. Fronsdal and W. F. Heidenreich, J. Math. Phys. **28**, 215 (1987).

[5] W. Heidenreich, Nuovo Cimento A **80**, 220 (1984).

[6] W. F. Heidenreich, Phys. Rev. D **36**, 1685 (1987).

[7] V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. (USA) **34**, 211 (1948); M. A. Rodriguez and M. Lorente, Nuovo Cimento A **83**, 249 (1984); G. P. Collins and N. A. Doughty, J. Math. Phys. **28**, 448 (1987); W. F. Heidenreich and M. Lorente, *ibid.* **29**, 1698 (1988).

$N=2$ super Riemann surfaces and algebraic geometry

Gregorio Falqui and Cesare Reina
S.I.S.S.A, (International School for Advanced Studies), Strada Costiera 11, 34014 Trieste, Italy

(Received 3 May 1989; accepted for publication 4 October 1989)

The geometric framework for $N=2$ superconformal field theories are described by studying *susy$_2$ curves*—a nickname for $N=2$ super Riemann surfaces. It is proved that "single" susy$_2$ curves are actually split supermanifolds, and their local model is a Serre self-dual locally free sheaf of rank two over a smooth algebraic curve. Superconformal structures on these sheaves are then examined by setting up deformation theory as a first step in studying moduli problems.

I. INTRODUCTION

Supersymmetric extensions of algebraic curves have been recently studied in the physical and mathematical literature (see, e.g., Refs. 1 and 2). Among the physical motivations, a complete understanding of $N=1$ susy curves and their moduli spaces is needed in superstring theory in order to give meaning to computations in the Polyakov approach: Besides, they provide the natural arena for higher genus superconformal field theories.

From the mathematical point of view, superalgebraic curves are the simplest candidates for testing "new directions in geometry" in the spirit advocated by Manin.[3] Although studying "two supersymmetries" may seem the most obvious step beyond $N=1$, it is already a nontrivial matter, as noticed in some works.[4,5] In fact, for $N>1$ one is lead to consider locally free sheaves of rank greater than 1 on algebraic curves, a topic that is not completely under control as compared to the complete understanding of invertible sheaves. Luckily enough, for $N=2$, the superconformal structure to be imposed on such objects will bypass most of the subtleties related to moduli of vector bundles over variable curves: On the contrary, for $N\geqslant 3$, these enter the stage in a substantial way. Thus susy$_2$ curves are in a way the last "easy going" supersymmetric extension of algebraic curves, a fact that deserves special attention.

From the physical point of view, there is some "stringy" interest in the study of $N=2$ superconformal models; for instance, a recent work[6] has pointed out that space-time $N=1$ supersymmetry requires $N=2$ world-sheet supersymmetry. It is also widely believed that viewing $N=1$ supermoduli spaces as embedded in $N=2$ supermoduli spaces could be a keen standpoint for investigating the peculiarities of the former (provided that one has a good control of the latter).

The plan of this work is as follows. In Sec. II we investigate the geometry of susy$_2$ curves in connection with the theory of rank two locally free sheaves over an algebraic curve. Some nice features due to the existence of a superconformal structure, such as the splitness of "isolated" susy$_2$ curves, are proved. We also show that isolated susy$_2$ curves are the same thing as the datum of a Serre self-dual vector bundle on a curve and we classify such bundles completely. In Sec. III we set up a deformation theory and construct the local model for $N=2$ supermoduli spaces. Finally, Sec. IV is devoted to a detailed discussion of the global structure of the reduced moduli spaces of untwisted susy$_2$ curves.

II. SUSY$_2$ CURVES

This paper deals with susy$_2$ curves from the point of view of the theory of Kostant–Leites supermanifolds. In this framework, N supersymmetry is encoded in a $\mathbb{Z}_2$-graded extension $\mathscr{A}_X$ of the structure sheaf of a (complex) manifold such that $\mathscr{A}_X$ is locally isomorphic to the total wedge product of a rank N locally free analytic sheaf $\mathscr{E}$ [hereinafter called the *characteristic* sheaf of the supermanifold $(X, \mathscr{A}_X)$] over X (for a full definition, see, e.g., Ref. 2).

Recall that susy$_1$ curves are $1|1$-dimensional supercurves that come equipped with a distinguished distribution $\mathscr{D}$ in the tangent sheaf, spanned by the supersymmetry generator. In the same way the structure sheaf $\mathscr{A}_C$ of a susy$_2$ curve is quite special since it should embody the idea of the superconformal structure. In the physical literature this is realized in terms of coordinate transformations.[4,5] Here we give a definition that naturally extends that of susy$_1$ curves.[7,8]

Definition 1: A family of susy$_2$ curves $(C, \mathscr{A}_C)$ parametrized by the complex superspace $(S, \mathscr{A}_S)$—a susy$_2$ curve over S—is the datum of (i) a sheaf homomorphism $\pi^{\#}:\pi^{-1}\mathscr{A}_S\rightarrow\mathscr{A}_C$ of relative dimension $1|2$ over a proper surjective flat map $C\overset{\pi}{\rightarrow}S$ and (ii) a $0|2$-dimensional locally free distribution $\mathscr{D}_\pi$ in the relative tangent sheaf $\mathscr{T}_\pi$ such that the commutator mod $\mathscr{D}_\pi$,

$$\{,\}_{\mathscr{D}}:\mathscr{D}_\pi\otimes\mathscr{D}_\pi\rightarrow\mathscr{T}_\pi/\mathscr{D}_\pi,$$

is a symmetric nondegenerate bilinear map of sheaves of $\mathscr{A}_C$ modules.

In the following, a susy$_2$ curve over the trivial superspace $\{*\}$ will be called a single susy$_2$ curve. The connection between Definition 1 and the usual coordinate approach, as given, e.g., in Ref. 9, is a simple generalization of the $N=1$ case (see Refs. 7–10). Indeed, one can easily prove that there exist generators D^i for $\mathscr{D}_\pi$ and $\partial/\partial z$ for $\mathscr{T}_\pi/\mathscr{D}_\pi$ such that

$$\{D^i, D^j\}_{\mathscr{D}}=\delta^{ij}\frac{\partial}{\partial z}.$$

A simple computation then shows that $D^i=\partial/\partial\theta^i+\theta^i(\partial/\partial z)$.

Besides matching with physical applications, Definition 1 allows an immediate characterization of single susy$_2$ curves.

Proposition 1: Let $(C, \mathscr{A}_C)$ be a single susy$_2$ curve with reduced canonical sheaf ω. Then there exists a rank two lo-

0022-2488/90/040948-05$03.00

cally free sheaf $\mathscr{E}$ such that (i) $\mathscr{A}_C \simeq \Lambda\mathscr{E}$, i.e., $\mathscr{A}_C$ is split; and (ii) $\mathscr{E} \simeq \mathscr{E}^* \otimes \omega$, i.e., $\mathscr{E}$ is Serre self-dual.

Proof: Let $(U_\alpha, z_\alpha, \theta^i{}_\alpha)$ be a canonical atlas with transition functions

$$z_\alpha = f_{\alpha\beta}(z_\beta) + g_{\alpha\beta}\epsilon_{ij}\theta^i\theta^j,$$
$$\theta_\alpha{}^i = [m_{\alpha\beta}]^i_j\,\theta_\beta{}^j.$$

The existence of the distribution $\mathscr{D}_\pi$ is then equivalent to the superconformal condition

$$D^i{}_\beta z_\alpha = \theta^k{}_\alpha D^i{}_\beta \theta^k{}_\alpha$$

(sum over repeated latin indices), which gives

$$\epsilon_{ij}g_{\alpha\beta} + \delta_{ij}f'_{\alpha\beta} = [{}^t m_{\alpha\beta} m_{\alpha\beta}]_{ij},$$

where $f'_{\alpha\beta} = \partial f_{\alpha\beta}/\partial z_\beta$. Looking at the symmetric and antisymmetric parts of this equation we have (i) $g_{\alpha\beta} = 0$, so that $\mathscr{A}_C$ splits to $\Lambda\mathscr{E}$, where $\mathscr{E}$ is locally generated by the $\theta^i{}_\alpha$'s; and (ii) ${}^t m_{\alpha\beta}\, m_{\alpha\beta} = \mathbf{1}\cdot f'_{\alpha\beta}$, where $m_{\alpha\beta}$ are the transition functions of $\mathscr{E}$. Thus $m_{\alpha\beta} = {}^t m_{\alpha\beta}^{-1} f'_{\alpha\beta}$, i.e., $\mathscr{E} \simeq \mathscr{E}^* \otimes \omega$. ■

We want to remark at this point on the power of superconformal structures. Indeed, a generic supercurve of dimension $1|2$ is by no means split, as opposed to the trivial $1|1$ case. Nevertheless, susy$_2$ curves are split, a peculiarity that does not survive to higher supersymmetric extensions.

According to the physical literature, a susy$_2$ curve is called *twisted* whenever the $O(2)$ symmetry of the (anti) commutation relations for the local supersymmetry generators D^i_α cannot be reduced to an SO(2) symmetry.[4] This is related to the vanishing of a class in $H^2(C, \mathbf{Z}_2)$ obtained by taking the determinant of the transition functions for the locally free sheaf $\mathscr{E}$. Namely, since any rank two locally free sheaf can be represented as the extension of an invertible sheaf $\mathscr{L}_1$ by another $\mathscr{L}_2$ fitting the exact sequence $0 \to \mathscr{L}_2 \to \mathscr{E} \to \mathscr{L}_1 \to 0$, we have $\det \mathscr{E} = \mathscr{L}_1 \otimes \mathscr{L}_2$. However, Serre self-duality implies $\det \mathscr{E} = \omega \otimes \mathscr{N}_2$, where $\mathscr{N}_2$ is a point of order two on the Jacobian of C. Then a susy$_2$ curve is untwisted whenever $\mathscr{N}_2$ is trivial.

From the holomorphic point of view, Serre self-dual rank two locally free sheaves are quite simple objects.

Lemma 1: The characteristic sheaf $\mathscr{E}$ of untwisted susy$_2$ curves decomposes as the direct sum $\mathscr{E} = \mathscr{L}_1 \oplus \mathscr{L}_2$, with $\mathscr{L}_1 \otimes \mathscr{L}_2 \simeq \omega$.

Proof: In a superconformal gauge the transition functions $\mu_{\alpha\beta}(z_\beta)^i{}_j$ of $\mathscr{E}$ satisfy ${}^t m_{\alpha\beta} \cdot m_{\alpha\beta} = f'_{\alpha\beta}\cdot\mathbf{1}$ and hence can be given the form

$$m_{\alpha\beta} = \begin{pmatrix} a_{\alpha\beta} & b_{\alpha\beta} \\ -b_{\alpha\beta} & a_{\alpha\beta} \end{pmatrix},$$

with $a^2_{\alpha\beta} + b^2_{\alpha\beta} = f'_{\alpha\beta}$. A simple computation shows that there is a one-cochain λ_α with values in the sheaf of GL (2, C)-valued holomorphic functions that diagonalizes $m_{\alpha\beta}$, showing that actually, $\mathscr{E} = \mathscr{L}_1 \oplus \mathscr{L}_2$. Imposing the Serre self-duality condition in this gauge gives

$$\mathscr{L}_1 \oplus \mathscr{L}_2 \simeq \mathscr{L}_1^{-1} \otimes \omega \oplus \mathscr{L}_2^{-1} \otimes \omega.$$

This completes the proof since $\mathscr{L}_1 \oplus \mathscr{L}_2 \simeq \mathscr{L}'_1 \oplus \mathscr{L}'_2$ if and only if either $\mathscr{L}_1 \simeq \mathscr{L}'_1$ or $\mathscr{L}_1 \simeq \mathscr{L}'_2$. ■

Proposition 2: Any twisted Serre self-dual locally free sheaf $\mathscr{E}$ of rank two on C is holomorphically isomorphic to the direct sum of two different θ characteristics, i.e., $\mathscr{E} = \mathscr{L}_1 \oplus \mathscr{L}_2$, with $\mathscr{L}_i^2 = \omega$.

Proof: Recall that a rank two locally free sheaf $\mathscr{E}$ is called semistable if for any invertible subsheaf $\mathscr{L} \subset \mathscr{E}$,

$$c_1(\mathscr{L}) < c_1(\det \mathscr{E})/2:$$

it is stable if the above inequality holds in the strict sense. A classical result of the theory of locally free sheaves over algebraic curves[11] states that a stable locally free sheaf cannot be decomposable (i.e., it cannot be isomorphic to the direct sum of two invertible subsheaves).

We first prove that a twisted semistable Serre self-dual locally free sheaf is strictly semistable, i.e., it admits only degree $g-1$ invertible subsheaves. Notice that if $\mathscr{E}$ is untwisted, by Lemma 1 it is not stable. If $\mathscr{E}$ is twisted, there is a point $\mathscr{M}$ of order 4 on the Jacobian of C such that $\mathscr{E} \otimes \mathscr{M}$ is untwisted in the sense that $\det(\mathscr{E} \otimes \mathscr{M}) = \omega$. Since $\mathscr{E} \otimes \mathscr{M}$ is stable if and only if $\mathscr{E}$ is stable, we are again in the above situation.

Second, an unstable Serre self-dual locally free sheaf is strictly semistable as well. In fact, suppose that $\mathscr{E}$ is given as $0 \to \mathscr{L}_1 \to \mathscr{E} \to \mathscr{L}_2 \to 0$, with $c_1(\mathscr{L}_1) \geqslant g-1$. Serre dualizing, we obtain $0 \to \mathscr{L}_2^\vee \to \mathscr{E} \to \mathscr{L}_1^\vee \to 0$. Then, supposing $c_1(\mathscr{L}_1) > g-1$, Lemma 15 of Ref. 12 shows that $\mathscr{L}_1 \simeq \mathscr{L}_2^\vee$ and hence as $\det \mathscr{E} = \mathscr{L}_1 \otimes \mathscr{L}_2 = \omega$ we obtain a contradiction with the assumption of the twisting of $\mathscr{E}$.

We have only to discuss the case $0 \to \mathscr{L}_1 \to \mathscr{E} \to \mathscr{L}_2 \to 0$, with $c_1(\mathscr{L}_i) = g-1$, $\mathscr{L}_1 \otimes \mathscr{L}_2 \neq \omega$. If $\mathscr{E}$ is not decomposable, again Lemma 15 of Ref. 12 tells us that there would be a unique invertible subsheaf $\mathscr{L} \subset \mathscr{E}$ of degree $g-1$, contradicting the assumption that $\mathscr{L}_1 \not\simeq \mathscr{L}_2^\vee$. Finally, given $\mathscr{E} \simeq \mathscr{L}_1 \oplus \mathscr{L}_2$ the Serre self-duality condition implies that $\mathscr{L}_i^2 \simeq \omega$. ■

In summary, we have that superconformal structures force the characteristic sheaf $\mathscr{E}$ to be, in the twisted case, the direct sum of two nonisomorphic square roots of the canonical bundle. The untwisted case has a richer structure since here $\mathscr{E}$ decomposes as $\mathscr{L} \oplus \omega \otimes \mathscr{L}^{-1}$, $\mathscr{L} \in \operatorname{Pic} C$. As pointed out in Ref. 13, this fact has interesting consequences both from the mathematical and physical standpoints. We simply notice that to ensure semistability of the sheaf $\mathscr{E}$ also in the untwisted sector, one has to be restricted to the case $\deg \mathscr{L} = g-1$. Here a convenient and physically reasonable parametrization of $\mathscr{E}$ is $\mathscr{E} = \mathscr{L} \otimes \mathscr{N} \oplus \omega \otimes (\mathscr{L} \otimes \mathscr{N})^{-1}$, with $\mathscr{N} \in \operatorname{Pic}_0(C)$ and $\mathscr{L}$ a θ characteristic on C.

Actually, this is not the whole story since for susy$_2$ curves the above analysis is somewhat blind. Indeed, we have to work in a finer category than the holomorphic one because two susy$_2$ curves may very well be holomorphically equivalent, but by no means superconformally equivalent. This finer classification is entirely an outspring of physics and we wish to uncover it in full detail by studying deformation theory of susy$_2$ curves.

III. DEFORMATIONS OF SUSY$_2$ CURVES

The first step in studying moduli space of algebraic objects is to find their local structure, as given by the base spaces of versal deformations.

Definition 2: A deformation of a susy$_2$ curve C over a pointed superspace $(B, \{*\})$ is a family $\mathscr{C}\overset{\pi}{\to}B$ of susy$_2$ curves together with an isomorphism ψ of C with the "central fiber" $\pi^{-1}(\{*\})$ fitting the commutative diagram

$$\begin{tikzcd} C \arrow[r, hook] \arrow[d] & \mathscr{C} \arrow[d, "\pi"] \\ \{*\} \arrow[r, hook] & B \end{tikzcd}.$$

As usual, the starting point for setting up a deformation theory is to identify the sheaf of infinitesimal automorphisms of the object to be deformed. In our case this is the subsheaf $T_\pi^{\mathscr{D}}$ of the relative tangent sheaf whose elements are germs of vector fields along the fibers which preserve $\mathscr{D}$:

$$T_\pi^{\mathscr{D}} := \{X \in T_\pi \mid [D, X] \in \mathscr{D}\ \forall D \in \mathscr{D}\}.$$

In perfect analogy with the case of $N = 1$ susy curves we find the following lemma.

Lemma 2: There is an isomorphism $T_\pi^{\mathscr{D}} \simeq (T_\pi)_{\rm red} \otimes \mathscr{A}_C$ as sheaves of $\pi^{-1}(\mathscr{A}_B)$ modules.

Proof: The condition for X to belong to $T_\pi^{\mathscr{D}}$ reads as $[D^i, X] \in \mathscr{D}$, where D^i are generators of $\mathscr{D}$. Introducing the canonical coordinates (z, θ^i), so that $D^i = \partial/\partial\theta^i + \theta^i(\partial/\partial z)$, and setting $X = a\cdot\partial/\partial z + b_i\cdot D^i$, one has

$$[D^i, X] = D^i a\frac{\partial}{\partial z} - (-1)^{|X|}b_k\delta^{ki}\frac{\partial}{\partial z} + D^i b_k\cdot D^k.$$

Therefore, $X \in T_\pi^{\mathscr{D}}$ if and only if $b_i = (-1)^{|a|}D^i a$ and the isomorphism is given by $a\cdot\partial/\partial z \rightsquigarrow a(\partial/\partial z) + (-1)^{|a|}D^i a\cdot D^i$. ■

Thanks to this lemma we have, for $\mathscr{E}$ semistable, the following proposition.

Proposition 3: Versal deformations of susy$_2$ curves exist. The dimension of the base of such deformations is $3g-3+g-a|4g-4$, with $a = 0, 1$ in the untwisted and twisted cases, respectively.

Proof: From the Kodaira–Spencer deformation theory, we know that possible obstructions lie in the second cohomology group of the sheaf of infinitesimal automorphisms $T_\pi^{\mathscr{D}}$. By Lemma 2 one obtains

$$T_\pi^{\mathscr{D}} = \omega^{-1} \oplus \Pi(\omega^{-1}\otimes\mathscr{E}) \oplus \det\mathscr{E}\otimes\omega^{-1},$$

where the above sum is the direct sum of sheaves of $\mathscr{O}_C$ modules. (The parity change operator Π has, strictly speaking, no effective meaning in this context; we just use it as a parity bookkeeper.) Then $H^2(T_\pi^{\mathscr{D}}) = \{0\}$, showing the existence of versal deformations. The second part of Proposition 3 follows from Serre self-duality of $\mathscr{E}$ and Proposition 2. Indeed, $\dim H^1(\omega^{-1}\oplus\det\mathscr{E}\otimes\omega^{-1}) = \dim H^1(\omega^{-1}\oplus\mathscr{N})$, where $\mathscr{N} = \det\mathscr{E}\otimes\omega^{-1} = \mathscr{O}$ for untwisted susy$_2$ curves, while it is a point of order 2 in the Jacobian of C in the twisted case. As for the odd dimension, notice that $\dim H^1(\omega^{-1}\otimes(\mathscr{L}_1\oplus\mathscr{L}_2)) = \dim H^1(\mathscr{L}_1^{-1}) + \dim H^1(\mathscr{L}_2^{-1})$. ■

Remark: As for the computation of the odd dimension q of the would-be moduli space of susy$_2$ curves in the general untwisted case, one can argue as follows. Since $\mathscr{E} \simeq \mathscr{L} \oplus \omega\otimes\mathscr{L}^{-1}$, $H^1(C, \mathscr{E})$ is invariant under the Kummer map $\mathscr{L} \rightsquigarrow \omega\otimes\mathscr{L}^{-1}$. Hence one can be restricted to discussing the case $\deg\mathscr{L} \equiv d \geqslant g-1$ only. By the Riemann–Roch theorem one has (i) if $g-1 \leqslant d < 2g-2$, then $q = 4g-4$; (ii) if $2g-2 \leqslant d \leqslant 3g-3$ *and $\mathscr{L}$ is generic*, then $q = 4g-4$; (iii) if $3g-3 \leqslant d \leqslant 4g-4$ *and $\mathscr{L}$ is generic*, then $q = d+g-1$; and (iv) if $4g-4 < d$, then $q = d+g-1$. Notice that in cases (ii) and (iii) the odd dimension of "moduli space" jumps on analytic submanifolds of the reduced space, a fact that renders its structure quite subtle in the framework of Kostant–Leites supermanifold theory.

From a more computative point of view, one can consider infinitesimal deformations, i.e., deformations over the superspace $\hat{S} = (\{*\}; \mathbb{C}(t, \eta)/(t^2, t\zeta))$, as being given by deforming the clutching functions of the central fiber. From Sec. II, we learn that these are of the form

$$z_\alpha = f_{\alpha\beta}(z_\beta),$$
$$\theta_\beta^i = [m_{\alpha\beta}(z_\beta)]_j^i\,\theta_\beta^j,$$

where the matrix $\mu_{\alpha\beta}(z_\beta)^i{}_j$ is of the form

$$\mu_{\alpha\beta}(z_\beta)_j^i = \begin{pmatrix} g_{1\alpha\beta} & 0 \\ 0 & g_{2\alpha\beta} \end{pmatrix},$$

with either $g_{i_{\alpha\beta}}^2 = f'_{\alpha\beta}$ or $g_{1_{\alpha\beta}}\cdot g_{2_{\alpha\beta}} = f'_{\alpha\beta}$.

The most general deformation of such clutching functions, i.e., those generated by a vector field in the whole $\mathscr{T}_\pi$, over $\hat{S}$ is given by

$$z_\alpha = f_{\alpha\beta}(z_\beta) + t(b_{\alpha\beta}(z_\beta) + \tfrac{1}{2}g_{\alpha\beta}(z_\beta)\epsilon_{ij}\,\theta_\beta^i\,\theta_\beta^j)$$
$$+ \zeta\eta_{i_{\alpha\beta}}(z_\beta)\theta_\beta^i,$$
$$\theta_\beta^i = [m_{\alpha\beta}(z_\beta)]_j^i\,\theta_\beta^j + t\,[l_{\alpha\beta}(z_\beta)]_j^i\,\theta_\beta^j + \zeta(\psi_{\alpha\beta}^i(z_\beta)$$
$$+ \tfrac{1}{2}\nu_{\alpha\beta}^i\epsilon_{jk}\theta_\beta^j\,\theta_\beta^k).$$

Imposing the superconformal condition shows that $g_{\alpha\beta}(z_\beta) = 0$ (this fact can be also grasped by writing explicitly the superconformal vector fields which generate the deformations) and the only independent data are $b_{\alpha\beta}(z_\beta)$, $\psi_{\alpha\beta}^i(z_\beta)$, and $[l_{\alpha\beta}(z_{\alpha\beta})]_j^i$. The cocycle condition leads easily to the identification of $\{b_{\alpha\beta}(z_\beta)\cdot\partial/\partial z_\alpha\}$ as a one-cocycle with values in the relative tangent sheaf ω_π^{-1} and $\{\psi_{\alpha\beta}^i(z_\beta)\}$ as a one-cycle with values in $\mathscr{E}^*$.

As for the role of the matrix $[l_{\alpha\beta}(z_\beta)]_j^i$, one can argue as follows. Since the even and odd infinitesimal deformations give decoupled equations, one can be limited to discussing a deformation of the form

$$z_\alpha = f_{\alpha\beta}(z_\beta) + tb_{\alpha\beta}(z_\beta),$$
$$\theta_\beta^i = [m_{\alpha\beta}(z_\beta)]_j^i\cdot\{\delta_k^j + t\,[m^{-1}\cdot l_{\alpha\beta}(z_\beta)]_k^j\}\cdot\theta_\beta^k.$$

The superconformal condition translates into $O_{\alpha\beta} + {}^tO_{\alpha\beta} = (1/f'_{\alpha\beta})(\partial b_{\alpha\beta}/\partial z_\beta)\cdot\mathbf{1}$ for the matrix $O_{\alpha\beta} \equiv m_{\alpha\beta}^{-1}\cdot l_{\alpha\beta}$. Hence

$$O_{\alpha\beta} = \begin{pmatrix} \tau_{\alpha\beta} & \alpha_{\alpha\beta} \\ -\alpha_{\alpha\beta} & \tau_{\alpha\beta} \end{pmatrix}$$

and its only free part is the off diagonal

$$\tilde{O}_{\alpha\beta} = \begin{pmatrix} 0 & \alpha_{\alpha\beta} \\ -\alpha_{\alpha\beta} & 0 \end{pmatrix}.$$

This decomposition is obviously due to the fact that when deforming the underlying curve C according to $b_{\alpha\beta}$, line bundles on C are deformed as well. The cocyle condition for $O_{\alpha\beta}$ gives

$$m_{\alpha\beta}O_{\alpha\beta}m_{\beta\gamma}+m_{\alpha\beta}m_{\beta\gamma}O_{\beta\gamma}=m_{\alpha\gamma}O_{\alpha\gamma}-b_{\alpha\beta}\,m'_{\alpha\beta}\,m_{\beta\gamma}.$$

Looking once again at the off-diagonal part of the above equation one has (multiplying on the left by $m_{\alpha\gamma}^{-1}$)

$$m_{\beta\gamma}^{-1}\,\widetilde{O}_{\alpha\beta}m_{\beta\gamma}+\widetilde{O}_{\beta\gamma}=\widetilde{O}_{\alpha\gamma}.$$

A simple algebra shows that $\widetilde{O}_{\alpha\beta}m_{\beta\gamma}=f'_{\beta\gamma}/\det m_{\beta\gamma}\cdot\widetilde{O}_{\alpha\beta}$, yielding $(f'_{\beta\gamma}/\det m_{\beta\gamma})\widetilde{O}_{\alpha\beta}+O_{\beta\gamma}=\widetilde{O}_{\alpha\gamma}$. Then considering local generators $\{\varphi_\alpha\}$ of $\omega^{-1}\otimes\det\mathscr{E}$ one readily identifies the collection $\{\alpha_{\alpha\beta}\cdot\varphi_\beta\}$ as a one-cocycle with values in $\omega^{-1}\otimes\det\mathscr{E}$.

In summary, a complete infinitesimal deformation of a susy$_2$ curve consists of a deformation of the underlying algebraic curve and the couple of line bundles that define the "single" object plus the deformation specified by $l_{\alpha\beta}$. Since the latter is completely qualified by an element in $H^1(C,\omega^{-1}\otimes\det\mathscr{E})$ we find complete agreement with the results of Proposition 3. As a final remark, we notice that this latter space, which can also be thought of as the space of superconformally nonequivalent susy$_2$ structures on a fixed curve, coincides with the space of holomorphically nonequivalent extensions of a θ characteristic $\mathscr{L}_1$ by another one $\mathscr{L}_2$.

IV. THE REDUCED MODULI SPACE OF UNTWISTED SUSY$_2$ CURVES

We can now give a detailed description of the reduced moduli spaces of susy$_2$ curves, which turns out to be complete in the untwisted case. According to Proposition 2, holomorphic isomorphism classes of twisted susy$_2$ curves are in one-to-one correspondence with isomorphism classes of couples $(C,\mathscr{L}_{12})$, where $\mathscr{L}_{12}$ is an unordered couple of nonequivalent θ characteristics. This sits inside the second symmetric power $\Sigma^{(2)}$ of the spin covering $\Sigma\rightarrow\mathscr{M}$ of the moduli space (at some fixed genus), a space that has a nice mathematical status.[14] Besides, according to Lemma 1, the reduced moduli space of untwisted susy$_2$ curves with a semistable characteristic sheaf $\mathscr{E}$ can be identified with the universal degree $g-1$ Picard variety $\mathrm{Pic}_{g-1}\rightarrow\mathscr{M}_g$ over the moduli variety of genus g algebraic curves, modulo the Kummer map $\mathscr{L}\rightsquigarrow\omega\otimes\mathscr{L}^{-1}$.

We next want to parametrize superconformal structures on a fixed curve C and a couple $\mathscr{L}_1\oplus\mathscr{L}_2$ of invertible sheaves fitting a Serre self-dual rank two locally free sheaf. The basic observation here is that the sheaf $\mathscr{E}$ should be regarded not merely as a holomorphic sheaf because the superconformal structure amounts to saying that it is the sheaf of sections of a vector bundle E with, as its structure group, the conformal group

$$G=\{m\in\mathrm{GL}(2,\mathbb{C})\,|\,{}^t m\cdot m=\lambda\,\mathbf{1}\}\equiv G_0\cup\eta\cdot G_0,$$

where G_0 is the identity component and

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

The map $\varphi: G\rightarrow\mathbb{C}^*$ given by $\varphi(m)={}^t m\cdot m$ gives rise to the exact diagram of complex groups:

$$\begin{tikzcd}
 & 1 \arrow[d] & 1 \arrow[d] & 1 \arrow[d] & \\
1 \arrow[r] & \mathrm{SO}(2) \arrow[r] \arrow[d] & G_0 \arrow[r, "\varphi"] \arrow[d] & \mathbb{C}^* \arrow[r] \arrow[d] & 1 \\
1 \arrow[r] & \mathrm{O}(2) \arrow[r] \arrow[d] & G \arrow[r, "\varphi"] \arrow[d] & \mathbb{C}^* \arrow[r] \arrow[d] & 1 \\
1 \arrow[r] & \mathbb{Z}_2 \arrow[r] \arrow[d] & \mathbb{Z}_2 \arrow[r] \arrow[d] & 1 & \\
 & 1 & 1 & &
\end{tikzcd}\tag{1}$$

Notice that the first row is an exact sequence of central subgroups of the groups in the second row. This is vital at the level of exact sequences of sheaves of germs of group-valued functions associated to the above diagram. Indeed, pushing the induced cohomology sequences as far as possible, we obtain an exact diagram:

$$\begin{tikzcd}
 & 1 \arrow[d] & 1 \arrow[d] & 1 \arrow[d] & \\
1 \arrow[r] & H^1(C,\mathscr{S}\mathscr{O}_2) \arrow[r] \arrow[d] & H^1(C,\mathscr{G}_0) \arrow[r] \arrow[d] & H^1(C,\mathscr{O}^*) \arrow[r] \arrow[d] & 1 \\
1 \arrow[r] & H^1(C,\mathscr{O}_2) \arrow[r] \arrow[d] & H^1(C,\mathscr{G}) \arrow[r] \arrow[d] & H^1(C,\mathscr{O}^*) \arrow[d] & \\
1 \arrow[r] & H^1(C,\mathbb{Z}_2) \arrow[r] \arrow[d] & H^1(C,\mathbb{Z}_2) \arrow[r] \arrow[d] & 1 & \\
 & 1 & 1 & &
\end{tikzcd}\quad.\tag{2}$$

Here we used some results of non-Abelian sheaf cohomology and the following lemma.

Lemma 3: The cohomology groups $H^*(C,\mathscr{O}^*)$ and $H^*(C,\mathscr{S}\mathscr{O}_2)$ coincide.

Proof: The exact sheaf sequence

$$0\rightarrow\mathbb{Z}\rightarrow\mathscr{O}\xrightarrow{m}\mathscr{S}\mathscr{O}_2\rightarrow 1,$$

where the map m is defined by

$$m(f)=\begin{pmatrix}\cos(2\pi i f)&\sin(2\pi i f)\\-\sin(2\pi i f)&\cos(2\pi i f)\end{pmatrix},\quad\forall f\in\Gamma(U,\mathscr{O}),$$

fits together with the standard exponential sequence into the commutative diagram of sheaves (of Abelian groups)

$$\begin{tikzcd}
0 \arrow[r] & \mathbb{Z} \arrow[r] \arrow[d, "\mathrm{id}"] & \mathcal{O} \arrow[r, "\exp"] \arrow[d, "\mathrm{id}"] & \mathcal{O}^* \arrow[r] \arrow[d, "m^*"] & 1 \\
0 \arrow[r] & \mathbb{Z} \arrow[r] & \mathcal{O} \arrow[r, "m"] & \mathcal{S}\mathcal{O}_2 \arrow[r] & 1
\end{tikzcd}, \qquad (3)$$

where

$$m^*(\psi) = \begin{pmatrix} (\psi+\psi^{-1})/2 & (\psi-\psi^{-1})/2i \\ -(\psi-\psi^{-1})/2i & (\psi+\psi^{-1})/2 \end{pmatrix}, \quad \forall \psi \in \Gamma(U,\mathcal{O}^*).$$

This gives rise to a long commutative sequence of cohomology groups, proving Lemma 3. ■

Remark: Notice that the above sequence shows that the cohomology group $H^1(C,\mathcal{S}\mathcal{O}_2)$ is isomorphic to the group Pic C of invertible sheaves on C.

The basic fact for our concern is the following lemma.

Lemma 4: The action of $H_1(C,\mathcal{S}\mathcal{O}_2)$ is transitive and free on the fiber of the bundle $H^1(C,\mathcal{G}_0) \to H^1(C,\mathcal{O}^*)$ over each class $\tau \in H^1(C,\mathcal{O}^*)$. The same is true for the action of $H^1(C,\mathcal{G}_0)$ on the fiber of $H^1(C,\mathcal{G}) \to H^1(C,\mathbb{Z}_2)$ over $\tau' \in H^1(C,\mathbb{Z}_2)$.

Proof: Since $\mathcal{S}\mathcal{O}_2 \hookrightarrow \mathcal{G}_0$ and $\mathcal{G}_0 \hookrightarrow \mathcal{G}$ are central and Abelian, we can apply a (simplified) argument of non-Abelian sheaf cohomology (see, e.g., Lemma 2.4 of Ref. 10) to obtain the proof. This runs as follows. Given an exact sequence of sheaves of groups $0 \to \mathcal{P} \to \mathcal{Q} \to \mathcal{R} \to 0$ in which $\mathcal{P}$ is central and Abelian and $\mathcal{R}$ is Abelian, one has the following results.

(i) There is a connecting map $H^1(\mathcal{R}) \overset{\delta_1}{\to} H^2(\mathcal{P})$, so that the sequence

$$H^1(\mathcal{Q}) \to H^1(\mathcal{R}) \overset{\delta_1}{\to} H^2(\mathcal{P})$$

is exact.

(ii) Whenever $\tau \in \operatorname{Ker} \delta_1$, $H^1(\mathcal{P})$ acts transitively on the fiber of $H^1(\mathcal{Q})$ over τ, with the kernel given by the image of $H^0(\mathcal{R}) \overset{\delta_0}{\to} H^1(\mathcal{Q})$. In our case Lemma 4 follows from the fact that $H^2(C,\mathcal{S}\mathcal{O}_2) \simeq H^2(C,\mathcal{O}^*) = \{0\}$ and the observation that elements in $H^0(C,\mathbb{Z}_2)$ $((H^0(C,\mathcal{O}^*))$ are mapped into locally constant matrices by the connecting homomorphisms δ_0 and thus are clearly trivial cocycles. ■

Using Lemma 4, we can give the following description of the (reduced) moduli space of untwisted susy$_2$ structures over a fixed curve C.

Proposition 4: Nonequivalent untwisted susy$_2$ structures on a fixed (smooth) algebraic curve C are parametrized by the fiber of $H^1(C,\mathcal{S}\mathcal{O}_2)$ in $H^1(C,\mathcal{G}_0)$ over $[\omega] \in H^1(C,\mathcal{O}^*)$.

Proof: This proof follows at once by noticing that the map $H^1(C,\mathcal{G}) \to H^1(C,\mathcal{O}^*)$ in diagram (2) is surjective and the map $H^1(C,\mathcal{O}_2) \to H^1(C,\mathcal{G})$ is injective. The last assertion follows by applying Lemma 4. ■

Remark: While in the general theory of supermanifolds the first infinitesimal neighborhood of M_{red} is an "ordinary" vector bundle and thus its vertical automorphisms are automorphisms of the supermanifold structure, when considering supermanifolds with contact structure, which are the most relevant to physics (see, e.g., Ref. 2), extra structures must be taken into account. Thus the classification of $N=2$ superconformal structures over an algebraic curve C is quite different from the classification of rank two vector bundles over C, which corresponds, as is well known, to the classification of all *split* supermanifolds of odd dimension 2 over C.

V. CONCLUSIONS AND OUTLOOK

In this paper we have reconsidered some features of the geometry of $N=2$ superconformal field theories in a proper geometric framework. We showed that most of the definitions and properties of $N=1$ super Riemann surfaces carry over, with obvious modifications, to the $N=2$ case. In particular, we pointed out the relations of the theory of susy$_2$ curves with the theory of Serre self-dual rank two locally free sheaves over algebraic curves.

This approach gives a full proof of the results that are usually obtained in the physical literature by studying degrees of freedom and "gauge invariance" of the $N=2$ supersymmetric action in two dimensions (see, e.g., Ref. 15). Namely, some pecularities of $N>1$ supersymmetry, such as the existence of modular parameters for the $U(1)$ current mixing the gravitinos, have been given a sound geometrical picture. In addition, a complete description of the reduced moduli space of susy$_2$ curves in the untwisted sector was given.

A detailed study of the global aspects of $N=2$ supermoduli spaces, together with a setup of the theory of superconformal fields on susy$_2$ curves along the lines of Refs. 16 and 17, will be the subject of future work.

ACKNOWLEDGMENTS

We benefited from discussions with L. Girardello, A. Pasquinucci, and P. Teofilatto. We also thank M. Martellini for drawing our attention to this problem and A. S. Schwarz for useful remarks.

This work was partially supported by the national project "Geometria e Fisica" MPI.

[1] E. D'Hoker and D. H. Phong, Rev. Mod. Phys. **60**, 917 (1988).

[2] Yu. I. Manin, *Gauge Field Theory and Complex Geometry*, Grundlehren der Matematischen Wissenschaften 289 (Springer, Berlin, 1988).

[3] Yu. I. Manin, in *Arbeitstagung, Bonn 1984*, Lecture Notes in Mathematics, Vol. 1111, edited by F. Hirzebruch, J. Schwermer and S. Suter (Springer, Berlin, 1985), p. 59.

[4] J. D. Cohn, Nucl. Phys. B **284**, 349 (1987).

[5] E. Melzer, J. Math. Phys. **29**, 1555 (1988).

[6] D. Gepner, Nucl. Phys. B **296**, 757 (1988).

[7] P. Deligne, unpublished letter to Yu. I. Manin (1987).

[8] Yu. I. Manin, Funct. Anal. Appl. **20**, 244 (1987).

[9] D. Friedan, in *Unified String Theories*, edited by M. Green and D. Gross (World Scientific, Singapore, 1986).

[10] C. LeBrun and M. Rothstein, Commun. Math. Phys. **117**, 159 (1988).

[11] M. S. Narasihman and S. Ramanan, Ann. Math. **89**, 1201 (1969).

[12] R. C. Gunning, *Lectures on Vector Bundles Over Riemann Surfaces* (Princeton U.P., Princeton, 1967).

[13] S. N. Dolgikh, A. A. Rosly, and A. S. Schwarz, ICTP IC/89/210 preprint (1989).

[14] M. Cornalba, Universita' di Pavia preprint (1988).

[15] P. Di Vecchia, B. Duurhus, P. Oleson, and J. L. Petersen, Nucl. Phys. B **217**, 395 (1983).

[16] A. A. Rosly, A. S. Schwarz, and A. A. Voronov, Commun. Math. Phys. **119**, 129 (1988).

[17] S. B. Giddings and P. Nelson, Commun. Math. Phys. **116**, 607 (1988).

Finite-dimensional representations of the Lie superalgebra gl(2/2) in a gl(2)⊕gl(2) basis. II. Nontypical representations

Tchavdar D. Palev[a] and Nedjalka I. Stoilova[a]
Arnold-Sommerfeld Institute for Mathematical Physics, 3392 Clausthal-Zellerfeld, Federal Republic of Germany

(Received 25 July 1989; accepted for publication 25 October 1989)

All finite-dimensional irreducible representations of the general linear Lie superalgebra gl(2/2) are written down in matrix form. The basis within each representation space is chosen in such a way that it makes evident the decomposition of gl(2/2) into irreducible representations of its even subalgebra gl(2) ⊕ gl(2). Special attention is devoted to the analysis of all nontypical representations and some indecomposible representations.

I. INTRODUCTION

In Ref. 1 (hereafter referred to as I) the finite-dimensional irreducible representations of the general linear Lie superalgebra gl(2/2) were studied with an accent on the induced and the typical representations. In the present paper, we complete this investigation. Our final aim is to give explicit expressions for the transformations of all finite-dimensional irreducible modules (fidirmods) of gl(2/2) in a matrix form. In other words, we wish to introduce a basis within each (typical and nontypical) gl(2/2) fidirmod and to show how the basis transforms under the action of the algebra's generators.

The algebra gl(2/2) can be defined as the set of all squared four-dimensional complex matrices. As a convenient basis in it we choose the Weyl matrices e_{ij}, $i,j = 1,2,3,4$, where e_{ij} is a matrix with 1 on the ith row and the jth column and zero elsewhere. Assign to each index i a degree (i), which is zero for $i = 1,2$ and 1 for $i = 3,4$. Then e_{ij} is an even (resp. an odd) generator, if $(i) + (j)$ is an even (resp. an odd) number. The multiplication (= the supercommutator) between e_{ij} and e_{kl} is a commutator $e_{ij}e_{kl} - e_{kl}e_{ij}$, if at least one of both generators is even and it is an anticommutator $e_{ij}e_{kl} + e_{kl}e_{ij}$, if both generators are odd. The even subalgebra of gl(2/2) is isomorphic to the Lie algebra gl(2) ⊕ gl(2). For definiteness we set

$$\text{left: gl}(2) \equiv \text{gl}(2)_l = \text{lin.env.}\{e_{ij}\,|\,i,j = 1,2\}, \qquad (1)$$

$$\text{right: gl}(2) \equiv \text{gl}(2)_r = \text{lin.env.}\{e_{ij}\,|\,i,j = 3,4\}. \qquad (2)$$

An important subalgebra of gl(2/2), which is a 15-dimensional ideal in it, is the Lie superalgebra sl(2/2). The latter consists of all four-dimensional matrices, whose supertrace vanishes, i.e.,

$$\text{sl}(2/2) = \Big\{ a\,|\,a\in\text{gl}(2/2),\quad \text{str}(a) \equiv \sum_{i=1}^{4} (-1)^{(i)} a_{ii} = 0 \Big\}. \qquad (3)$$

The finite-dimensional irreducible modules of sl(2/2) [in the more general framework of the algebras sl(n/m)] were classified by Kac.[2] Each sl(2/2) fidirmod contains a highest weight vector. Its highest weight carries information about the fidirmod. In order to give more explicit description of the sl(2/2) fidirmods, introduce a basis in the Cartan subalgebra $H_{\rm sl}$ of sl(2/2) in the following way:

$$h_1 = e_{11} - e_{22}, \quad h_2 = e_{22} + e_{33}, \quad h_3 = e_{33} - e_{44} \qquad (4)$$

and denote by h^1, h^2, h^3 the conjugate basis in the dual to $H_{\rm sl}$ space $H^*_{\rm sl}$, i.e.,

$$h_i(h^j) = \delta_i^j.$$

Let Δ be a subset of $H^*_{\rm sl}$ defined as

$$\Delta = \{\alpha_1 h^1 + \alpha_2 h^2 + \alpha_3 h^3 | \alpha_1,\alpha_3 \in \mathbb{Z}_+, \alpha_2 \in \mathbb{C}\}. \qquad (5)$$

Proposition 1:[2] The vector $\Lambda \in H^*_{\rm sl}$ is a highest weight of a finite-dimensional irreducible sl(2/2) module, if and only if $\Lambda\in\Delta$.

Thus the sl(2/2) fidirmods can be labeled with three numbers $(\alpha_1,\alpha_2,\alpha_3)$, which satisfy the conditions, stated in (5); these numbers are the coordinates of the corresponding highest weight in the basis (4). Denote by $W(\alpha_1,\alpha_2,\alpha_3)$ a finite-dimensional irreducible sl(2/2) module with a highest weight

$$\Lambda = \alpha_1 h^1 + \alpha_2 h^2 + \alpha_3 h^3 \qquad (6)$$

and a highest weight vector x_Λ. The coordinates α_i of the highest weight Λ are the eigenvalue of h_i on x_Λ,

$$h_i x_\Lambda = \alpha_i x_\Lambda, \quad i = 1,2,3. \qquad (7)$$

In I, the finite-dimensional gl(2/2) modules, induced from the even subalgebra, have been considered. The construction goes as follows. Choose as an ordered basis in the Cartan subalgebra H of gl$(2)_l \oplus$ gl$(2)_r$ the generators e_{11}, e_{22}, e_{33}, e_{44}. Denote with

$$e^1,e^2,e^3,e^4, \quad e^i(e_{jj}) = \delta^i_j, \qquad (8)$$

the dual to its basis from H^* and let $V_0([m_{13},m_{23},m_{33},m_{43}])$ be a fidirmod of gl$(2)_l \oplus$ gl$(2)_r$ with a highest weight

$$\Lambda = m_{13}e^1 + m_{23}e^2 + m_{33}e^3 + m_{43}e^4. \qquad (9)$$

Extend $V_0([m_{13},m_{23},m_{33},m_{43}])$ to a module over the subalgebra

[a] On leave of absence from the Institute for Nuclear Research and Nuclear Energy, Blvd. Lenin 72, 1184 Sofia, Bulgaria.

$$P = \mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r \oplus P_+ ,$$
$$P_+ = \text{lin.env.}\{e_{13},e_{14},e_{23},e_{24}\} , \tag{10}$$

postulating that

$$P_+ V_0([m_{13},m_{23},m_{33},m_{43}]) = 0 .$$

The gl(2/2) module $W([m_{13},m_{23},m_{33},m_{43}])$, induced from $V_0([m_{13},m_{23},m_{33},m_{43}])$, is defined to be

$$W([m_{13},m_{23},m_{33},m_{43}])$$
$$= (U \otimes V_0([m_{13},m_{23},m_{33},m_{43}]))/I_0([m_{13},m_{23},m_{33},m_{43}]) ,$$

where U is the universal enveloping algebra of gl(2/2) and

$$I_0([m_{13},m_{23},m_{33},m_{43}])$$
$$= \text{lin.env.}\{up \otimes v - u \otimes pv \mid u \in U, p \in P \subset U,$$
$$v \in V_0([m_{13},m_{23},m_{33},m_{43}])\} .$$

Each induced module is either irreducible or indecomposible. It is important to point out that $W([m_{13},m_{23},m_{33},m_{43}])$ is gl(2/2) irreducible (resp. indecomposible), if and only if it is sl(2/2) irreducible (resp. indecomposible) [see Propositions 24 and 25]. Therefore, several of the definitions of the sl(2/2) modules $W([m_{13},m_{23},m_{33},m_{43}])$ can be generalized in a natural way to $W([m_{13},m_{23},m_{33},m_{43}])$, considered as gl(2/2) modules. In particular, we say that gl(2/2) module $W([m_{13},m_{23},m_{33},m_{43}])$ is typical (resp. nontypical), if considered as an sl(2/2) module it is typical (resp. nontypical). We recall that a given sl(2/2) module is typical[2] if it is isomorphic to an irreducible induced module. All other finite-dimensional irreducible sl(2/2) modules are called nontypical. It turns out that any nontypical gl(2/2) module [and hence nontypical sl(2/2) module] is a factor module of an induced module. More precisely, let $V([m_{13},m_{23},m_{33},m_{43}])$ be a nontypical gl(2/2) module with a highest weight (9). Then the induced module $W([m_{13},m_{23},m_{33},m_{43}])$ is indecomposible. It contains a maximal invariant submodule I so that up to an isomorphism

$$V([m_{13},m_{23},m_{33},m_{43}]) = W([m_{13},m_{23},m_{33},m_{43}])/I . \tag{11}$$

The following statement (I, Proposition 2) divides the induced modules into irreducible (and, hence, typical) modules and indecomposible modules.

Proposition 2: The induced module $W([m_{13},m_{23},m_{33},m_{43}])$ is gl(2/2) irreducible, if and only if

$$m_{i3} + m_{j3} - i - j + 5 \neq 0$$
$$\Leftrightarrow l_{i3} + l_{j3} + 3 \neq 0 \ \forall\, i = 1,2 \text{ and } j = 3,4 . \tag{12}$$

The authors of Ref. 1 have defined a basis within each module $W([m_{13},m_{23},m_{33},m_{43}])$, which was appropriate for a construction of both the typical and the nontypical modules. To this end they decomposed every induced module $W([m_{13},m_{23},m_{33},m_{43}])$ into a direct sum of irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules $V(i)$, $i = 1,2,\ldots$ and introduced a basis $\Gamma(i)$ in every $V(i)$. As a basis Γ in $W([m_{13},m_{23},m_{33},m_{43}])$, which was called a reduced basis, they took the union of all $\Gamma(1),\Gamma(2),\ldots$ and wrote down expressions for the transformations of the reduced basis under the action of the generators. Whenever the conditions (12) hold, the reduced basis is a basis in a typical gl(2/2) module. Therefore, the authors of I have found the transformations of all typical gl(2/2) modules in a matrix form.

In the present paper we solve the same problem for all nontypical modules. We hope this to be another small contribution, among the various other encouraging results,[3–28] toward the developing of a general representation theory of the basic Lie superalgebras [including the algebras gl(n/n), which are central extensions of basic Lie superalgebras]. We begin by recalling the notation, introduced in I, slightly modifying some of it. We write also the reduced basis and its transformations under the action of the generators in a somewhat more compact form.

II. INDUCED REPRESENTATIONS OF gl(2/2)

A. Some abbreviations and notation

In order to make the exposition self-consistent, we list here some of the abbreviations and notation that are used throughout the paper. Most of them are the same as in I.

LS, LS's—Lie superalgebra, Lie superalgebras,
fidirmod(s)—finite-dimensional irreducible module(s),
GZ basis—Gel'fand–Zetlin basis,
lin.env.$\{X\}$—the linear envelope of X,
$\mathbb{Z}$—all integers,
$\mathbb{Z}_+$—all non-negative integers,
$\mathbb{N}$—all positive integers,
$\mathbb{C}$—the complex numbers,
$[m] = [m_{13},m_{23},m_{33},m_{43}]$,
$l_{ij} = m_{ij} - i$, for $i = 1,2$ and $l_{ij} = m_{ij} - i + 2$, for $i = 3,4$,
$W([m])$—an induced gl(2/2) module,
I_k—the maximal invariant subspace in $W([m])$, corresponding to the class k, $k = 1,2,3,4,5$ [see (40)–(45)]; for $k = 1,2,3,4$ it is irreducible, whereas I_5 is indecomposible module,
$W_k([m]) = W([m])/I_k$—the nontypical module, corresponding to the class k, $k = 1,2,3,4,5$,
$\boxed{-a,-b,c,d}_{pq}$, $\boxed{-a,-b,c,d}_{s}$, $\boxed{-a,-b,c,d}_{s}$ —irreducible modules of $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ with a signature, $[m_{13} - a,m_{23} - b,m_{33} + c,m_{43} + d]$—see (47), (63), (64),

$$\langle i \rangle = \begin{cases} 0, & \text{if } i \text{ is an even number,} \\ 1, & \text{if } i \text{ is an odd number.} \end{cases} \tag{13}$$

By (m) we denote a table of 12 numbers, ordered as follows:

$$\begin{array}{cccc} m_{13}, & m_{23}, & m_{33}, & m_{43} \\ (m) \equiv m_{12}, & m_{22}, & m_{32}, & m_{42} \ . \\ m_{11}, & 0, & m_{31}, & 0 \end{array} \tag{14}$$

Then $(m)^{\pm i_1 j_1, \ldots, \pm i_k j_k}$ denotes a table, which is obtained from (m) by the replacements

$$m_{i_1 j_1} \to m_{i_1 j_1} \pm 1 ,$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \tag{15}$$
$$m_{i_k j_k} \to m_{i_k j_k} \pm 1 .$$

B. Transformations of the reduced basis

The induced gl(2/2) modules $W([m]) \equiv W([m_{13},m_{23},m_{33},m_{43}])$ are labeled with four in general complex numbers $m_{13},m_{23},m_{33},m_{43}$, which are the coordinates of the highest weight Λ in the basis (8), i.e., Eq. (9) holds. The coordinates $m_{13},m_{23},m_{33},m_{43}$ take all possible values, consistent with the conditions

$$m_{13},m_{23},m_{33},m_{43} \in \mathbb{C},\ m_{13}-m_{23}\in\mathbb{Z}_+,\ m_{33}-m_{43}\in\mathbb{Z}_+ . \tag{16}$$

Thus the values of $m_{13},m_{23},m_{33},m_{43}$ fix the induced representation space of the LS gl(2/2). The basis within $W([m])$ is given with the set of all possible patterns

$$[(m)]_{pq} \equiv \begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & m_{43}, \\ m_{12}, & m_{22}, & m_{32}, & m_{42}, \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{pq} , \tag{17}$$

which are compatible with the conditions

$$(1)\quad p,q\in\mathbb{Z}_+ ,\quad 2\geqslant p\geqslant q\geqslant 0 ; \tag{18}$$

(here are some new expressions for r, m_{12}, and m_{22})

$$\begin{aligned} (2)\quad & s=1,...,1+\min[p-q,m_{33}-m_{43}] , \\ & r=1,...,1+\min[\langle p\rangle+\langle q\rangle,m_{13}-m_{23}] ; \end{aligned} \tag{19}$$

$$\begin{aligned} (3)\quad & m_{12}=m_{13}-r-\theta(p-2)-\theta(q-2)+1 , \\ & m_{22}=m_{23}+r-\theta(p-1)-\theta(q-1)-1 , \\ & \qquad \text{where } \theta(0)=1 , \\ & m_{32}=m_{33}+p-s+1 , \\ & m_{42}=m_{43}+q+s-1 ; \end{aligned} \tag{20}$$

$$\begin{aligned} (4)\quad & m_{12}-m_{11}\in\mathbb{Z}_+ ,\quad m_{11}-m_{22}\in\mathbb{Z}_+ , \\ & m_{32}-m_{31}\in\mathbb{Z}_+ ,\quad m_{31}-m_{42}\in\mathbb{Z}_+ . \end{aligned} \tag{21}$$

The transformations of the reduced basis (17) under the action of the even generators read[1]:

$$\begin{aligned} & e_{11}[(m)]_{pq}=m_{11}[(m)]_{pq} , \\ & e_{22}[(m)]_{pq}=(m_{12}+m_{22}-m_{11})[(m)]_{pq} , \end{aligned} \tag{22}$$

$$\begin{aligned} & e_{33}[(m)]_{pq}=m_{31}[(m)]_{pq} , \\ & e_{44}[(m)]_{pq}=(m_{32}+m_{42}-m_{31})[(m)]_{pq} , \end{aligned} \tag{23}$$

$$e_{12}[(m)]_{pq}=|(l_{12}-l_{11})(l_{22}-l_{11})|^{1/2}[(m)^{11}]_{pq} , \tag{24}$$

$$\begin{aligned} e_{21}[(m)]_{pq}=&|(l_{12}-l_{11}+1)(l_{22}-l_{11}+1)|^{1/2} \\ &\times[(m)^{-11}]_{pq} , \end{aligned} \tag{25}$$

$$e_{34}[(m)]_{pq}=|(l_{32}-l_{31})(l_{42}-l_{31})|^{1/2}[(m)^{31}]_{pq} , \tag{26}$$

$$\begin{aligned} e_{43}[(m)]_{pq}=&|(l_{32}-l_{31}+1)(l_{42}-l_{31}+1)|^{1/2} \\ &\times[(m)^{-31}]_{pq} . \end{aligned} \tag{27}$$

The transformations of the reduced basis under the action of the odd generators e_{32} and e_{23} were given in I. Here, we write the result in a somewhat more compact form. Throughout the equations below $[(m)]_{pq}$ is a basis vector, written in the notations (17)–(21). In particular, each basis vector $[(m)]_{pq}$ depends also on the variables s and r, which take values as defined in Eqs. (19).

