Morphisms of crystallographic groups: Kernels and images

Marjorie Senechal
Smith College, Northampton, Massachusetts 01063

(Received 26 March 1983; accepted for publication 10 September 1984)

The crystallographic groups play an important role in solid state physics, where their representations are of particular interest. In this paper we classify the kernels of all possible representations by deriving necessary and sufficient conditions for a subgroup H of a crystallographic group G to be invariant. The structure of G/H is also discussed. A list of the one- and two-dimensional invariant subgroups of the two-dimensional crystallographic groups is appended as Table I; it includes the structural features of these subgroups needed for determining their settings, relative to the parent groups, and identifies the corresponding images. Table II is a list of the commutator subgroups of the two-dimensional groups.

I. INTRODUCTION

The crystallographic groups play an important role in many branches of mathematics, physics, and crystallography. Though these groups have been considered from many points of view in the nearly one hundred years since they were first studied, important problems still remain open. Among them is the structure of their homomorphic images.

In this paper we contribute toward the solution of this problem by deriving necessary and sufficient conditions for a subgroup H of a crystallographic group G to be invariant and for the image G/H to be a split extension of the image of the translation subgroup of G, and several related results. A table, hopefully complete, of the one- and two-dimensional invariant subgroups of the two-dimensional crystallographic groups, together with the corresponding images, is included. These tables show, surprisingly, that the invariance conditions are severe, in the sense that relatively few subgroups are invariant. This may explain the finding of Michel and Mozryzmas[1] that in the three-dimensional case there are only 37 "weakly inequivalent" images of "little groups" for representations with high-symmetry k vectors.

As far as we know, the only previous table of invariant subgroups of (two-dimensional) crystallographic groups is Table 4 of Ref. 2. That table is incomplete, since only the subgroups of least index are listed there, and they are identified only by subgroup type. For applications to representation theory and other problems, it is necessary to be able to find *all* the invariant subgroups of each group. The characterization of the space groups as extensions, which we exploit, permits such calculations more readily than their characterization by generators and relations, the method of Ref. 2.

In the next section, we discuss the basic properties of the n-dimensional crystallographic groups and their subgroups. In Sec. III we characterize the invariant subgroups and some of their properties. In Sec. IV we discuss their images and in Sec. V we explain how the tables were constructed.

II. THE CRYSTALLOGRAPHIC GROUPS AND THEIR SUBGROUPS

An n-dimensional crystallographic group G is an extension of $T=Z\times Z\times\cdots\times Z=Z^n$ by a finite subgroup P of $O(n)$. We write this as a "short exact sequence"

$$0\to T\to G\xrightarrow{\pi} P\to 1,$$

which means that T is invariant in G and P is isomorphic to the quotient group G/T. The crystallographic groups are distinguished from other extensions by the requirement that the mapping ϕ: $P\to\mathrm{Aut}\,T=\mathrm{GL}(n,Z)$ is an injection (i.e., is one-to-one). Every element of G can be written in the form $(t+\tau,p)$, where $t\in T$, $p\in\phi(P)$, and τ is a "fractional" translation, i.e., a translation not in T. If $p_1p_2=p_3$ in P, then from

$$\begin{aligned}(t_1+\tau_1,p_1)\cdot(t_2+\tau_2,p_2)&=(t_1+\tau_1+p_1t_2+p_1\tau_2,p_1p_2)\\&=((t_1+p_1t_2)+(\tau_1+p_1\tau_2),p_1p_2),\end{aligned}$$

it follows that $\tau_1+p_1\tau_2\equiv\tau_3$ (mod T); we write

$$\tau_1+p_1\tau_2-\tau_3=t^*_{p_1,p_2}. \qquad (1)$$

The set $\{t^*_{p_i,p_j}|p_i,p_j\in P\}$ is a factor set of the extension.

A subgroup H of G is an extension of $H\cap T=T_H=Z^r$, where $0\leqslant r\leqslant n$, by a subgroup P_H of P; $0\to T_H\to H\to P_H\to 1$. Thus every element of H is of the form $h=(t'+t+\tau,p)$, where $t'\in T_H$, $t\in T$, and $p\in P_H$. When $p_1p_2=p_3$ in P_H, then the translations t_1, t_2, t_3 satisfy the subgroup congruence[3]

$$t_1+p_1t_2\equiv t_3-t^*_{p_1,p_2}\quad(\mathrm{mod}\ T_H). \qquad (2)$$

ϕ induces an action ϕ_H of P_H on T_H; H is crystallographic if ϕ_H is an injection. H is always crystallographic if $r=n$ but need not be if $r<n$. For example, the seven frieze groups are all isomorphic to one-dimensional subgroups of two-dimensional crystallographic groups, but only three of them are themselves crystallographic. [There are two one-dimensional crystallographic groups (see Sec. V).]

Theorem 2.1: Let G be an n-dimensional crystallographic group, as described above. The following conditions are equivalent: (i) ϕ is an injection; (ii) T is a maximal abelian subgroup of G; and (iii) $C_G(T)=T$. [$C_G(T)$ is the centralizer of T in G, the subgroup of elements of G which commute with the elements of T.]

Proof: (i)→(iii). Suppose (τ,p), $p\neq 1$, is in $C_G(T)$. Then for every t in T, $(t+\tau,\ p)=(t,1)\cdot(\tau,p)=(\tau,p)\cdot(t,1)=(pt+\tau,p)$. Therefore $pt=t$ for every t, which contradicts (i). (iii)→(ii). If $T\subset H$ and H is abelian, then $H\subseteq C_G(T)$, which is impossible. (ii)→(i). Suppose $p\neq 1$ and $pt=t$ for

 0022-2488/85/020219-10$02.50

every t in T. Then $(\tau,p)^i \cdot (t,1) = (t,1)\cdot(\tau,p)^i$, whence $H = \cup_{i=1}^{k} T(\tau,p)^i$, where k is the order of p, is an abelian subgroup of G which properly contains T. This is a contradiction.

Corollary 1: $C(G) \subseteq T$.

Corollary 2: G is not a direct product of T and P.

The proofs of these corollaries are immediate.

Proposition 2.1 (Hermann's Theorem): If H is any subgroup of G, then there is a unique subgroup G^* of G such that $H \subseteq G^*$ and $0 \to T \to G^* \to P_H \to 1$.

Proof: $G^* = \Pi^{-1}(P_H)$; in other words, G^* is the subgroup generated by T and H.

Note: Theorem 2.1 and its corollaries, and Proposition 2.1 are all well known; we have included proofs for completeness and because these proofs are more elementary than any we have seen in the literature.

H is an invariant subgroup of G if and only if $g\,H\,g^{-1} = H$ for all g in G. This requirement leads to four conditions for invariance which will be derived in the following section.

III. INVARIANT SUBGROUPS

Theorem 3.1A: Let H be a subgroup of a crystallographic group G. Then H is invariant if and only if T_H is invariant in G and H/T_H is invariant in G/T_H.

Proof: Let H be invariant in G. Then $T_H = H \cap T$ is also invariant in G. Since the image of an invariant subgroup is invariant, H/T_H is invariant in G/T_H. Conversely, if T_H is invariant in G and H/T_H is invariant in G/T_H, then since $T_H \subseteq H \subseteq G$, H is invariant in G.

However, the following formulation of this theorem is more useful in computation. A computer program based on it has recently been developed by Engel, who is preparing tables of equivalence classes of invariant subgroups of the two- and three-dimensional crystallographic groups.[4]

Theorem 3.1B: Let $H = \cup_{p_i \in P_H} T_H(t_i + \tau_i, p_i)$ be a subgroup of G. Then H is invariant if and only if (i) P_H is invariant in P; (ii) T_H is invariant in G; (iii) $p_i\, t - t \equiv 0 \pmod{T_H}$ for every $p_i \in P_H$ and every $t \in T$; and (iv) if $p \in P$, $p_i \in P_H$, and $p p_i p^{-1} = p_j$, then

$$(pt_i - t_j) + (t^*_{p,p_i} - t^*_{p_j,p}) \equiv 0 \pmod{T_H}.$$

Proof: Let $g = (t + \tau, p) \in G$ and $h = (t' + t_i + \tau_i, p_i) \in H$, where $t' \in T_H$ and $p_i \in P_H$. Then

$$
\begin{aligned}
g\,h\,g^{-1} &= (t+\tau,p)\cdot(t' + t_i + \tau_i, p_i)\cdot(-p^{-1}t - p^{-1}\tau, p^{-1}) \\
&= (t + \tau + pt' + pt_i + p\tau_i - pp_i p^{-1} t \\
&\quad - pp_i p^{-1}\tau, pp_i p^{-1}) \\
&= (t'' + t_j + \tau_j, p_j),
\end{aligned}
$$

where $pp_i p^{-1} = p_j$ in P. H is invariant if and only if $(t'' + t_j + \tau_j, p_j) \in H$, or

(a) $pp_i p^{-1} = p_j \in P_H$,

and

(b) $(t - p_j\, t) + (pt_i - t_j) + (t^*_{p,p_i} - t^*_{p_j,p}) = t'' \in T_H$.

First assume that H is invariant. Then (i) follows immediately from (a). To establish (ii), let $t = 0$ and $p_i = 1$ in (b). Then only the expression in the second parentheses is nonzero. Similarly, (iii) follows if we set $p = 1$, $t' = 0$, and we obtain (iv) by setting $t = t' = 0$. Conversely, it is clear that if (i)–(iv) hold, so do (a) and (b).

Corollary: If H is a proper invariant subgroup of G, then $\dim T_H > 0$.

Proof: If $T_H = \{0\}$, then 3.1(iii) would imply that $p_i\, t = t$ and so, by Theorem 2.1(i), G would not be a crystallographic group.

The following useful theorem, while not new,[5] is not well known. We present an elementary proof.

Theorem 3.2: An invariant subgroup H of a crystallographic group is crystallographic.

Proof: We will show that $C_H(T_H) = T_H$. Let $h = (t_0 + \tau, p) \in H$, with $t_0 \in T$ and $p \in P_H$. If $h \in C_H(T_H)$ then, as in the proof of Theorem 2.1, $pt' = t'$ for every $t' \in T_H$. Let $t \in T$. Since H is invariant, $pt = t + t'$ for some $t' \in T_H$. Then $p^i t = t + it'$ for every positive integer i. Let k be the order of p; then $kt' = 0$. This implies $p = 1$.

Theorem 3.3: Let H be an invariant subgroup of G and let $r = \dim T_H$. Every element of P can be represented in $\mathrm{GL}(n,Z)$ by a matrix of the form $\begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$, where $A \in \mathrm{GL}(r,Z)$ and $C \in \mathrm{GL}(n-r,Z)$. If $p \in P_H$ then $C = I_{n-r}$.

Proof: We identify translations with vectors in E^n. A basis for the lattice T_H spans an r-dimensional hyperplane U. Let $T' = T \cap U$; then $\dim T' = r$ and $T_H \subseteq T' \subseteq T$. Choose a basis for T' and extend it to a basis of T. With respect to this basis, the elements of ϕ_P all have the form $\begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$; that $C = I_{n-r}$ if $p \in P_H$ follows from Theorem 3.1B(iii).

We group together some immediate consequences which are useful in computation.

Corollary: (i) If $r < n$ then $-I_n \notin \phi(P_H)$.

(ii) For $p_1, p_2 \in P_H$, $A_1 = A_2$ if and only if $p_1 = p_2$; in particular if $A = I_r$ then $B = 0$ and p is the identity.

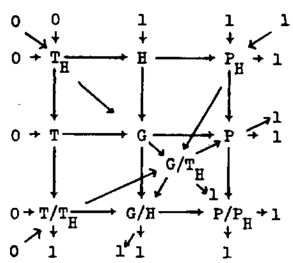
(iii) If $r = 1$ then p is either a reflection in an $(n-1)$-dimensional hyperplane or the identity.

(iv) If $\phi(P)$ is Z-irreducible then G has no invariant subgroups of dimension less than n.

IV. IMAGES OF THE CRYSTALLOGRAPHIC GROUPS

The preceding discussion implies that G and H satisfy a Michel diagram[6] of short exact sequences.

Theorem 4.1: If H is an invariant subgroup of G, then all aligned arrows in the diagram below are short exact sequences:



Proof: This follows immediately from Theorem 3.1 and the "3×3 lemma,"[7] which states that the exactness of two adjacent rows (or columns) in the diagram implies that the third is also exact.

TABLE I. Invariant subgroups and images for $n = 2$. In part A, $r = 1$. Nine of the 17 two-dimensional crystallographic groups have invariant subgroups of dimension 1. It follows from the corollary to Theorem 3.3 that such a subgroup either consists of translations alone—$l\,1$—or is a semidirect product $lm1$ of $l\,1$ and a group of order 2 generated by reflection in a line perpendicular to the direction of translation. In part B, $r = 2 = n$. In the first column we list the groups G, and in the second the isomorphism types of their invariant subgroups H. The H's are characterized in the next three columns: the generators of T_H appear in column 3 as a matrix of column vectors and in columns 4 and 5 we list the admissible t_i's corresponding to the generators of P_H. Here, P_H is cyclic or dihedral, with a single generator s or m, or a pair of generators $\{s,m\}$. The information in columns 3, 4, and 5 specifies the setting of H relative to b. A dagger in column 7 means that G/H is a split extension, an $\times$ means that it is not, while a dash means that the extension is trivial. We give the common name for G/H in column 8, if it has one. The index $[G:H]$ of H in G is the order of the image group, which is equal to the product of $[T:T_H]$ and $[P:P_H]$; it can easily be calculated by the reader.

A. Subgroups of infinite index.

G	H	T_H	t_m	T/T_H	P/P_H	Split
$p1$	$l\,1$	$\begin{bmatrix} a \\ b \end{bmatrix}$	–	$Z_k \times Z$ $k = gcd(a,b)$	1	–
$p2$	$l\,1$	$\begin{bmatrix} a \\ b \end{bmatrix}$	–	same	Z_2	†
pm	$l\,1$	$\begin{bmatrix} a \\ 0 \end{bmatrix}$	–	$Z_a \times Z$	Z_2	†
		$\begin{bmatrix} 0 \\ d \end{bmatrix}$	–	$Z \times Z_d$	Z_2	†
	$lm1$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	(0,0)	Z	1	–
		$\begin{bmatrix} 0 \\ 2 \end{bmatrix}$	(0,0) (0,1)	$Z \times Z_2$	1	–
pg	$l\,1$	$\begin{bmatrix} a \\ 0 \end{bmatrix}$	–	$Z_a \times Z$	Z_2	† iff a is odd
		$\begin{bmatrix} 0 \\ d \end{bmatrix}$	–	$Z \times Z_d$	Z_2	$\times$
cm	$l\,1$	$\begin{bmatrix} a \\ a \end{bmatrix}$	–	$Z_a \times Z$	Z_2	†
		$\begin{bmatrix} c \\ -c \end{bmatrix}$	–	$Z \times Z_c$	Z_2	†
	$lm1$	$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$	(0,0)	Z	–	–
pmm	$l\,1$	$\begin{bmatrix} a \\ 0 \end{bmatrix}$	–	$Z_a \times Z$	$Z_2 \times Z_2$	†
		$\begin{bmatrix} 0 \\ d \end{bmatrix}$	–	$Z \times Z_d$	$Z \times Z_2$	†
	$l\,1$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	(0,0)	Z	Z_2	†
		$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$	(0,0) (1,0)	$Z_2 \times Z$	Z_2	†
		$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	(0,0)	Z	Z_2	†
		$\begin{bmatrix} 0 \\ 2 \end{bmatrix}$	(0,0) (0,1)	$Z \times Z_2$	Z_2	†
pmg	$l\,1$	$\begin{bmatrix} a \\ 0 \end{bmatrix}$	–	$Z_a \times Z$	$Z_2 \times Z_2$	$\times$
		$\begin{bmatrix} 0 \\ d \end{bmatrix}$	–	$Z \times Z_d$	$Z_2 \times Z_2$	† iff d is odd
	$lm1$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	(0,0)	Z	Z_2	†
		$\begin{bmatrix} 0 \\ 2 \end{bmatrix}$	(0,0) (0,1)	$Z \times Z_2$	Z_2	$\times$
pgg	$l\,1$	$\begin{bmatrix} a \\ 0 \end{bmatrix}$	–	$Z_a \times Z$	$Z_2 \times Z_2$	$\times$
		$\begin{bmatrix} 0 \\ d \end{bmatrix}$	–	$Z \times Z_d$	$Z_2 \times Z_2$	$\times$
cmm	$l\,1$	$\begin{bmatrix} a \\ a \end{bmatrix}$	–	$Z_a \times Z$	$Z_2 \times Z_2$	†
		$\begin{bmatrix} c \\ -c \end{bmatrix}$	–	$Z \times Z_c$	$Z_2 \times Z_2$	†
	$lm1$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	(0,0)	Z	Z_2	†
		$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$	(0,0)	Z	Z_2	†

TABLE I (*Continued.*)

B. Subgroups of finite index.

G	H	T_H	t_s	t_m	T/T_H	P/P_H	Split	Name of image
$p1$	$p1$	$\begin{bmatrix} a & c \\ 0 & d \end{bmatrix}$	–	–	$Z_k \times Z_{ad/k}$ $k = \gcd a,c,d$	1	–	
$p2$	$p1$	$\begin{bmatrix} a & c \\ 0 & d \end{bmatrix}$	–	–	same	Z_2	†	
	$p2$	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	1	–	Z_2
		$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	1	–	Z_2
		$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (0,1)	–	Z_2	1	–	Z_2
		$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (1,0) (0,1) (1,1)	–	$Z_2 \times Z_2$	1	–	$Z_2 \times Z_2$ $= V_4 = D_2$
$p3$	$p1$	$k\begin{bmatrix} a & c \\ 0 & 1 \end{bmatrix}$ $c^2+c+1=ma$	–	–	$Z_k \times Z_{ak}$	Z_3	†	
	$p3$	$\begin{bmatrix} 3 & -1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0) (2,0)	–	Z_3	1	–	Z_3
$p4$	$p1$	$k\begin{bmatrix} a & -c \\ 0 & 1 \end{bmatrix}$ $c^2+1=ma$	–	–	$Z_k \times Z_{ak}$	Z_4	†	
	$p2$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
		$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	×	V_4 Z_4
		$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (1,1)	–	$Z_2 \times Z_2$	Z_2	†	D_4
	$p4$	$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	1	–	Z_2
$p6$	$p1$	$k\begin{bmatrix} a & -c \\ 0 & 1 \end{bmatrix}$ $c^2+c+1=ma$	–	–	$Z_{ka} \times Z_k$	Z_6	†	
	$p2$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	–	Z_3	–	Z_3
		$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0)	–	$Z_2 \times Z_2$	Z_3	†	A_4
	$p3$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
		$\begin{bmatrix} 3 & -1 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	Z_3	Z_2	–	D_3
pm	$p1$	$\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$	–	–	$Z_a \times Z_d$ $= Z_k \times Z_{ad/k}$	Z_2	†	
		$\begin{bmatrix} 2a & a \\ 0 & d \end{bmatrix}$	–	–	$Z_k \times Z_{2ad/k}$	Z_2	†	
	pm	$\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$	–	(0,0)	Z_a	1	–	Z_a
		$\begin{bmatrix} a & 0 \\ 0 & 2 \end{bmatrix}$	–	(0,0) (0,1)	$Z_a \times Z_2$	1	–	Z_{2a} if a is odd
	pg	$\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$ a even	–	$(a/2,0)$	Z_a	1	–	Z_a
		$\begin{bmatrix} a & 0 \\ 0 & 2 \end{bmatrix}$ a even	–	$(a/2,0)$ $(a/2,1)$	$Z_a \times Z_2$	1	–	$Z_a \times Z_2$
	cm	$\begin{bmatrix} 2a & a \\ 0 & 1 \end{bmatrix}$	–	(0,0) (0,1)	Z_{2a}	1	–	Z_{2a}
pg	$p1$	$\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$	–	–	$Z_a \times Z_d$ $= Z_k \times Z_{ad/k}$	Z_2	† iff a is odd	
		$\begin{bmatrix} 2a & a \\ 0 & d \end{bmatrix}$	–	–	$Z_k \times Z_{ad/k}$	Z_2	×	

TABLE I (*Continued.*)

B. Subgroups of finite index.

G	H	T_H	t_s	t_m	T/T_H	P/P_H	Split	Name of image
	pg	$\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}_{a \text{ odd}}$	–	$\left(\frac{a-1}{2},0\right)$	Z_a	1	–	Z_a
		$\begin{bmatrix} a & 0 \\ 0 & 2 \end{bmatrix}_{a \text{ odd}}$	–	$\left(\frac{a-1}{2},0\right)$ $\left(\frac{a-1}{2},1\right)$	Z_{2a}	1	–	Z_{2a}
cm	$p1$	$k\begin{bmatrix} a & c \\ 0 & 1 \end{bmatrix}_{c^2-1=ma}$	–	–	$Z_k \times Z_{ak}$	Z_2	†	
	pm	$\begin{bmatrix} a & -1 \\ 0 & 1 \end{bmatrix}_{a \text{ even}}$	–	(0,0)	Z_a	1	–	Z_a, a even
	pg	$\begin{bmatrix} a & -1 \\ 0 & 1 \end{bmatrix}_{a \text{ even}}$	–	$(a/2,0)$	Z_a	1	–	Z_a, a even
	cm	$\begin{bmatrix} a & -1 \\ 0 & 1 \end{bmatrix}_{a \text{ odd}}$	–	(0,0)	Z_a	1	–	Z_a, a odd
pmm	$p1$	$\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$	–	–	$Z_a \times Z_d$ $= Z_k \times Z_{ad/k}$	$Z_2 \times Z_2$	†	
		$\begin{bmatrix} 2a & a \\ 0 & d \end{bmatrix}$	–	–	$Z_k \times Z_{2ad/k}$	$Z_2 \times Z_2$	†	
	pm	$\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$	–	(0,0)	Z_a	Z_2	†	D_a
		$\begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}$	–	(0,0)	Z_d	Z_2	†	D_d
		$\begin{bmatrix} a & 0 \\ 0 & 2 \end{bmatrix}$	–	(0,0) (0,1)	$Z_a \times Z_2$	Z_2	†	$D_a \times Z_2 = D_{2a}$ if a odd
		$\begin{bmatrix} 2 & 0 \\ 0 & d \end{bmatrix}$	–	(0,0) (1,0)	$Z_2 \times Z_d$	Z_2	†	$D_d \times Z_2 = D_{2d}$ if d odd
	pg	$\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}_{a \text{ even}}$	–	$(a/2,0)$	Z_a	Z_2	†	D_a
		$\begin{bmatrix} a & 0 \\ 0 & 2 \end{bmatrix}_{a \text{ even}}$	–	$(a/2,0)$ $(a/2,1)$	$Z_a \times Z_2$	Z_2	†	$D_a \times Z_2$
		$\begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}_{d \text{ even}}$	–	$(0,d/2)$	Z_d	Z_2	†	D_d
		$\begin{bmatrix} 2 & 0 \\ 0 & d \end{bmatrix}_{d \text{ even}}$	–	$(0,d/2)$ $(1,d/2)$	$Z_2 \times Z_d$	Z_2	†	$D_d \times Z_2$
	cm	$\begin{bmatrix} 2a & a \\ 0 & 1 \end{bmatrix}$	–	(0,0) (0,1)	Z_{2a}	Z_2	†	D_{2a}
		$\begin{bmatrix} 2 & 1 \\ 0 & d \end{bmatrix}$	–	(0,0) (1,0)	$Z_2 \times Z_d$	Z_2	†	$D_d \times Z_2 = D_{2d}$ *if* d odd
	$p2$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
		$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (0,1)	–	Z_2	Z_2	†	V_4
		$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	†	V_4
		$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	†	V_4
		$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (1,0) (0,1) (1,1)	–	$Z_2 \times Z_2$	Z_2	†	$Z_2 \times Z_2 \times Z_2$
	pmm	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	(0,0) (0,0)	Z_2	1	–	Z_2
		$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (0,1)	(0,0) (0,1)	Z_2	1	–	Z_2

TABLE I (*Continued.*)

B. Subgroups of finite index.

G	H	T_H	t_s	t_m	T/T_H	P/P_H	Split	Name of image
		$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0) (1,0) (0,1) (1,1)	(0,0) (0,0) (0,1) (0,1)	$Z_2\times Z_2$	1		V_4
	pmg	$\begin{bmatrix}1&0\\0&2\end{bmatrix}$	(0,0) (0,1)	(0,1) (0,0)	Z_2	1	–	Z_2
		$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0) (1,0) (0,1) (1,1)	(0,1) (0,1) (0,0) (0,0)	$Z_2\times Z_2$	1	–	V_4
	pgm	$\begin{bmatrix}2&0\\0&1\end{bmatrix}$	(0,0) (1,0)	(1,0) (1,0)	Z_2	1	–	Z_2
		$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0) (1,0) (0,1) (1,1)	(1,0) (1,0) (1,1) (1,1)	$Z_2\times Z_2$	1	–	V_4
	pgg	$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0) (1,0) (0,1) (1,1)	(1,1) (1,1) (1,0) (1,0)	$Z_2\times Z_2$	1	–	V_4
	cmm	$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (1,0) (0,0) (1,0)	(0,0) (0,0) (1,0) (1,0)	Z_2	1	–	Z_2
pmg	*p*1	$\begin{bmatrix}a&0\\0&d\end{bmatrix}$	–	–	$Z_a\times Z_d$ $=Z_k\times Z_{ad/k}$	$Z_2\times Z_2$	† iff d odd	
		$\begin{bmatrix}2a&a\\0&d\end{bmatrix}$	–	–	$Z_2\times Z_{2ad/k}$	$Z_2\times Z_2$	×	
	pm	$\begin{bmatrix}a&0\\0&1\end{bmatrix}$	–	(0,0)	Z_a	Z_2	†	D_a
	pg	$\begin{bmatrix}a&0\\0&1\end{bmatrix}_{a\text{ even}}$	–	(a/2,0)	Z_a	Z_2	†	D_a
		$\begin{bmatrix}1&0\\0&d\end{bmatrix}_{d\text{ odd}}$	–	(0,(d – 1)/2)	Z_d	Z_2	†	D_d
		$\begin{bmatrix}2&0\\0&d\end{bmatrix}_{d\text{ odd}}$	–	(0,(d – 1)/2) (1,d – 1)/2)	$Z_2\times Z_d$	Z_2	†	D_{2d}
	*p*2	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	Z_2		Z_2
		$\begin{bmatrix}2&0\\0&1\end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	†	V_4
	pmg	$\begin{bmatrix}2&0\\0&1\end{bmatrix}$	(0,0) (1,0)	(0,0) (0,0)	Z_2	1	–	Z_2
	pgg	$\begin{bmatrix}2&0\\0&1\end{bmatrix}$	(0,0) (1,0)	(1,0) (1,0)	Z_2	1	–	Z_2
pgg	*p*1	$\begin{bmatrix}a&0\\0&d\end{bmatrix}$	–	–	$Z_a\times Z_d$	$Z_2\times Z_2$	† iff a,d odd	
		$\begin{bmatrix}2a&a\\0&d\end{bmatrix}$	–	–	$Z_k\times Z_{2ad/k}$	$Z_2\times Z_2$	×	
	*p*2	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	×	Z_4
	pg	$\begin{bmatrix}a&0\\0&1\end{bmatrix}_{a\text{ odd}}$	–	((a – 1)/2,0)	Z_2	Z_2	†	Z_{2a}
		$\begin{bmatrix}1&0\\0&d\end{bmatrix}_{d\text{ odd}}$	–	(0,(d – 1)/2)	Z_d	Z_2	†	Z_{2d}

TABLE I (*Continued.*)

B. Subgroups of finite index.

G	H	T_H	t_s	t_m	T/T_H	P/P_H	Split	Name of image
cmm	*p*1	$k\begin{bmatrix} a & -c \\ 0 & 1 \end{bmatrix}_{c^2-1=ma}$	–	–	$Z_{ka}\times Z_k$	$Z_2\times Z_2$	†	
	*p*2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	–	–	1	Z_2	–	Z_2
		$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	†	V_4
		$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (1,1)	–	$Z_2\times Z_2$	Z_2	†	D_4
	pm	$\begin{bmatrix} a & -1 \\ 0 & 1 \end{bmatrix}_{a \text{ even}}$	–	(0,0)	Z_a	Z_2	†	D_a
		$\begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix}_{4\mid a \text{ even}}$	–	(a/2,0)	Z_a	Z_2	†	D_a
	pg	$\begin{bmatrix} a & -1 \\ 0 & 1 \end{bmatrix}_{a \text{ even}}$	–	(a/2,0)	Z_a	Z_2	†	D_a
cmm	*pg*	$\begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix}_{4\nmid a \text{ even}}$	–	(a/2,0)	Z_a	Z_2	†	D_a
	cm	$\begin{bmatrix} a & -1 \\ 0 & \end{bmatrix}_{a \text{ odd}}$	–	(0,0)	Z_a	Z_2	†	D_a
		$\begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix}_{a \text{ odd}}$	–	(0,0)	Z_a	Z_2	†	D_a
	pmm	$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0)	(0,0)	Z_2	1	–	Z_2
	pmg	$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(1,0)	(0,0)	Z_2	1	–	Z_2
	pgm	$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(1,0)	(1,0)	Z_2	1	–	Z_2
	pgg	$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0)	(0,1)	Z_2	1	–	Z_2
*p*3*m*1	*p*1	$\begin{bmatrix} a & -0 \\ 0 & -a \end{bmatrix}$	–	–	$Z_a\times Z_a$	D_3	†	
		$\begin{bmatrix} 3a & -a \\ 0 & a \end{bmatrix}$	–	–	$Z_{3a}\times Z_a$	D_3	†	
	*p*3	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	–	–	1	Z_2	–	Z_2
		$\begin{bmatrix} 3 & -1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0) (2,0)	–	Z_3	Z_2	†	D_3
*p*31*m*	*p*1	$\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$	–	–	$Z_a\times Z_a$	D_3	†	
		$\begin{bmatrix} 3a & -a \\ 0 & a \end{bmatrix}$	–	–	$Z_{3a}\times Z_a$	D_3	†	
	*p*3	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
		$\begin{bmatrix} 3 & -1 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	Z_3	Z_2	†	Z_6
	*p*3*m*1	$\begin{bmatrix} 3 & -1 \\ 0 & 1 \end{bmatrix}$	(0,0)	(0,0)	Z_3	1	–	Z_3
*p*4*m*	*p*1	$\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$	–	–	$Z_a\times Z_a$	D_4	†	
		$\begin{bmatrix} 2a & a \\ 0 & a \end{bmatrix}$	–	–	$Z_{2a}\times Z_a$	D_4	†	
	*p*2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(0,0)	–	1	$Z_2\times Z_2$	–	V_4
		$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$	(0,0) (1,0)	–	Z_2	$Z_2\times Z_2$	† ×	$Z_2\times Z_2\times Z_2$ D_4
		$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	(0,0) (1,1)	–	$Z_2\times Z_2$	$Z_2\times Z_2$	†	$Z_2\times D_4$

TABLE I (*Continued.*)

B. Subgroups of finite index.

G	H	T_H	t_s	t_m	T/T_H	P/P_H	Split	Name of image
	$p4$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (1,0)	–	Z_2	Z_2	†	V_4
	pmm	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	(0,0)	1	Z_2	–	Z_2
		$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0) (1,1)	(0,0) (0,1)	$Z_2\times Z_2$	Z_2	†	D_4
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (0,0)	(0,0) (1,0)	Z_2	Z_2	†	V_4
	pgg	$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0) (1,1)	(1,1) (1,0)	$Z_2\times Z_2$	Z_2	†	D_4
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0)	(1,0)	Z_2	Z_2	†	V_4
	cmm	$\begin{bmatrix}1&0\\0&0\end{bmatrix}$	–	–	1	Z_2	–	Z_2
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0)	(0,0)	Z_2	Z_2	†	V_4
	$p4m$	$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (1,0)	(0,0) (1,0)	Z_2	1	–	Z_2
	$p4g$	$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (1,0)	(1,0) (0,0)	Z_2	1	–	Z_2
$p4g$	$p1$	$\begin{bmatrix}a&0\\0&a\end{bmatrix}$	–	–	$Z_a\times Z_a$	D_4	† iff a odd	
		$\begin{bmatrix}2a&a\\0&a\end{bmatrix}$	–	–	$Z_{2a}\times Z_a$	D_4	×	
	$p2$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	$Z_2\times Z_2$	–	V_4
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(0,0) (1,0)	–	Z_2	$Z_2\times Z_2$	× ×	$Z_4\times Z_2$ D_4
	$p4$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
	pmm	$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(1,0)	(1,0)	Z_2	Z_2	×	Z_4
	pgg	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	(0,0)	1	Z_2	–	Z_2
		$\begin{bmatrix}2&1\\0&1\end{bmatrix}$	(1,0)	(0,0)	Z_2	Z_2	×	Z_4
$p6m$	$p1$	$\begin{bmatrix}a&0\\0&a\end{bmatrix}$	–	–	$Z_a\times Z_a$	D_6	†	
		$\begin{bmatrix}3a&-a\\0&a\end{bmatrix}$	–	–	$Z_{3a}\times Z_a$	D_6	†	
	$p2$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	D_3	–	D_3
		$\begin{bmatrix}2&0\\0&2\end{bmatrix}$	(0,0)	–	$Z_2\times Z_2$	D_3	†	T_d
	$p3$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	$Z_2\times Z_2$	–	V_4
		$\begin{bmatrix}3&-1\\0&1\end{bmatrix}$	(0,0)	–	Z_3	D_2	†	D_6
	$p6$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	–	1	Z_2	–	Z_2
	$p3m1$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	(0,0)	1	Z_2	–	Z_2
		$\begin{bmatrix}3&-1\\0&-1\end{bmatrix}$	(0,0)	(0,0)	Z_3	Z_2	†	D_3
	$p31m$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	(0,0)	(0,0)	1	Z_2	–	Z_2

Although G/H is an extension of T/T_H by P/P_H, it does not follow that G/H is a finite version of a crystallographic group. For example, if H contains the commutator subgroup of G (see Table II) then G/H is abelian, which a crystallographic group cannot be unless $P=\{1\}$.

Let $P_M=\{p\in P\,|\,C(p)=I_{n-r}\}$ (see Theorem 3.3), and let $M=\Pi^{-1}(P_M)$. Then $H\subseteq M\subseteq G$ and $T_M=T$. Thus M is the maximal subgroup of G containing H such that $pt-t\equiv 0 \pmod{T_H}$ for each $p\in P_M$ and every $t\in T$. Write M as a union of cosets of H, $M=\cup H(t_j+\tau_j,p_j)$ and let $t_1,\ldots,t_r$ be a set of coset representatives of T_H in T. Then, modulo T_H, $(t_i,1)\cdot(t_j+\tau_j,p_j)=(t_j+\tau_j,p_j)\cdot(t_i,1)$, which shows that $M/H\subseteq C_{G/H}(T/T_H)$. Since M is the maximal subgroup with this property, $M/H=C_{G/H}(T/T_H)$. Summarizing, we have the following theorem.

Theorem 4.2: The centralizer of T/T_H in G/H is the image of the subgroup M of G defined above. $T_M=T$ and $P_H\subseteq P_M\subseteq P$, so $C_{G/H}(T/T_H)$ is an extension of T/T_H by P_M/P_H.

If $P_M=P_H$, then since Theorem 2.1 obviously holds for extensions of any abelian group, G/H can be regarded as a "finite crystallographic group."

In representation theory it is customary to "finitize" the crystallographic groups by replacing T by a direct product of n finite cyclic groups. This amounts to considering, instead of G, its image modulo a normal subgroup of finite index which is contained in T. If this index is $r_1 r_2\cdots r_n$, with each r_i sufficiently large, then the image will indeed be a finite crystallographic group in our sense. However, it is important to note that its structure may differ from that of G. In particular, G may be a semidirect product and the image not, or vice versa.

Theorem 4.3: G/H is a split extension of T/T_H (direct or semidirect product) if and only if there exists a subgroup H^* of G containing H such that $0\to T_H\to H^*\to P\to 1$.

Note: Unlike G^* of Proposition 2.1, H^* need not exist. If it does exist, then G/H is a direct product if $P_M=P$ (Theorem 4.2).

Proof: Let $\beta: G\to G/H$. Assume H^* exists. Then it contains H as an invariant subgroup and we can replace G by H^* in the diagram. Since $\beta H^*=P/P_H$, G/H contains a subgroup isomorphic to P/P_H which, by construction, has only the identity in common with T/T_H. Therefore G/H is a split extension. Conversely, if G/H is a split extension then there exists a subgroup H^* of G such that H^* is isomorphic to P/P_H and $\beta H^*\cap T/T_H=\{(0,1)\}$. This implies $T_{H^*}=T_H$ and therefore $P_{H^*}=P$.

Corollary: If G is a split extension of T and P is a split extension of P_H then G/H is a split extension of T/T_H.

Proof: By hypothesis, G contains a subgroup S isomorphic to P and S contains a subgroup Q isomorphic to P/P_H. Then $H^*=\cup_{q_i\in Q}H(0,q_i)$ is a subgroup of G containing H.

Table I shows that both hypotheses are necessary for $n=2$.

V. KERNELS AND IMAGES FOR $n=2$

The invariant subgroups of the seventeen two-dimensional crystallographic groups are listed in Table I. In the first column we list the groups G, and in the second the isomorphism types of their invariant subgroups H. The H's are characterized in the next three columns: the generators of T_H appear in column 3 as a matrix of column vectors, and in columns 4 and 5 we list the admissible (Theorem 3.1B) t_i's corresponding to the generators of P_H. Here, P_H is cyclic or dihedral, with a single generator s or m, or a pair of generators $\{s,m\}$. (The subgroups of the two-dimensional crystallographic groups are discussed in detail in Ref. 8.) A dagger in column 7 means that G/H is a split extension, an $\times$ means that it is not, while a dash means that the extension is trivial. We give the common name for G/H in column 8, if it has one. Reference 2 is suggested for a survey of discrete groups.

When G/H is finite, we have used coset charts to facilitate identification. (These charts should not be confused with the coset tables in Ref. 2.) Since $0\to T/T_H\to G/T_H\to P\to 1$, we can represent G/T_H by an array with T/T_H rows and $|P|$ columns. The columns are grouped into $|P|/|P_H|$ blocks: at the heads of the columns in the first block we write the elements of P_H, and then write the elements of the cosets of P_H in P at the heads of the columns in the remaining blocks. The elements of T/T_H are recorded at the beginnings of the rows. For example, let $G=p4$, $H_1=p2$ with lattice $\begin{bmatrix}2&1\\0&1\end{bmatrix}$ and $t_s=(0,0)$, and H_2 also $p2$, with the same lattice, but with $t_s=(1,0)$.

(a) In both cases the empty chart looks like this:

	1	s^2	s	s^3
(0,0)				
(1,0)				

To fill in the charts, first locate H_1/T_H and H_2/T_H in G/T_H by assigning the letter a to the squares representing the coset representatives $(t_i+\tau_i,p_i)$:

(b)

(i) $t_s=(0,0)$

	1	s^2	s	s^3
(0,0)	a	a		
(1,0)				

(ii) $t_s=(1,0)$

	1	s^2	s	s^3
(0,0)	a			
(1,0)		a		

We complete them by assigning letters to the squares corresponding to the cosets of H_1/T_H and H_2/T_H:

(c)

(i)

	1	s^2	s	s^3
(0,0)	a	a	c	c
(1,0)	b	b	d	d

(ii)

	1	s^2	s	s^3
(0,0)	a	b	c	d
(1,0)	b	a	d	c

Since $1\to P_H\to G/T_H\to 1$, the set of letters forms a group isomorphic to G/H. The chart thus records the structural information about G/H which is contained in the diagram, and also describes the way in which H is embedded in G. This information, together with the generators of the lattice T_H and the factor set of G, gives a complete description of G/H. From it, we can determine the orders of the elements of G/H and the relations among the generators, and construct a multiplication table for the group.

TABLE II. The commutator subgroups of the two-dimensional crystallographic groups. The commutator subgroups consist of proper motions. In each case $t_s = (0,0)$.

G	G'
$p1$	$\{0\}$
$p2$	$p1\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$
$p3$	$p1\begin{bmatrix}3 & -1\\0 & 1\end{bmatrix}$
$p4$	$p1\begin{bmatrix}2 & 1\\0 & 1\end{bmatrix}$
$p6$	$p1\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
pm	$l1\begin{bmatrix}0\\2\end{bmatrix}$
pg	$l1\begin{bmatrix}0\\2\end{bmatrix}$
cm	$l1\begin{bmatrix}1\\-1\end{bmatrix}$
pmm	$p1\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$
pmg	$p1\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$
pgg	$p1\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$
cmm	$p1\begin{bmatrix}2 & 1\\0 & 1\end{bmatrix}$
$p31m$	$p3\begin{bmatrix}3 & -1\\0 & 1\end{bmatrix}$
$p3m1$	$p3\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
$p4m$	$p2\begin{bmatrix}2 & 1\\0 & 1\end{bmatrix}$
$p4g$	$p2\begin{bmatrix}2 & 1\\0 & 1\end{bmatrix}$
$p6m$	$p3\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

Sometimes this information can be obtained rapidly by inspection. Thus in the example above, the chart (c)(ii) shows immediately that G/H is a cyclic group of order 4. An easy calculation shows that the elements b, c, and d in chart (c)(i) are all of order 2, from which it follows that G/H is the four-group V_4.

The computation is further simplified by the observation that the images of subgroups which are equivalent under an automorphism of the parent group are necessarily isomorphic.

The tables show that if P_H is nontrivial then the image G/H is one of the following types: Z_a (cyclic of order a), D_a (dihedral of order $2a$), $Z_a \times Z_2$, $D_a \times Z_2$, $Z_2 \times Z_2 \times Z_2$, $T(A_4)$, or T_d (S_4). If $P_H = \{1\}$ then $H = T_H$ and G/H is an extension of T/T_H by P. For the most part, these groups do not have "common" names.

In Table II we list the commutator subgroups of the two-dimensional groups, as this does not seem to be available elsewhere. Since the product of an even number of improper motions (negative determinant) is proper, the commutator subgroups, generated by elements of the form $g_1 g_2 g_1^{-1} g_2^{-1}$ certain proper motions only. For $n = 2$, $t_s = (0,0)$ in each case.

ACKNOWLEDGMENTS

I am indebted to Louis Michel and Marko Jarić for many stimulating conversations. I am also grateful to them and to J. J. Burckhardt, H. S. M. Coxeter, Peter Engel, and Richard Roth for very helpful comments on the preliminary version of this paper. Finally, I would also like to thank I. H. E. S., Bures-sur-Yvette, France for its hospitality during the spring of 1983.

[1]L. Michel and J. Mozrzymas, "Weak equivalence of irreducible representations of little space groups," Match **10**, 223 (1981).

[2]H. S. M. Coxeter and W. O. J. Moser, *Generators and Relations for Discrete Groups* (Springer, New York, 1972), 3rd edition.

[3]M. Senechal, Acta Crystallog. A **36**, 845 (1980).

[4]P. Engel and M. Senechal, "Normal subgroups of plane and space groups" (to be published).

[5]D. Farkas, Rocky Mountain J. Math. **11**, 511 (1981).

[6]L. Michel, "Relations entre symetries internes et invariance relativiste," in *Lectures in Theoretical Physics*, edited by Flurcat (Gordon and Breach, New York, 1965), p. 409.

[7]S. Maclane, *Homology* (Academic, New York, 1963), p. 49.

[8]M. Senechal, Discrete Appl. Math. **1**, 51 (1979).

An infinite hierarchy of conservation laws and nonlinear superposition principles for self-dual Einstein spaces

Charles P. Boyer[a)]
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México 20, D. F., Mexico
Jerzy F. Plebański
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México 20, D. F., Mexico and Centro de Investigación y Estudios Avanzados del IPN, México D. F. 14, Ap. Post. 14-740, Mexico

(Received 6 February 1984; accepted for publication 28 June 1984)

Self-dual Einstein spaces are shown to admit an infinite hierarchy of conservation laws, and this hierarchy is then used to derive a formal version of Penrose's twistor construction. The set of formal holomorphic bundles of fiber dimension 2 over the Riemann sphere P^1 is shown to form a formal infinite group which is used to derive nonlinear superposition principles. As an example of our methods a new self-dual Einstein space is obtained as the result of a "collision" of complex *pp*-waves "traveling in opposite directions."

I. INTRODUCTION

Several years ago the authors[1] showed that the complexified self-dual Einstein spaces (or *H*-spaces) have associated with them a hierarchy of closed one-forms or, in the language of partial differential equations, conservation laws (or first integrals). After checking the existence of 14 closed one-forms, we conjectured that the hierarchy was, in fact, infinite. In this article we use the intrinsic calculus of lifts to higher-order tangent bundles to prove this conjecture and then use the hierarchy to deduce formally Penrose's curved twistor construction.[2]

We construct a formal symplectic structure on the space of formal holomorphic curves. The existence of an infinite number of conservation laws then allows us to characterize self-dual structures as certain maximal isotropic submanifolds of complex dimension 4. Penrose's twistor construction in our formulation becomes the symplectic fact that maximal isotropic submanifolds can be described locally as the graph of certain formal twisted canonical transformations. It is then shown that these formal twisted canonical transformations form a formal infinite group which is used to derive nonlinear superposition principles for the nonlinear graviton.

The methods used in this paper offer several new insights into the curved twistor construction: First, since we avoid the use of infinitesimal deformation theory, our methods may be more amenable to developing a global theory which would entail a study of the global behavior of the maximal isotropic submanifolds. Furthermore, we give an independent proof of the fact that, under certain locality assumptions, the space of holomorphic curves of curved twistor space is a four-complex-dimensional manifold.

Second, since we represent self-dual structures as a certain formal infinite group, all of the power of group theoretical methods may be brought to bear on the problem. The relevant group which we call the group of twisted canonical transformations is isomorphic to an abelian extension of the formal group $G \otimes C[[t,t^{-1}]]$, where G is the formal group of volume-preserving formal diffeomorphisms of C^2.

The Lie algebra of $G \otimes C[[t,t^{-1}]]$ is a Kac–Moody type algebra but with the Lie algebra of G the Lie algebra of formal divergence-free vector fields on C^2. There is thus a formal analogy with Arnold's[3] description of hydrodynamics in terms of mechanics on G, and it would be interesting to see how the method of coadjoint orbits applies to our case.

Third and finally, our approach to the self-dual Einstein equations using conservation laws and infinite groups provides a much closer connection with other important problems of mathematical physics, namely the soliton evolution equations, the two-dimensional chiral models, the axial symmetric stationary Einstein equations, as well as the self-dual Yang–Mills equations. The main thrust of these examples including the twistor construction is that the problem of solving certain nonlinear partial differential equations is transformed into a problem involving patching together holomorphic data. Moreover, there appears to be a deep relationship between this transform and both the theory of infinite-dimensional Lie algebras and the theory of holomorphic curves.

It is hoped that the methods developed in this paper can eventually be applied to the problem of solving the real Einstein equations. Although we are still far from realizing this goal, we believe it quite plausible that some remnant of our nonlinear superposition principle will survive when constructing real Einstein spaces.

II. THE DIFFERENTIAL EQUATIONS AND CONSERVATIONS LAWS

In 1975 Plebański[4] showed that the self-dual Einstein equations could be reduced locally to solving the one nonlinear partial differential equation

[a)] Present address: Department of Mathematics and Computer Science, Clarkson College of Technology, Potsdam, NY 13676.

 0022-2488/85/020229-06$02.50

$$\tfrac{1}{2}\Omega_{q^A\tilde q^B}\Omega_{q_A\tilde q_B} = 1, \tag{1}$$

where subindices $q^A\tilde q^B$ denote partial derivatives with respect to these variables. Here we are employing spinor coordinates $(q^A,\tilde q^A)$, $A = 1,2$ for the complex space-time C^4. The indices for these coordinates as with all other spinor quantities are to be raised and lowered by using the totally antisymmetric Levi–Civita symbols ϵ_{AB}, ϵ^{AB}, where $\epsilon_{12} = \epsilon^{12} = 1$. Our convention is thus $q_A = \epsilon_{AB}q^B$, $\tilde q^A = \epsilon^{BA}\tilde q_B$, where we also use summation over repeated indices.

Equation (1) is equivalent to one of two conservation laws[5]

$$\partial_{\tilde q_B}(\Omega_{q^A\tilde q^B}\Omega_{q_A} - \tilde q_B) = 0, \tag{2a}$$

or

$$\partial_{q_A}(\Omega_{q^A\tilde q^B}\Omega_{\tilde q_B} - q_A) = 0. \tag{2b}$$

These equations give rise to locally defined potentials Σ and $\tilde\Sigma$ satisfying

$$\Omega_{q^A\tilde q^B}\Omega_{q_A} - \tilde q_B = \Sigma_{\tilde q^B},$$
$$\Omega_q A_{\tilde q^B}\Omega_{\tilde q_{AB}} - q_A = \tilde\Sigma_{q^A}.$$

It was also shown in Ref. 4 that if one makes the changes of variables $(q^A,\tilde q^B) \mapsto (q^A,p_B)$ by defining $p_A = \Omega_{q^A}$ then (1) is equivalent to[6]

$$\tfrac{1}{2}\Theta_{p^Ap^B}\Theta_{p_Ap_B} + \Theta_{p_Aq^A} = 0, \tag{3}$$

for some potential function Θ. The local function Θ also arises from a conservation law. For if we consider Θ as a function of $(q^A, \tilde q^B)$, then

$$\Omega_{q^A\tilde q^B}(\tfrac{1}{2}\Omega_{q^C}\Omega_{q_Cq_A} + \Sigma_{q_A}) = \Theta_{\tilde q^B},$$

and the left-hand side of this equation defines conserved densities. Equation (3) itself can be written as a conservation law, namely

$$\partial_{p_A}(\tfrac{1}{2}\Theta_{p^Ap^B}\Theta_{p_B} + \Theta_{q^A}) = 0. \tag{4}$$

In Ref. 1 the authors showed that this process of obtaining conserved quantities continues. By casting the differential equation (1) in terms of a closed ideal of differential forms we showed by prolonging this ideal to larger differential ideals how to obtain 14 new conservation laws in the form of closed one-forms.

Let us recall[1,4] how the metric is obtained from the potentials Ω or Θ:

$$ds^2 = \Omega_{q^A\tilde q^B}\, dq^A\, d\tilde q^B = dq^A(dp_A - \Theta_{p^Ap^B}\, dq^B). \tag{5}$$

The latter form using the potential Θ and Eq. (3) has proved to be very efficient in finding explicit self-dual Einstein metrics (cf. Ref. 7 and references therein). However, the former with the Ω-potential is more geometric. Indeed, it is immediate from the form of the metric that the two-dimensional surfaces defined by $q^A = \text{const}$ or $\tilde q^A = \text{const}$ are null surfaces. Furthermore, they are totally geodesic. These totally geodesic null two-surfaces will play an extremely important role in what follows. Penrose has shown[2] that through every point $(q^A,\tilde q^A)$ there is a complex projective plane's worth of totally geodesic null two-surfaces. This lies at the heart of the twistor construction.

We shall now introduce a working definition of self-dual structure on the complexified space-time M. We assume for simplicity[8] that M is the direct product of two-dimensional complex manifolds, i.e., $M \simeq M_2 \times \tilde M_2$. Assume M_2 and $\tilde M_2$ are both endowed with complex symplectic structures, that is, complex-valued closed nondegenerate two-forms, ω and $\tilde\omega$, respectively. Furthermore, suppose that M itself is a complexified Kähler manifold,[9] i.e., there is a complex-valued nondegenerate closed two-form Ω_0 on M which in local coordinates $(q^A,\tilde q^A)$ on $M_2 \times \tilde M_2$ can be written as

$$\Omega_0 = \tfrac{1}{2}\Omega_{q^A\tilde q^B}\, dq^A \wedge d\tilde q^B. \tag{6}$$

Here locally the triple $(\omega,\tilde\omega,\Omega_0)$ is a basis in the bundle of anti-self-dual two-forms on M. The differential equation (1) is equivalent to the quadratic relation

$$2\Omega_0 \wedge \Omega_0 + \omega \wedge \tilde\omega = 0. \tag{7}$$

We refer to the triple of closed two-forms $(\omega,\tilde\omega,\Omega_0)$ satisfying (7) as a *self-dual structure* on M. Notice that Ω_0 uniquely determines the metric (5).

III. HIGHER-ORDER TANGENT BUNDLES

Let M be any complex manifold and define the rth-order tangent bundle[10] of r-jets of holomorphic curves from the origin in C to anywhere in M. Let $T^\infty M$ denote the inverse limit of the T^rM. For any holomorphic function f on M, we can define the λ th lift[10] $f^{(\lambda)}$ to T^rM by

$$f^{(\lambda)}(j_r \circ \psi(0)) = \frac{1}{\lambda!}\left.\frac{d^\lambda f\circ\psi}{dt^\lambda}\right|_{t=0}, \tag{8}$$

where $j_r \circ \psi$ denotes the r-jet of ψ. Similarly, we can lift tensor fields to T^rM. For example, for vector fields, define $X^{(\lambda)}f^{(\nu)} = (Xf)^{(\lambda+\nu-r)}$ if $\lambda + \nu \geqslant r$ and zero otherwise, and one-forms by $\omega^{(\lambda)}(X^{(\nu)}) = \omega(X)^{(\lambda+\nu-r)}$ if $\lambda + \nu \geqslant r$ and zero otherwise. Extend this operation to the full exterior bundle on T^rM by C-linearity and the formula

$$(\omega_1 \wedge \omega_2)^{(\lambda)} = \sum_{\mu=0}^{\lambda} \omega_1{}^{(\mu)} \wedge \omega_2{}^{(\lambda-\mu)}. \tag{9}$$

We refer to $X^{(\lambda)}$ and $\omega^{(\lambda)}$ as the λ th lift of X and ω, respectively.

In particular, we will be interested in two-forms on T^rM and $T^\infty M$. Indeed, one easily verifies the following proposition.

Proposition 1: If (M,ω) is a symplectic manifold, then $(T^rM,\omega^{(r)})$ is a symplectic manifold.

Now consider a one-complex-parameter family of holomorphic two-forms on T^rM,

$$\omega^r(t) = \sum_{\lambda=0}^{r} \omega^{(\lambda)}t^\lambda, \quad t\in C. \tag{10}$$

It follows (cf. Ref. 8) from Proposition 1 that for all nonvanishing $t\in C$, $\omega^r(t)$ defines a symplectic two-form on T^rM, if ω is a symplectic from on M.

IV. SYMPLECTIC GEOMETRY ON THE SPACE OF CURVES

The space $T^\infty M$ can be thought of as the space of parametrized formal curves on M (in the sense of formal power series). Now suppose $M = M_2 \times \tilde M_2$ (or more generally M

has a local product structure), then there is a splitting of the rth-order tangent bundle. Indeed, in the inverse limit we have

$$T^\infty M \simeq \rho^* T^\infty M_2 \oplus \tilde{\rho}^* T^\infty \widetilde{M}_2, \tag{11}$$

where $\rho(\tilde{\rho})$ denotes the projection onto $M_2(\widetilde{M}_2)$, respectively. We can turn $T^\infty M$ into a formal symplectic manifold as follows. First, consider $T^\infty M_2$ and the formal two-form

$$\omega_2(t) = \lim_{\leftarrow} \omega_2^r(t) = \sum_{k=0}^{\infty} \pi_k^* \omega^{(k)} t^k, \tag{12}$$

where π^k: $T^\infty M_2 \to T^k M_2$ is the natural projection.

For any complex manifold M a holomorphic section ω of the bundle $\Lambda^2 T^\infty M \otimes C[[t_1,...,t_l]]$ is called a formal two-form on $T^\infty M$. It is closed if $d\omega = 0$ (d is the exterior derivative on $T^\infty M$) and nondegenerate if for every $p \in T^\infty M$, $\omega_p(u,v) = 0$ for all $v \in T_p T^\infty M$ implies $u = 0$. The pair $(T^\infty M, \omega(t_1,...,t_l))$, where ω is a closed nondegenerate formal two-form, is called a *formal symplectic manifold*. It is easy to see that $T^\infty M_2$ with $\omega_2(t)$ given by (12) is a formal symplectic manifold. Furthermore, one easily sees that if $(T^\infty M_2, \omega_2(t))$ and $(T^\infty \widetilde{M}_2, \tilde{\omega}_2(s))$ are formal symplectic manifolds, then so is

$$(T^\infty M_2 \times T^\infty \widetilde{M}_2, \pi^* \omega_2(t) + \tilde{\pi}^* \tilde{\omega}_2(s)),$$

where π and $\tilde{\pi}$ denote the projections onto $T^\infty M_2$ and $T^\infty \widetilde{M}_2$, respectively. Now we can identify $T^\infty M_2 \times T^\infty \widetilde{M}_2 \simeq T^\infty (M_2 \times \widetilde{M}_2) \simeq T^\infty M \simeq \rho^* T^\infty M_2 \times \tilde{\rho}^* T^\infty \widetilde{M}_2$. Let us define

$$\omega(t) = t^{-1} \rho^* \omega_2(t) - t \tilde{\rho}^* \omega_2(t^{-1}). \tag{13}$$

Then we have the following proposition.

Proposition 2: $(T^\infty M, \omega(t))$ is a formal symplectic manifold.

In order to condense the notation we shall no longer write π_r^k in the pullback of forms to higher-order tangent spaces. Thus we shall consider $\omega^{(\lambda)}$ to be "living" on any $T^r M_2$ with $r > \lambda$ including $r = \infty$. This should cause no confusion as it should be clear from the context which tangent bundle we are working on. We shall be interested in the two-form $\sigma^* \omega^{(1)}$ on M, for a holomorphic section $\sigma: M \to T^\infty M$. We remark that fixing the two-form $\sigma^* \omega^{(1)}$ on M there is still freedom in the choice of that part of σ which has its image in the fibers of $T^\infty M \to TM$.

We present our main result of this section.

Theorem 2: Let σ: $M \to T^\infty M$ be a holomorphic section. The triple $(\omega, \tilde{\omega}, \Omega_0) := (\sigma^* \omega^{(0)}, \sigma^* \tilde{\omega}^{(0)}, \frac{1}{2} \sigma^* \omega^{(1)})$ defines a self-dual structure on M if and only if there is a choice of holomorphic section σ such that $\sigma^* \omega(t) = 0$.

Remark: The theorem asserts that the self-dual structure on M is coded into $T^\infty M$ by the graph of σ being an isotropic submanifold. In fact, it can be shown that it is maximal isotropic in the sense that it is not contained in a larger isotropic submanifold.[11] Hence, every self-dual structure on M defines an isotropic immersion of M into the space of formal holomorphic curves.

Proof: Let σ: $M \to T^\infty M$ be a holomorphic section such that $\sigma^* \omega(t) = 0$. We will construct a self-dual structure on M. To do so we need only consider second-order objects, i.e., $T^2 M$. This corresponds to the coefficients of t^0, t^1, in $\sigma^* \omega(t) = 0$. Explicitly, we have

$$\sigma^*(\omega^{(1)} - \tilde{\omega}^{(1)}) = 0, \quad \sigma^*(\omega^{(2)} - \tilde{\omega}^{(0)}) = 0. \tag{14}$$

The two-form $\Omega_0 := \frac{1}{2} \sigma^* \omega^{(1)}$ is a closed two-form on M which by Eqs. (14) must have the form (6). We show that (7) holds by virtue of the second of Eqs. (14). From (9) we have the identity

$$0 = (\omega \wedge \omega)^{(2)} = 2\omega^{(2)} \wedge \omega^{(0)} + \omega^{(1)} \wedge \omega^{(1)}.$$

The second equation of (14) implies $\sigma^* \omega^{(2)} = \tilde{\omega}$. Thus using the identity above, we have $\omega \wedge \tilde{\omega} = -\frac{1}{2} \sigma^* (\omega^{(1)} \wedge \omega^{(1)}) = -2\Omega_0^2$.

Conversely (6) implies the first of Eqs. (14), and we can retrace our steps to show that (6) and (7) imply $\sigma^*(\omega^{(2)} - \omega^{(0)}) \wedge \omega = 0$. By the freedom of choice of σ along the fibers of $T^2 M_2 \to TM_2$ we can choose $\sigma^* \omega^{(2)} = \sigma^* \tilde{\omega}^{(0)}$. We need to show that we can choose σ such that the remaining coefficients of $\sigma^* \omega(t)$ vanish. We first notice that the map sending $q^A \to \tilde{q}^A$ and $t \to t^{-1}$ is an involution of our structure. Thus the tilded version of (14) follows from (14). To complete the proof of the theorem it suffices to prove the following lemma.

Lemma: If the holomorphic section σ satisfies (14) then there is a choice of σ such that $\sigma^* \omega^k = 0$ for all $k \geqslant 3$.

Proof: By induction on k. For $k = 3$ we have the identity

$$(\omega \wedge \omega)^{(3)} = 2(\omega^{(3)} \wedge \omega^{(0)} + \omega^{(2)} \wedge \omega^{(1)}) = 0.$$

Using this and (14) gives

$$\sigma^* \omega^{(3)} \wedge \omega = -\sigma^*(\omega^{(2)} \wedge \omega^{(1)}) = -\sigma^*(\tilde{\omega}^{(0)} \wedge \tilde{\omega}^{(1)}) = 0.$$

Again by choice of σ along the fibers of $T^3 M_2 \to T^2 M_2$ we obtain $\sigma^* \omega^{(3)} = 0$. Now assume we can choose $\sigma^* \omega^{(j)} = 0$ for all $j = 3,...,k$. Again by (9) we have

$$0 = (\omega \wedge \omega)^{(k+1)} = \sum_{j=0}^{k+1} \omega^{(j)} \wedge \omega^{(k+1-j)}.$$

Applying σ^* to this identity and using the induction hypothesis, we obtain for $k \geqslant 3$,

$$2\sigma^* \omega^{(k+1)} \wedge \omega = 0.$$

For $k = 3$ this equation follows from (14). So again we obtain $\sigma^* \omega^{(k+1)} = 0$ by choice of σ along the fibers of $T^{(k+1)} M_2 \to T^k M_2$. This proves the lemma and thus the theorem. Q.E.D.