Transformations under the action of e_{32}:

$$e_{32}[(m)]_{00}=-\sum_{i=1}^{2}\sum_{j=3}^{4}\left|\frac{(l_{i3}-l_{11})(l_{7-j,3}-l_{31})}{(l_{13}-l_{23})(l_{33}-l_{43})}\right|^{1/2}[(m)^{-i2,j2,31}]_{10} , \tag{28}$$

$$\begin{aligned} e_{32}[(m)]_{10}=&\sum_{i=1}^{2}\sum_{j=3}^{4}\sum_{\substack{k=0,1\\ \langle r+i\rangle\leqslant k\leqslant\langle s+j\rangle}}(-1)^{(1-k)i+k(j+1)}\left|\frac{(l_{i2}-l_{11})(l_{7-j,2}-l_{31})}{(l_{12}-l_{22})(l_{32}-l_{42})}\right|^{1/2}\left|\frac{l_{3-i,3}-l_{i3}+2k-1}{(1+k)(l_{13}-l_{23})}\right|^{(1/2)\langle i+r+1\rangle} \\ &\times\left|\frac{l_{7-j,3}-l_{j3}+2k-1}{(2-k)(l_{33}-l_{43})}\right|^{(1/2)\langle s+j\rangle}[(m)^{-i2,j2,31}]_{2-k,k} , \end{aligned} \tag{29}$$

$$\begin{aligned} \underset{p+q=2}{e_{32}[(m)]_{pq}}=&-\sum_{i=\max(1,r-q)}^{\min(2,r-q+1)}\ \sum_{j=\max(3,5-q-s)}^{\min(4,6-q-s)}(-1)^{qi+(1-q)j} \\ &\times\left|\frac{(l_{i2}-l_{11})(l_{7-j,2}-l_{31})}{(l_{12}-l_{22})(l_{32}-l_{42})}\right|^{1/2}\left|\frac{(l_{i2}-l_{3-i,2}-\langle r\rangle+1)}{(2-\langle r\rangle)[l_{i3}-l_{3-i,3}-(-1)^r]}\right|^{q/2} \\ &\times\left|\frac{(l_{j2}-l_{7-j,2}+\langle s\rangle-1)}{(2-\langle s\rangle)[l_{j3}-l_{7-j,3}+(-1)^s]}\right|^{(1-q)/2}[(m)^{-i2,j2,31}]_{21} , \end{aligned} \tag{30}$$

$$e_{32}[(m)]_{21}=-(-1)^{r+s}\left|\frac{(l_{r3}-l_{11}-1)(l_{s+2,3}-l_{31}+2)}{(l_{13}-l_{23})(l_{33}-l_{43})}\right|^{1/2}[(m)^{-r2,j2,31}]_{22} ,\quad j=5-s , \tag{31}$$

$$e_{32}[(m)]_{22}=0 . \tag{32}$$

Transformation under the action of e_{23}:

$$e_{23}[(m)]_{22}=\sum_{i=1}^{2}\sum_{j=3}^{4}(-1)^{i+j}(l_{i3}+l_{j3}+3)\left|\frac{(l_{i3}-l_{11}-1)(l_{7-j,3}-l_{31}+3)}{(l_{13}-l_{23})(l_{33}-l_{43})}\right|^{1/2}[(m)^{i2,-j2,-31}]_{21} , \tag{33}$$

$$e_{23}[(m)]_{21} = -\sum_{i=1}^{2}\sum_{j=3}^{4}\sum_{\substack{k=0,1\\ \langle s+j\rangle<k<\langle r+i\rangle}} (-1)^{(1-k)i+kj}[l_{i3}+l_{j3}-(-1)^k\langle i+j+s+r\rangle+3]$$

$$\times\left|\frac{(l_{i2}-l_{11}+1)(l_{7-j,2}-l_{31}+1)}{(l_{12}-l_{22})(l_{32}-l_{42})}\right|^{1/2}\left|\frac{l_{r2}-l_{i2}+2k-2}{(2-k)(l_{13}-l_{23})}\right|^{(1/2)\langle i+r\rangle}$$

$$\times\left|\frac{l_{5-s,2}-l_{j2}+2k}{(1+k)(l_{33}-l_{43})}\right|^{(1/2)\langle s+j+1\rangle}[(m)^{i2,-j2,-31}]_{1+k,1-k}, \tag{34}$$

$$e_{23}[(m)]_{\substack{pq\\ p+q=2}} = \sum_{i=\max(1,q-r+2)}^{\min(2,q-r+3)}\ \sum_{j=\max(3,q+s+1)}^{\min(4,q+s+2)} (-1)^{(q-1)i+q(j+1)}$$

$$\times(l_{i3}+l_{j3}+\langle s\rangle-\langle r\rangle+3)\left|\frac{(l_{i2}-l_{11}+1)(l_{7-j,2}-l_{31}+1)}{(l_{12}-l_{22})(l_{32}-l_{42})}\right|^{1/2}$$

$$\times\left|\frac{l_{i2}-l_{3-i,2}+\langle r\rangle-1}{(2-\langle r\rangle)[l_{i3}-l_{3-i,3}+(-1)^r]}\right|^{q/2}\left|\frac{(l_{j2}-l_{7-j,2}-\langle s\rangle+1)}{(2-\langle s\rangle)[l_{j3}-l_{7-j,3}-(-1)^s]}\right|^{(1-q)/2}$$

$$\times[(m)^{i2,-j2,-31}]_{10}, \tag{35}$$

$$e_{23}[(m)]_{10} = -(l_{3-r,3}+l_{s+2,3}+3)\left|\frac{(l_{3-r,3}-l_{11})(l_{5-s,3}-l_{31}+1)}{(l_{13}-l_{23})(l_{33}-l_{43})}\right|^{1/2}[(m)^{i2,-j2,-31}]_{00},$$

$$i=3-r,\ j=s+2, \tag{36}$$

$$e_{23}[(m)]_{00}=0. \tag{37}$$

The expressions for the action of the other odd generators can be derived from Eqs. (25)–(37) and the supercommutation relations.

We improve a mistake that was done in I: The sign in front of the last term on the right-hand side of Eq. (3.77) has to be +.

III. NONTYPICAL REPRESENTATIONS OF gl(2/2)

In Ref. 1 it was proved (see Proposition 2) that the induced module $W([m])$ is irreducible, i.e., it is a fidirmod, if and only if

$$l_{i3}+l_{j3}+3\neq 0\ \forall\ i=1,2 \text{ and } j=3,4. \tag{38}$$

If for certain $i=1,2$ and $j=3,4$ the conditions (38) are not fulfilled, then the representation of gl(2/2) in $W([m])$ is indecomposible, i.e., the gl(2/2) module $W([m])$ contains a maximal invariant subspace I_k and in the same time there exists no compliment to I_k subspace, which is gl(2/2) invariant. The factor module $W_k([m]) = W([m])/I_k$ carries an irreducible representation of gl(2/2), which is called nontypical. We now proceed to determine all nontypical modules, to introduce a basis within each such module and to write down expressions for the transformation of the basis under the action of the generators. The relevance of this construction stems from the observation already stated above, namely that the set of all representations of gl(2/2) realized in all irreducible (= typical) induced modules $W([m])$ and in all factor modules $W_k([m])$ of the indecomposible induced modules exhaust the finite-dimensional irreducible representations of gl(2/2).

Taking into account that [see (16)] $m_{13}-m_{23}\in\mathbf{Z}_+$ and $m_{33}-m_{43}\in\mathbf{Z}_+$ or, which is equivalent, that

$$l_{13}-l_{23}\in\mathbf{N},\quad l_{33}-l_{43}\in\mathbf{N}, \tag{39}$$

we conclude that there exist five classes of indecomposible and hence also of nontypical gl(2/2) modules, namely:

$$\text{class 1}\quad \begin{aligned} &l_{13}+l_{43}+3=0\Leftrightarrow m_{13}+m_{43}=0,\\ &l_{23}+l_{33}+3\neq 0\Leftrightarrow m_{23}+m_{33}\neq 0;\end{aligned} \tag{40}$$

$$\text{class 2}\quad \begin{aligned} &l_{13}+l_{43}+3\neq 0\Leftrightarrow m_{13}+m_{43}\neq 0,\\ &l_{23}+l_{33}+3=0\Leftrightarrow m_{23}+m_{33}=0;\end{aligned} \tag{41}$$

$$\text{class 3}\quad l_{23}+l_{43}+3=0\Leftrightarrow m_{23}+m_{43}-1=0; \tag{42}$$

$$\text{class 4}\quad l_{13}+l_{33}+3=0\Leftrightarrow m_{13}+m_{33}+1=0; \tag{43}$$

$$\text{class 5}\quad \begin{aligned} &l_{13}+l_{43}+3=0\Leftrightarrow m_{13}+m_{43}=0,\\ &l_{23}+l_{33}+3=0\Leftrightarrow m_{23}+m_{33}=0.\end{aligned} \tag{44}$$

We denote by $I_k([m])$ the maximal (nontrivial) gl(2/2) invariant subspace in the indecomposible induced module $W([m])$, corresponding to the class k, $k=1,2,3,4,5$; then

$$W_k([m]) = W([m])/I_k([m]),\quad k=1,2,3,4,5 \tag{45}$$

is the corresponding nontypical module, carrying an irreducible (nontypical) representation of gl(2/2).

The following proposition, which will be often used, is actually one of the definitions of the irreducibility.

Proposition 3: Let V be a finite-dimensional $gl(2)_l\oplus gl(2)_r$ module and U_0 be the universal enveloping algebra of $gl(2)_l\oplus gl(2)_r$. Then V is an irreducible $gl(2)_l\oplus gl(2)_r$ module if and only if

$$U_0 x = V \,\forall\, 0 \neq x \in V. \tag{46}$$

Consider an induced gl(2/2) module $W([m]) \equiv W([m_{13},m_{23},m_{33},m_{43}])$ and let $V_{pq}([m_{12},m_{22},m_{32},m_{42}]) \subset W([m])$, $p,q \in \mathbb{Z}_+$, $2 \geqslant p \geqslant q \geqslant 0$, be the linear span of all vectors (17) with a fixed second row $[m_{12},m_{22},m_{32},m_{42}]$. From Eqs. (20) we conclude that

$$\begin{aligned}&[m_{12},m_{22},m_{32},m_{42}]\\ &\quad = [m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d] ,\end{aligned}$$

where a,b,c,d are non-negative integers whose exact values are completely determined from p,q,r,s [see (20)]. Therefore, if $V_{pq}([m_{12},m_{22},m_{32},m_{42}]) = V_{pq}([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])$, we set $V_{pq}([m_{12},m_{22},m_{32},m_{42}]) \equiv \boxed{-a,-b,c,d}_{pq} = \text{lin.env.}$,

$$\left\{ \begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & m_{43} \\ m_{12}, & m_{22}, & m_{32}, & m_{42} \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{pq} \middle|\, m_{12}-m_{11}, m_{11}-m_{22}, m_{32}-m_{31}, m_{31}-m_{42} \in \mathbb{Z}_+ \right\} . \tag{47}$$

By construction $V_{pq}([m_{12},m_{22},m_{32},m_{42}])$ is a $gl(2)_l \oplus gl(2)_r$ fidirmod with a signature $[m_{12},m_{22},m_{32},m_{42}]$. The labeling for the same subspaces in I was different, namely $V_i([m_{12},m_{22},m_{32},m_{42}])$, $i=1,...,6$. The one to one correspondence between pq and i reads:

$$00 \Leftrightarrow 1, \quad 10 \Leftrightarrow 2, \quad 20 \Leftrightarrow 3, \quad 11 \Leftrightarrow 4, \quad 21 \Leftrightarrow 5, \quad 22 \Leftrightarrow 6. \tag{48}$$

The gl(2/2) module $W([m])$ is a direct sum of all possible $gl(2)_l \oplus gl(2)_r$ modules $V_{pq}([m_{12},m_{22},m_{32},m_{42}])$ (which are no more than 16). More precisely,

$$\begin{aligned} W([m_{13},m_{23},m_{33},m_{43}]) &= V_{00}([m_{13},m_{23},m_{33},m_{43}])\\ &\oplus \sum_{i=0}^{\min(1,m_{13}-m_{23})} \sum_{j=0}^{\min(1,m_{33}-m_{43})} \oplus V_{10}([m_{13}-i,m_{23}+i-1,m_{33}-j+1,m_{43}+j])\\ &\oplus \sum_{j=0}^{\min(2,m_{33}-m_{43})} \oplus V_{20}([m_{13}-1,m_{23}-1,m_{33}-j+2,m_{43}+j])\\ &\oplus \sum_{i=0}^{\min(2,m_{13}-m_{23})} \oplus V_{11}([m_{13}-i,m_{23}+i-2,m_{33}+1,m_{43}+1])\\ &\oplus \sum_{i=0}^{\min(1,m_{13}-m_{23})} \sum_{j=0}^{\min(1,m_{33}-m_{43})} \oplus V_{21}([m_{13}-i-1,m_{23}+i-2,m_{33}-j+2,m_{43}+j+1])\\ &\oplus V_{22}([m_{13}-2,m_{23}-2,m_{33}+2,m_{43}+2]) . \end{aligned} \tag{49}$$

Proposition 4: Let I_k be the maximal gl(2/2) invariant subspace in the indecomposible induced module $W([m])$, corresponding to the class k [see (40)–(44)], $k=1,...,5$. Then

$$\begin{aligned}&V_{22}([m_{13}-2,m_{23}-2,m_{33}+2,m_{43}+2])\\ &\quad \equiv \boxed{-2,-2,2,2}_{22} \subset I_k .\end{aligned} \tag{50}$$

Proof: We carry out the proof using the induced basis [I,(2.29)] in $W([m])$. Suppose $0 \neq x \in I_k$. Then

$$\begin{aligned} x = \sum_{\theta_1,\theta_2,\theta_3,\theta_4=0,1} \sum_{(m)} \alpha(\theta_1,\theta_2,\theta_3,\theta_4;(m))\\ \times (e_{31})^{\theta_1}(e_{32})^{\theta_2}(e_{41})^{\theta_3}(e_{42})^{\theta_4} \otimes (m) ,\end{aligned} \tag{51}$$

where the second sum is over all basis vectors (m) in $V_0([m_{13},m_{23},m_{33},m_{43}])$ [see I, (2.29)]. Suppose that all coefficients $\alpha(\theta_1,\theta_2,\theta_3,\theta_4;(m))$ are equal to zero, if $\theta_1+\theta_2+\theta_3+\theta_4<k$, and that for certain $\theta_1^0,\theta_2^0,\theta_3^0,\theta_4^0$, $\theta_1^0+\theta_2^0+\theta_3^0+\theta_4^0=k$, and (m^0) $\alpha(\theta_1^0,\theta_2^0,\theta_3^0,\theta_4^0;(m^0)) \neq 0$. Then the first sum in (51) is over all $\theta_1,\theta_2,\theta_3,\theta_4$ such that $\theta_1+\theta_2+\theta_3+\theta_4 \geqslant k$. One easily derives from [I,(2.29),(3.20)] and (49) that

$$\begin{aligned} 0 &\neq (e_{31})^{1-\theta_1^0}(e_{32})^{1-\theta_2^0}(e_{41})^{1-\theta_3^0}(e_{42})^{1-\theta_4^0}x\\ &\equiv y \in e_{31}e_{32}e_{41}e_{42} \otimes V_0([m])\\ &\equiv V_{22}([m_{13}-2,m_{23}-2,m_{33}+2,m_{43}+2]) . \end{aligned} \tag{52}$$

Thus

$$0 \neq y \in I_k \cap V_{22}([m_{13}-2,m_{23}-2,m_{33}+2,m_{43}+2])$$

and, therefore, according to Proposition 3, (50) holds.

Proposition 5: Any maximal invariant subspace I_k has zero intersection with $V_{00}([m_{13},m_{23},m_{33},m_{43}]) \equiv \boxed{0,0,0,0}_{00}$.

Proof: If

$$0 \neq x \in I_k \cap V_{00}([m_{13},m_{23},m_{33},m_{43}]) , \tag{53}$$

then, according to Proposition 3, we would have

$$V_{00}([m_{13},m_{23},m_{33},m_{43}]) \subset I_k . \tag{54}$$

Since [see I, (3.19)] $V_{00}([m_{13},m_{23},m_{33},m_{43}]) = 1 \otimes V_0([m])$, then also

$$\left[\sum_{\theta_1,\theta_2,\theta_3,\theta_4=0,1}(e_{31})^{\theta_1}(e_{32})^{\theta_2}(e_{41})^{\theta_3}(e_{42})^{\theta_4}\right](1\otimes V_0([m]))$$
$$=\sum_{\theta_1,\theta_2,\theta_3,\theta_4=0,1}(e_{31})^{\theta_1}(e_{32})^{\theta_2}(e_{41})^{\theta_3}(e_{42})^{\theta_4}\otimes V_0([m])$$
$$=W([m])\subset I_k\,. \tag{55}$$

Thus, if (53) was true, we would have obtained $W([m])=I_k$. This, however, is impossible, since I_k is a proper subspace of $W([m])$. ■

Let $W([m])$ be an indecomposible induced gl(2/2) module, corresponding to one of the nontypical classes (40)–(44) with a maximal invariant subspace I_k. For every equivalence class

$$\xi_x\in W([m])/I_k\,,\quad x\in W([m])\,, \tag{56}$$

the mapping $\pi: g\rightarrow\pi(g)$, $g\in$gl(2/2), where

$$\pi(g)\xi_x=\xi_{gx}\,, \tag{57}$$

defines an irreducible nontypical representation of gl(2/2) in $W_k([m])\equiv W([m])/I_k$. Let $W_k^0([m])$ be a compliment to I_k subspace in $W([m])$,

$$W([m])=W_k^0([m])\oplus I_k\,. \tag{58}$$

Choose

$$e_1,e_2,\ldots,e_n \text{ to be a basis in } W_k^0([m])\,, \tag{59}$$

$$f_1,f_2,\ldots,f_m \text{ to be a basis in } I_k\,. \tag{60}$$

Then for any $g\in$gl(2/2)

$$ge_i=\sum_{j=1}^{n}A_{ji}e_j+\sum_{k=1}^{m}B_{ki}f_k\,. \tag{61}$$

The equivalence classes $\xi_{e_1},\xi_{e_2},\ldots,\xi_{e_n}$ constitute a basis in $W_k([m])$. From Eq. (57) one has that

$$\pi(g)\xi_{e_i}=\sum_{j=1}^{n}A_{ji}\xi_{e_j}\,. \tag{62}$$

Identifying $W_i^0([m])$ with $W_k([m])$, $W_k^0([m])\equiv W_k([m])$, through the natural isomorphism $x\Leftrightarrow\xi_x$ $\forall\, x\in W_i^0([m])$ and comparing (61) with (62), we come to the following conclusion.

Proposition 6: In order to obtain the transformation of the factor space $W_k([m])\equiv W([m])/I_k$ under the action of the gl(2/2) generators one simply has to replace in (61) all basis vectors $f_1,f_2,\ldots,f_m$ of the maximal invariant subspace I_k by zero.

Consider an indecomposible induced module $W([m])$ with a maximal invariant submodule I_k and let

$$V_s([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])\subset W([m])$$

be an irreducible gl$(2)_l\oplus$gl$(2)_r$ submodule with a signature $[m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d]$, labeled with an index s (s could be pq, upper or lower case index or any other index). Then we write

$$V_s([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])$$
$$\equiv\boxed{\boxed{-a,-b,c,d}}_s\quad, \tag{63}$$

if

$$V_s([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])\subset I_k$$

and

$$V_s([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])$$
$$\equiv\boxed{-a,-b,c,d}_s\quad, \tag{64}$$

if

$$V_s([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])\cap I_k=0\,,$$

i.e., if the linear spaces $V_s([m_{13}-a,m_{23}-b,m_{33}+c,m_{43}+d])$ and I_k are linearly independent. In particular [see (47)], if

$$\boxed{-a,-b,c,d}_{pq}\subset I_k \text{ we set } \boxed{-a,-b,c,d}_{pq}=\boxed{\boxed{-a,-b,c,d}}_{pq} \tag{65}$$

and if

$$\boxed{-a,-b,c,d}_{pq}\cap I_k=0 \text{ then } \boxed{-a,-b,c,d}_{pq}=\boxed{-a,-b,c,d}_{pq}\,. \tag{66}$$

Certainly it could well be that for some $\boxed{-a,-b,c,d}_{pq}$ neither (65) nor (66) holds. So far we have shown [Propositions 4 and 5] that

$$\boxed{-2,-2,2,2}_{22}=\boxed{\boxed{-2,-2,2,2}}_{22}\,,$$
$$\boxed{0,0,0,0}_{00}=\boxed{0,0,0,0}_{00}\,. \tag{67}$$

A. The class 1 nontypical representations

In this section, we consider the indecomposible induced modules $W([m])$, corresponding to the case

$$l_{13}+l_{43}+3=0\Leftrightarrow m_{13}+m_{43}=0\,,$$
$$l_{23}+l_{33}+3\neq0\Leftrightarrow m_{23}+m_{33}\neq0\,. \tag{68}$$

The induced modules from this class have signatures $[m]=[m_{13},m_{23},m_{33},-m_{13}]$, i.e.,

$$W([m])=W([m_{13},m_{23},m_{33},-m_{13}])\,. \tag{69}$$

In the cases

$$m_{13}>m_{23} \text{ and } m_{33}>-m_{13}$$
$$\Leftrightarrow l_{13}-l_{23}-1>0 \text{ and } l_{13}+l_{33}+2>0 \tag{70}$$

we define two new irreducible gl$(2)_l\oplus$gl$(2)_r$ modules. To this end we set

$$[m_{11},m_{31}]^1 \equiv \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}^1$$

$$= \left| \frac{l_{13}-l_{23}+1}{l_{13}-l_{23}-1} \right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{11}$$

$$- \left| \frac{l_{13}+l_{33}+4}{l_{13}+l_{33}+2} \right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}, \tag{71}$$

$$[m_{11},m_{31}]^1_{\text{inv}} = \left| \frac{l_{13}-l_{23}+1}{l_{13}-l_{23}-1} \right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{11}$$

$$+ \left| \frac{l_{13}+l_{33}+4}{l_{13}+l_{33}+2} \right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}. \tag{72}$$

Let

$$V^1 \equiv V^1([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1])$$
$$= \text{lin.env.}$$
$$\{[m_{11},m_{31}]^1 | m_{13}-m_{11}-1, m_{11}-m_{23}+1,$$
$$m_{33}-m_{31}+1, m_{31}+m_{13}-1 \in \mathbb{Z}_+\}, \tag{73}$$

$$V^1_{\text{inv}} \equiv V^1_{\text{inv}}([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1])$$
$$= \text{lin.env.}$$
$$\{[m_{11},m_{31}]^1_{\text{inv}} | m_{13}-m_{11}-1, m_{11}-m_{23}+1,$$
$$m_{33}-m_{31}+1, m_{31}+m_{13}-1 \in \mathbb{Z}_+\}. \tag{74}$$

Each of the spaces V^1 and V^1_{inv} is an irreducible $gl(2)_l \oplus gl(2)_r$ module with a signature $[m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1]$.

Proposition 7: Let $W([m_{13},m_{23},m_{33},-m_{13}])$ be a class 1 indecomposible induced gl(2/2) module with a maximal invariant subspace I_1 such that

$$m_{13} > m_{23} \text{ and } m_{33} > -m_{13}. \tag{75}$$

Then

$$V^1_{\text{inv}}([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1])$$
$$= \boxed{-1,-1,1,1}^1_{\text{inv}}, \tag{76}$$

i.e.,

$$V^1_{\text{inv}}([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1]) \subset I_1$$

and

$$V^1([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1])$$
$$= \boxed{-1,-1,1,1}^1, \tag{77}$$

i.e.,

$$V^1([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1]) \cap I_1 = 0.$$

The decomposition of $W([m_{13},m_{23},m_{33},-m_{13}])$ into a direct sum of irreducible $gl(2)_l \oplus gl(2)_r$ modules reads:

$$W([m_{13},m_{23},m_{33},-m_{13}])$$
$$= \boxed{0,0,0,0}_{00}$$
$$\oplus \boxed{-1,0,1,0}_{10} \oplus \boxed{0,-1,1,0}_{10} \oplus \boxed{-1,0,0,1}_{10} \oplus \boxed{0,-1,0,1}_{10}$$
$$\oplus \boxed{-1,-1,2,0}_{20} \oplus \boxed{-1,-1,0,2}_{20} \oplus \boxed{-2,0,1,1}_{11} \oplus \boxed{0,-2,1,1}_{11}$$
$$\oplus \boxed{-1,-1,1,1}^1_{\text{inv}} \oplus \boxed{-1,-1,1,1}^1$$
$$\oplus \boxed{-2,-1,2,1}_{21} \oplus \boxed{-1,-2,2,1}_{21} \oplus \boxed{-2,-1,1,2}_{21} \oplus \boxed{-1,-2,1,2}_{21}$$
$$\oplus \boxed{-2,-2,2,2}_{22}, \tag{78}$$

where

$$\boxed{-1,-1,0,2}_{20} = 0,$$
$$\text{if } m_{13} > m_{23}+1 \text{ and } m_{33} = -m_{13}+1, \tag{79}$$

$$\boxed{-2,0,1,1}_{11} = 0,$$
$$\text{if } m_{13} = m_{23}+1 \text{ and } m_{33} > -m_{13}+1. \tag{80}$$

The maximal invariant subspace I_1 is an irreducible gl(2/2) module with a signature $[m_{13}-1,m_{23},m_{33},-m_{13}+1]$.

The proof of this proposition, which is of a rather technical nature, is given in the Appendix. From (63) and (64) it follows immediately that I_1 [resp. its compliment subspace $W_1([m])$] is given with the sum of all $\boxed{}$ terms (resp. of all $\boxed{}$ terms) in the decomposition (78).

In order to write the transformations of the nontypical modules $W_1([m])$ under the action of the generators of gl(2/2) one has (1) to express

$$\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{11}$$

and (81)

$$\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}$$

in terms of $[m_{11},m_{31}]^1_{\rm inv}$ and $[m_{11},m_{31}]^1$, (2) to insert everywhere in (22)–(37) $m_{43}=-m_{13}$, and (3) to apply Proposition 6, namely to replace all basis vectors from the maximal invariant subspace by zero. The action of the even generators on all nonzero vectors $[(m)]_{pq}$ and on $[m_{11},m_{31}]^1$ is given with the same relations (22)–(27).

Transformations under the action of e_{32}:

$$e_{32}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13}, \\ m_{13}, & m_{23}, & m_{33}, & -m_{13}, \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{00} = -\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23}, & m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{13} \\ m_{11} & ,\ 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10}$$

$$-\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{13} \\ m_{11}, & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10}$$

$$-\left|\frac{(l_{23}-l_{11})(l_{33}-l_{31})}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}, & m_{23}-1, & m_{33} & , \quad -m_{13}+1 \\ m_{11}, & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10}, \tag{82}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23}, & m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{13} \\ m_{11} & ,\ 0\ , & m_{31} & , \quad 0 \end{bmatrix}_{10}$$

$$=\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{13} \\ m_{11} & ,\ 0\ , & m_{31}+1, & 0 \end{bmatrix}_{20}$$

$$-\left|\frac{(l_{23}-l_{11})(l_{33}-l_{31}+1)(l_{13}+l_{33}+3)}{2(l_{13}-l_{23})(l_{13}+l_{33}+4)(l_{13}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & ,\ 0\ , & m_{31}+1, & 0 \end{bmatrix}^{1}, \tag{83}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{13} \\ m_{11}, & 0\ , & m_{31} & ,\ 0 \end{bmatrix}_{10}$$

$$=-\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{13} \\ m_{11} & ,\ 0\ , & m_{31}+1, & 0 \end{bmatrix}_{20}$$

$$-\left|\frac{(l_{23}-l_{11}-1)(l_{33}-l_{31}+1)}{(l_{13}-l_{23}+1)(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}, & m_{23}-2, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{11}$$

$$+\frac{(l_{23}+l_{33}+3)}{(l_{13}+l_{33}+4)(l_{13}-l_{23}+1)}\left|\frac{(l_{13}-l_{11})(l_{33}-l_{31}+1)}{2(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & ,\ 0\ , & m_{31}+1, & 0 \end{bmatrix}^{1}, \tag{84}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23} & , m_{33}, & -m_{13} \\ m_{13}, & m_{23}-1, & m_{33}, & -m_{13}+1 \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{10}$$

$$=\left|\frac{(l_{23}-l_{11}-1)(l_{13}+l_{31}+2)}{(l_{13}-l_{23}+1)(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}, & m_{23}-2, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{11}$$

$$+\frac{1}{l_{13}-l_{23}+1}\left|\frac{(l_{13}-l_{11})(l_{13}-l_{23})(l_{13}+l_{31}+2)}{2(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , \quad -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & ,\ 0\ , & m_{31}+1, & 0 \end{bmatrix}^{1}, \tag{85}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{13} \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{20}$$

$$= -\left|\frac{(l_{23}-l_{11}-1)(l_{33}-l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-2, & m_{33}+2, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{21}, \tag{86}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}^{1}$$

$$= -\left|\frac{2(l_{23}-l_{11}-1)(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-2, & m_{33}+2, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{21}, \tag{87}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}, & m_{23}-2, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{11}$$

$$= \left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23}+1)(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-2, & m_{33}+2, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{21}, \tag{88}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-2, & m_{33}+2, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{21} = 0. \tag{89}$$

Transformations under the action of e_{23}:

$$e_{23}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-2, & m_{33}+2, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{21}$$

$$= \left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+1)(l_{13}+l_{33}+3)}{l_{13}-l_{23}+1}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}, & m_{23}-2, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}-1, & 0 \end{bmatrix}_{11}$$

$$+ \left|\frac{(l_{23}-l_{11}-1)(l_{33}-l_{31}+3)(l_{13}-l_{23})}{l_{13}+l_{33}+4}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{13} \\ m_{11}, & 0, & m_{31}-1, & 0 \end{bmatrix}_{20}$$

$$- \frac{(l_{23}+l_{33}+3)|2(l_{23}-l_{11}-1)(l_{13}-l_{23})(l_{13}+l_{31}+1)(l_{13}+l_{33}+3)|^{1/2}}{2(l_{13}-l_{23}+1)(l_{13}+l_{33}+4)}$$

$$\times\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}-1, & 0 \end{bmatrix}^{1}, \tag{90}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}, & m_{23}-2, & m_{33}+1, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{11}$$

$$= (l_{13}-l_{23})\left|\frac{(l_{23}-l_{11}-1)(l_{33}-l_{31}+2)}{(l_{13}-l_{23}+1)(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{13} \\ m_{11}, & 0, & m_{31}-1, & 0 \end{bmatrix}_{10}$$

$$+ (l_{23}+l_{33}+3)\left|\frac{(l_{23}-l_{11}-1)(l_{13}+l_{31}+1)}{(l_{13}-l_{23}+1)(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}, & m_{23}-1, & m_{33}, & -m_{13}+1 \\ m_{11}, & 0, & m_{31}-1, & 0 \end{bmatrix}_{10}, \tag{91}$$

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}^{1}$$

$$= - \left|\frac{2(l_{13}-l_{11})(l_{33}-l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}-1, & & m_{33}+1, & & -m_{13} \\ m_{11}, & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{10}$$

$$+ (l_{13}-l_{23}+1)\left|\frac{2(l_{23}-l_{11})(l_{33}-l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-1 & & m_{23}, & m_{33}+1, & & -m_{13} \\ m_{11} & , & 0 \;, & m_{31}-1, & & 0 \end{bmatrix}_{10}$$

$$+ (l_{13}+l_{33}+4)\left|\frac{2(l_{13}-l_{11})(l_{13}+l_{31}+1)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}-1, & & m_{33} & , & -m_{13}+1 \\ m_{11}, & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{10}, \tag{92}$$

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+2, & & -m_{13} \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}$$

$$= - \; (l_{13}+l_{33}+3)\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}-1, & & m_{33}+1, & & -m_{13} \\ m_{11}, & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{10}$$

$$+ (l_{23}+l_{33}+3)\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}, & m_{33}+1, & & -m_{13} \\ m_{11} & & 0 \;, & m_{31}-1, & & 0 \end{bmatrix}_{10}, \tag{93}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23} & , & m_{33}, & -m_{13} \\ m_{13}, & m_{23}-1, & & m_{33}, & -m_{13}+1 \\ m_{11}, & 0 & , & m_{31}, & 0 \end{bmatrix}_{10}$$

$$= \left|\frac{(l_{13}-l_{23})(l_{23}-l_{11})(l_{33}-l_{31}+1)}{l_{13}+l_{33}+3}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}, & m_{33} & , & -m_{13} \\ m_{11}, & 0 \;, & m_{31}-1, & & 0 \end{bmatrix}_{00}, \tag{94}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}-1, & & m_{33}+1, & & -m_{13} \\ m_{11}, & 0 & , & m_{31} & , & 0 \end{bmatrix}_{10}$$

$$= - \,(l_{23}+l_{33}+3)\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}, & m_{33} & , & -m_{13} \\ m_{11}, & 0 \;, & m_{31}-1, & & 0 \end{bmatrix}_{00}, \tag{95}$$

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}, & m_{33}+1, & & -m_{13} \\ m_{11} & , & 0 \;, & m_{31} & , & 0 \end{bmatrix}_{10}$$

$$= - \left|\frac{(l_{13}-l_{11})(l_{13}+l_{33}+3)(l_{13}+l_{31}+2)}{l_{13}-l_{23}}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}, & m_{23}, & m_{33} & , & -m_{13} \\ m_{11}, & 0 \;, & m_{31}-1, & & 0 \end{bmatrix}_{00}, \tag{96}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{13}, & m_{23}, & m_{33}, & -m_{13} \\ m_{11}, & 0 \;, & m_{31}, & 0 \end{bmatrix}_{00} = 0. \tag{97}$$

Proposition 8: Let $W([m_{13},m_{23},m_{33},-m_{13}])$ be a class 1 indecomposible induced gl(2/2) module, corresponding to a signature $[m_{13},m_{23},m_{33},-m_{13}]$, such that

$$m_{13} = m_{23} \text{ and } m_{33} > -m_{13}\,. \tag{98}$$

The structure of $W([m_{13},m_{13},m_{33},-m_{13}])$ with respect to $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reads:

$$
\begin{aligned}
&W([m_{13},m_{13},m_{33},-m_{13}])\\
&\quad=\boxed{0,0,0,0}_{00}\\
&\qquad\oplus\boxed{0,-1,1,0}_{10}\quad\oplus\boxed{0,-1,0,1}_{10}\\
&\qquad\oplus\boxed{-1,-1,2,0}_{20}\quad\oplus\boxed{-1,-1,1,1}_{20}\quad\oplus\boxed{-1,-1,0,2}_{20}\\
&\qquad\oplus\boxed{0,-2,1,1}_{11}\quad\oplus\boxed{-1,-2,2,1}_{21}\quad\oplus\boxed{-1,-2,1,2}_{21}\\
&\qquad\oplus\boxed{-2,-2,2,2}_{22}\quad ,
\end{aligned}
\tag{99}
$$

where

$$\boxed{-1,-1,0,2}_{20}=0\,,\ \text{if}\ m_{33}=-m_{13}+1\,. \tag{100}$$

The proofs of this proposition and of all others that follow are similar to the one given in the Appendix. Therefore, we will skip them. To obtain the transformation of the basis under the action of e_{23} and of e_{32} one has to replace in Eqs. (28)–(37) [or in I, Eqs. (3.62)–(3.93)] all basis vectors of the maximal invariant subspace

$$I_1=\boxed{-1,-1,0,2}_{20}\oplus\ \boxed{-1,-2,1,2}_{21}\ \oplus\boxed{-2,-2,2,2}_{22} \tag{101}$$

by zero, m_{23} by m_{13} and m_{43} by $-m_{13}$.

Proposition 9: Let $W([m_{13},m_{23},m_{33},-m_{13}])$ be a class 1 indecomposible induced gl(2/2) module with a signature such that

$$m_{13}>m_{23}\ \text{and}\ m_{33}=-m_{13}\,. \tag{102}$$

The structure of $W([m_{13},m_{23},-m_{13},-m_{13}])$ with respect to $\text{gl}(2)_l\oplus\text{gl}(2)_r$ reads:

$$
\begin{aligned}
&W([m_{13},m_{23},-m_{13},-m_{13}])\\
&\quad=\boxed{0,0,0,0}_{00}\\
&\qquad\oplus\boxed{0,-1,1,0}_{10}\quad\oplus\boxed{-1,0,1,0}_{10}\\
&\qquad\oplus\boxed{-1,-1,2,0}_{20}\quad\oplus\boxed{-1,-1,1,1}_{11}\quad\oplus\boxed{-2,0,1,1}_{11}\\
&\qquad\oplus\boxed{0,-2,1,1}_{11}\quad\oplus\boxed{-1,-2,2,1}_{21}\quad\oplus\boxed{-2,-1,2,1}_{21}\\
&\qquad\oplus\boxed{-2,-2,2,2}_{22}\quad ,
\end{aligned}
\tag{103}
$$

where

$$\boxed{-2,0,1,1}_{11}=0\,,\ \text{if}\ m_{13}=m_{23}+1\,. \tag{104}$$

To obtain the transformation of the basis under the action of e_{23} and of e_{32} one has to replace in Eqs. (28)–(37) [or in I, Eqs. (3.62)–(3.93)] all basis vectors of

$$I_1=\boxed{-2,0,1,1}_{11}\oplus\boxed{-2,-1,2,1}_{21}\oplus\boxed{-2,-2,2,2}_{22} \tag{105}$$

by zero, m_{33} by $-m_{13}$ and m_{43} by $-m_{13}$.

We have not considered here the cases

$$
\begin{gathered}
m_{13}=m_{23},\ m_{33}=-m_{13}\\
\text{and}\\
m_{13}=m_{23}+1,\ m_{33}=-m_{13}+1\,,
\end{gathered}
\tag{106}
$$

since they belong to the class 5 nontypical representations.

B. The class 2 nontypical representations

In this section, we consider the indecomposible induced modules $W([m])$, corresponding to the case

$$l_{13}+l_{43}+3\neq 0\Leftrightarrow m_{13}+m_{43}\neq 0\,, \tag{107}$$

$$l_{23}+l_{33}+3=0\Leftrightarrow m_{23}+m_{33}=0\,. \tag{108}$$

The induced modules from this class have signatures $[m]=[m_{13},m_{23},-m_{23},m_{43}]$, i.e.,

$$W([m])=W([m_{13},m_{23},-m_{23},m_{43}])\,. \tag{109}$$

Proposition 10: Consider a class 2 indecomposible induced gl(2/2) module (109), corresponding to a signature $[m_{13},m_{23},-m_{23},m_{43}]$, such that

$$m_{13}>m_{23}\ \text{and}\ -m_{23}>m_{43}\,. \tag{110}$$

Introduce a new basis in the $\text{gl}(2)_l\oplus\text{gl}(2)_r$ reducible module

$$\boxed{-1,-1,1,1}_{20}\quad\oplus\ \boxed{-1,-1,1,1}_{11} \tag{111}$$

setting

$$
\begin{aligned}
[m_{11},m_{31}]^2\equiv&\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1\\ m_{11} & , \quad 0\ , & m_{31} & ;\quad 0\end{bmatrix}^2\\
=&\left|\frac{l_{13}-l_{23}-1}{l_{13}-l_{23}+1}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1\\ m_{11} & , \quad 0\ , & m_{31} & , \quad 0\end{bmatrix}_{11}\\
&-\left|\frac{l_{23}+l_{43}+4}{l_{23}+l_{43}+2}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1\\ m_{11} & , \quad 0\ , & m_{31} & , \quad 0\end{bmatrix}_{20}\ ,
\end{aligned}
\tag{112}
$$

$$
\begin{aligned}
[m_{11},m_{31}]^2_{\text{inv}}=&\left|\frac{l_{13}-l_{23}-1}{l_{13}-l_{23}+1}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1\\ m_{11} & , \quad 0\ , & m_{31} & , \quad 0\end{bmatrix}_{11}\\
&+\left|\frac{l_{23}+l_{43}+4}{l_{23}+l_{43}+2}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1\\ m_{11} & , \quad 0\ , & m_{31} & , \quad 0\end{bmatrix}_{20}\ .
\end{aligned}
\tag{113}
$$

Construct two new irreducible $\text{gl}(2)_l\oplus\text{gl}(2)_r$ modules with a signature $[m_{13}-1,m_{23}-1,-m_{23}+1,m_{43}+1]$, setting

$$V^2 \equiv V^2([m_{13}-1,m_{23}-1,-m_{23}+1,m_{43}+1]) = \text{lin.env.}$$
$$\{[m_{11},m_{31}]^2 | m_{13}-m_{11}-1, m_{11}-m_{23}+1, -m_{23}-m_{31}+1, m_{31}-m_{43}-1 \in \mathbb{Z}_+\}, \qquad (114)$$
$$V^2_{\text{inv}} \equiv V^2_{\text{inv}}([m_{13}-1,m_{23}-1,-m_{23}+1,m_{43}+1]) = \text{lin.env.}$$
$$\{[m_{11},m_{31}]^2_{\text{inv}} | m_{13}-m_{11}-1, m_{11}-m_{23}+1, -m_{23}-m_{31}+1, m_{31}-m_{43}-1 \in \mathbb{Z}_+\}. \qquad (115)$$

Then [see (63), (64)]

$$V^2_{\text{inv}}([m_{13}-1,m_{23}-1,-m_{23}+1,m_{43}+1]) = \boxed{-1,-1,1,1}^2_{\text{inv}} \quad , \qquad (116)$$

$$V^2([m_{13}-1,m_{23}-1,-m_{23}+1,m_{43}+1]) = \boxed{-1,-1,1,1}^2 \quad . \qquad (117)$$

The decomposition of $W([m_{13},m_{23},-m_{23},m_{43}])$ into a direct sum of irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules reads

$$\begin{aligned} &W([m_{13},m_{23},-m_{23},m_{43}]) \\ &= \boxed{0,0,0,0}_{00} \\ &\quad \oplus \boxed{-1,0,1,0}_{10} \quad \oplus \boxed{0,-1,1,0}_{10} \quad \oplus \boxed{-1,0,0,1}_{10} \quad \oplus \boxed{0,-1,0,1}_{10} \\ &\quad \oplus \boxed{-1,-1,2,0}_{20} \quad \oplus \boxed{-1,-1,0,2}_{20} \quad \oplus \boxed{-2,0,1,1}_{11} \quad \oplus \boxed{0,-2,1,1}_{11} \\ &\quad \oplus \boxed{-1,-1,1,1}^2_{\text{inv}} \quad \oplus \boxed{-1,-1,1,1}^2 \\ &\quad \oplus \boxed{-2,-1,2,1}_{21} \quad \oplus \boxed{-1,-2,2,1}_{21} \quad \oplus \boxed{-2,-1,1,2}_{21} \quad \oplus \boxed{-1,-2,1,2}_{21} \\ &\quad \oplus \boxed{-2,-2,2,2}_{22} \quad , \end{aligned} \qquad (118)$$

where

$$\boxed{-2,0,1,1}_{11} = 0, \quad \text{if } m_{13} = m_{23}+1 \text{ and } -m_{23} > m_{43}+1, \qquad (119)$$
$$\boxed{-1,-1,0,2}_{20} = 0, \quad \text{if } m_{13} > m_{23}+1 \text{ and } -m_{23} = m_{43}+1. \qquad (120)$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{}$ terms in (118), i.e., from the maximal invariant subspace I_2 by zero and m_{33} by $-m_{23}$, one obtains the transformations of the nontypical module $W_2([m_{13},m_{23},-m_{23},m_{43}])$ under the action of the superalgebra.

Transformations under the action of e_{32}:

$$e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23}, & m_{43} \\ m_{13}, & m_{23}, & -m_{23}, & m_{43} \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{00}$$
$$= -\left|\frac{(l_{13}-l_{11})(l_{43}-l_{31})}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & m_{43} \\ m_{11} & , 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10}$$
$$-\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}, & -m_{23} & , m_{43}+1 \\ m_{11} & , 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10}$$
$$-\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{23} & , m_{43} \\ m_{13}, & m_{23}-1, & -m_{23} & , m_{43}+1 \\ m_{11}, & 0 & , m_{31}+1, & 0 \end{bmatrix}_{10}, \qquad (121)$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & m_{43} \\ m_{11} & , 0\ , & m_{31} & , 0 \end{bmatrix}_{10}$$
$$= -\left|\frac{(l_{13}-l_{11}-1)(l_{23}+l_{31}+2)}{(l_{13}-l_{23}-1)(l_{23}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0\ , & m_{31}+1, & 0 \end{bmatrix}_{11}$$
$$-\frac{1}{l_{13}-l_{23}-1}\left|\frac{(l_{13}-l_{23})(l_{23}-l_{11})(l_{23}+l_{31}+2)}{2(l_{23}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0\ , & m_{31}+1, & 0 \end{bmatrix}^2, \qquad (122)$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23}, & m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}, & m_{43}+1 \\ m_{11} & , 0 , & m_{31}, & 0 \end{bmatrix}_{10}$$

$$= \left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{43}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23} & , m_{43}+2 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}_{20}$$

$$+ \left|\frac{(l_{13}-l_{11}-1)(l_{43}-l_{31}+1)}{(l_{13}-l_{23}-1)(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0 , & m_{31}+1, & 0 \end{bmatrix}_{11}$$

$$+ \frac{(l_{13}+l_{43}+3)}{(l_{13}-l_{23}-1)(l_{23}+l_{43}+4)}\left|\frac{(l_{23}-l_{11})(l_{43}-l_{31}+1)}{2(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2}$$

$$\times\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}^{2}, \tag{123}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23} & , -m_{23}, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{23}, & m_{43}+1 \\ m_{11}, & 0 & , m_{31}, & 0 \end{bmatrix}_{10}$$

$$= -\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{43}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23} & , m_{43}+2 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}_{20}$$

$$+ \left|\frac{(l_{13}-l_{11})(l_{43}-l_{31}+1)(l_{23}+l_{43}+3)}{2(l_{13}-l_{23})(l_{23}+l_{43}+4)(l_{23}+l_{43}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}^{2}, \tag{124}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23}, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}, & m_{43}+2 \\ m_{11} & , 0 & , m_{31}, & 0 \end{bmatrix}_{20}$$

$$= \left|\frac{(l_{13}-l_{11}-1)(l_{43}-l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}-1, & -m_{23}+1, & m_{43}+2 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}_{21}, \tag{125}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}^{2}$$

$$= \left|\frac{2(l_{13}-l_{11}-1)(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}-1, & -m_{23}+1, & m_{43}+2 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}_{21}, \tag{126}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}, & -m_{23}+1, & m_{43}+1 \\ m_{11} & , 0 , & m_{31} & , 0 \end{bmatrix}_{11}$$

$$= -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23}-1)(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}-1, & -m_{23}+1, & m_{43}+2 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}_{21}, \tag{127}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , m_{43} \\ m_{13}-2, & m_{23}-1, & -m_{23}+1, & m_{43}+2 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{21} = 0. \tag{128}$$

Transformations under the action of e_{23}:

$$e_{23}\begin{bmatrix} m_{13}\ , & m_{23}\ , & -m_{23}\ , & m_{43} \\ m_{13}-2, & m_{23}-1, & -m_{23}+1, & m_{43}+2 \\ m_{11}\ , & 0\ , & m_{31}\ , & 0 \end{bmatrix}_{21}$$

$$= -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+1)(l_{23}+l_{43}+3)}{l_{13}-l_{23}-1}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-2, & m_{23}, & -m_{23}+1, & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{11}$$

$$+\left|\frac{(l_{13}-l_{11}-1)(l_{43}-l_{31}+3)(l_{13}-l_{23})}{l_{23}+l_{43}+4}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}\ , & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}\ , & m_{43}+2 \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{20}$$

$$+\frac{(l_{13}+l_{43}+3)|2(l_{13}-l_{23})(l_{13}-l_{11}-1)(l_{23}+l_{31}+1)(l_{23}+l_{43}+3)|^{1/2}}{2(l_{13}-l_{23}-1)(l_{23}+l_{43}+4)}$$

$$\times\begin{bmatrix} m_{13}\ , & m_{23}\ , & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}^{2}, \qquad (129)$$

$$e_{23}\begin{bmatrix} m_{13}\ , & m_{23}\ , & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}\ , & 0 \end{bmatrix}^{2}$$

$$= -(l_{23}+l_{43}+4)\left|\frac{2(l_{23}-l_{11})(l_{23}+l_{31}+1)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & m_{43} \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}$$

$$+(l_{13}-l_{23}-1)\left|\frac{2(l_{13}-l_{11})(l_{43}-l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}\ , & -m_{23}\ , & m_{43} \\ m_{13}, & m_{23}-1, & -m_{23}\ , & m_{43}+1 \\ m_{11}, & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}$$

$$+\left|\frac{2(l_{23}-l_{11})(l_{43}-l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}\ , & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (130)$$

$$e_{23}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-2, & m_{23}, & -m_{23}+1, & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}\ , & 0 \end{bmatrix}_{11}$$

$$= -(l_{13}+l_{43}+3)\left|\frac{(l_{13}-l_{11}-1)(l_{23}+l_{31}+1)}{(l_{13}-l_{23}-1)(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & m_{43} \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}$$

$$+(l_{13}-l_{23})\left|\frac{(l_{13}-l_{11}-1)(l_{43}-l_{31}+2)}{(l_{13}-l_{23}-1)(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}\ , & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (131)$$

$$e_{23}\begin{bmatrix} m_{13}\ , & m_{23}\ , & -m_{23}, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{23}, & m_{43}+2 \\ m_{11}\ , & 0\ , & m_{31}, & 0 \end{bmatrix}_{20}$$

$$= -(l_{13}+l_{43}+3)\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+4)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}\ , & -m_{23}\ , & m_{43} \\ m_{13}, & m_{23}-1, & -m_{23}\ , & m_{43}+1 \\ m_{11}, & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}$$

$$+(l_{23}+l_{43}+3)\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+4)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & -m_{23}\ , & m_{43} \\ m_{13}-1, & m_{23}, & -m_{23}\ , & m_{43}+1 \\ m_{11}\ , & 0\ , & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (132)$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23} & , & -m_{23}, & m_{43} \\ m_{13}, & m_{23}-1, & & -m_{23}, & m_{43}+1 \\ m_{11}, & 0 & , & m_{31}, & 0 \end{bmatrix}_{10}$$

$$= -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+2)(l_{23}+l_{43}+3)}{l_{13}-l_{23}}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23} & , & m_{43} \\ m_{13}, & m_{23}, & -m_{23} & , & m_{43} \\ m_{11}, & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{00}, \qquad (133)$$

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23}, & -m_{23}, & m_{43} \\ m_{13}-1, & & m_{23}, & -m_{23}, & m_{43}+1 \\ m_{11} & , & 0\ , & m_{31}, & 0 \end{bmatrix}_{10}$$

$$= -(l_{13}+l_{43}+3)\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{43}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23} & , & m_{43} \\ m_{13}, & m_{23}, & -m_{23} & , & m_{43} \\ m_{11}, & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{00}, \qquad (134)$$

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23}, & -m_{23} & , & m_{43} \\ m_{13}-1, & & m_{23}, & -m_{23}+1, & & m_{43} \\ m_{11} & , & 0\ , & m_{31} & , & 0 \end{bmatrix}_{10}$$

$$= -\left|\frac{(l_{13}-l_{11})(l_{43}-l_{31}+1)(l_{13}-l_{23})}{l_{23}+l_{43}+3}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23} & , & m_{43} \\ m_{13}, & m_{23}, & -m_{23} & , & m_{43} \\ m_{11}, & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{00}, \qquad (135)$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23}, & m_{43} \\ m_{13}, & m_{23}, & -m_{23}, & m_{43} \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{00} = 0. \qquad (136)$$

Proposition 11: Let $W([m_{13},m_{23},-m_{23},m_{43}])$ be a class 2 indecomposible induced gl(2/2) module, for which

$$m_{13}=m_{23} \text{ and } -m_{23}>m_{43}. \qquad (137)$$

Then the structure of this module with respect to $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reads:

$$W([m_{13},m_{13},-m_{13},m_{43}])$$

$$\begin{aligned} &= \boxed{0,0,0,0}_{00} \\ &\oplus \boxed{0,-1,1,0}_{10} \quad \oplus \boxed{0,-1,0,1}_{10} \\ &\oplus \boxed{-1,-1,2,0}_{20} \quad \oplus \boxed{-1,-1,1,1}_{20} \quad \oplus \boxed{-1,-1,0,2}_{20} \\ &\oplus \boxed{0,-2,1,1}_{11} \quad \oplus \boxed{-1,-2,2,1}_{21} \quad \oplus \boxed{-1,-2,1,2}_{21} \\ &\oplus \boxed{-2,-2,2,2}_{22} \quad , \end{aligned} \qquad (138)$$

where

$$\boxed{-1,-1,0,2}_{20} = 0, \text{ if } -m_{23}=m_{43}+1. \qquad (139)$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{\qquad\quad}$ terms in (138) (= from I_2) by zero m_{23} by m_{13} and m_{33} by $-m_{13}$, one obtains the transformations of the nontypical module, corresponding to the case (137), under the action of the superalgebra.

Proposition 12: Let $W([m_{13},m_{23},-m_{23},m_{43}])$ be a class 2 indecomposible induced gl(2/2) module, for which

$$m_{13}>m_{23} \text{ and } -m_{23}=m_{43}. \qquad (140)$$

The structure of the module with respect to $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reads:

$$W([m_{13},m_{23},-m_{23},-m_{23}])$$

$$\begin{aligned} &= \boxed{0,0,0,0}_{00} \\ &\oplus \boxed{0,-1,1,0}_{10} \quad \oplus \boxed{-1,0,1,0}_{10} \\ &\oplus \boxed{-1,-1,2,0}_{20} \quad \oplus \boxed{-1,-1,1,1}_{11} \quad \oplus \boxed{-2,0,1,1}_{11} \\ &\oplus \boxed{0,-2,1,1}_{11} \quad \oplus \boxed{-1,-2,2,1}_{21} \quad \oplus \boxed{-2,-1,2,1}_{21} \\ &\oplus \boxed{-2,-2,2,2}_{22} \quad , \end{aligned} \qquad (141)$$

where

$$\boxed{-2,0,1,1}_{11} = 0, \text{ if } m_{13}=m_{23}+1. \qquad (142)$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{\qquad\quad}$ terms in (141), i.e., from I_2 by zero, m_{33} by $-m_{23}$ and m_{43} by $-m_{23}$, one obtains the transformations of the nontypical module, corresponding to the case (140), under the action of the superalgebra.

C. The class 3 nontypical representations

In this section, we consider the indecomposible induced modules $W([m])$, corresponding to the case

$$l_{23}+l_{43}+3=0 \Leftrightarrow m_{23}+m_{43}-1=0. \qquad (143)$$

The induced modules from this class have signatures $[m]=[m_{13},m_{23},m_{33},-m_{23}+1]$, i.e.,

$$W([m])=W([m_{13},m_{23},m_{33},-m_{23}+1]). \qquad (144)$$

Proposition 13: Consider a class 3 indecomposible induced gl(2/2) module (144), corresponding to a signature $[m_{13},m_{23},m_{33},-m_{23}+1]$ such that

$$m_{13}>m_{23} \text{ and } m_{33}>-m_{23}+1\,. \tag{145}$$

Introduce a new basis in the $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reducible module

$$\boxed{-1,-1,1,1}_{20} \quad \oplus \quad \boxed{-1,-1,1,1}_{11} \tag{146}$$

setting

$$[m_{11},m_{31}]^3 \equiv \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}^3$$

$$= \left|\frac{l_{13}-l_{23}-1}{l_{13}-l_{23}+1}\right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{11}$$

$$+ \left|\frac{l_{23}+l_{33}+4}{l_{23}+l_{33}+2}\right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}, \tag{147}$$

$$[m_{11},m_{31}]^3_{\text{inv}} = \left|\frac{l_{13}-l_{23}-1}{l_{13}-l_{23}+1}\right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{11}$$

$$- \left|\frac{l_{23}+l_{33}+4}{l_{23}+l_{33}+2}\right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}. \tag{148}$$

Construct two new irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules with a signature $[m_{13}-1,m_{23}-1,m_{33}+1,-m_{23}+2]$:

$$V^3 \equiv V^3([m_{13}-1,m_{23}-1,m_{33}+1,-m_{23}+2])$$
$$= \text{lin.env.}$$
$$\{[m_{11},m_{31}]^3 | m_{13}-m_{11}-1, m_{11}-m_{23}+1,$$
$$m_{33}-m_{31}+1, m_{31}+m_{23}-2 \in \mathbf{Z}_+\}, \tag{149}$$
$$V^3_{\text{inv}} = V^3_{\text{inv}}([m_{13}-1,m_{23}-1,m_{33}+1,-m_{23}+2])$$
$$= \text{lin.env.}$$
$$\{[m_{11},m_{31}]^3_{\text{inv}} | m_{13}-m_{11}-1, m_{11}-m_{23}+1,$$
$$m_{33}-m_{31}+1, m_{31}+m_{23}-2 \in \mathbf{Z}_+\}\,. \tag{150}$$

Then [see (63), (64)]

$$V^3_{\text{inv}}([m_{13}-1,m_{23}-1,m_{33}+1,-m_{23}+2])$$
$$= \boxed{-1,-1,1,1}^3_{\text{inv}}\,, \tag{151}$$
$$V^3([m_{13}-1,m_{23}-1,m_{33}+1,-m_{23}+2])$$
$$= \boxed{-1,-1,1,1}^3\,. \tag{152}$$

The decomposition of

$$W([m]) = W([m_{13},m_{23},m_{33},-m_{23}+1])$$

into a direct sum of irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules reads

$$W([m]) = W([m_{13},m_{23},m_{33},-m_{23}+1])$$
$$= \boxed{0,0,0,0}_{00}$$
$$\oplus \boxed{-1,0,1,0}_{10} \oplus \boxed{0,-1,1,0}_{10} \oplus \boxed{-1,0,0,1}_{10} \oplus \boxed{0,-1,0,1}_{10}$$
$$\oplus \boxed{-1,-1,2,0}_{20} \oplus \boxed{-1,-1,0,2}_{20} \oplus \boxed{-2,0,1,1}_{11} \oplus \boxed{0,-2,1,1}_{11}$$
$$\oplus \boxed{-1,-1,1,1}^3_{\text{inv}} \oplus \boxed{-1,-1,1,1}^3$$
$$\oplus \boxed{-2,-1,2,1}_{21} \oplus \boxed{-1,-2,2,1}_{21} \oplus \boxed{-2,-1,1,2}_{21} \oplus \boxed{-1,-2,1,2}_{21}$$
$$\oplus \boxed{-2,-2,2,2}_{22}\,, \tag{153}$$

where

$$\boxed{-2,0,1,1}_{11} = 0, \text{ if } m_{13}=m_{23}+1 \text{ and } m_{33}>-m_{13}+3\,, \tag{154}$$

$$\boxed{-1,-1,0,2}_{20} = 0, \text{ if } m_{13}>m_{23}+1 \text{ and } m_{33}=-m_{23}+2\,, \tag{155}$$

$$\boxed{-1,-1,0,2}_{20} = 0, \quad \boxed{-2,0,1,1}_{11} = 0, \text{ if } m_{13}=m_{23}+1 \text{ and } m_{33}=-m_{13}+3. \tag{156}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all ▭ terms in (153) (= from I_3) by zero and m_{43} by $-m_{23}+1$, one obtains the transformations of the nontypical module $W_3([m_{13}, m_{23}, m_{33}, -m_{23}+1])$ under the action of the Lie superalgebra.

Transformations under the action of e_{32}:

$$
\begin{aligned}
&e_{32}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{23}+1 \\ m_{13}, & m_{23}, & m_{33}, & -m_{23}+1 \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{00} \\
&\quad = -\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{23}+1 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10} \\
&\qquad -\left|\frac{(l_{13}-l_{11})(l_{33}-l_{31})}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}\ , & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10} \\
&\qquad -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{23}+1 \\ m_{11}, & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10},
\end{aligned}
\tag{157}
$$

$$
\begin{aligned}
&e_{32}\begin{bmatrix} m_{13}\ , & m_{23}, & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{23}+1 \\ m_{11}\ , & 0\ , & m_{31}\ , & 0 \end{bmatrix}_{10} \\
&\quad = \left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{23}+1 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{20} \\
&\qquad -\left|\frac{(l_{13}-l_{11}-1)(l_{33}-l_{31}+1)}{(l_{13}-l_{23}-1)(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & m_{33}\ , & -m_{23}+1 \\ m_{13}-2, & m_{23}, & m_{33}+1, & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{11} \\
&\qquad -\frac{l_{13}+l_{33}+3}{(l_{13}-l_{23}-1)(l_{23}+l_{33}+4)}\left|\frac{(l_{23}-l_{11})(l_{33}-l_{31}+1)}{2(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}^{3},
\end{aligned}
\tag{158}
$$

$$
\begin{aligned}
&e_{32}\begin{bmatrix} m_{13}\ , & m_{23}, & m_{33}, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}, & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}, & 0 \end{bmatrix}_{10} \\
&\quad = \left|\frac{(l_{13}-l_{11}-1)(l_{23}+l_{31}+2)}{(l_{13}-l_{23}-1)(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}, & m_{33}\ , & -m_{23}+1 \\ m_{13}-2, & m_{23}, & m_{33}+1, & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{11} \\
&\qquad +\frac{1}{l_{13}-l_{23}-1}\left|\frac{(l_{13}-l_{23})(l_{23}-l_{11})(l_{23}+l_{31}+2)}{2(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}^{3},
\end{aligned}
\tag{159}
$$

$$
\begin{aligned}
&e_{32}\begin{bmatrix} m_{13}, & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{23}+1 \\ m_{11}, & 0\ , & m_{31}\ , & 0 \end{bmatrix}_{10} \\
&\quad = -\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{23}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{23}+1 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}_{20} \\
&\qquad -\frac{1}{l_{23}+l_{33}+4}\left|\frac{(l_{13}-l_{11})(l_{33}-l_{31}+1)(l_{23}+l_{33}+3)}{2(l_{13}-l_{23})}\right|^{1/2}\begin{bmatrix} m_{13}\ , & m_{23}\ , & m_{33}\ , & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{23}+2 \\ m_{11}\ , & 0\ , & m_{31}+1, & 0 \end{bmatrix}^{3},
\end{aligned}
\tag{160}
$$

$$e_{32}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+1 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{20}$$

$$= -\left|\frac{(l_{13}-l_{11}-1)(l_{33}-l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31}+1, & & 0 \end{bmatrix}_{21}, \tag{161}$$

$$e_{32}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}^{3}$$

$$= \left|\frac{2(l_{13}-l_{11}-1)(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31}+1, & & 0 \end{bmatrix}_{21}, \tag{162}$$

$$e_{32}\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}, & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0\ , & m_{31} & , & 0 \end{bmatrix}_{11}$$

$$= -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23}-1)(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31}+1, & & 0 \end{bmatrix}_{21}, \tag{163}$$

$$e_{32}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{21} = 0. \tag{164}$$

Transformations under the action of e_{23}:

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{21}$$

$$= -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+1)(l_{23}+l_{33}+3)}{l_{13}-l_{23}-1}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{23}+1 \\ m_{13}-2, & & m_{23}, & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{11}$$

$$-\left|\frac{(l_{13}-l_{11}-1)(l_{13}-l_{23})(l_{33}-l_{31}+3)}{l_{23}+l_{33}+4}\right|^{1/2}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+2, & & -m_{23}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20}$$

$$+\frac{(l_{13}+l_{33}+3)|2(l_{13}-l_{11}-1)(l_{13}-l_{23})(l_{23}+l_{31}+1)(l_{23}+l_{33}+3)|^{1/2}}{2(l_{13}-l_{23}-1)(l_{23}+l_{33}+4)}$$

$$\times\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{23}+2 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}^{3}, \tag{165}$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} \quad, & m_{23} \quad, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{23}+2 \\ m_{11} \quad, & 0 \quad, & m_{31} \quad, & 0 \end{bmatrix}^{3} \\
&\quad = -(l_{13}-l_{23}-1)\left|\frac{2(l_{13}-l_{11})(l_{33}-l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} \quad, & m_{33} \quad, & -m_{23}+1 \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{23}+1 \\ m_{11}, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\qquad - \left|\frac{2(l_{23}-l_{11})(l_{33}-l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{23}+1 \\ m_{11} \quad, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\qquad + (l_{23}+l_{33}+4)\left|\frac{2(l_{23}-l_{11})(l_{23}+l_{31}+1)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33} \quad, & -m_{23}+2 \\ m_{11} \quad, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (166)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-2, & m_{23}, & m_{33}+1, & -m_{23}+2 \\ m_{11} \quad, & 0 \quad, & m_{31} \quad, & 0 \end{bmatrix}_{11} \\
&\quad = -(l_{13}-l_{23})\left|\frac{(l_{13}-l_{11}-1)(l_{33}-l_{31}+2)}{(l_{13}-l_{23}-1)(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{23}+1 \\ m_{11} \quad, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\qquad + (l_{13}+l_{33}+3)\left|\frac{(l_{13}-l_{11}-1)(l_{23}+l_{31}+1)}{(l_{13}-l_{23}-1)(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33} \quad, & -m_{23}+2 \\ m_{11} \quad, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (167)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} \quad, & m_{23} \quad, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}-1, & m_{33}+2, & -m_{23}+1 \\ m_{11} \quad, & 0 \quad, & m_{31} \quad, & 0 \end{bmatrix}_{20} \\
&\quad = -(l_{13}+l_{33}+3)\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23} \quad, & m_{33} \quad, & -m_{23}+1 \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{23}+1 \\ m_{11}, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\qquad + (l_{23}+l_{33}+3)\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+4)}\right|^{1/2}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}+1, & -m_{23}+1 \\ m_{11} \quad, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (168)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13}, & m_{23} \quad, & m_{33} \quad, & -m_{23}+1 \\ m_{13}, & m_{23}-1, & m_{33}+1, & -m_{23}+1 \\ m_{11}, & 0 \quad, & m_{31} \quad, & 0 \end{bmatrix}_{10} \\
&\quad = -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+2)(l_{23}+l_{33}+3)}{l_{13}-l_{23}}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{11}, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{00}, \qquad (169)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} \quad, & m_{23}, & m_{33}, & -m_{23}+1 \\ m_{13}-1, & m_{23}, & m_{33}, & -m_{23}+2 \\ m_{11} \quad, & 0 \quad, & m_{31}, & 0 \end{bmatrix}_{10} \\
&\quad = -\left|\frac{(l_{13}-l_{11})(l_{33}-l_{31}+1)(l_{13}-l_{23})}{l_{23}+l_{33}+3}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{13}, & m_{23}, & m_{33} \quad, & -m_{23}+1 \\ m_{11}, & 0 \quad, & m_{31}-1, & 0 \end{bmatrix}_{00}, \qquad (170)
\end{aligned}
$$

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{23}+1 \\ m_{13}-1, & & m_{23}, & m_{33}+1, & & -m_{23}+1 \\ m_{11} & , & 0\ , & m_{31} & , & 0 \end{bmatrix}_{10}$$

$$= -(l_{13}+l_{33}+3)\left|\frac{(l_{13}-l_{11})(l_{23}+l_{31}+2)}{(l_{13}-l_{23})(l_{23}+l_{33}+3)}\right|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & m_{33} & , & -m_{23}+1 \\ m_{13}, & m_{23}, & m_{33} & , & -m_{23}+1 \\ m_{11}, & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{00}, \tag{171}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & -m_{23}+1 \\ m_{13}, & m_{23}, & m_{33}, & -m_{23}+1 \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{00} = 0\,. \tag{172}$$

Proposition 14: Replacing in (153) the subspaces

$$\boxed{-1,0,1,0}_{10}\ ,\ \boxed{-1,0,0,1}_{10}\ ,\ \boxed{0,-1,0,1}_{10}\ ,\ \boxed{-1,-1,0,2}_{20}\ ,$$
$$\boxed{-2,0,1,1}_{11}\ ,\ \boxed{-2,-1,2,1}_{21}\ ,\ \boxed{-2,-1,1,2}_{21}\ ,\ \boxed{-1,-2,1,2}_{21}\ ,$$
$$\boxed{-1,-1,1,1}^{3}_{\rm inv}\ ,\ \boxed{-1,-1,1,1}^{3} \tag{173}$$

by zero, one obtains the decomposition of the class 3 indecomposible module $W([m_{13},m_{13},-m_{13}+1,m_{13}+1])$ with respect to $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$. Replacing in (157)–(172) the basis vectors of the subspaces (173) by zero, one obtains the transformation of the nontypical module $W_3([m_{13},m_{13},-m_{13}+1,m_{13}+1])$ under the action of e_{23} and e_{32}.

Proposition 15: Let $W([m]) = W([m_{13},m_{23},m_{33},-m_{23}+1])$ be a class 3 indecomposible induced gl(2/2) module, for which

$$m_{13}=m_{23} \text{ and } m_{33} > -m_{13}+1\,. \tag{174}$$

The structure of this module with respect to $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reads:

$$\begin{aligned} W([m]) &= W([m_{13},m_{13},m_{33},-m_{13}+1]) \\ &= \boxed{0,0,0,0}_{00} \\ &\quad \oplus \boxed{0,-1,1,0}_{10} \quad \oplus \boxed{0,-1,0,1}_{10} \\ &\quad \oplus \boxed{-1,-1,2,0}_{20} \quad \oplus \boxed{-1,-1,1,1}_{20} \quad \oplus \boxed{-1,-1,0,2}_{20} \quad \oplus \boxed{0,-2,1,1}_{11} \\ &\quad \oplus \boxed{-1,-2,2,1}_{21} \quad \oplus \boxed{-1,-2,1,2}_{21} \\ &\quad \oplus \boxed{-2,-2,2,2}_{22} \quad , \end{aligned} \tag{175}$$

where

$$\boxed{-1,-1,0,2}_{20} = 0\,, \quad \text{if } m_{33} = -m_{13}+2\,. \tag{176}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{}$ terms in (175), i.e., from I_3 by zero, m_{23} by m_{13} and m_{43} by $-m_{13}+1$, one obtains the transformations of the nontypical module, corresponding to the case (174), under the action of the superalgebra.

Proposition 16: Let $W([m]) = W([m_{13},m_{23},m_{33},-m_{23}+1])$ be a class 3 indecomposible induced gl(2/2) module, for which

$$m_{13} > m_{23} \text{ and } m_{33} = -m_{23}+1\,. \tag{177}$$

The structure of this module with respect to $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reads:

$$\begin{aligned} W([m]) &= W([m_{13},m_{23},-m_{23}+1,-m_{23}+1]) \\ &= \boxed{0,0,0,0}_{00} \\ &\quad \oplus \boxed{0,-1,1,0}_{10} \quad \oplus \boxed{-1,0,1,0}_{10} \\ &\quad \oplus \boxed{-1,-1,2,0}_{20} \quad \oplus \boxed{-1,-1,1,1}_{11} \quad \oplus \boxed{-2,0,1,1}_{11} \quad \oplus \boxed{0,-2,1,1}_{11} \\ &\quad \oplus \boxed{-1,-2,2,1}_{21} \quad \oplus \boxed{-2,-1,2,1}_{21} \\ &\quad \oplus \boxed{-2,-2,2,2}_{22} \quad , \end{aligned} \tag{178}$$

where

$$\boxed{-2,0,1,1}_{11} = 0\,, \quad \text{if } m_{13} = m_{23}+1\,. \tag{179}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{}$ terms in (178) (= from I_3) by zero, m_{33} by $-m_{23}+1$ and m_{43} by $-m_{23}+1$, one obtains the transformations of the nontypical module, corresponding to the case (177), under the action of the superalgebra.

D. The class 4 nontypical representations

In this section, we consider the indecomposible induced modules $W([m])$, corresponding to the case

$$l_{13}+l_{33}+3=0 \Leftrightarrow m_{33}=-m_{13}-1\,. \qquad (180)$$

The induced modules from this class have signatures $[m]=[m_{13},m_{23},-m_{13}-1,m_{43}]$, i.e.,

$$W([m])=W([m_{13},m_{23},-m_{13}-1,-m_{43}])\,. \quad (181)$$

Proposition 17: Consider a class 4 indecomposible induced gl(2/2) module (181), whose signature satisfies, in addition, the conditions

$$m_{13}>m_{23} \text{ and } -m_{13}-1>m_{43}\,. \qquad (182)$$

Introduce a new basis in the $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ reducible module

$$\boxed{-1,-1,1,1}_{20} \oplus \boxed{-1,-1,1,1}_{11} \qquad (183)$$

setting

$$[m_{11},m_{31}]^4 \equiv \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}^4$$

$$= -\left|\frac{l_{13}-l_{23}+1}{l_{13}-l_{23}-1}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{11}$$

$$-\left|\frac{l_{13}+l_{43}+4}{l_{13}+l_{43}+2}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{20}, \qquad (184)$$

$$[m_{11},m_{31}]^4_{\mathrm{inv}} = -\left|\frac{l_{13}-l_{23}+1}{l_{13}-l_{23}-1}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{11}$$

$$+\left|\frac{l_{13}+l_{43}+4}{l_{13}+l_{43}+2}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{20}, \qquad (185)$$

Construct two new irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules with a signature $[m_{13}-1,m_{23}-1,-m_{13},m_{43}+1]$:

$$V^4 \equiv V^4([m_{13}-1,m_{23}-1,-m_{13},m_{43}+1])$$

$$= \text{lin.env.}\,\{[m_{11},m_{31}]^4 | m_{13}-m_{11}-1, m_{11}-m_{23}+1, -m_{13}-m_{31}, m_{31}-m_{43}-1 \in \mathbb{Z}_+\}\,, \qquad (186)$$

$$V^4_{\mathrm{inv}} \equiv V^4_{\mathrm{inv}}([m_{13}-1,m_{23}-1,-m_{13},m_{43}+1])$$

$$= \text{lin.env.}\,\{[m_{11},m_{31}]^4_{\mathrm{inv}} | m_{13}-m_{11}-1, m_{11}-m_{23}+1, -m_{13}-m_{31}, m_{31}-m_{43}-1 \in \mathbb{Z}_+\}\,. \qquad (187)$$

Then [see (63), (64)]

$$V^4_{\mathrm{inv}}([m_{13}-1,m_{23}-1,-m_{13},m_{43}+1]) = \boxed{-1,-1,1,1}^4_{\mathrm{inv}}\,, \qquad (188)$$

$$V^4([m_{13}-1,m_{23}-1,-m_{13},m_{43}+1]) = \boxed{-1,-1,1,1}^4\,, \qquad (189)$$

The decomposition of $W([m])=W([m_{13},m_{23},-m_{13}-1,m_{43}])$ into a direct sum of irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules reads

$$
\begin{aligned}
W([m\}) = {} & W([m_{13},m_{23},-m_{13}-1,m_{43}]) \\
= {} & \boxed{0,0,0,0}_{00} \\
& \oplus \boxed{-1,0,1,0}_{10} \oplus \boxed{0,-1,1,0}_{10} \oplus \boxed{-1,0,0,1}_{10} \oplus \boxed{0,-1,0,1}_{10} \\
& \oplus \boxed{-1,-1,2,0}_{20} \oplus \boxed{-1,-1,0,2}_{20} \oplus \boxed{-2,0,1,1}_{11} \oplus \boxed{0,-2,1,1}_{11} \\
& \oplus \boxed{-1,-1,1,1}^{4}_{1\mathrm{nv}} \oplus \boxed{-1,-1,1,1}^{4} \\
& \oplus \boxed{-2,-1,2,1}_{21} \oplus \boxed{-1,-2,2,1}_{21} \oplus \boxed{-2,-1,1,2}_{21} \oplus \boxed{-1,-2,1,2}_{21} \\
& \oplus \boxed{-2,-2,2,2}_{22} \quad ,
\end{aligned}
\tag{190}
$$

where

$$\boxed{-1,-1,0,2}_{20} = 0, \text{ if } m_{13} > m_{23}+1 \text{ and } -m_{13} = m_{43}+2\,, \tag{191}$$

$$\boxed{-2,0,1,1}_{11} = 0, \text{ if } m_{13} = m_{23}+1 \text{ and } -m_{13} > m_{43}+2\,, \tag{192}$$

$$\boxed{-1,-1,0,2}_{20} = 0, \quad \boxed{-2,0,1,1}_{11} = 0, \text{ if } m_{13} = m_{23}+1 \text{ and } -m_{13} = m_{43}+2\,. \tag{193}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{}$ terms in (190) (= from I_4) by zero and m_{33} by $-m_{13}-1$, one obtains the transformations of the nontypical module $W_4([m_{13},m_{23},-m_{13}-1,m_{43}])$ under the action of the Lie superalgebra.

Transformations under the action of e_{32}:

$$
\begin{aligned}
& e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{11}, & 0\ , & m_{31} & , \ 0 \end{bmatrix}_{00} \\
& = -\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13} & , \ m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}, & -m_{13}-1, & m_{43}+1 \\ m_{11} & , \ 0\ , & m_{31}+1, & 0 \end{bmatrix}_{10} \\
& \quad -\left|\frac{(l_{23}-l_{11})(l_{43}-l_{31})}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , \ -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13} & , \ m_{43} \\ m_{11}, & 0 & , \ m_{31}+1, & 0 \end{bmatrix}_{10} \\
& \quad -\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , \ -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}-1, & m_{43}+1 \\ m_{11}, & 0 & , \ m_{31}+1, & 0 \end{bmatrix}_{10} ,
\end{aligned}
\tag{194}
$$

$$
\begin{aligned}
& e_{32}\begin{bmatrix} m_{13} & , \ m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}, & -m_{13}-1, & m_{43}+1 \\ m_{11} & , \ 0, & m_{31} & , \ 0 \end{bmatrix}_{10} \\
& = \left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{43}+4)}\right|^{1/2} \begin{bmatrix} m_{13} & , \ m_{23} & , \ -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}-1, & m_{43}+2 \\ m_{11} & , \ 0 & , \ m_{31}+1, & 0 \end{bmatrix}_{20} \\
& \quad -\frac{1}{l_{13}+l_{43}+4}\left|\frac{(l_{23}-l_{11})(l_{43}-l_{31}+1)(l_{13}+l_{43}+3)}{2(l_{13}-l_{23})}\right|^{1/2} \begin{bmatrix} m_{13} & , \ m_{23} & , \ -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , \ m_{43}+1 \\ m_{11} & , \ 0 & , \ m_{31}+1, & 0 \end{bmatrix}^{4} ,
\end{aligned}
\tag{195}
$$

$$
\begin{aligned}
e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}, & m_{43} \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{10}
&= -\left|\frac{(l_{23}-l_{11}-1)(l_{13}+l_{31}+2)}{(l_{13}-l_{23}+1)(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-2, & -m_{13}, & m_{43}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{11} \\
&\quad + \frac{1}{l_{13}-l_{23}+1}\left|\frac{(l_{13}-l_{11})(l_{13}-l_{23})(l_{13}+l_{31}+2)}{2(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}, & m_{43}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}^{4},
\end{aligned}
\tag{196}
$$

$$
\begin{aligned}
e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}-1, & m_{43}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{10}
&= -\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}+l_{43}+4)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}-1, & m_{43}+2 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{20} \\
&\quad + \left|\frac{(l_{23}-l_{11}-1)(l_{43}-l_{31}+1)}{(l_{13}-l_{23}+1)(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-2, & -m_{13}, & m_{43}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{11} \\
&\quad + \frac{(l_{23}+l_{43}+3)}{(l_{13}+l_{43}+4)(l_{13}-l_{23}+1)}\left|\frac{(l_{13}-l_{11})(l_{43}-l_{31}+1)}{2(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}, & m_{43}+1 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}^{4},
\end{aligned}
\tag{197}
$$

$$
\begin{aligned}
e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}-1, & m_{43}+2 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{20}
&= \left|\frac{(l_{23}-l_{11}-1)(l_{43}-l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+4)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-2, & -m_{13}, & m_{43}+2 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{21},
\end{aligned}
\tag{198}
$$

$$
\begin{aligned}
e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}, & m_{43}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}^{4}
&= \left|\frac{2(l_{23}-l_{11}-1)(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-2, & -m_{13}, & m_{43}+2 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{21},
\end{aligned}
\tag{199}
$$

$$
\begin{aligned}
e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-2, & -m_{13}, & m_{43}+1 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{11}
&= \left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23}+1)(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-2, & -m_{13}, & m_{43}+2 \\ m_{11}, & 0, & m_{31}+1, & 0 \end{bmatrix}_{21},
\end{aligned}
\tag{200}
$$

$$
e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-2, & -m_{13}, & m_{43}+2 \\ m_{11}, & 0, & m_{31}, & 0 \end{bmatrix}_{21} = 0 .
\tag{201}
$$

Transformations under the action of e_{23}:

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-2, & -m_{13} & , m_{43}+2 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{21} \\
&= \left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+1)(l_{13}+l_{43}+3)}{l_{13}-l_{23}+1}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-2, & -m_{13} & , m_{43}+1 \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{11} \\
&\quad - \left|\frac{(l_{23}-l_{11}-1)(l_{43}-l_{31}+3)(l_{13}-l_{23})}{l_{13}+l_{43}+4}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}-1, & m_{43}+2 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{20} \\
&\quad + \frac{(l_{23}+l_{43}+3)|2(l_{23}-l_{11}-1)(l_{13}-l_{23})(l_{13}+l_{31}+1)(l_{13}+l_{43}+3)|^{1/2}}{2(l_{13}-l_{23}+1)(l_{13}+l_{43}+4)} \\
&\quad \times \begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}^{4}, \qquad (202)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13} & , m_{43}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}^{4} \\
&= (l_{13}+l_{43}+4)\left|\frac{2(l_{13}-l_{11})(l_{13}+l_{31}+1)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13} & , m_{43} \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\quad - \left|\frac{2(l_{13}-l_{11})(l_{43}-l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}-1, & m_{43}+1 \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\quad + (l_{13}-l_{23}+1)\left|\frac{2(l_{23}-l_{11})(l_{43}-l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}, & -m_{13}-1, & m_{43}+1 \\ m_{11} & , 0 , & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (203)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-2, & -m_{13} & , m_{43}+1 \\ m_{11}, & 0 & , m_{31} & , 0 \end{bmatrix}_{11} \\
&= -(l_{23}+l_{43}+3)\left|\frac{(l_{23}-l_{11}-1)(l_{13}+l_{31}+1)}{(l_{13}-l_{23}+1)(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13} & , m_{43} \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\quad + (l_{23}-l_{13})\left|\frac{(l_{23}-l_{11}-1)(l_{43}-l_{31}+2)}{(l_{13}-l_{23}+1)(l_{13}+l_{43}+3)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}-1, & m_{43}+1 \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (204)
\end{aligned}
$$

$$
\begin{aligned}
&e_{23}\begin{bmatrix} m_{13} & , m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}-1, & -m_{13}-1, & m_{43}+2 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{20} \\
&= -(l_{13}+l_{43}+3)\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+4)}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}-1, & m_{43}+1 \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{10} \\
&\quad + (l_{23}+l_{43}+3)\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+4)}\right|^{1/2} \begin{bmatrix} m_{13} & , m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}, & -m_{13}-1, & m_{43}+1 \\ m_{11} & , 0 , & m_{31}-1, & 0 \end{bmatrix}_{10}, \qquad (205)
\end{aligned}
$$

$$e_{23}\left[\begin{matrix} m_{13}, & m_{23} \;\;, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13}-1, & m_{43}+1 \\ m_{11}, & 0 \;\;, & m_{31} \;\;, & 0 \end{matrix}\right]_{10}$$

$$= -(l_{23}+l_{43}+3)\left|\frac{(l_{23}-l_{11})(l_{13}+l_{31}+2)}{(l_{13}-l_{23})(l_{13}+l_{43}+3)}\right|^{1/2}\left[\begin{matrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{11}, & 0 \;, & m_{31}-1, & 0 \end{matrix}\right]_{00}, \tag{206}$$

$$e_{23}\left[\begin{matrix} m_{13}, & m_{23} \;\;, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}-1, & -m_{13} \;\;, & m_{43} \\ m_{11}, & 0 \;\;, & m_{31} \;\;, & 0 \end{matrix}\right]_{10}$$

$$= \left|\frac{(l_{23}-l_{11})(l_{43}-l_{31}+1)(l_{13}-l_{23})}{l_{13}+l_{43}+3}\right|^{1/2}\left[\begin{matrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{11}, & 0 \;, & m_{31}-1, & 0 \end{matrix}\right]_{00}, \tag{207}$$

$$e_{23}\left[\begin{matrix} m_{13} \;\;, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}-1, & m_{23}, & -m_{13}-1, & m_{43}+1 \\ m_{11} \;\;, & 0 \;, & m_{31} \;\;, & 0 \end{matrix}\right]_{10}$$

$$= -\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+2)(l_{13}+l_{43}+3)}{l_{13}-l_{23}}\right|^{1/2}\left[\begin{matrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{11}, & 0 \;, & m_{31}-1, & 0 \end{matrix}\right]_{00}, \tag{208}$$

$$e_{23}\left[\begin{matrix} m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{13}, & m_{23}, & -m_{13}-1, & m_{43} \\ m_{11}, & 0 \;, & m_{31} \;\;, & 0 \end{matrix}\right]_{00} = 0. \tag{209}$$

Proposition 18: Replacing in (190) the subspaces

$$\boxed{-1,0,1,0}_{10}\;,\;\boxed{-1,0,0,1}_{10}\;,\;\boxed{0,-1,0,1}_{10}\;,\;\boxed{-1,-1,0,2}_{20}\;,$$
$$\boxed{-2,0,1,1}_{11}\;,\;\boxed{-2,-1,2,1}_{21}\;,\;\boxed{-2,-1,1,2}_{21}\;,\;\boxed{-1,-2,1,2}_{21}\;,$$
$$\boxed{-1,-1,1,1}^{4}_{\rm inv}\;,\;\boxed{-1,-1,1,1}^{4} \tag{210}$$

by zero, one obtains the decomposition of the class 4 indecomposible module $W([m_{13},m_{13},-m_{13}-1,-m_{13}-1])$ with respect to $\mathrm{gl}(2)_l\oplus\mathrm{gl}(2)_r$. Replacing in (194)–(209) the basis vectors of the subspaces (210) by zero, one obtains the transformations of the nontypical module $W_4([m_{13},m_{13},-m_{13}-1,-m_{13}-1])$ under the action of e_{23} and e_{32}.

Proposition 19: Let $W([m_{13},m_{23},-m_{13}-1,m_{43}])$ be a class 4 indecomposible induced gl(2/2) module, for which

$$m_{13}=m_{23} \text{ and } -m_{13}-1>m_{43}\,. \tag{211}$$

The structure of this module with respect to $\mathrm{gl}(2)_l\oplus\mathrm{gl}(2)_r$ reads:

$$\begin{aligned} W([m]) &= W([m_{13},m_{13},-m_{13}-1,m_{43}]) \\ &= \boxed{0,0,0,0}_{00} \\ &\quad\oplus\boxed{0,-1,1,0}_{10}\quad\oplus\boxed{0,-1,0,1}_{10} \\ &\quad\oplus\boxed{-1,-1,2,0}_{20}\quad\oplus\boxed{-1,-1,1,1}_{20}\quad\oplus\boxed{-1,-1,0,2}_{20}\quad\oplus\boxed{0,-2,1,1}_{11} \\ &\quad\oplus\boxed{-1,-2,2,1}_{21}\quad\oplus\boxed{-1,-2,1,2}_{21} \\ &\quad\oplus\boxed{-2,-2,2,2}_{22}\quad, \end{aligned} \tag{212}$$

where

$$\boxed{-1,-1,0,2}_{20}\quad=0, \text{ if } -m_{13}=m_{43}+2\,. \tag{213}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{\qquad\qquad}$ terms in (212) (= from I_3) by zero, m_{23} by m_{13} and m_{33} by $-m_{13}-1$, one obtains the transformations of the nontypical module, corresponding to the case (211), under the action of the superalgebra gl(2/2).

Proposition 20: Let $W([m])=W([m_{13},m_{23},-m_{13}-1,m_{43}])$ be a class 4 indecomposible induced gl(2/2) module, for which

$$m_{13}>m_{23} \text{ and } -m_{13}-1=m_{43}\,. \tag{214}$$

The structure of this module with respect to $\mathrm{gl}(2)_l\oplus\mathrm{gl}(2)_r$ reads:

$$
\begin{aligned}
&W([m_{13},m_{23},-m_{13}-1,-m_{13}-1])\\
&\quad=\boxed{0,0,0,0}_{00}\\
&\qquad\oplus\boxed{0,-1,1,0}_{10}\quad\oplus\boxed{-1,0,1,0}_{10}\\
&\qquad\oplus\boxed{-1,-1,2,0}_{20}\quad\oplus\boxed{0,-2,1,1}_{11}\quad\oplus\boxed{-1,-1,1,1}_{11}\quad\oplus\boxed{-2,0,1,1}_{11}\\
&\qquad\oplus\boxed{-1,-2,2,1}_{21}\quad\oplus\boxed{-2,-1,2,1}_{21}\\
&\qquad\oplus\boxed{-2,-2,2,2}_{22}\quad,
\end{aligned}
\tag{215}
$$

where

$$\boxed{-2,0,1,1}_{11}\quad=0,\ \text{if}\ m_{13}=m_{23}+1\,. \tag{216}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{}$ terms in (215) (= from I_4) by zero, m_{43} by $-m_{13}-1$ and m_{33} by $-m_{13}-1$, one obtains the transformations of the nontypical module, corresponding to the case (214), under the action of the Lie superalgebra.

E. The class 5 nontypical representations

In all nontypical cases considered so far, the maximal invariant subspaces I_i, $i=1,2,3,4$, were irreducible. In the class 5 induced modules this is no more the case. The maximal invariant subspaces are indecomposible; each I_5 contains several invariant subspaces. We first recall that the class 5 induced modules $W([m])$ are defined with the equations

$$
\begin{aligned}
&l_{13}+l_{43}+3=0\Leftrightarrow m_{43}=-m_{13}\,,\\
&l_{23}+l_{33}+3=0\Leftrightarrow m_{33}=-m_{23}\,.
\end{aligned}
\tag{217}
$$

The induced modules from this class have signatures $[m]=[m_{13},m_{23},-m_{23},-m_{13}]$, i.e.,

$$W([m])=W([m_{13},m_{23},-m_{23},-m_{13}])\,. \tag{218}$$

Proposition 21: Consider a class 5 indecomposible induced gl(2/2) module (218), whose signature satisfies, in addition, the condition

$$m_{13}>m_{23}+1\,. \tag{219}$$

Introduce as in the previous cases a new basis in the $\mathrm{gl}(2)_l\oplus\mathrm{gl}(2)_r$ reducible module

$$\boxed{-1,-1,1,1}_{20}\quad\oplus\ \boxed{-1,-1,1,1}_{11}\quad, \tag{220}$$

setting

$$
\begin{aligned}
[m_{11},m_{31}]^5\equiv&\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1\\ m_{11} & , 0 & , m_{31} & , 0\end{bmatrix}^5\\
=&\left|\frac{l_{13}-l_{23}-1}{l_{13}-l_{23}+1}\right|^{1/2}\left\{\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1\\ m_{11} & , 0 & , m_{31} & , 0\end{bmatrix}_{11}\right.\\
&\left.-\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1\\ m_{11} & , 0 & , m_{31} & , 0\end{bmatrix}_{20}\right\},
\end{aligned}
\tag{221}
$$

$$
\begin{aligned}
[m_{11},m_{31}]^5_{\mathrm{inv}}=&\left|\frac{l_{13}-l_{23}-1}{l_{13}-l_{23}+1}\right|^{1/2}\\
&\times\left\{\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1\\ m_{11} & , 0 & , m_{31} & , 0\end{bmatrix}_{11}\right.\\
&\left.+\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13}\\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1\\ m_{11} & , 0 & , m_{31} & , 0\end{bmatrix}_{20}\right\}.
\end{aligned}
\tag{222}
$$

Construct two new irreducible $\mathrm{gl}(2)_l\oplus\mathrm{gl}(2)_r$ modules with a signature $[m_{13}-1,m_{23}-1,-m_{23}+1,-m_{13}+1]$:

$$
\begin{aligned}
&V^5\equiv V^5([m_{13}-1,m_{23}-1,-m_{23}+1,-m_{13}+1])\\
&\quad=\text{lin.env.}\,\{[m_{11},m_{31}]^5\,|\,m_{13}-m_{11}-1,m_{11}-m_{23}+1,-m_{23}-m_{31}+1,m_{31}+m_{13}-1\in\mathbb{Z}_+\}\,,
\end{aligned}
\tag{223}
$$

$$\begin{aligned} V^5_{\rm inv} &\equiv V^5_{\rm inv}([m_{13}-1,m_{23}-1,-m_{23}+1,-m_{13}+1]) \\ &= \text{lin.env. } \{[m_{11},m_{31}]^5_{\rm inv} \,|\, m_{13}-m_{11}-1, m_{11}-m_{23}+1, -m_{23}-m_{31}+1, m_{31}+m_{13}-1 \in \mathbb{Z}_+\}\,, \end{aligned} \tag{224}$$

Then [see (63), (64)]

$$V^5_{\rm inv} \equiv V^5_{\rm inv}([m_{13}-1,m_{23}-1,-m_{23}+1,-m_{13}+1]) = \boxed{-1,-1,1,1}^5_{\rm inv}\,, \tag{225}$$

$$V^5([m_{13}-1,m_{23}-1,-m_{23}+1,-m_{13}+1]) = \boxed{-1,-1,1,1}^5\,. \tag{226}$$

The decomposition of $W([m]) = W([m_{13},m_{23},-m_{23},-m_{13}])$ into a direct sum of irreducible $\text{gl}(2)_l \oplus \text{gl}(2)_r$ modules reads

$$\begin{aligned} W([m]) &= W([m_{13},m_{23},-m_{23},-m_{13}]) \\ &= \boxed{0,0,0,0}_{00} \\ &\quad \oplus \boxed{-1,0,1,0}_{10} \oplus \boxed{0,-1,1,0}_{10} \oplus \boxed{-1,0,0,1}_{10} \oplus \boxed{0,-1,0,1}_{10} \\ &\quad \oplus \boxed{-1,-1,2,0}_{20} \oplus \boxed{-1-1,0,2}_{20} \oplus \boxed{-2,0,1,1}_{11} \oplus \boxed{0,-2,1,1}_{11} \\ &\quad \oplus \boxed{-1,-1,1,1}^5_{\rm inv} \oplus \boxed{-1,-1,1,1}^5 \\ &\quad \oplus \boxed{-2,-1,2,1}_{21} \oplus \boxed{-1,-2,2,1}_{21} \oplus \boxed{-2,-1,1,2}_{21} \oplus \boxed{-1,-2,1,2}_{21} \\ &\quad \oplus \boxed{-2,-2,2,2}_{22}\,. \end{aligned} \tag{227}$$

The maximal gl(2/2) invariant subspace I_5 is a sum of all $\boxed{}$ terms in (227). It is indecomposible and contains the following invariant subspaces:

$$\begin{aligned} I_5^0 = {} & \boxed{-1,-1,1,1}^5_{\rm inv} \\ & \oplus \boxed{-2,-1,2,1}_{21} \oplus \boxed{-1,-2,1,2}_{21} \\ & \oplus \boxed{-2,-2,2,2}_{22}\,, \end{aligned} \tag{228}$$

$$\begin{aligned} I_5^1 = I_5^0 & \oplus \boxed{0,-1,1,0}_{10} \\ & \oplus \boxed{-1,-1,2,0}_{20} \oplus \boxed{0,-2,1,1}_{11} \\ & \oplus \boxed{-1,-2,2,1}_{21}\,, \end{aligned} \tag{229}$$

$$\begin{aligned} I_5^2 = I_5^0 & \oplus \boxed{-1,0,0,1}_{10} \\ & \oplus \boxed{-1,-1,0,2}_{20} \oplus \boxed{-2,0,1,1}_{11} \\ & \oplus \boxed{-2,-1,1,2}_{21}\,. \end{aligned} \tag{230}$$

The subspace I_5^0 is irreducible and nontypical, I_5^0 has a signature

$$[m_{13}-1,m_{23}-1,-m_{23}+1,-m_{13}+1]\,. \tag{231}$$

The subspace I_5^1 and I_5^2 are indecomposible; each one contains as a maximal invariant subspace I_5^0. The factor spaces I_5^1/I_5^0 and I_5^2/I_5^0 carry nontypical (irreducible) representations of the LS gl(2/2), namely

$$I_5^1/I_5^0 \text{ has a signature } [m_{13},m_{23}-1,-m_{23}+1,-m_{13}]\,, \tag{232}$$

$$I_5^2/I_5^0 \text{ has a signature } [m_{13}-1,m_{23},-m_{23},-m_{13}+1]\,. \tag{233}$$

Replacing everywhere in (22)–(37) [or in I, (3.56)–(3.93)] the basis vectors from all $\boxed{}$ terms in (227) (= from I_5) by zero, m_{33} by $-m_{23}$ and m_{43} by $-m_{13}$, one obtains the transformations of the nontypical module $W([m_{13},m_{23},-m_{23},-m_{13}])$ under the action of the Lie superalgebra.

Transformations under the action of e_{32}:

$$\begin{aligned} e_{32}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23}, & -m_{13} \\ m_{13}, & m_{23}, & -m_{23}, & -m_{13} \\ m_{11}, & 0\ , & m_{31}, & 0 \end{bmatrix}_{00} &= -\left|\frac{(l_{13}-l_{11})(l_{13}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}-l_{23})}\right|^{1/2} \begin{bmatrix} m_{13} & , & m_{23}, & -m_{23} & , & -m_{13} \\ m_{13}-1, & & m_{23}, & -m_{23}+1, & & -m_{13} \\ m_{11} & , & 0\ , & m_{31}+1, & & 0 \end{bmatrix}_{10} \\ &\quad -\left|\frac{(l_{23}-l_{11})(l_{23}+l_{31}+3)}{(l_{13}-l_{23})(l_{13}-l_{23})}\right|^{1/2} \begin{bmatrix} m_{13}, & m_{23} & , & -m_{23} & , & -m_{13} \\ m_{13}, & m_{23}-1, & & -m_{23} & , & -m_{13}+1 \\ m_{11}, & 0 & , & m_{31}+1, & & 0 \end{bmatrix}_{10}\,, \end{aligned} \tag{234}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & -m_{13} \\ m_{11} & , 0 , & m_{31} & , 0 \end{bmatrix}_{10}$$

$$= -\frac{|2(l_{23}-l_{11})(l_{23}+l_{31}+2)|^{1/2}}{2(l_{13}-l_{23}-1)}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}^{5}, \tag{235}$$

$$e_{32}\begin{bmatrix} m_{13}, & m_{23} & , -m_{23}, & -m_{13} \\ m_{13}, & m_{23}-1, & -m_{23}, & -m_{13}+1 \\ m_{11}, & 0 & , m_{31}, & 0 \end{bmatrix}_{10}$$

$$= \frac{|2(l_{13}-l_{11})(l_{13}+l_{31}+2)|^{1/2}}{2(l_{13}-l_{23}-1)}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}+1, & 0 \end{bmatrix}^{5}, \tag{236}$$

$$e_{32}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}^{5} = 0. \tag{237}$$

Transformations under the action of e_{23}:

$$e_{23}\begin{bmatrix} m_{13} & , m_{23} & , -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & -m_{23}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}^{5}$$

$$= \frac{(l_{13}-l_{23}-1)|2(l_{13}-l_{11})(l_{13}+l_{31}+1)|^{1/2}}{l_{13}-l_{23}}\begin{bmatrix} m_{13}, & m_{23} & , -m_{23} & , -m_{13} \\ m_{13}, & m_{23}-1, & -m_{23} & , -m_{13}+1 \\ m_{11}, & 0 & , m_{31}-1, & 0 \end{bmatrix}_{10}$$

$$+ \frac{(l_{13}-l_{23}-1)|2(l_{23}-l_{11})(l_{23}+l_{31}+1)|^{1/2}}{l_{13}-l_{23}}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & -m_{13} \\ m_{11} & , 0 , & m_{31}-1, & 0 \end{bmatrix}_{10} \tag{238}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23} & , -m_{23}, & -m_{13} \\ m_{13}, & m_{23}-1, & -m_{23}, & -m_{13}+1 \\ m_{11}, & 0 & , m_{31}, & 0 \end{bmatrix}_{10}$$

$$= |(l_{23}-l_{11})(l_{23}+l_{31}+2)|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23} & , -m_{13} \\ m_{13}, & m_{23}, & -m_{23} & , -m_{13} \\ m_{11}, & 0 , & m_{31}-1, & 0 \end{bmatrix}_{00}, \tag{239}$$

$$e_{23}\begin{bmatrix} m_{13} & , m_{23}, & -m_{23} & , -m_{13} \\ m_{13}-1, & m_{23}, & -m_{23}+1, & -m_{13} \\ m_{11} & , 0 , & m_{31} & , 0 \end{bmatrix}_{10}$$

$$= -|(l_{13}-l_{11})(l_{13}+l_{31}+2)|^{1/2}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23} & , -m_{13} \\ m_{13}, & m_{23}, & -m_{23} & , -m_{13} \\ m_{11}, & 0 , & m_{31}-1, & 0 \end{bmatrix}_{00}, \tag{240}$$

$$e_{23}\begin{bmatrix} m_{13}, & m_{23}, & -m_{23}, & -m_{13} \\ m_{13}, & m_{23}, & -m_{23}, & -m_{13} \\ m_{11}, & 0 , & m_{31}, & 0 \end{bmatrix}_{00} = 0. \tag{241}$$

Proposition 22: Consider class 5 indecomposible induced 16-dimensional gl(2/2) modules (218), whose signatures satisfy the condition

$$m_{13}=m_{23} \Rightarrow m_{33}=m_{43}= -m_{13}. \tag{242}$$

The decomposition of

$$W([m]) = W([m_{13}, m_{13}, -m_{13}, -m_{13}])$$

into a direct sum of irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules reads

$$
\begin{aligned}
W([m]) &= W([m_{13},m_{13},-m_{13},-m_{13}])\\
&= \boxed{0,0,0,0}_{00}\\
&\quad \oplus \boxed{0,-1,1,0}_{10}\\
&\quad \oplus \boxed{-1,-1,2,0}_{20} \quad \oplus \boxed{0,-2,1,1}_{11}\\
&\quad \oplus \boxed{-1,-2,2,1}_{21}\\
&\quad \oplus \boxed{-2,-2,2,2}_{22} \quad .
\end{aligned} \tag{243}
$$

The maximal gl(2/2) invariant subspace I_5 is of dimension 15 and is a sum of all $\boxed{}$ terms in (243). It is indecomposible and contains as an invariant subspace the one-dimensional nontypical module $\boxed{-2,-2,2,2}_{22}$. The 14-dimensional factor module $I_5/\ \boxed{-2,-2,2,2}_{22}$ is also nontypical and

$$
I_5/\ \boxed{-2,-2,2,2}_{22} \text{ has a signature } [m_{13},m_{13}-1,-m_{13}+1,-m_{13}] . \tag{244}
$$

The factor module

$$
W_5([m_{13},m_{13},-m_{13},-m_{13}]) = W([m_{13},m_{13},-m_{13},-m_{13}])/I_5
$$

carries a trivial one-dimensional representation of gl(2/2).

Proposition 23: Consider class 5 indecomposible induced 64-dimensional gl(2/2) modules (218), whose signatures satisfy the condition

$$
m_{13}=m_{23}+1 .
$$

The decomposition of $W([m_{13},m_{13}-1,-m_{13}+1,-m_{13}])$ into a direct sum of irreducible $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ modules reads

$$
\begin{aligned}
W([m_{13},m_{13}&-1,-m_{13}+1,-m_{13}])\\
= \boxed{0,0,0,0}_{00}&\\
\oplus \boxed{-1,0,1,0}_{10} &\quad \oplus \boxed{0,-1,1,0}_{10} \quad \oplus \boxed{-1,0,0,1}_{10}\\
\oplus \boxed{0,-1,0,1}_{10} &\quad \oplus \boxed{-1,-1,2,0}_{20}\bullet \quad \oplus \boxed{0,-2,1,1}_{11}\\
\oplus \boxed{-1,-1,1,1}^{5}_{\mathrm{inv}} &\quad \oplus \boxed{-1,-1,1,1}^{5}\\
\oplus \boxed{-2,-1,2,1}_{21} &\quad \oplus \boxed{-1,-2,2,1}_{21} \quad \oplus \boxed{-2,-1,1,2}_{21}\\
\oplus \boxed{-1,-2,2,1}_{21} &\quad \oplus \boxed{-2,-2,2,2}_{22} \cdot
\end{aligned} \tag{245}
$$

The maximal gl(2/2) invariant subspace I_5 is of dimension 50 and is given with the sum of all $\boxed{}$ terms in (245). It is indecomposible and contains the following invariant subspaces:

$$
\begin{aligned}
I_5^0 = \boxed{-1,-1,1,1}^{5}_{\mathrm{inv}}&\\
\oplus \boxed{-2,-1,2,1}_{21} &\quad \oplus \boxed{-2,-1,1,2}_{21} \quad \oplus \boxed{-1,-2,1,2}_{21}\\
\oplus \boxed{-2,-2,2,2}_{22} &\quad ,
\end{aligned} \tag{246}
$$

$$
\begin{aligned}
I_5^1 = I_5^0 \oplus \boxed{0,-1,1,0}_{10}&\\
\oplus \boxed{-1,-1,2,0}_{20} &\quad \oplus \boxed{0,-2,1,1}_{11}\\
\oplus \boxed{-1,-2,2,1}_{21} &\quad ,
\end{aligned} \tag{247}
$$

$$
I_5^2 = I_5^0 \oplus \boxed{-1,0,0,1}_{10} \quad . \tag{248}
$$

The module I_5^0 is a 15-dimensional nondecomposible subspace. It contains a 14-dimensional nontypical subspace with a signature

$$
[m_{13}-1,m_{13}-2,-m_{13}+2,-m_{13}+1] . \tag{249}
$$

The subspaces I_5^1 and I_5^2 are indecomposible of dimensions 49 and 16, respectively; they contain as a maximal invariant subspace I_5^0. Each of the factor spaces I_5^1/I_5^0 and I_5^2/I_5^0 carries a nontypical (irreducible) representation of the LS gl(2/2), namely,

$$
I_5^1/I_5^0 \text{ has a signature } [m_{13},m_{13}-2,-m_{13}+2,-m_{13}] , \tag{250}
$$

$$
I_5^2/I_5^0 \text{ has a signature } [m_{13}-1,m_{13}-1,-m_{13}+1, -m_{13}+1] . \tag{251}
$$

Replacing everywhere in Eqs. (234)–(241) m_{23} by $m_{13}-1$, one obtains the transformations of the nontypical 14-dimensional modules

$$
\begin{aligned}
W_5([m_{13},m_{13}-1,-m_{13}+1,-m_{13}])\\
= W([m_{13},m_{13}-1,-m_{13}+1,-m_{13}])/I_5
\end{aligned} \tag{252}
$$

under the action of the odd generators e_{32} and e_{23}.