We shall briefly indicate how the theorem gives rise to an infinite set of closed one-forms on M, and thus describes the conservation laws of Sec. II. We work locally in an open set $S \subset M$ so that the closed two-form $\omega(t)$ can be written as

$$\omega(t) = d\tau(t),$$

for some section τ over S of $\Lambda^1 T^\infty M \otimes C[[t,t^{-1}]]$. Then our theorem implies that $\sigma^* \tau(t)$ is closed if and only if σ represents a self-dual structure on S. Notice that $\tau(t)$ is not unique, for if we add to $\tau(t)$ the exact differential $d\lambda(t)$, $\lambda \in \mathscr{O}(T^\infty S) \otimes C[[t,t^{-1}]]$, where $\mathscr{O}(T^\infty S)$ denotes the ring of holomorphic functions on $T^\infty S$, then $\omega(t)$ is left unaltered. Again locally we obtain a generating function $\Omega \in \mathscr{O}(S) \otimes C[[t,t^{-1}]]$ such that

$$\sigma^* \tau(t) = d\Omega(t). \tag{15}$$

By an appropriate choice of $\tau(t)$, Eq. (15) gives Eq. (4.39) of

Ref. 1, in particular we reproduce Eqs. (4.17), (4.26), and (4.33) of Ref. 1 (there are a few erroneous minus signs in this reference). With this choice of $\tau(t)$, $\Omega(0) = \Omega$, $\Omega(1) = \Sigma$, $\Omega(-1) = \widetilde{\Sigma}$, $\Omega(2) = \Theta$, and $\Omega(-2) = \widetilde{\Theta}$.

V. THE CURVED TWISTOR CONSTRUCTION

We shall show how the theorem of the previous section can be used to derive Penrose's curved twistor or nonlinear graviton construction.[2] (A good reference for the twistor construction is Wells.[12]) This is based on the following fact from symplectic geometry[11]: Consider a symplectic manifold (M,ω) and define $\omega^- = -\omega$. Here (M,ω^-) is also a symplectic manifold. Furthermore, if $\pi_1(\pi_2)$ denotes the projection of $M\times M$ onto the first (second) factor, respectively, then $(M\times M, \pi_1^*\omega + \pi_2^*\omega^-)$ is a symplectic manifold. Let us consider Lagrangian submanifolds of $M\times M$ which project under π_1 onto open submanifolds S of M. Such Lagrangian submanifolds can be identified with the graph of a canonical transformation $\phi: S\to M$. Now let M be complexified space-time and consider the formal symplectic manifold $(T^\infty M, \omega(t))$ and suppose that M_2 and $\widetilde{M}_2$ are diffeomorphic. So we have $M \simeq M_2\times M_2$ and $T^\infty M \simeq T^\infty M_2 \times T^\infty M_2$. The formal symplectic two-form $\omega_2(t)$ on $T^\infty M_2$ given by Eq. (12) can be viewed as a presymplectic two-form on $T^\infty M_2\times C^*$, where C^* denotes the nonvanishing complex numbers. Consider holomorphic maps $\widehat{F}: T^\infty M_2\times C^* \to T^\infty M_2\times C^*$ of the form $\widehat{F} = (F,I)$, where $I(t) = t^{-1}$. Now the graph of F, gr F annihilates $\omega(t)$, i.e., $(\text{gr } F)^*\omega(t) = 0$ if and only if

$$F^*\omega_2(t^{-1}) = t^{-2}\omega_2(t). \tag{16}$$

A computation in local coordinates then shows that gr $F \xrightarrow{i} T^\infty M \xrightarrow{\pi} M$ is an embedding of gr F onto an open submanifold S of M. Thus gr F can be identified with a local section $\sigma(S)$, for some holomorphic section $\sigma: S\to T^\infty M|S$ satisfying $\sigma^*\omega(t) = 0$.

Following Penrose[2] let us now construct a class of three-dimensional fibrations over the Riemann sphere $P^1 = C\cup\{\infty\}$, $\nu: \mathscr{T}\to P^1$. Cover $\mathscr{T}$ with two coordinate charts $(\mathscr{N}, z)$ and $(\widetilde{\mathscr{N}}, \tilde{z})$ with $\mathscr{N}\cap\widetilde{\mathscr{N}}$ an open subset of $\mathscr{T}$. Let (z^A,t) and $(\tilde{z}^A,s)$ denote coordinates on $\mathscr{N}$ and $\widetilde{\mathscr{N}}$, respectively. Now define the transition function $\mathrm{F}: z(\mathscr{N}\cap\widetilde{\mathscr{N}}) \to \tilde{z}(\mathscr{N}\cap\widetilde{\mathscr{N}})$ by

$$s = t^{-1}, \quad \tilde{z}^A = F^A(z^B,t), \tag{17}$$

where F^A are holomorphic functions on $z(\mathscr{N}\cap\widetilde{\mathscr{N}})$. Consider the natural injection $\nu^*: \nu^*T^*P^1\to T^*\mathscr{T}$ and let $\mathscr{Q}^*$ denote the quotient bundle on $\mathscr{T}$. Let $\mathscr{O}(1)$ denote the hyperplane bundle on P^1 and $\mathscr{O}(n)$ its nth-order tensor product.[13] If we require that there exists a global section μ of the bundle $\Lambda^2\mathscr{Q}^*\otimes\nu^*\mathscr{O}(2)$ which is closed under exterior differentiation in $\mathscr{Q}^*$ and nondegenerate there, then the transition functions $F^A(z^B,t)$ must satisfy

$$\det\frac{\partial F^A}{\partial z^B} = \frac{1}{2}\frac{\partial F^A}{\partial z^B}\frac{\partial F_A}{\partial z_B} = t^{-2}. \tag{18}$$

Conversely, if (18) is satisfied then we can construct a global nondegenerate closed holomorphic section μ of $\Lambda^2\mathscr{Q}^*\otimes\nu^*\mathscr{O}(2)$. Finally, we must demand that the normal bundle N_ψ to a section ψ of ν is isomorphic to the direct sum of two copies of the hyperplane bundle, i.e., $N_\psi \simeq \mathscr{O}(1)\oplus\mathscr{O}(1)$. The *curved twistor space* is the pair $(\mu,\mathscr{T})$.

To make contact with (16) consider the set $\Gamma(\mathscr{T})$ of global holomorphic sections of ν, i.e., compact holomorphic curves in $\mathscr{T}$. In local coordinates a holomorphic section $\psi\in\Gamma(\mathscr{T})$ sends $t\in U_0\subset P^1$ to $(t, z^A = \psi^A(t))$ and $s\in U_\infty$ to $(s,\tilde{z}^A = \tilde{\psi}^A(s))$, where U_0, U_∞ are open disks containing the points 0 and ∞ in P^1, respectively, and satisfying $U_0\cap U_\infty \neq \phi$. If we restrict μ to the local holomorphic section ψ: $U_0\to\mathscr{T}|_{\nu^{-1}U_0} \simeq U_0\times M_2$ with $N_\psi\simeq\mathscr{O}(1)\oplus\mathscr{O}(1)$, we obtain

$$\mu|_{\psi(t)} = \tfrac{1}{2}d\psi^A(t)\wedge d\psi_A(t) = \sigma^*\omega_2(t). \tag{19}$$

Thus (18) implies (16) and $\sigma = \text{gr } F$ defines a convergent holomorphic section of $T^\infty M$ which annihilates $\omega(t)$.

Conversely, suppose σ is a holomorphic section of $T^\infty M$ over the open set $S\subset M$ which annihilates $\omega(t)$. For $t\in C^*$ we write $\sigma = (\psi^A(t), \tilde{\psi}^B(t^{-1}))$ and assume that $\psi^A(t)$, $\tilde{\psi}^B(t^{-1})$ converge on the open disks U_0 and U_∞, respectively, with $U_0\cap U_\infty\neq\phi$. Then $t\mapsto\psi^A(t)$ and $s\mapsto\tilde{\psi}^B(s)$ define local holomorphic sections of $\mathscr{T}$ which by (16) patch together globally to give the holomorphic section $\psi\in\Gamma(\mathscr{T})$. Moreover, since σ has rank 4 the normal bundle N_ψ is isomorphic to $\mathscr{O}(1)\oplus\mathscr{O}(1)$. This is Penrose's curved twistor construction.[2] The local nature of the solutions of Eq. (1) is encoded in the global holomorphic structure of $\mathscr{T}$. Furthermore the set of holomorphic sections of $\mathscr{T}$ is parametrized by the points of S; therefore, as sets, we can identify $\Gamma(\mathscr{T})$ with S.

Our previous discussion suggests that we should consider the twistor construction formally, i.e., we consider the holomorphic functions F^A of (17) to be understood in the sense of formal power series (formal Laurent series in t). Thus corresponding to every such formal transition function F satisfying (18) we construct a *formal holomorphic bundle* $\mathscr{T}$ on P^1. As mentioned previously we can identify gr F with a formal holomorphic section $\sigma: S\to T^\infty M|S$ over some open submanifold $S\subset M$ such that $\sigma^*\omega(t) = 0$. This corresponds formally to a self-dual structure on S.

VI. THE GROUP OF FORMAL TWISTED CANONICAL TRANSFORMATIONS

In this section we shall give a brief description of a group theoretical treatment of formal twistor theory. This appears to be an important first step in constructing a viable *nonlinear superposition principle* for the nonlinear graviton. In order not to entangle ourselves in problems of convergence we work formally. At this stage algebraic properties are of foremost importance.

Consider C^3 with complex Cartesian coordinates $(z^A,t) = (z^1,z^2,t)$. Denote by $C^3 - C^2$ the complex submanifold of C^3 obtained by deleting the hyperplane $t = 0$. Let $\mathscr{D}$ denote the set of all formal diffeomorphisms from $C^3 - C^2$ into itself, and $\mathscr{C}$ the subset of $\mathscr{D}$ satisfying (17) and (18), where the F^A's are understood as formal power series on $C^3 - C^2$ (thus formal Laurent series in t). An element $F\in\mathscr{C}$ is called a *twisted canonical transformation*. Notice that $\mathscr{C}$ is not a subgroup of $\mathscr{D}$ since it does not contain the identity diffeomorphism. However, in $\mathscr{C}$ there is a distinguished element $\mathscr{J}$ defined by

$$\mathscr{J}(z^A,t) = (t^{-1}z^A,t^{-1}). \tag{20a}$$

Clearly, $\mathscr{J}^2 = id \in \mathscr{D}$. We give $\mathscr{C}$ the structure of a formal group by defining the composition of two elements F', $F \in \mathscr{C}$ by

$$F' * F = F' \circ \mathscr{J} \circ F, \tag{20b}$$

where $\circ$ denotes composition as formal diffeomorphisms. One easily checks that the relation (20) is associative and that $F * \mathscr{J} = \mathscr{J} * F = F$, i.e., $\mathscr{J}$ is the identity element of the group. The inverse element of F is $\mathscr{J} \circ F^{-1} \circ \mathscr{J}$.

Each element $F \in \mathscr{C}$ determines a formal holomorphic bundle $\mathscr{T}$ on P^1 of rank 2, i.e., fiber dimension 2. Let $\mathscr{B}$ denote the set of formal rank 2 holomorphic bundles on P^1. We can give $\mathscr{B}$ the structure of a group by defining $T' * T$ to be the bundle determined by the formal transition function $F' * F \in \mathscr{C}$. The identity in $\mathscr{B}$ is the bundle $\mathscr{T}_0$ determined by $\mathscr{J} \in \mathscr{C}$ whose total space is the direct sum of two copies of the hyperplane bundle, i.e., $\mathscr{T}_0 = \mathscr{O}(1) \oplus \mathscr{O}(1)$. Clearly as groups, $\mathscr{C}$ and $\mathscr{B}$ are isomorphic.

Our next result characterizes $\mathscr{C}$ in terms of a "Kac–Moody-type" group. Consider the formal group $\mathrm{GL}(2,C) \otimes C[[z^A,t,t^{-1}]]$ and the subset $\mathrm{SL}(2,C)_t \otimes C[[z^A,t,t^{-1}]]$ consisting of all $A \in GL(2,C) \otimes C[[z^A,t,t^{-1}]]$ such that $\det A = t^{-2}$. This is not a subgroup of $\mathrm{GL}(2,C) \otimes C[[z^A,t,t^{-1}]]$; however, we can give $\mathrm{SL}(2,C)_t \otimes C[[z^A,t,t^{-1}]]$ the structure of a formal group by defining group multiplication by $A * B = tA \cdot B$, A, $B \in \mathrm{SL}(2,C)_t \otimes C[[z^A,t,t^{-1}]]$, where $A \cdot B$ means matrix multiplication as matrices of formal series. We can easily verify the following proposition.

Proposition 3: The map $\rho : \mathrm{SL}(2,C)_t \otimes C[[z^A,t,t^{-1}]] \to \mathrm{SL}(2,C) \otimes C[[z^A,t,t^{-1}]]$ defined by $\rho(A) = tA$ is a group isomorphism.

Now consider the "Jacobian map" $J : \mathscr{C} \to \mathrm{SL}(2,C)_t \otimes C[[z^A,t,t^{-1}]]$ defined by sending $F \in \mathscr{C}$ to the Jacobian matrix

$$JF^A_B = \frac{\partial F^A}{\partial z^B}(z,t).$$

Here, J is a group homomorphism, for

$$J(F' * F) = JF' \cdot J\mathscr{J} \cdot JF = tJF' \cdot JF = JF' * JF.$$

The kernel of J consists of the "translations" defined by

$$\ker J = \{ T \in \mathscr{C} : T(z^B,t) = t^{-1}z^B + C^B(t), C^B(t) \in C[[t,t^{-1}]] \}.$$

For $l \in \mathrm{SL}(2,C) \otimes C[[z^A,t,t^{-1}]]$ consider the formal local coframe $l^A_B\, dz^B$. By exterior differentiation we have $d(l^A_B\, dz^B) = -\frac{1}{2}(\partial l^A_B/\partial z_B) dz^C C \wedge dz_C$ We denote by d the map sending l^A_B to $-\frac{1}{2}(\partial l^A_B/\partial z_B)$. By the Poincaré lemma $\ker d$ is just the set of Jacobian matrices with unit determinant. We have thus arrived at the following theorem.

Theorem 2: There is an exact sequence

$$1 \to \ker \mathscr{J} \xrightarrow{i} \mathscr{C} \xrightarrow{\rho \circ \mathscr{J}} \mathrm{SL}(2,C) \otimes C[[z^A,t,t^{-1}]] \xrightarrow{d}$$

of formal groups where i is the natural injection.

Remark: (1) The translation subgroup $\ker J$ is trivial in the sense that if $F' = T * F$, where $T \in \ker J$, then F' and F determine equivalent self-dual structures. Therefore the group $\mathrm{SL}(2,C) \otimes [[z^A,t,t^{-1}]]$ is the essential part.

(2) As mentioned previously for any $F \in \mathscr{C}$ whose holomorphic curves have $N_\psi \simeq \mathscr{O}(1) \oplus \mathscr{O}(1)$, we can associate a formal holomorphic section $\sigma : S \to T^\infty M|_S$ by $\sigma = \mathrm{gr}\, F$ which annihilates $\omega(t)$. Composition of canonical transformations corresponds to composition of canonical relations, i.e., the maximal isotropic submanifolds, see Ref. 11. But to each point of S, σ associates a holomorphic curve, thus the composition of canonical relations can be interpreted as a composition of holomorphic curves. The advantage of canonical relations over canonical transformations is that they make sense even when σ is not a graph (as long as a certain transversality condition holds[11]). Canonical relations should be important for constructing a global theory.

(3) The group $\mathscr{C}$ can be thought of as a group of formal curves as follows: The map defined by $\tilde{z}^A = F^A(z^B,t)$ associates to each $t \in U_0 \cap U_\infty$ a formal diffeomorphism of C^2 with Jacobian determinant t^{-2}. We thus have a map $f : U_0 \cap U_\infty \to \text{Form Diff}\, C^2$ whose image is isomorphic to the group of formal volume-preserving diffeomorphisms (or canonical transformations) of C^2 which we denote by G. Thus $C \simeq G \otimes C[[t,t^{-1}]]$, whose formal Lie algebra $g \otimes [[t,t^{-1}]]$ is of the Kac–Moody type, where g is the infinite-dimensional Lie algebra of formal symplectic vector fields (infinitesimal canonical transformations).

(4) The differential of the formal diffeomorphisms F^A, i.e., the map $(t,z^A) \mapsto (\partial F^A/\partial z^B)(z^B,t) \in \mathrm{SL}(2,C)_t$ is the transition function for the bundle $\mathscr{D}^*$. Similarly if $\psi : P^1 \to \mathscr{T}$ is a holomorphic curve then the map $t \to (\partial F^A/\partial z^B)(\psi^B(t),t) \in \mathrm{SL}(2,C)_t$ is the transition function for the normal bundle to $\psi(P^1)$. Nonsingular self-dual structures (i.e., $\dim S = 4$) have $N_\psi \simeq \mathscr{O}(1) \oplus \mathscr{O}(1)$ and this condition is preserved under group composition.

VII. AN EXAMPLE: COLLIDING *pp*-WAVES

Let us illustrate the ideas of the preceding sections by studying a simple but nontrivial example—the complex *pp*-waves.[4,14,15] Consider the subgroup $\mathscr{A}_L$ of $\mathscr{C}$ consisting of all twisted canonical transformations whose Jacobian matrices have the form

$$\frac{\partial F^A}{\partial z^I} = \begin{pmatrix} t^{-1} & 0 \\ F'(z^1,t) & t^{-1} \end{pmatrix}, \tag{21}$$

where F' is an arbitrary function. It is a simple task to compute the holomorphic curves

$$\begin{aligned} \tilde{\psi}^1(t^{-1}) &= t^{-1}\psi^1(t), \\ \tilde{\psi}^2(t^{-1}) &= t^{-1}\psi^2(t) + F(\psi^1(t),t), \end{aligned} \tag{22}$$

where F' is the integral of $\tilde{F}$ with respect to the first argument. The first equation is the projective line

$$\psi^1(t) = q^1 + \tilde{q}^1 t.$$

Substituting this into the second equation and equating powers we can compute the function (using some gauge freedom)

$$\Omega = \epsilon_{AB}\, q^A \tilde{q}^B + H(q^1, \tilde{q}^1), \tag{23}$$

where H is an arbitrary function determined from F. This gives the metric for the complex pp-waves.

The group $\mathscr{A}_L$ is abelian and amounts to addition of functions F in Eqs. (21) and (22), and addition of arbitrary functions H in the function Ω, and hence in the metric. Furthermore, the variables $q^1, \tilde{q}^1$ are the complex analogs of waves traveling, say, from left to right. Thus the abelian subgroup $\mathscr{A}_L$ describes the linear superposition of these waves. They are "noninteracting" plane waves.

Now there is another representation of the pp-waves by the abelian subgroup $\mathscr{A}_U$ of upper triangular matrices of the form

$$\frac{\partial G^A}{\partial z^B} = \begin{pmatrix} t^{-1} & G(z^2,t) \\ 0 & t^{-1} \end{pmatrix}. \tag{24}$$

Similarly one obtains the Ω function (and hence the metric)

$$\Omega = \epsilon_{AB}\, q^A \tilde{q}^B + K(q^2, \tilde{q}^2). \tag{25}$$

This represents plane waves traveling from right to left. Again they are noninteracting and have a linear superposition principle.

However our theorems guarantee that the composition $F * G$ describes another self-dual Einstein space, namely that determined by the holomorphic curve

$$\begin{aligned} \tilde{\psi}^1(t^{-1}) &= t^{-1}\psi^1(t) + G(\psi^2(t), t), \\ \tilde{\psi}^2(t^{-1}) &= t^{-1}\psi^2(t) + tF(\psi^1(t^{-1}), t). \end{aligned} \tag{26}$$

We have not yet studied the detailed structure of these spaces, but the interpretation is clear; they are the spaces obtained by the collision of impinging plane waves. This represents a *nonlinear superposition principle* for nonlinear gravitons.

In order to convince ourselves that we do indeed obtain something new by nonlinear superposition (i.e., not a pp-wave), let us consider the special case where

$$G(z^2,t) = t^{-2}G(z^2), \quad F(\tilde{z}^1,t) = F(\tilde{z}^1).$$

Expanding G and F in a power series and equating coefficients in (26), we can obtain the potential function Θ of Eq. (3) explicitly, namely

$$\Theta = \tfrac{1}{24} G'''(q^2)(p^2)^4 \Phi(\zeta, q^2), \tag{27}$$

where $\zeta = p^1 + \frac{1}{2} G'''(q^2)(p^2)^2$ and Φ satisfies the Hamilton–Jacobi equation

$$\tfrac{1}{2} G''' \Phi_\zeta^2 - \Phi_{q^2} = 0. \tag{28}$$

The metric and curvature coefficients[1,4,5] C_{ABCD} can easily be computed and, in general, depend on two arbitrary functions of one variable each. The second-order curvature invariant,

$$\begin{aligned} I &= C_{ABCD} C^{ABCD} \\ &= 3G'''\Phi_{\zeta\zeta\zeta}\Phi_{\zeta\zeta} + G'''\Phi_{\zeta\zeta\zeta\zeta} + 3(G'')^2(\Phi_{\zeta\zeta\zeta})^2, \end{aligned}$$

is, in general, nonvanishing; hence the resulting space is not a pp-wave. These spaces have a Killing vector field ∂_{q^1} and are special cases of those given in Ref. 1 as solutions to the three-dimensional Laplace equation.

There are several interesting related questions which arise from our results.

(1) What is the most general self-dual space that can be represented by a (possibly infinite) superposition of pp-waves?

(2) More generally, is there a spectral theory for self-dual spaces?

(3) Which real Euclidean signature spaces can be represented as a superposition of pp-waves and what is their singularity structure? Are there any nonsingular ones?

(4) How do gravitational instantons relate to the general theory?

After this work was completed we discovered a recent work[16] where the nonlinear superposition of pp-waves using the nonlinear graviton was discussed. The nonlinear superposition principle treated in Ref. 16 concerns only pp-waves and no group theoretical treatment is given. On the other hand, we have shown the general validity of the nonlinear superposition principle for self-dual Einstein spaces as arising from the underlying group theoretical nature of the nonlinear graviton.

ACKNOWLEDGMENT

We would like to thank J. D. Finley for some stimulating discussions.

[1]C. P. Boyer and J. F. Plebański, J. Math. Phys. **18**, 1022 (1977).

[2]R. Penrose, Gen. Relativ. Gravit. 7, 31 (1976).

[3]V. Arnold, Ann. Inst. Fourier Grenoble **16**, 319 (1966).

[4]J. F. Plebański, J. Math. Phys. **16**, 2395 (1975).

[5]By a conservation law we mean any equation written in the form of a divergence, $\partial_{q^A} Q^{AB_1 \cdots B_k} = 0$.

[6]Equivalently $\tilde{p}_B = -\Omega_{\tilde{q}^B}$ gives rise to a potential $\tilde{\Theta}$.

[7]C. P. Boyer, J. D. Finley, III, and J. F. Plebański, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 2, pp. 241–281.

[8]What follows can readily be generalized to local product structures, see C. P. Boyer, "The Geometry of Complex Self-dual Einstein Spaces," in *Nonlinear Phenomena*, edited by *Lecture Notes in Physics 189*, edited by K. B. Wolf (Springer, New York, 1983).

[9]In the real Euclidean situation M is a Kähler manifold and all of our results will hold if we interpret M_2 as the complex structure on M and $\tilde{M}_2$ as the complex conjugate structure. In this case the coordinates are taken as $\tilde{q}_A = \bar{q}^A$.

[10]K. Yano and S. Ishihara, *Tangent and Cotangent Bundles* (Marcel Dekker, New York, 1973).

[11]For basic notions of finite-dimensional symplectic geometry, see V. Guillemin and S. Sternberg, *Geometric Asymptotics, Mathematical Surveys 14* (A. M. S., Providence, RI, 1977). In the formal infinite-dimensional case treated here maximal isotropic submanifolds are not necessarily Lagrangian. The maximal isotropic submanifolds arising from our theorem have four complex dimensions, whereas Lagrangian submanifolds are infinite dimensional.

[12]R. O. Wells, *Complex Geometry in Mathematical Physics* (Les Presses de l'Universite de Montreal, Montreal, 1982).

[13]It is more common to use $\mathscr{O}(1)$ to denote the sheaf of germs of holomorphic sections of the hyperplane bundle, so we make no distinction between line bundles and their sheaf of sections $\mathscr{O}(n)$.

[14]W. D. Curtis, D. E. Lerner, and F. R. Miller, J. Math. Phys. **19**, 2024 (1978).

[15]R. S. Ward, Proc. R. Soc. London Ser. A **363**, 289 (1978).

[16]J. R. Porter, Gen. Relativ. Gravit. **14**, 1023 (1982).

Elementary representations and intertwining operators for SU(2, 2). I

V. K. Dobrev[a)]
International Centre for Theoretical Physics, Trieste, Italy

(Received 29 December 1983; accepted for publication 28 June 1984)

The structure of the group SU(2, 2) and of its Lie algebra is studied in detail. The results will be applied in subsequent parts devoted to the explicit construction of elementary representations of SU(2, 2) induced from different parabolic subgroups and of the intertwining operators between these representations. A summary of some results of Parts II and III is given.

I. INTRODUCTION

The group $G = \mathrm{SU}(2, 2)$ is of physical interest because it is locally isomorphic to the conformal group of Minkowski space-time, to the group $\mathrm{SO}_e(4, 2)$, and to the group of holomorphic automorphisms of the tube domain over the forward (or backward) light cone (for reviews see Refs. 1 and 2).

With this paper we start the systematic construction and study of the elementary representations of G. The importance of the elementary representations (ER) comes from the fact that every irreducible admissible representation of any semisimple Lie group is equivalent either to an irreducible elementary representation or to an irreducible component of a reducible ER of the group in consideration. The first statement of this type was the fundamental subquotient theorem of Harish-Chandra.[3,4] This result was refined by Lepowski[5] and improved by Casselman's subrepresentation theorem.[6] Combining these results with Langlands classification,[7] Knapp and Zuckermann[8] have formulated the most informative result. Following it we see that the elementary representations are those induced from the cuspidal parabolic subgroups (see Sec. IV for definitions).

There is not much work done on the elementary representations of G. In the mathematical literature we can single out the work by Knapp and Speh[9] which contains many useful facts and gives the complete classification of the irreducible unitary representations of G. However, it does not give explicit construction of the ER and of the intertwining operators between them, and there is no statement on the reducibility of the ER (only a few examples are graphically displayed). These facts are needed in the physical applications along with the facts on unitarity as we know from earlier experience.[10–12] In the mathematical physics literature (see, e.g., Refs. 13–17) representations of G are usually induced from finite-dimensional representations of the only noncuspidal parabolic subgroup P_2 of G, which is isomorphic to the 11-dimensional Weyl subgroup. Another type of induction is from the maximal compact subgroup K of G (cf. Refs. 15 and 18 and references therein). (For the unitary representations of the universal covering group of G see Ref. 19.)

The outline of this work which we now suppose to be in four parts is as follows. Part I (this paper) is devoted to the group G and its Lie algebra. Part II deals with the explicit construction of the elementary representations and the Knapp–Stein integral intertwining operator. Among other things we give a constructive proof that the usually used P_2-induced representations are equivalent to some ER. In fact we prove more. We show that if we are not restricted to finite-dimensional representations of P_2 we build a 1–1 correspondence with P_0-induced ER. This correpondence is given here (Sec. VI D). Part III deals with the reducible ER and the differential intertwining operators between them (see also Secs. VI B and VI C here). Part IV (unlike Parts II and III) is at a preliminary stage. We shall deal there with some questions for which the consideration of the universal covering group of G is essential. In particular, the question of unitary ray representations of G with positive energy outside those induced by the noncuspidal parabolic subgroup[14] shall be studied. We shall also establish there the relation between ER and those induced from the maximal compact subgroup K. We also study the homogeneous space structure of the complex flag manifolds corresponding to the induction from different parabolic subgroups.[20] There we shall come at last to some physical applications.

The organization of this paper, Part I, is as follows. Section II is devoted to the study of the Lie algebra of G (for G we give four different realizations—one of them not used in the literature). We display the Cartan decomposition and the three nonconjugate Cartan subalgebras (one of them is not used usually). Then we have the restricted root system, the Iwasawa decomposition of the Lie algebra, and the restricted Weyl group. The parabolic subalgebras are introduced in Sec. II E. Section III deals with the compactified Lie algebra for which we record some formulas because of the special basis we choose. We introduce the important notion of noncompact roots. Section IV takes up the structure of the group G. We list the important subgroups of G for all parabolics. In Sec. IV B we give explicit matrix representations of the Weyl groups. Section V studies the Iwasawa and the Bruhat decompositions of G.

The Iwasawa decomposition is given in two forms (Sec. V A and the Appendix). The (Gel'fand–Naimark)–Bruhat decomposition requires more care and is studied in detail for the P_0 parabolic subgroup (Sec. V B) and for the nonminimal parabolics (Sec. V C and Ref. 14). An important result is the connection between the Iwasawa and Bruhat decompositions for all parabolics (Sec. V D and the Appendix). Several partial cases and three decompositions of the maximal compact subgroup K (corresponding to the different parabolics)

[a)] On leave of absence from Institute of Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Sofia 1194, Bulgaria.

 0022-2488/85/020235-17$02.50

are given. Section VI is devoted to a summary of some of the results of Parts II and III concerning the intertwining operators and the reducible elementary representations.

II. STRUCTURE OF THE LIE ALGEBRA OF SU(2, 2)

A. Realizations of SU(2, 2)

The standard definition of the group SU(2, 2) is[21]

$$G \equiv \mathrm{SU}(2,2) \equiv \Big\{ g' \in \mathrm{GL}(4,\mathbb{C}) \,|\, g^+ \beta_0 g = \beta_0, \quad \beta_0 \equiv \begin{pmatrix} \mathbf{1}_1 & 0 \\ 0 & -\mathbf{1}_2 \end{pmatrix}, \quad \det g = 1 \Big\}. \tag{2.1}$$

Here G leaves invariant the Hermitian form

$$\phi_0(Z, Z') \equiv Z^+ \beta_0 Z', \quad Z, Z' \in \mathbb{C}^4. \tag{2.2}$$

We shall use also other realizations of G differing from (2.1) by unitary transformations [(2.4) is used in Ref. 14]:

$$\beta_0 \mapsto \beta_1 \equiv U_1 \beta_0 U_1^{-1} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

$$U_1 \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & \sqrt{2} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & \sqrt{2} \end{pmatrix}; \tag{2.3}$$

$$\beta_0 \mapsto \beta_2 \equiv U_2 \beta_0 U_2^{-1} = \begin{pmatrix} 0 & \mathbf{1}_2 \\ \mathbf{1}_2 & 0 \end{pmatrix},$$

$$U_2 \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{1}_2 & \mathbf{1}_2 \\ -\mathbf{1}_2 & \mathbf{1}_2 \end{pmatrix}. \tag{2.4}$$

In the realization with $\beta = \beta_0$ (β_1, β_2) the Cartan subalgebra with zero (one, two, respectively) noncompact generators is diagonal (see below). So each realization is natural for one of the three nonconjugate Cartan subalgebras of the Lie algebra of G.

Another realization is useful when studying the holomorphic representations of G[18,20]:

$$\beta_0 \mapsto \beta_3 \equiv U_3 \beta_0 U_3^{-1} = i \begin{pmatrix} 0 & \mathbf{1}_2 \\ -\mathbf{1}_2 & 0 \end{pmatrix},$$

$$U_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{1}_2 & -i\mathbf{1}_2 \\ -i\mathbf{1}_2 & \mathbf{1}_2 \end{pmatrix}. \tag{2.5}$$

The corresponding Hermitian forms are unchanged under the respective unitary transformations

$$\phi_j(Z, Z') \equiv Z^+ \beta_j Z' \quad (j = 1,2,3), \tag{2.6a}$$

$$\phi_j(U_j Z, U_j Z') = \phi_0(Z, Z'). \tag{2.6b}$$

B. The Lie algebra of G

It is known that the Lie algebra of G consists of all complex 4×4 matrices X satisfying

$$\operatorname{tr} X = 0, \quad X^+ \beta + \beta X = 0 \quad (\beta = \beta_0, \ldots, \beta_3). \tag{2.7}$$

Next is defined the Cartan involution θ

$$\theta X \equiv \beta X \beta^{-1} \tag{2.8}$$

by which we obtain the Cartan decomposition of $\mathfrak{g}$

$$\mathfrak{g} = \mathfrak{k} + \mathfrak{p}. \tag{2.9}$$

Here $\mathfrak{k}$ is the maximal compact subalgebra of $\mathfrak{g}$, and $\mathfrak{p}$ is a vector space so that

$$X \in \mathfrak{k} \Rightarrow \theta X = X, \quad X \in p \Rightarrow \theta X = -X. \tag{2.10}$$

Explicitly we have for the basis of $\mathfrak{g}(\beta = \beta_2)$

$$\frac{i}{2}\begin{pmatrix} 0 & \sigma_a \\ \sigma_2 & 0 \end{pmatrix}, \quad \frac{i}{2}\begin{pmatrix} 0 & e_a \\ e_a & 0 \end{pmatrix}, \quad \frac{i}{2}\begin{pmatrix} \sigma_k & 0 \\ 0 & \sigma_k \end{pmatrix} \quad (a = 1,2, \quad k = 1,2,3), \tag{2.11}$$

$$\frac{i}{2}\begin{pmatrix} 0 & \sigma_a \\ -\sigma_2 & 0 \end{pmatrix}, \quad \frac{i}{2}\begin{pmatrix} 0 & e_a \\ -e_a & 0 \end{pmatrix}, \quad \frac{i}{2}\begin{pmatrix} \sigma_a & 0 \\ 0 & -\sigma_a \end{pmatrix}, \quad \begin{pmatrix} e_a & 0 \\ 0 & -e_a \end{pmatrix}, \tag{2.12}$$

where σ_k are the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$e_1 \equiv \frac{1}{2}(\sigma_0 + \sigma_3) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

$$e_2 \equiv \frac{1}{2}(\sigma_0 - \sigma_1) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_0 = \mathbf{1}_2. \tag{2.13}$$

It is easy to check that the basis elements (2.11) span $\mathfrak{k}$, and (2.12) span $\mathfrak{p}$.

Next we single out important subalgebras of $\mathfrak{g}$. Let $\mathfrak{a}$ be the subspace of $\mathfrak{p}$, which is maximal subject to the condition $[X, X'] = 0$ if $X, X' \in \mathfrak{g}$. The split rank e of $\mathfrak{g}$ is defined to be $\dim \mathfrak{a}$ which equals 2 in our case. We choose for the basis of $\mathfrak{a}$ $(\beta = \beta_2)$

$$\hat{e}_a \equiv \begin{pmatrix} e_a & 0 \\ 0 & -e_a \end{pmatrix} \quad (a = 1,2). \tag{2.14}$$

(In the notation of Ref. 9, $\mathfrak{a} = \mathfrak{a}_{\min}$.) The centralizer $\mathfrak{m}_0$ of $\mathfrak{a}_0$ in $\mathfrak{k}$ is spanned by the generator

$$H = \frac{i}{2}\begin{pmatrix} \sigma_3 & 0 \\ 0 & \sigma_3 \end{pmatrix}. \tag{2.15}$$

The Cartan subalgebra $\mathfrak{h}_2$ consisting of all diagonal matrices in $\mathfrak{g}$ for $\beta = \beta_2$ is spanned by $\hat{e}_1$, $\hat{e}_2$, and H. It is the most noncompact Cartan subalgebra and is usually displayed.[14,9] The other noncompact nonconjugate Cartan subalgebra $\mathfrak{h}_1$ is diagonal for $\beta = \beta_1$ and is spanned by

$$\hat{e}_1, \ \operatorname{diag}(i/2, -i, i/2, 0), \ \operatorname{diag}(i/2, 0, i/2, -i). \tag{2.16}$$

The most commonly used compact Cartan subalgebra $\mathfrak{h}_0$ is diagonal for $\beta = \beta_0$ and is spanned by[14]

$$H_0 = \frac{i}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad H_1 = \frac{i}{2}\begin{pmatrix} \sigma_3 & 0 \\ 0 & 0 \end{pmatrix}, \quad H_2 = \frac{i}{2}\begin{pmatrix} 0 & 0 \\ 0 & \sigma_3 \end{pmatrix}. \tag{2.17}$$

The algebra $\mathfrak{g}$ is of course isomorphic to so(4, 2), the isomorphism given explicitly by $(\beta = \beta_2)$:

$$X_{jk} = -\epsilon_{jkl} \frac{i}{2}\begin{pmatrix} \sigma_l & 0 \\ 0 & \sigma_l \end{pmatrix}, \quad X_{j5} = \frac{i}{2}\begin{pmatrix} 0 & \sigma_j \\ \sigma_j & 0 \end{pmatrix},$$

$$X_{06} = \frac{i}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{2.18a}$$

$$X_{k0}=\frac{1}{2}\begin{pmatrix}\sigma_k & 0\\ 0 & -\sigma_k\end{pmatrix},\quad X_{56}=\frac{1}{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix},$$
$$X_{j6}=\frac{i}{2}\begin{pmatrix}0 & \sigma_j\\ -\sigma_j & 0\end{pmatrix},\quad X_{05}=\frac{i}{2}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}. \tag{2.18b}$$

Indeed

$$[X_{AB},X_{CD}]=\eta_{AC}X_{BD}+\eta_{BD}X_{AC}-\eta_{AD}X_{BC}-\eta_{BC}X_{AD}, \tag{2.18c}$$

where $A,B,C,D=0,1,2,3,5,6$; $\eta_{11}=\cdots=\eta_{55}=-\eta_{00}=-\eta_{66}=1$, $\eta_{AB}=0$, for $A\neq B$.

Note that the subalgebra $\mathfrak{k}$ [spanned by (2.18a)] is isomorphic to $so(4)\oplus so(2)$ and $\mathfrak{p}$ is spanned by (2.18b).

C. The restricted root system and the Iwasawa decomposition

To construct the Iwasawa decomposition[22] of $\mathfrak{g}$ we use the restricted root system of $\mathfrak{g}$ relative to $\mathfrak{a}_0$. Let $\mathfrak{a}_0^*$ be the space of linear functionals over $\mathfrak{a}_0$. They are determined by their values on $\hat{e}_a$. We define for $\lambda\in\mathfrak{a}_0^*$, $\lambda\neq 0$,

$$\mathfrak{g}_\lambda\equiv\{X\in\mathfrak{g}\,|\,[\hat{e}_a,X]=\lambda(\hat{e}_a)X\}, \tag{2.19a}$$
$$\Lambda\equiv\{\lambda\in\mathfrak{a}_0^*\,|\,\lambda\neq 0,\ \mathfrak{g}_\lambda\neq\{0\}\}. \tag{2.19b}$$

It is easily obtained that

$$\Lambda=\{\pm\lambda_k,\ k=1,2,3,4\}, \tag{2.20a}$$

where the set Λ^+ of positive roots is chosen to be[9]

$$\lambda_1(\hat{e}_1,\hat{e}_2)=(2,0),\quad \lambda_2(\hat{e}_1,\hat{e}_2)=(1,1),$$
$$\lambda_3(\hat{e}_1,\hat{e}_2)=(0,2),\quad \lambda_4(\hat{e}_1,\hat{e}_2)=(1,-1),$$
$$\lambda_1=\lambda_3+2\lambda_4,\quad \lambda_2=\lambda_3+\lambda_4 \tag{2.20b}$$

($\pm\lambda_2$, $\pm\lambda_4$ have multiplicity 2) and the simple roots with this ordering are λ_3 and λ_4. We display the corresponding root spaces basis vectors [denoting $\mathfrak{g}_k^\pm\equiv\mathfrak{g}_{\pm\lambda_k}$, $\sigma_\pm\equiv\frac{1}{2}(\sigma_1\pm i\sigma_2)$] for $\beta=\beta_2$:

$$\mathfrak{g}_1^+=\text{l.s.}\begin{pmatrix}0 & ie_1\\ 0 & 0\end{pmatrix},\quad \mathfrak{g}_1^-=\text{l.s.}\begin{pmatrix}0 & 0\\ ie_1 & 0\end{pmatrix};$$
$$\mathfrak{g}_2^+=\text{l.s.}\left\{\begin{pmatrix}0 & i\sigma_a\\ 0 & 0\end{pmatrix}\right\},$$
$$\mathfrak{g}_2^-=\text{l.s.}\left\{\begin{pmatrix}0 & 0\\ i\sigma_a & 0\end{pmatrix}\right\}\quad (a=1,2);$$
$$\mathfrak{g}_3^+=\text{l.s.}\begin{pmatrix}0 & ie_2\\ 0 & 0\end{pmatrix},\quad \mathfrak{g}_3^-=\text{l.s.}\begin{pmatrix}0 & 0\\ ie_2 & 0\end{pmatrix};$$
$$\mathfrak{g}_4^+=\text{l.s.}\left\{\begin{pmatrix}\sigma_+ & 0\\ 0 & -\sigma_-\end{pmatrix},\ i\begin{pmatrix}\sigma_+ & 0\\ 0 & \sigma_-\end{pmatrix}\right\},$$
$$\mathfrak{g}_4^-=\text{l.s.}\left\{\begin{pmatrix}-\sigma_- & 0\\ 0 & \sigma_+\end{pmatrix},\ i\begin{pmatrix}\sigma_- & 0\\ 0 & \sigma_+\end{pmatrix}\right\}, \tag{2.21}$$

where l.s. stands for the linear span.

Using standard notation we introduce the positive and negative root spaces

$$\tilde{\mathfrak{n}}_0\equiv\oplus_k\mathfrak{g}_k^+,\quad \mathfrak{n}_0\equiv\oplus_k\mathfrak{g}_k^-. \tag{2.22}$$

Obviously $\tilde{\mathfrak{n}}_0=\theta\mathfrak{n}_0$ and we can write the decomposition (valid generally for semisimple Lie algebras)

$$\mathfrak{g}=\tilde{\mathfrak{n}}_0\oplus\mathfrak{n}_0\oplus\mathfrak{g}_0\oplus\mathfrak{m}_0. \tag{2.23}$$

We also note that the map

$$J:\mathfrak{m}_0\oplus\tilde{\mathfrak{n}}_0\to\mathfrak{k},$$
$$J(X+X')=X+X'+\theta X'\quad (X\in\mathfrak{m}_0,\ X'\in\tilde{\mathfrak{n}}_0) \tag{2.24}$$

is bijective and that

$$\mathfrak{g}=\mathfrak{k}\oplus\mathfrak{g}_0\oplus\mathfrak{n}_0 \tag{2.25}$$

is the Iwasawa decomposition of $\mathfrak{g}$.

D. The restricted Weyl group $W(\mathfrak{g},\mathfrak{a}_0)$

For future reference we define for every $\lambda_k\in\Lambda^+$ a vector $\hat{H}_k\in\mathfrak{a}_0$ by

$$B(\hat{H}_k,\hat{e}_a)=\lambda_k(\hat{e}_a)\quad (a=1,2),$$
$$B(X,Y)\equiv\operatorname{tr}XY,\quad X,Y\in\mathfrak{g}, \tag{2.26}$$

where B is the Killing form on $\mathfrak{g}$. It is easily seen that

$$\hat{H}_3=\hat{e}_2,\quad \hat{H}_4=\tfrac{1}{2}(\hat{e}_1-\hat{e}_2),\quad \hat{H}_1=\hat{H}_3+2\hat{H}_4=\hat{e}_1,$$
$$\hat{H}_2=\hat{H}_3+\hat{H}_4=\tfrac{1}{2}(\hat{e}_1+\hat{e}_2). \tag{2.27}$$

We also introduce the restricted Weyl reflections s_k in $\mathfrak{a}_0$ standardly by

$$s_k(\hat{e}_a)\equiv\hat{e}_a-2(\lambda_k(\hat{e}_a)/\lambda_k(\hat{H}_k))\hat{H}_k, \tag{2.28}$$

which explicitly take the form (note $s_k^2=\text{id}$)

$$s_1(\hat{e}_1,\hat{e}_2)=(-\hat{e}_1,\hat{e}_2),\quad s_2(\hat{e}_1,\hat{e}_2)=-(\hat{e}_2,\hat{e}_1), \tag{2.29}$$
$$s_3(\hat{e}_1,\hat{e}_2)=(\hat{e}_1,-\hat{e}_2),\quad s_4(\hat{e}_1,\hat{e}_2)=(\hat{e}_2,\hat{e}_1). \tag{2.30}$$

It is well known that the restricted Weyl reflections generate the finite restricted Weyl group $W(\mathfrak{g},\mathfrak{a}_0)$

$$W(\mathfrak{g},\mathfrak{a}_0)=\{\text{id},s_1,\dots,s_7\}, \tag{2.31a}$$

where

$$s_5\equiv s_2s_1,\quad s_6\equiv s_1s_2,\quad s_7\equiv s_5^2=s_6^2,$$
$$s_7(\hat{e}_1,\hat{e}_2)=-(\hat{e}_1,\hat{e}_2)=\theta(\hat{e}_1,\hat{e}_2), \tag{2.31b}$$

and we have chosen s_1,s_2 as the generating elements and then $s_3=s_2s_1s_2$, $s_4=s_1s_2s_1$. (The other possible choice is s_3,s_4.) We also define the induced action on the roots by the formula

$$s_k^*\lambda_j\equiv\lambda_j\circ s_k, \tag{2.32a}$$
$$s_1^*(\lambda_1,\lambda_2,\lambda_3,\lambda_4)=(-\lambda_1,-\lambda_4,\lambda_3,-\lambda_2),$$
$$s_2^*(\lambda_1,\lambda_2,\lambda_3,\lambda_4)=(-\lambda_3,-\lambda_2,-\lambda_1,\lambda_4), \tag{2.32b}$$

and we have displayed as examples the action of the generating elements.

E. The parabolic subalgebras

In the above constructions

$$\mathfrak{p}_0\equiv\mathfrak{m}_0\oplus\mathfrak{g}_0\oplus\mathfrak{n}_0 \tag{2.33}$$

is a *minimal parabolic subalgebra* of $\mathfrak{g}$. A *standard parabolic subalgebra* is any subalgebra of $\mathfrak{g}$ containing $\mathfrak{p}_0$. It is known that the number of standard parabolic subalgebras is 2^l ($l\equiv\dim\mathfrak{g}_0$) which equals 4 here. One of the other three is $\mathfrak{g}$ itself. The remaining two are also given by the form

$$\mathfrak{p}_a=\mathfrak{m}_a\oplus\mathfrak{a}_a\oplus\mathfrak{n}_a\quad (a=1,2) \tag{2.34}$$

and are characterized as follows:

$$s_aX=-X,\quad X\in\mathfrak{a}_a\subset\mathfrak{a}_0, \tag{2.35a}$$

from which we see that $\mathfrak{a}_1,\mathfrak{a}_2$ are spanned by

$$\mathfrak{a}_1 = \text{l.s.}\ \hat{e}_1, \quad \mathfrak{a}_2 = \text{l.s.}(\hat{e}_1 + \hat{e}_2). \tag{2.35b}$$

Then $\mathfrak{m}_a \supset \mathfrak{m}_0$ is the centralizer of $\mathfrak{a}_a$ in $\mathfrak{g}$ explicitly given by [cf. (2.15) and (2.21)]

$$\mathfrak{m}_1 = \text{l.s.}\left\{\hat{e}_2, H, \begin{pmatrix} 0 & i\hat{e}_2 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ i\hat{e}_2 & 0 \end{pmatrix}\right\} \supset \mathfrak{g}_3^+ \oplus \mathfrak{g}_3^-, \tag{2.36a}$$

$$\mathfrak{m}_2 = \text{l.s.}\left\{\hat{e}_1 - \hat{e}_2, H, \frac{i}{2}\begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_a \end{pmatrix}, \begin{pmatrix} \sigma_a & 0 \\ 0 & -\sigma_a \end{pmatrix}\right\} \supset \mathfrak{g}_4^+ \oplus \mathfrak{g}_4^-. \tag{2.36b}$$

It is easy to see (as noted in Ref. 9) that

$$\mathfrak{m}_1 \cong \mathfrak{m}_0 \oplus \text{sl}(2, \mathbb{R}), \tag{2.37a}$$

$$\mathfrak{m}_2 \cong \text{sl}(2, \mathbb{C}). \tag{2.37b}$$

For each $\mathfrak{a}_a$ we define the roots of $(\mathfrak{g}, \mathfrak{a}_a)$ to be the nonzero restrictions to $\mathfrak{a}_a$ of the restricted roots. Explicitly the roots Λ_a of $(\mathfrak{g}, \mathfrak{a}_a)$ are

$$\Lambda_1 = \{\pm\lambda, \pm 2\lambda\}, \quad \lambda = \lambda_2|_{\mathfrak{a}_1} = \lambda_4|_{\mathfrak{a}_1}, \quad 2\lambda = \lambda_1|_{\mathfrak{a}_1}; \tag{2.38a}$$

$$\Lambda_2 = \{\pm\lambda'\}, \quad \lambda' = \lambda_1|_{\mathfrak{a}_2} = \lambda_2|_{\mathfrak{a}_2} = \lambda_3|_{\mathfrak{a}_2}. \tag{2.38b}$$

Obviously $\pm\lambda$, $\pm\lambda'$ have multiplicity 4, while 2λ has multiplicity 1. Defining as usual $\tilde{\mathfrak{n}}_a, \tilde{\mathfrak{n}}_a$ to be the positive and negative (resp.) root spaces we obtain [cf. (2.21)]

$$\tilde{\mathfrak{n}}_1 = \mathfrak{g}_1^+ \oplus \mathfrak{g}_2^+ \oplus \mathfrak{g}_4^+ \quad (\dim \tilde{\mathfrak{n}}_1 = 5),$$

$$\mathfrak{n}_1 = \mathfrak{g}_1^- \oplus \mathfrak{g}_2^- \oplus \mathfrak{g}_4^- = s_1 \tilde{\mathfrak{n}}_1, \tag{2.39a}$$

$$\tilde{\mathfrak{n}}_2, = \mathfrak{g}_1^+ \oplus \mathfrak{g}_2^+ \oplus \mathfrak{g}_3^+ \quad (\dim \tilde{\mathfrak{n}}_2 = 4),$$

$$\tilde{\mathfrak{n}}_2 = \mathfrak{g}_1^- \oplus \mathfrak{g}_2^- \oplus \mathfrak{g}_3^- = s_2 \tilde{\mathfrak{n}}_2. \tag{2.39b}$$

Thus we have given explicitly the factors in (2.34). Also we note two parallels of (2.23)

$$\mathfrak{g} = \tilde{\mathfrak{n}}_a \oplus \mathfrak{n}_a \oplus \mathfrak{a}_a \oplus \mathfrak{m}_0 = \tilde{\mathfrak{n}}_a \oplus \mathfrak{p}_a \quad (a = 1,2) \tag{2.40}$$

and of (2.24)

$$J_a: \mathfrak{m}_a^k \oplus \tilde{\mathfrak{n}}_a \to \mathfrak{k}, \quad \mathfrak{m}_a^k \equiv \mathfrak{k} \wedge \mathfrak{n}_a, \tag{2.41a}$$

$$J_a(X + X') = X + X' + s_a X' \quad (X \in \mathfrak{m}_a^k, X' \in \mathfrak{n}_a), \tag{2.41b}$$

$$\mathfrak{m}_1^k = \text{l.s.}\left\{H, \frac{i}{2}\begin{pmatrix} 0 & \hat{e}_1 \\ \hat{e}_2 & 0 \end{pmatrix}\right\}, \tag{2.42a}$$

$$\mathfrak{m}_2^k = \text{l.s.}\left\{\frac{i}{2}\begin{pmatrix} \sigma_j & 0 \\ 0 & \sigma_j \end{pmatrix}; \ j = 1,2,3\right\} \cong \text{SU}(2). \tag{2.42b}$$

III. THE COMPLEXIFIED LIE ALGEBRA

A. The root system of the complexified Lie algebra

Let $\mathfrak{g}^{\mathbb{C}}$ be the complexification of $\mathfrak{g}$, and let $\mathfrak{h}^{\mathbb{C}}$ be its Cartan subalgebra. Since $\mathfrak{g}^{\mathbb{C}} \cong \text{sl}(4, \mathbb{C})$ is complex, $\mathfrak{h}^{\mathbb{C}}$ is unique (up to conjugacy) and is the complexification of any of the Cartan subalgebras of $\mathfrak{g}$ displayed above. It is useful to choose the basis in $\mathfrak{h}^{\mathbb{C}}$ so that the roots of the pair $(\mathfrak{g}^{\mathbb{C}}, \mathfrak{h}^{\mathbb{C}})$ have real values on the subspaces $\mathfrak{g} \oplus i\mathfrak{b}$, where $\mathfrak{b}$ is a Cartan subalgebra of $\mathfrak{m}$. Also ordering of roots must be compatible with their restriction on $\mathfrak{g}$. For these reasons we do not use the standard basis in $\mathfrak{h}^{\mathbb{C}}$ consisting of

$$\begin{pmatrix} \sigma_3 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} \hat{e}_2 & 0 \\ 0 & -\hat{e}_1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & \sigma_3 \end{pmatrix}, \tag{3.1}$$

but rather the basis comprised of $\hat{e}_a$ and $\hat{H} \equiv -iH$. Then the root system is

$$\Delta = \{\pm\alpha_k, k = 1,...,6\}; \tag{3.2}$$

$$\begin{aligned} &\alpha_1(\hat{e}_1, \hat{e}_2, \hat{H}) = (2, 0, 0), \quad \alpha_2(\hat{e}_1, \hat{e}_2, \hat{H}) = (1, 1, 1), \\ &\alpha_3(\hat{e}_1, \hat{e}_2, \hat{H}) = (0, 2, 0), \\ &\alpha_4(\hat{e}_1, \hat{e}_2, \hat{H}) = (1, -1, 1), \\ &\alpha_5(\hat{e}_1, \hat{e}_2, \hat{H}) = (1, 1, -1), \\ &\alpha_6(\hat{e}_1, \hat{e}_2, \hat{H}) = (1, -1, -1); \end{aligned} \tag{3.3a}$$

$$\alpha_1 = \alpha_3 + \alpha_4 + \alpha_6, \quad \alpha_2 = \alpha_3 + \alpha_4, \quad \alpha_5 = \alpha_3 + \alpha_6. \tag{3.3b}$$

The corresponding root spaces $\mathfrak{g}_k^{\pm} (\equiv \mathfrak{g}_{\pm\alpha_k}^{\mathbb{C}})$ are (complexly) spanned by the root vectors $X_k^{\pm}$

$$\begin{aligned} &X_1^+ = \begin{pmatrix} 0 & i\hat{e}_1 \\ 0 & 0 \end{pmatrix}, \quad X_2^+ = \begin{pmatrix} 0 & i\sigma_+ \\ 0 & 0 \end{pmatrix}, \\ &X_3^+ = \begin{pmatrix} 0 & i\hat{e}_2 \\ 0 & 0 \end{pmatrix}, \quad X_4^+ = \begin{pmatrix} \sigma_+ & 0 \\ 0 & 0 \end{pmatrix}, \\ &X_5^+ = \begin{pmatrix} 0 & i\sigma \\ 0 & 0 \end{pmatrix}, \quad X_6^+ = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_- \end{pmatrix}, \\ &X_k^- = (X_k^+)^\dagger, \end{aligned} \tag{3.4}$$

and the normalization is chosen so that $\alpha_k(Z_k) = 2$, where

$$Z_k \equiv [X_k^+, X_k^-]. \tag{3.5}$$

Explicitly [cf. (2.27), (2.17)],

$$\begin{aligned} &Z_1 = \hat{e}_1 = \hat{H}, \\ &Z_2 = \frac{1}{2}(\hat{e}_1 + \hat{e}_2) + \hat{H} = \hat{H}_2 + \hat{H} = \begin{pmatrix} \hat{e}_1 & 0 \\ 0 & -\hat{e}_2 \end{pmatrix}, \\ &Z_3 = \hat{e}_2 = \hat{H}_3, \\ &Z_4 = \frac{1}{2}(\hat{e}_1 - \hat{e}_2) + \hat{H} = \hat{H}_4 + \hat{H} = \begin{pmatrix} \sigma_3 & 0 \\ 0 & 0 \end{pmatrix} = -2iH_1, \\ &Z_5 = \frac{1}{2}(\hat{e}_1 + \hat{e}_2) - \hat{H} = \hat{H}_2 - \hat{H} = \begin{pmatrix} \hat{e}_2 & 0 \\ 0 & -\hat{e}_1 \end{pmatrix}, \\ &Z_6 = \frac{1}{2}(\hat{e}_1 - \hat{e}_2) - \hat{H} = \hat{H}_4 + \hat{H} = \begin{pmatrix} 0 & 0 \\ 0 & -\sigma_3 \end{pmatrix} = 2iH_2. \end{aligned} \tag{3.6}$$

Equivalently, Z_k is defined so that $B(Z_k, X) = \alpha_k(x)$, $(X = \hat{e}_a, \hat{H})$. We define also the Weyl reflections

$$w_k(X) \equiv X - 2\frac{\alpha_k(X)}{\alpha_k(Z_k)} Z_k = X - \alpha_k(X) Z_k \quad (X = \hat{e}_a, \hat{H}), \tag{3.7a}$$

with the explicit actions given by

$$\begin{aligned} &w_1(\hat{e}_1, \hat{e}_2, \hat{H}) = (-\hat{e}_1, \hat{e}_2, \hat{H}), \\ &w_2(\hat{e}_1, \hat{e}_2, \hat{H}) = (Z_6, -Z_4, -\tfrac{1}{2}(\hat{e}_1 + \hat{e}_2)), \\ &w_3(\hat{e}_1, \hat{e}_2, \hat{H}) = (\hat{e}_1, -\hat{e}_2, \hat{H}), \\ &w_4(\hat{e}_1, \hat{e}_2, \hat{H}) = (Z_5, Z_2, -\tfrac{1}{2}(\hat{e}_1 - \hat{e}_2)), \\ &w_5(\hat{e}_1, \hat{e}_2, \hat{H}) = (Z_4, -Z_6, \tfrac{1}{2}(\hat{e}_1 + \hat{e}_2)), \\ &w_6(\hat{e}_1, \hat{e}_2, \hat{H}) = (Z_2, Z_5, \tfrac{1}{2}(\hat{e}_1 - \hat{e}_2)), \\ &w_k(Z_k) = -Z_k. \end{aligned} \tag{3.7b}$$

The induced action is naturally defined as

$$w_k^* \alpha_j \equiv \alpha_j \circ w_k, \tag{3.8}$$

which is explicitly given by

$$\begin{aligned}
&w_1^*(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)\\
&\quad = (-\alpha_1, -\alpha_6, \alpha_3, -\alpha_5, -\alpha_4, -\alpha_2),\\
&w_2^*(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)\\
&\quad = (\alpha_6, -\alpha_2, -\alpha_4, -\alpha_3, \alpha_5, \alpha_1),\\
&w_3^*(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)\\
&\quad = (\alpha_1, \alpha_4, -\alpha_3, \alpha_2, \alpha_6, \alpha_5),\\
&w_4^*(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)\\
&\quad = (\alpha_5, \alpha_3, \alpha_2, -\alpha_4, \alpha_1, \alpha_6),\\
&w_5^*(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)\\
&\quad = (\alpha_4, \alpha_2, -\alpha_6, \alpha_1, -\alpha_5, -\alpha_3),\\
&w_6^*(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)\\
&\quad = (\alpha_2, \alpha_1, \alpha_5, \alpha_4, \alpha_3, -\alpha_6).
\end{aligned} \tag{3.9}$$

We note from (2.9)

$$w_k^* \alpha_k = -\alpha_k,$$

$$w_2^* = w_3^* w_4^* w_3^* = w_4^* w_3^* w_4^*, \tag{3.10}$$

$$w_5^* = w_3^* w_6^* w_3^* = w_6^* w_3^* w_6^*,$$

$$w_1^* = w_4^* w_5^* w_4^* = w_6^* w_2^* w_6^*. \tag{3.11}$$

We have chosen the generating elements of the Weyl group $W(\mathfrak{g}^{\mathbb{C}}, \mathfrak{h}^{\mathbb{C}})$ as w_3, w_4, w_6 corresponding to the simple roots [cf. (3.3b)]. We do not give explicitly $W(\mathfrak{g}^{\mathbb{C}}, \mathfrak{h}^{\mathbb{C}})$ since we shall need only the action of $w_k (k = 1,...,6)$ given in (3.9)

B. Compact and noncompact roots

In the previous subsections matters have been arranged so that the root systems (2.20) and (3.3) are compatible

$$\lambda_1 = \alpha_1|_{\mathfrak{a}_0}, \quad \lambda_2 = \alpha_2|_{\mathfrak{a}} = \alpha_5|_{\mathfrak{a}},$$

$$\lambda_2 = \alpha_3|_{\mathfrak{a}_0}, \quad \lambda_4 = \alpha_4|_{\mathfrak{a}} = \alpha_6|_{\mathfrak{a}}. \tag{3.12}$$