IV. FINITE-DIMENSIONAL IRREDUCIBLE REPRESENTATIONS OF gl(2/2)

Denote by $\mathfrak{F}$ the class of finite-dimensional irreducible gl(2/2) modules $W([m_{13},m_{23},m_{33},m_{43}])$, which were determined in I and in the present paper. The modules from this class are labeled with all possible complex numbers

$$
\begin{aligned}
&m_{13},m_{23},m_{33},m_{43}, \text{ such that } m_{13}-m_{23}\in\mathbf{Z}_+ ,\\
&m_{33}-m_{43}\in\mathbf{Z}_+ .
\end{aligned} \tag{253}
$$

If $m_{13},m_{23},m_{33},m_{43}$ obey one of the conditions (40)–(44), then the corresponding module is nontypical: it carries a nontypical representation of gl(2/2). The transformations of all nontypical modules under the action of gl(2/2) are completely defined from the Eqs. (22)–(27) and the action of the odd generators e_{23}, e_{32} [see Propositions 7–23]. If $m_{13},m_{23},m_{33},m_{43}$ satisfy none of the conditions (40)–(44), then the corresponding module is typical. The transformations of these modules are determined from Eqs. (22)–(37). In all cases the numbers $m_{13},m_{23},m_{33},m_{43}$ give the signature of the gl(2/2) fidirmod $W([m_{13},m_{23},m_{33},m_{43}])$, i.e., these numbers are the coordinates of the highest weight Λ in the basis e^1,e^2,e^3,e^4 [see (9)],

$$
\Lambda = m_{13}e^1 + m_{23}e^2 + m_{33}e^3 + m_{43}e^4 . \tag{254}
$$

The highest weight vector corresponding to $W([m_{13},m_{23},m_{33},m_{43}])$ is

$$
x_\Lambda = \begin{bmatrix} m_{13}, & m_{23}, & m_{33}, & m_{43}\\ m_{13}, & m_{23}, & m_{33}, & m_{43}\\ m_{13}, & 0\ , & m_{33}, & 0 \end{bmatrix}
\in V_{00}([m_{13},m_{23},m_{33},m_{43}]) \tag{255}
$$

and it is simultaneously a highest weight vector of the $\mathrm{gl}(2)_l \oplus \mathrm{gl}(2)_r$ fidirmod $V_{00}([m_{13},m_{23},m_{33},m_{43}])$. Each coordinate m_{i3} of Λ is an eigenvalue of e_{ii} on x_Λ,

$$
e_{ii}x_\Lambda = m_{i3}x_\Lambda , \quad i=1,2,3,4 . \tag{256}
$$

We now proceed to show that the class $\mathfrak{F}$ contains all fidirmods of the LS gl(2/2).

Proposition 24: Let W be a finite-dimensional irreducible module of gl(2/2). Then $W\in\mathfrak{F}$.

Proof: The Cartan subalgebra

$$H = \text{lin.env.}\{e_{11},e_{22},e_{33},e_{44}\} \tag{257}$$

of gl(2/2) is a Cartan subalgebra of $\text{gl}(2)_l \oplus \text{gl}(2)_r$. The basis $e_1,...,e_N$ in any finite-dimensional $\text{gl}(2)_l \oplus \text{gl}(2)_r$ module and in particular in W can be chosen in such a way that H is diagonal,

$$he_i = \lambda_i(h)e_i\ ,\ \lambda_i \in H^*\ ,\ \forall\ h\in H\ . \tag{258}$$

Take any two elements $x, y\in W$,

$$x = \alpha_1 e_1 + \cdots + \alpha_N e_N\ . \tag{259}$$

The irreducibility of W implies that there exists a polynomial Q of the gl(2/2) generators, such that

$$y = Qx\ . \tag{260}$$

According to the Poincaré–Birkhoff–Witt theorem[29] Q can be represented as

$$Q = \sum_{j=1}^{M} Q_j H_j\ , \tag{261}$$

where Q_j (resp. H_j) is a monomial of the gl(2/2) root vectors e_{ij}, $i\neq j = 1,2,3,4$ (resp. of $e_{11},e_{22},e_{33},e_{44}$). From (258)–(261) one derives that

$$y = Qx = Q_0 x\ , \tag{262}$$

where Q_0 is a polynomial only of e_{ij}, $i\neq j = 1,2,3,4$. These generators are also root vectors of sl(2/2). Therefore, (262) yields that W is also an sl(2/2) fidirmod. Each sl(2/2) fidirmod has a unique (up to a multiplicative constant) highest weight.[2] Let $x_\Lambda \in W$ be the sl(2/2) highest weight vector, i.e. [see (4), (7)],

$$e_{ij}x_\Lambda = 0\ \forall\ i<j = 1,2,3,4\ , \tag{263}$$

$$h_i x_\Lambda = \alpha_i x_\Lambda\ . \tag{264}$$

Consider the vector

$$y = e_4 x_\Lambda\ ,\quad e_4 = e_{11} + e_{22} + e_{33} + e_{44} \in \text{gl}(2/2)\ . \tag{265}$$

From the supercommutation relations and (263) one concludes that y is also an sl(2/2) highest weight vector. Therefore, $y = \alpha x_\Lambda$, i.e., $e_4 x_\Lambda = \alpha x_\Lambda$. Thus, x_Λ is an eigenvector of the Cartan subalgebra $H = \text{lin.env.}\{e_4,h_1,h_2,h_3\} = \text{lin.env.}\{e_{11},e_{22},e_{33},e_{44}\}$ of gl(2/2), i.e.,

$$e_{ii}x_\Lambda = m_{i3}x_\Lambda\ ,\quad i = 1,2,3,4\ . \tag{266}$$

From (263) and (266) it follows that x_Λ is a gl(2/2) highest weight vector of W with a highest weight of

$$\Lambda = m_{13}e^1 + m_{23}e^2 + m_{33}e^3 + m_{43}e^4\ . \tag{267}$$

Hence W is a gl(2/2) fidirmod with a signature $[m_{13},m_{23},m_{33},m_{43}]$. Let U_0 be the universal enveloping algebra of $\text{gl}(2)_l \oplus \text{gl}(2)_r$. Then $V_{00} = U_0 x_\Lambda$ is an irreducible $\text{gl}(2)_l \oplus \text{gl}(2)_r$ module with the same highest weight vector x_Λ and with the same signature $[m_{13},m_{23},m_{33},m_{43}]$. The $\text{gl}(2)_l \oplus \text{gl}(2)_r$ module V_{00} is a tensor product

$$V_{00} = V_l([m_{13},m_{23}]) \otimes V_r([m_{33},m_{43}]) \tag{268}$$

of gl(2) fidirmods $V_l([m_{13},m_{23}])$ and $V_r([m_{33},m_{43}])$, each one with a signature $[m_{13},m_{23}]$ and $[m_{33},m_{43}]$, respectively. It is well known that a gl(2) irreducible module $V([m_{12},m_{22}])$ with a signature $[m_{12},m_{22}]$ is finite dimensional if and only if m_{12} and m_{22} are complex numbers, such that $m_{12} - m_{22}\in\mathbb{Z}_+$. Therefore,

$$m_{13},m_{23},m_{33},m_{43}\in\mathbb{C}\ ,\quad m_{13} - m_{23}\in\mathbb{Z}_+,\quad m_{33} - m_{43}\in\mathbb{Z}_+\ . \tag{269}$$

We have shown that any finite-dimensional irreducible gl(2/2) module W is a module with a signature $[m_{13},m_{23},m_{33},m_{43}]$, for which (269) holds. Hence [see (253)] $W\in\mathfrak{F}$. ■

According to Proposition 24, every gl(2/2) fidirmod $W([m_{13},m_{23},m_{33},m_{43}])$ is also an sl(2/2) fidirmod. The inverse also follows from the proof of Proposition 24: every sl(2/2) fidirmod W can be extended to (several inequivalent) gl(2/2) modules, simply setting $e_4 x_\Lambda = \alpha x_\Lambda$ for the highest weight vector x_Λ. From (4), (7), and (256) one concludes that the gl(2/2) fidirmod $W([m_{13},m_{23},m_{33},m_{43}])$, considered as a sl(2/2) fidirmod, corresponds to labels (see Proposition 1)

$$\alpha_1 = m_{13} - m_{23},\quad \alpha_2 = m_{23} - m_{33},\quad \alpha_3 = m_{33} - m_{43}\ . \tag{270}$$

Therefore, whenever the gl(2/2) labels $m_{13},m_{23},m_{33},m_{43}$ take all values consistent with (253), then the triple $(\alpha_1,\alpha_2,\alpha_3)$ runs over all labels for the sl(2/2) fidirmods. Thus we have the following result.

Proposition 25: The sl(2/2) modules from the class $\mathfrak{F}$ contain all finite-dimensional irreducible sl(2/2) modules.

From the results obtained so far one can go further and write down all irreducible representations of the basic LS $A(1/1)$ [see I, (1.7)] in a matrix form. We shall return to this problem elsewhere.

ACKNOWLEDGMENT

The authors are thankful to Professor H. D. Doebner for the kind hospitality at the Arnold Sommerfeld Institute for Mathematical Physics, where the main part of the results contained in this paper have been obtained.

APPENDIX: PROOF OF PROPOSITION 7

First of all we observe that the statements (79) and (80) follow immediately from the decomposition (49).

The subspace $V_{22}([m_{13} - 2,m_{23} - 2,m_{33} + 2, - m_{13} + 2])$ is in all cases different from zero and according to Proposition 4

$$V_{22}([m_{13} - 2,m_{23} - 2,m_{33} + 2, - m_{13} + 2]) = \boxed{-2,-2,2,2}_{22} \subset I_1\ .$$

Therefore, also

$$e_{23}V_{22}([m_{13} - 2,m_{23} - 2,m_{33} + 2, - m_{13} + 2]) \subset I_1\ .$$

Equation (33) [see also I, (3.77)] yields

$$x(m_{11},m_{31}) \equiv e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-2, && m_{23}-2, && m_{33}+2, && -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{22}$$

$$= a_1(m_{11},m_{31})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-2, && m_{23}-1, && m_{33}+2, && -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, && 0 \end{bmatrix}_{21}$$

$$+ a_2(m_{11},m_{31})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-2, && m_{23}-1, && m_{33}+1, && -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31}-1, && 0 \end{bmatrix}_{21}$$

$$+ a_3(m_{11},m_{31})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, && m_{23}-2, && m_{33}+1, && -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31}-1, && 0 \end{bmatrix}_{21} \in I_1, \tag{A1}$$

where the coefficients $a_1(m_{11},m_{31})$, $a_2(m_{11},m_{31})$, and $a_3(m_{11},m_{31})$ depend (apart from m_{13},m_{23},m_{33}, which are fixed numbers) on m_{11} and m_{31}. In particular,

$$a_1(m_{11},m_{31}) = 0, \quad \text{iff } m_{11} = m_{23}-2, \tag{A2}$$

$$a_2(m_{11},m_{31}) = 0, \quad \text{iff at least one of the equalities } m_{11} = m_{23}-2, \quad m_{31} = -m_{13}+2 \text{ hold}, \tag{A3}$$

$$a_3(m_{11},m_{31}) = 0, \quad \text{iff } m_{31} = -m_{13}+2\,. \tag{A4}$$

Setting in (A1) $m_{11} = m_{23}-2$ and taking into account (A2) and (A3), we have

$$x(m_{23}-2,m_{31}) = a_3(m_{23}-2,m_{31})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, && m_{23}-2, && m_{33}+1, && -m_{13}+2 \\ m_{23}-2, && 0 & , & m_{31}-1, && 0 \end{bmatrix}_{21} \in I_1\,. \tag{A5}$$

Since $m_{33} > -m_{13}$, for $m_{31} = m_{33}+2$ we have $m_{31} > -m_{13}+2$. Therefore, [see (A4)] $a_3(m_{23}-2,m_{33}+2) \neq 0$. Thus

$$0 \neq x(m_{23}-2,m_{33}+2) \in I_1 \cap V_{21}([m_{13}-1,m_{23}-2,m_{33}+1,-m_{13}+2]) \tag{A6}$$

and according to Proposition 3

$$V_{21}([m_{13}-1,m_{23}-2,m_{33}+1,-m_{13}+2]) \equiv \boxed{-1,-2,1,2}_{21} \subset I_1\,. \tag{A7}$$

In the case $m_{31} = -m_{13}+2$ (A1) reduces to

$$x(m_{11},-m_{13}+2) = a_1(m_{11},-m_{13}+2)\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-2, && m_{23}-1, && m_{33}+2, && -m_{13}+1 \\ m_{11} & , & 0 & , & -m_{13}+1, && 0 \end{bmatrix}_{21} \in I_1\,. \tag{A8}$$

Since $m_{13} > m_{23}$, if $m_{11} = m_{13}-2$, then $m_{11} > m_{23}-2$ and according to (A2) $a_1(m_{13}-2,-m_{13}+2) \neq 0$. Therefore,

$$\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-2, && m_{23}-1, && m_{33}+2, && -m_{13}+1 \\ m_{13}-2, && 0 & , & -m_{13}+1, && 0 \end{bmatrix}_{21}$$

$$\in V_{21}([m_{13}-2,m_{23}-1,m_{33}+2,-m_{13}+1]) \cap I_1\,. \tag{A9}$$

Applying Proposition 3 to (A9), we conclude that

$$V_{21}([m_{13}-2,m_{23}-1,m_{33}+2,-m_{13}+1]) \equiv \boxed{-2,-1,2,1}_{21} \subset I_1\,. \tag{A10}$$

Under the condition (75) $V_{21}([m_{13}-2,m_{23}-1,m_{33}+1,-m_{13}+2]) \neq 0$. For we have that

$$m_{11} > m_{23}-2, \quad m_{31} > -m_{13}+2 \tag{A11}$$

and, therefore [see (A3)], the coefficient $a_2(m_{13}-2,m_{33}+2) \neq 0$. Then Eqs. (A1), (A7), and (A10) yield

$$\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-2, && m_{23}-1, && m_{33}+1, && -m_{13}+2 \\ m_{13}-2, && 0 & , & m_{33}+1, && 0 \end{bmatrix}_{21}$$

$$\in V_{21}([m_{13}-2,m_{23}-1,m_{33}+1,-m_{13}+2]) \cap I_1\,. \tag{A12}$$

Hence (Proposition 3),

$$V_{21}([m_{13}-2,m_{23}-1,m_{33}+1,-m_{13}+2]) \equiv \boxed{-2,-1,1,2}_{21} \subset I_1\,. \tag{A13}$$

Suppose

$$V_{11}([m_{13}-2,m_{23},m_{33}+1,-m_{13}+1]) \neq 0 \Leftrightarrow m_{13}-2 \geqslant m_{23}. \qquad \text{(A14)}$$

Then also

$$V_{21}([m_{13}-2,m_{23}-1,m_{33}+2,-m_{13}+1]) \neq 0. \qquad \text{(A15)}$$

From (34) [or I, (3.73)] and (A10) we compute

$$\begin{aligned}
y_1(m_{11},m_{31}) \equiv & \frac{1}{|l_{13}-l_{11}-1|^{1/2}} e_{23} \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-2, & m_{23}-1, & m_{33}+2, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31} & , 0 \end{bmatrix}_{21} \\
= & -(l_{23}+l_{33}+3) \left| \frac{(l_{23}-l_{11})(l_{13}+l_{31}+1)}{(l_{13}-l_{11}-1)(l_{13}-l_{23}-1)(l_{13}+l_{33}+3)} \right|^{1/2} \\
& \times \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-2, & m_{23} & , m_{33}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{11} \\
& + b_1(m_{31}) \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{11} \\
& + b_2(m_{31}) \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{20} \in I_1, \qquad \text{(A16)}
\end{aligned}$$

where $|l_{13}-l_{11}-1|^{1/2}$ cannot vanish and

$$b_1(m_{31}) = (l_{13}+l_{33}+2) \left| \frac{(l_{13}-l_{23}+1)(l_{13}+l_{31}+1)}{2(l_{13}-l_{23})(l_{13}+l_{33}+3)(l_{13}-l_{23}-1)} \right|^{1/2}, \qquad \text{(A17)}$$

$$b_2(m_{31}) = (l_{13}+l_{33}+4) \left| \frac{(l_{13}+l_{33}+2)(l_{13}+l_{31}+1)}{2(l_{13}-l_{23})(l_{13}+l_{33}+3)(l_{13}+l_{33}+4)} \right|^{1/2}. \qquad \text{(A18)}$$

The coefficients $b_1(m_{31})$ and $b_2(m_{31})$ are independent on m_{11} and

$$b_1(m_{31}) = b_2(m_{31}) = 0 \quad \text{iff}\ m_{31} = -m_{13}+1 \Leftrightarrow l_{13}+l_{31}+1 = 0. \qquad \text{(A19)}$$

Similarly, from (34) and (24) we obtain

$$\begin{aligned}
y_2(m_{11},m_{31}) \equiv & \frac{1}{(l_{13}-l_{11})|l_{23}-l_{11}|^{1/2}} e_{12}e_{23} \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-2, & m_{23}-1, & m_{33}+2, & -m_{13}+1 \\ m_{11}-1, & 0 & , m_{31} & , 0 \end{bmatrix}_{21} \\
= & -\frac{(l_{23}+l_{33}+3)(l_{23}-l_{11}+1)}{l_{13}-l_{11}} \left| \frac{(l_{13}+l_{31}+1)(l_{13}-l_{11}-1)}{(l_{23}-l_{11})(l_{13}-l_{23}-1)(l_{13}+l_{33}+3)} \right|^{1/2} \\
& \times \begin{bmatrix} m_{13} & , m_{23}, & m_{33} & , -m_{13} \\ m_{13}-2, & m_{23}, & m_{33}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{11} \\
& + b_1(m_{31}) \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{11} \\
& + b_2(m_{31}) \begin{bmatrix} m_{13} & , m_{23} & , m_{33} & , -m_{13} \\ m_{13}-1, & m_{23}-1, & m_{33}+1, & -m_{13}+1 \\ m_{11} & , 0 & , m_{31}-1, & 0 \end{bmatrix}_{20} \in I_1, \qquad \text{(A20)}
\end{aligned}$$

where $(l_{13}-l_{11})|l_{23}-l_{11}|^{1/2}\neq 0$. Evaluating $y_1(m_{11},m_{31})-y_2(m_{11},m_{31})$ for

$$m_{11}=m_{13}-2 \Leftrightarrow l_{11}=l_{13}-2\,, \qquad\qquad (\text{A21})$$
$$m_{31}=m_{33}+2 \Leftrightarrow l_{31}=l_{33}+2\,,$$

we have

$$y_1(m_{13}-2,m_{33}+2)-y_2(m_{13}-2,m_{33}+2)$$

$$=\alpha\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-2, & & m_{23}, & m_{33}+1, & & -m_{13}+1 \\ m_{13}-2, & & 0\ \ , & m_{33}+1, & & 0 \end{bmatrix}_{11} \in I_1\,, \qquad (\text{A22})$$

where

$$\alpha=-(l_{23}+l_{33}+3)\left|\frac{l_{13}-l_{23}-2}{l_{13}-l_{23}-1}\right|^{1/2}$$
$$\times\left(1+\frac{l_{13}-l_{23}-3}{2(l_{13}-l_{23}-2)}\right). \qquad (\text{A23})$$

In view of the conditions (40) and (A14) $\alpha\neq 0$. Therefore,

$$\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-2, & & m_{23}, & m_{33}+1, & & -m_{13}+1 \\ m_{13}-2, & & 0\ \ , & m_{33}+1, & & 0 \end{bmatrix}_{11}$$
$$\in V_{11}([m_{13}-2,m_{23},m_{33}+1,-m_{13}+1])\cap I_1\,, \qquad (\text{A24})$$

and applying Proposition 3 we conclude that

$$V_{11}([m_{13}-2,m_{23},m_{33}+1,.-m_{13}+1])$$
$$\equiv \boxed{-2,0,1,1}_{11}\ \subset I_1\,. \qquad (\text{A25})$$

From (A16) and (A25) it follows also that

$$v(m_{11},m_{31})\equiv b_1(m_{31})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{11}$$
$$+b_2(m_{31})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20} \in I_1\,. \qquad (\text{A26})$$

Inserting the expressions for $b_1(m_{31})$ and $b_2(m_{31})$ in (A26) one easily shows that up to a multiple $v(m_{11},m_{31})$ is equal to [see (72)] $[m_{11},m_{31}]^1_{\text{inv}}$, i.e., that

$$V^1_{\text{inv}}([m_{13}-1,m_{23}-1,m_{33}+1,-m_{13}+1])\equiv\boxed{-1,-1,1,1}^1_{\text{inv}}\ \subset I_1\,. \qquad (\text{A27})$$

If

$$V_{20}([m_{13}-1,m_{23}-1,m_{33},-m_{13}+2])\neq 0\,,\ \text{i.e., if}\ m_{33}\geqslant -m_{13}+2, \qquad (\text{A28})$$

then also

$$V_{21}([m_{13}-1,m_{23}-2,m_{33}+1,-m_{13}+2])\neq 0\,. \qquad (\text{A29})$$

From (34) and (A7) we obtain

$$z_1(m_{11},m_{31})\equiv\frac{1}{|l_{33}-l_{31}+2|^{1/2}}e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-2, & & m_{33}+1, & & -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31} & , & 0 \end{bmatrix}_{21}$$
$$=(l_{23}+l_{33}+3)\left|\frac{(l_{23}-l_{11}-1)(l_{13}+l_{31})}{(l_{13}+l_{33}+2)(l_{13}-l_{23})(l_{33}-l_{31}+2)}\right|^{1/2}$$
$$\times\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33} & , & -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20}$$
$$+c_1(m_{11})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{11}$$
$$+c_2(m_{11})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20} \in I_1\,, \qquad (\text{A30})$$

where always $|l_{33}-l_{31}+2|^{1/2}\neq 0$ and

$$c_1(m_{11})=(l_{13}-l_{23}+1)\left|\frac{(l_{13}-l_{23}-1)(l_{23}-l_{11}-1)}{2(l_{13}-l_{23})(l_{13}+l_{33}+3)(l_{13}-l_{23}+1)}\right|^{1/2}, \tag{A31}$$

$$c_2(m_{11})=(l_{13}-l_{23}-1)\left|\frac{(l_{13}+l_{33}+4)(l_{23}-l_{11}-1)}{2(l_{13}-l_{23})(l_{13}+l_{33}+3)(l_{13}+l_{33}+2)}\right|^{1/2}. \tag{A32}$$

The coefficients $c_1(m_{11})$ and $c_2(m_{11})$ are independent on m_{31} and

$$c_1(m_{11})=c_2(m_{11})=0 \quad \text{iff}\, m_{11}=m_{23}-2\,. \tag{A33}$$

Similarly, from (34) and (26) we have

$$\begin{aligned}
z_2(m_{11},m_{31})\equiv{}&\frac{1}{(l_{33}-l_{31}+3)|l_{13}+l_{31}|^{1/2}}\,e_{34}e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-2, & & m_{33}+1, & & -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{21}\\
={}&\frac{(l_{23}+l_{33}+3)(l_{13}+l_{31}-1)}{(l_{33}-l_{31}+3)}\left|\frac{(l_{23}-l_{11}-1)(l_{33}-l_{31}+2)}{(l_{13}+l_{33}+2)(l_{13}-l_{23})(l_{13}+l_{31})}\right|^{1/2}\\
&\times\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33} & , & -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20}\\
&+c_1(m_{11})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{11}\\
&+c_2(m_{11})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20}\in I_1\,,
\end{aligned} \tag{A34}$$

where $(l_{33}-l_{31}+3)|l_{13}+l_{31}|^{1/2}\neq 0$. Setting in (A30) and (A34)

$$\begin{aligned} &m_{11}=m_{13}-1\Leftrightarrow l_{11}=l_{13}-1,\\ &m_{31}=m_{33}+1\Leftrightarrow l_{31}=l_{33}+1, \end{aligned} \tag{A35}$$

we get

$$z_1(m_{13}-1,m_{33}+1)-z_2(m_{13}-1,m_{33}+1)=\beta\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33}, & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}, & -m_{13}+2 \\ m_{13}-1, & & 0 & , & m_{33}, & 0 \end{bmatrix}_{20}. \tag{A36}$$

From (40) and (A28) one concludes that the constant β never vanishes, i.e.,

$$2\beta\equiv(l_{23}+l_{33}+3)\left|\frac{l_{13}+l_{33}+2}{l_{13}+l_{33}+1}\right|^{1/2}\neq 0\,. \tag{A37}$$

Therefore

$$\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33}, & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}, & -m_{13}+2 \\ m_{13}-1, & & 0 & , & m_{33}, & 0 \end{bmatrix}_{20}\in V_{20}([m_{13}-1,m_{23}-1,m_{33},-m_{13}+2])\cap I_1\,. \tag{A38}$$

Hence (Proposition 3),

$$V_{20}([m_{13}-1,m_{23}-1,m_{33},-m_{13}+2])\equiv\boxed{-1,-1,0,2}_{20}\quad\subset I_1\,. \tag{A39}$$

Now from (A30) and (A38) it follows that the maximal invariant subspace contains all possible vectors

$$w(m_{11},m_{31})=c_1(m_{11})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{11}$$

$$+c_2(m_{11})\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}+1, & & -m_{13}+1 \\ m_{11} & , & 0 & , & m_{31}-1, & & 0 \end{bmatrix}_{20} \in I_1\,. \tag{A40}$$

Inserting in (A40) the expressions for $c_1(m_{11})$ and $c_2(m_{11})$ one concludes also in this case that $w(m_{11},m_{31})$ is proportional to $[m_{11},m_{31}]^1_{\rm inv}$, i.e., that $w(m_{11},m_{31})\in \boxed{-1,-1,1,1}^{1}_{\rm inv}$.

In a similar way one shows that

$$e_{23}\ \boxed{-2,-1,1,2}_{21}\ \subset\ \boxed{-1,-1,1,1}^{1}_{\rm inv}\oplus\boxed{-1,-1,0,2}_{20}\ \oplus\ \boxed{-2,0,1,1}_{11}\quad . \tag{A41}$$

In order to show that $\boxed{-1,0,0,1}_{10}\ \in I_1$ we observe that the equalities

$$m_{33}=-m_{13}+1 \text{ and } m_{13}=m_{23}+1 \tag{A42}$$

cannot be fulfilled simultaneously. Indeed, if both Eqs. (A42) hold, since also $m_{43}=-m_{13}$, we would obtain an induced module with a signature

$$[m_{13},m_{23},m_{33},m_{43}]=[m_{13},m_{13}-1,-m_{13}+1,-m_{13}]\,,$$

which belongs to the class 5 representations. Suppose that $m_{33}>-m_{13}+1$. Then [see (A28) and (A39)]

$$0\neq V_{20}([m_{13}-1,m_{23}-1,m_{33},-m_{13}+2])\subset I_1 \tag{A43}$$

and, therefore, also

$$e_{23}V_{20}([m_{13}-1,m_{23}-1,m_{33},-m_{13}+2])\subset I_1\,. \tag{A44}$$

From (35) we have

$$e_{23}\begin{bmatrix} m_{13} & , & m_{23} & , & m_{33}, & -m_{13} \\ m_{13}-1, & & m_{23}-1, & & m_{33}, & -m_{13}+2 \\ m_{11} & , & 0 & , & m_{31}, & 0 \end{bmatrix}_{20}$$

$$=k(m_{11},m_{31})$$

$$\times\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}, & m_{33} & , & -m_{13}+1 \\ m_{11} & , & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{10}\ , \tag{A45}$$

where the coefficient

$$k(m_{11},m_{31})$$

$$=(l_{23}-l_{13})\left|\frac{(l_{23}-l_{11})(l_{33}-l_{31}+1)}{(l_{13}-l_{23})(l_{13}+l_{33}+2)}\right|^{1/2}$$

$$=0\quad\text{iff}\quad m_{11}=m_{23}-1\,. \tag{A46}$$

Since $m_{13}>m_{23}$ [see (70)], for $m_{11}=m_{13}-1$ we have that $m_{11}>m_{23}-1$. Hence $k(m_{13}-1,m_{31})\neq 0$ and from (A45) we obtain that

$$\begin{bmatrix} m_{13} & , & m_{23}, & m_{33} & , & -m_{13} \\ m_{13}-1, & & m_{23}, & m_{33} & , & -m_{13}+1 \\ m_{13}-1, & & 0\ , & m_{31}-1, & & 0 \end{bmatrix}_{10}$$

$$\in V_{10}([m_{13}-1,m_{23},m_{33},-m_{13}+1])\cap I_1\,. \tag{A47}$$

Hence (Proposition 3),

if $m_{33}>-m_{13}+1$,

$$V_{10}([m_{13}-1,m_{23},m_{33},-m_{13}+1])\subset I_1\,.$$

If $m_{33}=-m_{13}+1$, then $m_{13}>m_{23}+1$ and, therefore, also

$$0\neq V_{11}([m_{13}-2,m_{23},m_{33}+1,-m_{13}+1])\subset I_1\,.$$

In a similar way as before one shows that

$$0\neq e_{23}V_{11}([m_{13}-2,m_{23},m_{33}+1,-m_{13}+1])$$

$$\subset V_{10}([m_{13}-1,m_{23},m_{33},-m_{13}+1])\subset I_1\,.$$

Therefore, in all cases

$$V_{10}([m_{13}-1,m_{23},m_{33},-m_{13}+1])$$

$$\equiv\boxed{-1,0,0,1}_{10}\ \subset I_1\,. \tag{A48}$$

From (50), (A7), (A10), (A13), (A25), (A27), (A39), and (A48) we conclude that

$$\begin{aligned} I\equiv\ &\boxed{-1,0,0,1}_{10} \\ &\oplus\boxed{-1,-1,0,2}_{20}\ \oplus\boxed{-1,-1,1,1}^{1}_{\rm inv}\ \oplus\boxed{-2,0,1,1}_{11} \\ &\oplus\boxed{-1,-2,1,2}_{21}\ \oplus\boxed{-2,-1,1,2}_{21}\ \oplus\boxed{-2,-1,2,1}_{21} \\ &\oplus\boxed{-2,-2,2,2}_{22}\quad\subset I_1\,. \end{aligned} \tag{A49}$$

Now it is a matter of a direct computation to show that:

(1) The subspace I is gl(2/2) invariant. This follows from the results obtained so far and the observation that

$$e_{32}\ \boxed{}\ \subset I \tag{A50}$$

for every term $\boxed{}$ on the right-hand side of (A49).

(2) The subspace I is an irreducible gl(2/2) module with a signature (= with coordinates of its highest weight vector [see (9)]) $[m_{13}-1,m_{23},m_{33},-m_{13}+1]$. Note that this representation is also nontypical.

(3) Acting appropriate times with e_{23} on each $\boxed{}$ term in the right-hand side of the decomposition (78) one ends in $\boxed{0,0,0,0}_{00}$.

From (1)–(3) it follows immediately that (a) I is the maximal gl(2/2) invariant subspace in

$$W([m_{13},m_{23},m_{33},-m_{13}]),\ I=I_1\,, \tag{A51}$$

(b) The compliment 'to I_1 subspace $W_1([m_{13},m_{23},m_{33},-m_{13}])$ in $W([m_{13},m_{23},m_{33},-m_{13}])$ [which is isomorphic to the factor space $W([m_{13},m_{23},m_{33},-m_{13}])/I_1$] is given with the sum of all $\boxed{}$ terms in the right-hand side of the decomposition (78). (A52)

[1] A. H. Kamupingene, N. A. Ky, and T. D. Palev, J. Math. Phys. **30**, 553 (1989).
[2] V. G. Kac, Lect. Notes Math. **626**, 597 (1978).
[3] A. Pais and V. Rittenberg, J. Math. Phys. **16**, 2062 (1975).
[4] M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977).
[5] F. A. Berezin, Sov. J. Nucl. Phys. **29**, 857 (1979); **30**, 605 (1979).
[6] P. D. Jarvis and H. S. Green, J. Math. Phys. **20**, 2115 (1979).
[7] M. Marcu, J. Math. Phys. **21**, 1277 (1980).
[8] P. H. Dondi and P. D. Jarvis, Z. Phys. C **4**, 201 (1980); J. Phys. (Paris) A **14**, 547 (1981).
[9] J. W. B. Hughes, J. Math. Phys. **22**, 245 (1981).
[10] A. H. Balantekin and I. Bars, J. Math. Phys. **22**, 1149 (1981); **22**, 1810 (1981); **23**, 1239 (1982).
[11] D. A. Leites, Theor. Math. Phys. **52**, 764 (1982) (in Russian); Sov. Prob. Math. **25**, 3 (1984) (in Russian).
[12] J.-P. Hurni and B. Morel, J. Math. Phys. **23**, 2236 (1982); **24**, 157 (1983).
[13] I. Bars, B. Morel, and H. Ruegg, J. Math. Phys. **24**, 2253 (1983).
[14] R. J. Farmer and P. D. Jarvis, J. Phys. (Paris) A **16**, 473 (1983); **17**, 2365 (1984).
[15] B. Gruber, T. S. Santhanam, and R. Wilson, J. Math. Phys. **25**, 1253 (1984).
[16] J. Thierry-Mieg, Phys. Lett. B **138**, 393 (1984).
[17] J. Van der Jeugt, J. Math. Phys. **25**, 3334 (1984); **28**, 292 (1987); J. Phys. A **20**, 809 (1987).
[18] F. Delduc and M. Gourdin, J. Math. Phys. **25**, 1651 (1984); **26**, 1865 (1985).
[19] I. Bars, Physica D **15**, 42 (1985).
[20] T. D. Palev, J. Math. Phys. **26**, 1640 (1985); **27**, 1994 (1986); **28**, 272 (1987); **28**, 2280 (1987).
[21] T. D. Palev, Funct. Analysis its Appl. **21**, 85 (1987) (in Russian), English translation: Funct. Anal. Appl. **21**, 245 (1987); Funct. Analysis its Appl. **23**, 69 (1989) (in Russian).
[22] A. Berele and A. Regev, Adv. Math. **64**, 118 (1987).
[23] C. J. Cummins and R. C. King, J. Phys. (Paris) A **20**, 3103 and 3121 (1987).
[24] J. P. Hurni, J. Phys. (Paris) A **20**, 5755 (1987).
[25] M. D. Gould, J. Austral. Math. Soc. Ser. B **28**, 310 (1987).
[26] T. D. Palev, "Irreducible finite-dimensional representations of the Lie superalgebra $gl(n/1)$ in a Gel'fand–Zetlin basis," preprint ICTP, IC/88/207, 1988 (to appear in J. Math. Phys.).
[27] R. Le Blanc and D. J. Rowe, J. Math. Phys. **30**, 1415 (1989).
[28] M. D. Gould, A. J. Bracken, and J. W. B. Hughes, J. Phys. (Paris) A **22**, 2879 (1989).
[29] J. Milnor and J. Moore, Ann. Math. **81**, 211 (1965).

Fermionic determinants for chiral-bag-like two-dimensional systems

H. Falomir
Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina

M. A. Muschietti
Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina

E. M. Santangelo
Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de la Plata, La Plata, Argentina

(Received 23 June 1989; accepted for publication 18 October 1989)

This paper evaluates the determinant of a Dirac-like operator, for a system of fermions confined to a region of two-dimensional Euclidean space and subject to bag-like boundary conditions. In the framework of the ζ-function regularization method, use is made of Seeley's developments for the resolvent of such operators. A relation is derived between the Green's function and the finite part of the diagonal element of the kernel of the power z of the operator for $z = -1$, which amounts to having information about the whole Seeley's series.

I. INTRODUCTION

Chiral bags[1] are well known to model confinement without violating chiral symmetry. They employ quarks and gluons at short distances and light mesons at larger ones, the bag wall dividing both descriptions in space. The observed approximate independence of the radius of the bag shown by some physical observables in four dimensions, together with the exact equivalence of bosonic and fermionic theories in two dimensions, gave rise to the Cheshire cat hypothesis,[2] according to which the bag wall has no physical significance, merely separating two regions where different descriptions of the same physics are used. Testing this hypothesis in the path integral formalism amounts to the calculation of fermionic determinants in a region of space (Ω), with local (bag-like) conditions at its boundary ($\partial\Omega$).

While the definition of determinants of Dirac operators on boundaryless manifolds (compactified space-time) has been extensively studied,[3,4] the presence of boundaries poses some extra difficulties: in particular, the knowledge of the Green's function is needed.

In previous work,[5,6] we evaluated the Dirac determinant for massless fermions coupled to a bosonic background field, confined to a region of two-dimensional space-time, and satisfying bag boundary conditions. In doing so, we used an ad hoc regularization scheme aiming at avoiding possible singularities on $\partial\Omega$. There, we found that the quotient of determinants of the operator with and without a background field was given by the volume integral over Ω of the same density as in the boundaryless case, plus an integral on $\partial\Omega$. This last contribution was seen to vanish for static bosonic configurations while, in the general case, it could be related to the determinant of the free Dirac operator, with "chirally transformed" boundary conditions. However, this interpretation was established by combining the previously mentioned regularization with a point splitting one for boundary terms.

In this paper, we evaluate the same quotient, both in the Abelian and non-Abelian cases, consistently using a unique regularization scheme: the ζ function one. We make use of Seeley's[7] definition of complex powers of differential operators acting on functions that fulfill given boundary conditions. We look for the kernel of the relevant operator, satisfying the right conditions on $\partial\Omega$, which amounts to obtaining two sets of Seeley's coefficients, those common to the case with no boundaries plus a new set adjusting the behavior of the kernel on $\partial\Omega$. The main ingredient in our calculation will be the derivation of an expression for the finite part of $D_B^z(x,x)|_{z=-1}$, the diagonal element of the kernel of the inverse of an elliptic invertible operator of order ω on a ν-dimensional compact manifold with boundary. This expression generalizes the one presented in Ref. 8 for the case with no boundaries. This method also allows us to compare our result to the determinant of the free Dirac operator with chiral boundary conditions in the framework of a consistent regularization, thus confirming the relationship remarked in Refs. 5 and 6.

II. THE FERMIONIC DETERMINANT

We consider a theory of massless fermions, confined to a region Ω of two-dimensional Euclidean space, interacting with a scalar background field φ, taking values in the Lie algebra of a compact Lie group, whose action is given by

$$S = \int_\Omega d^2x\, \bar\psi e^{\gamma_5\varphi} i\not\partial e^{\gamma_5\varphi}\psi + \frac{i}{2}\int_{\partial\Omega} d^1x\, \bar\psi(1-\not n)\psi . \qquad (1)$$

[Our convention is

$$\gamma_0 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \gamma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma_5 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and $\not n$ is the exterior normal to $\partial\Omega$.]

In order to obtain the effective action for φ, resulting from its interaction with fermions, we evaluate the functional integral:

$$Z = \int D\bar\psi\, D\psi\, e^{-S} . \qquad (2)$$

This model is classically equivalent to a free fermion model with chiral bag boundary conditions, through the transformation $\chi = e^{\gamma_5\varphi}\psi$; $\bar{\chi} = \bar{\psi}e^{\gamma_5\varphi}$; however, this analogy is no longer true at the quantum level due to the noninvariance of the measure.

Since the integral over $\bar{\psi}(\partial\Omega)$ selects those trajectories satisfying bag boundary conditions, we must consider the differential operator defined as:

$$D_B\psi = Ui\not\partial U\psi = e^{\gamma_5\varphi}i\not\partial e^{\gamma_5\varphi}\psi\,, \quad \text{in } \Omega \tag{3}$$

if

$$B\psi = (1 - i\not n)\psi = 0\,, \quad \text{on } \partial\Omega\,. \tag{4}$$

By developing ψ in eigenfunctions (generalized eigenfunctions, if necessary) of D_B, and $\bar{\psi}$ in eigenfunctions of D_B^*, which constitute a biorthogonal basis in Hilbert space,[9] we define the functional integral as

$$-\Gamma(\varphi) = \log Z = \log \operatorname{Det} D_B = \operatorname{Tr}\log D_B\,. \tag{5}$$

These are just formal equations, which require a regularization; we choose the Riemann ζ-function scheme:

$$\log \operatorname{Det} D_B = -\left.\frac{d}{ds}\right|_{s=0} \operatorname{Tr}\{D_B^{-s}\}\,, \tag{6}$$

the trace converging for $\operatorname{Re}(s)$ large enough, and the result being analytically continued to $s = 0$.

As usual, we introduce a parameter τ by changing $\varphi \rightarrow \tau\varphi$. Notice that the resulting $D_B(\tau)$ acts on a domain that is τ independent. We evaluate the τ derivative of (5) and recover Γ by integration:

$$\frac{d}{d\tau}\Gamma[\varphi] = -\left.\frac{d}{ds}\right|_{s=0} s\operatorname{Tr}\{[\gamma_5\varphi(Ui\not\partial U) + (Ui\not\partial U)\gamma_5\varphi]D_B^{-s-1}\}\,. \tag{7}$$

An integration by parts of the second term allows us to write this expression as:

$$\frac{d}{d\tau}\Gamma[\varphi] = A(\tau) + B(\tau)\,, \tag{8}$$

with

$$A(\tau) = -\left.\frac{d}{ds}\right|_{s=0} s\int_\Omega d^2x\,\operatorname{tr}\{2\gamma_5\varphi(x)D_B^{-s}(x,x)\} \tag{9}$$

and

$$B(\tau) = -\left.\frac{d}{ds}\right|_{s=0} s\int_{\partial\Omega} d^1x\,\operatorname{tr}\{i\not n\gamma_5\varphi(x)D_B^{-s-1}(x,x)\}\,, \tag{10}$$

where tr means sum over group and Lorentz indices.

Although Eq. (9) has the same form as in the boundaryless case, the integral is limited to Ω and, as we will see in the next section, the boundary conditions will give rise to some extra terms.

Equation (10) introduces an additional complication: one needs the diagonal element of the kernel for powers of the operator close to minus one.

III. EVALUATION OF $A(\tau)$

The power z of D_B for $\operatorname{Re}(z) < 0$ is defined as

$$D_B^z = \frac{i}{2\pi}\int d\lambda\,\lambda^z(D_B - \lambda)^{-1}\,, \tag{11}$$

where the integral is evaluated on a path encircling the spectrum of D_B. We use Seeley's expansion for the resolvent in terms of pseudodifferential operators.[7,10] It can be verified that D_B and B in Eqs. (3) and (4) constitute an elliptic boundary problem, satisfying Agmon's condition on the positive imaginary λ axis.

In terms of Seeley's coefficients, the resolvent is approximated by

$$\begin{aligned}(D_B - \lambda)^{-1}(x,y) &\sim \sum_{j=0}^{N}\frac{1}{(2\pi)^\nu}\int d^\nu\xi\, e^{i\xi(x-y)} \\ &\quad\times c_{-\omega-j}(x,\xi,\lambda)\theta_2(\xi,\lambda) - \sum_{j=0}^{N}\frac{1}{(2\pi)^\nu}\int d^{\nu-1}\xi_\parallel \\ &\quad\times e^{i\xi_\parallel(x-y)_\parallel}\tilde{d}_{-\omega-j}(x_\parallel,x_\perp,\xi_\parallel,y_\perp,\lambda)\theta_1(\xi,\lambda)\,,\end{aligned} \tag{12}$$

where $\parallel$ ($\perp$) stand for directions parallel (orthogonal) to the boundary, ν is the dimension of the space-time, and ω is the order of the differential operator (in our case, $\nu = 2$ and $\omega = 1$). Here, $\theta_1(\xi_\parallel,\lambda)((\theta_2(\xi,\lambda))$ is a smooth function, vanishing for small $|\xi_\parallel|^2 + |\lambda|^{2/\omega}(|\xi|^2 + |\lambda|^{2/\omega})$, and identically 1 when this quantity is greater than one.

The coefficients in this expansion are defined through iterative relations. In the case of the c's, they are algebraic equations, while the evaluation of d in

$$\tilde{d}_{-\omega-j}(x_\parallel,x_\perp,\xi_\parallel,y_\perp,\lambda) = \int_{-\infty}^{\infty} d\xi_\perp\, e^{-i\xi_\perp y_\perp}d_{-\omega-j}(x_\parallel,x_\perp,\xi_\parallel,\xi_\perp,\lambda) \tag{13}$$

amounts to solving a differential equation in the normal coordinates.[7]

In order to evaluate $A(\tau)$ in (9), one needs the diagonal element of the kernel $D_B^z(x,x)$, for z near zero. As shown in Ref. 7, the asymptotic expansion obtained by replacing (12) in (11) has only a single pole at $z = 0$; therefore, the presence of $d/ds|_{s=0}s$ amounts to taking the finite part of this analytic extension at $z = 0$. From the same reference, it is easy to see that (due to its homogeneity properties) the only c coefficient contributing to $A(\tau)$ is c_{-3}. Taking into account the extension of Seeley's method given in Ref. 10, which is necessary due to the arbitrary dependence of φ on $x_\perp$, it turns out that one needs also d_{-1} and d_{-2}.

The resulting expression for $A(\tau)$ is

$$\begin{aligned}A(\tau) = -\frac{1}{2\pi^2}FP_{z=0}\frac{1}{z}\Big\{&\operatorname{tr}\int_\Omega d^2x\,\gamma_5\varphi(x) \\ &\times\int_{|\xi|=1} d^1\xi\, C_{-3}(x,\xi,z) - \operatorname{tr}\int_{\partial\Omega} d^1x\,\gamma_5 \\ &\times\sum_{\xi_\parallel=\pm1}\int_0^\infty dx_\perp\,[\varphi(x_\parallel,0)D_{-2}(x_\parallel,x_\perp,\xi_\parallel,x_\perp;z) \\ &+ \partial_\perp\varphi(x_\parallel,0)x_\perp D_{-1}(x_\parallel,x_\perp,\xi_\parallel,x_\perp;z)]\Big\}\,,\end{aligned} \tag{14}$$

where

$$C_{-3}(x,\xi,z) = \frac{i}{2\pi}\int d\lambda\,\lambda^z c_{-3}(x,\xi,\lambda)\theta_2(\xi,\lambda)\,, \tag{15}$$

$$D_{-1-j}(x_{\|},x_{\perp},\xi_{\|},z)$$
$$=\frac{i}{2\pi}\int d\lambda\,\lambda^{z}\tilde{d}_{-1-j}(x_{\|},x_{\perp}\xi_{\|},x_{\perp},\lambda)\theta_1(\xi_{\|},\lambda)\,, \quad (16)$$

and the integral is taken on a curve coming from ∞ along the positive imaginary axis to a small circle around the origin, then clockwise around the circle and back to infinity along the same axis.

The c coefficients, which are independent of the presence of boundaries, have been evaluated for a similar problem in Ref. 4, from which we take

$$c_{-1}(x,\xi,\lambda)=(\lambda-\not{\xi})/(\xi^2-\lambda^2)\,, \quad (17)$$
$$c_{-2}(x,\xi,\lambda)=-[1/(\xi^2-\lambda^2)^2](\lambda-\not{\xi})\not{A}(\lambda-\not{\xi})\,, \quad (18)$$
$$c_{-3}(x,\xi,\lambda)=[1/(\xi^2-\lambda^2)^3](\lambda-\not{\xi})(\not{A}+i\not{\partial})$$
$$\times(\lambda-\not{\xi})\not{A}(\lambda-\not{\xi})\,, \quad (19)$$

where

$$\not{A}=U(i\not{\partial}U)=A_\mu\gamma_\mu\,. \quad (20)$$

In order to simplify our calculations, we will consider that the region Ω is the half-plane $x_1>0(x_\perp=x_1;x_{\|}=x_0)$.

The coefficient d_{-1} is the solution of the differential equation:

$$\left(\frac{\partial}{\partial x_1}-M\right)d_{-1}(x,\xi,\lambda)=0\,, \quad (21)$$

satisfying

$$d_{-1}\underset{x_1\to+\infty}{\to}0\,,$$
$$(1,1)(d_{-1}-c_{-1})|_{x_1=0}=0\,. \quad (22)$$

In Eq. (21), M is the matrix

$$M=\xi_0\gamma_5-i\lambda\gamma_1\,. \quad (23)$$

Solving (18) and (19), one gets for $\tilde{d}_{-1}$ [see Eq. (13)],

$$\tilde{d}_{-1}(x,\xi_0,s,\lambda)=\frac{i\pi e^{-\rho(x_1+s)}}{\rho(i\lambda+\xi_0+\rho)}\begin{pmatrix} i\lambda \\ \rho+\xi_0\end{pmatrix}$$
$$\otimes(1,1)(M+\rho)\gamma_1\,, \quad (24)$$

where

$$\rho=+(\xi_0^2-\lambda^2)^{1/2}\,. \quad (25)$$

Concerning d_{-2}, it satisfies

$$\left(\frac{\partial}{\partial x_1}-M\right)d_{-2}(x,\xi,\lambda)=i\gamma_1\not{A}d_{-1}(x,\xi,\lambda)\,, \quad (26)$$

and

$$d_{-2}\underset{x_1\to+\infty}{\to}0\,,$$
$$(1,1)(d_{-2}-c_{-2})|_{x_1=0}=0\,. \quad (27)$$

From these two equations, one obtains for $\tilde{d}_{-2}$:

$$\tilde{d}_{-2}(x,\xi_0,s,\lambda)=e^{-\rho(2+x_1)}\frac{i\pi(i\lambda+\xi_0-\rho)}{4\lambda\xi_0\rho^2}$$
$$\times\left\{\begin{pmatrix} i\lambda \\ \rho+\xi_0\end{pmatrix}\otimes(1,1)\left[\gamma_1\not{A}(M+\rho)\gamma_1+(M+\rho)\gamma_1\not{A}\gamma_1-\left(s+\frac{1}{\rho}\right)(M+\rho)\gamma_1\not{A}(M+\rho)\gamma_1\right]\right.$$
$$\left.+\left[x_1(\rho-M)+\frac{(i\lambda-\xi_0-\rho)}{2i\lambda\rho}\begin{pmatrix} 1 \\ -1\end{pmatrix}\otimes(i\lambda,\xi_0-\rho)\right]\gamma_1\not{A}\begin{pmatrix} i\lambda \\ \rho+\xi_0\end{pmatrix}\otimes(1,1)(M+\rho)\gamma_1\right\}. \quad (28)$$

The presence of the trace over Lorentz indices greatly simplifies the awful expression resulting from the replacement of (19), (24), and (28) in Eq. (14) (giving, in particular, a vanishing contribution from $\tilde{d}_{-1}$). The final result is

$$A(\tau)=-\frac{1}{4\pi}\int_\Omega d^2x\,\mathrm{tr}\{\gamma_5\varphi\not{\partial}(U^2\not{\partial}U^2)\}$$
$$+\frac{1}{\pi}\int_{\partial\Omega}d^1x\,\mathrm{tr}_g\{\varphi A_0\}\,, \quad (29)$$

where tr_g means trace over group indices.

In the last term, coming from $\tilde{d}_{-2}$,

$$\mathrm{tr}_g\{\varphi A_0\}=\tau\,\mathrm{tr}_g\{\varphi\partial_1\varphi\}\,, \quad (30)$$

leading to a trivial τ integration.

With respect to the first term in (29), taking into account that

$$\mathrm{tr}\{\gamma_5\varphi\not{\partial}(U^2\not{\partial}U^2)\}=\mathrm{tr}\{2\gamma_5\varphi U\not{\partial}(U^2\not{\partial}U)\}\,, \quad (31)$$

its τ integral is as given in Ref. 6. The final result is

$$\int_0^1 d\tau\,A(\tau)$$
$$=W_\Omega[u=e^{2\varphi}]-\frac{1}{2\pi}\int_{\partial\Omega}dx_0\,\mathrm{tr}_g\{\varphi\partial_1\varphi\}_{(x_0,0)}$$
$$+\frac{1}{2\pi}\int_{\partial\Omega}dx_0\,\mathrm{tr}_g\{\varphi\partial_1\varphi\}_{(x_0,0)}\,, \quad (32)$$

where

$$W_\Omega[u]=\frac{1}{8\pi}\int_\Omega d^2x\,\mathrm{tr}_g[(u^{-1}\partial_\mu u)(u^{-1}\partial_\mu u)]$$
$$-\frac{i}{4\pi}\int_0^1 d\tau\int_\Omega d^2x\epsilon_{\mu\nu}\,\mathrm{tr}_g[(u_\tau^{-1}\partial_\tau u_\tau)$$
$$\times(u_\tau^{-1}\partial_\mu u_\tau)(u_\tau^{-1}\partial_\nu u_\tau)]\,, \quad (33)$$

with $u_\tau=e^{2\tau\varphi}$, has the form of the Wess–Zumino–Witten action,[11] but restricted to the region Ω of space-time.

Notice that there is a cancellation between the d contribution and a boundary term coming from an integration by parts in the c contribution.

In the simple Abelian case, Eq. (33) reduces to

$$W_\Omega[u] = \frac{1}{2\pi}\int_\Omega d^2x(\partial\varphi)^2 . \tag{34}$$

One sees that the final expression is the volume integral over Ω of the same density as in the boundaryless case.

IV. EVALUATION OF $B(\tau)$

Since the kernel $D_B^z(x,x)$ has only a simple pole at $z = -1$ (Ref. 7),

$$B(\tau) = -\underset{z=-1}{FP}\int d^1x\, \mathrm{tr}\{i\gamma_1\gamma_5\varphi D_B^z(x,x)\} . \tag{35}$$

In calculating $B(\tau)$ one is faced with an extra complication: it cannot be expressed in terms of a finite number of Seeley's coefficients; rather, the knowledge of the Green's function is needed. In the Appendix, we show the relationship between the finite part of the relevant kernel and the Green's function of D_B (G). When replaced in (35), it gives

$$\begin{aligned} B(\tau) = &-\frac{1}{(2\pi)^2}\,\mathrm{tr}\int_{\partial\Omega} dx_0\,\gamma_0\varphi(x_0,0)\Big\{(2\pi)^2G(x_0,0;y_0,0) - \int_{\mathbf{R}^2} d^2\xi\, e^{i\xi_0(x-y)_0}|\xi|^{-1}C_{-1}(x_0,0,\xi/|\xi|;z=-1) \\ &+ \int_{\mathbf{R}} d\xi_0\, e^{i\xi_0(x-y)_0}D_{-1}(x_0,0,\xi_0/|\xi_0|,0;z=-1) - \int_{|\xi|>1} d^2\xi\, e^{i\xi_0(x-y)_0}C_{-2}(x_0,0,\xi;z=-1) \\ &+ \int_{|\xi_0|>1} d\xi_0\, e^{i\xi_0(x-y)_0}D_{-2}(x_0,0,\xi_0,0;z=-1) - \int_{|\xi|=1} d^1\xi\,\frac{d}{dz}C_{-2}(x_0,0,\xi;z)\big|_{z=-1} \\ &+ \sum_{\xi_0=\pm1}\frac{d}{dz}D_{-2}(x_0,0,\xi_0,0;z)\big|_{z=-1}\Big\}\Big|_{y_0=x_0} . \end{aligned} \tag{36}$$

The evaluation of the C terms is similar to that of Ref. 8. One gets

$$\frac{1}{(2\pi)^2}\int_{\mathbf{R}^2} d^2\xi\, e^{i\xi(x-y)}|\xi|^{-1}C_{-1}(x,\xi/|\xi|;z=-1) = (1/2\pi i)(\not{x}-\not{y})/|x-y|^2 , \tag{37}$$

$$\left(\frac{1}{2\pi}\right)^2\int_{|\xi|>1} d^2\xi\, e^{i\xi(x-y)}C_{-2}(x,\xi;z=-1) = -1/4\pi[\not{A} - 2A_\mu(x-y)_\mu(\not{x}-\not{y})/|x-y|^2] , \tag{38}$$

$$\left(\frac{1}{2\pi}\right)^2\int_{|\xi|=1} d^1\xi\, C_{-2}(x,\xi;z) = (z+1)\not{A}/4\pi\{1+o(z+1)\} . \tag{39}$$

Notice that, between the last two contributions coming from C_{-2}, there will be a cancellation of the continuous parts, leading to the subtraction of a "minimal" nonregular zero order term to G.

Concerning D contributions, it is easy to obtain

$$D_{-1}(x_0,x_1,\xi_0,x_1;z=-1) = i\pi e^{-2}|\xi_0|x_1\begin{pmatrix} 1-\mathrm{sgn}\,\xi_0 & 0 \\ 0 & 1+\mathrm{sgn}\,\xi_0 \end{pmatrix} , \tag{40}$$

from which it follows that

$$\mathrm{tr}\{\gamma_0 D_{-1}(x_0,x_1,\xi_0,x_1;z=-1)\} = 0 . \tag{41}$$

In the case of D_{-2}, an otherwise complicated calculation is greatly simplified by first evaluating

$$\mathrm{tr}\{\gamma_0\tilde{d}_{-2}(x_0,0,\xi_0,0,\lambda)\} = 2\pi A_0/\xi_0^2[-\lambda + \lambda^4/\rho^3 + 2\xi_0^2\lambda^2/\rho^3] , \tag{42}$$

which leads to

$$\mathrm{tr}\{\gamma_0 D_{-2}(x_0,0,\xi_0,0;z)\} = 0 . \tag{43}$$

Replacing (37) to (43) in (36) one obtains

$$\begin{aligned} B(\tau) = -\lim_{y_0\to x_0}\Big\{&\mathrm{tr}\int dx_0\,\gamma_0\varphi(x_0,0)G(x_0,0;y_0,0) \\ &+\frac{1}{\pi}\,\mathrm{tr}_g\int dx_0\,\varphi(x_0,0)[i/(x_0-y_0) \\ &-A_0 + o(x_0-y_0)]\Big\} . \end{aligned} \tag{44}$$

In order to evaluate (44), we need the explicit expression for G:

$$G(x,y) = U^{-1}(x)g(x,y)U^{-1}(y) . \tag{45}$$

For the simple Abelian case, we found in Ref. 5:

$$g(x,y) = V(x)[g_0(x_0,x_1;y) - \gamma_1 g_0(x_0,-x_1;y)] \times\gamma_1 V^{-1}(y)\gamma_1 , \tag{46}$$

where

$$g_0(x,y) = (1/2\pi i)(\not{x}-\not{y})/|x-y|^2 \tag{47}$$

and

$$V(x) = \mathrm{diag}(e^{2\tau\varphi_+(x_0-ix_1)}, e^{-2\tau\varphi_+^*(x_0+ix_1)}) , \tag{48}$$

with $\varphi_+(x_0)$ being the positive frequency part of $\varphi(x_0,x_1=0)$ such that $\varphi_+(x_0)+\varphi_+^*(x_0) = \varphi(x_0,0)$.

It can be seen that the second term in (46) gives a vanishing contribution to the trace. By developing $U(y)$ and $V(y)$ in powers of $(x–y)$ one easily obtains

$$\int_0^1 d\tau\, B(\tau) = \frac{1}{2\pi}\int dx_0\,\varphi\partial_1\varphi(x_0,0) + \frac{i}{2\pi}\int dx_0\,\varphi(x_0,0)\partial_0(\varphi_+^* - \varphi_+)(x_0) . \tag{49}$$

The first term in this expression can be written as the volume integral over Ω of a total divergence, while the second one is a complicated nonlocal function of $\varphi(x_0,x_1=0)$, vanishing for static configurations.

The effective action Γ is

$$\Gamma[\varphi] = -\frac{1}{2\pi}\int_\Omega d^2x\,\varphi\partial^2\varphi + \frac{1}{2\pi}\int_{\partial\Omega} dx_0\,\varphi\partial_0 \times(\varphi_+^* - \varphi_+) - \log \operatorname{Det}(i\not{\partial})_B\,, \tag{50}$$

where the last term is the "Casimir energy" of free fermions inside the bag, being $(i\not{\partial})_B$ defined by Eqs. (3) and (4), with $\varphi = 0$.

This result coincides with that found in Ref. 5, up to the volume integral of a total divergence (due to the different regularization method employed), so that the same comments concerning the relevance of the model for testing the Cheshire cat hypothesis made in that reference still apply.

For the non-Abelian case, one can employ a multiple reflection expansion[12] for g, given in detail in Ref. 6

$$g(x,y) = g_0(x,y) + g_1(x,y) + \Delta g(x,y)\,, \tag{51}$$

where g_0 is given by (47), g_1 is the one-reflection contribution (explicitly evaluated in Ref. 6), and Δg (the multiple reflection contribution) is regular for $x_0 \to y_0$.

When replacing (51) into (44), g_1 is easily seen to give no contribution due to the null value of the trace over Lorentz indices. By developing $U^{-1}(y)$ in powers of $(x-y)$, the remaining terms in (44) add to

$$\int_0^1 d\tau\, B(\tau) = \frac{1}{2\pi}\int dx_0\,\operatorname{tr}_g\{\varphi\partial_1\varphi\}(x_0,0) - \int_0^1 d\tau \int dx_0\,\operatorname{tr}\{\varphi(x_0,0)\gamma_0\Delta g(x_0,0)\}\,. \tag{52}$$

As in the Abelian case, the first term can be cast in the form of a volume integral over Ω of a total divergence, while an interpretation of the second one (a nonlocal function of φ on $\partial\Omega$) will be given in the next section.

The resulting effective action is

$$\Gamma[\varphi] = W_\Omega[e^{2\varphi}] - \frac{1}{2\pi}\int_\Omega d^2x\,\operatorname{tr}_g[\partial_\mu(\varphi\partial_\mu\varphi)] - \int_0^1 d\tau\int_{\partial\Omega} dx_0\,\operatorname{tr}[\varphi\gamma_0\Delta g] - \log\operatorname{Det}(i\not{\partial})_B\,. \tag{53}$$

Again, we find the same result as in Ref. 6 even though, as a consequence of the different regularization employed, there is an extra volume integral of a total divergence.

V. COMPARISON WITH THE CHIRALLY TRANSFORMED OPERATOR

As mentioned in the Introduction, we suggested in previous work,[5,6] that the nonlocal expressions under boundary integrals in Eqs. (50) and (53) could be related to the determinant of the operator chirally transformed from D_B:

$$(i\not{\partial})_{BU^{-1}}\psi = i\not{\partial}\psi\,, \quad \text{in } \Omega \tag{54}$$

for functions satisfying

$$BU^{-1}\psi = 0, \quad \text{on } \partial\Omega\,. \tag{55}$$

It is easy to establish the relation[13]

$$\operatorname{Det}(i\not{\partial})_{BU^{-1}} = \operatorname{Det}(U^{-1}i\not{\partial}U)_B\,, \tag{56}$$

from which one can proceed as earlier, introducing the parameter τ and taking the τ derivative. The difference is that, in this case, there is no volume contribution. One gets

$$\frac{d}{d\tau}\log\operatorname{Det}(i\not{\partial})_{BU^{-1}} = \frac{d}{ds}\bigg|_{s=0} s\,\operatorname{tr}\int_{\partial\Omega} dx_\parallel\, U^{-1}i\not{n}U\gamma_5\varphi(x) \times(U^{-1}i\not{\partial}U)_B^{-1-s}(x,x)\,. \tag{57}$$

It is not difficult to establish the following relation between the resolvents of the operators appearing in (56);

$$[(U^{-1}i\not{\partial}U)_B - \lambda]^{-1} = U^{-1}[(i\not{\partial})_{BU^{-1}} - \lambda]^{-1}U\,. \tag{58}$$

By means of Eq. (11),

$$(U^{-1}i\not{\partial}U)_B^z = U^{-1}(i\not{\partial})_{BU^{-1}}^z U\,. \tag{59}$$

Replacing (59) in (57), it follows that

$$\log\left\{\frac{\operatorname{Det}(i\not{\partial})_{BU^{-1}}}{\operatorname{Det}(i\not{\partial})_B}\right\} = \int_0^1 d\tau\, FP_{z=-1}\,\operatorname{tr}\int_{\partial\Omega} dx_\parallel\, i\not{n}\gamma_5\varphi(x) \times(i\not{\partial})_{BU^{-1}}^z(x,x)\,, \tag{60}$$

where we have used the fact that the asymptotic expansion of this kernel has only a single pole at $z = -1$.

One can check that (54) and (55) define an elliptic boundary problem satisfying Agmon's condition on the negative imaginary λ axis.[7]

In order to evaluate (60), we have to look for the Seeley's coefficients appearing in Eq. (A10) in the Appendix. While c_{-1} has the same form as before [Eq. (17)], $c_{-2} = c_{-3} = \cdots = 0$.

Concerning d_{-1}, it satisfies

$$\left(\frac{\partial}{\partial x_1} - M\right)d_{-1} = 0\,, \tag{61}$$

$$\lim_{x_1\to\infty} d_{-1} = 0\,,$$

$$(u^{-1},1)[d_{-1} - c_{-1}]|_{x_1=0} = 0\,, \tag{62}$$

where $u = e^{2\tau\varphi(x_0,0)}$. This leads to

$$\tilde{d}_{-1}(x_0,x_1,\xi_0,s=x_1;\lambda) = i\pi e^{-2\rho x_1}[i\lambda u^{-1} + \xi_0 + \rho]^{-1} \times\begin{pmatrix} i\lambda \\ \xi_0+\rho\end{pmatrix}\otimes(u^{-1},1)(M+\rho)/\rho\gamma_1\,, \tag{63}$$

with M and ρ defined in Eqs. (23) and (25), respectively. It follows that

$$D_{-1}(x_0,x_1=0,\xi_0,s=0;z=-1) = i\pi\begin{pmatrix} u(1-\operatorname{sgn}\xi_0) & 0 \\ 0 & u^{-1}(1+\operatorname{sgn}\xi_0)\end{pmatrix}, \tag{64}$$

which gives a vanishing contribution to the trace in (60). The remaining coefficient is determined by

$$\left(\frac{\partial}{\partial x_1} - M\right)d_{-2} = -i\gamma_5\frac{\partial d_{-1}}{\partial x_0}\,, \tag{65}$$

$$\lim_{x_1\to\infty} d_{-2} = 0\,,$$

$$(u^{-1},1)d_{-2}|_{x_1=0} = 0\,, \tag{66}$$

from which it follows that

$$\begin{aligned}&\mathrm{tr}\{\gamma_0\varphi(x_0,0)\tilde{d}_{-2}(x_0,0,\xi_0,0;\lambda)\}\\&\quad=-i\pi\frac{\partial}{\partial x_0}\left\{\frac{\rho+\xi_0}{\rho}\tau\frac{d}{d\tau}\right.\\&\qquad\left.\times\frac{1}{\tau}\mathrm{tr}_g[i\lambda u^{-1}+\rho+\xi_0]^{-1}\right\}.\end{aligned}\tag{67}$$

Being a total x_0 derivative, this expression gives no contribution when integrated over the boundary. Therefore, we obtain

$$\begin{aligned}&\log\left\{\frac{\mathrm{Det}(i\not\partial)_{BU^{-1}}}{\det(i\not\partial)_B}\right\}\\&\quad=\int_0^1 d\tau\int dx_0\lim_{y_0\to x_0}\{\mathrm{tr}[\gamma_0\varphi(x_0,0)g(x_0,0;y_0,0)]\\&\qquad+i/\pi\,\mathrm{tr}_g[\varphi(x_0,0)/(x_0-y_0)]\}.\end{aligned}\tag{68}$$

We will compare this result with Eq. (44). Making use of (45), we easily get

$$\begin{aligned}&\mathrm{tr}\{\gamma_0\varphi(x_0,0)[G-g](x_0,0;y_0,0)\}\\&\quad=\mathrm{tr}\{\gamma_0\varphi(x_0,0)g(x_0,0;y_0,0)(y_0-x_0)\\&\qquad\times[\partial_0U^{-1}U](x_0,0)\}+O(x_0-y_0).\end{aligned}\tag{69}$$

Both in the Abelian [Eq. (46)] and non-Abelian [Eq. (51)] cases, (69) equals to

$$i/2\pi\,\mathrm{tr}\{\varphi\partial_0U^{-1}U\}(x_0,0)+o(y_0-x_0)\underset{y_0\to x_0}{\to}0.\tag{70}$$

This allows us to give a nice interpretation to nonlocal boundary terms in Eqs. (50) and (53), thus confirming our conjecture of Refs. 5 and 6: In the framework of the ζ-function regularization we can write

$$\begin{aligned}&\log\left\{\frac{\mathrm{Det}(Ui\not\partial U)_B}{\mathrm{Det}(i\not\partial)_{BU^{-1}}}\right\}\\&\quad=W_\Omega[e^{2\varphi}]-\frac{1}{2\pi}\int_\Omega d^2x\,\partial_\mu\,\mathrm{tr}\{\varphi\partial_\mu\varphi\}.\end{aligned}\tag{71}$$

That is, by chirally changing both the differential operator and boundary conditions, this quotient is given by the volume integral, restricted to Ω, of the same density as in the boundaryless case, up to a total divergence.

VI. CONCLUSIONS

In summary, we have developed a regularization method for the definition of determinants of Dirac-like operators with local (bag-like) boundary conditions. In the framework of the ζ function regularization, we have made use of Seeley's development for complex powers of elliptic boundary systems. We have shown that, while volume contributions can be written in terms of a finite number of Seeley's coefficients, there are also boundary contributions, which require the knowledge of the whole series, or equivalently, of the Green's function (G) of the problem. In the Appendix we have established the relation (which also involves a finite number of Seeley's coefficients) between the finite part of the z power of the operator for $z=-1$, and this Green's function. For that reason, we restricted fermions to be confined to a half-plane, a simple situation for which we have derived G in previous work, exactly in the Abelian case and in a multiple reflection expansion in the non-Abelian one.

In this way, by evaluating the relevant Seeley's coefficients, we have obtained for the logarithm of the fermionic determinant, a Wess–Zumino–Witten-like functional in Ω (up to the volume integral of a total divergence), plus a complicated nonlocal term, depending only on the values taken by the background field at the boundary. We have established, in the framework of this regularization, that this last term (which vanishes for static configurations of the background field) can in general be related to the determinant of the free Dirac operator with chiral bag boundary conditions [Eqs. (68)–(70)]. This allowed us to give a simple interpretation to our result, confirming the suggestion made in previous work: by chirally changing both the differential operator and boundary conditions, the logarithm of the quotient of determinants turns out to be the volume integral over Ω of the same density as in the boundaryless case, up to a total divergence. From the path integral point of view, this amounts to saying that, by means of the present regularization scheme, the Jacobian due to the noninvariance of the measure under a chiral change of fermionic variables is given by a Wess–Zumino–Witten-like functional (up to the integral of a total divergence), with no additional boundary contributions.

The relevance of this results in modeling a Cheshire cat behavior has already been discussed in Refs. 5 and 6, and we will not go over to those arguments here.

Our result in the Appendix was established for arbitrary dimension of the space-time (and order of the differential operator), so this method could be directly applied to a more realistic four dimensional case, although that would require (an approximation to) the Green's function of the problem.

ACKNOWLEDGMENTS

This work was partially supported by CONICET, CICPBA, and SECYT (Republica Argentina).

APPENDIX

In order to deduce an expression for the finite part of the kernel of the operator D_B^z for $z=-1$, we must remark on the following properties of such a kernel[7]:

For $\mathrm{Re}(z)<0$, it can be developed as

$$\begin{aligned}D_B^z(x,y)=\frac{1}{(2\pi)^\nu}\sum_{j=0}^{\nu-\omega}\bigg[&\int_{\mathbf{R}^\nu}d^\nu\xi\,C_{-\omega-j}(x,\xi;z)e^{i\xi(x-y)}\\&-\int_{\mathbf{R}^{\nu-1}}d^{\nu-1}\xi_\parallel\,D_{-\omega-j}(x_\parallel,x_\perp,\xi_\parallel,y_\perp;z)\\&\times e^{i\xi_\parallel(x-y)_\parallel}\bigg]+R(x,y;z),\end{aligned}\tag{A1}$$

with $C_{-\omega-j}$ and $D_{-\omega-j}$ defined as in Eqs. (15) and (16).

Moreover,

(i) For $\mathrm{Re}(z)<-1+1/\omega$, $R(x,y;z)$ is a continuous function of its arguments, even at the boundary $\partial\Omega$ of the manifold Ω.

(ii) $C_{-\omega-j}(x,\xi;z)$ is a homogeneous function of ξ of degree $(\omega z-j)$, for $|\xi|\geqslant 1$.

(iii) $D_{-\omega-j}(x_\|,0,\xi_\|,0;z)$ is a homogeneous function of $\xi_\|$ of degree $(\omega z-j+1)$, for $|\xi_\||\geqslant 1$. (Observe that it is so only at $\partial\Omega$.)

Then, we have

$$R(x_\|,0,x_\|,0;z)=D_B^z(x_\|,0,x_\|,0)-\frac{1}{(2\pi)^\nu}\times\sum_{j=0}^{\nu-\omega}\left[\int_{\mathbf{R}^\nu}d^\nu\xi\, C_{-\omega-j}(x_\|,0,\xi;z)-\int_{\mathbf{R}^{\nu-1}}d^{\nu-1}\xi_\| D_{-\omega-j}(x_\|,0,\xi_\|,0;z)\right], \tag{A2}$$

and, for $y_\|\neq x_\|$,

$$R(x_\|,0,y_\|,0;z=-1)=G(x_\|,0,y_\|,0)-\frac{1}{(2\pi)^\nu}\times\sum_{j=0}^{\nu-\omega}\left[\int_{\mathbf{R}^\nu}d^\nu\xi\, C_{-\omega-j}(x_\|,0,\xi;z=-1)e^{i\xi_\|(x-y)_\|}-\int_{\mathbf{R}^{\nu-1}}d^{\nu-1}\xi_\|\, D_{-\omega-j}(x_\|,0,\xi_\|,0;z=-1)\times e^{i\xi_\|(x-y)_\|}\right], \tag{A3}$$

where the Green's function is defined as

$$D_B G(x,y)=\delta(x-y)\,,\qquad BG(x,y)=0\,,\quad \text{for } x\in\partial\Omega\,. \tag{A4}$$

Because of the homogeneity properties of $C_{-\omega-j}$ and $D_{-\omega-j}$, we can write, for $0\leqslant j<\nu-\omega$:

$$\begin{aligned}\lim_{y_\|\to x_\|}&\int_{\mathbf{R}^\nu}d^\nu\xi\,[C_{-\omega-j}(x_\|,0,\xi;z=-1)-|\xi|^{-\omega-j}C_{-\omega-j}(x_\|,0,\xi/|\xi|;z=-1)]e^{i\xi_\|(x-y)_\|}\\&=\int_{|\xi|<1}d^\nu\xi\, C_{-\omega-j}(x_\|,0,\xi;z=-1)-\frac{1}{(-\omega-j+\nu)}\int_{|\xi|=1}d^{\nu-1}\xi\, C_{-\omega-j}(x_\|,0,\xi;z=-1)\\&=\lim_{z\to-1}\int_{\mathbf{R}^\nu}d^\nu\xi\, C_{-\omega-j}(x_\|,0,\xi;z)\,,\end{aligned}\tag{A5}$$

where the limit is taken from values of $\mathrm{Re}(z)<-1$.

A similar equation holds for $D_{-\omega-j}$;

$$\begin{aligned}\lim_{y_\|\to x_\|}&\int_{\mathbf{R}^{\nu-1}}d^{\nu-1}\xi_\|\,[D_{-\omega-j}(x_\|,0,\xi_\|,0;z=-1)-|\xi_\||^{1-\omega-j}D_{-\omega-j}(x_\|,0,\xi_\|/|\xi_\||,0;z=-1)]e^{i\xi_\|(x-y)_\|}\\&=\int_{|\xi_\||<1}d^{\nu-1}\xi\, D_{-\omega-j}(x_\|,0,\xi_\|,0;z=-1)-\frac{1}{(1-\omega-j+\nu)}\int_{|\xi_\||=1}d^{\nu-2}\xi_\|\, D_{-\omega-j}(x_\|,0,\xi_\|,0;z=-1)\\&=\lim_{z\to-1}\int_{\mathbf{R}^{\nu-1}}d^{\nu-1}\xi_\|\, D_{-\omega-j}(x_\|,0,\xi_\|,0;z).\end{aligned}\tag{A6}$$

For $j=\nu-\omega$:

$$\begin{aligned}\lim_{y_\|\to x_\|}&\int_{\mathbf{R}^\nu}d^\nu\xi\, C_{-\nu}(x_\|,0,\xi;z=-1)e^{i\xi_\|(x-y)_\|}\\&=\int_{|\xi|<1}d^\nu\xi\, C_{-\nu}(x_\|,0,\xi;z=-1)+\lim_{y_\|\to x_\|}\int_{|\xi|>1}d^\nu\xi\, C_{-\nu}(x_\|,0,\xi;z=-1)e^{i\xi_\|(x-y)_\|}\end{aligned}\tag{A7}$$

where the last term includes a $(x-y)$ homogeneous function of degree zero and (possibly) a logarithmic term. (For a little more explicit expression, see Ref. 8).

On the other hand,

$$\begin{aligned}\int_{\mathbf{R}^\nu}d^\nu\xi\, C_{-\nu}(x_\|,0,\xi;z)&=\int_{|\xi|<1}d^\nu\xi\, C_{-\nu}(x_\|,0,\xi;z=-1)-\frac{1}{\omega}\int_{|\xi|=1}d^{\nu-1}\xi\frac{d}{dz}C_{-\nu}(x_\|,0,\xi;z)|_{z=-1}\\&\quad-\frac{1}{\omega(z+1)}\int_{|\xi|=1}d^{\nu-1}\xi\, C_{-\nu}(x_\|,0,\xi;z=-1)+o(z+1)\,.\end{aligned}\tag{A8}$$

An analogous analysis holds for the term involving $D_{-\nu}$ in (A2) and (A3).

Finally, taking into account that

$$\lim_{z\to-1}R(x,y;z)=\lim_{y\to x}R(x,y;z=-1) \tag{A9}$$

[property (i)], we have

$$\begin{aligned}
&\lim_{z\to -1}\Big\{D_B^z(x_\|,0,y_\|,0)\\
&\quad -\frac{1}{(z+1)}\Big[-\frac{1}{[\omega(2\pi)^\nu]}\int_{|\xi|=1} d^{\nu-1}\xi\, C_{-\nu}(x_\|,0,\xi;-1)+\frac{1}{[\omega(2\pi)^\nu]}\int_{|\xi_\||=1} d^{\nu-2}\xi\, D_{-\nu}(x_\|,0,\xi_\|,0;-1)\Big]\Big\}\\
&\quad = -\frac{1}{[\omega(2\pi)^\nu]}\int_{|\xi|=1} d^{\nu-1}\xi\,\frac{d}{dz}C_{-\nu}(x_\|,0,\xi;z)|_{z=-1}+\frac{1}{[\omega(2\pi)^\nu]}\int_{|\xi_\||=1} d^{\nu-2}\xi\, D_{-\nu}(x_\|,0,\xi_\|,0;z)|_{z=-1}\\
&\qquad +\lim_{y\to x}\Big\{G(x_\|,0,y_\|,0)-\sum_{j=0}^{\nu-\omega-1}\frac{1}{(2\pi)^\nu}\Big[\int_{\mathbf{R}^\nu} d^\nu\xi\, C_{-\omega-j}(x_\|,0,\xi/|\xi|;-1)|\xi|^{-\omega-j}e^{i\xi_\|(x-y)_\|}\\
&\qquad -\int_{\mathbf{R}^{\nu-1}} d^{\nu-1}\xi\, D_{-\omega-j}(x_\|,0,\xi_\|/|\xi_\||,0;-1)|\xi_\||^{-\omega-j+1}e^{i\xi_\|(x-y)_\|}\Big]\\
&\qquad -\frac{1}{(2\pi)^\nu}\int_{|\xi|>1} d^\nu\xi\, C_{-\nu}(x_\|,0,\xi;-1)e^{i\xi_\|(x-y)_\|}+\frac{1}{(2\pi)^\nu}\int_{|\xi_\||>1} d^{\nu-1}\xi\, D_{-\nu}(x_\|,0,\xi_\|,0;-1)e^{i\xi_\|(x-y)_\|}\Big\}, \qquad \text{(A10)}
\end{aligned}$$

which is the required expression relating the finite part of the diagonal element of the kernel of D_B^z for $z=-1$ with the Green function of D_B.

[1] A. Chodos and C. B. Thorn, Phys. Rev. D **12**, 2733 (1975); T. Inowe and T. Maskawa, Progr. Theor. Phys. **54**, 1833 (1975); G. E. Brown and M. Rho, Phys. Lett. B **82**, 177 (1979).

[2] See, e.g., M. Rho, "Cheshire cat phenomena and quarks in nuclei," Saclay preprint, SPth/86-159 (November 1986), and references therein; S. Nadkarni, H. B. Nielsen, and I. Zahed, Nucl. Phys. B **253**, 308 (1985); S. Nadkarni and H. B. Nielsen, *ibid.* **263**, 1 (1986); S. Nadkarni and I. Zahed, *ibid.* **263**, 23 (1986).

[3] R. E. Gamboa Saravi, F. A. Schaposnik, and J. E. Solomin, Nucl. Phys. B **181**, 239 (1981); A. Polyakov and P. V. Weigmann, Phys. Lett. B **141**, 128 (1983); **141**, 223 (1984).

[4] R. E. Gamboa Saravi, M. A. Muschietti, F. A. Schaposnik, and J. E. Solomin, Ann. Phys. **157**, 360 (1984).

[5] H.Falomir, M. A. Muschietti, and E. M. Santangelo, Phys. Lett. B **205**, 93 (1988).

[6] H. Falomir, M. A. Muschietti, and E. M. Santangelo, Phys. Rev. D **37**, 1677 (1988).

[7] R. T. Seeley, Am. J. Math. **91**, 889 (1969); **91**, 963 (1969).

[8] R. E. Gamboa Saravi, M. A. Muschietti, F. A. Schaposnik, and J. E. Solomin, J. Math. Phys. **26**, 2045 (1985).

[9] S. Agmon, Comm. Pure Appl. Math. **15**, 119 (1962).

[10] R. Durhuus, P. Olesen, and J. L. Petersen, Nucl. Phys. B **198**, 157 (1982).

[11] J. Wess and B. Zumino, Phys. Lett. B **37**, 95 (1971); E. Witten, Commun. Math. Phys. **92**, 455 (1984).

[12] J. Goldstone and R. L. Jaffe, Phys. Rev. Lett. **51**, 1518 (1983); T. H. Hanson and R. L. Jaffe, Phys. Rev. D **28**, 882 (1983); I. Zahed, *ibid.* **30**, 2221 (1984).

[13] R. Forman, Invent. Math. **88**, 447 (1987).

Generalized Luneburg canonical varieties and vector fields on quasicaustics

S. Janeczko[a),b)]
Max-Planck-Institut für Mathematic, Gottfried-Claren-Strasse 26, 5300 Bonn, West Germany

(Received 2 December 1988; accepted for publication 22 November 1989)

Some aspects of a particular class of bifurcation varieties which are provided by simple and unimodal boundary singularities are studied. Their correspondence to diffraction theory is established. The generic caustics by diffraction on apertures are derived and their generating families for the corresponding Lagrangian varieties are calculated. It is proved that the quasicaustics associated to simple singularities are smooth hypersurfaces or Whitney's cross-caps. The procedure for calculating the modules of logarithmic vector fields is given, and the minimal sets of the corresponding generators are explicitly calculated. The general boundary singularities are constructed and the structure of quasicaustics defined by parabolic singularities is investigated.

I. INTRODUCTION

Let F: $(\mathbb{C}^{n+1}\times\mathbb{C}^p,0)\rightarrow(\mathbb{C},0)$ be a germ of a holomorphic function. By $(S,0)\subset(\mathbb{C}^{n+1},0)$ we denote a germ of a some hypersurface in $(\mathbb{C}^{n+1},0)$. The quasicaustic $Q(F)$ of F is defined as

$$Q(F)=\{a\in\mathbb{C}^p;\ F(\cdot,a)\ \text{ has a critical point on } S\}.$$

Let F represent the distance function from the general wave front in the presence of an obstacle formed by an aperture (cf. Refs. 1 and 2) with boundary S. The corresponding quasicaustic $Q(F)$ is build up from the rays orthogonal to the given wave front and touching the boundary of the aperture (see the example of the quasicaustic illustrated in Fig. 1). The quasicaustic is a subvariety of the usual caustic (also called the bifurcation set[3,4])

$$\{a\in\mathbb{C}^p;\ F(\cdot,a)\ \text{ or }\ F|_{s\times C^p}(\cdot,a)\ \text{ have a critical point}\},$$

and represents the structure of shadows formed by the common, pecular positions of aperture and incident wave front.

In this paper we investigate the structure of generic caustics and quasicaustics by diffraction on smooth obstacle curves and apertures (optical instruments). We use for this the classical phase space for general optical instruments, i.e., the space of pairs of rays $(l,\tilde{l})$, where l is an incident ray and $\tilde{l}$ is a transformed ray (produced by l and the optical instrument), endowed with the canonical symplectic structure. This space was first introduced by Luneburg[5] in his mathematical theory of optics and then revived by Guillemin and Sternberg[6] in their symplectic approach to various physical theories. To each optical instrument, in the mentioned phase space, there corresponds a Lagrangian subvariety, say A, defining all physical properties (from the point of view of the geometrical theory of optics[7]) of the system. So when A is fixed we can obtain all transformed wave fronts by taking the symplectic images $A(L)$ of all Lagrangian subvarieties L of incident rays (i.e., optical sources). (See, also, Ref. 8.)

The plan of the paper is as follows. In Sec. II we give preliminary results about the basic phase spaces and construct representative examples in the symplectic approach to general optical systems. The geometrical structure of caustics by diffraction on apertures, as well as their generic classification in the case of half-line aperture on the plane and half-plane aperture in Euclidean three-space, is investigated in Sec. III. We compute the normal forms for generating families of the generic canonical varieties in the case of diffraction on smooth curves in Sec. IV. When considering the caustics by diffraction on apertures, the quasicaustic component becomes important. In Sec. V we generalize the methods for ordinary caustics initiated by Bruce[9,10] to investigate the structure of logarithmic vector fields on quasicaustics. In Sec. VI we derive the generators for the modules of tangent vector fields to the quasicaustics corresponding to simple

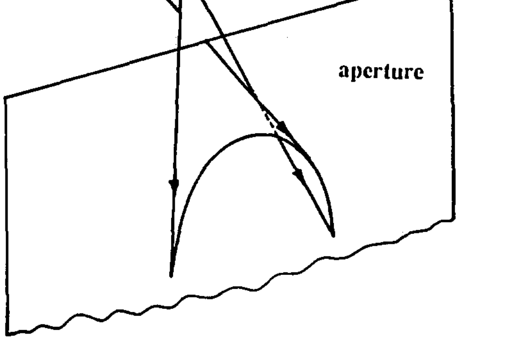


FIG. 1. Whitney's cross-cap quasicaustic.

[a)] On leave of absence from Institute of Mathematics, Technical University of Warsaw, Pl. Jednosci Robotniczej 1, 00-661 Warszawa, Poland.

[b)] This paper was partially written while the author was a SERC visitor to the Mathematics Institute, University of Warwick.

0022-2488/90/040997-13$03.00

boundary singularities and prove that they are not free. Finally in Sec. VI we analyze the structure of quasicaustics and the reduction of functional moduli in normal forms of Lagrangian pairs.

II. SINGULARITIES IN ACTION OF OPTICAL INSTRUMENTS

Let (M, ω) be the symplectic manifold of all oriented lines in $V \cong \mathbb{R}^3$. We look on V as the configurational space of geometrical optics with refraction index n: $V \to \mathbb{R}$, $n \equiv 1$. Here (M, ω) is given by the standard symplectic reduction

$$\pi_M : H^{-1}(0) \to M \cong T^*S^2,$$

where the hypersurface $H^{-1}(0)$ is defined by the Hamiltonian

$$H: T^*V \to \mathbb{R}, \quad H(p,q) := \tfrac{1}{2}(\|p\|^2 - 1),$$

and π_M is the projection along characteristics of the associated Hamiltonian system.

Let (p,q) be coordinates on (T^*V, ω_V), where ω_V is an associated Liouville two-form. By (U,ω) we denote the local chart on (M,ω) described as an image $\pi_M(H^{-1}(0) \cap \{p_1 > 0\})$ with restricted symplectic form ω. The (p,q) form Darboux coordinates on (T^*V, ω_V). In corresponding Darboux coordinates (r,s) on (U,ω) we can write

$$(r,s) = \pi_M(p_2,p_3;q_1,q_2,q_3)$$
$$= \left(p_2,p_3;\, q_2 - \frac{q_1 p_2}{\sqrt{1 - p_2^2 - p_3^2}}, q_3 - \frac{q_1 p_3}{\sqrt{1 - p_2^2 - p_3^2}}\right),$$

where the unique reduced symplectic structure ω is given by the formula

$$\omega_V|_{H^{-1}(0)} = \pi_M^* \omega, \quad \omega|_U = \sum_{i=1}^{2} dr_i \wedge ds_i.$$

In the introduced coordinates on M, to each point $(r,s) \in U$ we can uniquely associate the corresponding ray (in parametric form):

$$(q_1,q_2,q_3)$$
$$= (0,s_1,s_2) + u\left(1, \frac{r_1}{\sqrt{1 - r_1^2 - r_2^2}}, \frac{r_2}{\sqrt{1 - r_1^2 - r_2^2}}\right),$$
$$u \in \mathbb{R}.$$

By the above formula one can translate the concrete optical problems into the language of the phase space (M,ω) and vice versa (cf. Refs. 5, 6, and 11).

Let (U,ω) and $(\widetilde{U},\widetilde{\omega})$ be two examples of the symplectic space of optical rays or its open subsets. Usually these manifolds denote the spaces of incident and transformed rays of an optical instrument.

Definition 2.1: The phase space of optical instruments is the following product symplectic manifold:

$$\Pi = (U \times \widetilde{U};\, \pi_2^* \widetilde{\omega} - \pi_1^* \omega),$$

where $\pi_{1,2} : U \times \widetilde{U} \to U, \widetilde{U}$ are canonical projections (this was first introduced by Luneburg[5]).

The process of optical transformation (say, reflection, refraction, or diffraction, etc., of the incident rays) is governed by the subvariety of Π, which is Lagrangian, i.e., it is stratified onto isotropic submanifolds of Π where maximal strata are Lagrangian (cf. Refs. 8, 12, and 13).

Definition 2.2: We define the general optical instrument to be a Lagrangian subvariety of Π (generalized symplectic relation[8,14]).

Remark 2.3: It is easily seen that reflecting or refracting optical instruments (cf. Ref. 15) correspond to graphs of symplectomorphisms between $(U;\omega)$ and $(\widetilde{U},\widetilde{\omega})$. But, for example, the diffraction process is described by a quite general Lagrangian subvariety of Π (cf. Ref. 1). In fact, let $(a,b,x,y,u,v,w) \to F(a,b,x,y,u,v,w)$ be the optical distance function (cf. Refs. 2 and 16) from the wave front

$$\{z = \varphi(x,y) = \lambda_1 x^2 + \lambda_2 xy + \lambda_3 y^2 + O_3(x,y)\}$$

in the presence of the aperture $\{a \geqslant 0, z = mb - 1\}$, where $m \geqslant 0$. If the incident ray goes from $(x,y) = (0,0)$ to $(a,b) = (0,0)$, then the transformed rays from $(a,b) = (0,0)$ to (u,v,w) are given by

$$\frac{\partial F}{\partial b}(0,u,v,w) = \quad 0,$$

$$\overline{F}(b,x,y,u,v,w) := F(0,b,x,y,u,v,w),$$

which, for the distance function

$$F = [(x-a)^2 + (y-b)^2 + (\varphi(x,y) - mb + 1)^2]^{1/2}$$
$$+ ((u-a)^2 + (v-b)^2 + (w - mb + 1)^2)^{1/2},$$

reads

$$m^2u^2 + v^2(m^2 - 1) - 2mv(1 + w) = 0$$

and

$$v + m(1 + w) \leqslant 0.$$

These conditions define the half-cone of diffracted rays (see Refs. 1 and 7).

Example 2.4: Reflection from the curve: Let the mirror be defined by $\{q_1 = 0\}$. Let (U,ω), the space of incident rays, be defined as $\pi_M(H^{-1}(0) \cap \{p_1 > 0\})$ and the corresponding space of reflected rays be defined as $\widetilde{U} = \pi_M(H^{-1}(0) \cap \{p_1 < 0\})$. Then this reflecting optical instrument is equivalent to the Lagrangian subvariety of Π,

$$\Pi \supset \{((r,s),(\tilde{r},\tilde{s})) \in U \times \widetilde{U};\, r = \tilde{r}, s = \tilde{s}\} = :A,$$

and its corresponding generating family (cf. Refs. 17–19)

$$G(\lambda,s,\tilde{s}) = \lambda(s - \tilde{s}),$$

where $\lambda \in \mathbb{R}$, is a Morse parameter.

In our approach the sources of radiation produce rays in the space denoted by (U,ω). Thus we have the following definition.

Definition 2.5: We define the general source of light as a Lagrangian subvariety $L \subset (U,\omega)$ of the space of incident rays. If $A \subset \Pi$ is an optical instrument, then the transformed system of rays [or equivalently the transformed wave front (cf. Ref. 18)] is a symplectic image L' of L by means of A, i.e.,

$$L' := A(L) := \{\tilde{p} \in \widetilde{U};\text{ there exists } p \in L \text{ such that } (p,\tilde{p}) \in A\},$$

which is usually a Lagrangian subvariety of $(\widetilde{U},\widetilde{\omega})$ (cf. Ref. 8).

Example 2.6: Reflection of a parallel beam of rays: The beam of parallel rays is given in (U,ω) by $L=\{r=0\}$ (a point source of light at infinity). By reflection in the mirror, $x\rightarrow(\varphi(x),x)\in\mathbb{R}^2, \varphi(0)=\varphi'(0)=0, \varphi''(0)\neq 0$, the canonical variety $A\subset\Pi$ (defining the reflection process) brings into L some focusing property and produces the well known caustic. The reflected beam of rays $A(L)$ has the form

$$(\tilde{r},\tilde{s})=\left(\frac{2\varphi'(x)}{\varphi'(x)^2+1},x-\frac{\varphi(x)\varphi'(x)(1+\varphi'(x)^2)^2}{\varphi'(x)^2-1}\right).$$

Remark 2.7: Local genericity of the wave front produced by $L\subset(U,\omega)$ is preserved during the process of reflection or refraction (cf. Ref. 15) because the corresponding canonical variety is a graph of symplectomorphism. Thus the caustics, produced by reflection or refraction, are classified by the simple singularities of type A_k, D_k, E_k.[20] It may not be so in a diffraction process, where $A\subset\Pi$ is no longer the graph of symplectomorphism. In this case the differentiable structure of L is drastically changed by A and $A(L)$ is no longer smooth. Its singular locus brings a completely new type of caustic responsible for the structure of shadows and half-shadows of an obstacle as well.

III. CAUSTICS AND QUASICAUSTICS BY DIFFRACTION

Let L be a source of light or transformed wave front in (M,ω). Now we recall the geometric construction that allows us to define caustic or wave front evolution in V, corresponding to L (cf. Refs. 12 and 13). Let Ξ be the product symplectic manifold

$$\Xi=(M\times T^*V,\pi_2^*\omega_V-\pi_1^*\omega),$$

where $\pi_{1,2}: M\times T^*V\rightarrow M, T^*V$ are the canonical projections. One can check that $\widetilde{K}:=\text{graph }\pi_M\subset\Xi$ is a Lagrangian submanifold of Ξ. Thus there exists its local generating Morse family (cf. Ref. 17), say,

$$K: \mathbb{R}^k\times\widetilde{X}\times V\rightarrow\mathbb{R}\ (\mu,\tilde{x},q)\rightarrow K(\mu,\tilde{x},q),$$

where $T^*\widetilde{X}$ is an appropriate local cotangent bundle structure (special symplectic structure,[12–14] on (M,ω). The transformed system of rays forms a Lagrangian subvariety of (T^*V,ω_V) given as an image

$$\widetilde{L}=(\widetilde{K}\circ A)(L)\subset(T^*V,\omega_V),$$

where $\widetilde{K}\circ A\subset\Xi$ is a composition of symplectic relations (cf. Refs. 12 and 17). If

$$G: \mathbb{R}^l\times X\times\widetilde{X}\rightarrow\mathbb{R},\ (\nu,x,\tilde{x})\rightarrow G(\nu,x,\tilde{x}),\quad X,\widetilde{X}\cong\mathbb{R}^n,$$

is a generating family for $A\subset\Pi$ and F: $\mathbb{R}^m\times X\rightarrow\mathbb{R}$, $(\lambda,x)\rightarrow F(\lambda,x)$ is a generating family for L, then the transformed Lagrangian subvariety $\widetilde{L}\subset(T^*V,\omega_V)$ is generated by (not necessarily a Morse family)

$$\widetilde{F}: \mathbb{R}^{k+l+m+2n}\times V\rightarrow\mathbb{R},$$
$$\widetilde{F}(\lambda,\nu,\mu,x,\tilde{x};q):=G(\nu,x,\tilde{x})+K(\mu,\tilde{x},q)+F(\lambda,x),$$

where $\mathbb{R}^{k+l+m+2n}$ is a parameter space.

In optical arrangements the source of light is usually a smooth Lagrangian submanifold of (U,ω). Only after the transformation process through an optical instrument does it become singular.

Definition 3.1: Let $L\subset(U,\omega)$ be an initial source variety. We define the caustic by an optical instrument $A\subset\Pi$, to be a hypersurface of V formed by two components: (1) singular values of $\pi_V|_{\widetilde{L}-\text{Sing}\widetilde{L}}$; and (2) π_V (Sing $\widetilde{L}$); where $\widetilde{L}=(\widetilde{K}\circ A)(L)$ and Sing $\widetilde{L}$ denotes the singular locus of $\widetilde{L}$.

Remark 3.2: In reflection or refraction we do not go beyond the smooth category of L (at least in this paper) so the associate caustics, in transformed wave fronts $\widetilde{L}$, are those realizable by smooth generic sources (cf. Refs. 15 and 21). Thus in what follows we will be interested in caustics caused by diffraction, which will enrich substantially the list of optical events (cf. Ref. 22) and complete the correspondence between singularities of functions and groups generated by reflections.[16,23]

Diffracted rays are produced, for example, when an incident ray hits an edge of an impenetrable screen [i.e., an edge of a boundary or interface (cf. Ref. 1)]. In this case the incident ray produces infinitely many diffracted rays, which have the same angle with the edge as does the incident ray (see Remark 2.3.) This is so if both incident and diffracted rays lie in the same medium. Otherwise, the angles between the two rays and the plane normal to the edge are related by Snells law.[7] Furthermore, the diffracted ray lies on the opposite side of the normal plane from the incident ray; that is, all rules and laws of geometrical optics correspond exactly to the Lagrangian properties of the corresponding varieties $A\subset\Pi$.

Let I be the diagonal in Π. By Ω we denote the set of oriented lines in (U,ω) that do not intersect the screen. Thus we have the following proposition.

Proposition 3.3: In the edge diffraction in an arbitrary Euclidean space, the canonical variety $A\subset\Pi$ has two components

$$A=A^I\cup A^D,$$

where $A^I=\Omega\times\Omega\subset I$ and A^D is a pure diffraction of rays passing through the edge of an aperture, defined in Remark 2.3.

Corollary 3.4: Let $L\subset(U,\omega)$ be an incident system of rays. Then the edge diffracted system of rays,

$$\widetilde{L}=(\widetilde{K}\circ A)(L),$$

is a regular intersection (cf. Ref. 24) of two smooth components: $\widetilde{L}_1=(\widetilde{K}\circ A^I)(L)$ and $\widetilde{L}_2=(\widetilde{K}\circ A^D)(L)$, i.e.,

$$\widetilde{L}=\widetilde{L}_1\cup\widetilde{L}_2,\quad \dim\widetilde{L}_1\cap\widetilde{L}_2=\dim\widetilde{L}_1-1,$$
$$T_x(\widetilde{L}_1\cap\widetilde{L}_2)=T_x\widetilde{L}_1\cap T_x\widetilde{L}_2.$$

Thus we see that the caustic caused by the edge diffraction has three components: (1) the caustic of $\widetilde{L}_1$, which is a part of the caustic in incident wave front L; (2) the caustic, purely by diffraction on the edge, i.e., the caustic of $\widetilde{L}_2$; and (3) the image $\pi_V(\widetilde{L}_1\cap\widetilde{L}_2)$ of the rays passing exactly through an edge.

Definition 3.5: The set $\pi_V(\widetilde{L}_1\cap\widetilde{L}_2)\subset V$ is called the quasicaustic by diffraction on aperture. The rays belonging to the quasicaustic that are contained in the aperture plane we will call the rays at infinity.

Usually the quasicaustics describe the structure of shadows and half-shadows in configurational space V (see Fig. 1).

Proposition 3.6: (1) Generic caustics by diffraction on the half-line aperture on the plane are diffeomorphic to the $\widetilde{A}_2$, $\widetilde{A}_3$, $B_2 \cong C_2$, B_3 boundary caustics. Normal forms for their generating families as images $A(L)$ [or pairs (A,L) in general position] are the following:

$$\widetilde{A}_2\colon\ -\tfrac{1}{3}\lambda^3 + \lambda(q_2 - a) - \tfrac{1}{2}q_1\lambda^2,\quad a>0,\qquad \text{and } A\colon = \{q_1 = 0,\ q_2 \leqslant 0\};$$

$$\widetilde{A}_3\colon\ -\tfrac{1}{4}\lambda^4 + \lambda(q_2 - a) - \tfrac{1}{2}q_1\lambda_2,\quad a>0,\qquad \text{and } A\colon = \{q_1 = 0,\ q_2 \leqslant 0\};$$

$$B_2\colon\ -\tfrac{1}{2}\lambda^2 + q_2\lambda - \tfrac{1}{2}q_1\lambda^2,\quad \{\lambda \geqslant 0\},\qquad \text{and } A\colon = \{q_1 = 0,\ q_2 \leqslant 0\};$$

$$B_3\colon\ -\tfrac{1}{3}\lambda^3 - \tfrac{1}{2}q_1\lambda^2 + \lambda(q_2 - q_1 a),\quad \{\lambda \geqslant 0\},\qquad \text{and } A\colon = \{q_1 = 2a,\ q_2 \leqslant 2a^2\},\quad a>0;$$

where λ is a Morse parameter and a is the moduli of the common position.

(2) In generic one-parameter families of caustics by diffraction on the half-line aperture, which do not pass through infinity, the only possible configurations are those described in metamorphoses of optical caustics (see Ref. 21, p. 113, and Ref. 25) and the additional cases illustrated in Fig. 2.

Proof: It is easily seen that $\widetilde{K} = \text{graph}\ \pi_M \subset \Xi$ is generated locally by

$$K(r,q_1,q_2) = q_2 r - \tfrac{1}{2}q_1 r^2.$$

The only stable systems of rays $\widetilde{K}(L) \subset (T^*V,\omega_V)$ are generated in (M,ω) by $L\colon = \{(r,s);\ s = -(\partial F/\partial r)(r)\}$, where

$$A_1\colon\ F_1(r) = -\tfrac{1}{2}r^2,$$

$$A_2\colon\ F_2(r) = -\tfrac{1}{3}r^3,$$

$$A_3\colon\ F_3(r) = -\tfrac{1}{4}r^4 \quad \text{(cf. Refs. 18 and 21)}.$$

Let the aperture be defined in its normal form by $q_1 = 0$, $q_2 \leqslant 0$ (so $A \subset \Pi$). Thus we have the boundary singularities (cf. Ref. 26) $A(L)$ defined in $(\widetilde{M},\widetilde{\omega})$ by the following generating functions:

$$\widetilde{A}_1\colon\ \widetilde{F}_1(\widetilde{r}) = -\tfrac{1}{2}\widetilde{r}^2,\quad \{\widetilde{r} \geqslant 0\};$$

$$\widetilde{A}_2\colon\ \widetilde{F}_2(\widetilde{r}) = -\tfrac{1}{3}\widetilde{r}^3,\quad \widetilde{r} \in \mathbb{R};$$

$$\widetilde{A}_3\colon\ \widetilde{F}_3(\widetilde{r}) = -\tfrac{1}{4}\widetilde{r}^4,\quad \{\widetilde{r} \geqslant 0\}.$$

Taking A_i in the general position with respect to A we obtain part (1) of Proposition 3.6. Part (2) follows by checking all the possible one-parameter evolutions (where the quasicaustic is not passing through infinity) of the stable caustic on the plane and in the presence of the half-line aperture. Two possible directions of intersection of the A_2 caustic by an edge of the aperture give us the cases (a) and (b) in Fig. 2. The evolution of an edge of the aperture passing through the ray tangent to the cusp caustic A_3 is illustrated in Fig. 2(c). Finally an evolution through the intersection point of the $A_2 + A_2$ caustic gives us the case of Fig. 2(d). This completes the proof of Proposition 3.6. □

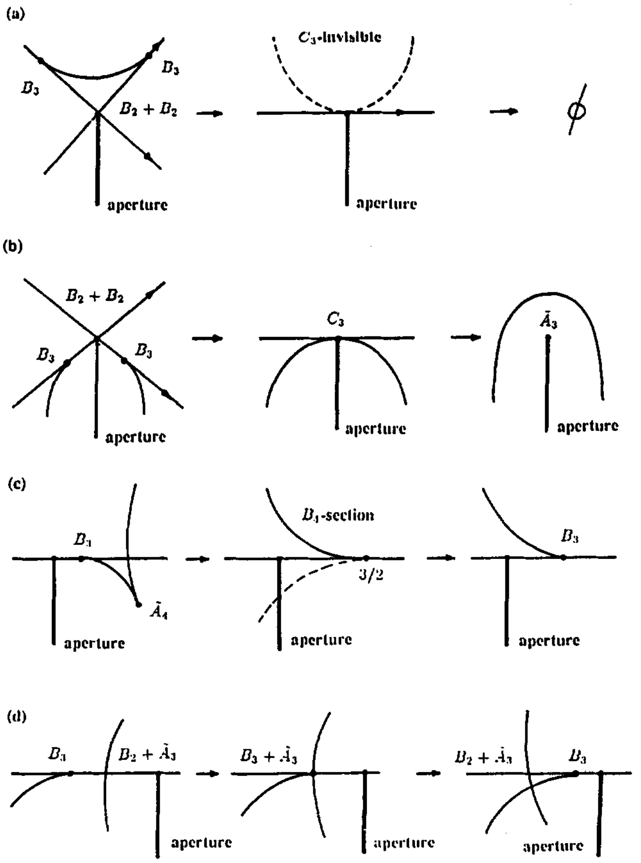


FIG. 2. Transformations of caustics in the presence of aperture.

Looking at the position of the quasicaustic in the diffraction problem with a half-plane aperture in $\mathbb{R}^3$ we can eliminate the C_4-boundary caustic. Thus we have the following proposition.

Proposition 3.7: Generic caustics by diffraction on the half-plane aperture in $\mathbb{R}^3$ are diffeomorphic to the $\widetilde{A}_2$, $\widetilde{A}_3$, $\widetilde{A}_4$, B_2, B_3, B_4, F_4 boundary caustics.

Remark 3.8: (1) For the general linear hyperbolic system of first order (cf. Ref. 7),

$$\mathscr{L}u = u_t + \sum_{\nu=1}^{3} A^{\nu}\frac{\partial u}{\partial x_\nu} + Bu = 0,$$

where u represents, say, in the case of crystal optics, the pair of vectors (E,H), and $\mathscr{L}u = 0$ corresponds to Maxwell's equations. In the geometrical optics approximation, we obtain another characteristic equation (eikonal equation)

$$\det\left(\Phi_t + \sum_{\nu=1}^{3} A^{\nu}\frac{\partial\Phi}{\partial x_\nu}\right) = 0,$$

for the phase function $\Phi(x,t)$; $u \sim e^{i\omega\Phi(x,t)}a^0(x,t)$. In this case the conical refraction in crystal optics is an example of a

Lagrangian variety quite generally situated in the associated phase space (cf. Refs. 6 and 7).

(2) In the edge diffraction on system of apertures (mentioned in Ref. 1) the singularities of the distance function are classified by the singularities on the many-dimensional corners.[27] In very constrained systems of apertures the classification is obtained using the methods of the theory of singularities of functions on singular varieties (cf. Refs. 9 and 16).

(3) The generic quasicaustic in the edge diffraction in $\mathbb{R}^3$, corresponding to the F_4 singularity of the distance function (cf. Ref. 25), is realized geometrically (see Fig. 1) when the curve of rays passing through the edge on the incident wave front is tangent to a constant curvature line on the wave front. This situation is generic (cf. Ref. 21).

IV. DIFFRACTION ON SMOOTH OBSTACLES

Now we can apply an introduced symplectic framework to describe the diffraction on smooth closed surfaces in $\mathbb{R}^3$. The problem is connected to the Riemannian obstacle problem (cf. Ref. 28), i.e., determination of geodesics on a Riemannian manifold with smooth boundary. Any geodesic on such a manifold is C^1 and consists of generically finitely many so-called switchpoints, where the geodesic has an initial or end point according to whether it lies in the interior part of the manifold or on the boundary. Cauchy uniqueness for manifolds with a boundary states that every boundary point (point of an obstacle) has a neighborhood in which, if two geodesic segments with the same initial point, initial tangent vector, and length do not coincide, then one of them has its right end point in the interior part of the manifold and is an involute of the other (in the planar case it lies on an appropriate involute of the obstacle curve). A geodesic γ' that has the same initial point, initial tangent vector, and length as γ is called an involute of a geodesic γ. The reformulation of the above obstacle problem in terms of geometrical optics of diffraction needs a definition of a surface diffracted ray. A surface diffracted ray is produced when a ray is incident tangentially on a smooth boundary or interface. It is a geodesic on the surface in the metric nds, where n is the refractive index of the medium on the side of the surface containing the incident ray. At every point it sheds a diffracted ray along its tangent (cf. Refs. 1 and 22). A surface diffracted ray is also produced on the second side of an interface by a ray incident from the first side at the critical angle [arcsin (n_1/n_2)]. In this case at every point it sheds rays back toward the first side at the critical angle. However, in what follows we will neglect these rays.

Let us consider an open subset S of an obstacle surface in $\mathbb{R}^3$. Let l_1 be the initial tangent line to the geodesic segment γ on S, and let l_2 be a tangent line to S. We say that l_2 is subordinate to l_1 with respect to an obstacle S if l_2 [or its piece in $(\mathbb{R}^3,S)$] belongs to the geodesic segment with the same initial point and the same tangent vector as γ has. By simple checking we have the following (cf. Ref. 18).

Proposition 4.1: Let γ be a geodesic flow on S. Then the set

$$A=\{(l,\tilde{l})\in\Pi;\ \tilde{l}\text{ is subordinate to }l$$
$$\text{with respect to }S\text{ and geodesic flow }\gamma\}$$

is a Lagrangian subvariety of Π defining the diffraction process on an obstacle S.

Now we look for the generic pairs (A,L). At first we consider the planar case.