We introduce notation for the simple roots of Δ and Λ [cf. (3.3b) and (2.20b)]

$$\Delta_S \equiv \{\alpha_3, \alpha_4, \alpha_6\}, \quad \Lambda_S \equiv \{\lambda_3, \lambda_4\} = \Delta_S|_{\mathfrak{a}}. \tag{3.13}$$

Next we define the set of *compact roots* Δ_a *with respect to the parabolic subalgebra* $\mathfrak{p}_a$ by

$$\Delta_a \equiv \{\alpha \in \Delta \,|\alpha|_{\mathfrak{a}} = 0\} \quad (a = 0,1,2). \tag{3.14}$$

Explicitly, we have

$$\Delta_0 = \varnothing, \quad \Delta_1 = \{\pm\alpha_3\}, \quad \Delta_2 = \{\pm\alpha_4, \pm\alpha_6\}. \tag{3.15}$$

This notion was introduced first in Ref. 23 for the case of the minimal parabolic subalgebra. The extension of the definition is justified because it is natural for Δ_a to be the root systems of $(\mathfrak{m}_a^{\mathbb{C}}, \mathfrak{b}_a^{\mathbb{C}})$, where $\mathfrak{m}_a^{\mathbb{C}}$ is the complexification of $\mathfrak{m}_a$ (2.36), $\mathfrak{b}_a^{\mathbb{C}}$ is the complexified Cartan subalgebra of $\mathfrak{m}_a^{\mathbb{C}}$ given by

$$\mathfrak{b}_1^{\mathbb{C}} = \text{c.l.s.}\{\hat{e}_2, \hat{H}\}, \tag{3.16a}$$

$$\mathfrak{b}_2^{\mathbb{C}} = \text{c.l.s.}\{\hat{e}_1 - \hat{e}_2, \hat{H}\}, \tag{3.16b}$$

where c.l.s. stands for complex linear span. For this reason in the case of the parabolic subalgebra being $\mathfrak{g}$ itself the relevant notion $\Delta_{\mathfrak{g}}$ is of course

$$\Delta_{\mathfrak{g}} = \Delta. \tag{3.17}$$

Then we define the *noncompact roots* Δ_a^n *with respect to the parabolic subalgebra* $\mathfrak{p}_a$ by

$$\Delta_a^n \equiv \Delta \setminus \Delta_a, \tag{3.18}$$

obtaining

$$\Delta_0^n = \Delta, \quad \Delta_1^n = \{\pm\alpha_k, k = 1,2,4,5,6\}, \tag{3.19}$$

$$\Delta_2^n = \{\pm\alpha_k, k = 1,2,3,5\}, \quad \Delta_{\mathfrak{g}}^k = \varnothing.$$

Analogously to (2.22), (2.39) in the case of the real algebra we define noncompact positive and negative root spaces [cf. (3.4)]

$$\tilde{\mathfrak{n}}_a^{\mathbb{C}} \equiv \bigoplus_{\alpha_k \in \Delta_a^n} \tilde{\mathfrak{g}}_k^+, \quad \mathfrak{n}_a^{\mathbb{C}} \equiv \bigoplus_{\alpha_k \in \Delta_a^n} \tilde{\mathfrak{g}}_k^-, \tag{3.20}$$

$$\tilde{\mathfrak{n}}_0^{\mathbb{C}} = \bigoplus_k \tilde{\mathfrak{g}}_k^+, \quad \tilde{\mathfrak{n}}_1^{\mathbb{C}} = \bigoplus_{k \neq 3} \mathfrak{g}_k^+,$$

$$\tilde{\mathfrak{n}}_2^{\mathbb{C}} = \bigoplus_{k \neq 4,6} \tilde{\mathfrak{g}}_k^+, \quad \tilde{\mathfrak{n}}_{\mathfrak{g}}^{\mathbb{C}} = \{0\}. \tag{3.21}$$

The complex parallels of (2.23), (2.40) are

$$\mathfrak{g}^{\mathbb{C}} = \tilde{\mathfrak{n}}_a^{\mathbb{C}} \oplus \mathfrak{n}_a^{\mathbb{C}} \oplus \mathfrak{g}_a^{\mathbb{C}} \oplus \mathfrak{m}_a^{\mathbb{C}} \quad (a = 0,1,2), \tag{3.22}$$

which in the case of the minimal parabolic $(a = 0)$ takes the well-known form

$$\mathfrak{g}^{\mathbb{C}} = \tilde{\mathfrak{n}}_0^{\mathbb{C}} \oplus \tilde{\mathfrak{n}}_0^{\mathbb{C}} \oplus \mathfrak{h}^{\mathbb{C}} \quad (\mathfrak{h}^{\mathbb{C}} = \mathfrak{g}_0^{\mathbb{C}} \oplus \mathfrak{m}_0^{\mathbb{C}}). \tag{3.23}$$

IV. THE STRUCTURE OF THE GROUP G

A. Important subgroups of G

We shall usually write the elements of G as $g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$, where $\alpha, \beta, \gamma, \delta$ are 2×2 complex matrices constrained by defining conditions (2.1). Explicitly, for $\beta = \beta_2$ [cf. (2.4)] we have

$$\alpha^+\gamma + \gamma^+\alpha = 0, \quad \beta^+\delta + \delta^+\beta = 0, \quad \alpha^+\delta + \gamma^+\beta = \mathbf{1}_2. \tag{4.1}$$

The maximal compact subgroup K of G is given usually[21] by $(\beta = \beta_0)$,

$$K \equiv \{g \in G \,|g^+ = g^{-1}\} = \left\{\begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}, \ k_i \in \mathrm{U}(2), \ \det k_1 k_2 = 1\right\},$$

$$K \cong \mathrm{S}(\mathrm{U}(2)\times\mathrm{U}(2)) \cong \mathrm{U}(1)\times\mathrm{SU}(2)\times\mathrm{SU}(2), \tag{4.2}$$

and its Lie algebra is $\mathfrak{k}$. Let us define (below in this section $\beta = \beta_2$)

$$A_0 \equiv \exp \mathfrak{a}_0 = \left\{a = \begin{pmatrix} \hat{a} & 0 \\ 0 & \hat{a}^{-1} \end{pmatrix}, \ \hat{a} = e_1 e^s + e_2 e^t, s, t \in \mathbb{R}\right\}, \tag{4.3a}$$

$$\tilde{N}_0 \equiv \exp \tilde{\mathfrak{n}}_0 = \left\{\tilde{n} = \exp\begin{pmatrix} z\sigma_+ & i x \\ 0 & -\bar{z}\sigma_- \end{pmatrix}, \right.$$

$$\left. z \in \mathbb{C}, x \equiv x_0 \mathbf{1}_2 + \sum_{k=1}^{3} x_k \sigma_k, x_0, x_k \in \mathbb{R}\right\}$$

$$= \left\{\begin{pmatrix} x_z & iX \\ 0 & x_z^{+\,-1} \end{pmatrix}, x_z \equiv 1 + z\sigma_+, \right.$$

$$\left. X \equiv x + \frac{1}{2}(z\sigma_+ x - \bar{z} x \sigma_-) - \frac{|z|^2}{6}\sigma_+ x \sigma_-\right\}; \tag{4.3b}$$

$$N_0 \equiv \exp \mathfrak{n}_0 = \left\{n = \exp\begin{pmatrix} -\bar{w}\sigma_- & 0 \\ i\bar{b} & w\sigma_+ \end{pmatrix}, \right.$$

$$w\in \mathbb{C},\ \tilde{b}\equiv b_0 1_2 - \sum_k b_k\sigma_k, b_0,\ b_k\in\mathbb{R}\Big\}$$

$$= \left\{\begin{pmatrix} b_w & 0 \\ i\tilde{B} & b_w^{+\,-1}\end{pmatrix},\quad b_w\equiv 1-\bar{w}\sigma_-,\right.$$

$$\left.\tilde{B}\equiv\tilde{b}+\frac{1}{2}(w\sigma_+\tilde{b}-\bar{w}\tilde{b}\sigma_-)-\frac{|w|^2}{6}\sigma_+\tilde{b}\sigma_-\right\}. \tag{4.3c}$$

Note that

$$x_-\equiv x_0-x_3=\mathrm{tr}\, \mathfrak{X}e_2,\quad u\equiv x_2+ix_1=i\,\mathrm{tr}\,\mathfrak{X}\sigma_-,$$
$$x_+\equiv x_0+x_3=\ \mathrm{tr}\,\mathfrak{X}e_1-(i/2)(z\bar{u}+\bar{z}u)+x_-|z|^2/6;$$
$$b_+\equiv b_0+b_3=\mathrm{tr}\,\tilde{B}\hat{e}_2,\quad v\equiv b_2+ib_1=\ -i\,\mathrm{tr}\,\tilde{B}\sigma_-,$$
$$b_-\equiv b_0-b_3=\mathrm{tr}\,\tilde{B}\hat{e}_1+(i/2)(\bar{w}v+w\bar{v})+(|w|^2/6)b_+. \tag{4.3d}$$

Let M_0 be the centralizer of A_0 in K. Then[9]

$$M_0=\{i^k\tau|k=0,1,2,3;\ \tau\in T\}$$
$$=\{m=\tau(\theta)\gamma_3^N,\ \tau\in T, N=0,1\}=T\oplus\{1,\gamma_3\},$$
$$T\equiv\left\{\tau(\theta)=\begin{pmatrix}\hat{\tau}(\theta) & 0\\ 0 & \hat{\tau}(\theta)\end{pmatrix}, \hat{\tau}(\theta)=\mathrm{diag}(e^{i\theta/2},e^{-i\theta/2})\right\},$$
$$\gamma_3\equiv\begin{pmatrix}\sigma_3 & 0\\ 0 & \sigma_3\end{pmatrix}=2\hat{H}. \tag{4.4}$$

and the Lie algebra of M_0 (and of T) is $\mathfrak{m}_0$. Let M'_0 be the normalizer of A_0 in K. Then of course M'_0/M_0 is equal to the restricted Weyl group $W(\mathfrak{g},\mathfrak{a}_0)$ which we displayed in Sec. II D.

In complete parallel to the algebraic discussion $P_0=M_0A_0N_0$ is a minimal parabolic subgroup of G and a standard parabolic subgroup of G is any closed subgroup of G containing P_0. The parabolic subgroups are displayed in the following way.[4] Let $\Psi=\{s_1,s_2\}$ be the set of generating elements of $W(\mathfrak{g},\mathfrak{g}_0)$ [cf. Sec. II D]. Then to each subset $\psi\in\Psi$ corresponds a parabolic subgroup of G

$$P_\psi\equiv\bigcup_{\sigma\in s\in\psi}P_0\sigma P_0, \tag{4.5}$$

where σ is a representative of s. Thus we have

$$P_\varnothing=P_0,\quad P_a=P_0\sigma_aP_0,\quad \sigma_a\in s_a\quad (a=1,2),$$
$$P_\Psi=G=M_\Psi,\quad A_\Psi=N_\Psi=\{1\}. \tag{4.6}$$

Explicitly we obtain

$$P_a=M_aA_aN_a\quad (a=1,2), \tag{4.7}$$

$$M_1=T\times\mathrm{SL}(2,\mathbb{R})$$
$$=\left\{\begin{pmatrix} e_1e^{i\theta/2}+\lambda e^{-i\theta/2}e_2 & i\mu e^{-i\theta/2}e_2\\ -i\nu e^{-i\theta/2}e_2 & e_1e^{i\theta/2}+\rho e_2e^{-i\theta/2}\end{pmatrix},\right.$$
$$\left.\lambda,\mu,\nu,\rho\in\mathbb{R},\lambda\rho-\mu\nu=1\right\}; \tag{4.8a}$$

$$A_1=\exp(\mathfrak{a}_1)=\left\{a_1=\begin{pmatrix} e_1e^s+e_2 & 0\\ 0 & e_1e^{-s}+e_2\end{pmatrix},\quad s\in\mathbb{R}\right\}, \tag{4.8b}$$

$$N_1=\exp(\mathfrak{n}_1)=\left\{n_1=\begin{pmatrix} b_w & 0\\ i\tilde{B} & b_w^{+\,-1}\end{pmatrix}\in N_0,\ b_+=0\right\}, \tag{4.8c}$$

$$\tilde{N}_1=\exp(\tilde{\mathfrak{n}}_1)=\left\{\tilde{n}_1=\begin{pmatrix} x_z & i\mathfrak{X}\\ 0 & x_z^{+\,-1}\end{pmatrix}\in\tilde{N}_0,\ x_-=0\right\}, \tag{4.8d}$$

$$M_2=\mathrm{SL}(2,\mathbb{C})\times\{1,\gamma_3\}$$
$$=\left\{m_2=\begin{pmatrix} l\sigma_3^N & 0\\ 0 & l^{+\,-1}\sigma_3^N\end{pmatrix}, l\in\mathrm{SL}(2,\mathbb{C}), N=0,1)\right\}, \tag{4.9a}$$

$$A_2=\exp(\mathfrak{a}_2)=\left\{a_2=\begin{pmatrix}\sqrt{|a|}1_2 & 0\\ 0 & \sqrt{|a|}^{-1}1_2\end{pmatrix},\quad |a|\in\mathbb{R}_+\right\}, \tag{4.9b}$$

$$N_2=\exp(\mathfrak{n}_2)$$
$$=\left\{n_2=\begin{pmatrix}1_2 & 0\\ i\tilde{b} & 1_2\end{pmatrix}\in N_0,\ w=0\right\}\quad (\tilde{B}(w=0)=\tilde{b}), \tag{4.9c}$$

$$\tilde{N}_2=\exp(\tilde{\mathfrak{n}}_2)$$
$$=\left\{\tilde{n}_2=\begin{pmatrix}1_2 & i\mathfrak{x}\\ 0 & 1_1\end{pmatrix}\in\tilde{N}_0,\ z=0\right\}\quad (\mathfrak{X}(z=0)=\mathfrak{x}). \tag{4.9d}$$

In the notation of Ref. 9 $P_1=P_{2f_2}$, $P_2=P_{f_1-f_2}$ and analogously for $M_a\ A_a\ N_a$. Here $M_a\,(a=1,2)$ is the centralizer of A_a in G and let M'_a be the normalizer of A_a in G. Then $M'_a/M_a=W(\mathfrak{g},\mathfrak{a}_a)$, which are given [cf. Secs. II D and II E and (4.6)] by construction[9]

$$W(\mathfrak{g},\mathfrak{a}_a)=\{1,s_a\}. \tag{4.10}$$

B. Matrix representation of the Weyl groups

First we give an explicit expression for the representative elements of the restricted Weyl group $W(\mathfrak{g},\mathfrak{a}_0)$. As matrices these belong to the maximal compact subgroup K. We display one-parameter families of representatives so we actually have a parametrization of $M'_0\subset K$:

$$\sigma(s_1)=\begin{pmatrix} le_2 & -le_1\\ -le_1 & le_2\end{pmatrix},$$
$$l=[(1-i)/\sqrt{2}](l_0+il_3\sigma_3),\quad l_0^2+l_3^2=1, \tag{4.11a}$$

$$\sigma(s_2)=\begin{pmatrix}0 & l'\\ l' & 0\end{pmatrix},$$
$$l'=i(l'_1\sigma_1+l'_2\sigma_2),\quad l_1'^2+l_2'^2=1, \tag{4.11b}$$

$$\sigma(s_3)=\begin{pmatrix} l''e_1 & -l''e_2\\ -l''e_2 & l''e_1\end{pmatrix},$$
$$l''=[(1+i)/\sqrt{2}](-l''_0+il''_3\sigma_3),\quad l_0''^2+l_3''^2=1, \tag{4.11c}$$

$$\sigma(s_4)=\begin{pmatrix} l''' & 0\\ 0 & l'''\end{pmatrix},$$
$$l'''=i(l'''_1\sigma_1+l'''_2\sigma_2),\quad l_1'''^2+l_2'''^2=1. \tag{4.11d}$$

We can explicitly check that as matrices [cf. (2.29) and (2.30)]

$$\sigma(s_k)\hat{e}_a\sigma(s_k)^{-1}=s_k(\hat{e}_a). \tag{4.12}$$

In such calculations the phase factors in l, l', l'', l''' of course cancel. They are needed however in other calculations, since without them $\sigma(s_k)$ would not belong to K [and (4.11a), and (4.11c) not even to G]. Note also that l, l', l'', l''' are isomorphic to M_0.

As we know it is important to choose representatives (in the study of Knapp–Stein intertwining operators) so that [cf. (2.14), (2.31b)]

$$\beta_2=\sigma(s_2)=\sigma(s_1)\sigma(s_3)=\sigma(s_2)\sigma(s_4). \tag{4.13}$$

This is, of course, possible because the action of θ and s_7 coincide. To satisfy (4.13) we take $l''_{0,3} = l_{0,3}, l''' = -l'$. In order to simplify calculations we shall most often additionally set

$$l = [(1-i)/\sqrt{2}]\,1_2, \quad l' = -i\sigma_2 \equiv \epsilon. \tag{4.14}$$

Using these choices we also have $l'' = -[(1+i)/\sqrt{2}]\,1_2$ and

$$\sigma(s_5) = \sigma(s_2)\sigma(s_1) = \frac{i-1}{\sqrt{2}}\begin{pmatrix}\sigma_- & \sigma_+\\ \sigma_+ & \sigma_-\end{pmatrix},$$
$$\sigma(s_6) = \sigma(s_1)\sigma(s_2) = \frac{1-i}{\sqrt{2}}\begin{pmatrix}\sigma_+ & \sigma_-\\ \sigma_- & \sigma_+\end{pmatrix}. \tag{4.15}$$

Next we consider the restricted Weyl groups $W(\mathfrak{g}, \mathfrak{a}_1)$ and $W(\mathfrak{g}, \mathfrak{a}_2)$. Their nontrivial elements are s_1 and s_2, respectively [cf. (4.10)], and a natural choice is to take for the representatives (4.11a) and (4.11b), respectively. However we have more choices since we only need to fulfill [cf. (2.35b)]

$$\sigma_1(s_1)\hat{e}_1\sigma_1(s_1)^{-1} = -\hat{e}_1, \quad \sigma_2(s_2)\hat{e}_2\sigma_2(s_2)^{-1} = -(\hat{e}_1+\hat{e}_2), \tag{4.16}$$

instead of (4.12). We shall not write the full solutions of (4.16). What is important and we shall use it below is that we can take (as expected on general grounds)

$$\sigma_1(s_1) = \beta_2, \quad \sigma_2(s_2) = \beta_2. \tag{4.17}$$

Last we turn to the question of the representatives of the (full) Weyl group $W(\mathfrak{g}^{\mathbb{C}}, \mathfrak{h}^{\mathbb{C}})$. Analogously to the case of the restricted Weyl group we find the following representatives $\omega(w_k)$, w_k in (3.7):

$$\omega(w_1) = \sigma(s_1) = \frac{1-i}{\sqrt{2}}\begin{pmatrix}e_2 & -e_1\\ -e_1 & e_2\end{pmatrix}, \tag{4.18a}$$

$$\omega(w_2) = \begin{pmatrix}e_2 & \sigma_+\\ \sigma_- & e_1\end{pmatrix}, \tag{4.18b}$$

$$\omega(w_3) = \sigma(s_3) = -\frac{1+i}{\sqrt{2}}\begin{pmatrix}e_1 & -e_2\\ -e_2 & e_1\end{pmatrix}, \tag{4.18c}$$

$$\omega(w_4) = \begin{pmatrix}\sigma_1 & 0\\ 0 & 1_2\end{pmatrix}, \tag{4.18d}$$

$$\omega(w_s) = \begin{pmatrix}e_1 & \sigma_-\\ \sigma_+ & e_2\end{pmatrix}, \tag{4.18e}$$

$$\theta(w_6) = = \begin{pmatrix}1_2 & 0\\ 0 & \sigma_1\end{pmatrix}. \tag{4.18f}$$

All are given up to phase factors, which are chosen so that

$$\omega(w_k) \in \mathrm{SL}(4,\mathbb{C}) = G^{\mathbb{C}}. \tag{4.19}$$

Actually $\omega(w_1), \omega(w_3) \in K$ since they coincide with $\sigma(s_1), \sigma(s_3)$. This simple fact shall have very important consequences later. Namely, the Knapp–Stein integral intertwining operators $\mathscr{A}(s_1)$, $\mathscr{A}(s_3)$ shall reduce to the differential intertwining operators $d^p(w_1)$, $d^q(w_3)$ whenever the latter are defined. The $\mathscr{A}(s_k)$ shall be defined in Part II and $d(w_k)$ in Part III. In Sec. VI below this connection is explained in some detail.

V. THE IWASAWA AND BRUHAT DECOMPOSITIONS OF G

A. The Iwasawa decomposition

In this section $\beta = \beta_2$ unless specified otherwise. It is well known that every element of G may be represented uniquely in the factorized form[22] ($k_I \in K$, $n_I \in N_0$, $a_I \in A_0$)

$$g = \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = k_I n_I a_I, \tag{5.1}$$

where for $\beta = \beta_2$ [compared with (4.2)]

$$k = \begin{pmatrix}p & q\\ q & p\end{pmatrix}, \quad p \pm q \in \mathrm{U}(2), \quad \det(p+q)(p-q) = 1, \tag{5.2}$$

$p^+p + q^+q = pp^+ + qq^+ = 1_2$, $p^+q + q^+p = 0$ and n_I, a_I are parametrized as in (4.3). For a 2×2 matrix α we shall write the decomposition

$$\alpha = \alpha_1 e_1 + \alpha_2 e_2 + \alpha_+\sigma_+ + \alpha_-\sigma_-. \tag{5.3}$$

Then for the parameters of the Iwasawa decomposition we obtain

$$e^{s_I} = 1/\sqrt{D_1}, \quad D \equiv \beta^+\beta + \delta^+\delta \quad (D_1 > 0);$$
$$e^{t_I} = \sqrt{\frac{D_1}{\det D}}, \quad w_I = \frac{D_+}{\sqrt{\det D}} \quad (\det D > 0); \tag{5.4a}$$
$$p_I = \delta(b_{w_I}a_I)^+, \quad q_I = \beta(b_{w_I}\hat{a}_I)^+,$$
$$b_{w_I}a_I = \frac{1}{\sqrt{D_1 \det D}}\begin{pmatrix}\sqrt{\det D} & 0\\ -D_+ & D_1\end{pmatrix}; \tag{5.4b}$$
$$iB_I = (b_{w_I}\hat{a}_I)^{+-1}\delta^{-1}\gamma\hat{a}_I^{-1} + b_{w_I}\hat{a}_I\beta^+\delta^{+-1}\hat{a}_I^{-1}$$
$$(\det \delta \neq 0);$$
$$i\widetilde{B}_I = (b_{w_I}\hat{a}_I)^{+-1}\beta^{-1}\alpha\hat{a}_I^{-1} + b_{w_I}\hat{a}_I\delta^+\beta^{+-1}\hat{a}_I^{-1}$$
$$(\det \beta \neq 0). \tag{5.4c}$$

From (5.4c) we obtain b_I using (4.3d). Note that $D_1 > 0$, $\det D > 0$ always. The order of the factors in the Iwasawa decomposition is a matter of convenience [however the expressions (5.4) depend on it]. In the Appendix we display the Iwasawa decomposition in the form $\widetilde{N}_0 A_0 K$, which is convenient in the study of K-induced representations.[15,18]

B. The (Gel'fand–Naimark)–Bruhat decomposition

It is well known that up to some submanifolds of lower dimension every element of G may be written in a unique way as a product[24]

$$g = \tilde{n}nam, \tag{5.5}$$

where $\tilde{n} \in N_0$, $n \in N_0$, $a \in A_0$, and $m \in M_0$. We give the precise statement for $G = \mathrm{SU}(2, 2)$ in the following proposition.

Proposition 5.1: Let $g = \left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right) \in \mathrm{SU}(2.2)$. Let also $\det \delta \neq 0$ and $\delta_1 = g_{33} \neq 0$. Then formula (5.5) holds and if $\tilde{n}$, n, a, and m are parametrized as in (4.3), (4.4) the following formulas for the parameters hold:

$$e^s = 1/|\delta_1|, \quad e^t = |\delta_1|/|\det \delta|; \tag{5.6a}$$

$$e^{i\theta/2} = \delta_1/|\delta_1|, \quad m = \tau(\theta)\gamma_3^N, \quad (-1)^N = \mathrm{sgn}\,\det \delta; \tag{5.6b}$$

$$\bar{z} = -\delta_-/\delta_1, \quad x_- = i(\delta_+\beta_- - \delta_1\beta_2)/\det \delta, \tag{5.6c}$$

$$\bar u = -\underline{\beta}_-/\underline{\delta}_1 - ix_-\bar z/2, \quad x_+ = |z|^2 x_-/6 + \mathrm{Im}(\underline{\beta}_1/\underline{\delta}_1);$$
$$w = \underline{\delta}_+\underline{\delta}_1/\det \underline{\delta}, \quad b_+ = i(\underline{\gamma}_+\underline{\delta}_- - \underline{\gamma}_2\underline{\delta}_1)\det \underline{\delta}/|\underline{\delta}_1|^2,$$
$$\bar v = \underline{\gamma}_+ \det \underline{\delta}/\underline{\delta}_1 + ib_+\bar w/2, \quad b_- = |w|^2 b_+/6 + \mathrm{Im}(\underline{\gamma}_1\underline{\delta}_1). \tag{5.6d}$$

Proof: By straightforward matrix multiplication we first obtain

$$x_z b_w \hat a \hat r \sigma_3^N = \underline{\delta}^{+\,-1}, \tag{5.7a}$$
$$i\underline{X} = \underline{\beta}\underline{\delta}^{-1}x_z^{+\,-1}, \quad i\tilde{\underline{B}} = x_z^+ \underline{\delta}\underline{\delta}^+ x_z b_w. \tag{5.7b}$$

From (5.7a) we easily obtain (5.6a), (5.6b), and the expressions for $\bar z$ and w. Then we substitute x_z and b_w in (5.7b) and using (4.3d) obtain the rest of (5.6c) and (5.6d).

Remark: Note that x_- and b_+ are real. For convenience we have written down the formulas for the complex conjugated variables $\bar z$, $\bar u$, $\bar v$ and to make expressions shorter we use z, x_- in the formulas for $\bar u$, x_+ and those for w, b_+ in the formulas for $\bar v$, b_-.

When the conditions of Proposition 5.1 are not met, there are decompositions of the form

$$g = \sigma(s_k)\tilde n^k nam \quad (k = 1,...,7), \tag{5.8}$$

where $\sigma(s_k)$ are representatives of the restricted Weyl reflections [cf. (4.11)] and $\tilde n^k \in \widetilde N_0^{(k)}$ which is the submanifold of $\widetilde N_0$ that remains invariant under the action of s_k,

$$\widetilde N_0^{(k)} \equiv \sigma(s_k)\widetilde N_0 \sigma(s_k)^{-1} \cap \widetilde N_0. \tag{5.9}$$

If we set $d_{nk} = \dim \widetilde N_0^{(k)}$, then (5.8) is describing a submanifold $G^{(k)} \subset G$

$$G^{(k)} = \sigma(s_k)\widetilde N_0^{(k)} N_0 A_0 M_0, \quad d_k \equiv \dim G^{(k)} = d_{nk} + 9. \tag{5.10}$$

Note $\widetilde N_0^{(7)} = \{1\}$, $d_7 = 9$.

We collect the various cases in the following proposition.

Proposition 5.2: Let $g = \begin{pmatrix} \underline{\alpha} & \underline{\beta} \\ \underline{\gamma} & \underline{\delta} \end{pmatrix} \in \mathrm{SU}(2,2)$. Then either $\det \underline{\delta} \neq 0$ and $\underline{\delta}_1 \neq 0$ or the elements of g meet one of the seven conditions below. Further in each of these seven cases a decomposition formula for g holds as indicated by the arrows and g describes a submanifold of G with dimension as specified:

(1) $\det \underline{\delta} \neq 0,\ \underline{\delta}_1 = 0 \Rightarrow g = \sigma(s_4)\tilde n(z=0)nam \in G^{(4)},$
$$d_4 = 13; \tag{5.11}$$
(2) $\det \underline{\delta} = 0, \underline{\delta}_1 \neq 0 \Rightarrow g = \sigma(s_3)\tilde n(x=0)nam \in G^{(3)},$
$$d_3 = 14; \tag{5.12}$$
(3) $\det \underline{\delta} = 0, \underline{\delta}_1 = 0 \neq \underline{\delta}_+$
$\Rightarrow g = \sigma(s_6)\tilde n(x_+ = z = 0)nam \in G^{(6)},$
$$d_6 = 12; \tag{5.13}$$
(4) $\det \underline{\delta} = 0,\ \underline{\delta}_1 = \underline{\delta}_+ = 0 \neq \underline{\delta}_-$
$\Rightarrow g = \sigma(s_5)\tilde n(x = z = 0)nam \in G^{(5)},$
$$d_5 = 12; \tag{5.14}$$
(5) $\det \underline{\delta} = 0,\ \underline{\delta}_1 = \underline{\delta}_+ = \underline{\delta}_- = 0 \neq \underline{\delta}_2$
$\Rightarrow g = \sigma(s_1)\tilde n(x_+ = u = z = 0)nam \in G^{(1)}$
$$d_1 = 10; \tag{5.15}$$
(6) $\underline{\delta} = 0,\ \underline{\gamma}_+ \neq 0 \Rightarrow g = \sigma(s_2)\tilde n(x_\pm = u = 0)nam \in G^{(2)},$
$$d_2 = 11; \tag{5.16}$$
(7) $\underline{\delta} = 0, \underline{\gamma}_+ = 0 \Rightarrow g = \sigma(s_7)nam \in G^{(7)},$
$$d_7 = g. \tag{5.17}$$

Let also $\sigma(s_k)$, $\tilde n$, n, a, m be given as in (4.11), (4.13), (4.15), (4.2), and (4.4). Then the following formulas for the parameters in $\tilde n$, n, a, and m hold [case by case in the compact form (5.7)]:

(1) $b_w \hat a \hat r \sigma_3^N = \epsilon(\underline{\delta}^+)^{-1}$ (5.18a)
$$i\underline{X} = i\underline{x} = -\epsilon\underline{\beta}\underline{\delta}^{-1}\epsilon, \quad i\tilde{\underline{B}} = -\epsilon\underline{\gamma}\underline{\delta}^+\epsilon b_w; \tag{5.18b}$$
(2) $x_z b_w \hat a \hat r \sigma_3^N = -[(1+i)/\sqrt{2}](e_1\underline{\delta} - e_2\underline{\beta})^{+\,-1},$ (5.19a)
$$i\underline{X}(x_- = 0) = (e_1\underline{\beta} - e_2\underline{\delta})(e_1\underline{\delta} - e_2\underline{\beta})^{-1}x_z^{+\,-1},$$
$$i\tilde{\underline{B}} = x_z^+(e_1\underline{\gamma} - e_2\underline{\alpha})(\underline{\delta}^+ e_1 - \underline{\beta}^+ e_2)x_z b_w \tag{5.19b}$$
(3) $b_w \hat a \hat r \sigma_3^N = [(1+i)/\sqrt{2}](\sigma_+\underline{\beta} + e_2\underline{\delta}_+)^{+\,-1},$ (5.20a)
$$i\underline{X}(x_+ = 0) = (\sigma_-\underline{\beta} + \sigma_+\underline{\delta}_2)(\sigma_+\underline{\beta} + e_2\underline{\delta}_+)^{-1},$$
$$i\tilde{\underline{B}} = (\sigma_+\underline{\alpha} + \sigma_-\underline{\gamma})(\underline{\beta}^+\sigma_- + e_2\bar{\underline{\delta}}_+)b_w; \tag{5.20b}$$
(4) $x_z b_w \hat a \hat r \sigma_3^N = -[(1+i)/\sqrt{2}](\sigma_-\underline{\beta} + \sigma_+\underline{\delta})^{+\,-1},$ (5.21a)
$$i\underline{X} = ix_+e_1 = [-\det \underline{\beta}/(\underline{\delta}_-\underline{\beta}_+ - \underline{\beta}_1\underline{\delta}_2)]e_1,$$
$$i\tilde{\underline{B}} = x_z^+(\sigma_-\underline{\alpha} + \sigma_+\underline{\gamma})(\underline{\beta}^+\sigma_+ + \underline{\delta}^+\sigma_-)x_z b_w; \tag{5.21b}$$
(5) $b_w \hat a \hat r \sigma_3^N = [(1+i)/\sqrt{2}](e_2\underline{\delta}_2 - e_1\underline{\beta})^{+\,-1},$ (5.22a)
$$i\underline{X} = ix_-e_2 = e_2\underline{\beta}_2/\underline{\delta}_2,$$
$$i\tilde{\underline{B}} = (e_2\underline{\gamma} - e_1\underline{\alpha})(\bar{\underline{\delta}}_2 e_2 - \underline{\beta}^+ e_1)b_w; \tag{5.22b}$$
(6) $x_z b_w \hat a \hat r \sigma_3^N = -\epsilon\underline{\beta}^{+\,-1},$ (5.23a)
$$i\underline{X} = 0, \quad i\tilde{\underline{B}} = -x_z^+\epsilon\underline{\alpha}\underline{\beta}^+\epsilon x_z b_w; \tag{5.23b}$$
(7) $b_w \hat a \hat r \sigma_3^N = \underline{\beta}^{+\,-1}$ (5.24a)
$$i\underline{\beta} = \underline{\alpha}\underline{\beta}^+ b_w. \tag{5.24b}$$

Proof: First we must show that formulas (5.18)–(5.24) are meaningful. For this we prove that all inverses of matrices exist. In case (1) this is true by supposition. In cases (6) and (7) we note that $\det \underline{\beta} \neq 0$ because we cannot have $\det \underline{\beta} = 0$ *and* $\underline{\delta} = 0$. In cases (2)–(5) suppose the opposite, i.e., that the relevant matrix (e.g., $e_1\underline{\delta} - e_2\underline{\beta}$) in case (2) has determinant zero. Then exploiting $\underline{\beta}^+\underline{\delta} + \underline{\delta}^+\underline{\beta} = 0$ [see (4.1)] in each case we conclude that also $\det \underline{\beta} = 0$, which is a contradiction since the three matrices in consideration cannot have zero determinants simultaneously. In case (4) we also note that the denominator is actually $\det(\sigma_-\underline{\beta} + \sigma_+\underline{\delta})$. Again using $\underline{\beta}^+\underline{\delta} + \underline{\delta}^+\underline{\beta} = 0$ (4.1) we show that the expressions for $x_\pm$ in cases (4) and (5) are real. In the same manner we observe that when we recover the most explicit expressions for the parameters using (4.3d) these are meaningful (i.e., real for $x_\pm$, $b_\pm$, positive for e^t and e^s, and of absolute value 1 for $e^{i\theta/2}$).

The next step is to substitute the expressions (5.18)–(5.24) on the right-hand sides of (5.11)–(5.17), respectively, and to obtain $g = g$ by straightforward matrix multiplication.

The last step is to show that either $\det \underline{\delta} \neq 0$ and $\underline{\delta}_1 \neq 0$ or the elements of g meet one of the seven conditions above. For this note that in the above we showed that these conditions can be met by the elements of g and by construction—formulas (5.18)–(5.24)—that the manifolds $G^{(k)}$ are not empty (and have the right dimensions as indicated). It remains to note that the condition $\det \underline{\delta} \neq 0$ and $\underline{\delta}_1 \neq 0$ plus the above seven conditions form a complete set of conditions. This completes the proof of Proposition 5.2.

C. The Bruhat decomposition for the nonminimal parabolics

We have written in the previous subsection the Bruhat decomposition in the case of the minimal parabolic subgroup P_0. We shall also display the Bruhat decomposition for P_1 and P_2 [the parallel of the algebra decomposition (2.40)].

Proposition 5.3: Let $g = \begin{pmatrix} \underline{\alpha} & \underline{\beta} \\ \underline{\gamma} & \underline{\delta} \end{pmatrix} \in \mathrm{SU}(2,2)$. Let also $\delta_1 = g_{33} \neq 0$. Then formula (5.5) holds with $\tilde{n} \in \tilde{N}_1$, $n \in N_1$, $a \in A_1$, $m \in M_1$, and if $\tilde{n}$, n, a, m are parametrized as in (4.8) then the following formulas for the parameters hold:

$$e^{i\theta/2} = \underline{\delta}_1/|\underline{\delta}_1|, \quad \lambda = (\underline{\alpha}_2\underline{\delta}_1 - \underline{\beta} - \underline{\gamma}_+)/|\delta_1|,$$

$$\mu = i(\underline{\beta}_-\underline{\delta}_+ - \underline{\beta}_2\underline{\delta}_1)/|\delta_1|, \tag{5.25a}$$

$$\nu = i(\underline{\delta}_1\underline{\gamma}_2 - \underline{\delta}_-\underline{\gamma}_+)/|\underline{\delta}_1|, \quad \rho = \det \underline{\delta}/|\underline{\delta}_1|$$

$$(\lambda\rho - \mu\nu = 1), \quad e^s = 1/|\delta_1|; \tag{5.25b}$$

$$w = \underline{\delta}_1^2(\underline{\alpha}_2\underline{\delta}_+ - \underline{\beta}_2\underline{\gamma}_+)/|\underline{\delta}_1|^2, \quad b_- = \mathrm{Im}(\underline{\gamma}_1\,\delta_1),$$

$$v = \underline{\delta}_1^2(\underline{\gamma}_+\underline{\delta}_2 - \underline{\delta}_+\underline{\gamma}_2)/|\underline{\delta}_1|^2; \tag{5.25c}$$

$$\bar{z} = -\underline{\delta}_-/\underline{\delta}_1, \quad x_+ = \mathrm{Im}(\underline{\beta}_1/\underline{\delta}_1), \quad \bar{u} = -\underline{\beta}_-/\underline{\delta}_1. \tag{5.25d}$$

Proposition 5.4 (Mack [14]): Let $g = \begin{pmatrix} \underline{\alpha} & \underline{\beta} \\ \underline{\gamma} & \underline{\delta} \end{pmatrix} \in \mathrm{SU}(2,2)$. Let also $\det \underline{\delta} \neq 0$. Then formula (5.5) holds with $\tilde{n} \in \tilde{N}_2$, $n \in N_2$, $a \in A_2$, $m \in M_2$, and if $\tilde{n}$, n, a, m are parametrized as in (4.9) then the following formulas for the parameters hold:

$$l = \underline{\delta}^{+-1}\sigma_3^N|\det \underline{\delta}|^{1/2}, \quad (-1)^N = \mathrm{sgn}\det \underline{\delta}, \tag{5.26a}$$

$$|a| = |\det \underline{\delta}|^{-1}, \tag{5.26b}$$

$$i\tilde{b} = \underline{\gamma}\underline{\delta}^+, \tag{5.26c}$$

$$i\underline{x} = \underline{\beta}\underline{\delta}^{-1}. \tag{5.26d}$$

Proof of Propositions 5.3 and 5.4: Straightforward matrix multiplication.

Remarks: Note that we do not have compact 2×2 matrix expressions in the case of the P_1 formulas (5.25). [Compare with (5.7) for P_0 and (5.26) for P_2.] Proposition 5.4 is given in Ref. 14 (not in the same form) and concerns the universal covering group of SU(2, 2) (that complicates the expression for the phase factor).

Next we display the connection between Bruhat decompositions for different parabolic subgroups.

Proposition 5.5: Let us have (for $\det \underline{\delta} \neq 0 \neq \underline{\delta}_1$), $g = \tilde{n}_0 n_0 a_0 m_0 = \tilde{n}_1 n_1 a_1 m_1$, where $\tilde{n}_k \in N_k$, $n_k \in N_k$, $a_k \in N_k$, $m_k \in M_k (k = 1, 2)$. Then the following formulas expressing the connection between the parameters in (5.6) and in (5.25) hold:

$$s_0 = s_1, \quad e^t = 1/|\rho| \quad (\rho = \det \underline{\delta}/|\underline{\delta}_1| \neq 0), \tag{5.27a}$$

$$e^{i\theta_0/2} = e^{i\theta_1/2}, \quad (-1)^N = \mathrm{sgn}\,\rho, \tag{5.27b}$$

$$z_0 = z_1, \quad x = \mu/\rho, \tag{5.27c}$$

$$\bar{u}_0 = \bar{u}_1 - i\mu z_1/2\rho, \quad x_{0+} = x_{1+} + \mu|z_1|^2/6\rho,$$

$$w_0 = w_1 + i\mu v_1/\rho, \quad b_+ = -\nu\rho,$$

$$v_0 = -\rho(i\nu w_1 - \lambda v_1),$$

$$b_{0-} = b_{1-} - \nu|\rho w_1 + i\mu v_1|^2/6\rho, \tag{5.27d}$$

$$\mu = (-1)^N e^{-t}x_-, \quad \nu = (-1)^{N+1}b_+e^t, \tag{5.28a}$$

$$\rho = (-1)^N e^{-t},$$

$$w_1 = (1 - xb_+)w_0 - ix_-v_0, \quad v_1 = v_0 + i\nu\rho w_0, \tag{5.28b}$$

$$b_{1-} = b_{0-} - |w_0|^2 b_+/6,$$

$$u_1 = u_0 - ix_-\bar{z}_0/2, \quad x_{1+} = x_{0+} - x_-|z_0|^2/6; \tag{5.28c}$$

where the parameters for the P_0 and P_1 decompositions are distinguished with lower indices 0 and 1 (when necessary).

Proof: Formulas (5.27) are obtained by substituting in (5.6) the inverse formulas of (5.25). Formulas (5.28) together with $s_1 = s_0$, $z_1 = z_0$, and $\theta_1 = \theta_0$ are the inverse of (5.27).

Proposition 5.6: Let us have (for $\det \underline{\delta} \neq 0 \neq \underline{\delta}_1$) $g = \tilde{n}_0 n_0 a_0 m_0 = \tilde{n}_2 n_2 a_2 m_2$, where $\tilde{n}_k \in N_k$ $n_k \in N_k$, $a_k \in A_k$, $m_k \in M_k (k = 0, 2)$. Then hold the following formulas expressing the connection between the matrices (5.7) and (5.26):

$$x_z b_w \hat{a}_0 \hat{\tau} \sigma_3^{N_0} = \sqrt{|a_2|}\, l \sigma_3^{N_2}, \tag{5.29a}$$

$$\underline{X} x_t^+ = \underline{x}_2, \quad x_t^{+\,-1}\tilde{\underline{B}}(x_z b_w)^{-1} = \tilde{b}_2. \tag{5.29b}$$

Also hold the connection between the parameters in (5.29a):

$$\sigma_3^{N_0} = \sigma_3^{N_2}, \quad l = \begin{pmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{pmatrix} = \exp\left(-\frac{s+t}{2}\right) x_z b_w \hat{a}_0 \hat{\tau}$$

$$= x_z b_w \hat{a}_- \hat{\tau} \quad (\alpha'\delta' - \beta'\gamma' = 1); \tag{5.30a}$$

$$|a_2| = e^{s+t}, \; \hat{a}_- \equiv \exp\left(-\frac{s+t}{2}\right)\hat{a}_0 = \begin{pmatrix} e^{(s-t)/2} & 0 \\ 0 & e^{(t-s)/2} \end{pmatrix}; \tag{5.30b}$$

$$z = \beta'/\delta', \quad \overline{w} = -\gamma'\delta', \quad e^{-i\theta/2} = \delta'/|\delta'| \quad (\delta' \neq 0); \tag{5.31a}$$

$$e^s = |a_2|/|\delta'|, \quad e^t = |a_2||\delta'|, \quad e^{(s-t)/2} = 1/|\delta'|. \tag{5.31b}$$

Proof: Formulas (5.29) are obtained by simply comparing (5.7) and (5.26), while formulas (5.30) and their inverses (5.31) are obtained by straightforward matrix multiplication.

Remark 1: Note that (5.30) and (5.31) give actually the P_0-Bruhat decomposition of m_2 [cf. (5.6)]

$$m_2 = \begin{pmatrix} l\sigma_3^N & 0 \\ 0 & l^{+-1}\sigma_3^N \end{pmatrix} = \tilde{n}_0(m_2) n_0(m_2) a_0(m_2) m_0(m_2), \tag{5.32}$$

where

$$\tilde{n}_0(m_2) = \begin{pmatrix} x_z & 0 \\ 0 & x_t^{+\,-1} \end{pmatrix}, \quad z = \beta'/\delta';$$

$$n_0(m_2) = \begin{pmatrix} b_w & 0 \\ 0 & b_w^{+\,-1} \end{pmatrix}, \quad \overline{w} = -\gamma'\delta'; (\delta' \neq 0),$$

$$a_0(m_2) = \begin{pmatrix} \hat{a}_- & 0 \\ 0 & \hat{a}_-^{-1} \end{pmatrix}, \quad \hat{a}_- = \begin{pmatrix} 1/|\delta'| & 0 \\ 0 & |\delta'| \end{pmatrix},$$

$$m_0(m_2) = \tau(\theta)\sigma_3^N, \quad e^{-i\theta/2} = \delta'/|\delta'|. \tag{5.33}$$

Even more directly (5.30) and (5.31) give the Bruhat decomposition for the group $\mathrm{SL}(2, \mathbb{C}) \ni l$.

Remark 2: Prompted by formulas (5.29) and (5.30) we shall use also the following parametrization for the elements of A_0, $\tilde{N}_0$, N_0 instead of (4.3):

$$\hat{a}_0=\sqrt{|a_0|}\,\hat{a}_-=\sqrt{|a_0|}\begin{pmatrix} e^r & 0\\ 0 & e^{-r}\end{pmatrix},\quad \tilde{n}_0=\begin{pmatrix} x_z & i\underset{\sim}{x}_0 x_z^{+-1}\\ 0 & x_z^{+-1}\end{pmatrix},$$

$$n_0=\begin{pmatrix} b_w & 0\\ i\tilde{b}_0 b_w & b_w^{+-1}\end{pmatrix}, \tag{5.34}$$

where

$$r\equiv(s-t)/2,\quad |a_0|\equiv e^{s+t},\quad \underset{\sim}{x}_0\equiv\underset{\sim}{X}x_t^+,\quad \tilde{b}_0\equiv\tilde{B}b_w^{-1}. \tag{5.35}$$

With this parametrization (5.29b), (5.30b), and (5.31b) now become

$$\underset{\sim}{x}_0=\underset{\sim}{x}_2,\quad x_z^{+-1}\tilde{b}_0x_z^{-1}=\tilde{b}_2,\quad |a_0|=|a_2|,\quad e^r=1/|\delta'|. \tag{5.36}$$

D. Relationship between the Iwasawa and Bruhat decompositions

Proposition 5.7: Let us have [for $(\det\delta\neq0\neq\delta_1)$] $g=kn_Ia_I=\tilde{n}\,n_Ba_B\,m$, where $k\in K$, $n_I, n_B\in N_0$, $a_I, a_B\in A_0$, $\tilde{n}\in\tilde{N}_0$, $m\in M_0$. Then hold the following formulas expressing the relationship between the parameters in (5.4) and (5.6) using (5.34):

$$|a_I|=\exp(s_I+t_I)=|a_B|/\sqrt{1+2x_E^2+x_M^4}$$

$$=\exp(s_B+t_B)/\sqrt{1+2x_E^2+x_M^4},$$

$$x_E^2\equiv x_0^2+x_1^2+x_2^2+x_3^2,\quad x_M^2\equiv x_0^2-x_1^2-x_2^2-x_3^2,$$

$$e^{r_I}=\exp\left(\frac{s_I-t_I}{2}\right)=e^{r_B}/\sqrt{f}\,(1+2x_E^2+x_m^4)^{1/4},$$

$$1+2x_E^2+x_M^4=\det(1+\underset{\sim}{x}^2), \tag{5.37a}$$

$$f\equiv(1+|u|^2)(1+|z|^2)+\ x_+^2+|z|^2x_-^2+2ix_0(u\bar{z}-\bar{u}z)>0,$$

$$w_I=(-1)^Ne^{-i\theta}\sqrt{1+2x_E^2+x_M^4}\,[fw_B$$

$$-2ix_0u-z(1+x_-^2+|u|^2)], \tag{5.37b}$$

$$\tilde{b}_I=\Omega^+[\tilde{b}_B-x_z^+\underset{\sim}{x}(1+\underset{\sim}{x}^2)^{-1}x_z]\Omega, \tag{5.37c}$$

$$\Omega\equiv b_{w_B}\hat{a}_{B-}\hat{r}\sigma_3^N(b_{w_I}\hat{a}_{I-})^{-1}=b_{w'}\hat{a}'\hat{r}\sigma_3^N$$

$$=\begin{pmatrix}\omega & 0\\ \omega w' & (-1)^Ne^{-i\theta/2}/\sqrt{f'}\end{pmatrix},$$

$$w'\equiv[2ix_0\bar{u}-\bar{z}(1+x_-^2+|u|^2)]/f,$$

$$\omega\equiv e^{i\theta/2}\sqrt{f(1+2x_E^2+x_M^4)},$$

$$\hat{a}'\equiv\hat{a}_{B-}\hat{a}_{I-}^{-1}=\begin{pmatrix}\sqrt{f(1+2x_E^2+x_m^4)} & 0\\ 0 & 1/\sqrt{f}\end{pmatrix};$$

$$\underset{\sim}{p}=(x_+\Omega)^{+-1},\quad \underset{\sim}{q}=i\underset{\sim}{x}(x_+\Omega)^{+-1},$$

$$\underset{\sim}{p}\underset{\sim}{p}^+=(1+\underset{\sim}{x}^2)^{-1},$$

$$|\det\underset{\sim}{p}|=1/\sqrt{1+2x_E^2+x_M^4}; \tag{5.37d}$$

$$x_zb_w\hat{a}_{B-}\hat{r}=\underset{\sim}{p}^{+-1}b_{w_I}\hat{a}_{I-}\sigma_3^N|\det p|^{1/2}, \tag{5.38a}$$

$$|a_B|=|a_I|/|\det\underset{\sim}{p}|,\quad(-1)^N=\operatorname{sgn}\det\underset{\sim}{p}, \tag{5.38b}$$

$$i\underset{\sim}{x}=\underset{\sim}{q}\underset{\sim}{p}^{-1},\quad i\tilde{b}_B=x_z^+(\underset{\sim}{q}+\underset{\sim}{p}i\tilde{b}_I)\underset{\sim}{p}^+x_+. \tag{5.38c}$$

Proof: Formulas (5.37) are obtained by substituting in (5.4) formulas (5.34) and the inverse of (5.6) [also using (5.34)]. Formulas (5.38) are the inverse of (5.37) [in (5.38c) x_z is understood as obtained from (5.38a)].

Another type of relationship when the Iwasawa decomposition is in the form $\tilde{N}_0A_0K$ is given in the Appendix.

We mention several special cases. First the Iwasawa decomposition of $\tilde{n}\in\tilde{N}_0$:

$$\tilde{n}=\begin{pmatrix}x_t & i\underset{\sim}{x}x_z^{+-1}\\ 0 & x_z^{+-1}\end{pmatrix}=k(\tilde{n})n(\tilde{n})a(\tilde{n}), \tag{5.39}$$

where instead of (5.4a) we have, using Proposition 5.7,

$$|a(\tilde{n})|=\exp(s(\tilde{n})+t(\tilde{n}))=1/\sqrt{1+2x_E^2+x_M^4}, \tag{5.40a}$$

$$e^{r(\tilde{n})}=\exp(s(\tilde{n})-t(\tilde{n}))/2$$

$$=1/\sqrt{f}(1+2x_E^2+x_M^4)^{1/4},$$

$$w(\tilde{n})=-\sqrt{1+2x_E^2+x_M^4}\,(z(1+x_-^2$$

$$+|u|^2)+iu(x_++x_-)), \tag{5.40b}$$

$$b(\tilde{n})=-(x_z\Omega)^+\underset{\sim}{x}(1+\underset{\sim}{x}^2)^{-1}x_z\Omega,$$

$$\Omega\equiv\Omega(\tilde{n})\equiv\begin{pmatrix}\omega(\tilde{n}) & 0\\ \overline{w}(\tilde{n})/\sqrt{f} & 1/\sqrt{f}\end{pmatrix},$$

$$\omega(\tilde{n})=\sqrt{f(1+2x_E^2+x_M^4)}, \tag{5.40c}$$

$\underset{\sim}{p}(\tilde{n})$ and $\underset{\sim}{q}(\tilde{n})$ are given as in (5.37d) with $\Omega=\Omega(\tilde{n})$, f is from (5.37a).

Next we consider the P_0-Bruhat decomposition of K

$$k=\begin{pmatrix}\underset{\sim}{p} & \underset{\sim}{q}\\ \underset{\sim}{q} & \underset{\sim}{p}\end{pmatrix}=\tilde{n}(k)n(k)a(k)m(k)\quad(\det\underset{\sim}{p}\neq0\neq\underset{\sim}{P}_1), \tag{5.41}$$

where instead of (4.6) we have

$$x_z(k)b_w(k)\hat{a}_-(k)\hat{r}(k)=|\det\underset{\sim}{p}|^{1/2}\underset{\sim}{p}^{+-1}\sigma_3^N, \tag{5.42a}$$

$$|a(k)|=1/|\det\underset{\sim}{p}|,\quad(-1)^N=\operatorname{sgn}\det\underset{\sim}{p}, \tag{5.42b}$$

$$i\underset{\sim}{x}(k)=\underset{\sim}{q}\underset{\sim}{p}^{-1},\quad i\tilde{b}(k)=x_z^+(k)\underset{\sim}{q}\underset{\sim}{p}^+x_z(k). \tag{5.42c}$$

Note also the Bruhat decomposition of $k(\tilde{n})$ [given by (5.37d) with $\Omega=\Omega(\tilde{n})$]:

$$k(\tilde{n})=\tilde{n}(k(\tilde{n}))n(k(\tilde{n}))a(k(\tilde{n}))m(k(\tilde{n})), \tag{5.43a}$$

$$\tilde{n}(k(\tilde{n}))=\tilde{n},\quad a(k(\tilde{n}))=a(\tilde{n})^{-1},\quad m(k(\tilde{n}))=1_4, \tag{5.43b}$$

$$n(k(\tilde{n}))=a(\tilde{n})^{-1}n(\tilde{n})^{-1}a(\tilde{n}), \tag{5.43c}$$

$$w(k(\tilde{n}))=e^{-2r(\tilde{n})}w(\tilde{n}),$$

$$\tilde{b}(k(\tilde{n}))=-\Omega(\tilde{n})^{+-1}\tilde{b}(\tilde{n})\Omega(\tilde{n})^{-1}. \tag{5.43d}$$

Formulas (5.43b) and (5.43c) are just a restatement of (5.39), while (5.43d) is (5.43c) written in detail. We are now ready to prove the following proposition.

Proposition 5.8: Let $k=\left(\begin{smallmatrix}\underset{\sim}{p} & \underset{\sim}{q}\\ \underset{\sim}{q} & \underset{\sim}{p}\end{smallmatrix}\right)$ and $\det\underset{\sim}{p}\neq0\neq\underset{\sim}{p}_1$. Then k can be decomposed uniquely in the form

$$k=k(\tilde{n}(k))m(k), \tag{5.44}$$

where $k(\tilde{n})$ is from (5.39), $\tilde{n}(k)$, $m(k)$ from (5.41).

Proof: We apply the Bruhat decomposition of $k(\tilde{n}(k))$ and obtain from (5.42)

$$k(\tilde{n}(k))=\tilde{n}(k)a(\tilde{n}(k))^{-1}n(\tilde{n}(k))^{-1}. \tag{5.45}$$

It remains to see by direct computation that

$$a(\tilde{n}(k))^{-1}=a(k),\quad n(\tilde{n}(k))=a(k)^{-1}n(k)^{-1}a(k), \tag{5.46}$$

which completes the proof.

Remark: Proposition 5.8 is the group structure parallel of the algebra map (2.24).

Next we consider quantities connected with the parabolic subgroup P_2 starting with the Iwasawa decomposition of $\tilde{n}_2 \in \tilde{N}_2$

$$\tilde{n}_2 = \begin{pmatrix} 1 & i\underset{\sim}{x} \\ 0 & 1 \end{pmatrix} = k(x)n(x)a(x), \tag{5.47}$$

where $a(x) = a(\tilde{n} = \tilde{n}_2)$ is given by (5.40a) with

$$f = f(x) = 1 + |u|^2 + x_+^2 = 1 + x_E^2 + 2x_0x_3, \tag{5.48a}$$

$$w(x) = -\sqrt{1 + 2x_E^2 + x_M^4}\, 2iux_0, \tag{5.48b}$$

$$\tilde{b}(x) = -\Omega_2^+ \underset{\sim}{x}(1 + \underset{\sim}{x}^2)^{-1}\Omega_2, \tag{5.48c}$$

$$\underset{\sim}{p}(x) = \Omega_2^{+-1}, \quad \underset{\sim}{q}(x) = i\underset{\sim}{x}\Omega_2^{+-1}, \tag{5.48d}$$

$$\Omega_2 \equiv \Omega_2(x) \equiv \Omega(\tilde{n} = \tilde{n}_2). \tag{5.48e}$$

Then we write down the P_2-Bruhat decomposition of k:

$$k = \begin{pmatrix} \underset{\sim}{p} & \underset{\sim}{q} \\ \underset{\sim}{q} & \underset{\sim}{p} \end{pmatrix} = \tilde{n}_2(k)n_2(k)a_2(k)m_2(k) \quad (\det \underset{\sim}{p} \neq 0), \tag{5.49}$$

$$l(k) = |\det \underset{\sim}{p}|^{1/2}\underset{\sim}{p}^{+-1}\sigma_3^N, \quad (-1)^N = \operatorname{sgn}\det \underset{\sim}{p}, \tag{5.50a}$$

$$|a_2(k)| = 1/|\det \underset{\sim}{p}|, \tag{5.50b}$$

$$i\underset{\sim}{x}_2(k) = \underset{\sim}{q}\underset{\sim}{p}^{-1}, \quad i\tilde{b}_2(k) = \underset{\sim}{q}\underset{\sim}{p}^{+}. \tag{5.50c}$$

Next the P_2-Bruhat decomposition of $k(x)$:

$$k(x) = \tilde{n}_2(k(x))n_2(k(x))a_2(k(x))m_2(k(x)), \tag{5.51}$$

$$l(k(x)) = \frac{1}{(1 + 2x_E^2 + x_M^4)^{1/4}}\Omega_2(x) = \frac{1}{(1 + 2x_E^2 + x_M^4)^{1/4}} \times \begin{pmatrix} \sqrt{f(x)(1 + 2x_E^2 + x_M^4)} & 0 \\ \overline{w}(x)/\sqrt{f(x)} & 1/\sqrt{f(x)} \end{pmatrix}, \tag{5.52a}$$

$$|a_2(k(x))| = \sqrt{1 + 2x_E^2 + x_M^4} = |a_2(x)|^{-1}, \tag{5.52b}$$

$$\underset{\sim}{x}_2(k(x)) = \underset{\sim}{x}, \quad \tilde{b}_2(k(x)) = \underset{\sim}{x}_2(1 + \underset{\sim}{x}_2^2)^{-1}, \tag{5.52c}$$

$$l(k(x)) = (b_{w(x)}\hat{a}_-(x))^{-1},$$

$$a_2(k(x))m_2(k(x)) = a_0(x)^{-1}n_0(w(x))^{-1}, \tag{5.52d}$$

$$\tilde{n}_2(k(x)) = \tilde{n}_2,$$

$$n_2(k(x)) = a_0(x)^{-1}n_0(x)^{-1}n_0(w(x))a_0(x). \tag{5.52e}$$

In (5.52) $a_0(x)$, $n_0(x)$ are from (5.47), $n_0(w(x))$ is $n_0(x)$ with $\tilde{b}(x) = 0$. Then we see that (5.51) is a restatement of (5.47).

We shall also need the Iwasawa decomposition of m_2:

$$m_2 = \begin{pmatrix} l\sigma_3^N & 0 \\ 0 & l^{+-1}\sigma_3^N \end{pmatrix} = k(m_2)n(m_2)a(m_2), \tag{5.53}$$

where

$$|a(m_2)| = 1, \quad e^{r(m_2)} = 1/\sqrt{|\beta'|^2 + |\delta'|^2}, \tag{5.54a}$$

$$l = \begin{pmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{pmatrix}, \quad \alpha'\delta' - \beta'\gamma' = 1,$$

$$w(m_2) = (-1)^{N+1}(\bar{\alpha}'\beta' + \bar{\gamma}'\delta'), \quad \tilde{b}(m_2) = 0, \tag{5.54b}$$

$$\underset{\sim}{q}(m_2) = 0, \quad \underset{\sim}{p}(m_2) = \underset{\sim}{p}(m_2)^{+-1} = \frac{1}{\sqrt{|\beta'|^2 + |\delta'|^2}}\begin{pmatrix} \bar{\delta}' & (-1)^N\beta' \\ -\bar{\beta}' & (-1)^N\delta' \end{pmatrix}. \tag{5.54c}$$

Now we are ready to prove the following which parallels the algebraic map (5.41a) for P_2.

Proposition 5.9: Let $k = \left(\begin{smallmatrix} \underset{\sim}{p} & \underset{\sim}{q} \\ \underset{\sim}{q} & \underset{\sim}{p} \end{smallmatrix}\right)$ and $\det \underset{\sim}{p} \neq 0$. Then k can be decomposed uniquely in the form

$$k = k(x(k))k(m_2(k)), \tag{5.55}$$

where $k(x)$ is from (5.47), and $x(k)$ and $m_2(k)$ are from (5.49).

Proof: We apply the Bruhat decomposition of $k(x(k))$ and obtain from (5.47)

$$k(x(k)) = \tilde{n}_2(k)a_0(x(k))^{-1}n_0(x(k))^{-1}. \tag{5.56}$$

Then we note

$$(n_0(x(k))a_0(x(k)))^{-1} = \begin{pmatrix} \Omega_k & 0 \\ \underset{\sim}{q}\underset{\sim}{p}^{+}\Omega_k & \Omega_k^{+-1} \end{pmatrix}, \tag{5.57a}$$

$$\Omega_k \equiv \Omega_2(x(k)) = \begin{pmatrix} \sqrt{\underset{\sim}{F}_1}/|\det \underset{\sim}{p}| & 0 \\ \underset{\sim}{F}_-/\sqrt{\underset{\sim}{F}_1}|\det \underset{\sim}{p}| & 1/\sqrt{\underset{\sim}{F}_1} \end{pmatrix},$$

$$\underset{\sim}{F} \equiv (\underset{\sim}{p}\underset{\sim}{p}^{+})^{-1}, \tag{5.57b}$$

$$l(k) = \sqrt{|\det \underset{\sim}{p}|}\,\Omega_2(x(k))\underset{\sim}{p}(m_2(k))\sigma_3^N, \tag{5.57c}$$

from which we immediately obtain

$$(n_0(x(k))a_0(x(k)))^{-1}k(m_2(k)) = n_2(k)a_2(k)m_2(k), \tag{5.58}$$

which completes the proof.

Finally we state without proof the analog of Propositions 5.8 and 5.9 for the parabolic subgroup P_1.

Proposition 5.10: Let $k = \left(\begin{smallmatrix} \underset{\sim}{p} & \underset{\sim}{q} \\ \underset{\sim}{q} & \underset{\sim}{p} \end{smallmatrix}\right)$ and $\underset{\sim}{p}_1 \neq 0$. Then k can be decomposed uniquely in the form

$$k = k(\tilde{n}_1(k))k(m_1(k)), \tag{5.59}$$

where $k(\tilde{n}_1)$, $\tilde{n}_1(k)$ and $m_1(k)$, $k(m_1)$ are, respectively, from

$$\tilde{n}_1 = k(\tilde{n}_1)n(\tilde{n}_1)a(\tilde{n}_1), \tag{5.60}$$

$$k = \tilde{n}_1(k)n_1(k)a_1(k)m_1(k) \quad (\underset{\sim}{p}_1 \neq 0), \tag{5.61}$$

$$m_1 = k(m_1)n(m_1)a(m_1). \tag{5.62}$$

Remark: Here $k(\tilde{n}_1)$ is given by formulas (4.37) with $\tilde{n} = \tilde{n}_1(x_- = 0)$.

VI. INTERTWINING OPERATORS AND REDUCIBLE ELEMENTARY REPRESENTATIONS (SUMMARY)

A. Knapp–Stein intertwining operators and octets of elementary representations

We shall announce some of the results which shall be explicitly proved in Parts II and III concerning the minimal parabolic subgroup $P_0 = M_0A_0N_0$. We parametrize the P_0-induced representations by

$$\chi = [n, \epsilon, c_1, c_2], \tag{6.1}$$

where $n \in \mathbb{Z}$ is indexing a character of T and $\epsilon = 0, 1$ indexing a character of $(1, \gamma_3)$ [see (4.4)]; c_1 and c_2 are two complex numbers [(a, b) in the notation of Ref. 9] which are the values characterizing the linear functional over $\mathfrak{a}_0$ that enters the inducing representation.

In Part II we give explicit construction of the representation space of χ and of the action of the representation operators. There are given three pictures depending on whether the representation space consists of C^∞ functions on G, on K or on $\tilde{N}_0$ (the so-called general, compact, and noncompact pictures, respectively). This is also done for the other parabolic $P_i = M_iA_iN_i$ subgroups and their representation spaces may consist also of C^∞ functions with values in the corresponding representation spaces of M_i.

We also find the exact relations between representations induced from different parabolic subgroups. For example, take the only noncuspidal parabolic $P_2 = M_2A_2N_2$. (Recall that a parabolic subgroup is cuspidal iff M has nonempty discrete series of representations.[4])

Let m_1, m_2 be two complex numbers ($n_1 - 1, n_2 - 1$ in the notation of Gel'fand *et al.*[25]) whose difference is an integer $m_1 - m_2 \in \mathbb{Z}$. These numbers fix a representation of SL(1,$\mathbb{C}$) (See Ref. 25). Let also $\epsilon' = 0, 1$ index a character of $(1, \gamma_3)$ as for (6.1) and $c \in \mathbb{C}$ fixes a representation of A_2. Then the P_2-induced representations are labeled

$$\chi_{(2)} = (\epsilon', m_1/2, m_2/2, c). \tag{6.2}$$

In Part II we give a constructive proof that the representations χ and $\chi_{(2)}$ are equivalent iff the following connection between the labels hold:

$$\begin{aligned} &\epsilon' = \epsilon, \quad m_1 = (n + c_2 - c_1)/2 - 1, \\ &m_2 = (c_2 - c_1 - n)/2 - 1, \quad c = (c_1 + c_2)/2; \\ &n = m_1 - m_2, \quad c_1 = c - 1 - (m_1 + m_2)/2, \\ &c_2 = c + 1 + (m_1 + m_2)/2. \end{aligned} \tag{6.3}$$

We must point out that this is not the usual construction of P_2-induced representations since we do not restrict to finite-dimensional representations of SL(2, C). These are obtained in the case

$$m_{1,2} + 1 = (c_2 - c_1 + n)/2 \in \mathbb{Z}_+. \tag{6.4}$$

Usually the finite-dimensional representations of SL(2, C) are labeled by the positive half-integers $j_k = m_k/2$ (see Ref. 14). We cited this particular example of the connection between induction from different parabolics for the benefit of the readers with the usual mathematical physics background.

We construct the Knapp–Stein[26] integral intertwining operators $\mathscr{A}$ which correspond to the seven nontrivial elements s_k of the restricted Weyl group [cf. (2.33)]. These operators group the ER's into octets of representations (each in one of the eight restricted Weyl chambers) by intertwining each member of a given octet with the other seven members. Explicitly the action of $\mathscr{A}(s_k)$ on χ is given by (we denote the representation spaces also by χ)

$$\mathscr{A}(s_1): \chi \equiv [n, \epsilon, c_1, c_2] \to [n, \epsilon, -c_1, c_2] \quad (c_1 \neq 0), \tag{6.5a}$$

$$\mathscr{A}(s_2): \chi \to [-n, (\epsilon + n)(\text{mod } 2), -c_2, -c_1] \quad (c_1 + c_2 \neq 0 \text{ or } n \neq 0), \tag{6.5b}$$

$$\mathscr{A}(s_3): \chi \to [n, \epsilon, c_1, -c_2] \quad (c_2 \neq 0), \tag{6.5c}$$

$$\mathscr{A}(s_4): \chi \to [-n, (\epsilon + n)(\text{mod } 2), c_2, c_1] \quad (c_1 - c_2 \neq 0 \text{ or } n \neq 0), \tag{6.5d}$$

$$\mathscr{A}(s_5): \chi \to [-n, (\epsilon + n)(\text{mod } 2), -c_2, c_1] \quad (c_1 \neq 0, c_2 \neq 0, \text{ or } n \neq 0), \tag{6.5e}$$

$$\mathscr{A}(s_6): \chi \to [-n, (\epsilon + n)(\text{mod } 2), c_2, -c_1] \quad (c_1 \neq 0, c_2 \neq 0, \text{ or } n \neq 0), \tag{6.5f}$$

$$\mathscr{A}(s_7): \chi \to [n, \epsilon, -c_1, -c_2] \quad (c_1 \neq 0 \text{ or } c_2 \neq 0), \tag{6.5g}$$

and the octet is explicitly parametrized by the following eight representations [$n \cdot c_1 \cdot c_2 \neq 0$ or $c_1 \cdot c_2 \cdot (c_1 \pm c_2) \neq 0$ is required]:

$$\chi_1 = [n, \epsilon, c_1, c_2], \tag{6.6a}$$

$$\chi_2 = [n, \epsilon, -c_1, c_2], \tag{6.6b}$$

$$\chi_3 = [-n, (\epsilon + \mathrm{n})(\text{mod } 2), -c_2, c_1], \tag{6.6c}$$

$$\chi_4 = [-n, (\epsilon + \mathrm{n})(\text{mod } 2), c_2, c_1], \tag{6.6d}$$

$$\chi_5 = [n, \epsilon, -c_1, -c_2], \tag{6.6e}$$

$$\chi_6 = [n, \epsilon, c_1, -c_2], \tag{6.6f}$$

$$\chi_7 = [-n, (\epsilon + \mathrm{n})(\text{mod } 2), c_2, -c_1], \tag{6.6g}$$

$$\chi_8 = [-n, (\epsilon + \mathrm{n})(\text{mod } 2), -c_2, -c_1]. \tag{6.6h}$$

This way of parametrizing the octet is dictated by the following: (6.6a) and (6.6b) are interconnected by the Knapp–Stein intertwining operator $\mathscr{A}(s_1)$; the same is true for the pairs (6.6c) and (6.6d), (6.6e) and (6.6f) and (6.6g) and (6.6h) while the pairs (6.6b) and (6.6c), (6.6d) and (6.6e), (6.6f), and (6.6g), and (6.6h) and (6.6a) are interconnected by the operator corresponding to the other simple reflection—$\mathscr{A}(s_2)$. This shall be our way, in general, of parametrizing the octets and often they shall be referred to by pointing out the first member [(6.6a) in the case above]. A nice way of graphically depicting the octet is to assign to each member one of the vertices of a cube. Then the three links and the four diagonals (connecting a given vertex with the other vertices) can be assigned the intertwining operators $\mathscr{A}(s_k)$ in a way consistent with the connections between these operators. This picture (which we do not include here for the lack of space) shall appear in Part II.

The symmetry of the octet under the restricted Weyl reflections allows us to choose the first member of the octet to be in the closed positive restricted Weyl chamber, i.e.,[9]

$$\operatorname{Re} c_1 \geqslant \operatorname{Re} c_2 \geqslant 0,$$

and also to choose $n \geqslant 0$. However this choice shall make our classification clumsy and because of this we shall make only the convention

$$\chi_1 = [n, \epsilon, c_1, c_2], \quad \operatorname{Re} c_1 \geqslant 0, \quad \operatorname{Re} c_2 \geqslant 0. \tag{6.7}$$

We stress that this condition is imposed only on the first member of an octet.

Next we mentioned the exceptional cases. In the case $c_1 = 0 \neq c_2$ the octet reduces to a quadruplet

$$\chi_1 = \chi_2 = [n, \epsilon, 0, c_2], \tag{6.8a}$$

$$\chi_3 = \chi_8 = [-n, \epsilon + n(\text{mod } 2), -c_2, 0], \tag{6.8b}$$

$$\chi_4 = \chi_7 = [-n, \epsilon + n(\text{mod } 2), c_2, 0], \tag{6.8c}$$

$$\chi_5 = \chi_6 = [n, \epsilon, 0, -c_2]. \tag{6.8d}$$

It is trivial to see how the seven operators from each member reduce to three and how the quadruplet can be arranged on the vertices of a quadrangle. This is spelled out in Part II where also the other, exceptional cases (the relevant one being $c_1 = c_2 \neq 0 = n$ — quadruplet and $c_1 = c_2 = 0 \neq n$ — doublet) are given together with graphical representation.

In Part II we give explicit construction of the operators for all induction pictures and all parabolics.

B. Reducible ER and differential intertwining operators

It is well known (cf. Refs. 8 and 9) that almost all ER are topologically irreducible, i.e., they contain no closed invariant subspaces. In our case the representation χ is reducible *only if either* at least one of the following four conditions is true:

$$(c_1 \pm c_2 \pm n)/2 \in \mathbf{Z}' \equiv \mathbf{Z} \setminus \{0\} \tag{6.9a}$$

or

$$c_1 \in \mathbf{Z},\ \epsilon = (c_1 + n + 1)_{\mathrm{mod}\,2} \ \text{ or } \ c_2 \in \mathbf{Z},\ \epsilon = (c_2 + 1)_{\mathrm{mod}\,2}. \tag{6.9b}$$

We derived the exact statement of the epsilon part of (6.9b) (which is not essential for the classification into multiplets in Sec. VI C) using the general (and not explicit) criterion of Ref. 27 concerning arbitrary linear semisimple Lie groups. (We become aware of Ref. 27 after the results announced here were obtained. Note that our elementary representations are called generalized principle series representations in Ref. 27.) It is easy to see that the ER's in a given octet are simultaneously irreducible or reducible. When they are irreducible the restricted Weyl group $W(\mathfrak{g}, \mathfrak{a}_0)$ is isomorphically mapped to the set of $\mathscr{A}(s_k)$ which then are expressed in terms of $\mathscr{A}(s_1)$, $\mathscr{A}(s_2)$. This fact is known in general[26] for any semisimple Lie group and because of it in the mathematical literature only one of the closed Weyl chambers is considered (usually the positive one in some ordering). However, when the ER's in the octet are reducible the intertwining operators realize only partial equivalence (being neither injective or surjective) and one should use their intrinsic definitions. This is one reason to consider (contrary to the usual use in the mathematical literature) all Weyl chambers at the same time.

In Part III we introduce and construct differential intertwining operators between reducible ER's. There are six basic operators corresponding to the positive noncompact roots of the complexified Lie algebra gl(4, $\mathbb{C}$) (cf. Sec. III B). It is not necessary here to consider all elements of the (nonrestricted) Weyl group, however again it is not enough to consider only the simple roots, or correspondingly, the generating elements of the Weyl group.

We shall give as an example the definition of these operators in the case of compex-valued C^∞-functions $f(\tilde{n})$ over $\widetilde{N}_0$. (They also satisfy some asymptotic conditions spelled out in Part II.) Recall that $\widetilde{N}_0$ is six dimensional and parametrized by z, $u \in \mathbb{C}$ and $x_\pm \in \mathbb{R}$ [cf. (5.43) $\underset{\sim}{x} = x_0 + x_k \sigma_k$, $u \equiv x_2 + ix_1$, $x_\pm \equiv x_0 \pm x_3$]. First we give the expression of the operators corresponding to the simple roots.

The operator corresponding to w_3 is defined when

$$c_2 = -\nu \in \mathbf{Z}_-, \tag{6.10a}$$

and is a mapping

$$d^{\nu}(w_3): [n, \epsilon, c_1, -\nu] \to [n, \epsilon, c_1, \nu], \tag{6.10b}$$

explicitly given by

$$d^{\nu}(w_3) f(\tilde{n}) \equiv \left[\frac{\partial}{\partial x_-} + |z|^2 \frac{\partial}{\partial x_+} + i\left(z\frac{\partial}{\partial u} - \bar{z}\frac{\partial}{\partial \bar{u}}\right)\right]^{\nu} f(\tilde{n}). \tag{6.10c}$$

The operator corresponding to w_4 is accordingly given as

$$\tfrac{1}{2}(c_1 - c_2 + n) = -P \in \mathbf{Z}_-, \tag{6.11a}$$

$$\begin{aligned} d^{p}(w_4): &[n, \epsilon, c_1, c_1 + n + 2p] \\ &\to [n + 2p, (\epsilon + p)_{(2)}, c_1 + p, c_1 + n + p] \\ &\qquad (\varkappa_{(2)} \equiv \varkappa (\mathrm{mod}\ 2)). \end{aligned} \tag{6.11b}$$

$$d^{p}(w_4) f(\tilde{n}) \equiv \left(\frac{\partial}{\partial z}\right)^{p} f(\tilde{n}), \tag{6.11c}$$

and the operator corresponding to w_6 as

$$\tfrac{1}{2}(c_1 - c_2 - n) = -p \in \mathbf{Z}_-, \tag{6.12a}$$

$$\begin{aligned} d^{p}(w_6): &[n, \epsilon, c_1, c_1 - n + 2p] \\ &\to [n - 2p, (\epsilon + p)_{(2)}, c_1 + p, c_1 - n + p], \end{aligned} \tag{6.12b}$$

$$d^{p}(w_6) f(\tilde{n}) \equiv \left(-\frac{\partial}{\partial \bar{z}}\right)^{p} f(\tilde{n}). \tag{6.12c}$$

[In (6.12c) we have introduced a normalization for convenience.] For the operators corresponding to the nonsimple roots the definitions are not so explicit. For w_1 we have

$$c_1 = -\nu \in \mathbf{Z}_-, \tag{6.13a}$$

$$d^{\nu}(w_1): [n, \epsilon, -\nu, c_2] \to [n, \epsilon, \nu, c_2], \tag{6.13b}$$

$$\begin{aligned} d^{\nu}(w_1) f(\tilde{n}) \equiv \prod_{k=1}^{\nu} &[(m_1 + 1 - k)(m_2 + 1 - k)\partial_+ \\ &+ (m_1 + 1 - k)(i\partial_{\bar{u}} - z\partial_+) \\ &- (m_2 + 1 - k)(i\partial_u - \bar{z}\partial_+) + \partial_- \partial_z \partial_{\bar{z}} \\ &+ i(\bar{z}\partial_{\bar{u}} - z\partial_u)\partial_z \partial_{\bar{z}}] f(\tilde{n}), \end{aligned} \tag{6.13c}$$

$$m_{1,2} \equiv \frac{c_2 + \nu \pm n}{2} - 1, \quad (6.3),$$

$$\partial_\pm \equiv \frac{\partial}{\partial x_\pm}, \quad \partial_z \equiv \frac{\partial}{\partial z}, \quad \partial_u \equiv \frac{\partial}{\partial u}.$$

In the cases of w_2 and w_5 we have, respectively,

$$(c_1 + c_2 + n)/2 = -q \in \mathbf{Z}_-, \tag{6.14a}$$

$$\begin{aligned} d^{q}(w_2): &[n, \epsilon, c_1, -c_1 - n - 2q] \\ &\to [n + 2q, (\epsilon + q)_{(2)}, c_1 + q, -c_1 - n - q], \end{aligned} \tag{6.14b}$$

$$\begin{aligned} d^{q}(w_2) f(\tilde{n}) \equiv \prod_{k=1}^{q} &[|z|^2 \partial_z \partial_+ + \partial_- \partial_z \\ &+ i(z\partial_u - \bar{z}\partial_{\bar{u}})\partial_z - (m_2 + 1 - k) \\ &\times (i\partial_u + \bar{z}\partial_+)] f(\tilde{n}) \\ (m_2 = &-(c_1 + n + q + 1)); \end{aligned} \tag{6.14c}$$

$$(c_1 + c_2 - n)/2 = -q \in \mathbf{Z}_-, \tag{6.15a}$$

$$\begin{aligned} d^{q}(w_5): &[n, \epsilon, c_1, -c_1 + n - 2q] \\ &\to [n - 2q, (\epsilon + q)_{(2)}, c_1 + q, -c_1 + n - q], \end{aligned} \tag{6.15b}$$

$$\begin{aligned} d^{q}(w_5) f(\tilde{n}) \equiv \prod_{k=1}^{q} &[|z|^2 \partial_+ \partial_{\bar{z}} + \partial_- \partial_{\bar{z}} \\ &+ i(z\partial_u - \bar{z}\partial_{\bar{u}})\partial_{\bar{z}} + (m_1 + 1 - k) \\ &\times (i\partial_{\bar{u}} - z\partial_+)] f(\tilde{n}), \\ (m_1 = &-(c_1 - n + q + 1)). \end{aligned} \tag{6.15c}$$

It should be stressed that formulas (6.10b)–(6.15b) are immediate consequences of the action of w_k on $\mathfrak{h}^{\mathbb{C}}$ (3.7) or alternatively on the roots (3.8). This is simply derived even in general[4] once the induction parameters are fixed. (This fact is also utilized in Ref. 28 which we read after the results announced here were obtained.) Then formulas (6.10a)–(6.15a) may be deduced by exploiting the differential character of the operators and compatibility with the labeling of the representations. The most difficult task of the actual construction [formulas (6.10c)–(6.15c)] is achieved in Part III in two ways. The first uses the fact that analogous operators appear in the study[12] of another real form of SL(4, C), namely $G' = \mathrm{SU}^*(4)$ [the double covering of $\mathrm{SO}_e(5,1)$]. This group is simpler being of split rank 1 and having only one parabolic different from G', say $P' = M'A'N'$, $M' = \mathrm{SU}(2)\times\mathrm{SU}(2)$ being compact. The differential intertwining operators are very simple—see formulas (6.8)–(6.11) of Ref. 12. Our operators $d^{\nu}(w_1)$, $d^{p}(w_2)$, $d^{\nu}(w_3)$, and $d^{p}(w_5)$ correspond to d'^{ν}, ∂^{p}, d^{ν}, and ∂'^{p} of Ref. 12, respectively. [The operators d^{ν}, d'^{ν} were first constructed for $\mathrm{SO}_e(n+1,1)$ in the case of symmetric tensor representations of $M' = \mathrm{SO}(n)$ (see Refs. 11 and 29).] In order to use this correspondence and derive (6.10c)–(6.15c) several steps are needed. These include passage from G', M' to G, M_2 by the Weyl unitary trick. This is explained in detail in Part III where also an independent derivation utilizing the properties of w_k as in Ref. 23 is given.

We must note that the operators $d^{\nu}(w_1)$, $d^{\nu}(w_3)$ may be obtained as reductions from the Knapp–Stein integral operators $\mathscr{A}(s_1)$, $\mathscr{A}(s_3)$, respectively, and they replace these last in the octets when they act on the left-hand sides of (6.13b) and (6.10b), respectively. This happens when the octets contain reducible representations. In some more degenerate cases (see Types IIa, IIIa, IIIb, IIId of reducible representations in Sec. VI C) also the action of $d^{p}(w_2)$, $d^{p}(w_5)$ may coincide with the action of $\mathscr{A}(s_2)$, while the action of $d^{q}(w_4)$, $d^{q}(w_6)$ with that of $\mathscr{A}(s_4)$.

C. On the classification of the reducible elementary representations

One of the main results of Part III is the classification of the reducible representations. This classification is twofold. First the reducible representations are classified according to the way they are grouped and which differential intertwining operators are defined. In this way three types emerge. Type I consists of octets and quadruplets and only the operators $d^{\nu}(w_1)$ and $d^{\nu}(w_3)$ appear. Type II consists of 16-plets (two octets connected) and their reductions and only the operators $d^{\nu}(w_2)$, $d^{\nu}(w_4)$, $d^{\nu}(w_5)$, and $d^{\nu}(w_6)$ appear. Type III (which could be viewed as intersection of the other types) consists of 24-plets (three octets connected) and their reductions and all differential operators appear. The second way of classification is according to the irreducible composition factors of the reducible elementary representations.

We now turn to the first way of classification.

Type I (*octet*): It is characterized by the following conditions on the first member of the octet χ_1:

$$c_1 = \nu \in \mathbb{Z}_+,\quad c_2 \notin \mathbb{Z}_+,\quad (c_2 \pm \nu \pm n)/2 \notin \mathbb{Z},$$

$$\epsilon = (\nu + n + 1)_{(2)}, \tag{6.16a}$$

$$\chi_1 = [n, (\nu + n + 1)_2, \nu, c_2]. \tag{6.16b}$$

This octet supports only four differential intertwining operators defined from the second and the fifth members (to the first and sixth, respectively) $-\, d^{\nu}(w_1)$ and from the seventh and eighth (to the fourth and the third, respectively) $-\, d^{\nu}(w_3)$.

If we start from the requirements (6.16a) with $c_1 \notin \mathbb{Z}_+$, $c_2 = \nu = (1+\epsilon)_{(2)}$, we obtain the same octet.

Type Ia (*octet*): It may be viewed as a subtype of type I; the conditions are

$$c_1 = \nu \in \mathbb{Z}_+,\quad c_2 = p \in \mathbb{Z}_+,\quad (p \pm \nu \pm n)/2 \notin \mathbb{Z}, \tag{6.17a}$$

$$\chi_1 = [n, \epsilon, \nu, p]. \tag{6.17b}$$

This type supports the same operators as above plus another *four* defined from the third and eighth members (to the fourth and seventh, respectively) $-\, d^{p}(w_1)$, and from the fifth and sixth (to the second and first, respectively) $-\, d^{p}(w_3)$.

Type Ib (*quadruplet*): This subtype parallels also the general case (6.8)

$$c_1 = 0,\quad c_2 = p \in \mathbb{Z}_+,\quad (p \pm n)/2 \notin \mathbb{Z}, \tag{6.18a}$$

$$\chi_1 = \chi_2 = [n, \epsilon, 0, p]. \tag{6.18b}$$

It supports only two operators $d^{p}(w_1)$ from $\chi_3 = \chi_8$ to $\chi_4 = \chi_7$ and $d^{p}(w_3)$ from $\chi_5 = \chi_6$ to $\chi_1 = \chi_2$.

Type II (16-plet): It is characterized by the following conditions:

$$(c_1 + c_2 - n)/2 = \nu \in \mathbb{Z}_+,\quad (c_1 - c_2 + n)/2 \notin \mathbb{Z},\quad \nu + n > 0. \tag{6.19a}$$

The 16-plet is a pairing of two octets connected with the operators $d^{\nu}(w_k)$ $k = 2,4,5,6$. The first and eighth members of the first octet are given by

$$\chi_1 = [n, \epsilon, c_1, -c_1 + n + 2\nu], \tag{6.19b}$$

$$\chi_8 = [-n, \epsilon + (n)_{(2)}, c_1 - n - 2\nu, -c_1].$$

Now one differential intertwining operator defined on χ_8 is $d^{\nu+n}(w_2)$ and it intertwines with the first member of the second octet or we shall call it the ninth member of the 16-plet

$$\chi'_1 = \chi_9 = [2\nu + n, \epsilon + (\nu)_{(2)}, c_1 - \nu, -c_1 + n + \nu]. \tag{6.19c}$$

Thus we have given the whole 16-plet ($\chi'_2 = \chi_{10}$, etc.). In this case we have 16 operators: $d^{\nu}(w_2)$ acts from the 5th and 12th member *to* the 13th and 4th, respectively $d^{\nu+n}(w_2)$ acts from the 8th and 16th *to* the 9th and 1st, respectively; $d^{\nu}(w_5)$ acts from the 8th and 9th *to* the 16th and 1st, respectively; $d^{\nu+n}(w_5)$ acts from the 5th and 13th *to* the 12th and 4th, respectively; $d^{\nu}(w_4)$ acts from the 2nd and 15th *to* the 10th and 7th, respectively; $d^{\nu+n}(w_4)$ acts from the 3rd and 11th *to* the 14th and 6th respectively; $d^{\nu}(w_6)$ acts from the 3rd and 14th to the 11th and 6th, respectively; $d^{\nu+n}(w_6)$ acts from the 2nd and 10th *to* the 15th and 7th, respectively.

If we require $\nu + n < 0$ in (6.19a) we shall obtain the same 16-plet after the change $n \to -n - 2\nu$.

The same 16-plet is obtained if we start with $(c_1+c_2+n)/2=\nu\in\mathbf{Z}_+$, $\nu-n\neq0$; or if we change (6.19a) to

$$(c_1-c_2-n)/2=\nu\in\mathbf{Z}_+,\quad \nu+n>0,\quad (c_1+c_2\pm n)/2\notin\mathbf{Z}. \tag{6.19a'}$$

Type IIa (*octet*): It may be viewed as a subtype in the case when $\nu+n=0$ and (6.19a) is replaced by

$$c_1+c_2=\nu\in\mathbf{Z}_+,\quad n=-\nu,\quad (c_1-c_2\pm\nu)/2\notin\mathbf{Z}. \tag{6.20a}$$

Now the 16-plet collapses to an octet

$$\chi_1=\chi_{16}=[-\nu,\epsilon,c_1,-c_1+\nu], \tag{6.20b}$$

$\chi_2=\chi_{15}$, etc. Only four intertwining differential operators remain $d^{\,\nu}(w_2)$, $d^{\,\nu}(w_5)$, $d^{\,\nu}(w_4)$, $d^{\,\nu}(w_6)$ acting from the 5th, 8th, 2nd, and 3rd, respectively, *to* the 4th, 1st, 7th, and 6th, respectively. So the operators act now inside the octet and actually the action of $d^{\,\nu}(w_2)$, $d^{\,\nu}(w_5)$ coincides with the action of $\mathscr{A}(s_2)$, while the action of $d^{\,\nu}(w_4)$, $d^{\,\nu}(w_6)$ with that of $\mathscr{A}(s_4)$.

This type can also be obtained when $\nu=0$ and (6.19a) is replaced by

$$c_1+c_2=n\in\mathbf{Z}_+,\quad (c_1-c_2\pm n)/2\notin\mathbf{Z}. \tag{6.20a'}$$

Type III (24-plet): This type may be characterized by requiring all four numbers for the first member

$$(c_1\pm c_2\pm n)/2\in\mathbf{Z}_+. \tag{6.21}$$

In this case we shall choose a parametrization which shall display explicitly the connection with previous work on a split-rank 1 real form of SL(4, C) – SU*(4) (see Ref. 12) and with work on P_2-finitely dimensionally induced representations.[17] Namely for the first member of the first octet we set

$$\chi_1=[2l-2n+2,\epsilon,2l+\nu+2,\nu], \tag{6.22a}$$

where

$$l=0,\tfrac12,1,\ldots,\quad \nu\in\mathbf{Z}_+,$$
$$n=1,2,\ldots,2l+1,\quad \epsilon=(2l+\nu+1)_{(2)},$$

and several other members are as follows:

$$\chi_1=[2l-2n+2,\epsilon,-(2l+\nu+2),\nu]\equiv\chi^{1-}_{l\nu n}. \tag{6.22b}$$

$$\chi_3=[-2l+2n-2,(\epsilon+2l)_{(2)},-\nu,2l+\nu+2]\equiv\chi^{1+}_{l\nu n}. \tag{6.22c}$$

$$\chi_4=[-2l+2n-2,(\epsilon+2l)_{(2)},\nu,2l+\nu+2]\equiv\chi^{+}_{l\nu n}. \tag{6.22d}$$

$$\chi_5=[2l-2n+2,\epsilon,-(2l+\nu+2),-\nu]\equiv\chi^{-}_{l\nu n}. \tag{6.22e}$$

$$\vdots$$

$$\chi_8=[-2l+2n-2,(\epsilon+2l)_{(2)},-\nu,-(2l+\nu+2)], \tag{6.22f}$$

$$\chi_9=[2l+2\nu+2,(\epsilon+\nu+n)_{(2)},2l-n+2-n], \tag{6.22g}$$

$$\vdots$$

$$\chi_{16}=[-(2l+2\nu+2),(\epsilon+2l+\nu+n)_{(2)},\ n,-(2l-n+2)], \tag{6.22h}$$

$$\chi_{17}=[-(2l+2),(\epsilon+2l+n)_{(2)},\ n+\nu,-(2l+\nu+2-n)], \tag{6.22i}$$

$$\vdots$$

$$\chi_{20}=[(l+2,(\epsilon+n)_{(2)},-(2l+\nu+2-n),\ n+\nu]\equiv\chi''^{+}_{l\nu n}, \tag{6.22j}$$

$$\chi_{21}=[-(2l+2),(\epsilon+2l+n)_{(2)},\ -(n+\nu),2l+2+\nu-n]\equiv\chi''^{-}_{l\nu n}. \tag{6.22k}$$

In the above members inside the octets χ_1 to χ_8, χ_9 to χ_{10}, and χ_{17} to χ_{24} are connected standardly; from χ_8 to χ_9 we go by $d^{\,p}(w_2)$, $p\equiv2l+\nu+2-n$ and from χ_{16} to χ_{17} we go by $d^{\,\nu}(w_4)$. We have singled out the P_2-finitely dimensionally induced representations and introduced for them the notation used before in the SU*(4) case—compare formulas (3.1) and (3.3) of Ref. 12 with (6.22) with ϵ suppressed.

Thus the whole 24-plet is given. In this case we have 72 operators, 12 of each type, so 24 are acting inside the octets (8 in each) and 48 are acting between the octets (8 in each direction). We give in Table I the action of the operators. We postpone the detailed comment till Part III where also a graphical representation of this table shall appear. We only note that for a given member of the 24-plet the sum of the number of operators it supports plus the number of the operators which map into it is constant and equals 6.

Type IIIa (12-plet): This would be obtained from III by putting $\nu=0$; then the 24-plet reduces to a 12-plet with

$$\chi_1=\chi_6=[2l-2n+2,\epsilon,2l+2,0],\quad \epsilon=(2l+1)_{(2)}, \tag{6.23}$$

$$\chi_2=\chi_5,\ \chi_3=\chi_4,\ \chi_7=\chi_8,$$
$$\chi_9=\chi_{24},\ \chi_{10}=\chi_{23},\ldots,\chi_{16}=\chi_{17}.$$

The reduction is as if the first octet of type III is reduced to a quadruplet, while the second and third octets have coincided. With the identification above all operators can be viewed from Table I where setting $\nu=0$ is equivalent to a blank space since there is no operator in this case. (The reduction can be viewed also as if the "vanishing" operator has merged together the spaces it intertwined.) Altogether there are 30 operators (five of each type), 14 of which act inside the quadruplet containing χ_1 (two operators) and the octet containing χ_9 (12 operators), while 16 are connecting the quadruplet and the octet (eight from each). We also note that in four cases the operators $d\,(w_k)$ $k=2,4,5,6$ act inside the octet and (as in Type IIa) their action coincides with $\mathscr{A}(s_2)$ ($k=2$, 5) and $\mathscr{A}(s_4)$ ($k=4$, 6).

Type IIIb (12-plet): This would be obtained from III by setting $n=0$

$$\chi_1=\chi_{19}=[2l+2,\epsilon,2l+\nu+2,\nu],\quad \epsilon=(2l+\nu+1)_{(2)}, \tag{6.24}$$

$$\chi_2=\chi_{20},\ldots,\chi_6=\chi_{24},\ \chi_7=\chi_{17},\ \chi_8=\chi_{18},\ \chi_9=\chi_{14},$$
$$\chi_{10}=\chi_{13},\ \chi_{11}=\chi_{12},\ \chi_{15}=\chi_{16}.$$

Here the reduction is as if the first and the third octets have coincided and the second is reduced to a quadruplet. It has many common features with type IIIa. The main difference is in the arrangement of the operators. For instance, χ_1, which supports no operators, is now in the octet, while in IIIa it was in the quadruplet. However, the same type would be obtained from III if we set $n=2l+2$.

Type IIIc (*sextet*): This would be obtained from III by

TABLE I. Action of the differential intertwining operators in the case of type III (24-plet) reducible representations. Column 0 gives the numbers k of the representations χ_k in the 24-plet. Columns 1a–6a give the numbers of the representations of the 24-plet to which the corresponding operators are mapping, and the degrees p of these operators are given in columns 1b–6b. Whenever an operator of a given type is not defined there are blank spaces in columns 1a, b–6a, b.

0 from	1a to	1b by $d^p(w_1)$	2a to	2b by $d^p(w_2)$	3a to	3b by $d^p(w_3)$	4a to	4b by $d^p(w_4)$	5a to	5b by $d^p(w_5)$	6a to	6b by $d^p(w_6)$
1												
2	1	$2l+\nu+2$	20	n			10	$n+\nu$	21	$2l+2-n$	15	$2l+2-n+\nu$
3	4	ν					14	$2l+2-n+\nu$			11	$n+\nu$
4							19	$2l+2-n$			22	n
5	6	$2l+\nu+2$	13	$n+\nu$	2	ν	23	n	12	$2l+2-n+\nu$	18	$2l+2-n$
6					1	ν						
7			24	$2l+2-n$	4	$2l+\nu+2$			17	n		
8	7	ν	9	$2l+2-n+\nu$	3	$2l+\nu+2$			16	$n+\nu$		
9					14	n			1	$n+\nu$	24	ν
10	9	$2l+2-n$			13	n			22	$2l+2+\nu$	7	$2l+2+\nu-n$
11			22	ν			6	$2l+\nu+2-n$				
12	11	n	4	$n+\nu$			24	$2l+2+\nu$				
13	14	$2l+2-n$							4	$2l+2-n+\nu$	17	$2l+2+\nu$
14									19	ν	6	$\nu+n$
15	16	n	19	$2l+\nu+2$	12	$2l+2-n$	7	$\nu+n$				
16			1	$2l+\nu+2-n$	11	$2l+2-n$	17	ν				
17			6	$2l+2-n$	22	$2l+2-n+\nu$						
18	17	$\nu+n$	14	$2l+2+\nu$	21	$2l+2-n+\nu$	8	n	15	ν		
19											1	n
20	19	$2l+\nu+2-n$					13	ν	3	$2l+2-n$	16	$2l+2+\nu$
21	22	$\nu+n$	3	n			9	$2l+2+\nu$			12	ν
22							1	$2l+2-n$				
23	24	$2l+\nu+2-n$	10	ν	20	$n+\nu$			11	$2l+\nu+2$	8	$2l+2-n$
24					19	$n+\nu$			6	n		

setting $n=0$, $l=-1$, $(\epsilon=(\nu+1)_{(2)})$,

$$\chi_1=\chi_4=\chi_{19}=\chi_{22}=[0,\epsilon,\nu,\nu], \tag{6.25a}$$

$$\chi_2=\chi_3=\chi_{20}=\chi_{21}=[0,\epsilon,-\nu,\nu], \tag{6.25b}$$

$$\chi_5=\chi_8=\chi_{18}=\chi_{23}=[0,\epsilon,-\nu,-\nu], \tag{6.25c}$$

$$\chi_6=\chi_7=\chi_{17}=\chi_{24}=[0,\epsilon,\nu,-\nu], \tag{6.25d}$$

$$\chi_9=\chi_{10}=\chi_{13}=\chi_{14}=[2\nu,(\epsilon+\nu)_{(2)},0,0], \tag{6.25e}$$

$$\chi_{11}=\chi_{12}=\chi_{15}=\chi_{16}=[-2\nu,(\epsilon+\nu)_{(2)},0,0], \tag{6.25f}$$

This could be viewed also as a subtype of IIIb where the octet has reduced to a quadruplet (6.25a)–(6.25d) and the quadruplet to a doublet. Twelve operators remain (two of each type), of which four act inside the quadruplet and eight connect the quadruplet and the doublet (four in each direction).

Type IIId (*quadruplet*): This would be obtained from III by setting $n=\nu=0$

$$\chi_1=\chi_6=\chi_9=\chi_{14}=\chi_{19}=\chi_{24}=[2l+2,\epsilon,2l+2,0],$$
$$\epsilon=(2l+1)_{(2)}, \tag{6.26}$$

and χ_2, χ_3, and χ_7 are the relevant spaces different from χ_1. There remain six operators acting inside this reduced octet (one of each type).

This completes the classification of the reducible ER according to the way they are grouped in multiplets. On the composition factors content of the reducible ER we shall now announce only an upper bound on the number of composition factors $-\kappa$. This number cannot exceed the number of all intersections and unions of the kernels and images of all intertwining operators acting from and to a fixed representation. Thus we have

$$\begin{aligned}&\text{type I},\ \ \kappa\leqslant 4;\quad \text{type Ia},\ \ \kappa\leqslant 6;\quad \text{type Ib},\ \ \kappa\leqslant 5;\\ &\text{type II},\ \ \kappa\leqslant 7;\quad \text{type IIa},\ \ \kappa=2;\\ &\text{type III},\ \ \kappa\leqslant 111;\quad \text{type IIIa},\ \ \kappa\leqslant 85;\\ &\text{type IIIb},\ \ \kappa\leqslant 78.\end{aligned} \tag{6.27}$$

Remark: All finite-dimensional irreducible representations of SU(2, 2) appear as the irreducible subrepresentations of the type III representations $\chi_{\bar{l\nu n}}$ [cf. (6.22e)]. The dimension of the finite-dimensional representation $E_{l\nu n}\subset\chi_{\bar{l\nu n}}$ [as in the SU*(4) case] is

$$\dim E_{l\nu n}=\nu n(\nu+n)(2l+\nu+2)$$
$$\times(2l+\nu+2-n)(2l+2-n)/12. \tag{6.28}$$

ACKNOWLEDGMENTS

The author would like to thank Professor Abdus Salam, the International Atomic Energy Agency, and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste. He wishes to thank Professor I. T. Todorov and Professor N. S. Craigie for the support of this program and Dr. V. B. Petkova for illuminating discussions.

APPENDIX: THE IWASAWA DECOMPOSITION IN THE FORM $\tilde{N}_0A_0K$

Another form of the Iwasawa decomposition (5.1) is

$$g=\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}=\tilde{n}ak, \tag{A1}$$

where $\tilde{n}\in\tilde{N}_0$, and $a\in A_0$ shall be parametrized as in (5.34) and $k\in K$. Then for the parameters in $\tilde{n}$, a, k we obtain

$$|a| = e^{s+t} = 1/\sqrt{\det \underset{\sim}{E}}, \quad e^r = e^{(s-t)/2} = (\det \underset{\sim}{E})^{1/4}/\sqrt{\underset{\sim}{E}_1}; \tag{A2a}$$

$$\underset{\sim}{E} \equiv \underset{\sim}{\gamma}\underset{\sim}{\gamma}^+ + \underset{\sim}{\delta}\underset{\sim}{\delta}^+ \quad (\underset{\sim}{E}_1 > 0, \quad \det \underset{\sim}{E} > 0); \tag{A2b}$$

$$z = -\underset{\sim}{E}_+/\underset{\sim}{E}_1, \tag{A2c}$$

$$\underset{\sim}{p} = \hat{a}x_z^+\underset{\sim}{\delta}, \quad \underset{\sim}{q} = \hat{a}x_z^+\underset{\sim}{\gamma}; \tag{A2d}$$

$$\hat{a}x_z^+ = \frac{1}{\sqrt{\underset{\sim}{E}_1 \det \underset{\sim}{E}}}\begin{pmatrix} \sqrt{\det \underset{\sim}{E}} & 0 \\ -\underset{\sim}{E}_+ & \underset{\sim}{E}_1 \end{pmatrix} \quad x_z\hat{a}^2x_z^+ = \underset{\sim}{E}^{-1};$$

$$i\underset{\sim}{x} = (\underset{\sim}{\beta} - \underset{\sim}{E}^{-1}\underset{\sim}{\gamma})\underset{\sim}{\delta}^{-1} \quad (\det \underset{\sim}{\delta} \neq 0), \tag{A2e}$$

$$i\underset{\sim}{x} = (\underset{\sim}{\alpha} - \underset{\sim}{E}^{-1}\underset{\sim}{\delta})\underset{\sim}{\gamma}^{-1} \quad (\det \underset{\sim}{\gamma} \neq 0).$$

Note that $\underset{\sim}{E}_1 > 0$, $\det \underset{\sim}{E} > 0$ always. We shall also give the relationship of this Iwasawa decomposition with the Bruhat decomposition. Let now $\det \underset{\sim}{\delta} \neq 0 \neq \underset{\sim}{\delta}$ and

$$g = \tilde{n}_I a_I k = \tilde{n}_B n a_B m, \tag{A3}$$

where $\tilde{n}_I, \tilde{n}_B \in \tilde{N}_0$, $n \in N_0$, $a_I, a_B \in A_0$, $m \in M_0$, and $k \in K$. [Parameters of $\tilde{n}_I, \tilde{n}_B, n, a_I, a_B$ are as in (5.34).] Then we obtain

$$x_{Iz}\hat{a}_I^2x_{Iz}^+ = x_{Bz}(\underset{\sim}{\Delta}^{-1} + \tilde{b}\underset{\sim}{\Delta}\tilde{b})^{-1}x_{Bz}^+, \tag{A4a}$$

$$\underset{\sim}{\Delta} \equiv b_w\hat{a}_B^2b_w^+,$$

$$\underset{\sim}{p} = \hat{a}_Ix_{Iz}^+(x_{zB}b_w\hat{a}_B\hat{r}\sigma_3^N)^{+-1}, \tag{A4b}$$

$$\underset{\sim}{q} = \hat{a}_Ix_{Iz}^+x_{Bz}^{+-1}i\tilde{b}b_w\hat{a}_B\hat{r}\sigma_3^N,$$

$$\underset{\sim}{x}_I = \underset{\sim}{x}_B - x_{Bz}[(\underset{\sim}{\Delta}\tilde{b}\underset{\sim}{\Delta})^{-1} + \tilde{b}]^{-1}x_{Bz}^+ \quad (\det \tilde{b} \neq 0), \tag{A4c}$$

$$x_I = \underset{\sim}{x}_B - x_{Bz}[\underset{\sim}{\Delta}^{-1} + \tilde{b}\underset{\sim}{\Delta}\tilde{b}]^{-1}\tilde{b}\Delta x_{Bz}^+ \quad (\det \tilde{b} \sim),$$

where in (A4b) $\hat{a}_Ix_{Iz}^+$ are understood as obtained from (A4a). The inverse formulas are

$$x_{Bz}b_w\hat{a}_{B-}\hat{r} = x_{Iz}\hat{a}_{I-}\underset{\sim}{p}^{+-1}\sigma_3^N|\det \underset{\sim}{p}|, \tag{A5a}$$

$$|a_B| = |a_I|/|\det \underset{\sim}{p}|, \quad (-1)^N = \operatorname{sgn} \det \underset{\sim}{p} \quad (\det \underset{\sim}{p} \neq 0), \tag{A5b}$$

$$i\tilde{b} = x_{Bz}^+(x_{Iz}\hat{a}_I)^{+-1}\underset{\sim}{q}\underset{\sim}{p}^+(x_{Iz}\hat{a}_I)^{-1}x_{Bz}, \tag{A5c}$$

$$i\underset{\sim}{x}_B = i\underset{\sim}{x}_I + x_{Iz}\hat{a}_I\underset{\sim}{q}\underset{\sim}{p}^{-1}(x_{zI}\hat{a}_I)^+ \tag{A5d}$$

[x_{Bz} in (A5c) is from (A5a)].

Consider next the P_2-Bruhat decomposition, namely let in (A3)$\tilde{n}_B \in \tilde{N}_2$, $n \in N_2$, $a_B \in A_2$, $m \in M_2$ with parametrization as in (4.9). Then we obtain

$$x_{Iz}\hat{a}_I^2x_{Iz}^+ = (\underset{\sim}{\Delta}_2^{-1} + \tilde{b}\underset{\sim}{\Delta}_2\tilde{b})^{-1} = \underset{\sim}{\Delta}_2[1 + (\tilde{b}\underset{\sim}{\Delta}_2)^2]^{-1}, \tag{A6a}$$

$$\underset{\sim}{\Delta}_2 \equiv |a_B|ll^+,$$

$$\underset{\sim}{p} = \hat{a}_Ix_{Iz}^+l^{+-1}\sigma_3^N/\sqrt{|a_B|}, \quad \underset{\sim}{q} = \hat{a}_Ix_{Iz}^+i\tilde{b}l\sigma_3^N\sqrt{|a_B|}, \tag{A6b}$$

$$\underset{\sim}{x}_I = \underset{\sim}{x}_B - [(\underset{\sim}{\Delta}_2\tilde{b}\underset{\sim}{\Delta}_2)^{-1} + \tilde{b}]^{-1} \quad (\det \tilde{b} \neq 0), \tag{A6c}$$

$$\underset{\sim}{x}_I = \underset{\sim}{x}_B - [\underset{\sim}{\Delta}_2^{-1} + \tilde{b}\underset{\sim}{\Delta}_2\tilde{b}]^{-1}\tilde{b}\underset{\sim}{\Delta}_2 \quad (\det \tilde{b} \sim)$$

[$\hat{a}_Ix_{Iz}^+$ in (A6b) is from (A6a)]. Note $\tilde{b} = \tilde{b}_2$ in (A6) is connected with $\tilde{b} = \tilde{b}_0$ in (A4.36). The inverse formulas are (A5b) and (A5d) [recall (5.30a) and (5.36)] and

$$l = x_{Iz}\hat{a}_-\underset{\sim}{p}^{+-1}\sigma_3^N|\det \underset{\sim}{p}|, \tag{A7a}$$

$$i\tilde{b} = (x_{Iz}\hat{a}_I)^{+-1}\underset{\sim}{q}\underset{\sim}{p}^+(x_{Iz}a_I)^{-1}. \tag{A7b}$$

[1] I. T. Todorov, M. C. Mintchev, and V. B. Petkova, "Conformal Invariance in OFT," Scuola Normale Superiore, Pisa, 1978.
[2] V. B. Petkova and I. T. Todorov, SISSA, Trieste, preprint, 14/83/E.P., 1983.
[3] Harish-Chandra, Trans. Am. Math. Soc. **76**, 26 (1954).
[4] G. Warner, *Harmonic Analysis on Semi-Simple Lie Groups, I, II* (Springer-Verlag, Berlin, 1972).
[5] J. Lepowski, Trans. Am. Math. Soc. **176**, 1 (1973).
[6] W. Casselman, "The differential equations satisfied by matrix coefficients," Princeton preprint, 1975; D. Miličić, "Notes on asymptotics of admissible representations of semi-simple Lie groups," Princeton preprint, 1976.
[7] R. Langlands, "On the classification of irreducible representations of real algebraic groups," mimeographic notes, IAS, Princeton, 1973.
[8] A. W. Knapp and G. J. Zuckermann, "Classification theorems for representations of semi-simple groups," in *Non-Commutative Harmonic Analysis, Lecture Notes in Mathematics*, Vol. 587 (Springer-Verlag, Berlin, 1977), pp. 138–159.
[9] A. W. Knapp and B. Speh, J. Funct. Anal. **45**, 41 (1982).
[10] V. K. Dobrev, G. Mack, V. B. Petkova, S. G. Petrov, and I. T. Todorov, *Lecture Notes in Physics*, Vol. 63 (Springer-Verlag, Berlin, 1977).
[11] V. K. Dobrev, V. B. Petkova, S. G. Petrova, and I. T. Todorov, Phys. Rev. D **13**, 887 (1976).
[12] V. K. Dobrev and V. B. Petkova, Rep. Math. Phys. **13**, 223 (1978).
[13] G. Mack and I. T. Todorov, J. Math. Phys. **10**, 2078 (1969).
[14] G. Mack, Commun. Math. Phys. **55**, 1 (1977).
[15] W. Ruhl and B. C. Yunn, Fortschr. Phys. **25**, 83 (1978).
[16] S. M. Paneitz and I. E. Segal, J. Funct. Anal. **47**, 78 (1982).
[17] V. B. Petkova and C. M. Sotkov, SISSA, Trieste, preprint, 15/83/E.P., 1983 and Lett. Math. Phys. **8**, 217 (1984).
[18] H. P. Jakobsen, Invent. Math. **62**, 67 (1980).
[19] A. Angelopoulos, Commun. Math. Phys. **89**, 41 (1983).
[20] R. O. Wells, Jr., Bull. Am. Math. Soc. (New Series) **1**, 296 (1979).
[21] S. Helgason, *Differential Geometry and Symmetric Spaces* (Academic, New York, 1962).
[22] K. Iawasawa, Ann. Math. **50**, 507 (1949).
[23] D. P. Zhelobenko, *Harmonic Analysis on Complex Semi-Simple Lie Groups* (Nauka, Moscow, 1974) (in Russian).
[24] I. M. Gel'fand and M. A. Naimark, Izv. AN SSSR, Math. **11**, 411 (1947); F. Bruhat, Bull. Soc. Math. France **84**, 97 (1956).
[25] I. M. Gel'fand, M. I. Graev, and N. Ya. Vilenkin, *Generalized Functions* (Academic, New York, 1966), Vol. 5.
[26] A. W. Knapp and E. M. Stein, Ann. Math. **93**, 489 (1971); Invent. Math. **60**, 9 (1980).
[27] B. Speh and D. A. Vogan, Jr., Acta Math. **145**, 227 (1980).
[28] B. Binegar, C. Fronsdal, and W. Heidenreich, J. Math. Phys. **24**, 2828 (1983).
[29] V. K. Dobrev, V. B. Petkova, S. G. Petrova, and I. T. Todorov, "On the exceptional integer points in the representation space of the pseudo-orthogonal group O(2h + 1, 1)," ICTP, Trieste, preprint IC/75/1, 1975.

Markov-type Lie groups in GL(n,R)

Joseph E. Johnson
Department of Physics, University of South Carolina, Columbia, South Carolina 29208

(Received 21 March 1984; accepted for publication 24 August 1984)

The general linear group GL(n,R) is decomposed into a Markov-type Lie group and an abelian scale group. The Markov-type Lie group basis is shown to generate all singly stochastic matrices which are continuously connected to the identity when non-negative parameters are used. A basis is found which shows that it in turn contains a Lie subgroup which corresponds to doubly stochastic matrices, which basis, over the complex field, is shown to give the symmetric group for certain discrete values of the complex parameters. The basis of the Markov algebra is shown to give the negative of the corresponding M-matrices with property "C" (for non-negative combinations). These stochastic Lie groups are shown to be isomorphic to the affine group and the general linear group in one less dimension. The basis generates transformations with a natural interpretation for physical applications.

I. INTRODUCTION

There is extensive literature[1–3] on the general linear group in n dimensions over the real (or complex) field, GL(n,R), which explores various subgroup chains and their representations. Usually these decompositions begin by removing the Lie algebra generator I, leaving the nonsingular unimodular group SL(n,R). Further restrictions requiring the invariance of some bilinear form leads to subsequent decomposition and in particular the determination of all simple Lie algebras. This paper will explore an alternative decomposition of GL(n,R) requiring the invariance of a linear form and resulting in a solvable (not semisimple) Lie group chain with Markov-type Lie groups and their associated Lie algebras down to the symmetric group. Butler and King[4] have extensively explored the symmetric group as a subgroup of the general linear group and have introduced two ideas which we explore more fully: (1) the invariance of a linear form in GL(n,R) and (2) the concept of the symmetric group S_n as a subgroup of GL(n,R).

Requiring the invariance of a linear form

$$\sum x_i$$

is closely related to singly (and doubly) stochastic processes which leave Σx_i invariant and $x_i \geqslant 0$. First studied by Markov[5] in 1907, a singly (row) stochastic or Markov process is a linear transformation $M_{ij} \geqslant 0$ with

$$\sum_i M_{ij} = 1, \tag{1.1}$$

which can be thought of as transforming a vector of probabilities (or occupation numbers) $x_i \geqslant 0$ into a new set $x_i' = M_{ij}x_j$ and is also doubly stochastic if

$$\sum_j M_{ij} = 1. \tag{1.2}$$

Markov processes only form a semigroup since, in general, they do not process an inverse.[6]

In Sec. II we will study the decomposition of GL(2,R) into a Markov-type Lie group and an abelian scale group. Specifically it will be shown that all Markov processes continuously connected to the identity are all generated by a certain basis for its Lie algebra with non-negative linear combinations. In Sec. III we will extend these ideas to n dimensions and discuss a connection to S_n illustrating that the permutations are Markov processes which can be reached from the identity with the same Lie algebra over the complex field.

In Sec. IV we briefly discuss the invariance of indefinite linear forms $\Sigma x_i - \Sigma y_j$. Section V is a general discussion of properties of the Markov Lie group. In particular it is shown that all analytic functions of the basis are linear and thus no Casimir operators exist. In Sec. VI a basis for a doubly stochastic Lie algebra is obtained and related to the symmetric group in Sec. VII. A close connection between the Markov Lie algebra and the M-matrices with property "C" is established in Sec. VIII with general conclusions following in Sec. IX.

II. NOTATION AND DEFINITION OF M(n,R) IN TWO DIMENSIONS

We define the "Markov" Lie group $M(n,R)$ to be the subgroup of GL(n,R) which preserves

$$\sum x_i,$$

where x_i are the vector components $i = 1 \cdots n$ acted upon by the $n \times n$ representation of GL(n,R). We define the vectors $\langle 1|$ and $|1\rangle$ to be row and column vectors, respectively, with all components equal to 1. It follows that $\langle 1|M|x\rangle = \langle 1|x\rangle$ is equivalent to

$$\sum_i M_{ij} = 1, \tag{2.1}$$

for all j. This is equivalent to the preservation of a linear rather than a bilinear form. The subset consisting of all $M_{ij} \geqslant 0$ would not be useful unless the M_{ij} are smoothly connected in the group space and have a useful form as we now show.

The infinitesimal transformation which takes a positive fraction $0 < \epsilon < 1$ of a component and adds it to the other component will preserve the sum and will always be positive when acting upon non-negative components. It also has the natural interpretation of a transition probability for a time ϵ. It can be written in two dimensions as

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = 1 + \epsilon \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

or

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = 1 + \epsilon \begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{2.2}$$

transferring the fraction ϵx_2 to x_1 and ϵx_1 to x_2, respectively. Defining $M(2,R)$ in terms of the basis

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

and

$$L^{21} = \begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix} \tag{2.3}$$

one verifies that

$$e^{\lambda L^{12}} = \begin{pmatrix} 1 & 1 - e^{-\lambda} \\ 0 & e^{-\lambda} \end{pmatrix}$$

and

$$e^{\lambda L^{21}} = \begin{pmatrix} e^{-\lambda} & 0 \\ 1 - e^{-\lambda} & 1 \end{pmatrix} \tag{2.4}$$

and that $\langle 1 | e^{\lambda L^{12}} = \langle 1 | e^{\lambda L^{12}} = \langle 1 |$ as required (λ real). (2.5)

One also verifies that $[L^{12}, L^{21}] = +L^{12} - L^{21}$, giving the structure constants. Closure of the group can be seen from closure of the L^{12} and L^{21} commutation rules or from sequences of infinitesimal transformations which individually and thus collectively preserve $\langle 1 | x \rangle$. Thus in two dimensions the most general form of $M(2,R)$ is

$$e^{\lambda \cdot L} = e^{\lambda_{12} L^{12} + \lambda_{21} L^{21}} \tag{2.6}$$

with the group inverse $e^{-\lambda \cdot L}$ and group unit with $\lambda_{ij} = 0$. GL(n,R) itself has the additional basis elements

$$L^{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$L^{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

no combination of which preserves $\langle 1 | x \rangle$. The Lie group $M(2,R)$ thus satisfies the requirement of preserving the linear form $\langle 1 | x \rangle$, but as λ ranges over the reals there is an unphysical region when either $\lambda_{i \neq j} < 0$, which will not give a Markov matrix, as well as a physical region with both $\lambda_{i \neq j} \geqslant 0$, which always gives an acceptable Markov matrix. Like GL(n,R), $M(n,R)$ is noncompact. The limit points at $\lambda = \infty$ give the singular transformations $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$ with λ_{12} and λ_{21}, respectively.

III. GENERALIZATION TO n DIMENSIONS

These results are easily generalized to n dimensions where we define

$$L^{ij}_{kl} \equiv \delta_{ik} \delta_{jl} - \delta_{jk} \delta_{jl} \tag{3.1}$$

for $i \neq j$ to be the kl element of the L^{ij} linear operator. Similarly $(L^{ii})_{kl} \equiv \delta_{ik} \delta_{il}$. The $(n^2 - n) L^{ij}$ matrices and the $(n)\, L^{ii}$ matrices form a basis for the Lie algebra which generates GL(n,R). This can be seen by forming the n^2 combinations which possess a 1 at only one position in the matrix with zeros elsewhere. We define $M(n,R)$ and $A(n,R)$ to be the matrices generated by the L^{ij} and the L^{ii}, respectively. Thus GL$(n,R) = A(n,R) \oplus M(n,R)$ for their respective Lie algebras.

That the L^{ii} generate an abelian subgroup of order $n, A(n,R)$, of GL(n,R) follows immediately from the general form

$$e^{\Sigma \lambda_{ii} L^{ii}} = \begin{pmatrix} e^{\lambda_{11}} & & & \\ & e^{\lambda_{22}} & & \\ & & \vdots & \\ & & & e^{\lambda_{nn}} \end{pmatrix} \tag{3.2}$$

which scales the ith coordinate by $e^{\lambda ii}$. It is closed, noncompact, has the inverse

$$e^{-\lambda_{ii} L^{ii}} \tag{3.3}$$

and a unit defined by $\lambda_{ii} = 0$. A unimodular subalgebra is obtained by redefining the basis as

$$I = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & & \\ & -1 & \\ & & 0 \end{pmatrix},$$

$$H_3 = \begin{pmatrix} 1 & & \\ & 1 & \\ & & -2 \end{pmatrix}, \quad H_4 = \cdots, \tag{3.4}$$

with H_i as a diagonal traceless basis with $i = 2,3,...,n$.

The L^{ij} $(i \neq j)$ in three dimensions take the form

$$L^{12} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad L^{23} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix},$$

$$L^{31} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad L^{21} = \begin{pmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{3.5}$$

$$L^{32} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad L^{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

which follows from writing the infinitesimal transformation which subtracts ϵx_j from the jth component and adds ϵx_j to the ith component. Thus these infinitesimal transformations preserve $\langle 1 | x \rangle$ individually and collectively and thus any group element

$$e^{\lambda_{ij} L^{ij}}$$

compounded from sequences of infinitesimal transformations also preserves $\langle 1 | x \rangle$. Conversely all linear transformations in GL(n,R) which preserve $\langle 1 | x \rangle$ are included in the basis since $\langle 1 | (1 + \Sigma \epsilon_{ij} L^{ij}) = \langle 1 |$ implies that $\Sigma \epsilon_{ij} L^{ij} = 0$ over a column and the $n - 1$ different linear combinations using L^{ij} for a fixed j spans all such possible combinations. Consequently $M(n,R)$ contains all those and only those transformations in GL(n,R) which preserve $\langle 1 | x \rangle$. Furthermore it is both necessary and sufficient that $\lambda_{ij} \geqslant 0$ for all i and j in order to guarantee that any vector with all nonnegative components is transformed into a vector with nonnegative components. This can be seen by looking at the most general infinitesimal transformation which is seen to be non-negative and thus all products of these are also. Thus for real λ_{ij}, all Markov transformations in GL(n,R) continuously connected to the identity are those elements in $M(n,R)$ formed with $\lambda_{ij} \geqslant 0$. The closure of $M(n,R)$ can be shown from the closure of the commutators of the generating Lie algebra

$$\sum_i \sum_j (L_{ij}^{lm} L_{jk}^{rs} - \Sigma L_{ij}^{rs} L_{jk}^{lm}) = 0, \tag{3.6}$$

which demonstrates that the commutator must be a combination of matrices with a zero row sum for each column. Thus the commutator is a linear combination of elements of the algebra. Also the product of two elements of $M(n,R)$ (with unit row sums) is

$$\sum_i \sum_j M_{ij}^a M_{jk}^b = \sum_j M_{jk} = 1 \tag{3.7}$$

and thus is a member of $M(n,R)$. The unit operator is produced with $\lambda_{ij} = 0$ and the inverse with $-\lambda_{ij}$. Thus $M(n,R)$ is a Lie group with L^{ij} $(i \neq j)$ forming the basis of its Lie algebra. (Antisymmetry and the Jacobi identity follow automatically from a matrix definition.)