Proposition 4.2: For the generic obstacle curve on the plane the only possible canonical varieties $A\subset\Pi$ have the following normal forms of generating families (or functions):

$$\widetilde{A}_2\text{: } G(r,\tilde{r})=-\tfrac{1}{12}(r^3+\tilde{r}^3),\ \ (\text{obstacle curve } q_2=-q_1^2),$$
$$\widetilde{H}_3\text{: } G(\lambda_1,\lambda_2,r,\tilde{r})=\tfrac{9}{10}(\lambda_1^5+\lambda_2^5)-r\lambda_1^3-\tilde{r}\lambda_2^3$$
$$+\tfrac{1}{2}r^2\lambda_1+\tfrac{1}{2}\tilde{r}^2\lambda_2,$$
$$(\text{obstacle curve } q_2=q_1^3),$$
$$\widetilde{A}_{2,2}\text{: } G(r,\tilde{r})=\tfrac{1}{2}(r|r|+\tilde{r}|\tilde{r}|),\ \ (\text{double tangent}).$$

Proof: Let us take the noninflection point of the generic curve. Parametrically the curve is given as $(q_1,q_2)=(v,-v^2)$, $v\in\mathbb{R}$, and the corresponding family of tangent lines corresponding to the given incident ray has the form

$$(q_1,q_2)=(0,v^2)+u(1,-2v),\quad u\in\mathbb{R}.$$

By identification

$$s=v^2,\quad r=4v/(1+4v^2),$$
$$\tilde{s}=\tilde{v}^2,\quad \tilde{r}=4\tilde{v}/(1+4\tilde{v}^2),$$

where $(v,\tilde{v})\in\mathbb{R}^2$ parametrize the variety A, we obtain the case $\widetilde{A}_2$ that corresponds to the Cartesian product of two ordinary folds. Taking the inflection point for an obstacle curve, we obtain, in the same way, the following parametrization for $A\subset\Pi$:

$$s=-2v^3,\quad r=\frac{3v^2}{\sqrt{1+9v^4}},\quad \tilde{s}=-2\tilde{v}^3,\quad \tilde{r}=\frac{3\tilde{v}^2}{\sqrt{1+9\tilde{v}^4}}.$$

After straightforward calculations we obtain the generating family for it, denoted by $\widetilde{H}_3$. Analogously we obtain the $\widetilde{A}_{2,2}$ case. □

Corollary 4.3: For (A,L) in the general position we have the possible stable images $A(L)\subset(\widetilde{M},\widetilde{\omega})$,

$$A_2\text{: } \widetilde{F}_1(\tilde{r})=-\tfrac{1}{12}\tilde{r}^3,$$
$$H_3\text{: } \widetilde{F}_2(\lambda,\tilde{r})=\tfrac{9}{10}\lambda^5-\tilde{r}\lambda^3+\tfrac{1}{2}\tilde{r}^2\lambda,$$
$$A_{2,2}\text{: } \widetilde{F}_3(\tilde{r})=\tfrac{1}{2}|\tilde{r}|\tilde{r},$$

and the generating families for their corresponding configurational images,

$$K(A_2)\text{: } F_1(\lambda,q_1,q_2)=-\tfrac{1}{12}\lambda^3+q_2\lambda-\tfrac{1}{2}q_1\lambda^2,$$
$$K(H_3)\text{: } F_2(\lambda_1,\lambda_2,q_1,q_2)=\tfrac{9}{10}\lambda_1^5-\lambda_2\lambda_1^3+\tfrac{1}{2}\lambda_2^2\lambda_1$$
$$+q_2\lambda_2-\tfrac{1}{2}q_1\lambda_2^2,$$
$$K(A_{2,2})\text{: } F_3(\lambda,q_1,q_2)=\tfrac{1}{2}\lambda|\lambda|+q_2\lambda-\tfrac{1}{2}q_1\lambda^2$$

[see Figs. 3(a)–3(c) and also the figures in Ref. 22].

Proof: In the general position of A and L, only one point

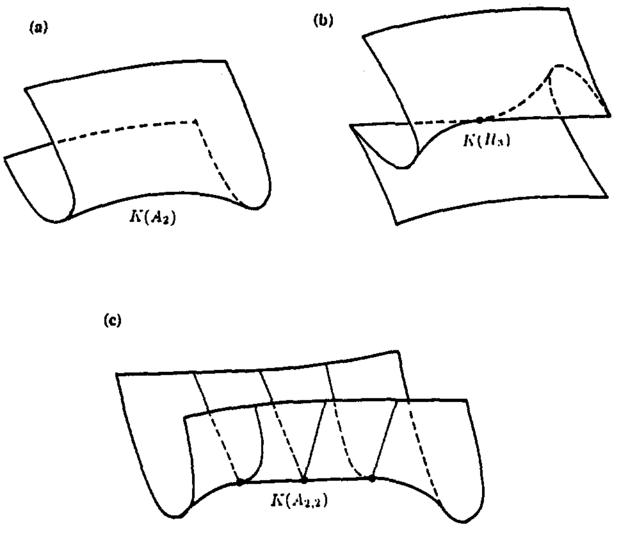


FIG. 3. Canonical varieties for the generic obstacle curve on the plane.

of L is tangent to an obstacle curve in the neighborhood of the considered point of this curve. Hence in the calculation of $(\widetilde{K}\circ A)(L)$ in all the cases $(\widetilde{A}_2, \widetilde{H}_3$, and $\widetilde{A}_{2,2})$ it is necessary to put $r=$ const in generating families of Proposition 4.2. □

Remark 4.4: (A) The first, most important, results in obstacle geometry and its correspondence to the structure of singular orbits of H_3 and H_4 group actions were discovered by Shcherbak.[16] The aim of the present paper is to show how singular wave front evolutions appear in the general setting of the mathematical theory of optics (cf. Refs. 5, 6, and 18) and to complete the investigations of the caustics and quasicaustics that appear there. As we see, the planar obstacle problem is connected to the studies of tangent developables. More degenerated singularities there can be described using the blowing-up construction (cf. Ref. 29).

(B) The $K(A_{2,2})$ singularity appeared as an adjacent to the higher singular one (see Fig. 4) in a generic one-parameter family of obstacles

$$q_2 = -\tfrac{1}{4}q_1^4 + \tfrac{1}{2}aq_1^2 - \tfrac{1}{4}a^2, \quad a\in\mathbb{R}_+,$$

i.e.,

$$r = -2av_\epsilon - 3\epsilon\sqrt{a}v_\epsilon^2 + (4a^3-1)v_\epsilon^3 + O(v_\epsilon^4),$$

$$s = 2\epsilon a^{3/2}v_\epsilon + 4av_\epsilon^2 + 3\epsilon\sqrt{a}v_\epsilon^3 + \tfrac{3}{4}v_\epsilon^4,$$

$\epsilon = \pm 1,\ v_+ \geqslant 0,\ v_- \leqslant 0.$

(C) We can see that by choosing the special symplectic structure fibered over (p_1,p_2) in the H_3 case, we can investigate only a cuspidal edge of $A(L)$. In fact, with its generating family

$$F_2'(\lambda,\mu,p) = F_2(\lambda,\mu) - \mu_1p_1 - \mu_2p_2,$$

after reduction of the μ_1, μ_2, and λ_2 parameters, we obtain the generating family for the H_2 singularity,

$$F_2'(\lambda,p) = \tfrac{9}{10}\lambda^5 - p_2\lambda^3 + \tfrac{1}{2}p_2^2\lambda,$$

and its level sets (wave fronts) as in Table 2 of Ref. 16. This observation is connected with the much more general feature of obstacle singular wave front evolutions; namely, all singularities in obstacle geometry as indicated in Table 2 of Ref. 16 are generated by the generalized open swallowtails [in $(\widetilde{M},\tilde{\omega})$ space] with generating family (see Ref. 8, p. 106)

$$\widetilde{A}_{2(k+1)}: \int_0^\lambda \left(x^{k+1} + \sum_{i=2}^{k+1} \bar{s}_{i-1}x_{k-i+1}\right)^2 dx.$$

The Ξ_l $(l\geqslant 1)$, Δ_l $(l\geqslant 2)$ (cf. Ref. 16) singular wave front evolutions are reconstructed from $\widetilde{A}_{2(k+1)}$ singularities by specifying appropriate common generic positions of $A\subset \Pi$ and $\widetilde{A}_{2(k+1)} \subset (\widetilde{M},\tilde{\omega})$.

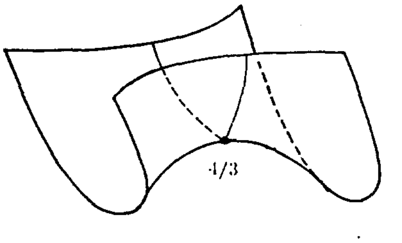


FIG. 4. The higher-order singularity of a canonical variety.

V. VECTOR FIELDS ON CAUSTICS AND QUASICAUSTICS

As we can see from the preceding sections, caustics in the wave front evolution, or in a diffracted wave front on the aperture, are defined as bifurcation sets for the corresponding generating family (Morse family[2,17]) of functions or the family of functions on the manifold with boundary, respectively (cf. Refs. 20 and 26). To investigate the structure of these sets and modules of tangent vector fields on them, in what follows we shall consider the real analytic or holomorphic functions (germs). For the ordinary caustics, defined as the critical values of the Lagrange projections (cf. Ref. 20) from the Lagrangian submanifolds, which are not necessarily fibered by optical rays, the procedure is the following.[3,4]

Let f: $(\mathbb{C}^n,0)\to(\mathbb{C},0)$ be a holomorphic function of finite codimension, i.e., the dimension of the quotient $\mathscr{O}_{(x)}/J(f)$ as a complex vector space is finite, where $\mathscr{O}_{(x)}$ denotes the ring of holomorphic functions h: $(\mathbb{C}^n,0)\to(\mathbb{C},0)$ and $J(f)$ is the ideal in $\mathscr{O}_{(x)}$ generated by the partial derivatives $\partial f/\partial x_1,\dots,\partial f/\partial x_n$. Let $\mathscr{M}_{(x)}$ denote the maximal ideal in $\mathscr{O}_{(x)}$. If $g_1,\dots,g_p$ is a basis for $\mathscr{M}_{(x)}/J(f)$, then

$$F: (\mathbb{C}^n\times\mathbb{C}^p,0)\to(\mathbb{C},0),$$

$$F(x,a) = f(x) + \sum_{i=1}^p a_ig_i(x)$$

is a miniversal unfolding of f (cf. Ref. 30).

The caustic of F [or bifurcation set of F (see Refs. 4 and 9)] is the following set (germ):

$$B(F) = \{a\in\mathbb{C}^p;\ F_a \text{ has a degenerate critical point}\}.$$

The set of critical values of π: $(\Sigma F,0)\to(\mathbb{C}^p,0)$ (π is a canonical projection on the second factor), where

$$\Sigma F = \left\{(x,a)\in\mathbb{C}^n\times\mathbb{C}^p;\ \frac{\partial F}{\partial x_1} = \cdots = \frac{\partial F}{\partial x_n} = 0\right\},$$

is the caustic. It appears to be important to know the modules of tangent vector fields to caustics (as well as to wave fronts,[9,20,31] which is easier). They are useful in the reduction of functional moduli in the classification of generic symmetric and nonsymmetric Lagrangian submanifolds (cf. Ref. 20, p. 344, and Ref. 32). We recall some necessary definitions from Refs. 3 and 33. The set of germs of holomorphic vector fields on $\mathbb{C}^p$, at 0, tangent to the nonsingular part of $B(F)$, is called the set of logarithmic vector fields of $B(F)$ at 0. It is denoted also by Derlog $B(F)$. In Refs. 4, 9, and 10 (see, also, Refs. 31 and 33) a general method for computing these vector fields was given. It was shown that A_k singularities are the only ones whose module of tangent vector fields to $B(F)$ is free (i.e., caustic is a free divisor[33]). Applying the method used in these papers we investigate the modules of vector fields tangent to the quasicaustics in diffraction on apertures (this is a first step in the investigation of the structure of caustics by diffraction).

Let $\mathscr{O}_{(y,x)}$ denote the ring of holomorphic functions h: $(\mathbb{C}\times\mathbb{C}^n,0)\rightarrow(\mathbb{C},0)$. The hypersurface $S=\{y=0\}$ corresponds to the boundary of an aperture. Following the general scheme used in Ref. 20 for boundary singularities, we shall consider holomorphic functions f: $(\mathbb{C}\times\mathbb{C}^n,0)\rightarrow(\mathbb{C},0)$ of finite codimension, i.e.,

$$\dim_{\mathbb{C}} \mathscr{O}_{(y,x)}/\Delta(f)<\infty,$$

where

$$\Delta(f)=\left\langle y\frac{\partial f}{\partial y},\frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n}\right\rangle$$

denotes the ideal in $\mathscr{O}_{y,x}$ generated by the partial derivatives $\partial f/\partial x_1,\dots,\partial f/\partial x_n$ and $y\,\partial f/\partial y$ (cf. Refs. 20 and 34). Let $g_0,\dots,g_{\mu-1}$ form a basis for $\mathscr{O}_{(y,x)}/\Delta(f)$ with $g_0=1$ and $g_i\in\mathscr{M}_{(y,x)}$. Then the miniversal deformation, in the category of deformations of functions on the manifold with a boundary, as a Morse family for the corresponding diffracted Lagrangian variety (cf. Refs. 13 and 24) is defined as follows:

$$F\colon (\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^{\mu-1},0)\rightarrow(\mathbb{C},0),$$

$$F(y,x,a)=f(y,x)+\sum_{i=1}^{\mu-1}a_ig_i(y,x).$$

Proposition 5.1: The caustic (or bifurcation set) from diffraction on the aperture, having the generating family F: $(\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^p,0)\rightarrow(\mathbb{C},0)$ (p is not necessarily minimal) of functions on the manifold with boundary (extended edge) has three components

(1) $B_1(F)=\{a\in\mathbb{C}^p;\ F(\cdot,\cdot,a)$ has a degenerate critical point$\}$,

(2) $B_2(F)=\{a\in\mathbb{C}^p;\ F(0,\cdot,a)$ has a degenerate critical point$\}$,

(3) $Q(F)=\{a\in\mathbb{C}^p;\ F(\cdot,\cdot,a)$ has a critical point on $S=\{y=0\}\}$.

Proof: By Corollary 3.4, we have the three isotropic submanifolds defining the system of diffracted rays $\widetilde{L}_1$, $\widetilde{L}_2$, and $\widetilde{L}_1\cap\widetilde{L}_2$. It is easily seen that in terms of the generating family/distance function F, the corresponding caustics can be written in forms (1)–(3) of Proposition 5.1. □

The set (germ)

$$(\Sigma_r\,F,0)=\left(\left\{(x,a)\in\mathbb{C}^n\times C^p;\ \left.\frac{\partial F}{\partial y}\right|_{s\times\mathbb{C}^p}=\left.\frac{\partial F}{\partial x_1}\right|_{s\times\mathbb{C}^p}=\cdots=\left.\frac{\partial F}{\partial x_n}\right|_{s\times\mathbb{C}^p}=0\right\},0\right)$$

is called the restricted critical set.

Using the Splitting Lemma[30] and the versality property of F, we have the following proposition.

Proposition 5.2: (A) The restricted critical set $(\Sigma_r\,F,0)$ is the germ of a smooth manifold of dimension $p-1$.

(B) The quasicaustic of F, $(Q(F),0)$, is an image of $(\Sigma_r\,F,0)$ by the natural projection π: $\Sigma_r\,F,0\rightarrow\mathbb{C}^p,0$ to the second factor.

The set of logarithmic vector fields of $Q(F)$ at 0 is defined (cf. Refs. 3 and 33) to be the set of germs of holomorphic vector fields on $\mathbb{C}^p$ at 0, tangent to the nonsingular part of $Q(F)$; it is an $\mathscr{O}_{(a)}$ module.

Proposition 5.3: Let $\xi\in$Derlog $Q(F)$, then it is π liftable, i.e., for some germ of a vector field $\tilde{\xi}$, on $\mathbb{C}^n\times\mathbb{C}^p$, tangent to $\Sigma_r\,F$ at 0, we have

$$\xi\circ\pi=d\pi\circ\tilde{\xi}.$$

Proof: ξ lifts uniquely by π at every point $a\in\mathbb{C}^p-\Gamma(\pi|_{\Sigma_r F})$. Hence ξ lifts to a holomorphic vector field $\tilde{\xi}_1$ on $\mathbb{C}^n\times\mathbb{C}^p$, tangent to $\Sigma_r\,F$ and defined off a set of codimension 2 in $\mathbb{C}^n\times\mathbb{C}^p$. By Hartog's theorem, $\tilde{\xi}_1$ extends to a holomorphic vector field $\tilde{\xi}$ tangent to $\Sigma_r\,F$.

Now using the π-lowerable vector fields $\tilde{\xi}$ tangent to $\Sigma_r\,F$ we will construct the module Derlog $Q(F)$. Let F be as above. We define the ideal

$$I(F)=\left\langle\psi(x,a),\frac{\partial\overline{F}}{\partial x_1}(x,a),\dots,\frac{\partial\overline{F}}{\partial x_n}(x,a)\right\rangle\mathscr{O}_{(x,a)},$$

where ψ and $\overline{F}$ are given by decomposition:

$$F(y,x,a)=F(0,x,a)+y\psi(x,a)+y^2g(y,x,a),$$

$$\overline{F}(x,a):=F(0,x,a).$$

Let

$$\tilde{\xi}=\sum_{i=1}^{n}\beta_i\frac{\partial}{\partial x_i}+\sum_{i=1}^{p}\gamma_i\frac{\partial}{\partial a_i},\quad \beta_i,\gamma_i\in\mathscr{O}_{(x,a)},$$

be the germ of a vector field at $0\in\mathbb{C}^n\times\mathbb{C}^p$, tangent to Σ_rF. Then we have

$$\tilde{\xi}\left(\frac{\partial F}{\partial y}(0,x,a)\right)\in I(F)$$

and

$$\tilde{\xi}\left(\frac{\partial F}{\partial x_i}(0,x,a)\right)\in I(F),\quad i=1,\dots,n.$$

For our

$$F(y,x,a)=f(y,x)+\sum_{i=1}^{\mu-1}a_ig_i(y,x),$$

we have

$$\psi(x,a)=\frac{\partial f}{\partial y}(0,x)+\sum_{i=1}^{\mu-1}a_i\frac{\partial g_i}{\partial y}(0,x).$$

So we need

$$\sum_{i=1}^{n}\beta_i\frac{\partial\psi}{\partial x_i}+\sum_{i=1}^{\mu-1}\gamma_i\frac{\partial g_i}{\partial y}\Big|_{0\times\mathbb{C}^n}\in I(F)$$

and

$$\sum_{i=1}^{n}\beta_i\frac{\partial^2\overline{F}}{\partial x_i\,\partial x_j}+\sum_{i=1}^{\mu-1}\gamma_i\frac{\partial\overline{g}_i}{\partial x_j}\in I(F),\quad 1\leqslant j\leqslant n,$$

where $\overline{g}(x):=g(0,x)$. Thus we obtain the following lemma.

Lemma 5.4: $\tilde{\xi}$ is a lifting of $\xi\in\mathrm{Derlog}\ Q(F)$,

$$\xi=\sum_{i=1}^{p}\alpha_i(a)\frac{\partial}{\partial a_i},$$

if and only if, for some $\beta_i\in\mathscr{O}_{(x,a)}$ $(i=1,\ldots,n)$, we have

$$\begin{gathered}\sum_{i=1}^{n}\beta_i\frac{\partial\psi}{\partial x_i}+\sum_{i=1}^{\mu-1}\alpha_i\frac{\partial g^i}{\partial y}\Big|_{0\times\mathbb{C}^n}\in I(F),\\ \sum_{i=1}^{n}\beta_i\frac{\partial^2\overline{F}}{\partial x_i\,\partial x_j}+\sum_{i=1}^{\mu-1}\alpha_i\frac{\partial\overline{g}_i}{\partial x_j}\in I(F).\end{gathered}\tag{5.1}$$

We have chosen the normal form for F in such a way that the variables $a_\mu,\ldots,a_p$ $(p\geqslant\mu-1)$ do not appear in F. Now, following the general scheme used in Refs. 3 and 10 for ordinary bifurcation sets, we can propose a procedure for constructing the tangent vector fields to quasicaustics.

By the Preparation Theorem,[20,30] the module

$$\mathscr{O}_{(y,x,a)}/\overline{\Delta}(F),$$

where

$$\overline{\Delta}(F)=\left\langle y\frac{\partial F}{\partial y},\frac{\partial F}{\partial x_1},\ldots,\frac{\partial F}{\partial x_n}\right\rangle\mathscr{O}_{(y,x,a)},$$

is a free $\mathscr{O}_{(a)}$ module[20] generated by $1,g_1,\ldots,g_{\mu-1}$. So, for any $h\in\mathscr{O}_{(y,x,a)}$, we can write

$$\begin{aligned}h(y,x,a)&=\beta(y,x,a)y\frac{\partial F}{\partial y}(y,x,a)\\&\quad+\sum_{i=1}^{n}\beta_i(y,x,a)\frac{\partial F}{\partial x_i}(y,x,a)\\&\quad+\sum_{j=1}^{\mu-1}a_j(a)g_j(y,x)+\alpha(a),\end{aligned}\tag{5.2}$$

for some $\beta_i\in\mathscr{O}_{(y,x,a)}$, $\alpha_j\in\mathscr{O}_{(a)}$, $\alpha\in\mathscr{O}_{(a)}$.

By straightforward checking we obtain the following proposition.

Proposition 5.5: Let $h\in\mathscr{O}_{(y,x,a)}$ satisfy

$$\frac{\partial h}{\partial y}\Big|_{0\times\mathbb{C}^n\times\mathbb{C}^p}\in I(F),\quad \frac{\partial h}{\partial x_i}\Big|_{0\times\mathbb{C}^n\times\mathbb{C}^p}\in I(F),\quad i=1,\ldots,n.$$

Then the vector field

$$\xi=\sum_{i=1}^{p}\alpha_i\frac{\partial}{\partial a_i},$$

where α_i, $i=1,\ldots,\mu-1$, are defined in (5.2) and α_i, $i=\mu,\ldots,p$, are arbitrary holomorphic functions from $\mathscr{O}_{(a)}$, is tangent to the quasicaustic $Q(F)=\pi(\Sigma_rF)$. Conversely, suppose

$$\xi=\sum_{i=1}^{p}\alpha_i\frac{\partial}{\partial a_i}$$

is tangent to $Q(F)$. Then there is some $h\in\mathscr{O}_{(y,x,a)}$ as above with

$$h=\sum_{i=1}^{n}\beta_i\frac{\partial F}{\partial x_i}+\beta y\frac{\partial F}{\partial y}+\sum_{i=1}^{\mu-1}\alpha_ig_i+\alpha,$$

and

$$\frac{\partial h}{\partial x_i}\Big|_{0\times\mathbb{C}^n\times\mathbb{C}^p}\in I(F),\quad \frac{\partial h}{\partial y}\Big|_{0\times\mathbb{C}^n\times\mathbb{C}^p}\in I(F).\qquad\square$$

We see that the set of all such h with $(\partial h/\partial y)|_{\overline{s}}\in I(F)$, $(\partial h/\partial x_i)|_{\overline{s}}\in I(F)$, $1\leqslant i\leqslant n$, form an $\mathscr{O}_{(a)}$ module: in fact, it is the kernel of the $\mathscr{O}_{(a)}$ module homomorphism,

$$\Phi:\ \mathscr{O}_{(y,x,a)}\ni h\to\left(\frac{\partial h}{\partial y},\frac{\partial h}{\partial x_1},\ldots,\frac{\partial h}{\partial x_n}\right)\in\left(\frac{\mathscr{O}_{(y,x,a)}}{I(F)+\langle y\rangle\mathscr{M}_{(y,x,a)}}\right)^{n+1}.$$

Here, $\overline{\Delta}(F)\subset I(F)+\langle y\rangle\mathscr{M}_{(y,x,a)}$ and clearly the set of tangent vector fields to $Q(F)$ is a finitely generated $\mathscr{O}_{(a)}$ module. We denote $\overline{s}=\mathscr{O}\times\mathbb{C}^n\times\mathbb{C}^p$.

VI. QUASICAUSTICS OF SIMPLE AND UNIMODAL BOUNDARY SINGULARITIES

The simple singularities of functions on the boundary $\{y=0\}$ of a manifold with a boundary were classified in Ref. 20, p. 281. Their miniversal unfoldings are

$\widetilde{A}_\mu$: $\pm y\pm x^{\mu+1}+\sum_{i=1}^{\mu-1}a_ix^i,\quad \mu\geqslant1;$

B_μ: $\pm y^\mu\pm x^2+\sum_{i=1}^{\mu-1}a_i\,y^{\mu-i},\quad \mu\geqslant2;$

C_μ: $yx\pm x^\mu+\sum_{i=1}^{\mu-1}a_ix^{\mu-i},\quad \mu\geqslant2;$

$\widetilde{D}_\mu$: $\pm y+x_1^2x_2\pm x_2^{\mu-1}+\sum_{i=1}^{\mu-2}a_ix_2^i+a_{\mu-1}x_1,\quad \mu\geqslant4;$

$\widetilde{E}_6$: $\pm y+x_1^3\pm x_2^4+a_1x_1+a_2x_2+a_3x_2^2+a_4x_1x_2+a_5x_1x_2^2;$

$\widetilde{E}_7$: $\pm y+x_1^3+x_1x_2^3+a_1x_1+a_2x_2+a_3x_2^2+a_4x_1x_2+a_5x_2^3+a_6x_2^4;$

$\widetilde{E}_8$: $\pm y+x_1^3+x_2^5+a_1x_1+a_2x_2+a_3x_2^2+a_4x_1x_2+a_5x_2^3+a_6x_1x_2^2+a_7x_1x_2^3;$

F_4: $\pm y^2+x^3+a_2y+a_3x+a_1xy.$

Thus we have, after direct checking, the following proposition.

Proposition 6.1: The quasicaustics for simple boundary singularities are

$\widetilde{A}_\mu,\ \widetilde{D}_\mu,\ \widetilde{E}_k$: $Q(F)=\varnothing,$

B_μ: $Q(F)=\{a\in\mathbb{C}^{\mu-1};\ a_{\mu-1}=0\},$

C_μ: $Q(F)=\{a\in\mathbb{C}^{\mu-1};\ a_{\mu-1}=0\},$

F_4: $Q(F)=\{a\in\mathbb{C}^3;\ a_2^2+\frac{1}{3}a_1^2a_3=0\}$

(i.e., Whitney's cross-cap, see Fig. 1).

Thus we need to calculate only the module of vector fields tangent to $Q(F_4)$. Let us define the germ, at zero, of the variety $X := Q(F_4)\cup\{a_1=0\}$. We see that the vector fields tangent to $(X,0)$ lie in Derlog $Q(F_4)$.

Proposition 6.2: The vector fields

$$V_1 = -\frac{1}{6}a_1^2\frac{\partial}{\partial a_2}+a_2\frac{\partial}{\partial a_3},$$

$$V_2 = a_1\frac{\partial}{\partial a_1}+a_2\frac{\partial}{\partial a_2},$$

$$V_3 = -\frac{1}{3}a_1\frac{\partial}{\partial a_1}+a_3\frac{2}{3}\frac{\partial}{\partial a_3}$$

form a free basis for the $\mathscr{O}_{(a)}$ module Derlog X.

Before we prove this theorem we need the following proposition.

Proposition 6.3: For corank-2 boundary singularities F: $(\mathbb{C}\times\mathbb{C}\times\mathbb{C}^p,0)\rightarrow(\mathbb{C},0)$, the space of functions $h\in\mathscr{O}_{(y,x,a)}$ reconstructing the $\mathscr{O}_{(a)}$ module of vector fields tangent to the quasicaustic $Q(F)$ has the form

$$h(y,x,a)=\int_0^x\left(\frac{\partial F}{\partial y}(0,s,a)\psi_1(s,a)+\frac{\partial F}{\partial x}(0,s,a)\psi_2(s,a)\right)ds+y^2\xi(y,x,a),$$

where $\psi_i\in\mathscr{O}_{(x,a)}$ $(i=1,2)$, $\xi\in\mathscr{O}_{(y,x,a)}$.

Proof: Every function $h\in\mathscr{O}_{(y,x,a)}$ can be written in the form

$$h(y,x,a)=\eta_2(x,a)+y\eta_1(x,a)+y^2\eta(y,x,a),$$

and thus

$$\frac{\partial h}{\partial y}(0,x,a)=\eta_1(x,a),\quad \frac{\partial h}{\partial x}(0,x,a)=\frac{\partial\eta_2}{\partial x}(x,a).$$

By Proposition 5.5, we can take

$$\eta_1(x,a)\in I(F)\ \text{ and }\ \eta_2(x,a)=\int_0^x g(s,a)ds,\quad g\in I(F),$$

obtaining all functions

$$\eta_2(x,a)+y\eta_1(x,a)+y^2\eta(v,x,a)(\text{mod}\ \overline{\Delta}(F)),$$

defining the $\mathscr{O}_{(a)}$ module of vector fields tangent to $Q(F)$. Now we see that

$$\eta_2(x,a)+y\eta_1(x,a)+y^2\eta(y,x,a)=\eta_2(x,a)+y^2\xi(y,x,a)\left(\text{mod}\left\langle y\frac{\partial F}{\partial y},y\frac{\partial F}{\partial x}\right\rangle\mathscr{O}_{(y,x,a)}\right),$$

where $\xi\in\mathscr{O}_{(y,x,a)}$. Adding an element of $\langle y\rangle\overline{J}(F)$ [$\overline{J}(F)$ is an ideal of $\mathscr{O}_{(y,x,a)}$ generated by $\partial F/\partial y,\partial F/\partial x_1,\ldots,\partial F/\partial x_n$] does preserve the space of functions and does not affect the resulting vector field.

Proof of Proposition 6.2: $I(F_4)=\langle a_1x+a_2,3x^2+a_3\rangle\mathscr{O}_{(x,a)}$. By Proposition 6.3, taking $\psi_1,\psi_2,\xi\equiv 1$, we have

$$\begin{aligned}h_1(x,a)&=\tfrac{1}{2}a_1x^2+a_2x\\&=-\tfrac{1}{6}a_1^2y+a_2x-\tfrac{1}{6}a_1a_3(\text{mod}\ \overline{\Delta}(F_4)),\\h_2(x,a)&=y^2=-a_1xy-a_2y(\text{mod}\ \overline{\Delta}(F_4)),\\h_3(x,a)&=x^3+xa_3=-\tfrac{1}{3}a_1xy+\tfrac{2}{3}a_3x(\text{mod}\ \overline{\Delta}(F_4)).\end{aligned}$$

Then the corresponding V_i belongs to Derlog $Q(F_4)$ $(i=1,2,3)$. By simple computation we obtain

$$V_1(a_1)=0,\quad V_2(a_1)=-a_1,\quad V_3(a_1)=-\tfrac{1}{3}a_1;$$

Thus $V_i\in$Derlog X as well. We also have that

$$\det(V_1(a),V_2(a),V_3(a))=-\tfrac{1}{3}a_1(a_2^2+\tfrac{1}{3}a_3a_1^2)$$

is a reduced equation for $(X,0)$; thus, by the results of Saito[33] (see, also, Ref. 3), we find that $(X,0)$ is a free divisor. □

We define the following ideals of $\mathscr{O}_{(y,x)}$ and $\mathscr{O}_{(y,x,a)}$, respectively:

$$\Theta(f)=\langle y\rangle J(f)+\left\langle\frac{\partial f}{\partial x_1},\ldots,\frac{\partial f}{\partial x_n}\right\rangle^2\mathscr{O}_{(y,x)}$$

and

$$\overline{\Theta}(F)=\langle y\rangle\overline{J}(F)+\left\langle\frac{\partial F}{\partial x_1},\ldots,\frac{\partial F}{\partial x_n}\right\rangle^2\mathscr{O}_{(y,x,a)}.$$

For determining all fields tangent to the quasicaustic we need the following lemma.

Lemma 6.4: The space $\mathscr{O}_{(y,x)}/\Theta(f)$ is finite dimensional. Its $\mathbb{C}$ basis also generates the quotient space $\mathscr{O}_{(y,x,a)}/\overline{\Theta}(F)$ as an $\mathscr{O}_{(a)}$ module.

Proof: $\Theta(f)\supset\Delta(f)$ and f is finitely determined as a boundary singularity. Thus $\mathscr{O}_{(y,x)}/\Theta(f)$ is $\mathbb{C}$-finite dimensional with the basis $\{g_1,\ldots,g_N\}$. Let us define the mapping

$$\Psi\colon(\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^p,0)\rightarrow(\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^{n(n+1)/2}\times\mathbb{C}^p,0),$$

$$\Psi(y,x,a)=\left(y\frac{\partial F}{\partial y}(y,x,a),y\frac{\partial F}{\partial x_1}(y,x,a),\ldots,y\frac{\partial F}{\partial x_n}(y,x,a),\frac{\partial F}{\partial x_i}(y,x,a)\frac{\partial F}{\partial x_j}(y,x,a),a\right),$$

with $1\leqslant i,j\leqslant n$, $i\leqslant j$, and ordered set of pairs (i,j). Thus we have

$$\mathscr{O}_{(y,x,a)}/\Psi^*(\mathscr{M}_{(y,x,a)})\mathscr{O}_{(y,x,a)}\cong\mathscr{O}_{(y,x)}/\Theta(f)\mathscr{O}_{(y,x)}.$$

By the Preparation Theorem,[30] every element h of $\mathscr{O}_{(y,x,a)}$ has the form

$$h(y,x,a)=\sum_{i=1}^N\phi_i\left(y\frac{\partial F}{\partial y}(y,x,a),y\frac{\partial F}{\partial x_1}(y,x,a),\ldots,y\frac{\partial F}{\partial x_n}(y,x,a),\frac{\partial F}{\partial x_i}(y,x,a)\frac{\partial F}{\partial x_j}(y,x,a),a\right)\times g_i(y,x).$$

Thus

$$\mathscr{O}_{(y,x,a)}/\overline{\Theta}(F)\cong\left\{\sum_{i=1}^N\psi_i(a)g_i(y,x)\right\},\quad\psi_i\in\mathscr{O}_{(a)},$$

which completes the proof of Lemma 6.4.

Let $\{g_1,\ldots,g_N\}$ be a $\mathbb{C}$ basis for $\mathscr{O}_{(y,x)}/\Theta(f)$. In general we have the following proposition.

Proposition 6.5: Functions $h\in\mathscr{O}_{(y,x,a)}$, which reconstruct the $\mathscr{O}_{(a)}$ module of vector fields tangent to $Q(F)$, can be written as

$$h(y,x,a)=\sum_{i=1}^N\alpha_i(a)g_i(y,x),$$

where

$$\sum_{i=1}^{N} \alpha_i(a) \frac{\partial g_i}{\partial y}(0,x) \in I(F),$$

$$\sum_{i=1}^{N} \alpha_i(a) \frac{\partial g_i}{\partial x_j}(0,x) \in I(F),$$

$1 \leqslant j \leqslant n.$

Proof: By Lemma 6.4, any $h \in \mathscr{O}_{(y,x,a)}$ can be written as

$$h(y,x,a) = \sum_{i=1}^{N} \alpha_i(a) g_i(y,x) + \beta(y,x,a) y \frac{\partial F}{\partial y}(y,x,a) + \sum_{j=1}^{n} \beta_j(y,x,a) y \frac{\partial F}{\partial x_j}(y,x,a) + \sum_{k,l=1}^{n} \beta_{k,l}(y,x,a) \frac{\partial F}{\partial x_k}(y,x,a) \frac{\partial F}{\partial x_l}(y,x,a),$$

where $\alpha_i \in \mathscr{O}_{(a)}$, $\beta, \beta_j, \beta_{kl} \in \mathscr{O}_{(y,x,a)}$. By simply checking the assumption of Proposition 5.5, we see that the three last terms in the above formula do not affect on the resulting vector field belonging to Derlog $Q(F)$. This proves Proposition 6.5. □

Proposition 6.6: The $\mathscr{O}_{(a)}$ module Derlog $Q(F_4)$, i.e., the module of holomorphic vector fields tangent to Whitnery's cross-cap, is generated by the vector fields

$$\begin{aligned} V_1 &= -\frac{1}{6} a_1^2 \frac{\partial}{\partial a_2} + a_2 \frac{\partial}{\partial a_3}, \\ V_2 &= a_1 \frac{\partial}{\partial a_1} + a_2 \frac{\partial}{\partial a_2}, \\ V_3 &= -\frac{1}{3} a_1 \frac{\partial}{\partial a_1} + \frac{2}{3} a_3 \frac{\partial}{\partial a_3}, \\ V_4 &= a_2 \frac{\partial}{\partial a_1} - \frac{1}{3} a_1 a_3 \frac{\partial}{\partial a_2}, \end{aligned} \tag{6.1}$$

which satisfy the relation

$$a_1 V_4 - 2a_3 V_1 + 3a_2 V_3 = 0.$$

Proof: We have

$$\Theta(f) = \langle y^2, x^2 y, x^4 \rangle \mathscr{O}_{(y,x)}$$

and

$$\mathscr{O}_{(y,x)} / \Theta(f) \cong [1, x, y, x^2, x^3, xy]_{\mathbb{C}}.$$

By Proposition 6.5 all functions $h \in \mathscr{O}_{(y,x,a)}$ leading to the construction of Derlog $Q(F_4)$ can be written in the form

$$h(y,x,a) = \alpha_1(a) + \alpha_2(a)x + \alpha_3(a)x^2 + \alpha_4(a)x^3 + \alpha_5(a)y + \alpha_6(a)xy,$$

where $\alpha_i \in \mathscr{O}(a)$, $i = 1,...,6$, are such that

$$\begin{aligned} &\alpha_5(a) + \alpha_6(a)x \in I(F_4), \\ &\alpha_2(a) + 2\alpha_3(a)x + 3\alpha_4(a)x^2 \in I(F_4) \quad \text{(see Sec. V)}. \end{aligned} \tag{6.2}$$

By simple calculations we check that V_i, $i = 1,...,4$, are tangent to Whitney's cross-cap. Calculations using power series or a homogeneous filtration show that these are the only vector fields generating Derlog $Q(F_4)$. In fact,

$$h = \alpha_1 - \tfrac{1}{3}\alpha_3 a_3 + (\alpha_2 - \tfrac{1}{3}\alpha_4 a_3)x + (\alpha_5 - \tfrac{1}{3}\alpha_3 a_1)y + (\alpha_6 - \tfrac{1}{3}\alpha_4 a_1)xy (\text{mod } \Delta(F_4)).$$

Hence all vector fields belonging to Derlog $Q(F_4)$ can be written in the form

$$V = \alpha_6 \frac{\partial}{\partial a_1} + \alpha_5 \frac{\partial}{\partial a_2} + \alpha_5' \frac{\partial}{\partial a_3} - \frac{1}{6} a_1 \alpha_6' \frac{\partial}{\partial a_2} + \alpha_4 V_3, \tag{6.3}$$

where $\alpha_4, \alpha_5, \alpha_6, \alpha_5', \alpha_6' \in \mathscr{O}(a)$ satisfy

$$\alpha_5 + \alpha_6 x \in I(F_4), \quad \alpha_5' + \alpha_6 x \in I(F_4), \tag{6.4}$$

which are a simple rewritten version of (6.2). Here we use the formula $x^2 = -\frac{1}{3} a_3 (\text{mod } I(F_4))$. Solving (6.4) using a power series, we obtain an expression for (6.3) that involves only V_i, $i = 1,2,3,4$. □

Proposition 6.5 gives an algorithm for calculating all vector fields tangent to quasicaustics corresponding to boundary singularities. Now we restrict our attention to quasicaustics corresponding to the unimodal boundary singularities.

Let us consider the miniversal deformations for parabolic boundary singularities[20]:

$$F_{1,0}: \; y^3 + x^3 + a_1 y^2 x + a_2 xy + a_3 y^2 + a_4 y + a_5 x,$$

$$K_{4,2}: \; y^2 + x^4 + a_1 y x^2 + a_2 xy + a_3 x^2 + a_4 x + a_5 y,$$

$$D_{4,1} (= L_6): \; \tfrac{1}{2} x_1^2 x_2 + \tfrac{1}{3} x_2^3 + y x_1 + a_1 y x_2 + \tfrac{1}{2} a_2 x_2^2 + a_3 y + a_4 x_1 + a_5 x_2,$$

where a_1 is a modulus parameter. The Milnor number of these deformations is 6 and the boundary is $\{y = 0\}$. We treat these three cases separately, starting with $F_{1,0}$.

Proposition 6.7: The module Derlog $Q(F_{1,0})$ is not free and is generated by the following vector fields:

$$\begin{aligned} V_1 &= -\frac{1}{6} a_2^2 \frac{\partial}{\partial a_4} + a_4 \frac{\partial}{\partial a_5}, \\ V_2 &= a_2 \frac{\partial}{\partial a_2} + a_4 \frac{\partial}{\partial a_4}, \\ V_3 &= -a_2 \frac{\partial}{\partial a_2} + 2a_5 \frac{\partial}{\partial a_5}, \\ V_4 &= a_4 \frac{\partial}{\partial a_2} - \frac{1}{3} a_2 a_5 \frac{\partial}{\partial a_4}, \\ V_5 &= \frac{\partial}{\partial a_1}, \quad V_6 = \frac{\partial}{\partial a_3}. \end{aligned} \tag{6.5}$$

Proof: We have

$$I(F) = \langle a_2 x + a_4, 3x^2 + a_5 \rangle \mathscr{O}_{(x,a)}$$

and

$$\mathscr{O}_{(y,x)} / \Theta(f) = [1, x, y, x^2, x^3, xy, y^2, xy^2]_{\mathbb{C}}.$$

Thus

$$h = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \alpha_4 x^3 + \alpha_5 y + \alpha_6 xy + \alpha_7 y^2 + \alpha_8 xy^2.$$

The equations

$$\left.\frac{\partial h}{\partial y}\right|_{y=0} = \alpha_5 + \alpha_6 x \in I(F),$$

$$\left.\frac{\partial h}{\partial x}\right|_{y=0} = \alpha_2 + 2\alpha_3 x + 3\alpha_4 x^2 \in I(F)$$

reduce the calculations to those in the proof of Proposition 6.6. □

In remaining cases we only need to calculate the one-jets of vector fields generating the module. We now treat the case $K_{4,2}$.

Proposition 6.8: All vector fields belonging to Derlog $Q(K_{4,2})$ have the following form:

$$\left(\alpha_9-\frac{1}{2}a_1\alpha_5+U\alpha_6\right)\frac{\partial}{\partial a_1}+\left(\alpha_8-\frac{1}{6}a_1\alpha_4'\right.$$
$$\left.-\frac{1}{4}a_2\alpha_5+\alpha_6\left(V-\frac{5}{12}a_1a_3\right)\right)\frac{\partial}{\partial a_2}$$
$$+\left(\frac{1}{2}\alpha_3'+\frac{1}{2}\alpha_5a_3+\frac{3}{8}a_4\alpha_6\right)\frac{\partial}{\partial a_3}$$
$$+\left(\alpha_2'-\frac{1}{6}a_3\alpha_4'+\frac{3}{4}\alpha_5a_4-\frac{1}{6}a_3^2\alpha_6\right)\frac{\partial}{\partial a_4}$$
$$+\left(\alpha_7-\frac{1}{12}a_2\alpha_4'+\alpha_6\left(W-\frac{5}{24}a_2a_3\right)\right)\frac{\partial}{\partial a_5}, \quad (6.6)$$

where

$\alpha_5,\alpha_6\in\mathscr{O}(a)$,

$U=-\frac{1}{16}a_1^2a_2((8+a_1^2)/(4-a_1^2))-\frac{1}{16}a_1^2a_2-\frac{1}{4}a_2$,

$V=-\frac{1}{2}a_1^2((a_5+\frac{1}{8}a_1a_2^2-\frac{1}{2}a_1a_3)/(4-a_1^2))$
$\quad -\frac{1}{16}a_1a_2^2+\frac{1}{2}a_1a_3$,

$W=-\frac{1}{16}a_1^3((a_2a_5-2a_4)/(4-a_1^2))-\frac{1}{16}a_1a_2a_5$
$\quad +\frac{1}{8}a_1a_4+\frac{1}{8}a_3a_2$,

$\alpha_7+\alpha_8x+\alpha_9x^2=A(x,a)(a_1x^2+a_2x+a_5)$
$\quad \times\mathrm{mod}(\langle 4x^3+2a_3x+a_4\rangle\mathscr{O}_{(x,a)})$,

$\alpha_2'+\alpha_3'x+\alpha_4'x^2=B(x,a)(a_1x^2+a_2x+a_5)$
$\quad \times\mathrm{mod}(\langle 4x^3+2a_3x+a_4\rangle\mathscr{O}_{(x,a)})$,

$A,B\in\mathscr{O}_{(x,a)}$.

Proof: We easily calculate

$$I(F)=\langle a_1x^2+a_2x+a_5,\ 4x^3+2a_3x+a_4\rangle\mathscr{O}_{(x,a)},$$
$$\Theta(f)=\langle y^2,yx^3,x^6\rangle\mathscr{O}_{(y,x)},$$

so we can write

$$h=\alpha_1+\alpha_2x+\alpha_3x^2+\alpha_4x^3+\alpha_5x^4$$
$$+\alpha_6x^5+\alpha_7y+\alpha_8xy+\alpha_9x^2y,$$
$$\left.\frac{\partial h}{\partial y}\right|_{y=0}=\alpha_7+\alpha_8x+\alpha_9x^2\in I(F),$$
$$\left.\frac{\partial h}{\partial x}\right|_{y=0}=\alpha_2+2\alpha_3x+3\alpha_4x^2+4\alpha_5x^3+5\alpha_6x^4\in I(F).$$

Introducing the functions

$$\alpha_2'=\alpha_2-\alpha_5a_4,$$
$$\alpha_3'=2\alpha_3-2\alpha_5a_3-\tfrac{5}{4}\alpha_6a_4,$$
$$\alpha_4'=3\alpha_4-\tfrac{5}{2}a_3\alpha_6,$$

and using the Malgrange preparation theorem

$$\mathscr{O}_{(x,a)}/\langle 4x^3+2a_3x+a_4\rangle\mathscr{O}_{(x,a)}\cong[1,x,x^2]_{\mathscr{O}_{(a)}},$$

we obtain the respective equations for α_7, α_8, α_9 and

α_2', α_3', α_4'. □

Now, taking (6.5) into account, we can calculate the one-jets of the corresponding module generators:

$$j^1V_1=a_1\frac{\partial}{\partial a_1}+a_2\frac{\partial}{\partial a_2}+a_5\frac{\partial}{\partial a_5},$$
$$j^1V_2=a_2\frac{\partial}{\partial a_1}+a_5\frac{\partial}{\partial a_2},$$
$$j^1V_3=a_5\frac{\partial}{\partial a_1},$$
$$j^1V_4=\frac{1}{2}a_2\frac{\partial}{\partial a_3}+a_5\frac{\partial}{\partial a_4}, \quad (6.7)$$
$$j^1V_5=\frac{1}{2}a_5\frac{\partial}{\partial a_3},$$
$$j^1V_6=-2a_1\frac{\partial}{\partial a_1}-a_2\frac{\partial}{\partial a_2}+2a_3\frac{\partial}{\partial a_3}+3a_4\frac{\partial}{\partial a_4},$$
$$j^1V_7=-\frac{1}{4}a_2\frac{\partial}{\partial a_1}+\frac{3}{8}a_4\frac{\partial}{\partial a_3}.$$

We now treat the last case $D_{4,1}$.

Proposition 6.9: The module of the logarithmic vector fields Derlog $Q(D_{4,1})$ has seven generators. Their one-jets are

$$j^1V_1=\frac{\partial}{\partial a_1}+a_2\frac{\partial}{\partial a_3}-a_5\frac{\partial}{\partial a_4}+a_4\frac{\partial}{\partial a_5},$$
$$j^1V_2=j^1V_1,$$
$$j^1V_3=-2\frac{\partial}{\partial a_2}-2a_1\frac{\partial}{\partial a_3}+a_3\frac{\partial}{\partial a_4}-a_2\frac{\partial}{\partial a_5},$$
$$j^1V_4=-3a_1\frac{\partial}{\partial a_1}+2a_2\frac{\partial}{\partial a_2}-a_3\frac{\partial}{\partial a_3}+a_4\frac{\partial}{\partial a_4}+4a_5\frac{\partial}{\partial a_5},$$
$$j^1V_5=-\frac{1}{2}a_3\frac{\partial}{\partial a_1}+2a_5\frac{\partial}{\partial a_2}-\frac{1}{2}a_4\frac{\partial}{\partial a_3},$$
$$j^1V_6=a_2\frac{\partial}{\partial a_1}+a_4\frac{\partial}{\partial a_2},$$
$$j^1V_7=a_2\frac{\partial}{\partial a_2}+a_3\frac{\partial}{\partial a_3}+2a_4\frac{\partial}{\partial a_4}+2a_5\frac{\partial}{\partial a_5}.$$

Proof: As in the preceding cases we follow the standard procedure:

$$I(F)=\langle x_1+a_1x_2+a_3,x_1x_2+a_4,\tfrac{1}{2}x_1^2+x_2^2$$
$$+a_2x_2+a_5\rangle\mathscr{O}_{(x,a)},$$
$$\mathscr{O}_{(y,x)}/\Theta(f)\cong[1,x_1,x_2,y,x_1^2,x_1x_2,x_2^2,x_2y,x_1^3,$$
$$x_1^2x_2,x_1x_2^2,x_2^3,x_2^4,x_1x_2^3]_{\mathrm{C}}.$$

Thus by Proposition 6.5 we have

$$h=\alpha_0+\alpha_1x_1+\alpha_2x_2+\alpha_3y+\alpha_4x_1^2+\alpha_5x_1x_2+\alpha_6x_2^2$$
$$+\alpha_7x_2y+\alpha_8x_1^3+\alpha_9F(0,x,a)+\alpha_{10}x_1x_2^2$$
$$+\alpha_{11}x_2^3+\alpha_{12}x_2^4+\alpha_{13}x_1x_2^3$$

and

$$\left.\frac{\partial h}{\partial y}\right|_{y=0}=\alpha_3+\alpha_7x_2\in I(F),$$

$$\left.\frac{\partial h}{\partial x_1}\right|_{y=0}$$
$$\begin{aligned}&= \alpha_1 - 2\alpha_4 a_3 + 3(a_1 a_4 + a_3^2)\alpha_8 - \alpha_{10}(a_5 + \tfrac{1}{2}a_1 a_4\\ &\quad + \tfrac{1}{2}a_3^2) + \alpha_{13}\nu + x_2(-2a_1\alpha_4 + 3a_1 a_3 \alpha_8 + \alpha_5\\ &\quad - \alpha_{10}(a_2 + \tfrac{1}{2}a_1 a_3) + \alpha_{13}\mu) \in I(F),\end{aligned}$$

$$\left.\frac{\partial h}{\partial x_2}\right|_{y=0}$$
$$\begin{aligned}&= \alpha_2 - \alpha_5 a_3 - 2a_4\alpha_{10} - 3\alpha_{11}(a_5 + \tfrac{1}{2}a_1 a_4 + \tfrac{1}{2}a_3^2)\\ &\quad + 4\alpha_{12}\nu + x_2(2\alpha_6 - a_1\alpha_5 + 4\mu\alpha_{12}\\ &\quad - 3\alpha_{11}(a_2 + \tfrac{1}{2}a_1 a_3) - 3\alpha_{13}a_4) \in I(F).\end{aligned}$$

Because 1, x_2 form a free basis for $\mathscr{O}_{(x,a)}/I(F)$, we immediately obtain

$$\begin{aligned}&\alpha_3 = 0, \quad \alpha_7 = 0,\\ &\alpha_1 = 2\alpha_4 a_3 - 3(a_1 a_4 + a_3^2)\alpha_8\\ &\qquad + \alpha_{10}(a_5 + \tfrac{1}{2}a_1 a_4 + \tfrac{1}{2}a_3^2) - \alpha_{13}\nu,\\ &\alpha_5 = 2a_1\alpha_4 - 3a_1 a_3\alpha_8 + \alpha_{10}(a_2 + \tfrac{1}{2}a_1 a_3) - \alpha_{13}\mu,\\ &\alpha_2 = a_3(2a_1\alpha_4 - 3a_1 a_3 \alpha_8 + \alpha_{10}(a_2 + \tfrac{1}{2}a_1 a_3) - \alpha_{13}\mu)\\ &\qquad + 2a_4\alpha_{10} + 3\alpha_{11}(a_5 + \tfrac{1}{2}a_1 a_4 + \tfrac{1}{2}a_3^2) - 4\alpha_{12}\nu,\\ &\alpha_6 = \tfrac{1}{2}a_1(2a_1\alpha_4 - 3a_1 a_3\alpha_8 + \alpha_{10}(a_2 + \tfrac{1}{2}a_1 a_3) - \alpha_{13}\mu)\\ &\qquad - 2\mu\alpha_{12} - \tfrac{3}{2}\alpha_{11}(a_2 + \tfrac{1}{2}a_1 a_3) + \tfrac{3}{2}\alpha_{13}a_4,\end{aligned}$$

where

$$\begin{aligned}&\mu = (a_2 + \tfrac{1}{2}a_1 a_3)^2 - a_5 - \tfrac{1}{2}a_1 a_4 - \tfrac{1}{2}a_3^2,\\ &\nu = (a_2 + \tfrac{1}{2}a_1 a_3)(a_5 + \tfrac{1}{2}a_1 a_4 + \tfrac{1}{2}a_3^2).\end{aligned}$$

Inserting these into h we obtain the formula for

$$h(\text{mod } \overline{\Delta}(F)) - h(\text{mod } \overline{\Delta}(F))|_{x=0,\, y=0}.$$

From this formula we can read off not only the one-jets but also all the generators of the module. □

Let ρ: $\mathbb{C}^p \to \mathbb{C}^k$ be a projection on $\mathbb{C}^k \subset \mathbb{C}^p$. We say that $Q(F) \subset \mathbb{C}^p$ is locally equisingular along $\mathbb{C}^k$ near $p_0 \in \mathbb{C}^k$ if, for all $p \in \mathbb{C}^k$ near p_0, the pairs $(\rho^{-1}(p),0)$ and $(\rho^{-1}(p) \cap Q(F),0)$ are all diffeomorphic. Checking the vector fields listed in Propositions 6.7 and 6.9, we have the following corollary.

Corollary 6.10: (1) The quasicaustic $Q(F_{1,0})$ is equisingular along the two-dimensional singular locus, parametrized by $\{a_1,a_3\}$.

(2) The quasicaustic $Q(D_{4,1})$ is equisingular along the two-dimensional singular locus, parametrized by $\{a_1,a_2\}$.

In both cases the fiber $(\rho^{-1}(p) \cap Q(F),0)$ is diffeomorphic to Whitney's cross-cap.

The logarithmic vector fields can also be used for the classification of the generic Lagrangian pairs (L_1,L_2) up to quasicaustic equivalence (cf. Refs. 24 and 32). The singular Lagrangian variety $L_1 \cup L_2$ is provided by generic families of functions on the manifold with boundary. In this sense, to determine the germ of the Lagrangian pair means to define the generating family of functions on a manifold with a boundary (cf. Sec. III).