Although we found all Markov matrices in GL(n,R) with real λ_{ij}, one can ask if there are acceptable real Markov matrices arising from complex λ_{ij}. It is easy to verify that none are in the neighborhood of the identity. However consider

$$e^{\lambda(L^{12}+L^{21})} = \frac{1}{2}\begin{pmatrix} 1+e^{-2\lambda} & 1-e^{-2\lambda} \\ 1-e^{-2\lambda} & 1+e^{-2\lambda} \end{pmatrix}, \tag{3.8}$$

for imaginary λ, which give real matrices. One can obtain $e^{-2\lambda} = -1$ with $-2\lambda = \pm i n_0 \pi$ or

$$\lambda = n_0 i\pi/2, \tag{3.9}$$

where n_0 is an odd integer. This gives

$$e^{\lambda L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \text{for } n_0 = 1$$

which is a permutation (transposition) of the two variables. Thus using these discrete imaginary values for λ with $(L^{ij} + L^{ji})$ one obtains the transpositions between any two pairs of variables and, by multiplication of these, any permutation. Thus the permutation (symmetric) group is contained in $M(n,C)$ for certain discrete complex values of the group parameters (that a transposition is continuously connected to the identity only with complex parameters, is easily proven by diagonalizing the transposition matrix).

IV. TRANSFORMATIONS PRESERVING $\Sigma x_i - \Sigma y_i$

Beginning with an example in two dimensions, we can ask for transformations in GL(n,R) which preserve $x - y$. The above results on the Markov matrices suggest infinitesimal transformations which add or subtract a fraction of either coordinate to the other. Thus we define

$$L^{-12} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = L^{12} + 2L^{22}, \qquad L^{-21} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = L^{21} + 2L^{11}, \tag{4.1}$$

which give

$$e^{\lambda L^{-12}} = \begin{pmatrix} 1 & -1+e^{+\lambda} \\ 0 & e^{+\lambda} \end{pmatrix}, \qquad e^{\lambda L^{-21}} = \begin{pmatrix} e^{+\lambda} & 0 \\ -1+e^{+\lambda} & 1 \end{pmatrix}, \tag{4.2}$$

respectively. In n dimensions these matrices give the correct prescription for the connection between the positive definite and negative definite subspaces. The invariant form can be written as $\langle 1|\eta|x\rangle$ where η is a metric which carries the sign for the negative definite portions of the space. We will refer to these transformations as indefinite Markov transformations $M(r+s,R)$, where r and s are the dimensions of the positive definite and negative definite subspaces. The $M(r+s,R)$ transformations also form a Lie group and give physically acceptable vectors $(x_i \geqslant 0)$ when any element $\lambda_{ij} \geqslant 0$ acts upon a physically acceptable vector.

V. GENERAL PROPERTIES OF *M(n,R)*

Geometrically, $M(n,R)$ can be viewed as giving all nonsingular linear transformations on the hyperplane perpendicular to the vector $\langle 1| = (1,1,1,...,1)$ since $\langle 1|M = \langle 1|$ or equivalently since $\Sigma x_i = \text{const}$ is the equation for the hyperplane and invariant. For non-negative λ_{ij}, $e^{\lambda \cdot L}$ maps the positive quadrant into itself. In fact, from an arbitrary point $x_i \geqslant 0$ any other point $x_i \geqslant 0$ can be reached with $M(n,R)$. A particular λ_{ij} determines the fraction of the jth sector which is added to the ith sector. If y_i are defined by $y_i^2 = x_i$ then $M(n,R)$ maps the sphere $\Sigma y_i^2 = \text{const}$ into itself for $\lambda_{ij} \geqslant 0$ and thus behaves like a nonlinear representation of the rotation group but without an inverse. Likewise in two dimensions, $M(1+1,R)$ preserves $y_0^2 - y_1^2$ and thus behaves like a nonlinear representation of the Lorentz group. The invariant hyperplane of $M(r+s,R)$ is

$$\sum_{i=1}^{r} x_i - \sum_{i=r+1}^{r+s} x_i = \text{const.} \tag{5.1}$$

All of the physical portion of the space can be covered with $M(r+s,R)$ from the initial state with $x_i^{\max} = c$, $x_{j \neq i} = 0$.

The group $M(n,R)$ is not unimodular (determinant $\neq 1$) since the basis of its algebra, L^{ij}, is not traceless. Consequently $M(n,R)$ is not contained in SL(n,C). By evaluating the Killing form, $g_{ij} = c_{ik}^{l} C_{jl}^{k}$ in two dimensions one obtains

$$g = |g_{ih}| = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0. \tag{5.2}$$

Since a Lie algebra is semisimple if and only if $g \neq 0$, it follows that $M(2,R)$ is not semisimple. Defining $L = L^{12} - L^{21}$, one can show $[L,L^{12}] = L = [L,L^{21}]$ and thus L forms an invariant subalgebra or ideal. Consequently S is not simple. $M(n,R)$ is also noncompact since the parameter space is unbounded.

Generally one can prove that $M(n,R)$ is isomorphic to the affine group in $n-1$ dimensions (consisting of the general linear group and translations). This follows from the result above that $M(n,R)$ consists of linear transformations in GL(n,R) which are restricted to transformations in the hyperplane perpendicular to $|1\rangle$, which is a space of dimension $(n-1)$. The actual isomorphism can be implemented by a coordinate transformation, R, which rotates the x_n axis into the vector $|1\rangle$ after which all of the $n^2 - n$ linear transformations which were previously in the hyperplane now become linear transformations on the subspace $x_1 x_2 \cdots x_{n-1}$, leaving the x_n axis invariant. The transformed $M(n,R)$ matrices then take the customary form for the affine group:

$$\begin{pmatrix} \mathrm{GL}(n-1,R) & T(n-1) \\ 0 & 1 \end{pmatrix}.$$

Thus all properties and representations of $M(n,R)$ are those of the affine group in $(n-1)$ dimensions.

For semisimple Lie groups, the irreducible representations are classified by the spectra of Casimir operators[7]

$$I_n = C^{\beta_2}_{\alpha_1\beta_1} C^{\beta_3}_{\alpha_2\beta_2} \cdots C^{\beta_1}_{\alpha_n\beta_n} L^{\alpha_1} L^{\alpha_2} \cdots L^{\alpha_3}, \tag{5.3}$$

which commute with all the elements of the algebra. Normally I_n is defined only for semisimple algebras but an interesting nonexistence proof is possible for $M(n,R)$ for representations of the form (3.1): For two elements L^a and L^b, in a representation of arbitrary order, we have

$$\sum_i \sum_j L^a_{ij} L^b_{jk} = \sum_j \sum_i L^a_{ij} L^b_{jk} = 0, \tag{5.4}$$

showing that the product of two matrices with

$$\sum_i L_{ij} = 0 \tag{5.5}$$

is again a matrix of this type. But since the L^{ij} are a complete basis of all such matrices it follows that any product is expressible as a linear combination:

$$L^{ij} L^{rs} = \Sigma \lambda_{lm} L^{lm}. \tag{5.6}$$

Consequently any analytic function of the L^{ij} is expressible as a linear combination of the L^{ij} and thus no operator like the Casimir operators exist for $M(n,R)$ for representations of the form (3.1). The generality of this proof rests upon the fact that the L generate an algebra of arbitrary order n. In fact the general group element

$$M = e^{\lambda \cdot L} = 1 + \lambda \cdot L + (1/2!)(\lambda \cdot L)^2 + \cdots \tag{5.7}$$

must therefore be repressible as $M = 1 + a_{ij}(\lambda) L^{ij}$ where the a_{ij} are functions of the λ_{ij} and must all satisfy $0 \leqslant a_{ij} \leqslant 1$.

It would be important to have a useful expression for the functions $a_{ij}(\lambda)$ as well as for the inverse functions because the $a_{ij}(\lambda)$ give the detailed connection between any particular Markov transformation and the element of the Lie algebra which generates it. In this paper we have only established existence and general properties of this connection.

VI. THE DOUBLY STOCHASTIC SUBGROUP

In certain applications of Markov or stochastic processes an additional requirement, $M|1\rangle = 1$, is imposed (in addition to $\langle 1|M = \langle 1|$). These transformations are termed doubly stochastic and have both unit row and unit column sums. We denote the collection of real nonsingular doubly stochastic transformations on an n-dimensional space as $M^D(n,R)$. By considering the infinitesimal transformations

$$M^D = 1 + \epsilon_{ij} L^{D_{ij}}, \tag{6.1}$$

it follows that it is necessary and sufficient that

$$\sum_m L^{D_{ij}}_{lm} = 0. \tag{6.2}$$

It can be seen that this imposes $n-1$ independent conditions on the L^{ij} since the nth row sum will follow from the zero column sums. That

$$L^{D_{ij}}$$

forms a Lie algebra follows from

$$\sum_r L^{D_{ij}}_{mn} L^{D_{kl}}_{nr} = 0, \tag{6.3}$$

thus the product of two elements must be a linear combination of a complete basis of L^D. That result is stronger than necessary for the commutator to be expressible in terms of the basis elements. As a consequence of the expression of the product as a member of the algebra it follows, as for singly stochastic processes, that any analytic function of

$$L^{D_{ij}}$$

is linearly expressible in terms of the L^D basis and thus is a member of the algebra. It also follows that

$$M^D = e^{\lambda \cdot L^D} = +\lambda L^D + \cdots = 1 + \alpha \cdot L^D, \tag{6.4}$$

where α is the linear combination is detemined by the λ. The proof follows from the products being expressible as elements of the algebra which gives a linear combination of elements which is an element of the algebra

$$\alpha \cdot L = \alpha_{ij} L^{D_{ij}}. \tag{6.5}$$

(Convergence is guaranteed for the exponential.) A basis for the Lie algebra L^D can be constructed by taking certain combinations of the L^{ij} generators which give vanishing row sums. The $(n^2 - n)L^{ij}$ must satisfy $n-1$ independent restrictions giving $(n-1)^2$ independent

$$L^{D_{ij}}.$$

We will absorb the $n-1$ constraints by using the $n-1$ elements on the diagonal just below the main diagonal. We define

$$L^{D_{ij}},$$

beginning with L^{ij}:

$$\begin{pmatrix} 0 & & & & & & \\ & 0 & & & 1 & & \\ & & 0 & & & & \\ & & & 0 & & & \\ & & & & -1 & & \\ & & & & & 0 & \\ & & & & & & 0 \end{pmatrix}, \tag{6.6}$$

where one observes that the row sums can always be made zero by adding the terms

$$L^{j,j-1} + L^{j-1,j-2} + \cdots + L^{i+1,i}, \tag{6.7}$$

which takes the form

$$\begin{pmatrix} 0 & & & & & & \\ & -1 & \cdots & & +1 & & \\ & 1 & -1 & & & & \\ & & 1 & -1 & & & \\ & & & 1 & -1 & & \\ & & & & 0 & 0 & \\ & & & & & & 0 \end{pmatrix} \tag{6.8}$$

(+ 1in i,j position) (case for $i<j$).

If $i>j$ then the sequence is

$$L^{D_{ij}} = L^{ij} + L^{j,j+1} + L^{j-1,j-2} \ldots L^{2,1} + L^{i+1,j} + L^{i+2,i+1} \ldots L^{n,n-1} + L^{n,n}, \quad (6.9)$$

which takes the form

$$\begin{pmatrix} -1 & & & \cdots & & 1 \\ 1 & -1 & & & & \\ & & 0 & & & \\ & & & 0 & & \vdots \\ & & & 0 & & \\ & +1 & & \cdots & -1 & \\ & & & & 1 & -1 \end{pmatrix}. \quad (6.10)$$

The basis for the Markov (singly stochastic process) could be taken as the

$$(n-1)^2 L^{D_{ij}},$$

along with the $(n-1)L^{i,i-1}$.

The Lie group $M^D(n,R)$ can be proved to be isomorphic to GL(n,R) by referring to the rotation R which transformed the x_n axis into the vector $|1\rangle$ in Sec. III. That transformation R showed that $M(n,R)$ was isomorphic to the affine group which contains the $(n-1)$-dimensional translation group on the remaining $x_1x_2\cdots x_{n-1}$ coordinates. A restriction of $M(n,R)$ to $M^D(n,R)$ imposes the requirement that the vector $|1\rangle$ is invariant ($n-1$ new constraints) and thus in the R transformed coordinates the origin must be invariant. The origin is left invariant by disallowing the translation portion of the affine group in $(n-1)$ dimensions, thus leaving the allowable transformations as GL$(n-1,R)$ which is thus isomorphic to $M^D(n,R)$.

VII. CONNECTION TO S_n

The symmetric (permutation) group S_n is nonsingular and thus is in GL(n,C) for certain values of the λ's in the generating Lie algebra. Furthermore, since S_n must permute each element into some new position, it must consist of exactly a single one in each row and each column (giving $n!$ possible matrices). Thus S_n must not only be Markov [in $M(n,C)$]; it must also be doubly stochastic [thus in $M^D(n,C)$]. Thus the $n!$ elements of S_n must be generated by some set of λ_{ij} in the Lie algebra $M^D(n,C)$. As $n! > (n-1)^2$ for all n it follows that some of the

$$L^{D_{ij}}$$

must generate several of the S_n elements. Furthermore, if

$$e^{\lambda \cdot L^D} \in S_n,$$

then, because of closure of S_n,

$$e^{m\lambda \cdot L^D} \in S_n, \quad (7.1)$$

for all integers m. Using the previous result that

$$e^{\lambda \cdot L^D} = 1 + \alpha \cdot L^D \quad (7.2)$$

and that S_n must be contained in

$$e^{\lambda \cdot L^D},$$

then it follows that S_n must be contained in

$$1 + \alpha \cdot L^D,$$

for selected values of α_{ij}. In particular when a single $\alpha_{ij} = 1$, others $= 0$, one obtains the permutations

$$\begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 0 & & 1 & & \\ & & 1 & 0 & & & \\ & & & 1 & 0 & & \\ & & & & 1 & 0 & \\ & & & & & & 1 \\ & & & & & & & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & & & & & & 1 \\ 1 & 0 & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & 1 & & & & & 0 \end{pmatrix}. \quad (7.3)$$

Thus

$$L^{D_{ij}}(i<j)$$

gives the permutation $x_1x_2(x_i\cdots x_j)\cdots x_n$ and

$$L^{D_{ij}}(i>j+1)$$

gives the permutation

$$x_1x_2\cdots x_j)x_{j+1}\cdots(x_i\cdots,$$

where terms outside the parentheses are unchanged and those inside are cyclically permuted to the right. The fundamental permutations can be simply represented by ordered pairs (i,j), which are defined $i,j = 1,\cdots n$ with $i \neq j$ and $i \neq j-1$. They are fundamental in the sense that there is a one-to-one correspondence between these $(n-1)^2$ permutations and the doubly stochastic Lie algebra basis which contains S_n.

VIII. CONNECTION TO M-MATRICES

M-matrices form an important class of matrices which are connected to the theory of Markov matrices. An M-matrix A can be defined by $A = sI - B$, where $s>0$, $B_{ij}>0$ and where $s \geqslant \rho(b)$ is the spectral radius of B. The form of A_{ij} is

$$\begin{pmatrix} a_{11} & -a_{12} & -a_{13} & \\ -a_{21} & a_{22} & -a_{23} & \vdots \\ & \cdots & & \end{pmatrix},$$

with $a_{ij} \geqslant 0$ (non-negative diagonal and nonpositive off diagonal terms). Extensive literature has developed relating M-matrices to Markov matrices and to non-negative matrices in general. In particular it can be shown that if B is a Markov matrix then $A = 1 - B$ is an M-matrix with "property C" (rank A = rank A^2).

We have previously proved that a Markov matrix $B = e^{\lambda \cdot L}(\lambda_{ij} \geqslant 0)$ has the representation $B = 1 + \alpha \cdot L$, where the $\alpha_{ij} \geqslant 0$ are determined by the λ_{ij}. Thus it follows from $-\alpha \cdot L = 1 - B$ that $-\alpha \cdot L$ is an M-matrix with property C

($\alpha_{ij} \geqslant 0$). Thus all those elements of the Markov Lie algebra, which are acceptable generators of Markov transformations, are the negative of an M-matrix with property C.

IX. CONCLUSIONS

We have studied a decomposition of the general linear group $\mathrm{GL}(n,R) = A(n,R) \oplus M(n,R)$, where $A(n,R)$ is the abelian scale transformation in n dimensions which naturally separates into the unit I and the $(n-1)H_i$ traceless generators. $M(n,R)$ was defined by $\langle 1|M = \langle 1|$, preserving

$$\sum_i x_i,$$

and was shown to give all Markov matrices continuously connected to the identity when the parameters in the associated Lie algebra were non-negative. Thus, even though Markov transformations do not form a group, they can be studied using much of the power and theorems available with Lie algebras. $M(n,R)$ was shown to contain a subgroup $M^D(n,R)$ of doubly stochastic processes and a basis of the $(n-1)^2$ generators of its Lie algebra were found. The M^D subalgebra was shown to contain the discrete symmetric group on n symbols, S_n, for certain values of the parameters over the complex field for which the transformations become real. Likewise the abelian group over the complex field $A(n,C)$ contains the real inversions. Thus the real transformations in $\mathrm{GL}(n,C)$ consist of those continuously connected to the identity through real parameters and the "discrete" groups which consist of those real transformations (inversions and the symmetric group) which can only be reached from the identity with complex parameters. Thus one can ask what restrictions are placed on behavior of representations of real Lie groups under the associated discrete groups which can be reached through complex parameters.

All subgroups of $\mathrm{GL}(n,C)$ can be viewed as a simultaneous implementation of

$$I, H_i, L^{D_{ij}}$$

and the $L^{i+1\,i}$ and thus as simultaneous scaling-, Markov-, and doubly stochastic-type transformations. In particular, the importance of classifying tensors under S_n can be seen here from a different point of view.

The Lie group approach to Markov processes allows one to formally use some alternative approaches: If the actual Markov transformation is uncertain but one knows the probability that a given transformation is correct then the transformation can be written

$$\int \eta(\lambda_{ij}) e^{\lambda_{ij} L^{ij}}\, d\lambda_{ij}, \tag{9.1}$$

where η represents a statistical weighting for different transformations. Since $e^{\lambda \cdot L} = 1 + \alpha \cdot L$ and since one requires

$$\int \eta\, d\lambda = 1, \tag{9.2}$$

then it follows that there exists a β such that

$$\int \eta(\lambda) e^{\lambda \cdot L}\, d\lambda = e^{\beta \cdot L}, \tag{9.3}$$

showing that statistical weightings of Markov processes are a single Markov process.

ACKNOWLEDGMENT

The author is indebted to the referee for a number of very useful comments and criticisms.

[1]B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[2]M. Hamermesh, *Group Theory And Its Application To Physical Problems* (Addison–Wesley, Reading, MA 1962).
[3]R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley, New York, 1974).
[4]P. H. Butler and R. C. King, J. Math. Phys. **14**, 1176 (1973).
[5]A. A. Markov, *Limiting Distribution Theorem for Computing the Probability of a Sum of Variables Forming a Chain* (in Russian, 1908) (Isbrannye Translations, 1951), pp. 365–397.
[6]A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences* (Academic, New York, 1979).
[7]G. Racah, Ergeb. Exakten Naturwiss. **37**, 28 (1965).

The Painlevé property and Bäcklund transformations for the sequence of Boussinesq equations

John Weiss
Center for Studies of Nonlinear Dynamics,[a] La Jolla Institute, 8950 Villa La Jolla Drive, Suite 2150, La Jolla, California 92037 and Institute for Pure and Applied Physical Science, University of California, San Diego, La Jolla, California 92093

(Received 27 March 1984; accepted for publication 24 August 1984)

We investigate the sequence of Boussinesq equations by the method of singular manifolds. For the Boussinesq equation, which is known to possess the Painlevé property, a Bäcklund transformation is defined. This Bäcklund transformation, which is formulated in terms of the Schwarzian derivative, obtains the system of modified Boussinesq equations and the resulting Miura-type transformation. The modified Boussinesq equations are found to be invariant under a discrete group of symmetries, acting on the dependent variables. By linearizing the Miura transformation (and modified equations) the Lax pair is readily obtained. Furthermore, by a result of Fokas and Anderson, the recursion operators defining the sequence of (higher-order) Boussinesq equations may be constructed from the Miura transformation. This allows the (recursive) definition of Bäcklund transformations for this sequence of equations. The recursion operator is found to preserve the discrete symmetries of the modified Boussinesq equations. This leads to the conclusion that the sequences of Boussinesq and modified Boussinesq equations identically possess the Painlevé property (are meromorphic). We also find that, by a simple reduction, the sequences of Caudrey–Dodd–Gibbon and Kuperschmidt equations are contained within the Boussinesq sequence. Rational solutions are iteratively constructed for the Boussinesq equation and a criterion is proposed for the existence of rational solutions of general integrable systems.

I. INTRODUCTION

Since this paper is one of several papers appearing recently concerning the Painlevé property for partial differential equations we spare the reader a formal definition of the Painlevé property, Bäcklund transformations, etc. For this see Refs. 1–6. Informally, when an equation possesses the Painlevé property the solutions are meromorphic functions of the independent variables. For a reasonably self-contained presentation we review, in Sec. II, the calculation of the Painlevé property and Bäcklund transformation for the Boussinesq equation.

In this paper we propose an extension of the methods of Refs. 1, 2, 4, and 5 for calculating Bäcklund transformations and Lax pairs. That is, when an equation is found to possess the Painlevé property, a certain Bäcklund transformation is defined. This Bäcklund transformation, when formulated in terms of the Schwarzian derivative, leads to an equation invariant under the Moebius group. From this equation, by a specific change of dependent variables (Miura transformations), both the original and a form modified equation are obtained. The resulting Miura transformation from modified to original equation is then linearized to obtain the Lax pair.

Now, when the equation/modified equation both have a Hamiltonian structure a result of Fokas and Anderson[7] may be used to construct the recusion operators defining the sequences of higher-order equations. (See also Ref. 8.) For these equations we can recursively define Bäcklund transformations and, in certain cases, by observing the effect of the discrete symmetries of the modified equations acting on the singularities prove that the entire sequence of equations possesses the Painlevé property.[4]

[a] Affiliated with the University of California, San Diego.

In Sec. II the Bäcklund transformation, modified equations, Miura transformations, and Lax pair for the Boussinesq equation are calculated by the above method. The modified Boussinesq equations are also found to be invariant under a discrete group of symmetries.

In Sec. III the sequences of higher-order equations are investigated. The recursion operators are shown to preserve the discrete symmetries of the modified equations. These discrete symmetries, when interpreted in terms of the underlying equation for the singular manifold, and combined with the invariance of this equation under the Moebius group, allows the conclusion that the sequences of higher-order Boussinesq and modified Boussinesq equations identically possess the Painlevé property. We also define Bäcklund transformations for both sequences of equations.

With the view toward understanding the generality of the above procedures we consider in Appendix A the nonlinear Schrödinger equation. Insofar as obtaining the Bäcklund transformation, modified equations, and a (scalar) Lax pair the method proceeds as before. However, the modified nonlinear Schrödinger equations, while similar to the modified Boussinesq equations, do not allow a group of discrete symmetries. Therefore, the agruments used to conclude that the Boussinesq sequence is identically Painlevé do not apply to the nonlinear Schrödinger sequence.

Finally, in Appendix B certain rational solutions connected with the discrete group of symmetries are obtained.

II. THE BOUSSINESQ EQUATION

The Boussinesq equation

$$U_{tt} = -\frac{\partial^2}{\partial x^2}\left(\frac{U_{xx}}{3} + U^2\right) \tag{2.1}$$

is known to possess the Painlevé property.[1,2] That is, about a

 0022-2488/85/020258-12$02.50

"manifold" of "movable" singularities determined by the expression

$$\varphi(x,t)=0, \tag{2.2}$$

the Boussinesq equation has the expansion

$$U=\varphi^{-2}\sum_{j=0}^{\infty}U_j\varphi^j, \tag{2.3}$$

where (φ,U_4,U_5,U_6) are "arbitrary," locally analytic functions of (x,t). In general, for the expansion (2.3) to be well defined about the manifold (2.2), it is required that (2.2) be "noncharacteristic" for the equation (2.1) (i.e., the Cauchy–Kovalevskaya theorem). In the present case, this requires that $\varphi_x\neq 0$ when $\varphi=0$. With this provision, (2.3) defines the general (meromorphic) expansion of the solution about (2.2).

From the recursion relations for U_j [substituting (2.3) into (2.1)] it is found that

$$U_0=-2\varphi_x^2, \tag{2.4}$$

$$U_1=2\varphi_{xx}, \tag{2.5}$$

$$\varphi_t^2-\varphi_{xx}^2+\tfrac{4}{3}\varphi_x\varphi_{xxx}+2\varphi_x^2U_2=0, \tag{2.6}$$

$$\varphi_{tt}+\tfrac{1}{3}\varphi_{xxxx}+2\varphi_{xx}U_2-2\varphi_x^2U_3=0, \tag{2.7}$$

and (U_4, U_5, U_6) are "arbitrary."[1] We note that the noncharacteristic condition is (essentially) $U_0\neq 0$ when $\varphi=0$.

We now attempt to define a Bäcklund transformation for Eq. (2.1) by truncating the expansion (2.3) at the "constant" level. That is, let

$$U=U_0\varphi^{-2}+U_+\varphi^{-1}+U_2, \tag{2.8}$$

and require, in the expressions defined by the recursion relations, that

$$U_j\equiv 0, \tag{2.9}$$

for $j\geqslant 3$. In general, we would expect to obtain an overdetermined system of equations for (φ, U_0, U_1, U_2). In this case, the system is *not* overdetermined. The (U_0, U_1) are determined by (2.4) and (2.5), and the (φ, U_2) are defined by (2.6) and (2.7), with $U_3=0$. Since (U_4, U_5, U_6) are arbitrary they may be set to zero without restriction. The system terminates at the condition $U_6=0$, obtaining that U_2 satisfies Eq. (2.1) as a (trivial) consequence of Eqs. (2.4)–(2.7) (with $U_3=0$). Solving for (U_2, φ), the Bäcklund transformation reads

$$U=2\frac{\partial^2}{\partial x^2}\ln\varphi+U_2, \tag{2.10}$$

where (U, U_2) satisfy Eq. (2.1),

$$2U_2+\frac{\varphi_t^2}{\varphi_x^2}-\frac{\varphi_{xx}^2}{\varphi_x^2}+\frac{4}{3}\frac{\varphi_{xxx}}{\varphi_x}=0, \tag{2.11}$$

and

$$\frac{\partial}{\partial t}\left(\frac{\varphi_t}{\varphi_x}\right)+\frac{1}{3}\left(\{\varphi;x\}+\frac{3}{2}\left(\frac{\varphi_t}{\varphi_x}\right)^2\right)=0. \tag{2.12}$$

The expression

$$\{\varphi;x\}=\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right)-\frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2 \tag{2.13}$$

is the Schwarzian derivative,[2] which is invariant under the Moebius group

$$\varphi=(a\psi+b)/(c\psi+d). \tag{2.14}$$

By this Eq. (2.12) is also invariant under (2.14). Note that (2.11) is a Miura-type transformation from Eq. (2.12) to Eq. (2.1). In effect, Eq. (2.12) is a form of "modified" Boussinesq equation. If we let

$$v=\varphi_{xx}/\varphi_x, \tag{2.15}$$

$$\omega=\varphi_t/\varphi_x \tag{2.16}$$

and use the identity

$$v_t=\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}+v\right)\omega, \tag{2.17}$$

then (v,ω) satisfy the system of modified Boussinesq equations[7]:

$$v_t=\frac{\partial}{\partial x}(\omega_x+v\omega),$$

$$\omega_t=-\frac{1}{3}\frac{\partial}{\partial x}\left(v_x-\frac{1}{2}v^2+\frac{3}{2}\omega^2\right). \tag{2.18}$$

The Miura transformation (2.11) is

$$2U_2+\omega^2+\tfrac{4}{3}(v_x+\tfrac{1}{4}v^2)=0. \tag{2.19}$$

Since (2.19) maps the system (2.18) into the scalar equation (2.1), it is convenient to reformulate (2.1) as the system of equations

$$U_t=H_x,\quad H_t=\frac{\partial}{\partial x}\left(-\frac{U_{xx}}{3}-U^2\right), \tag{2.20}$$

with the Miura transformation

$$2U+\omega^2+\tfrac{4}{3}(v_x+\tfrac{1}{4}v^2)=0,$$

$$3H+2\omega_{xx}-\omega^2+v_x\omega+3v\omega_x+v^2\omega=0. \tag{2.21}$$

Now, the modified Boussinesq equations (2.18) are invariant under the transformation

$$\begin{pmatrix}v\\ \omega\end{pmatrix}=A_\pm\begin{pmatrix}\theta\\ z\end{pmatrix}, \tag{2.22}$$

where

$$A_\pm=\begin{pmatrix}-\frac{1}{2} & \mp\frac{3}{2}\\ \pm\frac{1}{2} & -\frac{1}{2}\end{pmatrix}, \tag{2.23}$$

and

$$\begin{aligned}&\text{(i)}\quad |A_\pm|=1,\\ &\text{(ii)}\quad A_\pm^{-1}=A_\mp,\\ &\text{(iii)}\quad A_\pm^3=I.\end{aligned} \tag{2.24}$$

The Miura transformation (2.21) is

$$U=\pm z_x-z^2/2+\tfrac{1}{3}(\theta_x-\theta^2/2),$$

$$3H=\left(\frac{\partial}{\partial x}\mp 2z\right)\left(z_x\mp\frac{z^2}{2}\mp\theta_x\pm\frac{\theta^2}{2}\right), \tag{2.25}$$

where

$$\theta_t=\frac{\partial}{\partial x}(z_x+\theta z),$$

$$z_t=-\frac{1}{3}\frac{\partial}{\partial x}\left(\theta_x-\frac{1}{2}\theta^2+\frac{3}{2}z^2\right). \tag{2.26}$$

Equations (2.25) are linearized by the substitution

$$z = \mp \beta_x / \beta \tag{2.27}$$

to

$$4\beta_{xxx} + 6U\beta_x + 3(U_x \pm H)\beta = 0, \tag{2.28}$$

From Eqs. (2.26) there is found

$$\pm \beta_t = \beta_{xx} + (U + \lambda)\beta, \tag{2.29}$$

where λ is a constant of integration. Equations (2.28) and (2.29) are the Lax pair for Eq. (2.20).

We recall that, for Eq. (2.12),

$$v = \varphi_{xx}/\varphi_x, \quad \omega = \varphi_t/\varphi_x. \tag{2.30}$$

From the symmetry (2.22) of (2.18) we identify

$$\theta = \psi_{xx}/\psi_x, \quad z = \psi_t/\psi_x. \tag{2.31}$$

Thus,

$$\frac{\varphi_{xx}}{\varphi_x} = -\frac{1}{2}\frac{\psi_{xx}}{\psi_x} \mp \frac{3}{2}\frac{\psi_t}{\psi_x}, \quad \frac{\varphi_t}{\varphi_x} = \pm\frac{1}{2}\frac{\psi_{xx}}{\psi_x} - \frac{1}{2}\frac{\psi_t}{\psi_x}. \tag{2.32}$$

The compatibility condition

$$\varphi_{xxt} = \varphi_{txx} \tag{2.33}$$

is satisfied by Eqs. (2.32) if and only if ψ satisfies Eq. (2.12). Thus, Eqs. (2.32) constitute a Bäcklund transformation for (2.12). As previously noted Eq. (2.12) is also invariant under the Moebius group. This dual invariance allows certain rational solutions to be constructed iteratively for Eq. (2.12) (see Appendix B).

Equation (2.12) allows two types of singularities. For one,

$$\text{(i)} \quad \varphi = \epsilon^{-1}\sum_{j=0}^{\infty}\varphi_j\epsilon^j, \tag{2.34}$$

and for the other

$$\text{(ii)} \quad \varphi \simeq \varphi_0(t) + \varphi_2\epsilon^2 + \cdots, \tag{2.35}$$

where

$$\varphi_{0_t} = \pm 2\epsilon_x\varphi_2.$$

Singularities of the form (2.35) occur at point where $\varphi_x = 0$. By direct calculation both forms of singularity are single valued. As explained in Ref. 4 the form of Eq. (2.12) is sufficient to guarantee the meromorphic behavior of the singularity, (2.34). For instance, the invariance of Eq. (2.12) [under (2.14)]

$$\psi = 1/\varphi \tag{2.36}$$

throws the simple pole of φ into a simple zero of ψ:

$$\psi = \epsilon\sum_{j=0}^{\infty}\psi_j\epsilon^j, \tag{2.37}$$

where ψ is locally analtyic near $\epsilon = 0$. We note that, by the Cauchy–Kovalevsky theorem,[9] (2.37) converges in an open neighborhood at the mainfold ($\epsilon = 0$).

For simplicity, let

$$\epsilon \to x + \epsilon(t), \tag{2.38}$$

and find to leading order

$$\text{(i)} \quad v = \varphi_{xx}/\varphi_x \simeq -2/\epsilon, \quad \omega = \varphi_t/\varphi_x \simeq O(1),$$

$$\begin{pmatrix} v \\ \omega \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}\epsilon^{-1}, \tag{2.39}$$

for (2.34); and

$$\text{(ii)} \quad \begin{pmatrix} v \\ \omega \end{pmatrix} \simeq \begin{pmatrix} 1 \\ \pm 1 \end{pmatrix}\epsilon^{-1}, \tag{2.40}$$

for (2.35). From (2.22),

$$\begin{aligned} A_\pm\begin{pmatrix} -2 \\ 0 \end{pmatrix} &= \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \\ A_\pm\begin{pmatrix} 1 \\ 1 \end{pmatrix} &= \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \\ A_\pm\begin{pmatrix} 1 \\ -1 \end{pmatrix} &= \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ -0 \end{pmatrix}. \end{aligned} \tag{2.41}$$

Thus, the singularities of Eqs. (2.12) and (2.18) are permuted by the symmetry (2.22) and (2.32). A singularity of the form (2.35) can be transformed into the form (2.34). Therefore, by reconstruction from (2.30) and (2.22), (2.35) is single valued. In the next section it is found that *all* singularities of the Boussinesq sequence can be transformed into form (2.34) by a combination of the invariances (2.32) and (2.36).

At this point it is worth remarking that Eq. (2.12) is unique among equations of the form

$$\frac{\partial}{\partial t}\left(\frac{\varphi_t}{\varphi_x}\right) + \frac{\partial}{\partial x}\left(a\{\varphi;x\} + b\left(\frac{\varphi_t}{\varphi_x}\right)^2\right) = 0, \tag{2.42}$$

since only equations equivalent to (2.12) under scalings of (x,t) have a set of nontrivial discrete symmetries [when expressed in the form (2.18)]. This will be relevant to the analysis of the nonlinear Schrödinger equation in Appendix A.

III. THE BOUSSINESQ SEQUENCE

The Boussinesq and modified Boussinesq equations may be formulated as Hamiltonian systems.[7] That is,

$$\begin{pmatrix} U \\ H \end{pmatrix}_t = \Omega_1\begin{pmatrix} -U_{xx}/3 - U^2 \\ H \end{pmatrix}, \tag{3.1}$$

$$\begin{pmatrix} v \\ \omega \end{pmatrix}_t = \Omega_2\begin{pmatrix} \omega_x + v\omega \\ -v_x + \frac{1}{2}v^2 - \frac{3}{2}\omega^2 \end{pmatrix}, \tag{3.2}$$

where

$$D = \frac{\partial}{\partial x},$$

$$\Omega_1 = \begin{pmatrix} 0 & D \\ D & 0 \end{pmatrix}, \tag{3.3}$$

$$\Omega_2 = \begin{pmatrix} D & 0 \\ 0 & \frac{1}{3}D \end{pmatrix}, \tag{3.4}$$

are symplectic operators and

$$\begin{pmatrix} -U_{xx}/3 - U^2 \\ H \end{pmatrix} = \nabla H_1, \tag{3.5}$$

$$\begin{pmatrix} \omega_x + v\omega \\ -v_x + \frac{1}{2}v^2 - \frac{3}{2}\omega^2 \end{pmatrix} = \nabla H_2, \tag{3.6}$$

are the functional gradients of the Hamiltonians

$$H_1 = \int\left\{\frac{U_x^2}{6} - \frac{U^3}{3} + H^2\right\}, \tag{3.7}$$

$$H_2 = \int\left\{\frac{v\omega_x - \omega v_x}{2} + \frac{1}{2}v^2\omega - \frac{1}{2}\omega^3\right\}. \tag{3.8}$$

By the results of the previous section Eq. (3.2) is invariant under the transformation

$$\begin{pmatrix} v \\ \omega \end{pmatrix} = A_\pm \begin{pmatrix} \theta \\ z \end{pmatrix}, \tag{3.9}$$

where $A_\pm$ is defined by (2.23). The three Miura transformations from (3.2) to (3.1) are

(i) $$U = -\tfrac{1}{2}\left(\omega^2 + \tfrac{4}{3}(v_x + \tfrac{1}{4}v^2)\right),$$
$$H = -\tfrac{1}{3}(2\omega_{xx} - \omega^3 + v_x\omega + 3v\omega_x + v^2\omega), \tag{3.10}$$

(ii) $$U = z_x - z^2/2 + \tfrac{1}{3}(\theta_x - \theta^2/2),$$
$$H = \tfrac{1}{3}(D - 2z)(z_x - z^2/2 - (\theta_x - \theta^2/2)), \tag{3.11}$$

with A_+ in (3.9); and

(iii) $$U = -z_x - z^2/2 + \tfrac{1}{3}(\theta_x - \theta^2/2),$$
$$H = \tfrac{1}{3}(D + 2z)[z_x + z^2/2 + (\theta_x - \theta^2/2)], \tag{3.12}$$

with A_-.

By a theorem of Ref. 7 a Miura transformation between two systems with a Hamiltonian structure provides the means for constructing a second Hamiltonian structure for both equations, and, thereby, the recursion operators determining the sequences of higher-order equations. We have from (3.10)–(31.2) the operators

(i) $$B_1 = -\frac{1}{3}\begin{pmatrix} 2D + v & 3\omega \\ D\omega + 2(\omega_x + v\omega) & 2D^2 + 3Dv - 2v_x + v^2 - 3\omega^2 \end{pmatrix}, \tag{3.13}$$

(ii) $$B_2 = \frac{1}{3}\begin{pmatrix} D - \theta & 3(D - z) \\ -(D - 2z)(D - \theta) & D^2 - 3Dz + 3z^2 + 2(\theta_x - \tfrac{1}{2}\theta^2) \end{pmatrix}, \tag{3.14}$$

(iii) $$B_3 = \frac{1}{3}\begin{pmatrix} D - \theta & -3(D + z) \\ (D + 2z)(D - \theta) & D^2 + 3Dz + 3z^2 + 2(\theta_x - \tfrac{1}{2}\theta^2) \end{pmatrix}, \tag{3.15}$$

which determine the first variations of the respective Miura transformations about solutions of (3.2). From Ref. 7 the recursion operators (strong symmetries) of (3.1) and (3.2) are

$$M = B\Omega_2 B^* \Omega_1^{-1}, \tag{3.16}$$

$$L = \Omega_2 B^* \Omega_1^{-1} B, \tag{3.17}$$

where B is (3.13), (3.14), or (3.15), B^* is the adjoint operator, and

$$\Omega_1^{-1} = \begin{pmatrix} 0 & D^{-1} \\ D^{-1} & 0 \end{pmatrix}. \tag{3.18}$$

The sequences of Boussinesq and modified Boussinesq equations are

$$\begin{pmatrix} U \\ H \end{pmatrix}_t = M^n \Omega_1 \begin{pmatrix} -U_{xx}/3 - U^2 \\ H \end{pmatrix}, \tag{3.19}$$

$$\begin{pmatrix} \theta \\ z \end{pmatrix}_t = L^n \Omega_2 \begin{pmatrix} z_x + \theta z \\ -\theta_x + \tfrac{1}{2}\theta^2 - \tfrac{3}{2}z^2 \end{pmatrix}, \tag{3.20}$$

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

By direct calculation, using (3.10)–(3.15), we find that

$$M_1 = M_2 = M_3 = M, \quad L_1 = L_2 = L_3 = L, \tag{3.21}$$

where the subscript refers to the transformations (3.10), (3.11), (3.12), respectively. This result demonstrates that Eqs. (3.20) are invariant under (3.9), and (3.10) to (3.12) defines Miura transformations from (3.20) to (3.19). For reference,

$$B\Omega_2 B^* = -\frac{1}{9}\begin{pmatrix} 4D^3 + 6UD + 3U_x & 9HD + 6H_x \\ 9HD + 3H_x & -\tfrac{4}{3}D^5 - 10UD^3 - 15U_xD^2 - (9U_{xx} + 12U^2)D - (2U_{xxx} + 12UU_x) \end{pmatrix}, \tag{3.22}$$

$$B^*\Omega_1^{-1}B = \frac{1}{9}$$
$$\times \begin{pmatrix} \begin{matrix} -4zD - 2z_x + 2\theta D^{-1}(z_x + z\theta) \\ + 2(z_x + z\theta)D^{-1}\theta \end{matrix} & \begin{matrix} -4D^2 - 4\theta D + 2(-\theta_x + \theta^2/2 + \tfrac{3}{2}z^2). \\ + 2\theta D^{-1}(-\theta_x + \theta^2/2 - \tfrac{3}{2}z^2) + 6(z_x + \theta z)D^{-1}z \end{matrix} \\ \begin{matrix} 4D^2 - 4\theta D + 2(-\theta_x - \theta^2/2 - \tfrac{3}{2}z^2) \\ + 2(-\theta_x + \theta^2/2 - \tfrac{3}{2}z^2)D^{-1}\theta + 6zD^{-1}(z_x + \theta z) \end{matrix} & \begin{matrix} 12zD + 6z_x + 6zD^{-1}(-\theta_x + \theta^2/2 - \tfrac{3}{2}z^2) \\ + 6(-\theta_x + \theta^2/2 - \tfrac{3}{2}z^2)D^{-1}z \end{matrix} \end{pmatrix}. \tag{3.23}$$

At this point it is convenient to identify the following expressions:

$$s = \theta_x - \tfrac{1}{2}\theta^2, \tag{3.24}$$

$$\Omega = \begin{pmatrix} (D - \theta)D(D + \theta) & 0 \\ 0 & -\tfrac{1}{2}D \end{pmatrix}, \tag{3.25}$$

$$C = \frac{1}{3}\begin{pmatrix} 1 & 3(D - z) \\ -D + 2z & D^2 - 3Dz + 3z^2 + 2s \end{pmatrix}, \tag{3.26a}$$

$$R = \begin{pmatrix} D - \theta & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.26b}$$

$$M_2 = C\Omega C^* \Omega_1^{-1}, \tag{3.27}$$

$$L_2 = \Omega C^* \Omega_1^{-1} C, \tag{3.28}$$

where

$$U = z_x - \tfrac{1}{2}z^2 + \tfrac{1}{3}s, \quad H = \tfrac{1}{3}(D - 2z)(z_x - \tfrac{1}{2}z^2 - s). \tag{3.29}$$

We note the following identities:

$$M_2 = -M, \quad B = CR, \tag{3.30}$$

$$RL = -L_2 R, \tag{3.31}$$

$$\begin{pmatrix} U \\ H \end{pmatrix}_t = C\begin{pmatrix} s \\ z \end{pmatrix}_t = B\begin{pmatrix} \theta \\ z \end{pmatrix}_t, \tag{3.32}$$

$$\Omega_1\begin{pmatrix} -U_{xx}/3 - U^2 \\ H \end{pmatrix} = B\Omega_2\begin{pmatrix} z_x + \theta z \\ -s - \tfrac{3}{2}z^2 \end{pmatrix} = C\Omega\begin{pmatrix} z \\ s + \tfrac{3}{2}z^2 \end{pmatrix}. \tag{3.33}$$

We now formulate the following theorem.

Theorem: For the Boussinesq sequence

$$\begin{pmatrix} U \\ H \end{pmatrix}_t = M^n \Omega_1\begin{pmatrix} -U_{xx}/3 - U^2 \\ H \end{pmatrix} \tag{3.34}$$

and the modified Boussinesq sequence

$$\begin{pmatrix} \theta \\ z \end{pmatrix}_t = L^n \Omega_2\begin{pmatrix} z_x + \theta z \\ -s - \tfrac{3}{2}z^2 \end{pmatrix}, \tag{3.35}$$

when $n = 0,1,2,3,\ldots$,

$$s = \theta_x - \tfrac{1}{2}\theta^2,$$

there exists the Bäcklund transformation (BT)

$$U = 2\frac{\partial^2}{\partial x^2}\ln\varphi + U_2, \quad H = 2\frac{\partial^2}{\partial x\,\partial t}\ln\varphi + H_2, \tag{3.36}$$

$$\theta = -2\frac{\partial}{\partial x}\ln\varphi + \theta_1, \quad z = z_1, \tag{3.37}$$

where (U,H), (U_2,H_2) satisfy (3.34); (θ,z), (θ_1,z) satisfy (3.35);

$$\theta_1 = \varphi_{xx}/\varphi_x, \quad s = \{\varphi;x\}, \tag{3.38}$$

$$U_2 = -\tfrac{1}{2}\{z_1^2 + \tfrac{4}{3}(\theta_{1x} + \tfrac{1}{4}\theta_1^2)\}, \tag{3.39}$$

$$H_2 = -\tfrac{1}{3}\{2z_{1xx} - z_1^3 + z_1\theta_{1x} + 3z_{1x}\theta_1 + z_1\theta_1^2\};$$

and

$$\begin{pmatrix} \varphi_t/\varphi_x \\ z_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -\tfrac{1}{3}D \end{pmatrix} P^n \begin{pmatrix} z \\ s + \tfrac{3}{2}z^2 \end{pmatrix}, \tag{3.40}$$

$$P = -C^*\Omega_1^{-1}C\Omega. \tag{3.41}$$

Furthermore, Eqs. (3.35) are invariant under the transformations

$$\begin{pmatrix} \theta_1 \\ z_1 \end{pmatrix} = A_+\begin{pmatrix} \theta_2 \\ z_2 \end{pmatrix}, \tag{3.42}$$

$$\begin{pmatrix} \theta_1 \\ z_1 \end{pmatrix} = A_-\begin{pmatrix} \theta_3 \\ z_3 \end{pmatrix}, \tag{3.43}$$

where

$$A_\pm = \begin{pmatrix} -\tfrac{1}{2} & \mp\tfrac{3}{2} \\ \pm\tfrac{1}{2} & -\tfrac{1}{2} \end{pmatrix}. \tag{3.44}$$

In addition,

$$U_3 = z_{2x} - z_2^2/2 + \tfrac{1}{3}(\theta_{2x} - \theta_2^2/2), \tag{3.45}$$

$$H_3 = \tfrac{1}{3}(D - 2z_2)(z_{2x} - z_2^2/2 - (\theta_{2x} - \theta_2^2/2)),$$

$$U_4 = -z_{3x} - z_3^2/2 + \tfrac{1}{3}(\theta_{3x} - \theta_3^2/2), \tag{3.46}$$

$$H_4 = \tfrac{1}{3}(D + 2z_3)(z_{3x} + z_3^2/2 + (\theta_{3x} - \theta_3^2/2))$$

also define solutions (U_3,H_3), (U_4,H_4) of Eqs. (3.34).

Proof: By (3.21), Eqs. (3.35) are invariant under (3.42), (3.43), and the Miura transformations (3.39), (3.45), and (3.46) from (3.35) to (3.34) are well defined. Now, the identity (when $\theta_1 = \varphi_{xx}/\varphi_x$)

$$\begin{pmatrix} D(D+\theta_1) & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \varphi_t/\varphi_x \\ z_t \end{pmatrix} = \begin{pmatrix} \theta_1 \\ z \end{pmatrix}_t, \tag{3.47}$$

$$\begin{pmatrix} D(D+\theta_1) & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -\tfrac{1}{3}D \end{pmatrix}P^n\begin{pmatrix} z \\ s + \tfrac{3}{2}z^2 \end{pmatrix} = L^n\Omega_2\begin{pmatrix} z_x + \theta_1 z \\ -s - \tfrac{3}{2}z^2 \end{pmatrix} \tag{3.48}$$

establishes that (θ_1,z) is a solution of (3.35), with $\theta_1 = \varphi_{xx}/\varphi_x$.

By evaluation of (3.41)

$$P = -\frac{1}{9}\begin{pmatrix} 1 & D + 2z \\ -3(D+z) & D^2 + 3zD + 3z + 2s \end{pmatrix} \times\begin{pmatrix} 0 & D^{-1} \\ D^{-1} & 0 \end{pmatrix} \times\begin{pmatrix} 1 & 3(D-z) \\ -D + 2z & D^2 - 3Dz + 3z^2 + 2s \end{pmatrix} \times\begin{pmatrix} D^3 + 2sD + s_x & 0 \\ 0 & -\tfrac{1}{3}D \end{pmatrix}, \tag{3.49}$$

where $s = \{\varphi;x\}$. Thus, by the invariance of the derivative s under the Moebius group and the form of Eqs. (3.40), Eqs. (3.40) are invariant under the transformation

$$\varphi = (a\psi + b)/(c\psi + d), \quad z = z. \tag{3.50}$$

In particular, Eqs. (3.40) are invariant under

$$\varphi = 1/\psi, \quad z = z. \tag{3.51}$$

However,

$$\theta_1 = \frac{\varphi_{xx}}{\varphi_x} = \frac{\psi_{xx}}{\psi_x} - 2\frac{\partial}{\partial x}\ln\psi = \frac{\psi_{xx}}{\psi_x} + 2\frac{\partial}{\partial x}\ln\varphi,$$

which is the BT (3.37) with

$$\theta = \psi_{xx}/\psi_x, \tag{3.52}$$

and by (3.51) and the previous remarks (θ,z) is a solution of (3.35). Furthermore, from (3.36), (3.39), (3.51) and (3.52), we find that

$$U = -\tfrac{1}{2}(z^2 + \tfrac{4}{3}(\theta_x + \tfrac{1}{4}\theta^2)) \tag{3.53}$$

$$H = -\tfrac{1}{3}(2z_{xx} - z^3 + z\theta_x + 3z_x\theta + z\theta^2),$$

demonstrating, by the previous remarks, that (U,H) are solutions of (3.34), completing the proof.

Remark 1: In certain instances it is preferable to express the equation sequences in terms of the recursion operators of conserved covariants, rather than the "symmetries." We find for Eqs. (3.34), (3.35), and (3.40) that

$$\begin{pmatrix} U \\ H \end{pmatrix}_t = \Omega_1 J^n\begin{pmatrix} -U_{xx}/3 - U^2 \\ H \end{pmatrix}, \tag{3.54}$$

$$\begin{pmatrix} \theta \\ z \end{pmatrix}_t = \Omega_2 K^n\begin{pmatrix} z_x + \theta z \\ -\theta_x + \tfrac{1}{2}\theta^2 - \tfrac{3}{2}z^2 \end{pmatrix}, \tag{3.55}$$

$$\begin{pmatrix}\varphi_t/\varphi_x\\ z_t\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & -\frac{1}{3}D\end{pmatrix}C*J^{n-1}\begin{pmatrix}-U_{xx}/3-U^2\\ H\end{pmatrix}, \tag{3.56}$$

where

$$\Omega_1^{-1}C\Omega\begin{pmatrix}z\\ s+\frac{3}{2}z^2\end{pmatrix}=\begin{pmatrix}-U_{xx}/3-U^2\\ H\end{pmatrix}, \tag{3.57}$$

and

$$J=\Omega_1^{-1}B\Omega_2B*, \tag{3.58}$$

$$K=B*\Omega_1^{-1}B\Omega_2. \tag{3.59}$$

Remark 2: By applying the operator R (3.26b) to the sequence (3.35), using (3.24), (3.25), (3.28), (3.32), (3.33), the sequence of Hamiltonian systems,

$$\begin{pmatrix}s\\ z\end{pmatrix}_t=(-L_2)^n\Omega\begin{pmatrix}z\\ s+\frac{3}{2}z^2\end{pmatrix}, \tag{3.60}$$

is found. From (3.45) and (3.46) we have the Miura transformations

(i) $U=z_x-z^2/2+\frac{1}{3}s,\quad H=\frac{1}{3}(D-2z)(z_x-z^2/2-s);$ (3.61)

(ii) $U=-z_x-z^2/2+\frac{1}{3}s,\quad H=\frac{1}{3}(D+2z)(z_x+z^2/2+s);$ (3.62)

connecting (3.60) to (3.34). From (3.17), (3.23), and (3.35) it is easy to see that (3.35) is invariant under

$$z\rightarrow -z, \tag{3.63}$$

when $n=2j+1, j\geqslant 0$. By construction the same invariance applied to (3.40) and (3.60). Therefore, when

$$n=2j+1,\quad j\geqslant 0, \tag{3.64}$$

a consistent reduction of (3.35), (3.40), (3.60) is to let

$$z\equiv 0. \tag{3.65}$$

The Miura transformations (3.39), (3.45), and (3.46) are

(i) $U=-\frac{2}{3}(\theta_{1x}+\frac{1}{4}\theta_1^2),\quad H=0;$ (3.66)

(ii) $U=\frac{1}{3}(\theta_{2x}-\theta_2^2/2),\quad H=-U_x;$ (3.67)

(iii) $U=\frac{1}{3}(\theta_{3x}-\theta_3^2/2),\quad H=U_x.$ (3.68)

For (3.66) we let

$$\theta_1=-2a,\quad b=a_x-\tfrac{1}{2}a^2, \tag{3.69}$$

$$U=\tfrac{2}{3}b, \tag{3.70}$$

and find from Eq. (3.34) that

$$b_t=\tfrac{4}{27}m_3^j\Omega_3(b_{xx}+4b^2), \tag{3.71}$$

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

$$m_3=(\tfrac{4}{3})^2\Omega_3J_3,\quad \Omega_3=(D-a)D(D+a),$$
$$J_3=-\tfrac{1}{3}D^{-1}(D-2a)(D-a)D(D+a)(D+2a)D^{-1}. \tag{3.72}$$

For (3.67), (3.68), with

$$s=\theta_{2x}-\theta_2^2/2=\theta_{3x}-\theta_3^2/2$$

or

$$s=\theta_x-\theta^2/2, \tag{3.73}$$

Equation (3.60) obtains

$$s_t=\tfrac{4}{27}m_4^j\Omega_4(s_{xx}+\tfrac{1}{3}s^2), \tag{3.74}$$

for $j=0,1,2,3,\ldots$, where

$$m_4=(\tfrac{4}{3})^2\Omega_4J_4,\quad \Omega_4=(D-\theta)D(D+\theta),$$
$$J_4=-\frac{1}{3}D^{-1}\Big(D-\frac{\theta}{2}\Big)\Big(D+\frac{\theta}{2}\Big)\times D\Big(D-\frac{\theta}{2}\Big)\Big(D+\frac{\theta}{2}\Big)D^{-1}. \tag{3.75}$$

Equations (3.71) and (3.74) are the sequences of Kuperschmidt/Caudrey–Dobb–Gibbon equations, respectively.[4]

To continue the analysis of the Boussinesq sequence it is necessary to define the discrete symmetries of the modified Boussinesq equations (3.42) and (3.43), as Bäcklund transformations for the singular manifold equation (3.40). That is,

$$\begin{pmatrix}\varphi_{xx}/\varphi_x\\ z\end{pmatrix}=A_\pm\begin{pmatrix}\psi_{xx}/\psi_x\\ z_1\end{pmatrix}. \tag{3.76}$$

In this way the investigation of the singularities for the Boussinesq and the modified Boussinesq sequences is referred to an investigation of the singularities for the sequence (3.40), which, as in Sec. II, allows a simplified discussion. To begin for a solution (θ,z) of (3.35) we define variables (ψ,z) by

$$\psi_{xx}/\psi_x=\theta,\quad z=z. \tag{3.77}$$

Therefore, ψ is determined up to two arbitrary functions of t. On the other hand, with the identification (3.77), (ψ,z) satisfies Eq. (3.40) with the possible inclusions of a term from the null space of the operator,

$$T=\begin{pmatrix}D(D+\theta) & 0\\ 0 & 1\end{pmatrix}. \tag{3.78}$$

The general form of a null vector, when $\theta=\psi_{xx}/\psi_x$, is

$$\hat{n}=\begin{pmatrix}a/\psi_x+b\psi/\psi_x\\ 0\end{pmatrix}, \tag{3.79}$$

where (a,b) are functions of t. Therefore, for an arbitrary (ψ,z) satisfying (3.77),

$$\begin{pmatrix}\psi_t/\psi_x+a/\psi_x+b(\psi/\psi_x)\\ z_t\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & -\frac{1}{3}D\end{pmatrix}P^n\begin{pmatrix}z\\ s+\frac{3}{2}z^2\end{pmatrix}, \tag{3.80}$$

where $s=\{\psi;x\}$. Now, the right side of (3.80) is expressed entirely in terms of the variables (s,z), which implies that the right side is unchanged in form by the transformation

$$\psi\rightarrow e^{-\int^t b}\Big\{\psi_1-\int^t ae^{\int^s b}\,ds\Big\}, \tag{3.81}$$

where (ψ_1,z) satisfies (3.40). Thus for an appropriate choice of the time-dependent "constants" of integration there exists a solution of (3.77) [for "arbitrary" (θ,z)] so that (ψ,z) satisfies (3.40). From (3.81),

$$\psi_{xx}/\psi_x=\psi_{1xx}/\psi_{1x}=\theta. \tag{3.82}$$

Furthermore, (ψ_1,z) is uniquely determined up to transformations of the form

$$\psi_1=a\psi+b, \tag{3.83}$$

where (a,b) are (time-independent) constants, and [modulo (3.83)] the transformation $(\theta,z)\leftrightarrow(\psi_1,z)$ is one to one. Therefore, the Bäcklund transformation (3.76) is well defined for

Eqs. (3.40). Alternatively, let (ψ,z) be a known solution of (3.40) and, applying (3.76), substitute for $(\varphi_{xx}/\varphi_x,z)$ in the right side of (3.40). By the invariance of (3.35) the equation for z is satisfied identically, while φ_t/φ_x is a known function of (x,t), as is φ_{xx}/φ_x, which determined φ uniquely up to the equivalence (3.83). In a similar way it can be shown that

$$s=\{\varphi;x\},\quad z=z \tag{3.84}$$

define a transformation from (3.60) to (3.40) which determines an unique φ, [modulo (3.50)], as a solution of (3.40).

We next propose to classify the singularities of (3.40) according to their "leading-order" behavior and observe the effect of the transformations (3.50) and (3.76) on these singularities.

Recall from Sec. II that Eqs. (3.40) have, when $n=0$, two types of singularities, (2.34) and (2.35). With the notation

$$\varphi_{xx}/\varphi_x\simeq k\epsilon^{-1}+\cdots,\qquad z\simeq\beta\epsilon^{-1}+\cdots, \tag{3.85}$$

these are represented, to the leading order, by Table I, where $\alpha=k+1, \alpha_+=-\alpha_-$. To the leading order the symmetry (3.76) is represented by the transformation

$$k'=-\tfrac{1}{2}k\mp\tfrac{3}{2}\beta,\quad \beta'=\pm\tfrac{1}{2}k-\tfrac{1}{2}\beta, \tag{3.86}$$

and the inversion, $\varphi\rightarrow1/\varphi'$ by

$$\alpha'=-\alpha. \tag{3.87}$$

In the expansion of φ in (3.40) we have

$$\varphi=\varphi_0\epsilon^{\alpha}+\cdots, \tag{3.88}$$

hence, (3.87). Note (3.87) does not apply to singularities of the form

$$\varphi=\varphi_0+\varphi_1\epsilon^{\alpha}, \tag{3.89}$$

when real $(\alpha)>0$. [See (2.34).] Thus (3.87) does not apply to the last line of Table I. The entries in the left and right side of Table I are, however, separately closed under (3.86). The above remarks will apply to the entire Boussinesq sequence.

Now by a leading-order analysis it is possible to establish that all singularities of the sequence (3.40) are of the form (3.85), where k or β might vanish separately. Thus, it is required to find the values of (k,β) that are consistent with (3.85) for each equation in the sequence (3.40). With (3.85),

$$\hat{V}_0=\begin{pmatrix} z \\ \{\varphi;x\}+\tfrac{3}{2}z^2\end{pmatrix}\simeq\begin{pmatrix}\beta\epsilon^{-1} \\ \{\tfrac{3}{2}\beta^2-\tfrac{1}{2}((k+1)^2-1)\}\epsilon^{-2}\end{pmatrix}, \tag{3.90}$$

where $\epsilon=x+\epsilon(t)$. And, using (3.41),

$$p^j\hat{V}_0\simeq\hat{P}_j\begin{pmatrix}\epsilon^{-m} \\ \epsilon^{-m-1}\end{pmatrix}, \tag{3.91}$$

where $m=3j+1$, $j=0,1,2,3,...$ and $\hat{P}_j=\hat{P}_j(k,\beta,m)$, $\hat{P}_0=\hat{V}_{0j}$. Also,

$$\hat{P}_j=A_{j-1}\hat{P}_{j-1}, \tag{3.92}$$

where, from (3.41),

$$A_j=C_j^*\Omega_j^{-1}C_j\Omega_j, \tag{3.93}$$

TABLE I. $n=0$.