Let f: $(\mathbb{C}\times\mathbb{C}^n,0) \to (\mathbb{C},0)$ be a finitely determined boundary singularity. Let F: $(\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^{\mu-1},0) \to (\mathbb{C},0)$ be its miniversal unfolding. If G: $(\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^p,0) \to (\mathbb{C},0)$ is a generating family for a Lagrangian pair, then generically G is a pullback from the miniversal unfolding F of the finitely determined germ $f(y,x) = G(y,x,0)$, i.e.,

$$G(y,x,a) = F(\Phi(y,x,a),\phi(a)) + h(a),$$

where Φ: $(\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^p,0) \to (\mathbb{C}\times\mathbb{C}^n,0)$ is a family of biholomorphisms, germs preserving the hypersurface $\{y=0\}$. The pullback ϕ: $(\mathbb{C}^p,0) \to (\mathbb{C}^{\mu-1},0)$, $\phi \in \mathscr{O}^{\mu-1}_{(a)}$ and $h \in \mathscr{O}_{(a)}$. Thus analogously to the classification of generic Lagrangian submanifolds (see Ref. 20, p. 337), the classification of generic Lagrangian pairs is done by specifying the miniversal unfoldings of finitely determined boundary singularities and their generic pullbacks $\phi \in \mathscr{O}^{\mu-1}_{(a)}$.

Let us assume that Lagrangian pairs are modeled on unimodal singularities f: $(\mathbb{C}\times\mathbb{C}^n,0) \to (\mathbb{C},0)$, i.e., the generic generating family with such f has the following prenormal form:

$$G\colon\ (\mathbb{C}\times\mathbb{C}^n\times\mathbb{C}^p,0) \to (\mathbb{C},0), \quad p \geqslant \mu - 2,$$
$$G(y,x,a) = f(y,x) + \sum_{i=1}^{\mu-2} g_i(y,x)a_i + g_{\mu-1}(y,x)\lambda(a),$$

where $g_{\mu-1}(y,x)$ defines the modulus direction.

Generically, the pullback ϕ is transversal to this direction, so

$$\bar{\lambda} := \lambda|_{\{a_1 = \dots = a_{\mu-2} = 0\}}\colon\ (\mathbb{C}^{p-\mu+2},0) \to (\mathbb{C},0)$$

is a Morse function. Thus there are possible two generic normal forms for the generating families of Lagrangian pairs of unimodal type:

(1) $\lambda(a) = a_{\mu-1}$, when $p > \mu - 2$ and $D\bar{\lambda}(0) \neq 0$;

(2) $\lambda(a) = \eta(a_1,\dots,a_{\mu-2}) \pm a^2_{\mu-1} \pm \cdots \pm a^2_p$, when $D\bar{\lambda}(0) = 0$;

where $\eta \in \mathscr{O}_{(\bar{a})}$ $[\bar{a} = (a_1 \dots, a_{\mu-2})]$ is a functional modulus.

To obtain more information about classifying quasicaustics, we need to introduce a weaker equivalence relation in Lagrangian pairs (cf. Refs. 20 and 32 in the case of functional moduli in the standard classification of Lagrangian submanifolds). Let

$$\begin{aligned}&G_1(y,x,a) = F(y,x,\phi_1(a)) + f_1(a),\\ &G_2(y,x,a) = F(y,x,\phi_2(a)) + f_2(a)\end{aligned}$$

be two generating families for the corresponding Lagrangian pairs $\mathscr{L}_1$ and $\mathscr{L}_2$, respectively. We say that $\mathscr{L}_1$, $\mathscr{L}_2$ are quasicaustic equivalent if ϕ_1, ϕ_2 are right–left equivalent, i.e.,

$$\phi_1(a) = (\psi \circ \phi_2 \circ \xi)(a),$$

for some biholomorphism ξ: $(\mathbb{C}^p,0) \to (\mathbb{C}^p,0)$, and some biholomorphism ψ: $(\mathbb{C}^{\mu-1},0) \to (\mathbb{C}^{\mu-1},0)$ preserving the quasicaustic $(Q(F),0)$.

Proposition 6.11: For unimodal boundary singularities $F_{1,0}$, $D_{4,1}$, by quasicaustic equivalence, the functional modulus λ can be reduced to zero.

Proof: On the basis of Ref. 20, p. 343, we need to check only that

$$\mathscr{M}_{(a)} \subseteq \langle A^1_1(a),\dots,A^r_1(a)\rangle \mathscr{O}_{(a)}, \qquad (*)$$

which implies that

$$\mathscr{M}_{(a)}\mathscr{E}(\phi) \subseteq \phi^*\mathscr{E}(\mu - 1) + T\phi(\mathscr{E}(p)),$$

for ϕ: $(\mathbb{C}^p,0)\to(\mathbb{C}^{\mu-1},0)$ being in the general position to the modulus direction. Here by $\mathscr{E}(\phi)$ we denote the vector fields along ϕ (cf. Ref. 30). Let $\mathscr{E}(\mu-1)$ and $\mathscr{E}(p)$ be the spaces of vector fields on $(\mathbb{C}^{\mu-1},0)$ and $(\mathbb{C}^p,0)$, respectively. This enables us to apply the ordinary homotopic method to eliminate the functional modulus λ. Taking into account the vector fields listed in the Propositions 6.7 and 6.9,

$$V_l=\sum_{i=1}^{5}A_i^l\frac{\partial}{\partial a_i},$$

we immediately have fulfilled (*) for the parabolic singularities $F_{1,0}$ and $D_{4,1}$. □

ACKNOWLEDGMENTS

I am very grateful to Mark Roberts and Anthony Fletcher for useful discussions and their help in preparing this manuscript. I would also like to thank the referee for very helpful comments improving Sec. VI of the paper. I am also very grateful to Professor F. Hirzebruch and his Institute for their invitation and exceptional hospitality.

[1] J. B. Keller, "Rays, waves and asymptotics," Bull. Am. Math. Soc. **84**, 727 (1978).

[2] J. W. Bruce and S. Janeczko, "Classification of caustics by diffraction" (in preparation).

[3] J. W. Bruce, "Vector fields on discriminants and bifurcation varieties," Bull. London Math. Soc. **17**, 257 (1985).

[4] H. Terao, "The bifurcation set and logarithmic vector fields," Math. Ann. **263**, 313 (1983).

[5] R. K. Luneburg, *Mathematical Theory of Optics* (Univ. California Press, Berkeley, 1964).

[6] V. W. Guillemin and S. Sternberg, *Symplectic Techniques in Physics* (Cambridge U. P., London, 1984).

[7] M. Kline and I. W. Kay, *Electromagnetic Theory and Geometrical Optics* (Interscience, New York, 1965).

[8] S. Janeczko, "Generating families for images of Lagrangian submanifolds and open swallowtails," Math. Proc. Cambridge Philos. Soc. **100**, 91 (1986).

[9] J. W. Bruce, "Functions on discriminants," J. London Math. Soc. (2) **30**, 551 (1984).

[10] J. W. Bruce and D. L. Fidal, "Vector fields on caustics" (unpublished).

[11] T. Poston and I. Stewart, *Catastrophe Theory and its Applications* (Pitman, San Francisco, 1978).

[12] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin/Cummings, Reading, MA, 1978).

[13] S. Janeczko, "Covariant symplectic geometry of binary forms and singularities of systems of rays," Max-Planck-Institut für Mathematik, preprint, MPI, 1986–40.

[14] W. M. Tulczyjew, "The Legendre transformation," Ann. Inst. H. Poincaré **27**, 101 (1977).

[15] J. W. Bruce, P. J. Giblin, and C. G. Gibson, "On caustics by reflection," Topology **21**, 179 (1982).

[16] O. P. Shcherbak, "Wave fronts and groups generated by reflections," Usp. Mat. Nauk **43**, 125 (1988).

[17] A. Weinstein, "Lectures on symplectic manifolds," C. B. M. S. Conf. Ser. Am. Math. Soc. **29** (1977).

[18] S. Janeczko, "Singularities in the geometry of an obstacle," Suppl. Rend. Circolo Mat. Palermo, Ser. II **16**, 71 (1987).

[19] C. T. C. Wall, "Geometric properties of generic differentiable manifolds," in *Geometry and Topology, Lecture Notes in Mathematics*, Vol. **597**, edited by A. Dold and B. Eckmann (Springer, Berlin, 1977), pp. 707–774.

[20] V. I. Arnold, S. M. Gusein-Zade, and A. N. Varchenko, *Singularities of Differentiable Maps* (Birkhauser, Boston, 1985), Vol. 1.

[21] V. I. Arnold and A. B. Givental, "Symplectic geometry," Itogi Nauki Contemp. Probl. Math. Fund. Direct. **4**, 5 (1985).

[22] V. I. Arnold, "Singularities in variational calculus," Itogi Nauki Contemp. Probl. Math. **22**, 3 (1983).

[23] I. G. Shcherbak, "Focal set of a surface with boundary, and caustics of groups generated by reflections B_k, C_k, and F_4," Funct. Anal. Appl. **18**, 84 (1984).

[24] Nguyên hûu Dùc, Nguyên tiên Dai, "Stabilite de l'interaction geometrique entre deux composantes holonomes simples," C. R. Acad. Sci. Paris A **291**, 113 (1980).

[25] Yu. V. Chekanov, "Caustics in geometrical optics," Funct. Anal. Appl. **20**, 223 (1986).

[26] V. I. Matov, "Unimodal and bimodal germs of functions on manifold with boundary," Proc. Petrovski Sem. **7**, 174 (1981).

[27] D. Siersma, "Singularities of functions on boundaries, corners, etc.," Quart. J. Math. Oxford (2) **32**, 119 (1981).

[28] S. B. Alexander, I. D. Berg, and R. I. Bishop, "The Riemannian obstacle problem," Ill. J. Math. **31**, 167 (1987).

[29] D. Mond, "On the tangent developable of a space curve," Math. Proc. Camb. Philos. Soc. **91**, 351 (1982).

[30] J. Martinet, *Singularities of Smooth Functions and Maps* (Cambridge U. P., Cambridge, 1982).

[31] V. M. Zakalyukin, "Bifurcations of waveforms depending on one parameter," Funct. Anal. Appl. **10**, 139 (1976).

[32] S. Janeczko and M. Roberts, "Classification of symmetric caustics," preprint, University of Warwick, October (1983).

[33] K. Saito, "Theory of logarithmic differential forms and logarithmic vector fields," J. Fac. Sci. Univ. Tokyo Sec. I A **27**, 265 (1980).

[34] E. J. N. Looijenga, *Isolated Singular Points on Complete Intersections* (Cambridge U.P., London, 1984).

Radiative transfer on curved surfaces

J. Tessendorf
Climate System Research Program, College of Geosciences, Texas A&M University, College Station, Texas 77843

(Received 26 April 1989; accepted for publication 15 November 1989)

After a review of appropriate concepts in local surface geometry, a formally exact solution of the radiative transfer equation is constructed, for transfer from one surface of arbitrary shape to another. The solution is obtained from repeated application of the linear interaction principle to form a path integral over paths that cross many intermediate surfaces. Invariant imbedding in general geometries is presented and found to be manifest in the path integral solution as an invariance under local coordinate transformations of the intermediate surfaces. Aspects of possible numerical implementations of this geometrical approach are discussed.

I. INTRODUCTION

There are a number of problems of current interest in atmospheric remote sensing and ocean optics that have the common need for knowledge of the distribution of radiation propagated through a medium with curved or irregular boundaries. One such problem is the conversion of measured brightness temperatures of clouds into an estimate of the local rain rate,[1,2] in which the microwave emission by rainfall suffers absorption and multiple scattering in the volume of rain, ice particles, and cloud liquid water content. Plane-parallel models of rain fields with finite horizontal extent, for example, show that the brightness temperature–rain rate connection is significantly affected by the spatial extent of the rain field.[3]

There is a very active effort to calculate the radiance distribution emitted and reflected by clouds in the visible and IR regimes. Some Monte Carlo calculations have been used to study the effects of cloud geometry,[4] and a multimode technique exists for geometries that can be represented by a collection of cuboids.[5] The cloud geometry is more important in these regimes than in the microwave emission problem, because the cloud body itself is more attenuating in the visible and IR than at microwave wavelengths.

The resolution of underwater imaging systems is limited by blurring and contrast reduction, induced by scattering and absorption in the water. These effects can be suppressed somewhat by removal of the corresponding Mutual Transfer Function (MTF) from the image, or by range-gating the transceiver system.[6–8] Typically the formulation of the underwater imaging problem treats the imaged object as lying in a plane parallel to the camera plane. When the object has an extended structure within the field of view, however, it may be necessary to accommodate the range of scattering and absorption within the image by accounting for the object's three-dimensional shape.

These three examples illustrate radiative transfer problems with complicated spatial boundaries. In the first two examples the medium itself is bounded by irregular surfaces (the source of radiant power also has irregular bounds), while in the third the reflecting surface has some three-dimensional shape and the medium is effectively unbounded (ignoring for the moment any effects of the ocean surface in altering the light field). The geometric aspects of these examples are as important as the scattering and absorbing properties of the media. In general, any radiative transfer problem that involves an inhomogeneous medium and/or boundaries exhibits sensitivity to the geometry.

It seems worthwhile, therefore, to frame the solution of the radiative transfer equation in terms of the appropriate geometrical setting. This has been carried out in great detail and rigor for the special case of a medium composed of parallel planes, under the elegant formalism of reflection and transmission operators[9,10] in the context of the invariant imbedding relations. Preisendorfer also developed the full invariant imbedding relations for arbitrarily shaped media,[11] but, in that case, the emphasis was on developing relations with structure analogous to the flat surface case, and the geometric aspects were left implicit.

The purpose of this paper is to clarify the role of surface geometry in the solution of the radiative transfer equation and in the invariant imbedding relations. The approach taken is to construct an evolution operator for the general solution of the radiative transfer equation. In this way problems with constrained boundary conditions are restated as evolution problems with constrained initial conditions, and it is this latter form of the solution that yields most directly the invariant imbedding relations. A brief review is provided in Sec. III of invariant imbedding on flat surfaces, along with a generalization to curved surfaces.

The evolution operator approach has been used a number of times,[9,12,13] each distinguished by its own particular variations. In their basic form all of the variations assume the distribution is known on an initial plane, or assume complementary partial information on several planes, and use the evolution operator or transmission and reflection operators to obtain the distribution on the final plane(s) of interest. The more recent references construct the evolution operator in terms of a path integral[12] or a discretized matrix operator.[13] The two are related in the sense that the path integral solution is obtained in the limit of a very fine discretization for the matrix quantities.

However, as mentioned above, it is desirable to generalize the bounding planes to curved surfaces with potentially very complicated structure. The parametrization of the me-

0022-2488/90/041010-10$03.00

dium in terms of curved surfaces is accomplished in Sec. II. The linear interaction principle is used to begin the construction of the evolution operator in Sec. III. The result is the evolution operator for the transfer of radiance from the initial surface to the final surface, obtained from a sequence of transfers across many intermediate surfaces. When a large number of intermediate surfaces is used, the evolution operator for each transfer becomes an infinitesimal operator. The infinitesimal evolution operator is constructed in Sec. IV using the radiative transfer equation, and the full expression for the evolution operator in terms of a path integral is obtained. The path integral method for constructing the evolution operator has been used in the context of ocean optics,[12,14,15] and much of the notation and techniques used below can be found there.

One interesting consequence of incorporating surface geometry into the path integral formalism is that the principle of invariant imbedding is the natural consequence of the fact that the formal expression for the evolution operator is invariant under arbitrary local coordinate transformations of the intermediate surfaces. This invariance is demonstrated in Sec. V.

In Sec. VI, the explicit inclusion of the surface geometry is discussed as a possible method of improving the efficiency of numerical algorithms (such as finite-difference) that employ a spatial grid mesh.

The notation used below for the radiative transfer equation is

$$\{\hat{n}\cdot\nabla + c\}L(\mathbf{x},\hat{n}) = \int d\Omega'\,\beta(\hat{n}\cdot\hat{n}')L(\mathbf{x},\hat{n}'),$$

where L is the radiance, $\mathbf{x}$ is the position in the volume, $\hat{n}$ is the direction of propagation, c is the total extinction coefficient, and β is the volume scattering function. The dependence of the optical properties c and β on position in the volume is ignored, although all of the results can be extended to include a nonhomogeneous medium. For convenience, the volume scattering function is written as

$$\beta(\hat{n}\cdot\hat{n}') = b\,P(\hat{n}\cdot\hat{n}'),$$

where b is the scattering coefficient,

$$b = \int d\Omega\,\beta(\hat{n}\cdot\hat{n}'),$$

and the phase function P has unit normalization

$$\int d\Omega\,P(\hat{n}\cdot\hat{n}') = 1.$$

II. PARAMETRIZATION OF THE CURVED GEOMETRY

Suppose the volume of the medium is bounded by the two surfaces s_i and s_f, as in Fig. 1. Let $u \equiv (u^a) \equiv (u^1,u^2)$ be coordinates of a two-dimensional plane. A point $\mathbf{x}(u)$ on a surface is a mapping of a point u of the 2-D plane to the surface in the 3-D volume, indicated in Fig. 1 by the shape of the (u^1,u^2) mesh on the surface. All points in the volume of the medium can be parametrized by introducing the label s for each surface. The surface $s = s_i$ is the surface on which the distribution is known, and $s = s_f$ is the surface on which the distribution is to be obtained. The volume between these surfaces is the layered set of surfaces $s_i \leqslant s \leqslant s_f$. Each point in the volume is uniquely labeled by a triplet (u,s) with s labeling the particular surface, and u labeling the position on the surface. Points in the volume can be denoted $\mathbf{x}(u,s)$.

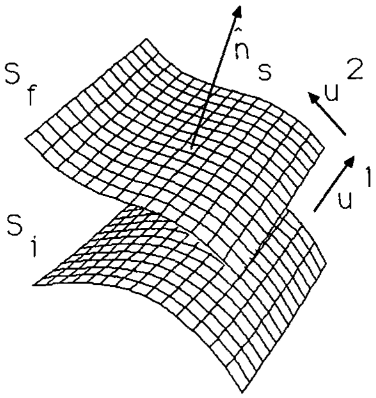


FIG. 1. Surface geometry showing (u^1,u^2) coordinates, surface normal, and layer of initial and final surfaces s_i, s_f.

Several concepts and quantities from differential geometry come into the formal calculations below (Ref. 16 is a good source). The primary quantity is the metric g. For a fixed surface s, the metric with components g_{ab} is defined as

$$g_{ab}(u,s) = \mathbf{x}_a(u,s)\cdot\mathbf{x}_b(u,s),$$

where $\mathbf{x}_a$ is

$$\mathbf{x}_a = \frac{\partial}{\partial u^a}\mathbf{x}.$$

The metric considered as a matrix has an inverse whose components are denoted g^{ab}, and which satisfies (implied summation over repeated indices is used throughout)

$$g_{ab}g^{bc} = g^{cb}g_{ba} = \delta^c_a,$$

where δ^c_a is the Kronecker delta function.

The two vectors $\mathbf{x}_a(u,s)$ define the local tangent plane to the surface, and are orthogonal to the surface normal (although they are not necessarily orthogonal to each other). The surface normal can be constructed from the cross product of the tangent vectors:

$$\hat{n}_s(u,s) = \frac{\mathbf{x}_1(u,s)\times\mathbf{x}_2(u,s)}{|\mathbf{x}_1(u,s)\times\mathbf{x}_2(u,s)|}.$$

A third vector defined at each point is $\dot{\mathbf{x}}(u,s)$, where, for convenience, derivatives with respect to s are denoted by

$$\dot{\mathbf{x}} \equiv \frac{\partial}{\partial s}\mathbf{x}.$$

Although this vector at each point on the surface is not necessarily orthogonal to either $\mathbf{x}_1$ or $\mathbf{x}_2$, it does not lie in the tangent plane, since it describes the layering of surfaces in the volume. Therefore the set of three vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dot{\mathbf{x}}\}$ could be used to construct a local basis of the three-dimensional space. However, as will be clear below, it is more convenient to use a basis in which $\dot{\mathbf{x}}$ is replaced by its component $\dot{\mathbf{x}}_\perp$ orthogonal to the tangent plane, given by

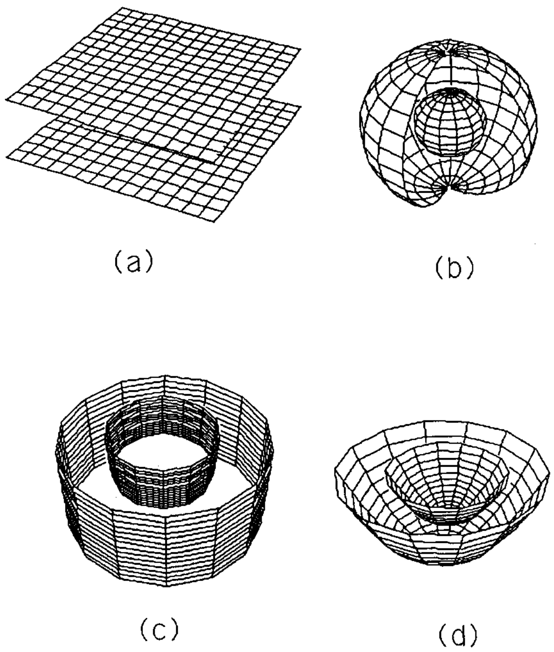


FIG. 2. Example volumes parametrized as layers of surfaces. (a) Layered flat planes; (b) imbedded concentric spheres; (c) imbedded concentric cylinders; (d) translated Monge patches.

$$\dot{\mathbf{x}}_\perp = \dot{\mathbf{x}} - f^a \mathbf{x}_a,$$

where

$$f^a = g^{ab}(\mathbf{x}_b \cdot \dot{\mathbf{x}}).$$

It can be verified directly from this definition that $\mathbf{x}_b \cdot \dot{\mathbf{x}}_\perp = 0$, and so $\dot{\mathbf{x}}_\perp$ is parallel to the surface normal, and we may write

$$\hat{n}_s(u,s) = \dot{\mathbf{x}}_\perp(u,s)/|\dot{\mathbf{x}}_\perp(u,s)|.$$

The local three-dimensional basis used below is the set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dot{\mathbf{x}}_\perp\}$.

As examples of this description of local geometry, we consider four examples illustrated in Fig. 2: layered flat planes, imbedded concentric spheres, imbedded concentric cylinders, and translated Monge patches. Each example is discussed below, and summarized in Table I.

Layered flat planes: Using the Cartesian coordinates (x,y,z), we take $u = (x,y)$ and $s = z$. Points $\mathbf{x}$ on each plane are given by $\mathbf{x} = (x,y,z)$, so that the basis $\{\mathbf{x}_1, \mathbf{x}_2, \dot{\mathbf{x}}_\perp\}$ is just the orthonormal set $\{(1,0,0), (0,1,0), (0,0,1)\}$.

Imbedded concentric spheres: Using the spherical coordinates (r,θ,ϕ), we assign $s = r$, and $u = (\theta,\phi)$. Positions on the surface of constant radius r are

$$\mathbf{x} = (r \sin\theta \cos\phi, r \sin\theta \sin\phi, r\cos\theta).$$

The derivative vectors are

$$\dot{\mathbf{x}} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta),$$

$$\mathbf{x}_1 = (r\cos\theta\cos\phi, r\cos\theta\sin\phi, -r\sin\theta),$$

$$\mathbf{x}_2 = (-r\sin\theta\sin\phi, r\sin\theta\cos\phi, 0).$$

The metric is

$$[g_{ab}] = r^2\begin{bmatrix} 1 & 0 \\ 0 & sin^2\,\theta \end{bmatrix}.$$

The perpendicular component of $\dot{\mathbf{x}}$ is

$$\dot{\mathbf{x}}_\perp = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta),$$

and this is also the surface normal $\hat{n}_s$.

Imbedded concentric cylinders: The cylindrical coordinates (ρ,φ,z) are assigned as $s = \rho$, $u = (\varphi,z)$. Positions on the surface of constant radius ρ are

$$\mathbf{x} = (\rho\cos\varphi, \rho\sin\varphi, z).$$

The derivative vectors are

$$\dot{\mathbf{x}} = (\cos\varphi, \sin\varphi, 0),$$

$$\mathbf{x}_1 = (-\rho\sin\varphi, \rho\cos\varphi, 0),$$

$$\mathbf{x}_2 = (0,0,1).$$

The metric is

$$[g_{ab}] = \begin{bmatrix} \rho^2 & 0 \\ 0 & 1 \end{bmatrix},$$

and the perpendicular component of $\dot{\mathbf{x}}$ is

$$\dot{\mathbf{x}}_\perp = (\cos\varphi, \sin\varphi, 0) = \hat{n}_s.$$

Translated Monge patches: A Monge patch is a surface of the form

$$\mathbf{x} = (x, y, h(x,y)),$$

where h is some well-behaved function and (x,y) are Cartesian coordinates. We can construct a volume by translating the Monge patch in the vertical direction, so that each Monte patch is given by $s = \text{const}$, and

$$\mathbf{x}(x,y,s) = (x, y, h(x,y) + s).$$

TABLE I. Summary of the example geometry coordinate systems.

	u^1	u^2	s	$\mathbf{x}$	$\dot{\mathbf{x}}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dot{\mathbf{x}}_\perp$
Plane	x	y	z	(x,y,z)	$(0,0,1)$	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$
Sphere	θ	ϕ	r	$(r\sin\theta\cos\phi,$ $r\sin\theta\sin\phi,$ $r\cos\theta)$	$(\sin\theta\cos\phi,$ $\sin\theta\sin\phi,$ $\cos\theta)$	$(r\cos\theta\cos\phi,$ $r\cos\theta\sin\phi,$ $-r\sin\theta)$	$(-r\sin\theta\sin\phi,$ $r\sin\theta\cos\phi,0)$	$(\sin\theta\cos\phi,$ $\sin\theta\sin\phi,$ $\cos\theta)$
Cylinder	φ	z	ρ	$(\rho\cos\varphi,$ $\rho\sin\varphi,z)$	$(\cos\varphi,$ $\sin\varphi,0)$	$(-\rho\sin\varphi,$ $\rho\cos\varphi,0)$	$(0,0,1)$	$(\cos\varphi,$ $\sin\varphi,0)$
Monge patch	z	y	s	$(x,y,h(x,y)+s)$	$(0,0,1)$	$(1,0,h_x)$	$(0,1,h_y)$	$(-h_x,-h_y,1)$

We take $u = (x,y)$, and the derivative vectors are

$$\dot{\mathbf{x}} = (0,0,1), \quad \mathbf{x}_1 = (1,0,h_x), \quad \mathbf{x}_2 = (0,1,h_y).$$

The metric is

$$[g_{ab}] = \begin{bmatrix} 1+h_x^2 & h_x h_y \\ h_x h_y & 1+h_y^2 \end{bmatrix},$$

and the perpendicular component of $\dot{\mathbf{x}}$ is

$$\dot{\mathbf{x}}_\perp = (-h_x, -h_y, 1).$$

The surface normal is

$$\hat{n}_s = (-h_x, -h_y, 1)/(1 + h_x^2 + h_y^2)^{1/2}.$$

These examples are used in the sections below to illustrate the geometric structure of the evolution operator.

III. THE EVOLUTION OPERATOR AND INVARIANT IMBEDDING

A functional definition for the evolution operator G can now be made. It is convenient to write G in the form $G(u,s_f,\hat{n};\, u',s_i,\hat{n}')$, and it is implicitly a function of the points on each of the two surfaces, and the intervening points. In this form the solution of the radiative transfer equation at points on the surface s_f is

$$L(\mathbf{x}(u,s_f),\hat{n}) = \int d^2u'\, d\Omega'\, G(u,s_f,\hat{n};\, u',s_i,\hat{n}') \times L(\mathbf{x}(u',s_i),\hat{n}'). \qquad (1)$$

For this solution the evolution operator must satisfy the radiative transfer equation with the initial condition

$$G(u,s_f,\hat{n};\, u',s_i,\hat{n}')|_{i\to f} = \delta(u-u')\delta(\hat{n}-\hat{n}').$$

The operator G is an evolution operator because, according to the linear interaction principle, it can be constructed from intermediate solutions. Suppose a particular intermediate surface s_1 between s_i and s_f is chosen. Using G, the radiance distribution at this intermediate surface is

$$L(\mathbf{x}(u,s_1),\hat{n}) = \int d^2u'\, d\Omega'\, G(u,s_1,\hat{n};\, u',s_i,\hat{n}') \times L(\mathbf{x}(u',s_i),\hat{n}').$$

Using the distribution at this intermediate surface, the distribution at the final surface s_f is

$$L(\mathbf{x}(u,s_f),\hat{n}) = \int d^2u'\, d\Omega'\, G(u,s_f,\hat{n};u',s_1,\hat{n}') \times L(\mathbf{x}(u',s_1),\hat{n}').$$

Combining these two results with the expression in Eq. (1), the operator G satisfies the convolution relationship

$$G(u,s_f,\hat{n};u_i,s_i,\hat{n}') = \int d^2u''\, d\Omega'' \times G(u,s_f,\hat{n};\, u'',s_1,\hat{n}'')\; G(u'',s_1,\hat{n}'';\, u',s_i,\hat{n}'). \qquad (2)$$

An alternate method of constructing a solution to the radiative transfer equation employs information about the radiance distribution on two parallel planes to obtain the distribution between them, using transmission and reflection operators. This method leads to the invariant imbedding principle, and is the source of the adding–doubling algorithm in some numerical methods of solution.[5] We review briefly this form of the invariant imbedding principle, and discuss how it is modified in a more general geometric setting.

Suppose surfaces s_1 and s_2 are flat parallel planes (see Fig. 3), with the normal of s_1 pointing ("upward") toward s_2. The portion of the radiance distribution on s_1 with components parallel to the normal (i.e., "upward") is denoted $L_U(s_1)$, and the portion of the distribution on s_2 with components antiparallel to the normal (i.e., "downward") is $L_D(s_2)$. The radiance distribution between s_1 and s_2 is given by

$$L(s) = F_U(s,s_1)L_U(s_1) + F_D(s,s_2)L_D(s_2), \qquad (3)$$

where (F_U,F_D) are the transmission and reflection operators, and we have suppressed the surface and angular convolutions. Explicitly,

$$F_{U,D}(s,s_i)L_{U,D}(s_i) \equiv \int d^2u' \int_{U,D} d\Omega' \times F_{U,D}(u,s,\hat{n};u',s_i,\hat{n}')L_{U,D}(u',s_i,\hat{n}'),$$

and the angular integrations are restricted to just the upward or downward direction, as appropriate. Equation (3) is the invariant imbedding equation. Its fundamental importance is that the radiance distribution on any plane s is determined by the distribution on the initial planes s_1 and s_2, but not by how the region between s_1 and s_2 is represented.

This solution can be iterated by choosing two planes s_3 and s_4, such as those shown in Fig. 3. From Eq. (3),

$$L(s_3) = F_U(s_3,s_1)L_U(s_1) + F_D(s_3,s_2)L_D(s_2),$$
$$L(s_4) = F_U(s_4,s_1)L_U(s_1) + F_D(s_4,s_2)L_D(s_2).$$

We can also write the invariant imbedding equation just in terms of the s_3 and s_4 surfaces:

$$L(s) = F_U(s,s_3)L_U(s_3) + F_D(s,s_4)L_D(s_4).$$

Denoting the upward component of F_U by F_{UU}, the downward component by F_{UD}, etc., we obtain

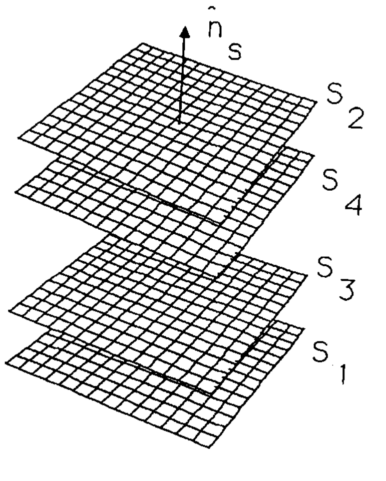


FIG. 3. Planar geometry used to describe the invariant imbedding principle.

$$F_U(s,s_1) = F_U(s,s_3)F_{UU}(s_3,s_1) + F_D(s,s_4)F_{UD}(s_4,s_1),$$
$$F_D(s,s_2) = F_U(s,s_3)F_{DU}(s_3,s_2) + F_D(s,s_4)F_{DD}(s_4,s_2). \tag{4}$$

Although this is an expression for $F_U(s,s_1)$ and $F_D(s,s_2)$ in terms of transmission and reflection operators at the intermediate planes s_3 and s_4, recall that s_3 and s_4 were introduced as convenient surfaces on which to iterate the invariant imbedding equation. The original solution is independent of any particular intermediate planes. Thus the invariant imbedding equation allows us to introduce additional conveniently located planes imbedded between s_1 and s_2, but guarantees that the solution obtained from such a decomposition is independent of the chosen imbedding. In Sec. V, this result is generalized to allow the imbedded surfaces to have arbitrary shape as well.

Despite the manipulations used above, the invariant imbedding principle can be stated in a simple, physically intuitive way: given a medium partitioned into regions, the operators in the full volume can be built up from the operators in the individual regions, and the result is independent of the choice of partition of the medium.

Preisendorfer considered the invariant imbedding problem in more general geometries. The basic change in the formalism arises from the fact that there is no longer a unique up and down orientation as in the planar geometry. Instead, up and down are defined locally according to the direction of the surface normal at each point. However, it should be possible to construct the invariant imbedding equation without choosing particular orientations. The necessity of the up and down directionality arises from the choice of the initial condition problem: the known distributions are $L_U(s_1)$ and $L_D(s_2)$, and so the problem is phrased in terms of these up and down directions.

We can, however, phrase a new problem: Suppose $\{s_1,...,s_N\}$ is a set of surfaces on which the radiance distribution $L(s_j)$ is known. We wish to find a solution $L(s)$ in the rest of the medium. From the linear interaction principle or the general form of the invariant imbedding equation, we might expect to write the solution as

$$L(s) = \sum_{j=1}^{N} F(s,s_j)L(s_j),$$

with

$$F(s_k,s_j) = \delta_{jk},$$

but we do not know yet what the F's are. Note that each of

$$L_j(s) = G(s,s_j)L(s_j),$$

are solutions of the radiative transfer equation, with $L_j(s_j) = L(s_j)$, but that the sum of these does not satisfy the conditions on each of the surfaces. However, a general solution using the evolution operator can be written

$$L(s) = \int ds'\, G(s,s')H(s').$$

The task is to find a function H that satisfies the initial condition on each surface. An ansatz for the solution is

$$H(s') = \sum_{j=1}^{N} \delta(s'-s_j)A_jL(s_j),$$

where the A_j are operators:

$$H(u,s,\hat{n}) = \sum_{j=1}^{N} \delta(s-s_j)\int d^2u'\, d\Omega' \times A_j(u,\hat{n};u',\hat{n}')L(u',s_j,\hat{n}').$$

The solution is found if the operators A_j satisfy the equation

$$\sum_j (G(s_k,s_j)A_j - \delta_{kj})L(s_j) = 0.$$

This is just the requirement that the functional determinant vanish:

$$\mathrm{Det}(G(s_k,s_j)A_j - \delta_{kj}) = 0.$$

Thus the A_j are related to the eigenvalues of $G(s_k,s_j)$. It is unclear under what conditions the A_j fail to exist. Presumably some choices of distributions and surfaces are incompatible, and a solution cannot exist. On the other extreme, we can reduce the geometry to the flat plane case discussed above, and only specify the appropriate up and down components, for which the solution is known to exist. The transition between these extremes is not understood, however.

Assuming for the moment the A_j exist, the solution is

$$L(s) = \sum_j G(s,s_j)A_jL(s_j),$$

which is the invariant imbedding equation with $F(s,s_j) = G(s,s_j)A_j$.

IV. CONSTRUCTION OF THE PATH INTEGRAL REPRESENTATION

The path integral representation constructed below follows from the linear interaction principle by iterating Eq. (2) over many intermediate surfaces. Suppose there are $N+1$ surfaces $s_j, j = 0,...,N$, with $s_0 = s_i$ and $s_N = s_f$; then successive iterations of Eq. (2) produce the result

$$G(u_f,s_f,\hat{n}_f;\, u_i,s_i,\hat{n}_i) = \int \prod_{j=1}^{N-1} d^2u_j\, d\Omega_j \times \prod_{j=1}^{N} G(u_j,s_j,\hat{n}_j;\, u_{j-1},s_{j-1},\hat{n}_{j-1}). \tag{5}$$

In the limit $N\rightarrow\infty$, this expression becomes the path integral representation.

To aid in understanding the path integral representation, we can think of G in terms of an effective attenuation coefficient τ_{eff} for an arbitrary path that starts at s_i and ends at s_f, summed over all such paths[12]:

$$G(s_f,s_i) \sim \sum_{\text{path}} \exp\{-\tau_{\text{eff}}(\text{path})\}.$$

Loosely speaking, when the number of intermediate surfaces N is large,

$$G(s_j,s_{j-1}) \sim \exp\{-\tau_{\text{eff}}(s_j,s_{j-1})\},$$

so that

$$\tau_{\text{eff}}(\text{path}) \sim \sum_j \tau_{\text{eff}}(s_j,s_{j-1}).$$

This statement is not rigorous, although it can be a useful way of picturing the physical content of the path integral. It is rigorous, however, to write

$$G(s_j,S_{j-1}) = \exp\{-\tau_c(s_j,s_{j-1})\}G_{\text{scatt}}(s_j,s_{j-1})$$

in the limit $N \to \infty$, in which τ_c accounts for total extinction, and G_{scatt} accounts for the distribution of scattering. Analogous to the construction of the evolution operator in quantum mechanics, a phase space is introduced below combining the set of all paths with the set of all directional modes of scattering at each point on a path, to yield a rigorous construction

$$G(s_f,s_i) = \sum_{\text{configurations}} \exp\{-\tau_{\text{eff}}(\text{configuration})\},$$

where $\Sigma_{\text{configurations}}$ means the sum over all phase space configurations of paths.

The first step in the construction is to parametrize an arbitrary path in the medium. This amounts to tracing a ray from the initial surface to the final surface, with its position on surface s denoted $\mathbf{l}(s)$. The local tangent of the path is a unit vector $\hat{\beta}$ pointing in the direction of propagation along the path, and is defined as[16]

$$\hat{\beta}(s) = \frac{d\mathbf{l}(s)/ds}{|d\mathbf{l}(s)/ds|}.$$

Alternatively, we can write this relationship in the differential form

$$d\mathbf{l}(s) = dl(s)\hat{\beta}(s).$$

The path $\mathbf{l}$ and its differential elements $d\mathbf{l}$ have been parametrized just in terms of the surface label s, without the use of the surface coordinates u. However, for the purpose of construction of the path integral representation, it is convenient to describe the path in terms of the local tangent vector $\hat{\beta}(s)$, and the points $\mathbf{l}(s_i)$ and $\mathbf{l}(s_f)$ on the initial and final surfaces. Thus we treat $\hat{\beta}(s)$ as the prescribed quantity of a path, and find an expression for the position along the path in terms of the surface coordinates u. Recalling that the triplet $\{\mathbf{x}_1, \mathbf{x}_2, \dot{\mathbf{x}}_\perp\}$ forms a local basis, the differential $d\mathbf{l}$ can be decomposed as (see Figs. 4 and 5)

$$\begin{aligned} d\mathbf{l} &= dl\hat{\beta} \\ &= \dot{\mathbf{x}}\, ds + \dot{\mathbf{x}}_a\, du^a \\ &= \dot{\mathbf{x}}_\perp\, ds + f^a\mathbf{x}_a\, ds + \mathbf{x}_a\, du^a. \end{aligned}$$

Because $\dot{\mathbf{x}}_\perp$ is normal to the local tangent plane, the distance traveled between s and $s+ds$ is

$$dl = [|\dot{\mathbf{x}}_\perp|/(\hat{\beta}\cdot\hat{n}_S)]ds. \tag{6}$$

This expression is obtained by taking the inner product of $d\mathbf{l}$ with $\dot{\mathbf{x}}_\perp$.

Taking the inner product with $\mathbf{x}_a$ and using the result for dl, du^a satisfies the constraint

$$du^a = \{(\mathbf{x}^a\cdot\hat{\beta})[|\dot{\mathbf{x}}_\perp|/(\hat{\beta}\cdot\hat{n}_S)] - f^a\}\, ds, \tag{7}$$

where $\mathbf{x}^a$ is related to $\mathbf{x}_a$ by

$$\mathbf{x}^a = g^{ab}\mathbf{x}_b,$$

and summation over the repeated index is implied. This constraint equation for du^a can be converted to the nonlinear differential equation

$$\frac{du^a(s)}{ds} = (\mathbf{x}^a(u,s)\cdot\hat{\beta}(s))\frac{|\dot{\mathbf{x}}_\perp(u,s)|}{\hat{\beta}(s)\cdot\hat{n}_s(u,s)} - f^a(u,s) \tag{8}$$

describing the path in terms of the surface coordinates of a ray having the direction of propagation $\hat{\beta}(s)$ at each surface s. Table II is a list of the path equations for the example geometries.

The effective attenuation factor τ_{eff} has a contribution due to the total extinction coefficient c, and one due to the distribution of scattering described by the phase function. For a ray from surface s to $s+ds$, the total extinction is

$$\tau_c(s+ds,s) = c\, dl.$$

This expression is the straightforward consequence of the exponential character of total extinction.

The redistribution due to scattering is more difficult to obtain. However, the procedure used by Tessendorf[12,14] is the same as is needed here. Simply stating the result,

$$G_{\text{scatt}}(s+ds,s)$$

$$\int_{-\infty}^{\infty} d^3p \exp\{i\, ds\, \mathbf{p}\cdot\dot{\hat{\beta}}(s) + b\, dl\, \Pi(\mathbf{p})\}, \tag{9}$$

where Π is the "pseudo-Fourier transform" of the phase function P:

$$P(\hat{n}\cdot\hat{n}') = \int \frac{d^3p}{(2\pi)^3}\Pi(\mathbf{p})\exp\{i\mathbf{p}\cdot(\hat{n}-\hat{n}')\}.$$

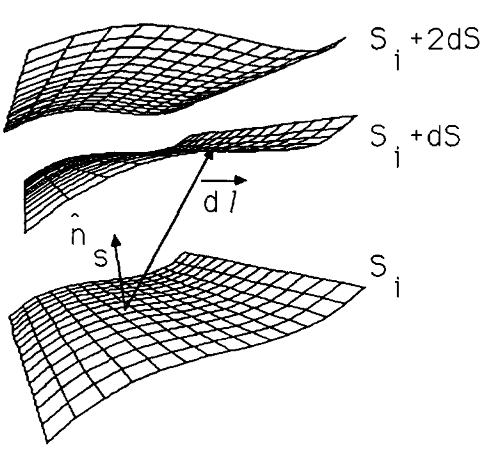


FIG. 4. Three-dimensional geometry showing the relationship between $\hat{n}_S$ and $d\mathbf{l}$.

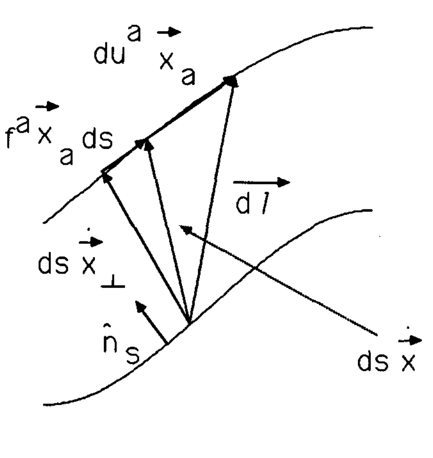


FIG. 5. The decomposition of $d\mathbf{l}$ in the $\{\mathbf{x}_1,\mathbf{x}_2,\dot{\mathbf{x}}_\perp\}$ basis.

TABLE II. Path equations for the example geometries. $\hat{\beta}(s)=(\cos\epsilon(s)\sin\lambda(s),\sin\epsilon(s)\sin\lambda(s),\cos\lambda(s))$.

Layered flat planes
$\frac{dx}{dz}=\cos\epsilon\tan\lambda$ $\frac{dy}{dz}=\sin\epsilon\tan\lambda$
Imbedded concentric spheres
$\frac{d\theta}{dr}=\left(\frac{1}{r}\right)\frac{\tan\lambda\cos(\epsilon-\phi)-\tan\theta}{1+\tan\lambda\tan\theta\cos(\epsilon-\phi)}$ $\frac{d\phi}{dr}=\left(\frac{1}{r}\right)\frac{\tan\lambda\sin(\epsilon-\phi)}{\cos\theta+\tan\lambda\sin\theta\cos(\epsilon-\phi)}$
Imbedded concentric cylinders
$\frac{d\varphi}{d\rho}=\frac{1}{\rho}\tan(\epsilon-\varphi)$ $\frac{dz}{d\rho}=\cot\lambda\sec(\epsilon-\varphi)$
Translated Monge patches
$\frac{dx}{ds}=\frac{(1+h_y^2)\cos\epsilon\tan\lambda-h_xh_y\sin\epsilon\tan\lambda+h_x}{1-h_x\cos\epsilon\tan\lambda-h_y\sin\epsilon\tan\lambda}-\frac{h_x}{1+h_x^2+h_y^2}$ $\frac{dy}{ds}=\frac{-h_xh_y\cos\epsilon\tan\lambda+(1+h_x^2)\sin\epsilon\tan\lambda+h_y}{1-h_x\cos\epsilon\tan\lambda-h_y\sin\epsilon\tan\lambda}-\frac{h_y}{1+h_x^2+h_y^2}$

The term "pseudo" refers to the fact that the representation of P in terms of Fourier amplitudes Π cannot be inverted in the usual sense of Fourier transforms to provide a unique expression for Π. We can, however, define Π as the inverse Fourier transform of a function $\overline{P}$:

$$\Pi(\mathbf{p})=\int d^3\sigma\,\overline{P}(\boldsymbol{\sigma})\exp\{-i\mathbf{p}\cdot\boldsymbol{\sigma}\},\tag{10}$$

where, for $\sigma\leqslant 2$,

$$\overline{P}(\boldsymbol{\sigma})=P(1-\sigma^2/2),$$

and, for $\sigma>2$, $\overline{P}$ converges to zero sufficiently fast to ensure the existence of Eq. (10). For example, if we choose $\overline{P}=0$ for $\sigma>2$, then

$$\Pi(\mathbf{p})=\frac{4\pi}{p}\int_0^2\sigma\,d\sigma\sin(p\sigma)P\left(1-\frac{\sigma^2}{2}\right).$$

A similar approach for handling the phase function is used in the small-angle approximation of the radiative transfer equation.[17] This "pseudo-Fourier" representation is not restricted to small-angle problems, however, and is sufficiently general to include backscatter.

Note that the scattering contribution G_{scatt} has the form of an integral over the "scattering modes" $\mathbf{p}$. This brings about the introduction of the phase space, consisting of points $(\hat{\beta},\mathbf{p})$. The configurations described earlier are combinations of paths $\hat{\beta}(s)$ and scattering modes $\mathbf{p}(s)$ at each surface.

In addition to the attenuation factors from extinction and scattering, there must be in G an additional factor to enforce the ray-path constraint in Eq. (8). This can be included by setting

$$\begin{aligned}G(s+ds,s)&\\&=\delta(\dot{u}^a(s)-(\mathbf{x}^a(u,s)\cdot\hat{\beta}(s))\\&\quad\times[|\dot{\mathbf{x}}_\perp(u,s)|/\hat{\beta}(s)\cdot\hat{n}_S(u,s)]+f^a(u,s))\\&\quad\times(1/ds)^2\exp\{-\tau_c(s+ds,s)\}G_{\text{scatt}}(s+ds,s).\end{aligned}\tag{11}$$

The full expression for G now follows from this expression placed in Eq. (5), in the limits $N\to\infty$ and $ds\to 0$ such that $N\,ds=s_f-s_i$. The notation for this solution is

$$\begin{aligned}G(u_f,s_f,\hat{n}_f;u_i,s_i,\hat{n}_i)&\\&=\int(D\beta)(Du)(Dp)\delta(\hat{\beta}(s_i)-\hat{n}_i)\delta(\hat{\beta}(s_f)-\hat{n}_f)\\&\quad\times\delta(u(s_i)-u_i)\delta(u(s_f)-u_f)\\&\quad\times\prod_s\delta\left(\dot{u}^a-(\mathbf{x}^a\cdot\hat{\beta})\frac{|\dot{\mathbf{x}}_\perp|}{\hat{\beta}\cdot\hat{n}_S}+f^a\right)\\&\quad\times\mathrm{Det}\left(\frac{\partial}{\partial s}\right)^2\exp\{-\tau_{\text{eff}}(\hat{\beta},\mathbf{p})\}.\end{aligned}\tag{12}$$

The integration measures $(D\beta)$, (Du), and (Dp) are

$$(D\beta)=\prod_s d\Omega(s),$$

$$(Du)=\prod_s d^2u(s),$$

$$(Dp)=\prod_s d^3p(s),$$

and the effective attenuation is

$$\begin{aligned}\tau_{\text{eff}}(\hat{\beta},\mathbf{p})&=\int_{s_i}^{s_f}ds\frac{c|\dot{\mathbf{x}}_\perp|}{\hat{\beta}(s)\cdot\hat{n}_S}-i\int_{s_i}^{s_f}ds\,\mathbf{p}(s)\cdot\dot{\hat{\beta}}(s)\\&\quad-\int_{s_i}^{s_f}ds\,b\Pi(\mathbf{p}(s))\frac{|\dot{\mathbf{x}}_\perp|}{\hat{\beta}(s)\cdot\hat{n}_S}.\end{aligned}$$

The constraint delta function introduced in Eq. (11) is a convenient method of obtaining the radiance distribution on each surface by following each path. The procedure for constraining integration variables in path integrals was introduced by Faddeev and Popov,[18] and it requires the inclusion of the factor

$$\mathrm{Det}\left(\frac{\partial}{\partial s}\right)^{-2}\mathrm{Det}\left\{\delta_{ab}\frac{\partial}{\partial s}-\frac{\delta}{\delta u^b}\left[(\mathbf{x}^a\cdot\hat{\beta})\frac{|\dot{\mathbf{x}}_\perp|}{\hat{\beta}\cdot\hat{n}_S}-f^a\right]\right\}.$$

Because the argument involves the first derivative in s, this term is equal to 1, and so is omitted. This was shown by Fried and Tessendorf in evaluation of a similar determinant in a fluid dynamics context.[19]

V. LOCAL COORDINATE TRANSFORMATIONS AND INVARIANT IMBEDDING

According to the invariant imbedding principle, the path integral expression for G should be independent of how the intermediate surfaces are parametrized. However, the expression in Eq. (12) for G clearly uses an explicit parametrization of the intermediate surfaces. In fact, the path integral representation is independent of the parametrization, in the sense that local coordinate transformations can be performed on the surfaces, and the expression for G is left invariant. This invariance is demonstrated below.

A local coordinate transformation $u \rightarrow \bar{u}(u)$ on a surface is characterized by a transformation matrix Λ with components

$$\Lambda^a_b = \frac{\partial u^a}{\partial \bar{u}^b}.$$

A physical position $\mathbf{x}$ on a surface is not altered by a coordinate transforation, although the local tangent plane is now characterized by the transformed vectors

$$\bar{\mathbf{x}}_a = \Lambda^b_a \mathbf{x}_b.$$

This transformation follows from the chain rule for a change of variables. Transformation of the tangent plane vectors also transforms the metric:

$$\bar{g}_{ab} = \Lambda^c_a g_{cd} \Lambda^d_b.$$

It also follows that

$$\bar{\mathbf{x}}^a = (\Lambda^{-1})^a_b \mathbf{x}^b.$$

The second type of local coordinate transformation is a rescaling transformation of the surface labeling: $s \rightarrow \bar{s}(s)$. Note that this transformation preserves the order of the surface, since $s = \text{const}$ implies $\bar{s} = \text{const}$ also, but this transformation allows the density of the surfaces to be changed. We assume, however, that $\bar{s}$ is a monotonic function of s, so that the order of the surfaces is preserved.

We wish to examine the behavior of Eq. (12) under the most general transformation $(u,s) \rightarrow (\bar{u}(u,s), \bar{s}(s))$, but leaving s_i and s_f fixed. This imposes the conditions

$$\bar{u}(u,s_i) = u_i, \quad \bar{u}(u,s_f) = u_f,$$
$$\bar{s}(s_i) = s_i, \quad \bar{s}(s_f) = s_f.$$

All of the terms in the exponential are invariant under the transformation. For example, the term

$$\int_{s_i}^{s_f} ds\, \mathbf{p}(s) \cdot \frac{\partial \hat{\beta}(s)}{\partial s}$$

becomes

$$\int_{s_i}^{s_f} d\bar{s} \frac{ds}{d\bar{s}} \mathbf{p} \cdot \frac{\partial \hat{\beta}}{\partial \bar{s}} \frac{d\bar{s}}{ds} = \int_{s_i}^{s_f} ds\, \mathbf{p} \cdot \hat{\beta},$$

and so is invariant. Similarly, the remaining two terms are invariant if

$$\frac{|\dot{\mathbf{x}}_\perp|}{\hat{\beta} \cdot \hat{\mathbf{n}}_S} = \frac{d\bar{s}}{ds} \frac{|\dot{\mathbf{x}}_\perp|}{\hat{\beta} \cdot \hat{n}_S}.$$

To show that this is the case, note that we can write

$$\dot{\mathbf{x}} = \frac{d\bar{s}}{ds} [\dot{\bar{\mathbf{x}}} + K^a \bar{\mathbf{x}}_a],$$

where

$$K^a = \frac{ds}{d\bar{s}} \frac{\partial \bar{u}^a}{\partial s}.$$

Using $\bar{\mathbf{x}}_a$ and $\dot{\bar{\mathbf{x}}}$, f^a can be written

$$f^a = \frac{d\bar{s}}{ds} \Lambda^a_b (\bar{f}^b + K^b),$$

with

$$\bar{f}^a = \bar{g}^{ab} (\dot{\bar{\mathbf{x}}} \cdot \bar{\mathbf{x}}_b).$$

Combining these expressions to construct $\dot{\mathbf{x}}_\perp$, the dependence on K^a cancels, and we have

$$\dot{\mathbf{x}}_\perp = \frac{d\bar{s}}{ds} \dot{\bar{\mathbf{x}}}_\perp.$$

Since $\dot{\mathbf{x}}_\perp$ is parallel to the surface normal in both coordinate systems, the normal is invariant, and we have the result that τ_{eff} is invariant under local coordinate transformations.