α_+	k	β	α_-	k	β
-1	-2	0	1	0	0
2	1	±1			

$$C_j^*=\begin{pmatrix}1 & 2\beta-m-2 \\ -3(\beta-m-3) & 3\beta^2-3(m+2)\beta+(m+2)(m+3)+1-(k+1)^2\end{pmatrix}, \tag{3.94}$$

$$\Omega_{ij}^{-1}=\begin{pmatrix}0 & 1/(m+4) \\ 1/(m+3) & 0\end{pmatrix}, \tag{3.95}$$

$$C_j=\begin{pmatrix}1 & -3(\beta+m+2) \\ 2\beta+m+3 & 3\beta^2+3(m+3)\beta+(m+2)(m+3)+1-(k+1)^2\end{pmatrix}, \tag{3.96}$$

$$\Omega_j=\begin{pmatrix}(k+m+2)(k-m) & 0 \\ 0 & \tfrac{1}{3}\end{pmatrix}, \tag{3.97}$$

and $m=3j+1$. Consider the $(j+1)$-th equation in sequence (3.40). We require that (i) the leading-order term

$$p^{j+1}\hat{V}_0\simeq\hat{P}_{j+1}\begin{pmatrix}\epsilon^{-m-3} \\ \epsilon^{-m-4}\end{pmatrix}=A_j\hat{P}_j\begin{pmatrix}\epsilon^{-m-3} \\ \epsilon^{-m-4}\end{pmatrix}, \tag{3.98}$$

$$m=3j+1$$

vanishes. Or, when

$$\varphi=\varphi_0+\varphi_{k+1}\epsilon^{k+1}+\cdots, \tag{3.99}$$

with $\varphi_0=\varphi_0(t)\neq0$

$$\frac{\varphi_t}{\varphi_x}\simeq\frac{\varphi_{0t}}{(k+1)\varphi_{k+1}}\epsilon^{-k}, \tag{3.100}$$

that

$$\text{(ii)}\quad\begin{pmatrix}[\varphi_{0t}/(k+1)\varphi_{k+1}]\epsilon^{-k} \\ 0\end{pmatrix}=\hat{P}_{j+1}\begin{pmatrix}\epsilon^{-m-3} \\ \epsilon^{-m-4}\end{pmatrix}. \tag{3.101}$$

In case (i), we have

$$A_j\hat{P}_j\equiv0, \tag{3.102}$$

which, by (3.92), includes the leading-order conditions of this type for all the preceding equations in the sequence. Therefore, it is sufficient (by recursion) to evaluate (3.102) when

$$\hat{P}_j\neq0,\quad \det|A_j|=0. \tag{3.103}$$

In case (ii) it can be shown that

$$\varphi_{0t}\simeq(k+1)\varphi_{k+1} \tag{3.104}$$

and (3.101) becomes

$$\hat{P}_{j+1}=\begin{pmatrix}c \\ 0\end{pmatrix}, \tag{3.105}$$

where

$$k=m+3. \tag{3.106}$$

In both cases (3.102) and (3.105) are polynomials in (k,β,m) that determine the allowed values of (k,β) in (3.85). The zeroth-order equation is evaluated in Table I. The first-order equation is

$$\begin{pmatrix}\varphi_t/\varphi_x \\ z_t\end{pmatrix}=\frac{1}{9}\begin{pmatrix}1 & 0 \\ 0 & -\tfrac{1}{3}D\end{pmatrix}\times\begin{pmatrix}-5z_x^2+\tfrac{5}{4}z^4+5sz^2-\tfrac{1}{3}s^2-\tfrac{4}{3}s_{xx} \\ 4z_{xxxx}-15z^2z_{xx}-15zz_x^2+\tfrac{21}{4}z^5+10sz_{xx}+10s_xz_x+5z^3s+5zs^2\end{pmatrix}. \tag{3.107}$$

For this equation we find the results in Table II, which is found by solving (3.102), (3.103), and (3.105) with $j = 0$. The complete list of singularities for this equation is found by striking the last line from Table I and adjoining it to Table II. The upper block of singularities (type 1) corresponds to solving (3.102) and (3.103) with

$$\det(C_0^* \Omega_{10}^{-1} C_0) = 0. \tag{3.108}$$

The middle block (type 2) corresponds to

$$\det(\Omega_0) = 0 \tag{3.109}$$

and the lower block (type 3) to (3.105) with $j = 0$.

We now claim that the solution of (3.102), (3.103), and (3.105) for the $(j+1)$-th equation is shown in Tables III and IV. In Tables III and IV the type 1, 2, and 3 blocks of singularities are identified as before. The following observations are straightforward to verify [using (3.86)]. Identifying blocks of singularities in Table III or IV as left (L) or right (R) and type 1, 2, or 3; then within a fixed table, we have the following.

(1) The values of (k, β) in the sets (i) (type 3, type 1L) and (ii) (type 1R, type 2R) are invariant under (3.86).

(2) (i) Any singularity of type 3 can be mapped into a singularity of type 1L by (3.86). (ii) Any singularity of type 1R can be mapped into a singularity of type 2R by (3.86).

(3) Under the transformation, $\varphi \to 1/\varphi$, (i) type 1L$\leftrightarrow$type 1R and (ii) type 2L$\leftrightarrow$type 2R.

(4) Since the singularities of type 2L correspond (with $m = 3j + 1$) identically to what would be the type 3 with $m = 3(j-1) + 1$, every singularity of type 2L(j) can (by observation 2) be mapped into a singularity of type 1L$(j-1)$. Recall that to obtain all the singularities of the $(j+1)$-th equation it is required to adjoin the types obtained from Tables III or IV with $m \to m - 3$, $m - 6$, etc., deleting in each instance the type 3 block.

(5) By a recursive application of observations (2)–(4) all the singularities described in Tables III and IV can be mapped into the first line of Table I.

Now it is easy to show that any singularity of Eq. (3.40) with $k = -2, \beta = 0$, is (1) meromorphic and (2) depends on the maximum number of arbitrary "constants" allowed for by the differential equation. (See Sec. II and Ref. 4.) By the obvious reconstructions, all the singularities mapped by (3.86) and (3.87) into the one with $(k = 2, \beta = 0)$ will be meromorphic. Therefore, if the claim that Tables III and IV represent the general forms of allowed singularities is valid, the above remarks demonstrate that the sequence (3.40), and, by implication, the Boussinesq sequence, identically posses the Painlevé property.

TABLE II. First-order equation.

α_+	k	β	α_-	k	β
-7	-8	0	7	6	0
-4	-5	± 1	4	3	∓ 1
-1	-2	± 2	1	0	∓ 2
2	1	± 3	-2	-3	∓ 3
2	1	± 1	-2	-3	± 1
5	4	0			
5	4	± 2			
5	4	± 4			

TABLE III. $(j+1)$-th equation, $m = 3j + 1$ even.

α_+	k	β	α_-	k	β
$-2m-5$	$-2m-6$	0	$2m+5$	$2m+4$	0
$-2m-2$	$-2m-3$	± 1	$2m+2$	$2m+1$	∓ 1
$-2m+1$	$-2m$	± 2	$2m-1$	$2m-2$	∓ 2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m+1$	m	$\pm(m+2)$	$-m-1$	$-m-2$	$\pm(m+2)$
$m+1$	m	0	$-m-1$	$-m-2$	0
$m+1$	m	± 2	$-m-1$	$-m-2$	∓ 2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m+1$	m	$\pm m$	$-m-1$	$-m-2$	$\mp m$
$m+4$	$m+3$	± 1			
$\vdots$	$\vdots$	$\vdots$			
$m+4$	$m+3$	$\pm(m+1)$			
$m+4$	$m+3$	$\pm(m+3)$			

The preceding remarks show that Tables III and IV contain allowed forms of singularities [values of (k, β)] for the $(j+1)$-th equation. We show now that, according to the degrees of various polynomials in β defined by conditions (3.102), (3.103), and (3.105), the tables contain every solution (k, β) of these conditions.

For singularities of type 1 it is found from (3.93)–(3.95) and (3.103) that $\det|C_j|$ vanishes when

$$k + 1 = \pm(3\beta + 2m + 5), \tag{3.110}$$

TABLE IV. $(j+1)$-th equation, $m = 3j + 1$ odd.

α_+	k	β	α_-	k	β
$-2m-5$	$-2m-6$	0	$2m+5$	$2m+4$	0
$-2m-2$	$-2m-3$	± 1	$2m+2$	$2m+1$	∓ 1
$-2m+1$	$-2m$	± 2	$2m-1$	$2m-2$	∓ 2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m+1$	m	$\pm(m+2)$	$-m-1$	$-m-2$	$\pm(m+2)$
$m+1$	m	± 1	$-m-1$	$-m-2$	∓ 1
$m+1$	m	± 3	$-m-1$	$-m-2$	∓ 3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m+1$	m	$\pm m$	$-m-1$	$-m-2$	$\mp m$
$m+4$	$m+3$	0			
$\vdots$	$\vdots$	$\vdots$			
$m+4$	$m+3$	$\pm(m+1)$			
$m+4$	$m+3$	$\pm(m+3)$			

and then $\det|C_j^*|$ vanishes when

$$k+1=\pm(3\beta-2m-5). \tag{3.111}$$

For singularities of type 2, $\det|\Omega_j|$ vanishes when

$$k+1=\pm(m+1), \tag{3.112}$$

and for singularities of type 3, by (3.106),

$$k+1=m+4. \tag{3.113}$$

Therefore k is either a linear or constant function of β and substitution for k determines (3.102), (3.103), and (3.105) as polynomial conditions for β, which depend on the index m. In all cases, by (3.90),

$$\hat{V}_{0j}\simeq\begin{pmatrix}\beta\\ \beta^2\end{pmatrix}, \tag{3.114}$$

where the equivalence indicates the highest power of β in an expression. For type 1, by (3.93) to (3.97); and (3.110), (3.111),

$$A_i\simeq\begin{pmatrix}\beta^3 & \beta^2\\ \beta^4 & \beta^3\end{pmatrix}, \tag{3.115}$$

for $i=0,1,...,j-1$. Now, for (3.110) with $\det|C_j|=0$,

$$C_j\Omega_j\simeq\begin{pmatrix}\beta^2 & \beta\\ \beta^3 & \beta^2\end{pmatrix}, \tag{3.116}$$

and by the above,

$$A_j\hat{P}_j\simeq\begin{pmatrix}\beta^{3j+3}\\ \beta^{3j+4}\end{pmatrix}. \tag{3.117}$$

When $\det|C_j^*|=0$, by (3.93) and (3.111),

$$C_j^*\Omega_{ij}^{-1}C_j\Omega_j\simeq\begin{pmatrix}\beta^3 & \beta^2\\ \beta^4 & \beta^3\end{pmatrix}, \tag{3.118}$$

and

$$A_j\hat{P}_j\simeq\begin{pmatrix}\beta^{3j+4}\\ \beta^{3j+5}\end{pmatrix}. \tag{3.119}$$

Now using the definition of m,

$$m=3j+1, \tag{3.120}$$

condition (3.117) determines $m+2$, and condition (3.119) $m+3$ solutions for β which equals the number $(2m+5)$ of (allowed) solutions of type 1 in a column of Table III or IV. The separate determinations of $k+1$ in (3.110) or (3.111) complete the left or right columns.

For singularities of type 2, by (3.112),

$$A_i\simeq\begin{pmatrix}\beta & \beta^2\\ \beta^2 & \beta^3\end{pmatrix}, \tag{3.121}$$

for $i=0,1,...,j-1$, and

$$\Omega_j\simeq\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}. \tag{3.122}$$

By the above,

$$A_j\hat{P}_j\simeq\begin{pmatrix}0\\ \beta^{3j+2}\end{pmatrix}, \tag{3.123}$$

which determines $m+1=3j+2$ solutions. This is equal to the number of type 2 solutions in Tables III or IV, where the separate determinations of $k+1$ in (3.112) complete the left or right columns.

For singularities of type 3, (3.121) is valid for $i=0, 1,2,...,j,j+1$ and

$$\hat{P}_{j+1}\simeq\begin{pmatrix}C\\ 0\end{pmatrix}\simeq\begin{pmatrix}\beta^{3j+4}\\ \beta^{3j+5}\end{pmatrix}. \tag{3.124}$$

This determines $3j+5=m+4$ solutions which equals the number of type 3 singularities in Table III or IV.

Therefore all singularities have been accounted for and the Boussinesq sequence has the Painlevé property.

ACKNOWLEDGMENTS

This work was supported by the Department of Energy Contract No. DOE-DE-AC03-81ER10923 and Air Force Office of Scientific Research Grant No. AFOSR 83-0095.

APPENDIX A: THE NONLINEAR SCHRÖDINGER EQUATION

The nonlinear Schrödinger (NLS) equation

$$iU_t+U_{xx}+2U|U|^2=0 \tag{A1}$$

may be written as the system[3]

$$iU_t+U_{xx}+2U^2V=0,$$
$$-iV_t+V_{xx}+2UV^2=0, \tag{A2}$$

which reduces to (A1) with the identification

$$V=U^*. \tag{A3}$$

The system (A2) has the Painlevé property[3] with expansions

$$U=\varphi^{-1}\sum_{j=0}^{\infty}U_j\varphi^j,\quad V=\varphi^{-1}\sum_{j=0}^{\infty}V_j\varphi^j, \tag{A4}$$

and resonances at

$$j=-1,0,3,4. \tag{A5}$$

The Bäcklund transformation is

$$U=U_0/\varphi+U_1,\quad V=V_0/\varphi+V_1 \tag{A6}$$

which determines the following system of equations for (φ, U_0, V_0, U_1, V_1):

$$\begin{aligned}
&U_0V_0=-\varphi_x^2,\\
&4\varphi_x^2U_1-2U_0^2V_1=-i\varphi_tU_0-2\varphi_xU_{0x}-\varphi_{xx}U_0,\\
&-2V_0^2U_1+4\varphi_x^2V_1=i\varphi_tV_0-2\varphi_xV_{0x}-\varphi_{xx}V_0;\\
&iU_{0t}+U_{0xx}+2V_0U_1^2+4U_0U_1V_1=0,\\
&-iV_{0t}+V_{0xx}+2U_0V_1^2+4V_0V_1U_1=0,\\
&iU_{1t}+U_{1xx}+2U_1^2V_1=0,\\
&-iV_{1t}+V_{1xx}+U_1V_1^2=0.
\end{aligned} \tag{A7}$$

Taking into account the resonances at $j=0$, 3, (A7) is, effectively, a system of "six" equations for the five variables (φ, U_0, V_0, U_1, V_1).

From (A7) it is found that

$$\begin{aligned}
&U_0V_0=-\varphi_x^2,\\
&U_0V_1+V_0U_1=\varphi_{xx},\\
&U_0V_1-V_0U_1=-i\varphi_t+(V_0U_{0x}-U_0V_{0x})/\varphi_x,\\
&2i(\varphi_t/\varphi_x)=(V_0U_{0x}-U_0V_{0x})/\varphi_x^2+\lambda,\\
&U_1V_1=-\frac{1}{4}\left\{\left(\frac{\varphi_t}{\varphi_x}\right)^2+\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2+2i\lambda\frac{\varphi_t}{\varphi_x}-\lambda^2\right\},
\end{aligned} \tag{A8}$$

and

$$\frac{\partial}{\partial t}\left(\frac{\varphi_t}{\varphi_x}\right)+\frac{\partial}{\partial x}\left[\{\varphi;x\}-\frac{3}{2}\left(\frac{\varphi_t}{\varphi_x}\right)^2-2i\lambda\frac{\varphi_t}{\varphi_x}+\frac{\lambda^2}{2}\right]=0, \tag{A9}$$

where λ is a constant of integration. The above system of equations were studied in Ref. 3 and further applied in Ref. 6 to derive the Hirota formulation of the NLS equation from the Bäcklund transformation (A6).

In this section we will find a scalar Lax pair for the NLS equation by "linearizing" the Miura transformation from the modified NLS to NLS equation. For this purpose it is convenient to let

$$\lambda = 2i\beta, \quad W=\varphi_{xx}/\varphi_x, \quad \Omega = \varphi_t/\varphi_x - 2\beta, \tag{A10}$$

which obtains from (A9) the system of modified NLS equations

$$W_t = \frac{\partial}{\partial x}(\Omega_x + W\Omega + 2\beta W),$$
$$\Omega_t = -\frac{\partial}{\partial x}(W_x - \frac{1}{2}W^2 - \tfrac{3}{2}\Omega^2 - 2\beta\Omega). \tag{A11}$$

By reduction of (A8),

$$-4U_1V_1 = W^2+\Omega^2,$$
$$\frac{U_{1x}}{U_1} = \frac{(W-i\Omega)_x}{W-i\Omega} - i(\Omega+\beta), \tag{A12}$$
$$\frac{V_{1x}}{V_1} = \frac{(W+i\Omega)_x}{W+i\Omega} + i(\Omega+\beta),$$

which is a Miura transformation from (A11) to (A2). Now let

$$G = W-i\Omega, \quad H=W+i\Omega, \tag{A13}$$

and find

$$-4U_1V_1 = GH,$$
$$\frac{U_{1x}}{U_1} = \frac{G_x}{G} + \frac{G-H}{2} - i\beta, \tag{A14}$$
$$\frac{V_{1x}}{V_1} = \frac{H_x}{H} + \frac{H-G}{2} + i\beta.$$

The substitutions

$$G = 2i(U_1/\alpha), \quad H = 2iV_1\alpha \tag{A15}$$

reduce (A14) to a Ricati-type equation

$$\alpha_x + iV_1\alpha^2 + i\beta\alpha - iU_1 = 0, \tag{A16}$$

that is linearized by

$$\alpha = -(i/V_1)(h_x/h) \tag{A17}$$

to

$$h_{xx} + (i\beta - V_{1x}/V_1)h_x + U_1V_1h = 0. \tag{A18}$$

Substitution of (A13), (A15), and (A17) into (A11) obtains

$$ih_t = h_{xx} + 2U_1V_1h + 2i\beta h_x. \tag{A19}$$

By (A18)

$$ih_t = (V_{ix}/V_1 + i\beta)h_x + U_1V_1h. \tag{A20}$$

Here, (A18) and (A20) constitute a Lax pair for the NLS system (A2) in the sense that

$$h_{txx} = h_{xxt} \tag{A21}$$

requires that

$$i(U_1V_1)_t + (U_1V_1)_{xx} = 2\frac{\partial}{\partial x}\left(U_1V_1\frac{V_{1x}}{V_1}\right),$$
$$i\left(\frac{V_{1x}}{V_1}\right)_t = \frac{\partial}{\partial x}\left(\left(\frac{V_{1x}}{V_1}\right)_x + \left(\frac{V_{1x}}{V_1}\right)^2 + 2U_1V_1\right), \tag{A22}$$

which is "equivalent" to the system (A2). With

$$A = V_{1x}/V_1, \quad B = U_1V_1, \tag{A23}$$

Eqs. (A22) are

$$iA_t = \frac{\partial}{\partial x}(A_x + A^2 + 2B),$$
$$iB_t = \frac{\partial}{\partial x}(-B_x + 2AB), \tag{A24}$$

and the Miura transformation from (A11) is

$$-4B = W^2 + \Omega^2,$$
$$A = (W+i\Omega)_x/(W+i\Omega) + i(\Omega+\beta). \tag{A25}$$

Now after a Galilean transformation,

$$t\to t, \quad x\to x-2\beta t, \quad \Omega=\varphi_t/\varphi_x, \quad W=\varphi_{xx}/\varphi_x, \tag{A26}$$

and

$$W_t = \frac{\partial}{\partial x}(\Omega_x + W\Omega),$$
$$\Omega_t = -\frac{\partial}{\partial x}\left(W_x - \frac{1}{2}W^2 - \frac{3}{2}\Omega^2\right). \tag{A27}$$

At first inspection Eq. (A27) would seen to be nearly the modified Boussinesq equations, (2.18). However, a simple calculation determines that Eqs. (A27) have no discrete invariances, i.e., no transformations of the form

$$\begin{pmatrix} W \\ \Omega \end{pmatrix} = A\begin{pmatrix} W' \\ \Omega' \end{pmatrix} \tag{A28}$$

that preserve the form of Eqs. (A27). Equations (A27) identically possess the Painlevé property with expansions

$$W = \epsilon^{-1}\sum_{j=0}^{\infty} W_j\epsilon^j, \quad \Omega = \epsilon^{-1}\sum_{j=0}^{\infty}\Omega_j\epsilon^j \tag{A29}$$

and resonances at

$$j = -1,2,2,3 \tag{A30}$$

From (A29):

$$\text{(i)} \quad \Omega_0 = 0, \quad W_0 = -2; \tag{A31}$$

$$\text{(ii)} \quad \Omega_0^2 = -1, \quad W_0 = 1. \tag{A32}$$

As was the case for the modified Boussinesq equations a transformation

$$A\pm = \begin{pmatrix} -\frac{1}{2} & \pm\frac{3}{2}i \\ \pm\frac{1}{2}i & -\frac{1}{2} \end{pmatrix} \tag{A33}$$

interchanges the "leading-order" vectors

$$\begin{pmatrix} W_0 \\ \Omega_0 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ i \end{pmatrix}, \begin{pmatrix} 1 \\ -i \end{pmatrix}. \tag{A34}$$

However, the substitution (A28) and (A33) is not invariant for (A27). Therefore, the method of analysis that was developed for the Boussinesq sequence is not directly applicable to the NLS sequence.

APPENDIX B: RATIONAL SOLUTIONS

One consequence of a discrete symmetry group (Bäcklund transformation) for the "modified" equations is the induced Bäcklund transformation for the "singular manifold" equation. [See (3.76).] This Bäcklund transformation [combined with the Moebius transformation (3.50)] determines a simple method for iteratively constructing rational and other special solutions of the equations under consideration. Therefore, discrete symmetries (of modified equations) are a sufficient condition (by construction) for the existence of sequences of "rational" solutions. We conjecture that a necessary condition (for rational solutions) is the occurrence of a nondegenerate Bäcklund transformation for the "modified" equations. This would imply, by the results of Appendix C, that the NLS equations (A2) have no (nontrivial) sequences of rational solutions. Effectively, the only direct (known) Bäcklund transformation for Eq. (A9) is the Moebius group (3.50), which is not sufficient for the iterative construction of solutions. In this section rational solutions are iteratively defined for the "Boussinesq" equation

$$\frac{\partial}{\partial t}\left(\frac{\varphi_t}{\varphi_x}\right)+\frac{1}{3}\frac{\partial}{\partial x}\left(\{\varphi;x\}+\frac{3}{2}\left(\frac{\varphi_t}{\varphi_x}\right)^2\right)=0, \tag{B1}$$

where

$$\{\varphi;x\}=\frac{\partial}{\partial x}\left(\frac{\varphi_{xx}}{\varphi_x}\right)-\frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x}\right)^2. \tag{B2}$$

Equation (B1) is invariant under the Moebius group

$$\varphi=(a\psi+b)/(c\psi+d) \tag{B3}$$

and the Bäcklund transformation

$$\begin{aligned}\frac{\varphi_{xx}}{\varphi_x}&=-\frac{1}{2}\frac{\psi_{xx}}{\psi_x}\mp\frac{3}{2}\frac{\psi_t}{\psi_x},\\ \frac{\varphi_t}{\varphi_x}&=\pm\frac{1}{2}\frac{\psi_{xx}}{\psi_x}-\frac{1}{2}\frac{\psi_t}{\psi_x}.\end{aligned} \tag{B4}$$

Now composing (B3) and (B4), where

$$\begin{aligned}\psi&=-1/\varphi_j,\\ \varphi&=\varphi_{j+1},\end{aligned} \tag{B5}$$

obtains

$$\begin{aligned}&\text{(i)}\quad \frac{\varphi_{j+1,xx}}{\varphi_{j+1,x}}=-\frac{1}{2}\frac{\partial}{\partial x}\ln\left(\frac{\varphi_{j,x}}{\varphi_j^2}\right)\mp\frac{3}{2}\frac{\varphi_{j,t}}{\varphi_{j,x}},\\ &\text{(ii)}\quad \frac{\varphi_{j+1,t}}{\varphi_{j+1,x}}=\pm\frac{1}{2}\frac{\partial}{\partial x}\ln\left(\frac{\varphi_{j,x}}{\varphi_j^2}\right)-\frac{1}{2}\frac{\varphi_{j,t}}{\varphi_{j,x}}.\end{aligned} \tag{B6}$$

From (B6) with lower sign and

$$\varphi_0=x, \tag{B7}$$

it is found that after normalization

$$\begin{aligned}\varphi_1&=x^2+2t,\\ \varphi_2&=x^4+4tx^2-4t^2,\\ \varphi_3&=(x^6+10tx^4+20t^2x^2+40t^3)/x.\end{aligned} \tag{B8}$$

By evaluation of (B6)

$$\begin{aligned}&(i)\quad \varphi_{j+1,x}=\varphi_j\varphi_{j,x}^{-1/2}\lambda_j,\\ &\text{(ii)}\quad \varphi_{j+1,t}=\pm\frac{1}{3}\varphi_j^4\varphi_{j,x}^{-2}\frac{\partial}{\partial x}(\varphi_j^{-3}\varphi_{j,x}^{3/2}\lambda_j),\end{aligned} \tag{B9}$$

where

$$\lambda_j=\left\{\prod_{k=1}^{j}\left(\frac{\varphi_{j-k,x}}{\varphi_{j-k}^2}\right)^{(-1/2)^k}\right\}^{3/2}. \tag{B10}$$

The identity

$$\lambda_j=(\varphi_{j-1,x}/\varphi_{j-1}^2)^{-3/4}\lambda_{j-1}^{-1/2} \tag{B11}$$

and recursive application of (B9) (i) obtains

$$\varphi_{j+1,x}=(\varphi_j\varphi_{j-1}/\varphi_{j-2}^2)\varphi_{j-2,x}. \tag{B12}$$

To simplify (B6), (B9), and (B12) let the meromorphic function

$$\varphi_j=P_j/Q_j, \tag{B13}$$

where (P_j, Q_j) are entire functions of (x,t). Substitutions into (B12) obtain

$$Q_{j+1}=P_{j-2}, \tag{B14}$$

$$P_{j-2}P_{j+1,x}-P_{j+1}P_{j-2,x}=P_jP_{j-1}, \tag{B15}$$

where, by (B14),

$$\varphi_j=P_j/P_{j-3}. \tag{B16}$$

Substitution of (B16) into (B6) (ii) obtains

$$P_{j-2}P_{j+1,t}-P_{j+1}P_{j-2,t}=\mp(P_{j-1}P_{j,x}-P_jP_{j-1,x}). \tag{B17}$$

Therefore, (B15) and (B17) define entire functions $P_j=P_j(x,t)$ and, from (B16), meromorphic φ_j. From (B8),

$$\begin{aligned}&P_0=x,\quad P_1=x^2+2t,\quad P_2=x^4+4tx^2-4t^2,\\ &P_3=x^6+10tx^4+20t^2x^2+40t^3.\end{aligned} \tag{B18}$$

By induction, using (B15), (B17), and (B18),

$$P_j=\sum_{k=0}^{j}C_kt^{j-k}x^{2k}, \tag{B19}$$

where (for $j>0$) the C_k are constant. By the results of Sec. II the above defines rational solutions for the Boussinesq and modified Boussinesq equations. The constructions (B15)–(B17) remain valid when, in (B6), φ_0 assumes other values than (B7). Say,

$$\varphi_0=xt$$

or

$$\varphi_0=e^{ax+bt}, \tag{B20}$$

which defines (P_0, P_1, P_2) and from (B15) to (B17), (φ_j, P_j) for $j\geqslant 3$.

Rational solutions of integrable partial differential equations have been studied for some time as "pole expansions" of the solution.[10–12] In Ref. 3 the pole expansions are derived from the (Painlevé) expansions about the singular manifold.

Our method is similar to that of Refs. 13 and 14 in that the solution is defined in terms of a polynomial in the independent variables. However, to us, the calculation based on the (Schwarzian) modified equation seems preferable in that the Bäcklund transformations apply to "general" forms of solutions and the "rational" solutions are found at the last stage of the analysis as "natural" special solutions. (See Ref. 5, Appendix B.)

APPENDIX C: DISCRETE SYMMETRIES AND REDUCTION OF MODIFIED EQUATIONS

When the modified equations have discrete symmetries they consititute a form of Bäcklund transformation that may be calculated in the following way. That is, for the modified Boussinesq equation (2.18), let

$$v = v_0\varphi^{-1} + v_1, \quad \omega = \omega_0\varphi^{-1} + \omega_1. \tag{C1}$$

The resonances of (2.18) occur at

$$j = -1,2,2,3. \tag{C2}$$

Therefore (C1) defines a system of five equations in the five unknowns, $(\varphi, v_0, \omega_0, v_1, \omega_1)$. It is found that

$$v_0 = \varphi_x, \quad \omega_0 = \sigma\varphi_x, \quad \sigma^2 = 1, \tag{C3}$$

$$v_1 = -\frac{1}{2}\frac{\varphi_{xx}}{\varphi_x} + \frac{3}{2}\sigma\frac{\varphi_t}{\varphi_x}, \quad \omega_1 = -\frac{\sigma}{2}\frac{\varphi_{xx}}{\varphi_x} - \frac{1}{2}\frac{\varphi_t}{\varphi_x}, \tag{C4}$$

and (v_1, ω_1), φ satisfy equations (2.18), (2.12), respectively. With the identification

$$\vartheta = \varphi_{xx}/\varphi_x, \quad z = \varphi_t/\varphi_x, \tag{C5}$$

(ϑ, z) satisfy (2.18) and (C4) is the symmetry (2.22).

On the other hand, the transformation

$$W = W_0\varphi^{-1} + W_1, \quad \Omega = \Omega_0\varphi^{-1} + \Omega_1 \tag{C6}$$

applied to the modified NLS equations (A27) [with resonances (C2)] obtains

$$W_0 = \varphi_x, \quad \Omega_0 = \delta\varphi_x, \quad \delta^2 = -1 \tag{C7}$$

$$W_1 = -\frac{1}{2}\left(\frac{\varphi_{xx}}{\varphi_x} + \delta\frac{\varphi_t}{\varphi_x}\right), \quad \Omega_1 = \frac{1}{2}\left(-\delta\frac{\varphi_{xx}}{\varphi_x} + \frac{\varphi_t}{\varphi_x}\right), \tag{C8}$$

where (W_1, Ω_1) satisfy (A27) and φ satisfies

$$\frac{\partial}{\partial t}\left(\frac{\varphi_t}{\varphi_x}\right) = \frac{\partial}{\partial x}\left(2\delta\frac{\partial}{\partial x}\left(\frac{\varphi_t}{\varphi_x}\right) + \{\varphi;x\} + \frac{1}{2}\left(\frac{\varphi_t}{\varphi_x}\right)^2\right). \tag{C9}$$

Now Eq. (C9) is not Eq. (A9) (with $\lambda = 0$) and (C8) is not a symmetry but a reduction of Eq. (A27), since by (C8),

$$W_1 = -\delta\Omega_1 \tag{C10}$$

and

$$\Omega_{1t} = \frac{\partial}{\partial x}(\delta\Omega_{1x} + \Omega_1^2). \tag{C11}$$

Equation (C11) is Burgers equation and Eq. (C9) is associated with (C11) by a Bäcklund transformation (see Ref. 2). Under the reduction (C8) and (C10) the Miura transformation becomes the Cole-Hopf transformation and the NLS equations (A2) are the linear diffusion equation

$$\delta U_t + U_{xx} = 0. \tag{C12}$$

Therefore the sequence of NLS/modified NLS equations contain the sequence of (higher-order) Burgers equations as a proper reduction.

[1] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983).
[2] J. Weiss, J. Math. Phys. **24**, 1405 (1983).
[3] D. V. Chudnovsky, G. V. Chudnovsky, and M. Tabor, Phys. Lett. **97**, 268 (1983).
[4] J. Weiss, J. Math. Phys. **25**, 13 (1984).
[5] J. Weiss, J. Math. Phys. **25**, 2226 (1984).
[6] P. Radmore, D. Wood, J. D. Gibbon, and M. Tabor, "The Painlevé Property and Hirota's Method," preprint Department of Applied Physics, Columbia University, 1984.
[7] A. S. Fokas and R. L. Anderson, J. Math. Phys. **23**, 1066 (1982).
[8] B. Kuperschmidt and G. Wilson, Invent. Math. **62**, 403 (1981).
[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1962), Vol. II.
[10] H. Airault, H. P. McKean, and J. Moser, Commun. Pure Appl. Math. **30**, 95 (1977).
[11] D. V. Chudnovsky and G. V. Chudnovsky, Il Nuovo Cimento B **40**, 339 (1977).
[12] I. M. Krichever, Stekl. Akad. Nauk. SSSR, **84**, 335 (1983).
[13] M. Adler and J. Moser, Commun. Math. Phys. **61**, 1 (1978).
[14] M. Ablowitz and H.Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981), Chap. 3, Sec. 4.

Solution of the Cauchy problem for a generalized sine–Gordon equation

Marco Boiti, Jérôme J. P. Leon,[a] and Flora Pempinelli
Dipartimento di Fisica, Università di Lecce, I-73100 Lecce, Italy

(Received 19 June 1984; accepted for publication 20 July 1984)

We consider a spectral problem generating a hierarchy of nonlinear evolution equations including the sine-Gordon equation and a physically interesting generalization in the laboratory coordinates. The direct and inverse problems are treated. The time evolution of the spectral data is explicitly given and, therefore, the Cauchy problem for the related equations is solved.

I. INTRODUCTION

The sine–Gordon equation in laboratory coordinates (SGE)

$$\omega_{tt} - \omega_{xx} + \sin\omega = 0, \quad \omega = \omega(x,t): \mathbb{R}\times\mathbb{R}\to\mathbb{C}, \qquad (1.1)$$

is one of those nonlinear evolution equations in $1+1$ dimensions that has attracted most interest in the last decades.

In a previous paper,[1] we have shown that the SGE belongs to a hierarchy of nonlinear evolution equations (NEE's) that is generated by a recursion operator and its inverse. Following the general method developed in Ref. 2, the canonical (geometrical) structure of the hierarchy, together with the Bäcklund transformation and permutability theorem for all equations in the hierarchy have been derived.[1] In the class of the NEE's that are isospectral deformation equations of a given spectral problem, the SGE is exceptional because the NEE's in the two hierarchies generated by the recursion operator and its inverse are both local.

The SGE is also exceptional amongst the class of one-dimensional nonlinear relativistic models[3,4] because its Cauchy problem can be solved by means of the inverse spectral transform (IST).[5–7]

For many physical problems (such as propagation of magnetic flux in Josephson junctions[3,8]) one needs to use a so-called "perturbed" SGE involving external arbitrary functions. One way to handle the problem is to apply generalizations of the IST to nonisospectral evolutions.[9] However, to solve explicitly the problem, one must make some *a priori* assumptions that have to be adapted to each specific case. Moreover, it is often difficult to justify the validity of the assumptions. In general one may only perform a consistency check (as, for instance, a numerical experiment).

So one is naturally led to ask the following question: is it possible to find generalizations of the SGE that involve arbitrary external fields and that are still solvable by IST? In Ref. 10 the generalized solvable SGE

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right)\left(\frac{\omega_t + \nu\omega_x}{1+\nu}\right) + 4\rho(1+\nu)\sin\omega = 0 \qquad (1.2)$$

has been derived. The external fields $\nu(x,t)$ and $\rho(x,t)$ belong to some space of functions that will be defined later [see (1.9) below] and must verify the conservation equation

$$\rho_t + (\nu\rho)_x = 0. \qquad (1.3)$$

The SGE in the form (1.1) is recovered from (1.2) by setting $\nu = 1$ and $\rho = \frac{1}{16}$. The general one-soliton solution of (1.2) has been studied.[10] For the choices $\nu(x,t) = 1$ and $16\rho(x,t) = 1 + \mathscr{V}(x-t)$, (1.2) becomes

$$\omega_{tt} - \omega_{xx} + (1 + \mathscr{V}(x-t))\sin\omega = 0 \qquad (1.4)$$

and has been shown,[10] under some conditions on $\mathscr{V}$, to be a model for the motion of a soliton in the external electromagnetic potential $4\mathscr{V}(x-t)$.

To provide a complete solution of (1.2), one has to solve explicitly the associated spectral transform, which is the purpose of the present work. More precisely, we solve the direct spectral problem (Sec. II), the inverse spectral problem (Sec. III), and the evolution of the spectrum (Sec. IV) for the spectral problem proposed in Ref. 11 by Boiti and Tu:

$$\mathsf{F}_x = \mathsf{U}\mathsf{F}, \quad \mathsf{U} = -i\lambda\sigma_3 + u\sigma_1 + i\lambda^{-1}(s\sigma_3 + iv\sigma_2), \qquad (1.5)$$

where the σ_i's are the Pauli matrices and λ is the spectral parameter. In (1.5), the three fields $u(x,t)$, $v(x,t)$, and $s(x,t)$ obey the behaviors

$$u(x,t)\to 0, \quad v(x,t)\to 0, \quad x\to\pm\infty, \qquad (1.6)$$

$$s(x,t)\rho^{-1}(x,t)\to 1, \quad x\to\pm\infty, \qquad (1.7)$$

$\rho(x,t)$ being an arbitrary function. Moreover (and this is crucial for our task) the fields v and s verify the reduction

$$s^2(x,t) - v^2(x,t) = \rho^2(x,t). \qquad (1.8)$$

One must notice that the asymptotic behavior (1.7) is much more general than what is usual in the context of direct and inverse spectral problems. An example of the deep changes induced by the modification of the asymptotic behavior can be found in Refs. 12 and 13, which deal with the Zakharov–Shabat spectral problem. Actually it will be seen that the spectral theory can still be constructed for [(1.5) $\div$ (1.8)] for $\rho(x,t)$ being a bounded, strictly positive, integrable but otherwise arbitrary, real function

$$\rho(x,t): \mathbb{R}\times\mathbb{R}\to\mathbb{R}/\exists\, M\in\mathbb{R}^+:$$

$$\forall(x,t)\in\mathbb{R}\times\mathbb{R}, \quad 0<\rho(x,t)<M. \qquad (1.9)$$

Let us also remark that, by performing on (1.2) the change of variables

$$x\to y = \tfrac{1}{2}[f(x,t) + x + t],$$

$$t\to\tau = \tfrac{1}{2}[f(x,t) - x - t], \qquad (1.10)$$

where

$$f(x,t) = 16\left[\int_0^x dz\,\rho(z,t) - \int_0^t d\xi\,\rho(0,\xi)\nu(0,\xi)\right], \qquad (1.11)$$

[a] Permanent address: Laboratoire de Physique Mathèmatique, U. S. T. L., 34060 Montpellier, France.

 0022-2488/85/020270-06$02.50

with $\rho(x,t)$ and $\nu(x,t)$ satisfying (1.3), we obtain the SGE in the form (1.1) in the variables (y,τ).

The map (1.10) transforms the Cauchy problem for (1.2)—initial data prescribed on $t=0$—into an initial-value problem for (1.1) with data prescribed on the curve $x+t=f(x,t)$.

The equivalence between two different initial-value problems for the SGE, the Cauchy and the Goursat problems has been studied in Refs. 14 and 15. The difficulties found in stating an exact equivalence between these two-initial value problems related by a trivial transformation suggest that the only way to handle the Cauchy problem for (1.2) is to study directly the spectral transform for the spectral problem (1.5).

II. THE DIRECT SPECTRAL PROBLEM

The problem consists in defining the so-called spectral data, say $\mathscr{S}(\lambda,t)$, associated to the potentials $u(x,t)$, $v(x,t)$, and $s(x,t)$, and to inspect their analytic properties in the λ-plane. A possible parametric t-dependence is understood everywhere.

We first write the spectral equation (1.5) in the following more convenient form:

$$\frac{\partial}{\partial x}\mathsf{F}(\lambda,x)=[\mathsf{A}(\lambda,x)+\mathsf{B}(\lambda,x)]\mathsf{F}(\lambda,x)\,, \tag{2.1}$$

where

$$\mathsf{A}(\lambda,x)=-i\sigma_3(\lambda-\lambda^{-1}\rho(x))\,, \tag{2.2}$$

$$\mathsf{B}(\lambda,x)=u(x)\sigma_1-2i\lambda^{-1}\rho(x)\sin\tfrac{1}{2}\omega(x)\mathsf{M}(x)\sigma_2\,, \tag{2.3}$$

with

$$\mathsf{M}(x)=\exp(-\tfrac{1}{2}i\sigma_1\omega(x))\,. \tag{2.4}$$

In the above formulas, the function $\omega(x)$ has been defined by the equations

$$s(x)=\rho(x)\cos\omega(x),\quad v(x)=i\rho(x)\sin\omega(x)\,, \tag{2.5}$$

that fit the relation (1.8). Taking into account the condition (1.9), the asymptotic behaviors (1.6) and (1.7) are equivalent to

$$u(x)\to 0,\quad \omega(x)\to 2n^{(\pm)}\pi,\quad x\to\pm\infty\,, \tag{2.6}$$

for arbitrary $n^{(\pm)}$ integers. So we have

$$\mathsf{B}(\lambda,x)\to 0\quad\text{as } x\to\pm\infty\,. \tag{2.7}$$

Then we seek a matrix solution Ψ of (2.1) that satisfies the following behaviors:

$$\mathsf{X}^{-1}(\lambda,x)\Psi(\lambda,x)\to\mathbf{1},\quad x\to+\infty\,, \tag{2.8}$$

with

$$\mathsf{X}(\lambda,x)=\exp[-\mathrm{i}\sigma_3\zeta(\lambda,\mathrm{x})]\,, \tag{2.9}$$

$$\zeta(\lambda,x)=\lambda x-\lambda^{-1}\int_0^x\rho(z)dz\,. \tag{2.10}$$

[Note that $\mathsf{X}(\lambda,x)$ is a solution of (2.1) for $\mathsf{B}\equiv 0$.] In order to prove the existence of the solution Ψ, and furthermore obtain its analytical properties, we use the fact the spectral problem (2.1), together with the behaviors (2.7) and (2.8), is equivalent to the integral equation

$$\Psi(\lambda,x)=\mathsf{X}(\lambda,x)-\int_x^\infty dy\,\mathsf{X}(\lambda,x)\mathsf{X}^{-1}(\lambda,y)\mathsf{B}(\lambda,y)\Psi(\lambda,y)\,. \tag{2.11}$$

Splitting the matrix Jost solution Ψ in its two column vectors

$$\Psi(\lambda,x)=(\psi_1(\lambda,x),\psi_2(\lambda,x)) \tag{2.12}$$

then (2.11) gives, say for ψ_2:

$$\begin{aligned}\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]&\\ =\begin{pmatrix}0\\1\end{pmatrix}-\int_x^\infty dy&\begin{pmatrix}\exp[2i(\zeta(\lambda,y)-\zeta(\lambda,x))] & 0\\ 0 & 1\end{pmatrix}\\ &\times\mathsf{B}(\lambda,y)\psi_2(\lambda,y)\exp[-i\zeta(\lambda,y)]\,.\end{aligned} \tag{2.13}$$

Following Ref. 7 we also write the following integral equation for the gauge-transformed matrix $\tilde{\Psi}$:

$$\tilde{\Psi}(\lambda,x)=\mathsf{M}^{-1}(x)\Psi(\lambda,x)\,, \tag{2.14}$$

$$\begin{aligned}\tilde{\Psi}(\lambda,x)=&(-1)^{n^{(+)}}\mathsf{X}(\lambda,x)\\ &-\int_x^\infty dy\,\mathsf{X}(\lambda,x)\mathsf{X}^{-1}(\lambda,y)\tilde{\mathsf{B}}(\lambda,y)\Psi(\lambda,y)\,,\end{aligned} \tag{2.15}$$

where

$$\tilde{\mathsf{B}}(\lambda,x)=(u(x)+\tfrac{1}{2}i\omega(x))\sigma_1-2\lambda\sin(\tfrac{1}{2}\omega(x))\sigma_2\mathsf{M}(x)\,. \tag{2.16}$$

We note that while $\mathsf{B}(\lambda,x)$ is linear in λ^{-1}, $\tilde{\mathsf{B}}(\lambda,x)$ is linear in λ, which will allow us to obtain from (2.13) and (2.15), respectively, the behaviors of $\Psi(\lambda,x)$ as $\lambda\to\infty$ and $\lambda\to 0$.

Using (1.9) we derive the property

$$\{y>x,\operatorname{Im}(\lambda)>0\}\Rightarrow\{\operatorname{Im}(\zeta(\lambda,y)-\zeta(\lambda,x))>0\}\,. \tag{2.17}$$

We define the norm of a vector V by

$$|V|=\sum_i|V_i|\,, \tag{2.18}$$

and the norm of a matrix C by

$$|\mathsf{C}|=\sum_{ij}|C_{ij}|\,. \tag{2.19}$$

Therefore, under the assumptions

$$\int_{-\infty}^{+\infty}dy|\mathsf{B}(\lambda,y)|<\infty,\quad\int_{-\infty}^{+\infty}dy|\tilde{\mathsf{B}}(\lambda,y)|<\infty\,, \tag{2.20}$$

the Volterra equations (2.13) for $\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]$ and (2.15) for $\tilde{\psi}_2(\lambda,x)\exp[-i\zeta(\lambda,x)]$ can be solved as Neumann series. Consequently, the following properties for the Jost solution $\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]$ are satisfied.

(i) It exists.

(ii) It obeys the following bound:

$$\left|\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]-\begin{pmatrix}0\\1\end{pmatrix}\right|<\exp\left[\int_x^\infty dy|\mathsf{B}(\lambda,y)|\right]. \tag{2.21}$$

(iii) It is continuous on the real λ-axis and possesses the following behavior as $\lambda\to 0$ ($\operatorname{Im}\lambda\geqslant 0$):

$$\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]=(-1)^{n^{(+)}}\mathsf{M}(x)\begin{pmatrix}0\\1\end{pmatrix}+O(\lambda)\,. \tag{2.22}$$

(iv) It is analytic in the upper half-λ-plane and possesses the following behavior as $|\lambda|\to\infty$ ($\operatorname{Im}\lambda\geqslant 0$):

$$\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]$$
$$=\begin{pmatrix}0\\1\end{pmatrix}-\frac{1}{2}i\lambda^{-1}\begin{pmatrix}u(x)\\ \int_x^\infty dy\left[u^2(y)-4\rho(y)\sin^2\frac{1}{2}\omega(y)\right]\end{pmatrix}$$
$$+O(\lambda^{-2}). \tag{2.23}$$

The same analysis can be made in the lower half-λ-plane for the Jost solution $\psi_1(\lambda,x)\exp[i\zeta(\lambda,x)]$. In particular, for Im $\lambda<0$, we get, as $\lambda\to 0$,

$$\psi_1(\lambda,x)\exp[i\zeta(\lambda,x)]=(-1)^{n^{(+)}}\mathsf{M}(x)\begin{pmatrix}1\\0\end{pmatrix}+O(\lambda), \tag{2.24}$$

and as $|\lambda|\to\infty$,

$$\psi_1(\lambda,x)\exp[i\zeta(\lambda,x)]$$
$$=\begin{pmatrix}1\\0\end{pmatrix}+\frac{1}{2}i\lambda^{-1}$$
$$\times\begin{pmatrix}\int_x^\infty dy\left[u^2(y)-4\rho(y)\sin^2\frac{1}{2}\omega(y)\right]\\ u(x)\end{pmatrix}+O(\lambda^{-2}). \tag{2.25}$$

Due to the zero-trace of $(\mathsf{A}+\mathsf{B})$, the determinant of $\Psi(\lambda,x)$ does not depend on x and therefore

$$\det\Psi(\lambda,x)=\det \mathsf{X}(\lambda,x)=1. \tag{2.26}$$

Consequently $\Psi(\lambda,x)$ is a fundamental matrix solution of the first-order differential equation (2.1).

We define as usual another Jost solution $\Phi(\lambda,x)$ through its asymptotic behavior, as $x\to-\infty$

$$\mathsf{X}^{-1}(\lambda,x)\,\Phi(\lambda,x)\to\mathbf{1},\quad x\to-\infty, \tag{2.27}$$

or equivalently as a solution of the integral equation

$$\Phi(\lambda,x)=\mathsf{X}(\lambda,x)$$
$$+\int_{-\infty}^x dy\,\mathsf{X}(\lambda,x)\mathsf{X}^{-1}(\lambda,y)\mathsf{B}(\lambda,y)\,\Phi(\lambda,y). \tag{2.28}$$

The procedure used preceedingly for $\psi_2(\lambda,x)\times\exp[-i\zeta(\lambda,x)]$ $(\psi_1(\lambda,x)\exp[i\zeta(\lambda,x)])$ applies to $\phi_1(\lambda,x)\exp[i\zeta(\lambda,x)]$ $(\phi_2(\lambda,x)\cdot\exp[-i\zeta(\lambda,x)])$ which is also continuous on the real axis, bounded and analytic in the upper (lower) half-λ-plane.

In particular, as $|\lambda|\to\infty$ in the upper half-plane,

$$\phi_1(\lambda,x)\exp[i\zeta(\lambda,x)]=\begin{pmatrix}1\\0\end{pmatrix}+O(\lambda^{-1}) \tag{2.29}$$

and as $|\lambda|\to\infty$ in the lower half-plane,

$$\phi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]=\begin{pmatrix}0\\1\end{pmatrix}+O(\lambda^{-1}). \tag{2.30}$$

Since they are not needed in the following, we have omitted the higher-order terms in the above expansions.

The spectral data are finally defined by expanding $\Phi(\lambda,x)$ on the basis $\Psi(\lambda,x)$, namely

$$\Phi(\lambda,x)=\Psi(\lambda,x)\mathsf{S}(\lambda),\quad \lambda\in\mathbb{R}. \tag{2.31}$$

The spectral matrix $\mathsf{S}(\lambda)$ satisfies the unitarity relation

$$\det\mathsf{S}(\lambda)=1. \tag{2.32}$$

Its diagonal elements can be expressed as

$$S_{11}(\lambda)=\det(\phi_1(\lambda,x),\psi_2(\lambda,x))$$
$$\equiv\det(\phi_1(\lambda,x)\exp[i\zeta(\lambda,x)],$$
$$\times\,\psi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]), \tag{2.33}$$
$$S_{22}(\lambda)=\det(\psi_1(\lambda,x),\phi_2(\lambda,x))$$
$$\equiv\det(\psi_1(\lambda,x)\exp[i\zeta(\lambda,x)],$$
$$\times\,\phi_2(\lambda,x)\exp[-i\zeta(\lambda,x)]). \tag{2.34}$$

Therefore $S_{11}(\lambda)$ ($S_{22}(\lambda)$) can be extended in the upper (lower) half-λ-plane and have the following asymptotic behaviors as $|\lambda|\to\infty$:

$$S_{11}(\lambda)=1+O(\lambda^{-1}),\quad \operatorname{Im}\lambda\geqslant 0, \tag{2.35}$$
$$S_{22}(\lambda)=1+O(\lambda^{-1}),\quad \operatorname{Im}\lambda\leqslant 0. \tag{2.36}$$

We assume in the following that the possible zeros of $S_{11}(\lambda)$ and $S_{22}(\lambda)$ are simple, of finite number, and not on the real axis. This is a rather strong assumption and it is in general difficult to find the conditions on the potentials which would induce such a property. However, it is in general sufficient that the potentials [in our case: $u(x)$, $v(x)$, and $s(x)\rho^{-1}(x)-1$] decrease exponentially at both ends (see, e.g., Ref. 16 for the Schrödinger spectral problem and Ref. 17 for the Zakharov–Shabat spectral problem).

Let us call

$$\lambda_n,\quad n=1,2,\ldots,N,\ \operatorname{Im}\lambda_n>0 \tag{2.37}$$

the zeroes of $S_{11}(\lambda)$ and

$$\tilde\lambda_n,\quad n=1,2,\ldots,\tilde N,\quad \operatorname{Im}\tilde\lambda_n<0 \tag{2.38}$$

the zeroes of $S_{22}(\lambda)$.

From (2.33) and (2.34), it follows then that the residues of ϕ_1/S_{11} and ϕ_2/S_{22} at $\lambda=\lambda_n$ and $\lambda=\tilde\lambda_n$ are proportional to ψ_2 and ψ_1, respectively,

$$\operatorname{res}\left\{\frac{\phi_1(\lambda,x)}{S_{11}(\lambda)}\right\}_{\lambda_n}=C_n\,\psi_2(\lambda_n,x). \tag{2.39}$$

$$\operatorname{res}\left\{\frac{\phi_2(\lambda,x)}{S_{22}(\lambda)}\right\}_{\tilde\lambda_n}=\tilde C_n\,\psi_1(\tilde\lambda_n,x), \tag{2.40}$$

For potentials $u(x)$ and $\omega(x)$ defined on a bounded support, the matrix elements S_{12} and S_{21} can be defined outside the real axis and the coefficients C_n and $\tilde C_n$ can be written as

$$C_n=\operatorname{res}\left\{\frac{S_{21}(\lambda)}{S_{11}(\lambda)}\right\}_{\lambda_n}, \tag{2.41}$$

$$\tilde C_n=\operatorname{res}\left\{\frac{S_{12}(\lambda)}{S_{22}(\lambda)}\right\}_{\tilde\lambda_n}. \tag{2.42}$$

The constants C_n and $\tilde C_n$ (that fix the values of the residues of ϕ_1/S_{11} and ϕ_2/S_{22}), the coefficients

$$R_1(\lambda)=S_{12}(\lambda)/S_{11}(\lambda),\quad \lambda\in\mathbb{R}, \tag{2.43}$$
$$R_2(\lambda)=S_{21}(\lambda)/S_{22}(\lambda),\quad \lambda\in\mathbb{R} \tag{2.44}$$

(that relate the values of ϕ_j/S_{jj} to those of ψ_j, $j=1,2$, on the real λ-axis), and the regularity on the appropriate half-plane including $\lambda=\infty$ of $\phi_j\exp[(-)^{j+1}i\zeta]S_{jj}^{-1}$ and $\psi_j\times\exp[(-)^{j+1}i\zeta]$, $j=1,2$, form a complete set of information about the analytic properties of the Jost solutions.[18]

Therefore, the complete spectral data for potentials satisfying (2.20) are

$$\mathscr{S}(\lambda) = \{R_1(\lambda), R_2(\lambda), \lambda \in \mathbb{R}; \lambda_n, C_n, N; \tilde{\lambda}_n, \tilde{C}_n, \tilde{N}\}. \tag{2.45}$$

The solution of the matrix Riemann–Hilbert problem, that is, the reconstruction of the Jost solutions from the knowledge of $\mathscr{S}(\lambda)$, is given in the subsequent section.

III. THE INVERSE SPECTRAL PROBLEM

The problem consists of obtaining the potentials $u(x)$ and $\omega(x)$ from a given set of spectral data $\mathscr{S}(\lambda)$.

We will not solve the problem of characterizing completely the spectral data, i.e., we will not give the necessary and sufficient conditions on $\mathscr{S}(\lambda)$ such that it does correspond univocally to potentials u and ω satisfying (2.20) (see, e.g., Refs. 16, 19, and 20 for the Schrödinger spectral problem on the line). We simply assume hereafter that these unspecified conditions are satisfied.

The inverse problem is solved via the following procedure. First, we introduce an orthonormal basis $\mathsf{X}(\lambda,x)$, $\lambda^{-1}\rho^{1/2}(x)\mathsf{X}(\lambda,x)$ in the space of distributions and use it to define via a Fourier-type transformation two matrices $\mathsf{K}(x,y)$ and $\mathsf{L}(x,y)$. This allows us to write the so-called triangular integral representation of the Jost solutions.

Second, by comparing the behaviors of the Jost solution $\Psi(\lambda,x)$ as $\lambda\to\infty$ and $\lambda\to 0$, on the real axis, obtained from the triangular representation with those obtained in the previous section [formulas (2.22)–(2.25)] we are able to provide explicit relations between $\mathsf{K}(x,x)$, $\mathsf{L}(x,x)$ and the two fields $u(x)$ and $\omega(x)$.

Third, we link the matrices $\mathsf{K}(x,y)$ and $\mathsf{L}(x,y)$ to the spectral data. This is done in the usual way by using the triangular representation and contour integration in the complex λ-plane. The kernel of the obtained Gel'fand–Levitan–Marchenko (GLM) integral equation is found to be the transform of the spectral data according to the previously introduced Fourier-type transformation.

Following the above outlined procedure let us, first, write the orthonormality relations

$$\int_{-\infty}^{+\infty} d\lambda\, \mathsf{X}(\lambda,x)\mathsf{X}^{-1}(\lambda,y) = 2\pi\delta(x-y)\mathbf{1}, \tag{3.1}$$

$$\int_{-\infty}^{+\infty} dy\, \lambda^{-1}\rho^{1/2}(x)\mathsf{X}(\lambda,x)\lambda^{-1}\rho^{1/2}(y)\mathsf{X}^{-1}(\lambda,y) = 2\pi\delta(x-y)\mathbf{1}, \tag{3.2}$$

$$\int_{-\infty}^{+\infty} d\lambda\, \lambda^{-1}\mathsf{X}(\lambda,x)\mathsf{X}^{-1}(\lambda,y) = 0, \tag{3.3}$$

and the completeness relation

$$\int_{-\infty}^{+\infty} dy(1+(\lambda k)^{-1}\rho(y))\mathsf{X}^{-1}(\lambda,y)\mathsf{X}(k,y) = 2\pi\delta(k-\lambda)\mathbf{1}. \tag{3.4}$$

Let us define

$$\mathsf{K}(x,y) = (2\pi)^{-1}\int_{-\infty}^{+\infty} d\lambda\,[\Psi(\lambda,x) - \mathsf{X}(\lambda,x)]\mathsf{X}^{-1}(\lambda,y) \tag{3.5}$$

and

$$\mathsf{L}(x,y) = (2\pi)^{-1}\int_{-\infty}^{+\infty} d\lambda\,\lambda^{-1}[\Psi(\lambda,x) - \mathsf{X}(\lambda,x)]\mathsf{X}^{-1}(\lambda,y), \tag{3.6}$$

From the property of $\Psi(\lambda,x)$ obtained in Sec. II it can be inferred that K and L are identically zero for $y<x$, continuous on both variables and that [use (2.17) and (2.21)]

$$\mathsf{K}(x,y)\to 0,\quad \mathsf{L}(x,y)\to 0 \quad \text{as } y\to\infty. \tag{3.7}$$

The above transformations can be inverted with the help of (3.4) and we obtain the so-called triangular representation for Ψ,

$$\Psi(\lambda,x) = \mathsf{X}(\lambda,x) + \int_x^{\infty} dy[\mathsf{K}(x,y) + \lambda^{-1}\rho(y)\mathsf{L}(x,y)]\,\mathsf{X}(\lambda,y). \tag{3.8}$$

Applying the Riemann–Lebesgue lemma to this representation, we obtain the behavior of Ψ on the real axis as $\lambda\to\infty$

$$\Psi(\lambda,x)\,\mathsf{X}^{-1}(\lambda,x) = \mathbf{1} - i\lambda^{-1}\mathsf{K}(x,x)\sigma_3 + O(\lambda^{-2}), \tag{3.9}$$

and as $\lambda\to 0$

$$\Psi(\lambda,x)\,\mathsf{X}^{-1}(\lambda,x) = \mathbf{1} + i\mathsf{L}(x,x)\sigma_3 + O(\lambda). \tag{3.10}$$

By comparing, on the real axis, (3.9) and (3.10) with (2.22) $\div$(2.25) we immediately obtain the sought for relations between K, L and the potentials u and ω:

$$\mathsf{K}(x,x) = \frac{1}{2}\int_x^{\infty}\{u^2(y) - 4\rho(y)\sin^2\tfrac{1}{2}\omega(y)\} \times dy\,\mathbf{1} - \tfrac{1}{2}u(x)\sigma_1, \tag{3.11}$$

$$\mathsf{L}(x,x) = i\sigma_3 - i(-1)^{n^{(+)}}\mathsf{M}(x)\sigma_3. \tag{3.12}$$

We are now ready to obtain the GLM integral equation which solves our Hilbert–Riemann problem. We start from (2.31), which we rewrite as

$$\left[\frac{1}{S_{11}(\lambda)}\phi_1(\lambda,x) - \begin{pmatrix}1\\0\end{pmatrix}\exp[-i\zeta(\lambda,x)]\right]\exp[i\zeta(\lambda,y)] = \left[\psi_1(\lambda,x) - \begin{pmatrix}1\\0\end{pmatrix}\exp[-i\zeta(\lambda,x)] + R_2(\lambda)\,\psi_2(\lambda,x)\right]\exp[i\zeta(\lambda,y)], \tag{3.13}$$

$$\left[\frac{1}{S_{22}(\lambda)}\phi_2(\lambda,x) - \begin{pmatrix}0\\1\end{pmatrix}\exp[i\zeta(\lambda,x)]\right]\exp[-i\zeta(\lambda,y)] = \left[\psi_2(\lambda,x) - \begin{pmatrix}0\\1\end{pmatrix}\exp[i\zeta(\lambda,x)] + R_1(\lambda)\,\psi_1(\lambda,x)\right]\exp[-i\zeta(\lambda,y)]. \tag{3.14}$$

We integrate both equations for $y>x$ along the real λ-axis. The left-hand sides, apart from poles, are analytic, respectively, on the upper and lower half-plane. Their contributions, thanks to the good behavior at $\lambda=\infty$, can be evaluated by closing the contour through infinity. The right-hand sides can be transformed by inserting in them the triangular representations for ψ_1 and ψ_2.

The final result is the desired GLM integral equation

$$K(x,y)+\theta(y-x)\Big\{G^{(0)}(x,y)+\int_x^\infty K(x,z)G^{(0)}(z,y)dz$$
$$+\int_x^\infty \rho(z)L(x,z)G^{(1)}(z,y)dz\Big\}=0\,, \tag{3.15}$$

where

$$G^{(l)}(x,y)=(2\pi)^{-1}\int_{-\infty}^{+\infty} d\lambda\,\lambda^{-l}X(\lambda,x)\,X(\lambda,y)\times\begin{pmatrix}0 & R_1(\lambda)\\ R_2(\lambda) & 0\end{pmatrix}$$
$$-i\sum_{n=1}^{N}\lambda_n^{-l}X(\lambda_n,x)X(\lambda_n,y)\begin{pmatrix}0&0\\ C_n&0\end{pmatrix}$$
$$+i\sum_{n=1}^{\tilde N}\tilde\lambda_n^{-l}X(\tilde\lambda_n,x)\,X(\tilde\lambda_n,y)\begin{pmatrix}0&\tilde C_n\\ 0&0\end{pmatrix},$$
$$l=0,1\,. \tag{3.16}$$

For real potentials $u(x)$ and $\omega(x)$ the spectral problem is invariant under the involution

$$F\to\sigma_1 F^*\sigma_1\,. \tag{3.17}$$

It follows that in this case K and L have the symmetry property

$$K=\sigma_1 K^*\sigma_1\,, \tag{3.18}$$
$$L=\sigma_1 L^*\sigma_1\,. \tag{3.19}$$

IV. EVOLUTION OF THE SPECTRAL DATA

As shown in Ref. 10, the generalized SGE (1.2) can be expressed as the compatibility condition between the spectral problem (2.1) where a parametric t-dependence is understood, and the following auxiliary spectral problem:

$$\frac{\partial}{\partial t}F(\lambda,x,t)=V(\lambda,x,t)F(\lambda,x,t)\,, \tag{4.1}$$

in which

$$V(\lambda,x,t)=-i\lambda\sigma_3+u(x,t)\sigma_1$$
$$-i\lambda^{-1}\nu(x,t)[s(x,t)\sigma_3+iv(x,t)\sigma_2]\,. \tag{4.2}$$

In (4.2), v and s satisfy the reduction condition (1.8) and ρ and ν the conservation law (1.3).

In fact, from the compatibility condition

$$U_t-V_x+[U,V]=0\,, \tag{4.3}$$

one gets the evolution equations

$$u_t=u_x+2(1+\nu)v\,, \tag{4.4}$$
$$v_t=-(\nu v)_x-2(1+\nu)us\,, \tag{4.5}$$
$$s_t=-(\nu s)_x-2(1+\nu)uv\,. \tag{4.6}$$

The reduction equation (1.8) and the conservation law (1.3) reduce the system (4.5) and (4.6), by means of the change of function (2.5), to the equation

$$2(1+\nu)u=-i(\omega_t+\nu\omega_x)\,, \tag{4.7}$$

while Eq. (4.4) furnishes (1.2).

Recalling the asymptotic behaviors of the Jost matrix solutions Ψ and Φ, one obtains for a generic V in (4.1)

$$\Psi_t=V\Psi-\Psi(X^{-1}V^{(+)}X-X^{-1}X_t)\,, \tag{4.8}$$
$$\Phi_t=V\Phi-\Phi(X^{-1}V^{(-)}X-X^{-1}X_t)\,, \tag{4.9}$$

where the $V^{(\pm)}$ stand for the asymptotic values of V as $x\to\pm\infty$.

Consequently, using definition (2.31), the spectral matrix evolves in time as follows:

$$S_t=(X^{-1}V^{(+)}X-X^{-1}X_t)S-S(X^{-1}V^{(-)}X-X^{-1}X_t)\,. \tag{4.10}$$

Inserting in (4.10) the explicit form (4.2) for V we obtain

$$S_t=i\Big[\lambda+\lambda^{-1}\nu(x,t)\rho(x,t)$$
$$+\lambda^{-1}\int_0^x dz\frac{\partial}{\partial t}\rho(z,t)\Big][S,\sigma_3]\,. \tag{4.11}$$

The independence of $S(\lambda)$ on x is ensured by the relation (1.3) which integrated on $[0,x]$ gives

$$\nu(x,t)\rho(x,t)+\int_0^x dz\frac{\partial}{\partial t}\rho(z,t)=\nu(0,t)\rho(0,t)\,. \tag{4.12}$$

From (4.11), in accordance with the isospectral character of the time evolution, we readily obtain

$$\frac{\partial}{\partial t}S_{11}(\lambda,t)=\frac{\partial}{\partial t}S_{22}(\lambda,t)=0\,. \tag{4.13}$$

Consequently, the positions and number of poles are time independent. The time evolution of the remaining spectral data reads

$$\frac{\partial}{\partial t}R_j(\lambda,t)=(-1)^j\tfrac{1}{2}i[\lambda+\lambda^{-1}\nu(0,t)\rho(0,t)]\,R_j(\lambda,t)\,,$$
$$j=1,2, \tag{4.14}$$

$$\frac{\partial}{\partial t}C_n(t)=\tfrac{1}{2}i[\lambda_n+\lambda_n^{-1}\nu(0,t)\rho(0,t)]C_n(t)\,,$$
$$n=1,\dots,N\,, \tag{4.15}$$

$$\frac{\partial}{\partial t}\tilde C_n(t)=\tfrac{1}{2}i[\tilde\lambda_n+\tilde\lambda_n^{-1}\nu(0,t)\rho(0,t)]\tilde C_n(t)\,,$$
$$n=1,\dots,\tilde N\,. \tag{4.16}$$

Therefore the Cauchy problem for the generalized SGE (1.2) together with the condition (1.3) can be solved by the spectral transform technique which can be schematically represented by

$$\{u(x,0),\omega(x,0)\}\xrightarrow{1}\mathscr{S}(\lambda,0)\xrightarrow{2}\mathscr{S}(\lambda,t)$$
$$\xrightarrow{3}\{u(x,t),\omega(x,t)\}\,.$$

Here 1 means solve the direct spectral problem for the potentials $u(x,0)$ and $\omega(x,0)$ in (2.1); 2 means obtain $\mathscr{S}(\lambda,t)$ from $\mathscr{S}(\lambda,0)$ by solving [(4.14) ÷ (4.16)]; and 3 means solve the inverse spectral problem for (2.1), that is, reconstruct $u(x,t)$ and $\omega(x,t)$ from the data of $\mathscr{S}(\lambda,t)$.

Let us conclude with some remarks. When $\rho(x,t)$ is a constant [then $\nu=\nu(t)$] the above technique allows us to solve all the equations of the hierarchy associated to (1.5) (see Refs. 1 and 11) obtained by choosing different traceless matrices $V(\lambda,x,t)$ with a polynomial dependence in λ and λ^{-1}.

Throughout the present work, we have assumed the asymptotic behaviors (2.6) and a strictly positive function $\rho(x,t)$ [(1.9)]. Of course the spectral problem is invariant under the transformation $\omega \to \omega + \pi$, $\rho \to -\rho$. However, for the asymptotic behaviors (2.6) in the case $\rho(x,t) < 0$, and more generally in the case of an unprescribed sign of $\rho(x,t)$, it seems to be much more difficult to control the analytic properties of the Jost solutions. Similar difficulties have been found in a Schrödinger-like spectral problem[21] and they seem to be related to the simultaneous presence in the spectral equation of terms with very different behaviors in λ.

ACKNOWLEDGMENTS

J. L. is grateful to the Dipartimento di Fisica dell'Università di Lecce for kind hospitality. M. B. and F. P. acknowledge enlighting discussions with Professor P. Caudrey about the spectral transform.

This work has been partially supported by M. P. I. and I. N. F. N., Sezione di Bari, Italy.

[1]M. Boiti, J. JP. Leon, and F. Pempinelli, J. Math. Phys. **25**, 1725 (1984).

[2]M. Boiti, F. Pempinelli, and G. Z. Tu, Nuovo Cimento B **79**, 231 (1984).

[3]See, e.g., D. W. McLaughlin and A. C. Scott, Phys. Rev. A **18**, 1652 (1978); A. R. Bishop, J. A. Krumhansl, and S. E. Trullinger, Physica D **1**, 1 (1980), and the references quoted there.

[4]T. H. R. Skyrme, Nucl. Phys. C **31**, 556 (1962); R. Dashen, B. Hasslacher, and A. Neveu, Phys. Rev. D **10**, 4114, 4130, 4138 (1974); V. E. Zakharov, L. A. Takhtadzhyan, and L. D. Faddeev, Sov. Phys. Dokl. **19**, 824 (1975); D. K. Campbell, and A. R. Bishop, Nucl. Phys. B **200**, 297 (1982).

[5]For a review on the IST one could refer to M. J. Ablowitz and H. Segur, "Solitons and the inverse scattering transform," in *SIAM Studies on Applied Mathematics* (SIAM, Philadelphia, 1981); and F. Calogero and A. Degasperis, "Spectral transform and solitons," in *Studies on Mathematical Applications* (North-Holland, Amsterdam, 1982), Vol. 13.

[6]G. L. Lamb, Jr., *Elements of Soliton Theory* (Wiley, New York, 1979); M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **30**, 1262 (1973); L. A. Takhtadzhyan and L. D. Faddeev, Theor. Math. Phys. **21**, 1046 (1974).

[7]D. J. Kaup, Stud. Appl. Math. **54**, 165 (1975).

[8]A. Barone and G. Paternò, *Physics and Applications of the Josephson Effect* (Wiley–Interscience, New York, 1982).

[9]D. J. Kaup and A. C. Newell, Proc. Soc. London A Ser. A **361**, 413 (1978); V. I. Karpman and E. M. Maslov, Sov. Phys. JETP **48**, 281 (1978); see also the first reference quoted in Ref. 3.

[10]J. JP. Leon, Phys. Rev. A **30**, 2830 (1984).

[11]M. Boiti and G. Z. Tu, Nuovo Cimento B **75**, 145 (1983).

[12]J. JP. Leon, J. Math. Phys. **21**, 2572 (1979).

[13]M. Boiti and F. Pempinelli, Nuovo Cimento B **69**, 213 (1982).

[14]D. J. Kaup and A. C. Newell, J. Appl. Math. **34**, 37 (1978).

[15]L. A. Taktadzhyan and L. D. Faddeev, Proc. Steklov Inst. Math **3**, 277 (1979).

[16]K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1977).

[17]N. Asano and Y. Kato, J. Math. Phys. **22**, 2780 (1981); H. M. A. de Castro and W. F. Wreszinki, J. Math. Phys. **24**, 1502 (1983).

[18]P. J. Caudrey, Physica D **6**, 51 (1982).

[19]P. Deift and E. Trubowitz, Pure Appl. Math. **32**, 121 (1979).

[20]P. C. Sabatier, "Rational reflection coefficients in one dimensional inverse scattering and applications," in *Proceedings of the TULSA Conference* (SIAM, Philadelphia, 1983); P. C. Sabatier, "Critical analysis of the mathematical methods used in electromagnetic inverse theories," to be published by J. Reidel in Nato workshops.

[21]M. Boiti, P. J. Caudrey, and F. Pempinelli, Nuovo Cimento B **83**, 71 (1984).

Weakly convergent expansions of a plane wave and their use in Fourier integrals

E. Joachim Weniger[a)]

Quantum Theory Group, Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

(Received 13 January 1984; accepted for publication 28 June 1984)

The Fourier transform of an irreducible spherical tensor is normally computed with the help of the Rayleigh expansion of a plane wave in terms of spherical Bessel functions and spherical harmonics. The angular integrations are then trivial. However, the remaining radial integral containing a spherical Bessel function may be so complicated that the applicability of Fourier transformation is severely restricted. As an alternative, the use of weakly convergent expansions of a plane wave in terms of complete orthonormal sets of functions is suggested. The weakly convergent expansions of a plane wave are constructed in such a way that their application in Fourier integrals leads to expansions of the Fourier or inverse Fourier transform that converge with respect to the norm of either the Hilbert space $L^2(\mathbb{R}^3)$ or the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$. Accordingly, these weakly convergent expansions may be viewed as distributions that are defined on either $L^2(\mathbb{R}^3)$ or $W_2^{(1)}(\mathbb{R}^3)$. The properties of some complete orthonormal sets of functions, in particular their Fourier transforms, are also studied. Shibuya and Wulfman [Proc. R. Soc. London Ser. A **286**, 376 (1965)] derived an expansion of a plane wave involving the four-dimensional spherical harmonics. It is shown that this Shibuya–Wulfman expansion is also a distribution which is defined on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$. Finally, as an application it is shown how weakly convergent expansions can be used profitably for the construction of addition theorems.

I. INTRODUCTION

The Fourier transform is undoubtedly a very powerful mathematical tool. Traditionally, it has been of considerable importance in classical analysis. But its realm could be enlarged greatly since it was possible to show that it makes sense to speak of the Fourier transforms of such nonclassical objects as the delta function or other distributions. In view of this wide applicability it is not surprising that the Fourier transform is also a very helpful device for the solution of numerous physical problems.

The main advantage of the Fourier transform is that its use quite often leads to a considerable formal simplification of the problem under consideration. A good example is the explanation of the accidental degeneracy of the nonrelativistic hydrogen atom by Fock.[1] Using Fourier transformation, Fock converted the Schrödinger equation of the hydrogen atom into an integral equation in momentum space. Now, only a relatively simple variable transformation in the integral equation had to be done and Fock was able to show that the accidental degeneracy of the hydrogen atom can be related to a rotational symmetry in four-dimensional space $\mathbb{R}^4$.

However, in spite of all the undisputed formal advantages and its formal elegance the use of Fourier transforms is quite often severely restricted by annoying technical problems. Unfortunately, it often turns out that the Fourier integrals one has to deal with are extremely complicated and sometimes they are even unmanageable.

In this article we shall treat Fourier transforms of irreducible spherical tensors,

$$F_l^m(\mathbf{r}) = f_l(r) Y_l^m(\mathbf{r}/r) . \tag{1.1}$$

Here, f_l is a radial function and Y_l^m is a spherical harmonic. Most functions that are of interest in atomic and molecular physics can be written in the form of Eq. (1.1). For the computation of the Fourier transform of such an irreducible tensor the well-known Rayleigh expansion of a plane wave in terms of spherical Bessel functions and spherical harmonics is the natural choice,

$$e^{\pm i\mathbf{p}\cdot\mathbf{r}} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} (\pm i)^l j_l(pr) Y_l^{m*}\left(\frac{\mathbf{r}}{r}\right) Y_l^m\left(\frac{\mathbf{p}}{p}\right) . \tag{1.2}$$

In this article we shall use the symmetric version of the Fourier transformation, i.e., a given function $f(\mathbf{r})$ and its Fourier transform $\bar{f}(\mathbf{p})$ are connected by the relationships

$$\bar{f}(\mathbf{p}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} f(\mathbf{r}) d^3\mathbf{r} , \tag{1.3}$$

$$f(\mathbf{r}) = (2\pi)^{-3/2} \int e^{i\mathbf{r}\cdot\mathbf{p}} \bar{f}(\mathbf{p}) d^3\mathbf{p} . \tag{1.4}$$

Because of the orthonormality of the spherical harmonics it follows immediately that the Fourier transform of an irreducible tensor, Eq. (1.1), is again an irreducible spherical tensor,

$$\bar{F}_l^m(\mathbf{p}) = \bar{f}_l(p) Y_l^m(\mathbf{p}/p) , \tag{1.5}$$

$$\bar{f}_l(p) = (-i)^l \left(\frac{2}{\pi}\right)^{1/2} \int_0^\infty r^2 j_l(pr) f_l(r) dr . \tag{1.6}$$

Hence, we see that in the case of irreducible spherical tensors

[a)] On leave of absence from Institut für Physikalische und Theoretische Chemie, Universität Regensburg, D-8400 Regensburg, Federal Republic of Germany. Please direct all correspondence to this address.

 0022-2488/85/020276-16$02.50

Eq. (1.1) the Rayleigh expansion Eq. (1.2) leads to trivial angular integrations. Unfortunately, the remaining radial integral in Eq. (1.6) need not be simple at all and may even prove to be prohibitively complicated. At that stage one has to emphasize that due to the highly oscillatory nature of the spherical Bessel functions a purely numerical evaluation of the radial integral (1.6) may not work either. Although some progress with oscillatory integrals has been reported recently[2-6] a numerical quadrature of the radial integral (1.6) still seems to be a formidable task. Hence, we see that a successful application of the Rayleigh expansion Eq. (1.2) in Fourier integrals depends to a large extent upon one's skill in handling integrals involving spherical Bessel functions.

In this article we want to develop some alternatives to the Rayleigh expansion of a plane wave Eq. (1.2). However, the basic ideas that we shall use here are neither restricted to spherical polar coordinates nor to the three-dimensional space $\mathbf{R}^3$. It is our aim to derive expansions of a plane wave that lead, when used in Fourier integrals, to radial integrals which are more manageable than those radial integrals involving spherical Bessel functions [Eq. (1.6)] which occur if the Rayleigh expansion Eq. (1.2) is used in Fourier integrals.

We shall show later that the Rayleigh expansion Eq. (1.2) is just a rearrangement of the defining power series of the exponential $e^{\pm i\mathbf{p}\cdot\mathbf{r}}$ which converges pointwise. However, in integrals the pointwise convergence of an expansion is not always needed. Therefore, a lot of flexibility and freedom for the construction of expansions can be gained if the condition of pointwise convergence is discarded and if one only requires that the expansions should converge weakly, i.e., in the sense of generalized functions or distributions. Accordingly, all expansions of a plane wave which we shall construct here are distributions that are defined on appropriate Hilbert spaces.

It has to be emphasized that the mathematical formalism was not developed for its own sake. All results which are presented in this article were derived with the intention of facilitating the computation of Fourier transforms of irreducible spherical tensors Eq. (1.1). In addition, the use of the weakly convergent expansions which shall be presented here in Fourier integrals leads to expansion of the Fourier transforms in terms of complete orthonormal sets of functions. This is very convenient if the Fourier transforms are to be used in integrals.

Since complete orthonormal sets of functions are used for the construction of our weakly convergent expansions of a plane wave we have to study some suitable sets of functions, particularly their orthonormality properties with respect to different scalar products and their Fourier transforms.

Some other expansion of a plane wave shall be analyzed also which involves the four-dimensional spherical harmonics. We shall show that this expansion converges also weakly, i.e., it is a distribution, and that it is a biorthogonal expansion which is closely related to some of the expansions which we derived.

Finally, as an application it is shown how the weakly convergent expansions can be used profitably for the construction of addition theorems.

II. DEFINITIONS

For the commonly occurring special functions of mathematical physics we shall use the notations and conventions of Magnus, Oberhettinger, and Soni[7] unless explicitly stated. Hereafter, this reference will be denoted as MOS in the text.

In this article we shall make extensive use of some classical orthogonal polynomials, namely Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ (MOS, pp. 209–217), Gegenbauer polynomials $C_n^{\lambda}(x)$ (MOS, pp. 218–227), and generalized Laguerre polynomials $L_n^{(\alpha)}(x)$ (MOS, pp. 239–249). These polynomials can all be expressed as terminating hypergeometric series (MOS, pp. 212, 220, and 240)

$$P_n^{(\alpha,\beta)}(x) = \binom{n+\alpha}{n}\, {}_2F_1(-n,\alpha+\beta+n+1;\alpha+1;(1-x)/2)\,, \tag{2.1}$$

$$C_n^{\lambda}(x) = ((2\lambda)_n/n!)\, {}_2F_1(-n,n+2\lambda;\lambda+\tfrac{1}{2};\ (1-x)/2)\,, \tag{2.2}$$

$$L_n^{(\alpha)}(x) = \binom{n+\alpha}{n}\, {}_1F_1(-n;\alpha+1;x)\,. \tag{2.3}$$

For the spherical harmonics $Y_l^m(\theta,\phi)$ we use the phase convention of Condon and Shortley,[8] i.e., they are defined by the expression

$$Y_l^m(\theta,\phi) = i^{m+|m|}\left[\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}\right]^{1/2} P_l^{|m|}(\cos\theta)e^{im\phi}\,. \tag{2.4}$$

Here, $P_l^{|m|}(\cos\theta)$ is an associated Legendre polynomial,

$$\begin{aligned} P_l^m(x) &= (1-x^2)^{m/2}\frac{d^{\,l+m}}{dx^{l+m}}\frac{(x^2-1)^l}{2^l l!} \\ &= (1-x^2)^{m/2}\frac{d^{\,m}}{dx^m}P_l(x)\,. \end{aligned} \tag{2.5}$$

For the regular solid harmonic we write

$$\mathscr{Y}_l^m(\mathbf{r}) = r^l Y_l^m(\theta,\phi)\,. \tag{2.6}$$

It is important to note that the regular solid harmonic is a homogeneous polynomial of degree l in the Cartesian components x, y, and z of $\mathbf{r}$ (Ref. 9),

$$\begin{aligned} \mathscr{Y}_l^m(\mathbf{r}) &= \left\{\frac{2l+1}{4\pi}(l+m)!(l-m)!\right\}^{1/2} \\ &\quad\times\sum_k \frac{(-x-iy)^{m+k}(x-iy)^k z^{l-m-2k}}{2^{m+2k}(m+k)!k!(l-m-2k)!}\,. \end{aligned} \tag{2.7}$$

For the integral over the product of three spherical harmonics, the so-called Gaunt coefficient, we write

$$\langle l_3m_3|l_2m_2|l_1m_1\rangle = \int Y_{l_3}^{m_3*}(\Omega)Y_{l_2}^{m_2}(\Omega)Y_{l_1}^{m_1}(\Omega)d\Omega\,. \tag{2.8}$$

The spherical Bessel function $j_n(z)$ is defined in terms of the Bessel function of the first kind (MOS, p. 65),

$$j_n(z) = \left(\frac{\pi}{2z}\right)^{1/2} J_{n+1/2}(z)\,,\quad n\in\mathbf{N}\,. \tag{2.9}$$

If $K_\nu(z)$ stands for the modified Bessel function of the second kind (MOS, p. 66), the reduced Bessel function $\hat{k}_\nu(z)$ is defined by[10]

$$\hat{k}_\nu(z) = (2/\pi)^{1/2} z^\nu K_\nu(z)\,. \tag{2.10}$$

In the case of half-integral orders $\nu = n + \frac{1}{2}$, $n \in \mathbb{N}$, these reduced Bessel functions can be expressed as an exponential multiplied by a terminating confluent hypergeometric series ${}_1F_1$,[11]

$$\hat{k}_{n+1/2}(z) = 2^n (\tfrac{1}{2})_n e^{-z} {}_1F_1(-n; -2n; 2z), \quad n \geqslant 0 . \tag{2.11}$$

The polynomial part of these reduced Bessel functions has been investigated quite extensively in the mathematical literature.[12] There, the notation

$$\Theta_n(z) = e^z \hat{k}_{n+1/2}(z) \tag{2.12}$$

is used. Together with some other, closely related polynomials these $\Theta_n(z)$ are called Bessel polynomials. They find applications in such diverse fields as number theory, statistics, and the analysis of complex electrical networks.[12]

As a nonscalar generalization of the reduced Bessel function, the so-called B function was introduced,[13]

$$B_{n,l}^m(\beta, \mathbf{r}) = [2^{n+l}(n+l)!]^{-1} \hat{k}_{n+1/2}(\beta r) \mathscr{Y}_l^m(\beta \mathbf{r}) . \tag{2.13}$$

In this article, we shall also need some concepts of functional analysis. For that purpose we define the following two scalar products for functions $f, g: \mathbb{R}^3 \to \mathbb{C}$:

$$(f, g) = \int f^*(\mathbf{r}) g(\mathbf{r}) d^3\mathbf{r} , \tag{2.14}$$

$$\langle f, g \rangle_\beta = \int f^*(\mathbf{r}) \frac{\beta^2 - \nabla^2}{2\beta^2} g(\mathbf{r}) d^3\mathbf{r} . \tag{2.15}$$

In Eq. (2.15), ∇ stands for the gradient. From now on we shall tacitly assume that the scaling parameter β is real and positive. The two scalar products (2.14) and (2.15) can be used to define the norms

$$\|f\|_2 = [(f, f)]^{1/2} , \tag{2.16}$$

$$\|f\|_{\beta,2} = [\langle f, f \rangle_\beta]^{1/2} . \tag{2.17}$$

Obviously, the norms (2.17) depend upon the scaling parameter β. However, it is easy to show that they are all equivalent if β is real and positive.

With the help of the norms (2.16) and (2.17) we introduce the Hilbert space of square integrable functions $L^2(\mathbb{R}^3)$ as well as the Sobolev space[14,15] $W_2^{(1)}(\mathbb{R}^3)$

$$L^2(\mathbb{R}^3) = \{ f: \mathbb{R}^3 \to \mathbb{C} | \ \|f\|_2 < \infty \} , \tag{2.18}$$

$$W_2^{(1)}(\mathbb{R}^3) = \{ f: \mathbb{R}^3 \to \mathbb{C} | \ \|f\|_{\beta,2} < \infty \} . \tag{2.19}$$

It is clear that the Sobolev space $W_2^{(1)}(\mathbf{r}^3)$ is a proper subset of the Hilbert space $L^2(\mathbb{R}^3)$.

It is not necessary to use the coordinate representation for the definition of the spaces $L^2(\mathbb{R}^3)$ and $W_2^{(1)}(\mathbb{R}^3)$. Instead, one could equally well have used the momentum representation. This is a consequence of the well-known fact that the Fourier transformation maps $L^2(\mathbb{R}^3)$ onto $L^2(\mathbb{R}^3)$ in a one-to-one manner such that scalar products are conserved.[16] This implies that for $f, g \in L^2(\mathbb{R}^3)$ their Fourier transforms $\bar{f}(\mathbf{p})$ and $\bar{g}(\mathbf{p})$ are also elements of $L^2(\mathbb{R}^3)$. In addition, one obtains for the scalar product (2.14)

$$(f, g) = \int \bar{f}^*(\mathbf{p}) \bar{g}(\mathbf{p}) d^3\mathbf{p} . \tag{2.20}$$

For the scalar product (2.15) which defines the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ a similar relationship can be derived. We use

$$\frac{\beta^2 + p^2}{2\beta^2} \bar{g}(\mathbf{p}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} \frac{\beta^2 - \nabla^2}{2\beta^2} g(\mathbf{r}) d^3\mathbf{r} , \tag{2.21}$$

and obtain for the scalar product (2.15)

$$\langle f, g \rangle_\beta = \int \bar{f}^*(\mathbf{p}) \frac{\beta^2 + p^2}{2\beta^2} \bar{g}(\mathbf{p}) d^3\mathbf{p} . \tag{2.22}$$

Hence, we see that the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ is conceptually much simpler in momentum representation. Instead of Eq. (2.19) one could also have used the definition

$$W_2^{(1)}(\mathbb{R}^3) = \{ f: \mathbb{R}^3 \to \mathbb{C} | [(\beta^2 + p^2)/(2\beta^2)]^{1/2} \bar{f}(\mathbf{p}) \in L^2(\mathbb{R}^3) \} . \tag{2.23}$$

Hence, in momentum representation $W_2^{(1)}(\mathbb{R}^3)$ is simply a L^2 space with the weight function $(\beta^2 + p^2)/(2\beta^2)$. In addition we may conclude that the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ is a Hilbert space with respect to the scalar product $\langle f, g \rangle_\beta$, defined in Eqs. (2.15) or (2.22).

In the theory of distributions the Schwartz space $\mathscr{S}$ of rapidly decreasing functions is of tantamount importance. The test functions $\phi: \mathbb{R}^3 \to \mathbb{C}$ belonging to $\mathscr{S}(\mathbb{R}^3)$ have to satisfy[17]

$$\sup_{\mathbf{r} \in \mathbb{R}^3} \left| r^k \left(\frac{\partial}{\partial x}\right)^l \left(\frac{\partial}{\partial y}\right)^m \left(\frac{\partial}{\partial z}\right)^n \phi(\mathbf{r}) \right| < \infty \tag{2.24}$$

for all integers $k, l, m, n \geqslant 0$.

The dual space of $\mathscr{S}(\mathbb{R}^3)$ which is denoted by $\mathscr{S}'(\mathbb{R}^3)$ is called the space of tempered distributions.[18] Obviously, the following inclusions hold:

$$\mathscr{S}(\mathbb{R}^3) \subset W_2^{(1)}(\mathbb{R}^3) \subset L^2(\mathbb{R}^3) \subset \mathscr{S}'(\mathbb{R}^3) . \tag{2.25}$$

III. WEAKLY CONVERGENT ORTHOGONAL EXPANSIONS

Before we proceed to construct alternative expansions for a plane wave we want to analyze the nature of the Rayleigh expansion Eq. (1.2) more carefully. In particular, we want to find out how it is related to the defining power series of an exponential,

$$e^{\pm i\mathbf{p}\cdot\mathbf{r}} = \sum_{n=0}^{\infty} \frac{(\pm i\mathbf{p}\cdot\mathbf{r})^n}{n!} = \sum_{n=0}^{\infty} \frac{(\pm ipr \cos \omega)^n}{n!} . \tag{3.1}$$

First, we use the well-known relationship[19]

$$\sum_{m=-l}^{l} Y_l^{m*}\left(\frac{\mathbf{r}}{r}\right) Y_l^m\left(\frac{\mathbf{p}}{p}\right) = \frac{2l+1}{4\pi} P_l(\cos \omega) \tag{3.2}$$

to rewrite the Rayleigh expansion Eq. (1.2) as follows:

$$e^{\pm i\mathbf{p}\cdot\mathbf{r}} = \sum_{l=0}^{\infty} (\pm i)^l (2l+1) j_l(pr) P_l(\cos \omega) . \tag{3.3}$$

It is now possible to show that Eq. (3.3) is just a rearrangement of the expansion (3.1). This can be demonstrated by expressing the powers of $\cos \omega$ in Eq. (3.1) in terms of Legendre polynomials $P_l(\cos \omega)$.[20]

The expansion (3.3) may also be viewed as the expansion of a plane wave in terms of the orthogonal polynomials $P_l(\cos \omega)$. However, orthogonal expansions are unique. Therefore, we have to conclude that we cannot achieve our aim—the derivation of alternative expansions which facilitate the analytical evaluation of Fourier integrals if spherical coordinates are used—by looking for other rearrangements of the power series (3.1).

Instead we shall construct expansions that converge weakly, i.e., in the sense of generalized functions. This means that we obtain expansions which converge to $\bar{f}(\mathbf{p})$ or $f(\mathbf{r})$, respectively, if we replace the plane wave in the Fourier integrals (1.3) or (1.4) by the distributions that we are going to construct.

In that context the following two questions have to be answered.

(i) For which class of functions should our distributions be defined.

(ii) In what sense should the resulting expansion converge to $\bar{f}(\mathbf{p})$ or $f(\mathbf{r})$. It is of course clear that these two questions cannot be answered independently.

The two extreme spaces where the Fourier transformation can be defined are the Schwartz space $\mathscr{S}(\mathbb{R}^3)$ and its dual $\mathscr{S}'(\mathbb{R}^3)$, the space of tempered distributions. We could require that our distributions should only be defined on the relatively small space $\mathscr{S}(\mathbb{R}^3)$. Due to the highly idealized nature of the element of $\mathscr{S}(\mathbb{R}^3)$ we would in return gain a lot of freedom in the construction of our distributions. Unfortunately, the space $\mathscr{S}(\mathbb{R}^3)$ is defined in such a restrictive way that most functions which are of interest in atomic and molecular physics do not belong to it. Therefore, it is necessary that our distributions should be defined on a less restrictive space. The largest space at our disposal would be $\mathscr{S}'(\mathbb{R}^3)$, the space of tempered distributions. However, the convergence of a sequence of elements of $\mathscr{S}'(\mathbb{R}^3)$ is generally only defined in the weak sense, i.e., in the sense of distributions, which would be somewhat inconvenient. Therefore, in this article we shall only consider distributions that are defined on the Hilbert space $L^2(\mathbb{R}^3)$ or on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ which is also a Hilbert space. Also, we exploit the topological properties of these Hilbert spaces by demanding that the application of our distributions in Fourier integrals should lead to expansions that converge in the sense of the norm of the corresponding Hilbert space. We shall show later in this section that this can be accomplished quite easily if we represent our distributions in terms of complete orthonormal sets of functions.

For the construction of distributions which are defined on the Hilbert space $L^2(\mathbb{R}^3)$ we shall extensively use the fact that the Fourier transformation does not change the scalar product in $L^2(\mathbb{R}^3)$.[16] Thus, for $f, g \in L^2(\mathbb{R}^3)$ their scalar product can be computed either in the coordinate or in the momentum representation,

$$(f,g) = \int f^*(\mathbf{r}) g(\mathbf{r}) d^3\mathbf{r} = \int \bar{f}^*(\mathbf{p}) \bar{g}(\mathbf{p}) d^3\mathbf{p} . \tag{3.4}$$

Let us now consider some complete orthonormal set $\{\phi_{nl}^m(\mathbf{r})\}$ in $L^2(\mathbb{R}^3)$. From Eq. (3.4) we may conclude that the set of their Fourier transforms $\{\bar{\phi}_{nl}^m(\mathbf{p})\}$ with

$$\bar{\phi}_{nl}^m(\mathbf{p}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} \phi_{nl}^m(\mathbf{r}) d^3\mathbf{r} \tag{3.5}$$

also forms a complete orthonormal set in $L^2(\mathbb{R}^3)$. Orthonormality means

$$\begin{aligned}(\phi_{nl}^m, \phi_{n'l'}^{m'}) &= \int \phi_{nl}^{m*}(\mathbf{r}) \phi_{n'l'}^{m'}(\mathbf{r}) d^3\mathbf{r} \\ &= \int \bar{\phi}_{nl}^{m*}(\mathbf{p}) \bar{\phi}_{n'l'}^{m'}(\mathbf{p}) d^3\mathbf{p} = \delta_{nn'}\delta_{ll'}\delta_{mm'} .\end{aligned} \tag{3.6}$$

Completeness means that every $f \in L^2(\mathbb{R}^3)$ can be expanded in terms of the set $\{\phi_{nl}^m(\mathbf{r})\}$, or equivalently, that its Fourier transform $\bar{f}(\mathbf{p})$ can be expanded in terms of the set $\{\bar{\phi}_{nl}^m(\mathbf{p})\}$,

$$f(\mathbf{r}) = \sum_{nlm} C_{nl}^m \phi_{nl}^m(\mathbf{r}) , \tag{3.7}$$

$$\bar{f}(\mathbf{p}) = \sum_{nlm} C_{nl}^m \bar{\phi}_{nl}^m(\mathbf{p}) , \tag{3.8}$$

$$C_{nl}^m = \int \phi_{nl}^{m*}(\mathbf{r}) f(\mathbf{r}) d^3\mathbf{r} = \int \bar{\phi}_{nl}^{m*}(\mathbf{p}) \bar{f}(\mathbf{p}) d^3\mathbf{p} . \tag{3.9}$$

The expansions (3.7) and (3.8) converge in the sense of the norm of the Hilbert space $L^2(\mathbb{R}^3)$ [Eq. (2.16)],

$$\begin{aligned}&\Big|\Big| f(\mathbf{r}) - \sum_{nlm} C_{nl}^m \phi_{nl}^m(\mathbf{r}) \Big|\Big|_2 \\ &= \Big|\Big| \bar{f}(\mathbf{p}) - \sum_{nlm} C_{nl}^m \bar{\phi}_{nl}^m(\mathbf{p}) \Big|\Big|_2 = 0 .\end{aligned} \tag{3.10}$$

Hence, we see that the expansion coefficients C_{nl}^m do not only determine a given function $f(\mathbf{r})$ but also its Fourier transform $\bar{f}(\mathbf{p})$. The only requirement is that one has to know the set of Fourier transforms $\{\bar{\phi}_{nl}^m(\mathbf{p})\}$.

We are now in the position to prove the following.

Theorem: The functions $\{\phi_{nl}^m(\mathbf{r})\}$ are a complete orthonormal set in $L^2(\mathbb{R}^3)$ and the functions $\{\bar{\phi}_{nl}^m(\mathbf{p})\}$ are their Fourier transforms according to Eq. (3.5). Then the equality

$$e^{i\mathbf{p}\cdot\mathbf{r}} = (2\pi)^{3/2} \sum_{nlm} \bar{\phi}_{nl}^{m*}(\mathbf{p}) \phi_{nl}^m(\mathbf{r}) \tag{3.11}$$

is valid as a distribution for all functions $f \in L^2(\mathbb{R}^3)$.

In order to prove this theorem we use Eq. (3.11) in the Fourier integral (1.3) and integrate termwise

$$\bar{f}(\mathbf{p}) = \sum_{nlm} \bar{\phi}_{nl}^m(\mathbf{p}) \int \phi_{nl}^{m*}(\mathbf{r}) f(\mathbf{r}) d^3\mathbf{r} \tag{3.12}$$

$$= \sum_{nlm} C_{nl}^m \bar{\phi}_{nl}^m(\mathbf{p}) . \tag{3.13}$$

However, Eq. (3.13) is identical with Eq. (3.8). We now use Eq. (3.11) in the Fourier integral (1.4) and integrate termwise again

$$f(\mathbf{r}) = \sum_{nlm} \phi_{nl}^m(\mathbf{r}) \int \bar{\phi}_{nl}^{m*}(\mathbf{p}) \bar{f}(\mathbf{p}) d^3\mathbf{p} \tag{3.14}$$

$$= \sum_{nlm} C_{nl}^m \phi_{nl}(\mathbf{r}) . \tag{3.15}$$

However, Eq. (3.15) is identical with Eq. (3.7).

For the construction of distributions which are defined on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ we can proceed in exactly the same way as in the case of the Hilbert space $L^2(\mathbb{R}^3)$. The only difference is that we now exploit the invariance of the scalar product of $W_2^{(1)}(\mathbb{R}^3)$,

$$\begin{aligned}\langle f, g\rangle_\beta &= \int f^*(\mathbf{r}) \frac{\beta^2 - \nabla^2}{2\beta^2} g(\mathbf{r}) d^3\mathbf{r} \\ &= \int \bar{f}^*(\mathbf{p}) \frac{\beta^2 + p^2}{2\beta^2} \bar{g}(\mathbf{p}) d^3\mathbf{p} .\end{aligned} \tag{3.16}$$

Let us now consider a complete orthonormal set $\{\psi_{nl}^m(\mathbf{r})\}$ in $W_2^{(1)}(\mathbb{R}^3)$. From Eq. (3.16) we may conclude that the set of their Fourier transforms $\{\bar{\psi}_{nl}^m(\mathbf{p})\}$ with

$$\bar{\psi}_{nl}^m(\mathbf{p}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} \psi_{nl}^m(\mathbf{r}) d^3\mathbf{r} \tag{3.17}$$

also forms a complete orthonormal set in $W_2^{(1)}$. Orthonormality means here

$$\begin{aligned} \langle \psi_{nl}^m, \psi_{n'l'}^{m'} \rangle_\beta &= \int \psi_{nl}^{m*}(\mathbf{r}) \frac{\beta^2 - \nabla^2}{2\beta^2} \psi_{n'l'}^{m'}(\mathbf{r}) d^3\mathbf{r} \\ &= \int \bar{\psi}_{nl}^{m*}(\mathbf{p}) \frac{\beta^2 + p^2}{2\beta^2} \bar{\psi}_{n'l'}^{m'}(\mathbf{p}) d^3\mathbf{p} \\ &= \delta_{nn'} \delta_{ll'} \delta_{mm'} . \end{aligned} \tag{3.18}$$

Completeness means that every $f \in W_2^{(1)}(\mathbb{R}^3)$ can be expanded in terms of the set $\{\psi_{nl}^m(\mathbf{r})\}$, or equivalently, that its Fourier transform $\bar{f}(\mathbf{p})$ can be expanded in terms of the set $\{\bar{\psi}_{nl}^m(\mathbf{p})\}$,

$$f(\mathbf{r}) = \sum_{nlm} \gamma_{nl}^m \psi_{nl}^m(\mathbf{r}) , \tag{3.19}$$

$$\bar{f}(\mathbf{p}) = \sum_{nlm} \gamma_{nl}^m \bar{\psi}_{nl}^m(\mathbf{p}) , \tag{3.20}$$

$$\begin{aligned} \gamma_{nl}^m &= \int \psi_{nl}^{m*}(\mathbf{r}) \frac{\beta^2 - \nabla^2}{2\beta^2} f(\mathbf{r}) d^3\mathbf{r} \\ &= \int \bar{\psi}_{nl}^{m*}(\mathbf{p}) \frac{\beta^2 + p^2}{2\beta^2} \bar{f}(\mathbf{p}) d^3\mathbf{p} . \end{aligned} \tag{3.21}$$

The expansions (3.19) and (3.20) converge in the sense of the norm of the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ [Eq. (2.17)]

$$\begin{aligned} &\left|\left| f(\mathbf{r}) - \sum_{nlm} \gamma_{nl}^m \psi_{nl}^m(\mathbf{r}) \right|\right|_{\beta,2} \\ &= \left|\left| \bar{f}(\mathbf{p}) - \sum_{nlm} \gamma_{nl}^m \bar{\psi}_{nl}^m(\mathbf{p}) \right|\right|_{\beta,2} = 0 . \end{aligned} \tag{3.22}$$

Just as in the case of the Hilbert space $L^2(\mathbb{R}^3)$ the expansion coefficients γ_{nl}^m determine not only a given function $f(\mathbf{r})$ but also its Fourier transform $\bar{f}(\mathbf{p})$. Again the only requirement is that one has to know the set of Fourier transforms $\{\bar{\psi}_{nl}^m(\mathbf{p})\}$.

We are now in the position to prove the following.

Theorem: The functions $\{\psi_{nl}^m(\mathbf{r})\}$ are a complete orthonormal set in $W_2^{(1)}(\mathbb{R}^3)$ and the functions $\{\bar{\psi}_{nl}^m(\mathbf{p})\}$ are their Fourier transforms according to Eq. (3.17). Then the equality

$$e^{i\mathbf{p}\cdot\mathbf{r}} = (2\pi)^{3/2} \sum_{nlm} \bar{\psi}_{nl}^{m*}(\mathbf{p}) \psi_{nl}^m(\mathbf{r}) \tag{3.23}$$

is valid as a distribution for all functions $f \in W_2^{(1)}(\mathbb{R}^3)$. Here, the scalar products are to be computed according to Eq. (3.16).

The proof of Eq. (3.23) can be done in the same way as the proof of Eq. (3.11). The only difference is that in the case of an integration over $\mathbf{r}$ the weight function $(\beta^2 - \nabla^2)/(2\beta^2)$ has to be included, whereas for an integration over $\mathbf{p}$ the weight function $(\beta^2 + p^2)/(2\beta^2)$ is needed. Accordingly, we use Eq. (3.23) in the Fourier integral (1.3) and integrate termwise

$$\bar{f}(\mathbf{p}) = \sum_{nlm} \bar{\psi}_{nl}^m(\mathbf{p}) \int \psi_{nl}^{m*}(\mathbf{r}) \frac{\beta^2 - \nabla^2}{2\beta^2} f(\mathbf{r}) d^3\mathbf{r} \tag{3.24}$$

$$= \sum_{nlm} \gamma_{nl}^m \bar{\psi}_{nl}^m(\mathbf{p}) . \tag{3.25}$$

However, Eq. (3.25) is identical with Eq. (3.20). We now use Eq. (3.23) in the Fourier integral (1.4) and integrate termwise

$$f(\mathbf{r}) = \sum_{nlm} \psi_{nl}^m(\mathbf{r}) \int \bar{\psi}_{nl}^{m*}(\mathbf{p}) \frac{\beta^2 + p^2}{2\beta^2} \bar{f}(\mathbf{p}) d^3\mathbf{p} \tag{3.26}$$

$$= \sum_{nlm} \gamma_{nl}^m \psi_{nl}^m(\mathbf{r}) . \tag{3.27}$$

However, Eq. (3.27) is identical with Eq. (3.19).

Formally, the distributions (3.11) and (3.23) look like the expansion of an element of some Hilbert space in terms of a complete orthonormal set. However, the plane wave $e^{i\mathbf{p}\cdot\mathbf{r}}$ is neither an element of $L^2(\mathbb{R}^3)$ nor an element of $W_2^{(1)}(\mathbb{R}^3)$. Therefore, it is not legitimate to conclude that the expansions (3.11) and (3.23) converge in the sense of the norms of L^2 or $W_2^{(1)}$, Eqs. (2.16) and (2.17), let alone pointwise.

What is the advantage if we use instead of the Rayleigh expansion Eq. (1.2) either one of the distributions (3.11) or (3.23) in Fourier transforms. With the help of these distributions the computation of a Fourier transform is reduced to the determination of the scalar products C_{nl}^m or γ_{nl}^m, respectively. In many cases the determination of expansion coefficients is much easier than the computation of Fourier integrals. In addition, the complete orthonormal systems which are used in the distributions (3.11) or (3.23) are so far completely unspecified and only subject to the restriction that one must know their Fourier transforms explicitly. Therefore, one can try to find some complete orthonormal system which has optimal properties for the problem under consideration.

If one wants to compute numerical values of the Fourier transform of a given function the use of the distributions (3.11) or (3.23) may not be possible. This is a consequence of the well-known fact that an orthogonal expansion does not necessarily converge pointwise. However, if one wants to use a Fourier transform in integrals the distributions (3.11) and (3.23) should have some distinct advantages in comparison with the Rayleigh expansion Eq. (1.2).

IV. EXPONENTIALLY DECLINING FUNCTION SETS AND THEIR FOURIER TRANSFORMS

In this section we shall analyze the properties of some complete orthonormal sets in $L^2(\mathbb{R}^3)$ and $W_2^{(1)}(\mathbb{R}^3)$. We shall only consider functions that can be written in the form

$$F_{nl}^m(\mathbf{r}) = R_{nl}(r) \mathcal{Y}_l^m(\mathbf{r}) . \tag{4.1}$$

Here, $R_{nl}(r)$ is a function that only depends upon the distance r and $\mathcal{Y}_l^m(\mathbf{r})$ is a regular solid harmonic defined in Eq. (2.6). The reason for this restriction is that almost all functions that are of interest in atomic and molecular physics are of the form of Eq. (4.1).

It is well known that the exact solutions of atomic and molecular Schrödinger equations decline exponentially for large distances.[21] Accordingly, in this section we only consider functions $F_{nl}^m(\mathbf{r})$, where the radial part $R_{nl}(r)$ can be written as an exponential multiplied by some polynomial. Due to their definition [Eq. (4.1)], the functions F_{nl}^m are automatically orthogonal with respect to an integration over the surface of the three-dimensional unit sphere and only their radial parts R_{nl} have to be orthogonalized. This can be done quite easily by exploiting the orthogonality relationship satisfied by the generalized Laguerre polynomials (MOS, p. 241)

$$\int_0^\infty e^{-x}x^\alpha L_m^{(\alpha)}(x)L_n^{(\alpha)}(x)dx = \frac{\Gamma(\alpha+n+1)}{n!}\delta_{mn} . \quad (4.2)$$

Hence, we shall only consider functions of the form

$$e^{-\beta r}L_{n-l-1}^{(\alpha)}(2\beta r)\mathscr{Y}_l^m(2\beta\mathbf{r}) . \quad (4.3)$$

We shall see that the parameter α determines whether these functions are orthonormal in the Hilbert space $L^2(\mathbf{R}^3)$ or in the Sobolev space $W_2^{(1)}(\mathbf{R}^3)$.

The following set of functions is complete[22] and orthonormal in $L^2(\mathbf{R}^3)$,

$$\Lambda_{n,l}^m(\beta,\mathbf{r}) = N_{n,l}(\beta)e^{-\beta r}L_{n-l-1}^{(2l+2)}(2\beta r)\mathscr{Y}_l^m(2\beta\mathbf{r}) , \quad (4.4a)$$

$$N_{n,l}(\beta) = (2\beta)^{3/2}[(n-l-1)!/(n+l+1)!]^{1/2} . \quad (4.4b)$$

These Λ functions satisfy the orthogonality relationship

$$\int \Lambda_{n,l}^{m*}(\beta,\mathbf{r})\Lambda_{n',l'}^{m'}(\beta,\mathbf{r})d^3\mathbf{r} = \delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (4.5)$$

The Λ functions were introduced into atomic and molecular calculations by Hylleraas[23] and by Shull and Löwdin.[24] Later they were used by Filter and Steinborn[25] for the derivation of addition theorems.

Closely related to the Λ functions is the following set of functions which were already in 1928 used by Hylleraas[26] and which are commonly called Coulomb Sturmians or simply Sturmians[27]:

$$\Psi_{n,l}^m(\beta,\mathbf{r}) = N_{n,l}(\beta)e^{-\beta r}L_{n-l-1}^{(2l+1)}(2\beta r)\mathscr{Y}_l^m(2\beta\mathbf{r}) , \quad (4.6a)$$

$$N_{n,l}(\beta) = (2\beta)^{3/2}\{(n-l-1)!/2n(n+l)!\}^{1/2} . \quad (4.6b)$$

From the orthogonality relation of the generalized Laguerre polynomials [Eq. (4.2)] we obtain immediately that the Sturmians satisfy the orthogonality relationship

$$\int \Psi_{n,l}^{m*}(\beta,\mathbf{r})\frac{1}{r}\Psi_{n',l'}^{m'}(\beta,\mathbf{r})d^3\mathbf{r} = \frac{\beta}{n}\delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (4.7)$$

This orthogonality relationship implies that the Sturmians are an orthogonal set in the Hilbert space $L_{1/r}^2(\mathbf{R}^3)$ which is defined by the scalar product

$$(f,g)_{1/r} = \int f^*(\mathbf{r})\frac{1}{r}g(\mathbf{r})d^3\mathbf{r} . \quad (4.8)$$

At that stage it must be emphasized that the Hilbert space $L_{1/r}^2(\mathbf{R}^3)$ is not suited for quantum mechanical applications since neither $L_{1/r}^2(\mathbf{R}^3)\subset L^2(\mathbf{R}^3)$ nor $L^2(\mathbf{R}^3)\subset L_{1/r}^2(\mathbf{R}^3)$ holds. Therefore, we cannot deduce from Eq. (4.7) alone that the Sturmians are of any use in atomic and molecular physics. However, if we combine the differential equation satisfied by the Sturmians,

$$[\nabla^2 + 2\beta n/r - \beta^2]\Psi_{n,l}^m(\beta,\mathbf{r}) = 0 , \quad (4.9)$$

with the orthogonality relationship (4.7), we find that the Sturmians are an orthonormal set in $W_2^{(1)}(\mathbf{R}^3)$,

$$\int \Psi_{n,l}^{m*}(\beta,\mathbf{r})\frac{\beta^2-\nabla^2}{2\beta^2}\Psi_{n',l'}^{m'}(\beta,\mathbf{r})d^3\mathbf{r} = \delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (4.10)$$

The completeness of the Sturmians in $W_2^{(1)}(\mathbf{R}^3)$ can also be proved.[22]

If we replace in the differential equation (4.9) the scaling parameter β by Z/n, we obtain the Schrödinger equation of a hydrogenlike ion with nuclear charge Z. Hence, the Sturmians must be closely related to the hydrogen eigenfunctions describing bound states with negative energies. These eigenfunctions are given by[28]

$$W_{n,l}^m(Z,\mathbf{r}) = N_{n,l}(Z)e^{-Zr/n}L_{n-l-1}^{(2l+1)}(2Zr/n)\mathscr{Y}_l^m(2Z\mathbf{r}/n) , \quad (4.11a)$$

$$N_{n,l}(Z) = (2Z/n)^{3/2}\{(n-l-1)!/2n(n+l)!\}^{1/2} . \quad (4.11b)$$

Comparison of Eqs. (4.6) and (4.11) shows that the Sturmians and the hydrogen eigenfunctions can be transformed into each other by exchanging β and Z/n,

$$\Psi_{n,l}^m(Z/n,\mathbf{r}) = W_{n,l}^m(Z,\mathbf{r}) . \quad (4.12)$$

It is in fact somewhat surprising that the normalization constants in Eqs. (4.6) and (4.11) are identical since the Sturmians are according to Eq. (4.10) orthonormal in $W_2^{(1)}(\mathbf{R}^3)$, whereas the hydrogen eigenfunctions are orthonormal in $L^2(\mathbf{R}^3)$ satisfying

$$\int W_{n,l}^{m*}(Z,\mathbf{r})W_{n',l'}^{m'}(Z,\mathbf{r})d^3\mathbf{r} = \delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (4.13)$$

There is another very important difference between the Sturmians and the hydrogen eigenfunctions. The Sturmians are complete in the Sobolev space $W_2^{(1)}(\mathbf{R}^3)$, whereas the hydrogen eigenfunctions are only complete in the Hilbert space $L^2(\mathbf{R}^3)$ if the eigenfunctions belonging to the continuous spectrum are included. Unfortunately, the complicated mathematical nature of the continuum eigenfunctions[29] effectively prevents their use in most applications. Therefore, one has to conclude that the completeness of the hydrogen eigenfunctions in $L^2(\mathbf{R}^3)$ is more of a formal nature and that one should try to avoid the use of these functions in expansions.

These inconvenient completeness properties of the hydrogen eigenfunctions have some unpleasant consequences in perturbation theory. If the unperturbed system is a hydrogen atom, perturbation theory involves not only a summation over discrete bound states but also an integration over continuum states. The last step may become extremely difficult. Fortunately, with the help of the Lie algebras so(2,1), so(4), and so(4,2) a nonunitary transformation can be constructed which reformulates the Hamiltonian in such a way that the solutions of the unperturbed system are Sturmians[30,31] which are complete and orthonormal in $W_2^{(1)}(\mathbf{R}^3)$. Since completeness in $W_2^{(1)}(\mathbf{R}^3)$ implies completeness in $L^2(\mathbf{R}^3)$ (Ref. 22) Sturmians can be used safely even in those expansions which occur in large order perturbation theory.[30]

According to Eqs. (4.7) and (4.10) Sturmians are either orthogonal with respect to the weight function $1/r$ or with respect to the differential operator $(\beta^2-\nabla^2)/(2\beta^2)$. We can use that fact to introduce a class of functions which are biorthogonal to the Sturmians. We define

$$\Phi_{n,l}^m(\beta,\mathbf{r}) = (n/\beta r)\Psi_{n,l}^m(\beta,\mathbf{r})$$

$$= (2\beta)^{3/2}\left\{\frac{n(n-l-1)!}{2(n+l)!}\right\}^{1/2}\frac{e^{-\beta r}}{\beta r}$$

$$\times L_{n-l-1}^{(2l+1)}(2\beta r)\mathscr{Y}_l^m(2\beta\mathbf{r}) . \quad (4.14)$$

Comparison with Eq. (4.7) yields the biorthogonality relation

$$\int \Psi_{n,l}^{m*}(\beta,\mathbf{r})\Phi_{n',l'}^{m'}(\beta,\mathbf{r})d^3\mathbf{r} = \delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (4.15)$$

Let us now consider the Fourier transforms of the functions which are discussed in this section. The Fourier transform of a Sturmian has been computed already in 1929 by Podolsky and Pauling.[32] In their derivation Podolsky and Pauling could not compute the Fourier transform of a Sturmian directly because the straightforward application of the Rayleigh expansion Eq. (1.2) lead to a radial integral which was not known. Instead, they computed the Fourier transform of

$$\frac{e^{-[(1+t)/(1-t)]\beta r}}{(1-t)^{2l+2}} \mathscr{Y}_l^m(2\beta\mathbf{r}) = \sum_{n=0}^{\infty} e^{-\beta r} L_n^{(2l+1)}(2\beta r)\mathscr{Y}_l^m(2\beta\mathbf{r})t^n , \quad (4.16)$$

which can be derived with the help of a generating function of the generalized Laguerre polynomials (MOS, p. 242),

$$\sum_{n=0}^{\infty} L_n^{(\alpha)}(x)t^n = (1-t)^{-\alpha-1} e^{tx/(t-1)}, \quad |t|<1 . \quad (4.17)$$

After integration, Podolsky and Pauling had to do a power series expansion in t in order to obtain the Fourier transform of a Sturmian, or in view of Eq. (4.12), the Fourier transform of a hydrogen eigenfunction.

Kaijser and Smith[33] showed that the generating function technique introduced by Podolsky and Pauling[32] can also be used for the computation of the Fourier transforms of Λ functions. However, all derivations which are based upon the generating function (4.17) are relatively complicated. Therefore, we want to present here a new method which allows a simple and unified computation of the Fourier transforms of all functions which are treated in this section.

Our new derivation is based upon the fact that generalized Laguerre polynomials can be expressed as finite sums of reduced Bessel functions with half-integral orders[34]

$$e^{-x} L_n^{(\alpha)}(2x) = (2n+\alpha+1) \times \sum_{t=0}^{n} \frac{(-2)^t \Gamma(n+\alpha+t+1)}{t!(n-t)!\Gamma(\alpha+2t+2)} \hat{k}_{t+1/2}(x) . \quad (4.18)$$

Consequently, it is possible to express Sturmians as well as Λ functions in terms of B functions which are defined in Eq. (2.13),

$$\Psi_{n,l}^m(\beta,\mathbf{r}) = (2\beta)^{3/2} \frac{2^{l+1}}{(2l+1)!!} \left\{ \frac{n(n+l)!}{2(n-l-1)!} \right\}^{1/2} \times \sum_{t=0}^{n-l-1} \frac{(-n+l+1)_t (n+l+1)_t}{t!(l+3/2)_t} \times B_{t+1,l}^m(\beta,\mathbf{r}) , \quad (4.19)$$

$$\Lambda_{n,l}^m(\beta,\mathbf{r}) = (2\beta)^{3/2} 2^l \frac{(2n+1)}{(2l+3)!!} \left\{ \frac{(n+l+1)!}{(n-l-1)!} \right\}^{1/2} \times \sum_{t=0}^{n-l-1} \frac{(-n+l+1)_t (n+l+2)_t}{t!(l+5/2)_t} \times B_{t+1,l}^m(\beta,\mathbf{r}) . \quad (4.20)$$

What is gained by expressing Sturmians and Λ functions in terms of B functions. From their definition [Eq. (2.13)], in connection with Eq. (2.11), we see that B functions are not particularly simple in coordinate representation. However, it could be shown that the Fourier transform of a B function is of exceptional simplicity,[35]

$$\bar{B}_{n,l}^m(\beta,\mathbf{p}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} B_{n,l}^m(\beta,\mathbf{r}) d^3\mathbf{r} = \left(\frac{2}{\pi}\right)^{1/2} \frac{\beta^{2n+l-1}}{[\beta^2+p^2]^{n+l+1}} \mathscr{Y}_l^m(-i\mathbf{p}) . \quad (4.21)$$

It is a direct consequence of this extremely compact Fourier transform that B functions have such advantageous properties in multicenter problems.[11,13,35,36] It can be seen from their definition [Eq. (2.13)] that B functions are classically defined only if the inequality $n+l>0$ holds. However, with the help of the Fourier transform (4.21) it could be shown that B functions with $n+l$ being a negative integer are derivatives of the delta function.[37]

It is now a simple matter to derive analytical expressions for the Fourier transforms of Sturmians and Λ functions. If we insert Eq. (4.21) into Eq. (4.19) we obtain

$$\bar{\Psi}_{n,l}^m(\beta,\mathbf{r}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} \Psi_{n,l}^m(\beta,\mathbf{r}) d^3\mathbf{r} = \left\{ \frac{2\beta n(n+l)!}{\pi(n-l-1)!} \right\}^{1/2} [(2l+1)!!]^{-1} \left\{ \frac{2\beta}{\beta^2+p^2} \right\}^{l+2} \times {}_2F_1\left(-n+l+1, n+l+1; l+\frac{3}{2}; \frac{\beta^2}{\beta^2+p^2}\right) \times \mathscr{Y}_l^m(-i\mathbf{p}) . \quad (4.22)$$

We now use Eq. (2.2) to express the terminating hypergeometric series ${}_2F_1$ as a Gegenbauer polynomial,

$${}_2F_1\left(-n+l+1, n+l+1; l+\frac{3}{2}; \frac{\beta^2}{\beta^2+p^2}\right) = \frac{(n-l-1)!(2l+1)!}{(n+l)!} C_{n-l-1}^{l+1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) . \quad (4.23)$$

Thus we finally obtain for the Fourier transform of a Sturmian

$$\bar{\Psi}_{n,l}^m(\beta,\mathbf{p}) = 2^l l! \left\{ \frac{2\beta n(n-l-1)!}{\pi(n+l)!} \right\}^{1/2} \left\{ \frac{2\beta}{\beta^2+p^2} \right\}^{l+2} \times C_{n-l-1}^{l+1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) \mathscr{Y}_l^m(-i\mathbf{p}) . \quad (4.24)$$

In the same way we can derive an expression for the Fourier transform of a Λ function

$$\bar{\Lambda}_{n,l}^m(\beta,\mathbf{p}) = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} \Lambda_{n,l}^m(\beta,\mathbf{r}) d^3\mathbf{r} = \frac{2}{(1/2)_n} \left\{ \frac{\beta(n+l+1)!(n-l-1)!}{\pi} \right\}^{1/2} \times \left\{ \frac{\beta}{\beta^2+p^2} \right\}^{l+2} \times P_{n-l-1}^{(l+3/2, l+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) \mathscr{Y}_l^m(-i\mathbf{p}) . \quad (4.25)$$

The only difference in the derivation of this relationship and of Eq. (4.24) is that here we have to use Eq. (2.1) to express a terminating hypergeometric series ${}_2F_1$ as a Jacobi polynomial.