From the expression for the measure, (Du) transforms as

$$(Du) \rightarrow (D\bar{u}) \prod \det(\Lambda),$$

while the determinant becomes

$$\text{Det}\left(\frac{\partial}{\partial s}\right)^2 \rightarrow \prod \left| \frac{ds}{d\bar{s}} \right|^{-2} \text{Det}\left(\frac{\partial}{\partial \bar{s}}\right)^2.$$

The delta function argument transforms as [using the fact that $\dot{\bar{\mathbf{x}}}_\perp = (ds/d\bar{s}) \dot{\mathbf{x}}_\perp$]

$$\dot{u}^a - (\mathbf{x}^a \cdot \hat{\beta}) \frac{|\dot{\mathbf{x}}_\perp|}{\hat{\beta} \cdot \hat{n}_S} + f^a = \frac{d\bar{s}}{ds} \Lambda^b_a \left\{ \bar{u}^a - (\bar{\mathbf{x}}^a \cdot \hat{\beta}) \frac{|\bar{\mathbf{x}}_\perp|}{\hat{\beta} \cdot \bar{n}_S} + \bar{f}^a \right\},$$

so that the delta function constraint becomes

$$\prod \det(\Lambda^{-1}) \prod \left| \frac{ds}{d\bar{s}} \right|^2 \prod \delta\left(\bar{u}^a - (\bar{\mathbf{x}}^a \cdot \hat{\beta}) \frac{|\bar{\mathbf{x}}_\perp|}{\hat{\beta} \cdot \bar{n}_S} + \bar{f}_a \right).$$

The $\det(\Lambda)$ and $\Pi|ds/d\bar{s}|$ factors in the delta function, determinant, and measure transformations cancel each other, leaving Eq. (12) invariant under transformations of the imbedded surface parametrization.

Invariance under local coordinate transformations is a generalization of the invariant imbedding principle, in that the imbedded surfaces can be arbitrary shape without altering the evolution operator.

VI. NUMERICAL CONSIDERATIONS

The geometrical formalism and invariant imbedding results described above potentially can influence the design of numerical algorithms and codes for integrating the radiative transfer equation. The purpose of this section is to speculate on avenues of exploiting these geometrical results in numerical schemes. We exclude from the discussion algorithms that explicitly trace ray paths through the entire medium, such as Monte Carlo algorithms, because it is not necessary to include geometry in them as we have done here.

An important class of numerical schemes are finite-difference methods such as those presented in Refs. 20 and 21. Such methods in fully three-dimensional problems are generally best suited for rectilinear geometries because the finite differences are along Cartesian coordinate axes. More complicated boundaries are handled by using a rectilinear grid of spatial points with sufficient resolution to include the desired features. However, if these schemes could be written in terms of finite differences in the (u,s) variables, i.e., by the replacement

$$\nabla \rightarrow \frac{\hat{n}_s}{|\dot{\mathbf{x}}_\perp|} \frac{\partial}{\partial s} + \left(\mathbf{x}^a - f^a \frac{\hat{n}_s}{|\dot{\mathbf{x}}_\perp|} \right) \frac{\partial}{\partial u^a},$$

two possible advantages may be realized. The first is that the boundary conditions are easier to specify on the spatial grid, since the boundaries correspond to $s = \text{const}$. The second could occur when the number of spatial points in a calculation exceeds the capacity of the computer core memory. In this situation a large fraction of the total execution time can be spent swapping portions of the spatial grid in and out of the core (the I/O operation is the slowest operation in many computers). However, in a geometrical formulation, the spatial grid could be replaced by a single set of points $\{u_i\}$ on the u plane, and points in space generated by a mapping formula for each surface. The time spent in I/O operations would be replaced by time for calculation of the geometric quantities, such as the metric, on the surfaces. The second potential advantage should be realizable when the total computational time for repeatedly executing a mapping formula is less than the total I/O transfer time for swapping grid points in and out of the core. The balance between these two approaches would depend on the machine, as well as on the particular geometry under consideration. This time-savings argument may be valid for other numerical schemes, also.

One possible numerical algorithm which is different from the typical finite-difference algorithms, yet has a similar structure, is based on the interaction principle in Eq. (2). Changing notation somewhat, Eq. (1) can be written

$$L(u,s,\hat{n}) = \int d^2u'\, d\Omega'\, G(u,s,\hat{n};u',s_i,\hat{n}') \times L(u',s_i,\hat{n}').$$

Using the interaction principle in its iterated form, we obtain the finite-difference equation for the distribution at s_j in terms of the distribution at s_{j-1}:

$$L(u,s_j,\hat{n}) = \int d^2u'\, d\Omega'\, G(u,s_j,\hat{n};u',s_{j-1},\hat{n}') \times L(u',s_{j-1},\hat{n}'). \tag{13}$$

This solution has the form of a finite difference. A numerical algorithm would follow if a suitable discretization scheme can be found. In fact, Eq. (13) is analogous to the starting point of a finite-difference algorithm constructed for time-dependent radiative transfer,[14] and the discretization steps used in that case can be applied to this problem as well. Those steps, as applied to this current problem, are summarized below.

The first step is to discretize the angular degrees of freedom by introducing a set of directions $\{\hat{n}_k\}$, $k = 1,\ldots,N$, which point in the directions of the centroids of a set of solid angles $\{\Delta\Omega_k\}$. Defining the averaged radiance

$$L_k(u,s) = \int_{\Delta\Omega_k} d\Omega\, L(u,s,\hat{n}) \Delta\Omega_k,$$

the angularly discretized finite difference equation is

$$L_k(u,s_j) = \sum_{k'} \int d^2u'\, G_{kk'}(u,s_j;u',s_{j-1}) \times L_{k'}(u',s_{j-1}),$$

where

$$G_{kk'}(u,s;u',s') = \frac{1}{\Delta\Omega_k} \int_k d\Omega \int_{k'} d\Omega' \times G(u,s,\hat{n};u',s',\hat{n}')$$

is the discretized version of the evolution operator.

The next step is to construct $G_{kk'}$ in terms of the discretized phase function. This will require the interpretation of the difference $s_j - s_{j-1} = ds$ as a small quantity. The precise criterion for smallness follows from examining the magnitude of the higher-order terms excluded in the approximation. In analogy with the time-dependent case, the criterion is essentially $b\, ds|\dot{\mathbf{x}}_\perp| < 1$, i.e., that the propagation distance between adjacent surfaces is less than a single scattering length. Assuming ds is small, we can use Eqs. (9) and (11) to write an approximate discretization (see Ref. 12 for the full derivation)

$$\begin{aligned} G_{kk'}&(u,s+ds;u',s) \\ &= \delta(u^a - u'^a - du^a_{k'}(s,u)) \exp\{-c\, dl(u,s,\hat{n}_{k'})\} \\ &\quad \times (\exp\{b\, |\dot{\mathbf{x}}_\perp|(u,s)| ds\, \mathbf{Q}\})_{kk'}, \end{aligned}$$

where

$$du^a_k(s,u) = ds \left\{ (\mathbf{x}^a(u,s) \cdot \hat{n}_k) \frac{|\dot{\mathbf{x}}_\perp(u,s)|}{\hat{n}_k \cdot \hat{n}_S(u,s)} - f^a(u,s) \right\},$$

$\mathbf{Q}$ is the matrix with elements

$$Q_{kk'} = P_{kk'}/\hat{n}_{k'} \cdot \hat{n}_S(u,s),$$

and $\mathbf{P}$ is

$$P_{kk'} = \frac{1}{\Delta\Omega_k} \int_k d\Omega \int_{k'} d\Omega'\, P(\hat{n}\cdot\hat{n}').$$

This method of discretizing the phase function has been used in both time-independent[20,22] and time-dependent[14] radiative transfer.

Assembling these steps, the numerical algorithm is the explicit finite-difference formulation

$$L_k(u,s+ds) = \sum_{k'} \exp\left\{ -c|\dot{\mathbf{x}}_\perp| \frac{ds}{\hat{n}_{k'} \cdot \hat{n}_S} \right\} \times \exp\{b\, |\dot{\mathbf{x}}_\perp| ds\, \mathbf{Q}\}_{kk'} L_{k'}(u - du_{k'},s).$$

In this form the computationally intensive elements of the algorithm are the exponentiation of the matrix and the spatial interpolations needed to estimate the distribution at the points $u - du_k$.

An alternative discretization is to expand in spherical harmonics, so that the elements of the matrix $\mathbf{Q}$ are obtained from the spherical harmonics expansion for P and $\hat{n}$. The numerical algorithm would follow by truncating the expansion to some finite number of harmonics, and exponentiating $\mathbf{Q}$ as before. The utility of each of these two methods of discretization should depend on the structure of the phase function and on the angular resolution necessary for a specific calculation, although this issue has not been examined in detail.

The primary test of the utility of any algorithm based on geometric methods, however, will be the actual construction and execution of a code to determine directly its computational resource usage.

VII. CONCLUSIONS

The path integral solution of the radiative transfer equation has been constructed for problems involving curved or irregular boundaries in a medium. Several concepts and quantities from differential geometry have been used to make the solution compact. The path integration is over paths through the surfaces intermediate (imbedded) between the initial and final surfaces. The principle of invariant imbedding is satisfied by this solution, in the form of explicit invariance of the path integral to local coordinate transformations of the imbedded surfaces. It is hoped that this form of invariant imbedding can be exploited efficiently in a numerical algorithm and code. Existing and new numerical algorithms could incorporate this geometrical formulation to exploit its convenient parametrization of boundaries, and possibly to save execution time in calculations involving large numbers of spatial points.

ACKNOWLEDGMENT

This work was partly supported by NASA Contract No. NAG5-869.

[1] J. Simpson, R. F. Adler, and G. R. North, "A proposed tropical rainfall measuring mission (TRMM) satellite," Am. Meterol. Soc. **69**, 278 (1988).

[2] J. A. Weinman and P. J. Guetter, "Determination of rainfall distributions from microwave radiation measured by the Nimbus 6 ESMR," J. Appl. Meterol. **16**, 437 (1977); R. W. Spencer, B. B. Hinton, and W. S. Olson, "Nimbus-7 37 GHz radiances correlated with radar rain rates over the Gulf of Mexico," J. Cli. Appl. Meterol. **22**, 2095 (1983); R. W. Spencer, "A satellite passive 37-GHz scattering-based method for measuring oceanic rain rates," J. Cli. Appl. Meterol. **25**, 754 (1986).

[3] J. A. Weinman and R. Davies, "Thermal microwave radiances from horizontally finite clouds of hydrometeors," J. Geophys. Res. **83** (C6), 3099 (1978); C. Kummerow and J. A. Weinman, "Determining microwave brightness temperatures from precipitating horizontally finite and vertically structured clouds," *ibid.* **93** (D4), 3720 (1988).

[4] T. B. McKee and S. K. Cox, "Scattering of visible radiation by finite clouds," J. Atmos. Sci. **31**, 1885 (1974); M. Aida, "Scattering of solar radiation as a function of cloud dimensions and orientation," J. Quant. Spectrosc. Radiat. Transfer **17**, 303 (1980).

[5] R. W. Preisendorfer and G. L. Stephens, "Multimode radiative transfer in finite optical media. I: Fundamentals," J. Atmos. Sci. **41**, 709 (1984); G. L. Stephens and R. W. Preisendorfer, "Multimode radiative transfer in finite optical media. II: Solutions," J. Atmos. Sci. **41**, 725 (1984); G. L. Stephens, "Radiative transfer through arbitrarily shaped optical media. Part I: A general method of solution," *ibid.* **45**, 1818 (1988); G. L. Stephens, "Radiative transfer through arbitrarily shaped optical media. Part II: Group theory and simple closures," *ibid.* **45**, 1837 (1988).

[6] N. Witherspoon, M. Strand, J. Holloway, B. Price, D. Brown, R. Mill, and L. Estrup, "Experimentally measured MTFs associated with imaging through turbid water," SPIE J. **925**, (Ocean Optics IX), 363 (1988).

[7] W. H. Wells, "Loss of resolution in water as a result of multiple small-angle scattering," J. Opt. Soc. Am. **59**, 686 (1969).

[8] J. S. Jaffe and C. Dunn, "A model-based comparison of underwater imaging systems," SPIE J. **925** (Ocean Optics IX), 344 (1988).

[9] R. W. Preisendorfer, *Radiative Transfer on Discrete Spaces* (Pergamon, New York, 1965).

[10] G. W. Kattawar, J. Quant. Spectrosc. Radiat. Transfer **13**, 145 (1973).

[11] R. W. Preisendorfer, *Hydrologic Optics* (Pacific Marine Environmental Laboratory, ERL/NOAA, Honolulu, 1976), Vol. 11.

[12] J. Tessendorf, "Radiative transfer as a sum over paths," Phys. Rev. A **35**, 872 (1987).

[13] P. J. Flatau and G. L. Stephens, "On the fundamental solution of the radiative transfer equation," J. Geophys. Res. **93** (D9), 11037 (1988); P. C. Waterman, "Matrix-exponential description of radiative transfer," J. Opt. Soc. Am. **71**, 410 (1981).

[14] J. Tessendorf, "Time-dependent radiative transfer and pulse evolution," J. Opt. Soc. Am. A **6**, 280 (1989); J. Tessendorf, C. Piotrowski, and R. L. Kelly, "Finite-difference evolution of a scattered laser pulse in ocean water," SPIE J. **925** (Ocean Optics IX), 22 (1988).

[15] J. Tessendorf, "Comparison between data and small-angle approximations for the in-water solar radiance distribution," J. Opt. Soc. Am. A **5**, 1410 (1988).

[16] R. S. Millman and G. D. Parker, *Elements of Differential Geometry* (Prentice–Hall, Englewood Cliffs, NJ, 1977).

[17] L. S. Dolin, Izv. Atmos. Oceanic Phys. **16**, 34 (1980).

[18] L. D. Faddeev and V. N. Popov, "Feynman diagrams for the Yang–Mills field," Phys. Lett. B **25**, 29 (1967).

[19] H. M. Fried and J. Tessendorf, "Green's functions at zero viscosity," J. Math. Phys. **25**, 1144 (1984).

[20] W. S. Helliwell, "Finite-difference solution to the radiative-transfer equation for in-water radiance," J. Opt. Soc. Am. A **2**, 1325 (1985).

[21] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems* (Interscience, New York, 1967).

[22] C. D. Mobley and R. W. Preisendorfer, "A numerical model for the computation of radiance distributions in natural waters with wind-roughened surfaces," NOAA Technical Memorandum ERL PMEL-75, 1988.

On the velocities of the barotropic perfect fluids

Bartolomé Coll
Département de Mécanique Relativiste, UA 766 CNRS, Université de Paris VI, Paris, France
Joan Josep Ferrando
Departament de Física Teòrica, Universitat de València, Burjassot, València, Spain

(Received 16 August 1989; accepted for publication 8 November 1989)

The conditions for a unit vector field to be the velocity of a relativistic barotropic perfect fluid are given. These conditions induce an eightfold classification of such fluids; for every class, the admissible barotropic variables are found. Some special cases, in particular polytropic fluids, are analyzed separately.

I. INTRODUCTION

In relativity, a perfect fluid is characterized by an energy tensor T of the form $T = (\rho + p) u \otimes u - pg$, where ρ is the total energy density, p is the pressure, and u is the (unit) velocity of the fluid, and g is the space-time metric. The conservation of T leads to a system of equations in (u,ρ,p), open from the evolutive point of view, which is usually closed by the adjunction of a barotropic relation $\rho = \rho(p)$. So completed, this system is called the *fundamental system* of barotropic hydrodynamics.

Thus in a given domain of the space-time, a barotropic perfect fluid is a solution $s \equiv (u,\rho(p),p)$ to the fundamental system. Let us denote by $\mathbf{U}$ the set of unit vector fields u, by $\mathbf{R}$ the set of functions of a single variable $\rho = \rho(p)$, and by $\mathbf{F}$ the set of functions p over the given domain of the space-time. In the total space $\mathbf{U}\times\mathbf{R}\times\mathbf{F}$, the space of solutions $\{s\}$ to the fundamental system defines, by circumscription, a parallelepiped $\mathbf{U}_b\times\mathbf{R}_b\times\mathbf{F}_b$.

The Cauchy problem for the fundamental system shows that $\mathbf{R}_b = \mathbf{R}$ or, in other words, that locally, *any* function of a single variable $\rho(p)$ is an element of a solution $(u,\rho(p),p)$ to the fundamental system.[1] Nevertheless, it can be shown that $\mathbf{U}_b$ is a *proper* subset of $\mathbf{U}$, $\mathbf{U}_b \neq \mathbf{U}$, that is, there does not exist, in general, a barotropic perfect fluid having as the velocity field an arbitrary unit vector field of $\mathbf{U}$. Thus it is natural to ask the following question: Is it possible to intrinsically define $\mathbf{U}_b$ or, more precisely, is it possible to express, solely in terms of u and its derivatives, the necessary and sufficient conditions for u to be the velocity field of a barotropic perfect fluid?

The answer, as we shall show, is affirmative. The search for the conditions on u leads to a classification of the unit vector fields in eight classes. For each class, we obtain the necessary and sufficient conditions on u and its differential concomitants for insuring that u is the velocity field of some barotropic perfect fluid. Furthermore, we give the holonomy potentials that allow us to determine the corresponding barotropic relations.

Similar problems to that of the intrinsic characterization of $\mathbf{U}_b$, but restricted to particular forms of the barotropic relation or to particular evolution laws, may be also considered. As an illustration, here we obtain the intrinsic characterization of the unit vector fields u that are the velocity fields of (i) perfect fluids with constant pressure, (ii) perfect fluids with constant total energy density, (iii) barotropic perfect fluids with u-invariant (i.e., constant along the streamlines) pressure, and (iv) polytropic fluids. For these cases, the conditions on u are simpler than those corresponding to the generic barotropic case.

From a formal point of view, the differential system in u defining the set $\mathbf{U}_b$ is nothing but the *conditional system* in the variable u associated to the fundamental system of barotropic hydrodynamics. In other very different contexts, such as thermodynamic perfect fluids,[2] electromagnetic fields,[3] and almost-product structures or Killing tensors,[4] we have already shown the conceptual interest of conditional systems.

Now, what is the interest of an intrinsic characterization of the barotropic velocities in hydrodynamics? We think that such a characterization may be of interest in many domains, as, for example, in the following.

(i) Our conditional systems allow one to divide the task of integration of the fundamental (test) system into two clearly defined steps: a first step in which, after selecting the desired class of velocities from our eightfold classification of the unit vector fields, one looks for a solution u to the corresponding conditional system and, once it is obtained, a second step in which, with the aid of our results on the holonomy potentials, one constructs the barotropic relations $\rho = \rho(p)$ associated to this u.

(ii) In the usual approach to the integration of the Einstein equations for barotropic perfect fluid space-times, one considers directly the Einstein system and its first integrability conditions; the problems of compatibility that appear because of the relation $\rho(p)$ are well known. Our characterization of $\mathbf{U}_b$ guarantees the existence of such a relation and allows one to relegate to a last, third step its computation: In a first step, taking local charts adapted to u, one translates the chosen conditional system in u into a system in the components of the space-time metric g; in a second step, for the corresending constrained form of g, one evaluates its Ricci tensor and imposes that u be an eigenvector; and finally, in a third step, one considers the remaining Einstein equations with respect to the barotropic relation(s) computed from the g obtained in the first step.

(iii) One of the few known results on the restrictions that the Einstein equations impose to the space of solutions of the fundamental (test) system is the Treciokas–Ellis[5] conjecture, recently reconsidered by Collins.[6] The conjecture

0022-2488/90/041020-06$03.00

states that a distortion-free barotropic perfect fluid space-time is either vorticity-free or expansion-free.[7] Because of its purely kinematical character, our associated conditional systems in u are well adapted to the study of this conjecture.

(iv) In given (vacuum, Robertson–Walker, etc.) space-times, it is sometimes interesting to know if some particular congruences may be interpreted as the streamlines of barotropic test perfect fluids (e.g., weak accretion in the neighbors of a star). The answer to this follows directly from our results by a simple, direct computation.

(v) Whatever its barotropic equation $\rho = \rho(p)$, a (test) barotropic perfect fluid may always evolve following any (static or stationary) Killing direction of any space-time. Nevertheless, the analog statement for conformally Killing directions is false: In fact, the only barotropic perfect fluid that may evolve following any conformally Killing direction of any space-time is that of isotropic radiation $\rho = 3p$ in equilibrium with dust of constant energy density. Properties such as these may be easily obtained from our characterization of the barotropic velocities.

(vi) Every barotropic velocity may be endowed with a barotropic relation $\rho(p)$ and, of course, also with other more general thermodynamic relations. We think that in the study of nonbarotropic perfect fluids or nonperfect fluids (anisotropy, viscosity, heat conduction), the hypothesis that their velocities are barotropic may be useful in the study of the behavior of such fluids. Either this hypothesis is incompatible (the actual motion of the fluid cannot be reproduced by any barotropic test fluid) or it is acceptable (one can compare the ideal barotropic variables to the actual thermodynamic ones). Both results constitute an interesting complement of information; in particular, the latter result may help us to better understand the limitations involved in the Eckart and Landau thermodynamic schemes.

(vii) For the taxonomy of the solutions of the fundamental (test) system and the Einstein equations, the eight classes of velocity vector fields not only allow one to label the known solutions, but also to play an heuristic role in the search of new solutions.

The paper is organized as follows. In order to make the proofs of the main result easier, in Sec. II the case $p =$ const is separate from the generic one, for which the data are reduced to a unit vector field and a holonomy potential. Section III contains the main results of this paper: the eightfold classification of the unit vector fields (Definition 1), the characterization of the barotropic velocities corresponding to each of these classes (Theorem 1), and the associated equations for the holonomy potentials (Theorem 2). Finally, in Sec. IV we characterize the velocities corresponding to some particular cases often found in the literature: constant pressure or density, u-invariant pressure, and polytropic fluids.

A portion of the present results (those leading to Theorem 1) with a sketch of the proof has been published elsewhere.[8]

II. THE BAROTROPIC PERFECT FLUID

Let (V_4,g) be the space-time sig$(g) = -2$. Vector and tensor fields and the expressions that relate them, unless otherwise stated, are given in their covariant form. The symbols $i(u)$, $*$, d, ∇, and δ denote, respectively, the interior product, Hodge dual, exterior derivative, covariant derivative, and divergence operators.

In a domain of (V_4,g), the conservation $\delta T = 0$ of the energy tensor T of a perfect fluid amounts to the system

$$dp = (\rho + p)a + p^0 u, \quad \theta + \rho^0/(\rho + p) = 0, \qquad (1)$$

where $a = i(u)\nabla u$ is the *acceleration vector*, $\theta \equiv -\delta u$ is the *expansion*, and $f^0 \equiv \pounds(u)f$ for any function f.

A *barotropic* relation is a functional relation between ρ and p of the form

$$d\rho \wedge dp = 0. \qquad (2)$$

When such a relation takes place (1) is called the *fundamental system* of barotropic hydrodynamics.

In the particular case of constant pressure $p = \bar{p} =$ const, system (1) becomes

$$a = 0, \quad \theta + \rho^0/(\rho + \bar{p}) = 0. \qquad (3)$$

Given u (and consequently, θ), the second of Eqs. (3) associates one solution ρ to every $\bar{p}$ and to every u-invariant function $f\,(f^0 = 0)$. Although simple, we explicitly state this result for completeness in the following proposition.

Proposition 1: Perfect fluids with constant pressure have geodesic velocities. Conversely, to every geodesic (unit) vector field u one can associate a family of perfect fluids with arbitrary constant pressure $\bar{p}$ and energy density $\rho = f\rho_0 + (f-1)\bar{p}$, where ρ_0 is a given solution to $\theta + \rho^0/(\rho + \bar{p}) = 0$ and f is any u-invariant function.

From here on, unless otherwise stated, we have consider $dp \neq 0$. Perfect fluids with a barotropic relation such that $dp \neq 0$ with be called *barotropic fluids*. Because of (2), there exists a (local) function π verifying

$$d\rho = (\rho + p)d\pi. \qquad (4)$$

This function is called the *holonomy potential.*[9] For a nonconstant p one has

$$p = p(\pi), \quad p'(\pi) = \rho + p \neq 0, \qquad (5)$$

where $p' \equiv dp/d\pi$. From (4) and (5), the first of Eqs. (1) may be written as

$$d\pi = a + \pi^0 u \neq 0, \qquad (6)$$

the scalar $\rho^0/(\rho + p)$ adopts the form

$$\rho^0/(\rho + p) = [\rho'(\pi)/(\rho + p)]\pi^0$$
$$= [p''(\pi) - p'(\pi)]/p'(\pi),$$

and the second of Eqs. (1) becomes

$$\theta = g(\pi)\pi^0, \qquad (7)$$

where

$$g(\pi) = 1 - (\ln p'(\pi))'. \qquad (8)$$

Conversely, let π be a function verifying (6), let $p(\pi)$ be an arbitrary function of π, and define $\rho(\pi)$ by

$$\rho(\pi) = p'(\pi) - p(\pi).$$

We have $dp = p'(\pi)d\pi = (\rho + p)d\pi$, $p^0 = (\rho + p)\pi^0$ and the first of Eqs. (1) follows. If in addition, $p(\pi)$ is a solution to (8), where $g(\pi)$ is determined by (7), the second of Eqs. (1) also follows. Thus we have shown the following proposition.

Proposition 2: The fundamental system for the barotropic perfect fluid is strictly equivalent to the system

$$d\pi = a + \pi^0 u \neq 0, \quad \theta = g(\pi)\pi^0 \qquad (9)$$

in the pair (u,π). Given such a pair, for every solution $p(\pi)$ to $g(\pi) = 1 - (\ln p'(\pi))'$ the triple (u,ρ,p) with $\rho = p'(\pi) - p$ is a barotropic fluid.

Let us note that if p and p_0 are two solutions to Eq. (8), one has $(\ln p(\pi)')' = (\ln p_0'(\pi))'$, so that $p_0 = k_1 p + k_2$ with $k_1 \geqslant 0$ and $\rho_0 = k_1\rho - k_2$. Thus if (u,ρ,p) is a barotropic fluid associated to the solution (u,π) to (9), all the other barotropic fluids associated to the same solution (u,π) are given by the biparametric family

$$(u,\ k_1\cdot\rho - k_2,\quad k_1\cdot p + k_2), \qquad (10)$$

where k_1 and k_2 are constants and $k_1 \geqslant 0$.

If (u,π) is such that $\pi^0 = 0$ one has, from (9), $da = 0$, $\theta = 0$. Thus for the solutions (u,π) to (9) that verify either $da \neq 0$ or $\theta \neq 0$, one has $\beta \equiv \pi^0 \neq 0$. The first of Eqs. (9) is (locally) equivalent to the equation expressing the closed character of the one-form $b = a + \beta u$ and the second equation implies that θ/β is a function of π, so that we have the following proposition.

Proposition 3: A unit vector field u such that $\theta\cdot da \neq 0$ is the velocity of a barotropic fluid if and only if there exists a function $\beta \neq 0$ such that

$$db = 0, \quad d(\theta/\beta) \wedge b = 0, \qquad (11)$$

where $b = a + \beta u$. For every such β, the holonomy potential π is determined, up to a constant, by $d\pi = b$.

III. CLASSIFICATION AND CHARACTERIZATION OF THE BAROTROPIC VELOCITIES

A vorticity-free unit vector field u is equivalently defined by $w \equiv *(u \wedge du) = 0$ or $du = u \wedge a$. If τ and σ are two integrating factors for u corresponding, respectively, to the potentials t and s,

$$u = \tau\, dt = \sigma\, ds, \qquad (12)$$

then the quotient τ/σ is a function of t; conversely, if τ is an integrating factor and τ/σ is a function of t, then σ is an integrating factor as well. Now, by differentiation and the interior product by u of the first equality in (12), one obtains

$$d\boldsymbol{\tau} = a + \boldsymbol{\tau}^0 u, \quad \boldsymbol{\tau} \equiv -\ln\tau, \qquad (13)$$

so that if π verifies the first of Eqs. (9), one has $d(\pi - \boldsymbol{\tau}) = (\pi^0 - \boldsymbol{\tau}^0)u$ or, equivalently, $d(\pi - \boldsymbol{\tau}) \wedge dt = 0$, that is, $\pi = \boldsymbol{\tau} + H(t)$: The function $\exp(-\pi)$ is an integrating factor. Thus we have obtained the following proposition.

Proposition 4: Let u be a vorticity-free unit vector field. The necessary and sufficient condition for π to verify $d\pi = a + \pi^0 u$ is that $\exp\{-\pi\}$ be an integrating factor for u.

Now, if $u = \sigma\, ds$ with $da = 0$, taking into account Proposition 3, one has $0 = d[(\ln\sigma)^0 u] = d[(\ln\sigma)'_s \cdot ds] \rightarrow (\ln\sigma)'_s = H(s) \rightarrow \sigma = h(s)\tau$, where τ is such that $\tau^0 = 0$: We have the following lemma characterization for the vector field u having a u-invariant integrating factor.

Lemma 1: A vorticity-free unit vector field admits a u-invariant integrating factor if and only if $da = 0$.

Let u be such that $w = 0$ and $\theta = 0$. Then according to Proposition 4, the holonomy potentials π that are solutions to $d\pi = a + \pi^0 u$ are determined by the integrating factors of u. Since $\theta = 0$, the second of Eqs. (9) is verified by taking $g(\pi) = 0$ (or $\pi^0 = 0$). Thus we have the following proposition.

Proposition 5: A unit vector fiend u verifying $w = 0$ and $\theta = 0$ is the velocity vector of the barotropic fluids having the holonomy potential π of the form $\pi = -\ln\tau$, where τ is an arbitrary integrating factor.

When $w = 0$ and $\theta \neq 0$, the hypothesis of Proposition 3 is verified. In the geodesic case $a = 0$, the integrating factors are constant, $u = dt$, and consequently, the first of Eqs. (11) reduces to $d\beta \wedge u = 0$ and, from it, the second equation becomes equivalent to $d\theta \wedge u = 0$: β and θ are of the form $\beta = \beta(t)$, $\theta = \theta(t)$. Then $d\pi = b = \beta\, dt$: The holonomy potential is a function of t as well. We have the following proposition.

Proposition 6: A unit vector field u verifying $w = 0$, $a = 0$, and $\theta \neq 0$ is the velocity of a barotropic fluid if and only if $d\theta \wedge u = 0$. The holonomy potentials π are the arbitrary functions $\pi(t)$ of the potential t of u, $u = dt$.

Let us now consider u verifying $w = 0$, $\theta\cdot a \neq 0$, and $æ \equiv a \wedge da = 0$; then one has

$$da = \alpha u \wedge a, \qquad (14)$$

where α is the scalar $\alpha \equiv i(a_*)i(u)da$ and a_* is the vector field $a_* = (1/a^2)a$. With the hypothesis of Proposition 3 being verified by the interior and exterior products by u (resp., a_*) of the first (resp., second) of Eqs. (11), we obtain, for this u,

$$(\theta/\beta)^0(a + \beta u) - \beta d(\theta/\beta) = 0, \qquad (15)$$

$$d(\theta/\beta) \wedge u \wedge a = 0, \qquad (16)$$

$$i(a_*)da + \beta^* u + \beta i(a_*)du = 0, \qquad (17)$$

$$d\beta \wedge u \wedge a = 0, \qquad (18)$$

where for any function f, $f^* = \pounds(a_*)f$. On account of (18), (16) becomes

$$d\theta \wedge u \wedge a = 0 \qquad (19)$$

and under our hypothesis, (17) is equivalent to

$$\beta^* = \beta + \alpha. \qquad (20)$$

By (19), (15) may be written as $(\theta/\beta)^0 = \beta(\theta/\beta)^*$, which in turn becomes $\beta^0 = \beta^2(1 - \theta^*/\theta) + \beta(\alpha + \theta^0/\theta)$ via (20). Thus we have the following proposition.

Proposition 7: A unit vector field u such that $w = 0$, $\theta\cdot a \neq 0$, and $a \wedge da = 0$ is the velocity of a barotropic fluid if and only if there exists a function β such that

$$d\theta \wedge u \wedge a = 0, \quad d\beta \wedge u \wedge a = 0, \qquad (21)$$

$$\beta^* = \beta + \alpha, \quad \beta^0 = \beta^2(1 - \Theta^*) + \beta(\alpha + \Theta^0), \qquad (22)$$

where

$$\alpha \equiv (1/a^2)i(a)i(u)da, \quad \Theta \equiv \ln\theta.$$

For a function f verifying $df \wedge u \wedge a = 0$, one has $df = f^0 u + f^* a$ and thus

$$\begin{aligned} &f^0 u\wedge a = df\wedge a, \quad f^* u\wedge a = -df\wedge u,\\ &df^0\wedge u\wedge a = 0, \quad df^*\wedge u\wedge a = 0, \end{aligned} \qquad (23)$$

so that if $du = u\wedge a$ and $da = \alpha u\wedge a$, one has

$$f^{0*} - f^{*0} = f^0 + \alpha f^*. \qquad (24)$$

Moreover, because of (21) and (23), the result is that all the scalars in (22) verify relation (24). From relation (24) it follows that a necessary integrability condition for Eqs. (22) is

$$\beta^{0*} - \beta^{*0} - \beta^0 - \alpha\beta^* = 0, \qquad (25)$$

which, according to (22), gives

$$\mu\beta^2 + \chi\beta + \gamma = 0, \qquad (26)$$

where

$$\mu \equiv -\Theta^{**}, \quad \chi \equiv \Theta^{*0} + \alpha^* - \alpha\Theta^*, \quad \gamma \equiv \alpha\Theta^0 - \alpha^0.$$

Let u be such that it verifies the hypothesis of Proposition 7 with $\mu^2 + \chi^2 = 0$. Equation (26) then says that γ vanishes also and (25) becomes an identity. In this case, there always exists at least one solution to Eqs. (22); a simple way to see the solution is to consider an evolution problem with the constraint equation $L \equiv \beta^* - \beta - \alpha = 0$. Taking into account the second of Eqs. (22) and (25), one finds

$$L^0 = [2\beta(1-\Theta^*) + \Theta^0]L + \mu\beta^2 + \chi\beta + \gamma,$$

so that since $\mu = \chi = \gamma = 0$, L^0 vanishes with L. Consequently, Eqs. (22) are in involution: If β is a solution of the second of Eqs. (22) in a neighborhood of a given instant and verifies the first of Eqs. (22) at that instant, then it is a solution to Eq. (22) in the neighborhood. Since the corresponding initial constraint admits a one-parametric family of solutions, we may state the following result.

Proposition 8: A unit vector field u such that $-w^2 + \ae^2 + \mu^2 + \chi^2 = 0$ and $\theta\cdot a \neq 0$ is the velocity of a barotropic fluid if and only if it verifies $d\theta\wedge u\wedge a = 0$ and $\gamma = 0$. Equations (22) admit a one-parametric family of solutions $\beta_\lambda = \beta_\lambda[u]$: For each of them, the one-form $b_\lambda = a + \beta_\lambda$ is closed and the holonomy potential π_λ is determined, up to a constant, by $d\pi_\lambda = b_\lambda$.

Suppose now that u verifies the hypothesis of Proposition 7 with $\mu^2 + \chi^2 \neq 0$. If $\mu \neq 0$ the result is that from (26) a necessary condition for (22) to admit a solution is

$$\Delta \equiv \chi^2 - 4\mu\gamma \geqslant 0. \qquad (27)$$

One then has $\beta = \beta_5$, where

$$\beta_5 = (1/2\mu)(-\chi \pm \Delta^{1/2}). \qquad (28)$$

On the other hand, if $\mu = 0$ (and, therefore, $\chi \neq 0$,), the result is that $\beta = \beta_4$, where

$$\beta_4 = -\gamma/\chi. \qquad (29)$$

Consequently, we have the following proposition.

Proposition 9: A unit vector field u such that $-w^2 + \ae^2 = 0$, $\theta\cdot a \neq 0$, and $\mu^2 + \chi^2 \neq 0$ is the velocity of the barotropic fluid if and only if it verifies either $\mu \neq 0$, $\chi \geqslant 4\mu\gamma$, (18), and (22) for $\beta = \beta_5$ as given by (28) or $\mu = 0$, (18), and (22) for $\beta = \beta_4$ as given by (29). In each case, the corresponding one-form $b_i \equiv a + \beta_i\, u$, $(i = 4,5)$ is closed and the holonomy potential π_i is determined, up to a constant, by $d\pi_i = b_i$.

Let u be such that $w = 0$ and $\theta\cdot\ae \neq 0$. In this case, taking into account that $du = u\wedge a$, the exterior product of Eqs. (11) by a implies that $a\wedge da + d\beta\wedge u\wedge a = 0$, $d(\theta/\beta)\wedge u\wedge a = 0$ and since $\theta\cdot\beta \neq 0$, it follows that

$$\beta\, d\theta\wedge u\wedge a = -\theta a\wedge da.$$

The one-form $z = -*(d\theta\wedge u\wedge a)$ does not vanish and is orthogonal to u. Consequently, $z^2 \neq 0$ and $\beta = \beta_6$, where

$$\beta_6 = (\theta/z^2)i(z)*(a\wedge da). \qquad (30)$$

Therefore, we may state the following proposition.

Proposition 10: A unit vector field u such that $w = 0$ and $\theta\cdot\ae \neq 0$ is the velocity of a barotropic fluid if and only if it verifies Eqs. (11) for $\beta = \beta_6$ as given by (30). Then the one-form $b_6 \equiv a + \beta_6 u$ is closed and the holonomy potential π_6 is determined, up to a constant, by $d\pi_6 = b_6$.

Consider now unit vector fields with $w \neq 0$ and $da = 0$. By differentiation and the exterior product by u of $d\pi = a + \pi^0 u$, one obtains $u\wedge da + \pi^0 u\wedge du = 0$, that is, $\pi^0 = 0$; thus on account of (7), $\theta = 0$. Conversely, since $da = 0$, let π be such that $d\pi = a$; then if $\theta = 0$, π is a solution to (9). Therefore, we have the following proposition.

Proposition 11: A unit vector field u such that $w \neq 0$ and $da = 0$ is the velocity of a barotropic fluid if and only if it verifies $\theta = 0$. Then the holonomy potential π is determined, up to a constant, by $d\pi = a$.

Finally, let us consider u such that $w \neq 0$ and $da \neq 0$. Since the hypothesis of Proposition 3 is verified, the result is that $u\wedge da + \beta u\wedge du = 0$ and since w is a nonvanishing spacelike vector field, one has $w^2 \neq 0$; consequently, $\beta = \beta_8$, where

$$\beta_8 \equiv -(1/w^2)i(w)*(u\wedge da). \qquad (31)$$

Thus we have the following result.

Proposition 12: A unit vector field u such that $w \otimes da \neq 0$ is the velocity of a barotropic fluid if and only if it verifies Eqs. (11) for $\beta = \beta_8$ as given by (31). Then the one-form $b_8 = a + \beta_8 u$ is closed and the holonomy potential π_8 is determined, up to a constant, by $d\pi_8 = b_8$.

In the above we have obtained conditional systems in u for the barotropic fluids. These systems depend on the nonvanishing of some differential quantities associated to u and do not admit a unique simple form valid for any unit field. On account of the above results, we are lead to introduce the following classification of unit vector fields.

Definition: A unit vector field u is said to be of class C_i $(i = 1,...,8)$ if it verifies the relations given in Table I, where we have written

$$\begin{aligned} &w = *(u\wedge du), \quad a = i(u)\nabla u, \quad \theta = -\delta u,\\ &\Theta = \ln\theta, \quad \alpha = (1/a^2)i(a)i(u)da,\\ &\mu \equiv -\Theta^{**}, \quad \chi \equiv \Theta^{*0} + \alpha^* - \alpha\Theta^*, \end{aligned}$$

and $f^0 = \pounds(u)f$, $f^* = (1/a^2)\pounds(a)f$ for any scalar f.

The results of this section may then be summarized in the following two theorems.

Theorem 1 (of characterization of barotropic velocities): A unit vector field u of class C_i $(i = 1,...,8)$ is the velocity of a barotropic perfect fluid if *and only if* it verifies the differential system B_i given in Table II, where the scalar β_j $(j = 4,5,6,8)$ is defined by

TABLE I. The eight classes of unit vector fields.

Class	Definition relations
C_1	$w=0, \theta=0$
C_2	$w=0, \theta\neq 0, a=0$
C_3	$w=0, \theta\neq 0, a\neq 0, a\wedge da=0, \mu^2+\chi^2=0$
C_4	$w=0, \theta\neq 0, a\neq 0, a\wedge da=0, \mu^2+\chi^2\neq 0, \mu=0$
C_5	$w=0, \theta\neq 0, a\neq 0, a\wedge da=0, \mu^2+\chi^2\neq 0, \mu\neq 0$
C_6	$w=0, \theta\neq 0, a\neq 0, a\wedge da\neq 0$
C_7	$w\neq 0, da=0$
C_8	$w\neq 0, da\neq 0$

TABLE III. Characterization of the holonomy potentials for a barotropic velocity.

Symbol	Characterization of π
P_1	$\pi=-\ln\tau+h(t)$, $(u=\tau\,dt)$
P_2	$\pi=\pi(t)$, $(u=\tau\,dt)$
P_3	$d\pi_\lambda=a+\beta_\lambda u$, where β_λ is the one-parametric family of solutions to the system $\beta^*=\beta+\alpha, \beta^0=\beta^2(1-\Theta^*)+\beta(\alpha+\Theta^0)$
P_4	$d\pi_4=a+\beta_4 u$
P_5	$d\pi_5=a+\beta_5 u$
P_6	$d\pi_6=a+\beta_6 u$
P_7	$d\pi=a$
P_8	$d\pi_8=a+\beta_8 u$

$$\beta_4\equiv(\alpha^0-\alpha\Theta^0)/\chi,\quad \beta_5\equiv(1/2\mu)(-\chi\pm\Delta^{1/2}),$$

$$\beta_6\equiv(\theta/z^2)i(z)*(a\wedge da),\quad \beta_8\equiv-(1/w^2)i(w)*(u\wedge da)$$

and we have written

$$\Delta\equiv\chi^2+4\mu(\alpha^0-\alpha\Theta^0),\quad z=-*(d\theta\wedge u\wedge a).$$

Theorem 2: The holonomy potential π associated to a barotropic velocity of class C_i $(i=1,...,8)$ is determined by the relations P_i given in Table III. Let $g(\pi)$ be the function such that $\theta=g(\pi)\pi^0$ and take

$$p(\pi)=\int\exp\left\{\int[1-g(\pi)]d\pi\right\}d\pi,\quad \rho(\pi)=p'(\pi)-p;$$

the triple (u,ρ,p) is then a barotropic perfect fluid.

IV. SOME SPECIAL BAROTROPIC MOTIONS: THE POLYTROPIC CASE

In many cases one may be interested in disclosing a more restricted character than that of barotropy. In this section, we study the following types of particular barotropic perfect fluids: (i) constant pressure $dp=0$; (ii) constant total energy density $d\rho=0$; (iii) u-invariant pressure (and density) $p^0=\rho^0=0$; and (iv) polytropic fluid, $p=(\lambda-1)\rho, \lambda\neq 1$.

We shall see that the characterization of these cases is easier than the general barotropic case.

Proposition 1 already characterized fluids of type (i); such fluids also belong to one of the types (ii)–(iv) if and only if $\theta=0$, so that (i) may be stated in form of the following proposition.

Proposition 13: The necessary and sufficient condition for a unit vector u to be the velocity of a perfect fluid with constant pressure and verifying one of the conditions (ii)–(iv) is that u be geodesic and expansion-free.

TABLE II. Differential systems characterizing the barotropic velocities of class C_1.

Symbol	Necessary and sufficient conditions
B_1	ϕ
B_2	$d\theta\wedge u=0$
B_3	$d\theta\wedge u\wedge a=0, \alpha\Theta^0-\alpha^0=0$
B_4	$d\theta\wedge u\wedge a=0$ $\beta_4^*=\beta_4+\alpha, \beta_4^0=\beta_4^2(1-\Theta^*)+\beta_4(\alpha+\Theta^0)$
B_5	$d\theta\wedge u\wedge a=0, \Delta\geqslant 0$ $\beta_5^*=\beta_5+\alpha, \beta_5^0=\beta_5^2(1-\Theta^*)+\beta_5(\alpha+\Theta^0)$
B_6	$d(a+\beta_6 u)=0, d(\theta/\beta_6)\wedge(a+\beta_6 u)=0$
B_7	$\theta=0$
B_8	$d(a+\beta_8 u)=0, d(\theta/\beta_8)\wedge(a+\beta_8 u)=0$

Now, let $dp\neq 0$. From Proposition 2, the barotropic relation $\rho=\rho(p)$ depends on the function $g(\pi)$ given by (7); indeed,

$$\rho'(p)=\rho'(\pi)/p'(\pi)=-g\{\pi(p)\}.$$

Thus one has $g(\pi)=$ const if and only if ρ is a linear function in p. It is then easy to see that cases (ii) and (iv) are characterized as in the following proposition.

Proposition 14: The necessary and sufficient condition for u to be the velocity of a barotropic fluid with $d\rho=0$ and $dp\neq 0$ is $\theta=0$ and $d\pi=a+\pi^0 u$ for some function π.

Proposition 15: The necessary and sufficient condition for u to be the velocity of a polytropic fluid with index λ is the existence of a function π such that $\{u,\pi\}$ is a solution to (9) with $g(\pi)=(1-\lambda)^{-1}$.

In case (iii), because of $\rho^0=p^0=0$, one has $\pi^0=0$, which by (9) leads to $\theta=0$ and $da=0$. Since the converse is also verified, one has the following proposition.

Proposition 16: The necessary and sufficient condition for u to be the velocity of a barotropic fluid with $\rho^0=p^0=0$ is $\theta=0$ and $da=0$. Then the holonomy index is determined, up to a constant, by $d\pi=a$.

When the conditions $\theta=0$, $da=0$ are verified for every function $p(\pi)$, the triple (u,ρ,p) with $\rho=p'(\pi)-p$ is a barotropic fluid verifying $\rho^0=p^0=0$. Consequently, every function $\rho=\rho(p)$ is admissible as a barotropic relation.

By additing suitable conditions to the systems B_i of Table II, one may associate barotropic relations of types (ii)–(iv) to unit vector fields of class C_i.

According to Proposition 16, the velocities of the classes C_1 and C_7 are of type (iii) if they verify $da=0$. Consequently, these velocities admit any function $\rho(p)$ as a barotropic relation and the velocities of class C_1 (with $da\neq 0$) and class C_7 (with $\theta=0$) are of constant energy density.

The velocities of classes C_8 (resp., C_6) with the additional conditions $\theta\neq 0$ and $\theta/\beta_8=$ const (resp., $\theta/\beta_6=$ const) admit polytropic barotropic relations.

The velocities of classes C_3, C_4, and C_5 admit a polytropic barotropic relation if $\beta=k\cdot\theta$ is a solution to the system (22), where k is a constant. One then has $\alpha/(\theta-\theta^0)=$ const.

Finally, the velocities of class C_2 admit any polytropic index because the holonomy potential is an arbitrary function of the potential t for u and one can always take it to be proportional, with an arbitrary constant, to a primitive of a given $\theta(t)$.

Propositions 1 and 16 characterize types (i) and (iii) in terms of u alone; meanwhile, Propositions 14 and 15 characterize types (ii) and (iv) in terms of u and π. Here we shall obtain the conditions in u ensuring the existence of π.

In case (ii) one has $\theta = 0$. When $w = 0$, Proposition 4 implies, for every integrant factor, the existence of a function π verifying $d\pi = a + \pi^0 u$. When $w \neq 0$ and $da = 0$, the potential π is such that $d\pi = a$ and if $w \neq 0$ and $da \neq 0$, according to the analysis given in Sec. III, u is a solution to $d(a + \beta_8 u) = 0$, where β_8 is given by (31). We thus have the following theorem.

Theorem 3: The necessary and sufficient conditions for u to be the velocity of a barotropic fluid with $dp \neq 0$ and $d\rho = 0$ are $\theta = 0$ and either $w = 0$ or $w \neq 0$ and $d(a + \beta_8 u) = 0$, where β_8 is given by $\beta_8 \equiv -(1/w^2) \times i(w) * (u \wedge da)$. In the first case, to every integrating factor τ corresponds a holonomy potential $\pi = -\ln \tau$; in the second case, the holonomy potential is determined, up to an additive constant, by $d\pi = a + \beta_8 u$. In both cases the triple (u, ρ_0, p) is a perfect fluid, where p is given by $p = k_0 \cdot \exp(\pi) - \rho_0$ and k_0 and ρ_0 are constants.

In case (iv), we know from Proposition 15 that u is the velocity of a polytropic fluid with index λ if and only if there exists a function π such that $d\pi = a + k\theta u$, $k = 1 - \lambda$; however, this is (locally) equivalent to

$$da + kd(\theta u) = 0, \tag{32}$$

so that $da = 0$ if and only if $d(\theta u) = 0$, where (32) then takes place for any constant k. If $da \neq 0$, for every two-form X such that $(X, da) \neq 0$, we have

$$k = -(X, da)/(X, d(\theta u)) \tag{33}$$

and by differentation

$$i(X)i'(X)\{d(\theta u) \otimes {}'\nabla\, da - da \otimes {}'\nabla d(\theta u)\} + \{i(d(\theta u))i'(da) - i(da)i'(d(\theta u))\}X \otimes {}'\nabla X = 0.$$

Since this equation is verified for every X, the two expressions inside the curly braces vanish and conversely, if they vanish, there exists a constant k such that (32) is verified. We have thus shown the following therorem.

Theorem 4: A unit vector fluid u is the velocity of a polytropic fluid if and only if it verifies either $da = d(\theta u) = 0$ or $da \otimes d(\theta u) = d(\theta u) \otimes da \neq 0$ and $da \otimes \nabla d(\theta u) = d(\theta u) \otimes \nabla\, da$. In the first case, any polytropic index $\lambda \neq 1$ is admitted; in the second case, the polytropic index $\lambda = 1 - k$ is uniquely determined by (33), where X is any two-form nonorthogonal to da. In both cases, the one-form $b \equiv a + k\theta u$ is closed and the holonomy potential associated to every k is determined, up to an additive constant, by $d\pi = b$. The triple (u, ρ, p) is a polytropic fluid of index $\lambda = 1 - k$, where $p(\pi) = k_0 \exp\{\pi \cdot \lambda / (\lambda - 1)\}$ if $k \neq 1$ and $p(\pi) = k_2 \pi$ if $k = 1$.

ACKNOWLEDGMENT

One of the authors (JJF) would like to express his thanks to the Conselleria de Cultura, Educació i Ciència de la Generalitat Valenciana for the partial support of this work.

[1] Of course, here we are not considering the natural restrictions usually added to the fundamental system for physical interpretation, namely the Plebanski energy conditions and the Lichnerowicz compressibility conditions.

[2] B. Coll, F. Fayos, and J. J. Ferrando, J. Math. Phys. **28**, 1075 (1987).

[3] B. Coll and J. J. Ferrando, J. Math. Phys. **30**, 2918 (1989).

[4] B. Coll and J. J. Ferrando, "Almost-product structures in relativity," contribution to *Relativity Meeting 89* (World Scientific, Singapore, 1990).

[5] R. Treciokas and G. F. R. Ellis, Commun. Math. Phys. **23**, 1 (1971).

[6] C. B. Collins, J. Math. Phys. **26**, 2009 (1985).

[7] Of course, there are barotropic *test* fluids which do not verify the conjecture.

[8] B. Coll and J. J. Ferrando, C. R. Acad. Sci. Paris **306**, Ser. I, 573 (1988).

[9] One has $\pi \equiv \ln F$, where F is the Synge *index function* or the Lichnerowicz *holonomy index*.

Erratum: Gauge transformations for the quadratic bundle [J. Math. Phys. 30,1744 (1989)]

Y. Vaklev
Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Boul. Lenin 72, Sofia 1784, Bulgaria

(Received 30 October 1989; accepted for publication 8 November 1989)

Formulas (2.2) for $C_j^{\pm}$ should read $\dot{a}_j^{\pm}$ instead of $a_j^{\pm}$.

Formulas (2.6) for A_p should read

$$\frac{i}{2}\left[\!\!\left[\begin{pmatrix}0\\q_1\end{pmatrix}, \Sigma\, \Lambda^p q\right]\!\!\right]$$

instead of

$$\left[\!\!\left[\frac{i}{2}\begin{pmatrix}0\\q_1\end{pmatrix} \Sigma_1\, \Lambda^p q\right]\!\!\right].$$

Formulas (2.8) and (2.11) should read:

$$\pi(\lambda)\colon = -\frac{1}{\pi}\ln\,(1+\rho^+\rho^-),\ \varkappa(\lambda,t)\colon = \frac{1}{2}\ln b^{\,+}/b^{\,-}, \tag{2.8}$$

$$\pi_j^{\pm}\colon = \pm 2i\lambda_j^{\pm},\ \varkappa_j^{\pm}\colon = \pm\ln b_j^{\pm},$$

$$\Omega^{(m)} = \cdots$$

$$+\frac{i}{2}\sum_{\substack{j=1\\ \epsilon=\pm}}^{N}(\lambda_j^{\epsilon})^m(\delta\pi_j^{\epsilon}\wedge\delta\varkappa_j^{\epsilon}) = \cdots. \tag{2.11}$$

Formulas (2.14) for e^- should read

$$\begin{pmatrix}1\\0\end{pmatrix} \text{ instead of } \begin{pmatrix}0\\1\end{pmatrix}.$$

Page 1747, the 22nd row on the left should read

$$\lim_{|\alpha|\to\infty} S = \sigma_3.$$

Formulas (3.16) for $\Omega_g^{(m)}$ and the seventh line on the right for $\Omega_{g_0}^{(m)}$ should read $[\hat{;}]$ instead of $[\ ,\ ]$. Formula (4.7) should read $\mathscr{F}(\Lambda_1')$ instead of $F(\Lambda_1')$. Formulas (4.12) for $\widetilde{M}$ should read iwS instead of iws. The left side of both formulas (4.16) and both sides of formulas (B3) should read " $,\wedge$ " instead of " $\hat{;}$ ". The fifth row of formulas (4.17) for A_{2n}' should read $\langle S_\nu,S\ldots\rangle$ instead of $\langle S_y,S\ldots\rangle$.

Page 1750, 10th and 11th lines from the bottom left below should read

$$\Omega'^{(2n)},\quad H_{\mathscr{F}}^{(2n)},\quad \widetilde{\Omega}^{(2n)},\ldots,$$

instead of

$$\Omega^{(2n)},\quad H_{\mathscr{F}}^{(2n)},\quad \Omega^{(2n)},\ldots.$$

Formulas on p. 1750, eighth line from the bottom should read $\Omega'^{(2n+1)}$ instead of $\Omega^{(2n+1)}$ and the sixth line from the bottom should read $[\,\hat{;}\,]$ instead of $[\,,\,]$.

Formulas on p. 1752 the 11th and 12th lines from the bottom left for Ω_{DLL} should read $\langle\,\hat{;}\,\rangle$ instead of $\langle\,,\,\rangle$.

Formulas (B2) for $\langle\widetilde{\widetilde{\Lambda}}'\rangle^{2n+1}X$ should read $(\lambda_0 - iwS)$ instead of $((\lambda_0) - iwS)$.

The equation number for formula (B4) is missing.

0022-2488/90/041026-01$03.00