For the computation of the Fourier transforms of the biorthogonal functions $\Phi^m_{n,l}(\beta,\mathbf{r})$ which were defined in Eq. (4.14) we use[38]

$$\frac{e^{-x}}{x} L_n^{(\alpha)}(2x) = \sum_{t=0}^{n} \frac{(-2)^t \Gamma(n+\alpha+t+1)}{t!(n-t)!\Gamma(\alpha+2t+1)} \hat{k}_{t-1/2}(x) \tag{4.26}$$

in order to express these functions in terms of B functions,

$$\Phi^m_{n,l}(\beta,\mathbf{r}) = (2\beta)^{3/2}\left\{\frac{n(n+l)!}{2(n-l-1)!}\right\}^{1/2} \frac{2^l}{(2l+1)!!} \times \sum_{t=0}^{n-l-1} \frac{(-n+l+1)_t(n+l+1)_t}{t!(l+3/2)_t} B^m_{t,l}(\beta,\mathbf{r}). \tag{4.27}$$

If we use the expression for the Fourier transform of a B function [Eq. (4.21)], we immediately obtain

$$\begin{aligned}\overline{\Phi}^m_{n,l}(\beta,\mathbf{r}) &= (2\pi)^{-3/2}\int e^{-i\mathbf{p}\cdot\mathbf{r}}\Phi^m_{n,l}(\beta,\mathbf{r})d^3\mathbf{r} \\ &= 2^l l!\left\{\frac{2n(n-l-1)!}{\pi\beta(n+l)!}\right\}^{1/2}\left\{\frac{2\beta}{\beta^2+p^2}\right\}^{l+1} \\ &\quad\times C^{l+1}_{n-l-1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right)\mathscr{Y}^m_l(-i\mathbf{p}).\end{aligned} \tag{4.28}$$

Comparison of Eqs. (4.24) and (4.28) yields

$$\overline{\Phi}^m_{n,l}(\beta,\mathbf{p}) = \frac{\beta^2+p^2}{2\beta^2}\overline{\Psi}^m_{n,l}(\beta,\mathbf{p}), \tag{4.29}$$

which is in agreement with the differential equation (4.9).

The Fourier transform (4.28) was in principle already derived by Rotenberg[39] in disguised form. Rotenberg had defined the Sturmians in such a way that the radial part of the infinitesimal volume element in spherical coordinates was absorbed in the functions, i.e., he dealt with functions that are proportional to $r\,\Psi^m_{n,l}(\beta,\mathbf{r})$. However, this definition makes it very hard to define the Fourier transform and the inverse Fourier transform consistently. Therefore, what Rotenberg called the Fourier transform of a function which is proportional to $r\,\Psi^m_{n,l}(\beta,\mathbf{r})$, is in the commonly used notation proportional to the Fourier transform of the function $\Phi^m_{n,l}(\beta,\mathbf{r}) = (n/\beta r)\Psi^m_{n,l}(\beta,\mathbf{r})$.

If we now use Eq. (4.12) we immediately obtain an expression for the Fourier transform of a hydrogen eigenfunction $W^m_{n,l}(Z,\mathbf{r})$ which is defined in Eq. (4.11),

$$\begin{aligned}\overline{W}^m_{n,l}(Z,\mathbf{p}) &= (2\pi)^{-3/2}\int e^{-i\mathbf{p}\cdot\mathbf{r}}W^m_{n,l}(Z,\mathbf{r})d^3\mathbf{r} \\ &= 2^l l!\left\{\frac{2Z(n-l-1)!}{\pi(n+l)!}\right\}^{1/2}\left\{\frac{2Zn}{n^2p^2+Z^2}\right\}^{l+2} \\ &\quad\times C^{l+1}_{n-l-1}\left(\frac{n^2p^2-Z^2}{n^2p^2+Z^2}\right)\mathscr{Y}^m_l(-i\mathbf{p}).\end{aligned} \tag{4.30}$$

If we compare Eq. (4.30) with formulas published by other authors we find some discrepancies. In the formula given by Podolsky and Pauling[32] a phase factor $(-i)^l$ is missing. The same error was reproduced by Bethe and Salpeter.[40] Also, the formula for the Fourier transform of a hydrogen eigenfunction given by Englefield[41] differs from Eq. (4.30) by a phase factor $(-1)^m$. The occurrence of this factor is due to different phase conventions for the hydrogen eigenfunctions Eq. (4.11). Englefield uses the phase convention of Condon and Shortley[8] for the spherical harmonics $Y^m_l(\theta,\phi)$. Therefore, he explicitly corrected the formula for the hydrogen eigenfunctions given by Bethe and Salpeter which use a different phase convention for the spherical harmonics. However, this phase factor $(-1)^m$ is relatively inconvenient and since it is not really necessary for our purposes it was simply suppressed in the definition of the hydrogen eigenfunctions Eq. (4.11). Finally, in the expression given by Biedenharn and Louck[42] for the Fourier transform of a hydrogen eigenfunction a factor $\pi^{-1/2}$ is missing.

We now want to study the orthogonality properties of the Fourier transforms (4.24) and (4.25). The Jacobi polynomials Eq. (2.1) satisfy (MOS, p. 212)

$$\int_{-1}^{1} P^{(\alpha,\beta)}_m(x)P^{(\alpha,\beta)}_n(x)(1-x)^\alpha(1+x)^\beta dx = \frac{\Gamma(\alpha+n+1)\Gamma(\beta+n+1)2^{\alpha+\beta+1}}{n!\Gamma(\alpha+\beta+n+1)(\alpha+\beta+2n+1)}\delta_{mn}. \tag{4.31}$$

With the help of the substitution $x = (p^2-\beta^2)/(p^2+\beta^2)$ we obtain after some algebra

$$\begin{aligned}&\int_0^\infty P^{(l+3/2,l+1/2)}_{n-l-1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) \times P^{(l+3/2,l+1/2)}_{n'-l-1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right)\frac{p^{2l+2}dp}{[p^2+\beta^2]^{2l+4}} \\ &= \frac{\pi}{4\beta^{2l+5}}\frac{[(1/2)_n]^2}{(n-l-1)!(n+l+1)!}\delta_{nn'}.\end{aligned} \tag{4.32}$$

Hence, we see that the Fourier transforms of Λ functions [Eq. (4.25)] are indeed orthonormal in $L^2(\mathbb{R}^3)$,

$$\int \overline{\Lambda}^{m*}_{n,l}(\beta,\mathbf{p})\overline{\Lambda}^{m'}_{n',l'}(\beta,\mathbf{p})d^3\mathbf{p} = \delta_{nn'}\delta_{ll'}\delta_{mm'}. \tag{4.33}$$

The Gegenbauer polynomials Eq. (2.2) satisfy the orthogonality relationship (MOS, p. 221)

$$\int_{-1}^{1} C^\lambda_m(x)C^\lambda_n(x)(1-x^2)^{\lambda-1/2}dx = \frac{\pi 2^{1-2\lambda}\Gamma(n+2\lambda)}{n!(n+\lambda)[\Gamma(\lambda)]^2}\delta_{mn}, \tag{4.34}$$

from which we obtain

$$\begin{aligned}&\int_0^\infty C^{l+1}_{n-l-1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) \times C^{l+1}_{n'-l-1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right)\frac{p^{2l+2}dp}{[p^2+\beta^2]^{2l+3}} \\ &= \frac{\pi(n+l)!2^{-4l-4}}{(n-l-1)!n[l!]^2\beta^{2l+3}}\delta_{nn'}.\end{aligned} \tag{4.35}$$

Hence, we see that the Fourier transforms of Sturmians Eq. (4.24) satisfy

$$\int \overline{\Psi}^{m*}_{n,l}(\beta,\mathbf{p})\frac{\beta^2+p^2}{2\beta^2}\overline{\Psi}^{m'}_{n',l'}(\beta,\mathbf{p})d^3\mathbf{p} = \delta_{nn'}\delta_{ll'}\delta_{mm'}, \tag{4.36}$$

which is obviously an orthonormality relationship in $W^{(1)}_2(\mathbb{R}^3)$. With the help of Eq. (4.29) the orthonormality relationship (4.36) can also be reformulated as the momentum representation of the biorthogonality relationship (4.15),

$$\int \bar{\Psi}_{n,l}^{m*}(\beta,\mathbf{p})\bar{\Phi}_{n',l'}^{m'}(\beta,\mathbf{p})d^3\mathbf{p} = \delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (4.37)$$

If we use the Λ functions Eq. (4.4) and their Fourier transforms Eq. (4.25) in Eq. (3.11), we can formulate the following weakly convergent expansion of a plane wave which is defined on the Hilbert space $L^2(\mathbb{R}^3)$:

$$\begin{aligned} e^{i\mathbf{p}\cdot\mathbf{r}} &= (2\pi)^{3/2} \sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} \bar{\Lambda}_{n,l}^{m*}(\beta,\mathbf{p})\Lambda_{n,l}^{m}(\beta,\mathbf{r}) \\ &= 4\pi \sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} i^l \frac{(n-l-1)!}{(1/2)_n}\left\{\frac{2\beta^2}{\beta^2+p^2}\right\}^{l+2} \\ &\quad \times P_{n-l-1}^{(l+3/2,l+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right)\mathscr{Y}_l^{m*}(\mathbf{p}) \\ &\quad \times e^{-\beta r}L_{n-l-1}^{(2l+2)}(2\beta r)\mathscr{Y}_l^m(\mathbf{r}) . \end{aligned} \quad (4.38)$$

In the same way we can formulate a weakly convergent expansion for a plane wave which is defined on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$. We only have to use Sturmians Eq. (4.6) and their Fourier transforms Eq. (4.24) in Eq. (3.23)

$$\begin{aligned} e^{i\mathbf{p}\cdot\mathbf{r}} &= (2\pi)^{3/2} \sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} \bar{\Psi}_{n,l}^{m*}(\beta,\mathbf{p})\Psi_{n,l}^{m}(\beta,\mathbf{r}) \\ &= 2\pi \sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} (2i)^l l! \frac{(n-l-1)!}{(n+l)!}\left\{\frac{4\beta^2}{\beta^2+p^2}\right\}^{l+2} \\ &\quad \times C_{n-l-1}^{l+1}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right)\mathscr{Y}_l^{m*}(\mathbf{p}) \\ &\quad \times e^{-\beta r}L_{n-l-1}^{(2l+1)}(2\beta r)\mathscr{Y}_l^m(\mathbf{r}) . \end{aligned} \quad (4.39)$$

At that stage it might be worth noting that in quantum mechanics Sobolev spaces are in some sense more important than Hilbert spaces. For instance, the Hilbert space $L^2(\mathbb{R}^3)$ contains elements like the Yukawa potential e^{-kr}/r which cannot be used as wave functions in atomic and molecular theory. On the other hand, it could be shown that the Rayleigh–Ritz variational procedure is closely related to approximation problems in Sobolev spaces.[22]

V. OSCILLATOR FUNCTIONS AND THEIR FOURIER TRANSFORMS

In this section we shall study a class of functions which form a complete orthonormal set in $L^2(\mathbb{R}^3)$ and which can again be written in the form

$$F_{nl}(\mathbf{r}) = R_{nl}(r)\mathscr{Y}_l^m(\mathbf{r}) . \quad (5.1)$$

Unlike the last section we now require that the radial part $R_{nl}(r)$ can be expressed as the product of a Gaussian function $e^{-\beta^2 r^2/2}$ and a polynomial in r. Again it is convenient to make use of the orthogonality properties of the generalized Laguerre polynomials. By the obvious substitution $x = y^2$ we obtain from Eq. (4.2) the orthogonality relationship

$$\int_0^\infty e^{-y^2}y^{2\alpha+1}L_m^{(\alpha)}(y^2)L_n^{(\alpha)}(y^2)dy = \frac{\Gamma(\alpha+n+1)}{2n!}\delta_{mn} . \quad (5.2)$$

Accordingly, we shall consider functions of the form

$$e^{-\beta^2 r^2/2}L_{n-l-1}^{(\alpha)}(\beta^2 r^2)\mathscr{Y}_l^m(\beta\mathbf{r}) . \quad (5.3)$$

The parameter α has to be chosen in such a way that these functions are orthogonal in $L^2(\mathbb{R}^3)$. Comparison of Eqs. (5.2) and (5.3) shows that $\alpha = l + \frac{1}{2}$ must hold and we find that the functions

$$\Omega_{n,l}^m(\beta,\mathbf{r}) = N_{n,l}(\beta)e^{-\beta^2 r^2/2}L_{n-l-1}^{(l+1/2)}(\beta^2 r^2)\mathscr{Y}_l^m(\beta\mathbf{r}) , \quad (5.4a)$$

$$N_{nl}(\beta) = \beta^{3/2}[2(n-l-1)!/\Gamma(n+\tfrac{1}{2})]^{1/2} \quad (5.4b)$$

are orthonormal in $L^2(\mathbb{R}^3)$ satisfying

$$\int \Omega_{n,l}^{m*}(\beta,\mathbf{r})\Omega_{n',l'}^{m'}(\beta,\mathbf{r})d^3\mathbf{r} = \delta_{nn'}\delta_{ll'}\delta_{mm'} . \quad (5.5)$$

The completeness of these functions in the Hilbert space $L^2(\mathbb{R}^3)$ can also be proved.[22] However, these functions $\Omega_{n,l}^m$ which were obtained by requiring that their radial part should be the product of a Gaussian function and of a generalized Laguerre polynomial and that they should be orthonormal in $L^2(\mathbb{R}^3)$ are also the solutions of the Schrödinger equation of a three-dimensional isotropic harmonic oscillator. Various applications of these oscillator functions in atomic, molecular, nuclear and elementary particle physics can be found in the book by Moshinsky.[43]

The oscillator functions $\Omega_{n,l}^m$ have another, very important property. One can show that for all integers $\kappa, \lambda, \mu, \nu \geq 0$ the inequality

$$\sup_{\mathbf{r}\in\mathbb{R}^3}\left|r^\kappa\left(\frac{\partial}{\partial x}\right)^\lambda\left(\frac{\partial}{\partial y}\right)^\mu\left(\frac{\partial}{\partial z}\right)^\nu\Omega_{n,l}^m(\beta,\mathbf{r})\right| < \infty \quad (5.6)$$

holds. Hence, the oscillator functions $\Omega_{n,l}^m$ are elements of the Schwartz space $\mathscr{S}(\mathbb{R}^3)$ which is of tantamount importance for the theory of distributions. It is important to note that the Λ functions [Eq. (4.4)] and the Sturmians [Eq. (4.6)] are not elements of $\mathscr{S}(\mathbb{R}^3)$. Although these functions decline faster than any power of r they do not possess continuous partial derivatives of all order at $\mathbf{r} = 0$. In fact one can show that an irreducible spherical tensor

$$F_l^m(\mathbf{r}) = f_l(r)Y_l^m(\mathbf{r}/r) \quad (5.7)$$

can only be analytic at $\mathbf{r} = 0$ and also an element of $\mathscr{S}(\mathbb{R}^3)$ if

$$F_l^m(\mathbf{r}) = R_l(r)\mathscr{Y}_l^m(\mathbf{r}) \quad (5.8)$$

with

$$R_l(r) = \sum_{n=0}^{\infty}\left\{\left(\frac{\partial}{\partial r}\right)^{2n}R_l(r)\right\}_{r=0}\frac{r^{2n}}{(2n)!} \quad (5.9)$$

holds in some neighborhood containing $r = 0$.

This statement can be proved by noting that according to Eq. (2.7) the solid harmonic $\mathscr{Y}_l^m$ is a homogeneous polynomial of degree l in x, y, and z and that odd powers of r cannot be differentiated arbitrarily often with respect to x, y, and z at $\mathbf{r} = 0$.

It is well known that the test function space $\mathscr{S}$ is invariant under Fourier transformation, i.e., the Fourier transform of a function $\phi \in \mathscr{S}$ again belongs to $\mathscr{S}$. Therefore, we may expect that the Fourier transforms of the oscillator functions should have a similar structure as the oscillator functions themselves. However, there are even much more far-reaching conclusions concerning the nature of the Fourier transforms of the oscillator functions possible. The oscillator functions $\Omega_{n,l}^m(\beta,\mathbf{r})$ are eigenfunctions of the differential operator

$$\beta^{-2}\nabla^2 - \beta^2 r^2 . \tag{5.10}$$

In the momentum representation this differential operator is replaced by

$$\beta^2\nabla_{\mathbf{p}}^2 - \beta^{-2}p^2 , \tag{5.11}$$

where $\nabla_{\mathbf{p}}$ is the gradient in momentum space. Since these two differential operators have the same structure we may conclude that the Fourier transforms of the oscillator functions should be proportional to $\Omega^m_{n,l}(\beta^{-1},\mathbf{p})$. Indeed, if we use the Rayleigh expansion [Eq. (1.2)] we obtain

$$\begin{aligned}\bar{\Omega}^m_{n,l}(\beta,\mathbf{p}) &= (2\pi)^{-3/2}\int e^{-i\mathbf{p}\cdot\mathbf{r}}\Omega^m_{n,l}(\beta,\mathbf{r})d^3\mathbf{r} \\ &= (-1)^{n-l-1}\beta^{-3/2}\left\{\frac{2(n-l-1)!}{\Gamma(n+1/2)}\right\}^{1/2} \\ &\quad\times e^{-p^2/2\beta^2}L^{(l+1/2)}_{n-l-1}\left(\frac{p^2}{\beta^2}\right)\mathcal{Y}^m_l\left(-\frac{i\mathbf{p}}{\beta}\right).\end{aligned} \tag{5.12}$$

Comparison of Eqs. (5.4) and (5.12) yields

$$\bar{\Omega}^m_{n,l}(\beta,\mathbf{p}) = (-1)^{n-1}i^l\Omega^m_{n,l}(\beta^{-1},\mathbf{p}) . \tag{5.13}$$

For the proof of Eq. (5.12) we need the integral

$$\int_0^\infty e^{-s^2/2}s^{\alpha+1}L^{(\alpha)}_n(s^2)J_\alpha(zs)ds = (-1)^n e^{-z^2/2}z^\alpha L^{(\alpha)}_n(z^2) , \tag{5.14}$$

which can be obtained by the substitutions $x=z^2$ and $t=s^2$ from the integral (MOS, p. 244)

$$\begin{aligned}\frac{1}{2}\int_0^\infty & e^{-t/2}t^{\alpha/2}L^{(\alpha)}_n(t)J_\alpha([xt]^{1/2})dt \\ &= (-1)^n e^{-x/2}x^{\alpha/2}L^{(\alpha)}_n(x) .\end{aligned} \tag{5.15}$$

If we now use the oscillator functions [Eq. (5.4)] and their Fourier transforms [Eq. (5.12)] in Eq. (3.11), we can formulate the following weakly convergent expansion of a plane wave which is defined on the Hilbert space $L^2(\mathbf{R}^3)$

$$\begin{aligned}e^{i\mathbf{p}\cdot\mathbf{r}} &= (2\pi)^{3/2}\sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l}\bar{\Omega}^{m*}_{n,l}(\beta,\mathbf{p})\Omega^m_{n,l}(\beta,\mathbf{r}) \\ &= 2(2\pi)^{3/2}\sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l}(-1)^{n-1}i^l\frac{(n-l-1)!}{\Gamma(n+\frac{1}{2})} \\ &\quad\times e^{-p^2/2\beta^2}L^{(l+1/2)}_{n-l-1}\left(\frac{p^2}{\beta^2}\right)\mathcal{Y}^{m*}_l(\mathbf{p}) \\ &\quad\times e^{-\beta^2r^2/2}L^{(l+1/2)}_{n-l-1}(\beta^2r^2)\mathcal{Y}^m_l(\mathbf{r}) .\end{aligned} \tag{5.16}$$

VI. THE SHIBUYA-WULFMAN EXPANSION OF A PLANE WAVE

In his famous article on the accidental degeneracy of the hydrogen atom, Fock[1] introduced the following set of variables:

$$\xi = \frac{2p_0p_x}{p_0^2+p^2} = \sin\alpha\sin\theta\cos\phi , \tag{6.1a}$$

$$\eta = \frac{2p_0p_y}{p_0^2+p^2} = \sin\alpha\sin\theta\sin\phi , \tag{6.1b}$$

$$\zeta = \frac{2p_0p_z}{p_0^2+p^2} = \sin\alpha\cos\theta , \tag{6.1c}$$

$$\chi = \frac{p_0^2-p^2}{p_0^2+p^2} = \cos\alpha . \tag{6.1d}$$

Here, p_0 is a scaling parameter. Obviously, these four variables introduced by Fock satisfy the relationship

$$\xi^2+\eta^2+\zeta^2+\chi^2 = 1 . \tag{6.2}$$

Hence, we see that the transformation (6.1) maps a point $\mathbf{p}$ of the three-dimensional momentum space onto a point on the surface of the four-dimensional unit sphere which is described by the angular variables α, θ, and ϕ. Consequently, every function $\psi(\mathbf{p})$ whose domain is the three-dimensional momentum space can be transformed into a function $\Psi(\alpha,\theta,\phi)$ which is defined on the surface of the four-dimensional unit sphere.

What is the motivation for such a transformation. Since the functions $\Psi(\alpha,\theta,\phi)$ are defined on the surface of the unit sphere in $\mathbf{R}^4$, one can try to relate them to the group of four-dimensional rotations O(4) hoping to detect additional symmetries which would not be obvious at all in three-dimensional momentum space, let alone in three-dimensional coordinate space.

For that purpose it is convenient to introduce the four-dimensional spherical harmonics. It is well known that the general n-dimensional spherical harmonics which are often called hyperspherical harmonics can be obtained by solving the homogeneous n-dimensional Laplace equation on the surface of the n-dimensional unit sphere.[44–46] In four-dimensional space the spherical harmonics are given by[46–50]

$$\begin{aligned}Y^m_{n,l}(\alpha,\theta,\phi) &= \pi_{n,l}2^{l+1}l!\left[\frac{n(n-l-1)!}{2\pi(n+l)!}\right]^{1/2} \\ &\quad\times\sin^l\alpha\, C^{l+1}_{n-l-1}(\cos\alpha)Y^m_l(\theta,\phi) .\end{aligned} \tag{6.3}$$

Here, $\pi_{n,l}$ is a phase factor with absolute value one. In the literature, different conventions for $\pi_{n,l}$ can be found. Stone[46] and Englefield[49] use $\pi_{n,l}=i^{n-l-1}$, Biedenharn[47] and Judd[48] use $\pi_{n,l}=(-i)^l$, and Sharp[48] uses $\pi_{n,l}=i^l$. The four-dimensional spherical harmonics Eq. (6.3) are orthonormal with respect to an integration over the surface of the four-dimensional unit sphere,

$$\begin{aligned}\int_0^\pi \sin^2\alpha\int_0^\pi\sin\theta\int_0^{2\pi} & Y^{m*}_{n,l}(\alpha,\theta,\phi)Y^{m'}_{n',l'}(\alpha,\theta,\phi)d\alpha\, d\theta\, d\phi \\ &= \delta_{nn'}\delta_{ll'}\delta_{mm'} .\end{aligned} \tag{6.4}$$

It should be noted that this normalization condition differs from the one originally introduced by Fock.[51] There, the functions

$$\begin{aligned}X^m_{n,l}(\alpha,\theta,\phi) &= \pi_{n,l}2^{l+1}l!\left[\frac{\pi n(n-l-1)!}{(n+l)!}\right]^{1/2} \\ &\quad\times\sin^l\alpha\, C^{l+1}_{n-l-1}(\cos\alpha)Y^m_l(\theta,\phi) ,\end{aligned} \tag{6.5}$$

are used which are normalized to give the surface of the unit sphere in $\mathbf{R}^4$,

$$\begin{aligned}\int_0^\pi \sin^2\alpha\int_0^\pi\sin\theta\int_0^{2\pi} & X^{m*}_{n,l}(\alpha,\theta,\phi)X^{m'}_{n',l'}(\alpha,\theta,\phi)d\alpha\, d\theta\, d\phi \\ &= 2\pi^2\delta_{nn'}\delta_{ll'}\delta_{mm'} .\end{aligned} \tag{6.6}$$

This normalization condition was used by Shibuya and Wulfman[52] who derived an expansion of a plane wave in terms of Sturmians Eq. (4.6) and the four-dimensional spherical harmonics Eq. (6.5) which in our notation reads

$$e^{i\mathbf{p}\cdot\mathbf{r}} = 4(\pi p_0)^{1/2} \sum_{n=1}^{\infty} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} (-1)^{n-l-1} \times i^l \Psi_{n,l}^{m}(p_0,\mathbf{r})[p_0^2 + p^2]^{-1} X_{n,l}^{m*}(\alpha,\theta,\phi). \tag{6.7}$$

In Eq. (6.7) Shibuya and Wulfman[53] used for the four-dimensional spherical harmonics the phase convention $\pi_{n,l} = 1$. It should be noted that in a later article on dynamical groups in atomic and molecular physics where expansion (6.7) was also treated, Wulfman[54] used the phase convention *of Biedenharn*,[47] $\pi_{nl} = (-i)^l$.

The Shibuya–Wulfman expansion Eq. (6.7) contains Sturmians just as the weakly convergent expansion (4.39) of a plane wave in terms of Sturmians and their Fourier transforms. Therefore, we want to find out how the expansions (4.39) and (6.7) are related. In particular, we want to know whether the Shibuya–Wulfman expansion is also a distribution and for which class of functions it is defined. For that purpose we first work out the connection between the four-dimensional spherical harmonics Eq. (6.5) and the Fourier transforms of Sturmians Eq. (4.24). From Eq. (6.1) we obtain immediately

$$\sin^l \alpha = [2p_0 p/(p_0^2 + p^2)]^l. \tag{6.8}$$

In addition, we use the fact that the Gegenbauer polynomials have either even or odd parity (MOS, p. 218),

$$C_n^{\lambda}(x) = (-1)^n C_n^{\lambda}(-x), \tag{6.9}$$

to express the Fourier transform of a Sturmian Eq. (4.24) in terms of functions depending upon the angular variables α, θ, and ϕ,

$$\frac{(p_0^2 + p^2)^2}{(2p_0)^{5/2}} \overline{\Psi}_{n,l}^{m}(p_0,\mathbf{p}) = (-1)^{n-1} i^l 2^l l! \left[\frac{n(n-l-1)!}{\pi(n+l)!}\right]^{1/2} \times \sin^l \alpha\, C_{n-l-1}^{l+1}(\cos\alpha) Y_l^m(\theta,\phi). \tag{6.10}$$

Comparison with Eq. (6.5) yields the relationship

$$X_{n,l}^{m}(\alpha,\theta,\phi) = (-1)^{n-1}(-i)^l \pi_{n,l} 2\pi \frac{(p_0^2 + p^2)^2}{(2p_0)^{5/2}} \overline{\Psi}_{n,l}^{m}(p_0,\mathbf{p}). \tag{6.11}$$

If we insert this relationship into the integral (6.6) and also use[55]

$$\sin^2\alpha \sin\theta\, d\alpha\, d\theta\, d\phi = [2p_0/(p_0^2 + p^2)]^3 d^3\mathbf{p}, \tag{6.12}$$

we find

$$\int_0^{\pi} \sin^2\alpha \int_0^{\pi} \sin\theta \int_0^{2\pi} X_{n,l}^{m*}(\alpha,\theta,\phi) X_{n',l'}^{m'}(\alpha,\theta,\phi) d\alpha\, d\theta\, d\phi = 2\pi^2 \int \overline{\Psi}_{n,l}^{m*}(p_0,\mathbf{p}) \frac{p_0^2 + p^2}{2p_0^2} \overline{\Psi}_{n',l'}^{m'}(p_0,\mathbf{p}) d^3\mathbf{p} = 2\pi^2 \delta_{nn'} \delta_{ll'} \delta_{mm'}. \tag{6.13}$$

Hence, the variable transformation (6.1) connects the four-dimensional spherical harmonics and the Sturmians in a one-to-one fashion. We also see that the orthogonality of the four-dimensional spherical harmonics with respect to an integration over the surface of the four-dimensional unit sphere and the orthogonality of the Sturmians with respect to the norm of the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ are equivalent.

We now insert Eq. (6.10) together with the phase convention $\pi_{n,l} = 1$ into the Shibuya–Wulfman expansion [Eq. (6.7)] and obtain

$$e^{i\mathbf{p}\cdot\mathbf{r}} = (2\pi)^{3/2} \sum_{n=1}^{\infty} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} \Psi_{n,l}^{m}(p_0,\mathbf{r}) \times \frac{p_0^2 + p^2}{2p_0^2} \overline{\Psi}_{n,l}^{m*}(p_0,\mathbf{p}). \tag{6.14}$$

The validity of this expansion can be checked in exactly the same way as we did it in the case of Eq. (3.23). Consider some function $f \in W_2^{(1)}(\mathbb{R}^2)$. Then $f(\mathbf{r})$ can be expanded in terms of Sturmians Eq. (4.6) or equivalently, its Fourier transform $\overline{f}(\mathbf{p})$ can be expanded in terms of the Fourier transforms of Sturmians [Eq. (4.24)],

$$f(\mathbf{r}) = \sum_{nlm} \gamma_{nl}^{m} \Psi_{n,l}^{m}(p_0,\mathbf{r}), \tag{6.15}$$

$$\overline{f}(\mathbf{p}) = \sum_{nlm} \gamma_{nl}^{m} \overline{\Psi}_{n,l}^{m}(p_0,\mathbf{p}), \tag{6.16}$$

$$\gamma_{nl}^{m} = \int \Psi_{n,l}^{m*}(p_0,\mathbf{r}) \frac{p_0^2 - \nabla^2}{2p_0^2} f(\mathbf{r}) d^3\mathbf{r} = \int \overline{\Psi}_{n,l}^{m*}(p_0,\mathbf{p}) \frac{p_0^2 + p^2}{2p_0^2} \overline{f}(\mathbf{p}) d^3\mathbf{p}. \tag{6.17}$$

The two expansions (6.15) and (6.16) converge both in the sense of the norm of the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ [Eq. (2.17)].

We check the correctness of expansion (6.14) by using it in the Fourier integral (1.4). If we integrate termwise we find an expansion which is identical with Eq. (6.15):

$$f(\mathbf{r}) = \sum_{nlm} \Psi_{n,l}^{m}(p_0,\mathbf{r}) \int \overline{\Psi}_{n,l}^{m*}(p_0,\mathbf{p}) \frac{p_0^2 + p^2}{2p_0^2} \overline{f}(\mathbf{p}) d^3\mathbf{p} \tag{6.18}$$

$$= \sum_{nlm} \gamma_{nl}^{m} \Psi_{n,l}^{m}(p_0,\mathbf{r}). \tag{6.19}$$

If we compare the expansions (4.39) and (6.14) we see that in Eq. (6.14) the weight function $[p_0^2 + p^2]/(2p_0^2)$ which is needed to make the Fourier transforms of Sturmians orthonormal is explicitly included. Therefore, it is not surprising that Eq. (6.14) gives the correct result in the case of an integration over $\mathbf{p}$. However, in the case of an integration over $\mathbf{r}$, the Sturmians alone, i.e., without the weight function $(p_0^2 - \nabla^2)/(2p_0^2)$, are no longer an orthonormal set. Thus, it is by no means obvious that expansion (6.14) is also correct in the case of an integration over $\mathbf{r}$. We check this question by computing the Fourier transform of a Sturmian. For that purpose it is advantageous to evaluate first the integral over two Sturmians alone, i.e., without the differential operator $(p_0^2 - \nabla^2)/(2p_0^2)$. We use the recurrence relationship of the generalized Laguerre polynomials (MOS, p. 241)

$$nL_n^{(\alpha)}(x) = (2n + \alpha - 1 - x)L_{n-1}^{(\alpha)}(x) - (n + \alpha + 1)L_{n-2}^{(\alpha)}(x) \tag{6.20}$$

to derive the following three-term recurrence formula of Sturmians:

$$\begin{aligned} p_0 r \Psi_{n,l}^m(p_0,\mathbf{r}) = & -[(n-l)(n+1)(n+l+1)/4n]^{1/2} \\ & \times \Psi_{n+1,l}^m(p_0,\mathbf{r}) + n\Psi_{n,l}^m(p_0,\mathbf{r}) \\ & -[(n-l-1)(n-1)(n+l)/4n]^{1/2} \\ & \times \Psi_{n-1,l}^m(p_0,\mathbf{r}) . \end{aligned} \tag{6.21}$$

The integral of two Sturmians can now be computed by combining this recurrence formula with the orthogonality relationship (4.7),

$$\begin{aligned} & \int \Psi_{n',l'}^{m'*}(p_0,\mathbf{r})\Psi_{n,l}^m(p_0,\mathbf{r})d^3\mathbf{r} \\ & = \left\{ -\left[\frac{(n-l)(n+l+1)}{4n(n+1)}\right]^{1/2} \delta_{n'n+1} + \delta_{n'n} \right. \\ & \left. -\left[\frac{(n-l-1)(n+l)}{4n(n-1)}\right]^{1/2} \delta_{n'n-1} \right\} \delta_{l'l}\delta_{m'm} . \end{aligned} \tag{6.22}$$

If expansion (6.14) is used in connection with this expression, we obtain for the Fourier transform of a Sturmian

$$\begin{aligned} \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) & = (2\pi)^{-3/2} \int e^{-i\mathbf{p}\cdot\mathbf{r}} \Psi_{n,l}^m(p_0,\mathbf{r})d^3\mathbf{r} \\ & = \sum_{n'l'm'} \frac{p_0^2 + p^2}{2p_0^2} \overline{\Psi}_{n',l'}^{m'}(p_0,\mathbf{p}) \\ & \times \int \Psi_{n',l'}^{m'*}(p_0,\mathbf{r})\Psi_{n,l}^m(p_0,\mathbf{r})d^3\mathbf{r} \end{aligned} \tag{6.23}$$

$$\begin{aligned} & = \frac{p_0^2 + p^2}{2p_0^2} \left\{ -\left[\frac{(n-l)(n+l+1)}{4n(n+1)}\right]^{1/2} \right. \\ & \times \overline{\Psi}_{n+1,l}^m(p_0,\mathbf{p}) + \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) \\ & \left. -\left[\frac{(n-l-1)(n+l)}{4n(n-1)}\right]^{1/2} \overline{\Psi}_{n-1,l}^m(p_0,\mathbf{p}) \right\} . \end{aligned} \tag{6.24}$$

With the help of Eq. (4.24) it can be shown that Eq. (6.24) is equivalent to the following relationship between Gegenbauer polynomials:

$$\begin{aligned} & \frac{p^2 - p_0^2}{p^2 + p_0^2} C_{n-l-1}^{l+1}\left(\frac{p^2 - p_0^2}{p^2 + p_0^2}\right) \\ & = \frac{n-l}{2n} C_{n-l}^{l+1}\left(\frac{p^2 - p_0^2}{p^2 + p_0^2}\right) \\ & + \frac{n+l}{2n} C_{n-l-2}^{l+1}\left(\frac{p^2 - p_0^2}{p^2 + p_0^2}\right). \end{aligned} \tag{6.25}$$

However, Eq. (6.25) is obviously identical with the homogeneous recurrence formula of the Gegenbauer polynomials (MOS, p. 222),

$$(n+1)C_{n+1}^\lambda(x) = 2(n+\lambda)xC_n^\lambda(x) - (n+2\lambda-1)C_{n-1}^\lambda(x) . \tag{6.26}$$

Thus we have shown that expansion (6.14) yields the correct result when used for the calculation of the Fourier transform of a Sturmian. Since all functions $f \in W_2^{(1)}(\mathbb{R}^3)$ can be expanded in terms of Sturmians we may conclude that Eq. (6.14) holds as a distribution on $W_2^{(1)}(\mathbb{R}^3)$ also for an integration over $\mathbf{r}$.

In Eq. (6.14) the weight function $(p_0^2 + p^2)/(2p_0^2)$ can be absorbed in the Fourier transform of a Sturmian according to Eq. (4.29). We then obtain

$$e^{i\mathbf{p}\cdot\mathbf{r}} = (2\pi)^{3/2} \sum_{n=1}^{\infty} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} \overline{\Phi}_{n,l}^{m*}(p_0,\mathbf{p})\Psi_{n,l}^m(p_0,\mathbf{r}) . \tag{6.27}$$

If written in this form the Shibuya–Wulfmann expansion looks like an expansion $e^{i\mathbf{p}\cdot\mathbf{r}}$ in terms of the biorthogonal sets $\{\Psi_{n,l}^m\}$ and $\{\Phi_{n,l}^m\}$. Now the question arises whether it would be possible in Eq. (6.27) to invert the role played by the sets $\{\Psi_{n,l}^m\}$ and $\{\Phi_{n,l}^m\}$. This means that we want to find out whether the following relationship which may be considered to be a kind of mirror image of Eq. (6.27) also holds:

$$e^{i\mathbf{p}\cdot\mathbf{r}} = (2\pi)^{3/2} \sum_{n=1}^{\infty} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} \overline{\Psi}_{n,l}^{m*}(p_0,\mathbf{p})\Phi_{n,l}^m(p_0,\mathbf{r}) . \tag{6.28}$$

The proof of Eq. (6.28) for an integration over $\mathbf{r}$ is trivial. From the definition of the biorthogonal functions Eq. (4.14) we may immediately deduce

$$\begin{aligned} \gamma_{nl}^m & = \int \Psi_{n,l}^{m*}(p_0,\mathbf{r}) \frac{p_0^2 - \nabla^2}{2p_0^2} f(\mathbf{r})d^3\mathbf{r} \\ & = \int \Phi_{n,l}^{m*}(p_0,\mathbf{r}) f(\mathbf{r})d^3\mathbf{r} . \end{aligned} \tag{6.29}$$

Hence, if we insert Eq. (6.28) into the Fourier integral (1.3), we obtain an expansion which is identical with Eq. (6.16)

$$\bar{f}(\mathbf{p}) = \sum_{nlm} \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) \int \Phi_{n,l}^{m*}(p_0,\mathbf{r}) f(\mathbf{r})d^3\mathbf{r} \tag{6.30}$$

$$= \sum_{nlm} \gamma_{nl}^m \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) . \tag{6.31}$$

We now have to find out whether Eq. (6.28) also holds for an integration over $\mathbf{p}$. For that purpose we use the recurrence formula of the Gegenbauer polynomials Eq. (6.26) to derive a homogeneous three-term recurrence formula for the Fourier transforms of Sturmians,

$$\begin{aligned} & \frac{2p_0^2}{p_0^2 + p^2} \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) \\ & = -\left[\frac{(n-l)(n+l+1)}{4n(n+1)}\right]^{1/2} \overline{\Psi}_{n+1,l}^m(p_0,\mathbf{p}) \\ & + \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) - \left[\frac{(n-l-1)(n+l)}{4n(n-1)}\right]^{1/2} \\ & \times \overline{\Psi}_{n-1,l}^m(p_0,\mathbf{p}) . \end{aligned} \tag{6.32}$$

If we use this relationship in connection with the orthogonality relationship (4.36) we find

$$\begin{aligned} & \int \overline{\Psi}_{n',l'}^{m'*}(p_0,\mathbf{p})\overline{\Psi}_{n,l}^m(p_0,\mathbf{p})d^3\mathbf{p} \\ & = \left\{ -\left[\frac{(n-l)(n+l+1)}{4n(n+1)}\right]^{1/2} \delta_{n'n+1} \right. \\ & \left. + \delta_{n'n} - \left[\frac{(n-l-1)(n+l)}{4n(n-1)}\right]^{1/2} \delta_{n'n-1} \right\} \delta_{ll'}\delta_{mm'} . \end{aligned} \tag{6.33}$$

The selection rules here and in Eq. (6.22) are identical. This is a consequence of the well-known fact that scalar products are invariant under Fourier transformation.

For the sake of simplicity we proceed as in the case of Eq. (6.14), i.e., we use expansion (6.28) in connection with Eq. (6.33) for the computation of the inverse Fourier transform of a Sturmian,

$$\Psi_{n,l}^m(p_0,\mathbf{r}) = (2\pi)^{-3/2} \int e^{i\mathbf{r}\cdot\mathbf{p}} \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) d^3\mathbf{p}$$

$$= \sum_{n'l'm'} \Phi_{n',l'}^{m'}(p_0,\mathbf{r}) \times \int \overline{\Psi}_{n',l'}^{m'*}(p_0,\mathbf{p}) \overline{\Psi}_{n,l}^m(p_0,\mathbf{p}) d^3\mathbf{p} \tag{6.34}$$

$$= -\left[\frac{(n-l)(n+l+1)}{4n(n+1)}\right]^{1/2} \Phi_{n+1,l}^m(p_0,\mathbf{r}) + \Phi_{n,l}^m(p_0,\mathbf{r}) - \left[\frac{(n-l-1)(n+l)}{4n(n-1)}\right]^{1/2} \Phi_{n-1}^m(p_0,\mathbf{r}). \tag{6.35}$$

In view of Eqs. (4.6) and (4.14) it can be shown that Eq. (6.35) is equivalent to the following relationship between generalized Laguerre polynomials:

$$p_0 r L_{n-l-1}^{(2l+1)}(2p_0 r) = [(n-l)/2] L_{n-l}^{(2l+1)}(2p_0 r) + n L_{n-l-1}^{(2l+1)}(2p_0 r) - [(n+l)/2] L_{n-l-2}^{(2l+1)}(2p_0 r). \tag{6.36}$$

However, Eq. (6.36) is equivalent to the homogeneous three-term recurrence formula of the generalized Laguerre polynomials Eq. (6.20).

Thus we have shown that expansion (6.28) yields the correct result when used for the calculation of the inverse Fourier transform of a Sturmian. Since the Fourier transforms of all functions $f \in W_2^{(1)}(\mathbf{R}^3)$ can be expanded in terms of Fourier transforms of Sturmians we may conclude that Eq. (6.28) holds as a distribution on $W_2^{(1)}(\mathbf{R}^3)$ also for an integration over $\mathbf{p}$.

It should be noted that the Shibuya–Wulfman expansion [Eq. (6.27)] as well as its mirror image [Eq. (6.28)] are not in general defined for functions $f \in L^2(\mathbf{R}^3)$. The restriction to elements of the Sobolev space $W_2^{(1)}(\mathbf{R}^3)$ is essential. This follows from the fact that for functions $f \in L^2(\mathbf{R}^3)$ the expansions (6.15) and (6.16) need not converge in $W_2^{(1)}(\mathbf{R}^3)$ even if the expansion coefficients [Eq. (6.17)] all exist. We have also proved the somewhat surprising result that for functions $f \in W_2^{(1)}(\mathbf{R}^3)$ the orthogonal expansion [Eq. (4.39)], the Shibuya-Wulfman expansion [Eq. (6.27)], and the biorthogonal expansion [Eq. (6.28)] are all identical as distributions on $W_2^{(1)}(\mathbf{R}^3)$ since they lead to the same expansions for either $f(\mathbf{r})$ or its Fourier transform $\overline{f}(\mathbf{p})$.

VII. ON THE DERIVATION OF ADDITION THEOREMS

In the theory of atoms, molecules, and solids one is often confronted with the problem of expressing a function $f(\mathbf{r}-\mathbf{R})$ which depends on two variables $\mathbf{r}$ and $\mathbf{R}$ in terms of functions that depend either upon $\mathbf{r}$ or upon $\mathbf{R}$. Expansions of that kind are usually called addition theorems. The probably best-known example of such an addition theorem is the Laplace expansion of the Coulomb potential in spherical coordinates,

$$\frac{1}{|\mathbf{r}-\mathbf{R}|} = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{4\pi}{2l+1} \frac{r_<^l}{r_>^{l+1}} Y_l^{m*}\left(\frac{\mathbf{r}}{r}\right) Y_l^m\left(\frac{\mathbf{R}}{R}\right), \tag{7.1a}$$

$$r_< = \min(r,R), \quad r_> = \max(r,R). \tag{7.1b}$$

How can such addition theorems be derived. One of the standard techniques is the use of Fourier transforms. This method was introduced independently by Ruedenberg[56] and Silverstone.[57] It is of course clear that the Fourier transform method is restricted to functions f where the Fourier integrals (1.3) and (1.4) are meaningful.

According to Eq. (1.4) a function $f(\mathbf{r}-\mathbf{R})$ can be represented as an inverse Fourier integral,

$$f(\mathbf{r}-\mathbf{R}) = (2\pi)^{-3/2} \int e^{i(\mathbf{r}-\mathbf{R})\cdot\mathbf{p}} \overline{f}(\mathbf{p}) d^3\mathbf{p}. \tag{7.2}$$

A separation of the variables $\mathbf{r}$ and $\mathbf{R}$ can be achieved if the Rayleigh expansion of a plane wave Eq. (1.2) is inserted twice into the integral, once for $e^{i\mathbf{r}\cdot\mathbf{p}}$ and once for $e^{-i\mathbf{R}\cdot\mathbf{p}}$,

$$f(\mathbf{r}-\mathbf{R}) = (32\pi)^{1/2} \sum_{l_1 m_1} \sum_{l_2 m_2} i^{l_1 - l_2} Y_{l_1}^{m_1}\left(\frac{\mathbf{r}}{r}\right) Y_{l_2}^{m_2}\left(\frac{\mathbf{R}}{R}\right) \times \int j_{l_1}(rp) Y_{l_1}^{m_1*}\left(\frac{\mathbf{p}}{p}\right) j_{l_2}(Rp) Y_{l_2}^{m_2*}\left(\frac{\mathbf{p}}{p}\right) \overline{f}(\mathbf{p}) d^3\mathbf{p}. \tag{7.3}$$

In most physical applications one is only interested in addition theorems of irreducible tensors

$$F_l^m(\mathbf{r}) = f_l(r) Y_l^m(\mathbf{r}/r). \tag{7.4}$$

The integral representation (7.3) can then be simplified further by introducing Gaunt coefficients which are defined in Eq. (2.8),

$$F_l^m(\mathbf{r}-\mathbf{R}) = (32\pi)^{1/2} \sum_{l_1 m_1} \sum_{l_2 m_2} i^{l_1 - l_2} \langle lm | l_1 m_1 | l_2 m_2 \rangle \times Y_{l_1}^{m_1}\left(\frac{\mathbf{r}}{r}\right) Y_{l_2}^{m_2}\left(\frac{\mathbf{R}}{R}\right) \times \int_0^\infty p^2 j_{l_1}(rp) j_{l_2}(Rp) \overline{f}_l(p) dp. \tag{7.5}$$

The function $\overline{f}_l(p)$ is defined in Eq. (1.6). Due to the selection rules satisfied by the Gaunt coefficient[58] the two infinite summations over l_1 and l_2 in Eq. (7.5) are no longer independent and one of them terminates after a finite number of terms.

We see that in Eq. (7.5) the angular parts of the variables $\mathbf{r}$ and $\mathbf{R}$ are already separated. Therefore, one has succeeded in deriving an addition theorem for the function $F_l^m(\mathbf{r}-\mathbf{R})$ if one is able to evaluate the remaining radial integral in Eq. (7.5) in such form that the radial variables r and R are separated. Unfortunately, this turns out to be a major obstacle. Compared with the radial parts of ordinary Fourier integrals the radial integral in Eq. (7.5) contains not one, but two spherical Bessel functions. In some cases such radial integrals could be evaluated. For instance, Silverstone[57] was able to derive an addition theorem for Slater-type functions with the help of the Fourier transform method. However, in many cases of interest it is virtually impossible to evaluate the remaining radial integrals involving two spherical Bessel functions.

There is also another annoying problem. The Fourier transform method in connection with the Rayleigh expansion leads to addition theorems that converge pointwise. Unfortunately, these addition theorems are often infinite series which only have a finite radius of convergence. This implies that different regions of space have to be distinguished where the addition theorem assumes different functional forms. A typical example is the two-range form of the Laplace expansion of the Coulomb potential Eq. (7.1), where the regions $r<R$ and $r>R$ have to be distinguished. Addition theorems are normally used in integrals. There, the two-range form of an addition theorem has some unpleasant consequences since indefinite integrals are now needed. This is a severe restriction of the applicability of an addition theorem because compared to definite integrals only a relatively small number of indefinite integrals is known. Thus we see that the knowledge of an addition theorem may not be sufficient for the evaluation of an integral if the use of the addition theorem leads to indefinite integrals which cannot be computed in a reasonable way.

Therefore, we want to propose a modification of the Fourier transform method for the derivation of addition theorems which avoids the two-range form of addition theorems completely. In Fourier integrals like Eq. (7.2) one should not use the Rayleigh expansion [Eq. (1.2)] but instead one of the weakly convergent expansions which were discussed in this article.

Of course, this approach is restricted to functions f that are either elements of $L^2(\mathbf{R}^3)$ or of $W_2^{(1)}(\mathbf{R}^3)$. However, this restriction is not very severe since most functions that are of interest in atomic or molecular physics belong to these spaces. Also, the addition theorems would then be expansions that in general would only converge with respect to some norms and not pointwise. But in integrals where addition theorems are normally used, pointwise convergence is in most cases not necessary.

In this article we shall only discuss addition theorems that are derived with the help of the weakly convergent expansion of a plane wave in terms of Λ functions [Eq. (4.37)]. But the general conclusions at which we shall arrive are equally valid if other weakly convergent expansions of a plane wave are to be used.

If we insert into the Fourier integral (7.2) twice the Λ function expansion of a plane wave [Eq. (4.38)] we obtain

$$f(\mathbf{r}-\mathbf{R}) = (2\pi)^{3/2} \sum_{n_1 l_1 m_1} \sum_{n_2 l_2 m_2} (-1)^{l_2} \Lambda_{n_1,l_1}^{m_1}(\beta,\mathbf{r}) \Lambda_{n_2,l_2}^{m_2}(\beta,\mathbf{R}) \times \int \overline{\Lambda}_{n_1,l_1}^{m_1^*}(\beta,\mathbf{p}) \overline{\Lambda}_{n_2,l_2}^{m_2^*}(\beta,\mathbf{p}) \bar{f}(\mathbf{p}) d^3\mathbf{p} . \tag{7.6}$$

We have already achieved a complete separation of the variables $\mathbf{r}$ and $\mathbf{R}$ since they only occur in the Λ functions and the remaining momentum space integrals depend only upon the indices n_1, l_1, m_1, n_2, l_2, and m_2 and upon the scaling parameter β.

A further simplification is possible if spherical tensors $F_l^m(\mathbf{r}-\mathbf{R})$ are considered. We then obtain with the help of Eq. (4.25),

$$F_l^m(\mathbf{r}-\mathbf{R}) = (32\pi)^{1/2}\beta^3 \sum_{n_1 l_1 m_1} \sum_{n_2 l_2 m_2} i^{l_1-l_2} \Lambda_{n_1,l_1}^{m_1}(\beta,\mathbf{r}) \Lambda_{n_2,l_2}^{m_2}(\beta,\mathbf{R}) \times \langle lm|l_1 m_1|l_2 m_2\rangle \frac{[(n_1-l_1-1)!(n_1+l_1+1)!(n_2-l_2-1)!(n_2+l_2+1)!]^{1/2}}{(1/2)_{n_1}(1/2)_{n_2}} \times \int_0^\infty P_{n_1-l_1-1}^{(l_1+3/2,l_1+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) P_{n_2-l_2-1}^{(l_2+3/2,l_2+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) \frac{(\beta p)^{l_1+l_2+2}}{[p^2+\beta^2]^{l_1+l_2+4}} \bar{f}_l(p) dp . \tag{7.7}$$

We see that we have derived an addition theorem for the irreducible spherical tensor $F_l^m(\mathbf{r}-\mathbf{R})$ which is given in the form of an expansion in terms of Λ functions as soon as we are able to compute the remaining radial integrals in momentum space. However, unlike the radial integrals in Eq. (7.5) which depend upon r and R and which involve spherical Bessel functions, the remaining radial integrals in Eq. (7.7) are simply numbers and can, if no better way is found, even be evaluated by numerical quadrature. This would not be possible in the case of the radial integrals in Eq. (7.5).

Let us now consider the addition theorem of Λ functions. With the help of Eq. (4.25) we obtain

$$\Lambda_{n,l}^m(\beta,\mathbf{r}-\mathbf{R}) = 8\beta^5 \sum_{n_1 l_1 m_1} \sum_{n_2 l_2 m_2} i^{l_1-l_2-l} \Lambda_{n_1,l_1}^{m_1}(\beta,\mathbf{r}) \Lambda_{n_2,l_2}^{m_2}(\beta,\mathbf{R}) \times \langle lm|l_1 m_1|l_2 m_2\rangle \frac{[2\beta(n_1-l_1-1)!(n_1+l_1+1)!(n_2-l_2-1)!(n_2+l_2+1)!(n-l-1)!(n+l+1)!]^{1/2}}{(1/2)_{n_1}(1/2)_{n_2}(1/2)_n} \times \int_0^\infty P_{n_1-l_1-1}^{(l_1+3/2,l_1+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) P_{n_2-l_2-1}^{(l_2+3/2,l_2+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) P_{n-l-1}^{(l+3/2,l+1/2)}\left(\frac{p^2-\beta^2}{p^2+\beta^2}\right) \times \frac{(\beta p)^{l_1+l_2+l+2}}{[p^2+\beta^2]^{l_1+l_2+l+6}} dp . \tag{7.8}$$

However, this result was already derived by Filter and Steinborn[59] in a completely different way. They also evaluated the remaining radial integrals in Eq. (7.8) and showed that due to the orthogonality properties of the functions involved, the momentum space radial integrals are different from zero only if certain selection rules are fulfilled. The numerical properties of this Λ function addition theorem were investigated by Trivedi and Steinborn.[60]

How do the addition theorems for irreducible spherical tensors $F_l^m(\mathbf{r}-\mathbf{R})$ that are derived using the Rayleigh expansion of a plane wave according to Eq. (7.5) compare with addition theorems that are derived with the help of weakly convergent expansions of a plane wave as the Λ function addition theorem Eq. (7.8). In all addition theorems that are based on weakly convergent expansions the variables **r** and **R** are completely separated. Consequently, it is not necessary to distinguish different regions of space in which the addition theorem assumes different functional forms. This is quite advantageous if such an addition theorem is used in an integral since indefinite integrals do not occur. Also, the fact that the variables **r** and **R** occur in an addition theorem like Eq. (7.8) only as arguments or orthogonal functions facilitates integrations greatly. We therefore believe that these structural advantages of addition theorems which are based on weakly convergent expansions make them superior in most applications.

It seems that these ideas should be pursued also for other functions beside Λ functions and one should also use other weakly convergent expansions of a plane wave. For instance, Novosadov[61] used the Shibuya–Wulfman expansion [Eq. (6.23)] for the derivation of addition theorems and the evaluation of multicenter integrals involving exponentially declining functions.

VIII. SUMMARY AND CONCLUSIONS

The standard way of computing the Fourier transform of an irreducible spherical tensor $F_l^m(\mathbf{r})$ consists in using the Rayleigh expansion of a plane wave in terms of spherical Bessel functions and spherical harmonics. Due to the orthonormality of the spherical harmonics the angular integration is then trivial and only a radial integral involving a spherical Bessel function remains to be done. However, the evaluation of integrals involving spherical Bessel functions is usually not at all easy and in some cases even impossible.

The Rayleigh expansion of a plane wave converges pointwise. However, when used in integrals the pointwise convergence of an expansion is not always needed and in many cases it is sufficient to use weakly convergent expansions.

As an alternative to the Rayleigh expansion we construct expansions of a plane wave in terms of complete orthonormal sets of functions and their Fourier transforms which may be viewed as distributions that are defined on either the Hilbert space $L^2(\mathbb{R}^3)$ or on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$. This means that the use of these distributions in Fourier integrals leads to orthogonal expansions of the (inverse) Fourier transforms which converge in the sense of the norm of either $L^2(\mathbb{R}^3)$ or $W_2^{(1)}(\mathbb{R}^3)$.

Complete orthonormal sets of functions and their Fourier transforms are used for the construction of the weakly convergent expansions. Accordingly, the properties of some complete orthonormal sets in $L^2(\mathbb{R}^3)$ and $W_2^{(1)}(\mathbb{R}^3)$ are studied and their Fourier transforms are calculated. It is demonstrated that the Fourier transforms of various exponentially declining functions (among them hydrogen eigenfunctions) can be computed in a unified way which is much simpler than the methods which were hitherto known. Beside exponentially declining functions the eigenfunctions of the three-dimensional isotropic harmonic oscillator are studied which decline like a Gaussian function. It is shown that the oscillator eigenfunctions are elements of the Schwartz space $\mathscr{S}(\mathbb{R}^3)$ of rapidly decreasing functions and that they have particular invariance properties under Fourier transformation.

Shibuya and Wulfman derived an expansion of a plane wave in terms of Sturmians and the four-dimensional spherical harmonics. However, the four-dimensional spherical harmonics are closely related to the Fourier transforms of the Sturmians and their orthogonality with respect to an integration over the four-dimensional unit sphere is equivalent to the orthogonality of the Sturmians in the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$. Accordingly, the Shibuya–Wulfman expansion as well as some other, closely related expansion is a distribution which is defined on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$. It seems that this fact as well as the intimate relationship between the four-dimensional spherical harmonics and the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ has so far been overlooked in the literature.

As a practical application it is suggested to use the weakly convergent expansions of a plane wave for the derivation of addition theorems. If addition theorems are derived via the Fourier transform method using the Rayleigh expansion of a plane wave it is often not possible to obtain a complete separation of the variables since the resulting expansions may assume different analytical forms in different regions of space. This is a consequence of the fact that the use of the Rayleigh expansion leads to addition theorems that converge pointwise. However, if weakly convergent expansions of a plane wave are used for the derivation of addition theorems, the resulting addition theorems converge only in the sense of the norm of either $L^2(\mathbb{R}^3)$ or $W_2^{(1)}(\mathbb{R}^3)$ but a complete separation of the variables is always possible. This fact facilitates the application of these addition theorems in integrals considerably since it is not necessary to distinguish different regions of space and no indefinite integrals are needed. As an example for these norm-convergent addition theorems we analyze the structure of the addition theorem of Λ functions which are exponentially declining and are a complete orthonormal set in $L^2(\mathbb{R}^3)$.

It may be concluded that in all cases where further mathematical manipulation of Fourier transforms, in particular integrations, have to be done, the weakly convergent expansions of a plane wave should have distinct advantages over the Rayleigh expansion. It should also be noted that the construction of weakly convergent expansions of a plane wave which may be viewed as distributions that are defined either on the Hilbert space $L^2(\mathbb{R}^3)$ or on the Sobolev space $W_2^{(1)}(\mathbb{R}^3)$ is not limited to the use of spherical polar coordinates. Hence, this approach may be generalized to other coordinate systems in $\mathbb{R}^3$ or even to the n-dimensional space $\mathbb{R}^n$.

ACKNOWLEDGMENTS

I would like to thank Professor J. Cizek and Professor J. Paldus for the invitation to work with them in the Quantum

Theory Group of the Department of Applied Mathematics at the University of Waterloo. Their hospitality, their generosity, and the inspiring atmosphere which they provided is highly appreciated. I would also like to thank Professor K. Davidson for helpful and clarifying discussions.

[1]V. Fock, Z. Phys. **98**, 145 (1935).
[2]J. D. Talman, J. Comput. Phys. **29**, 35 (1978), Comput. Phys. Commun. **30**, 93 (1983).
[3]D. R. Lehman, W. C. Parke, and L. C. Maximon, J. Math. Phys. **22**, 1399 (1981).
[4]R. Wong, Numer. Math. **39**, 351 (1982).
[5]A. Sidi, Math. Comput. **38**, 517 (1982).
[6]D. Levin, Math. Comput. **38**, 531 (1982).
[7]W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics* (Springer-Verlag, New York, 1966); hereafter, this reference will be denoted as MOS in the text.
[8]E. U. Condon and G. H. Shortley, *The Theory of Atomic Spectra* (Cambridge U.P., Cambridge, England, 1970), p. 48.
[9]L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison-Wesley, Reading, MA, 1981), p. 71, Eq. (3.153).
[10]E. O. Steinborn and E. Filter, Theoret. Chim. Acta **38**, 273 (1975), p. 275, Eq. (3.1).
[11]E. J. Weniger and E. O. Steinborn, Phys. Rev. A **28**, 2026 (1983), see p. 2028, Eq. (3.7).
[12]E. Grosswald, *Bessel Polynomials* (Springer-Verlag, Berlin, 1978), and references therein.
[13]E. Filter and E. O. Steinborn, Phys. Rev. A **18**, 1 (1978), see p. 2, Eq. (2.14).
[14]S. L. Sobolev, *Applications of Functional Analysis in Mathematical Physics* (American Mathematical Society, Providence, RI, 1963).
[15]R. A. Adams, *Sobolev Spaces* (Academic, New York, 1975).
[16]K. Yosida, *Functional Analysis* (Springer-Verlag, Berlin, 1971), p. 154, Corollary 2.
[17]L. Schwarz, *Théorie des distributions* (Hermann, Paris, 1966), p. 234.
[18]See Ref. 17, p. 237.
[19]See Ref. 9, p. 303, Eq. (6.137).
[20]F. D. Santos, Nucl. Phys. A **212**, 341 (1973), p. 358, Appendix 1.
[21]S. Agmon, *Lectures on Exponential Decay of Solutions of Second-Order Elliptic Equations: Bounds on Eigenfunctions of N-Body Schrödinger Operators* (Princeton University, Princeton, NJ, 1982), and references therein.
[22]B. Klahn and W. Bingel, Theoret. Chim. **44**, 9 (1977); **44**, 27 (1977).
[23]E. A. Hylleraas, Z. Phys. **54**, 347 (1929).
[24]H. Shull and P. O. Löwdin, J. Chem. Phys. **23**, 1392 (1955); P. O. Löwdin and H. Shull, Phys. Rev. **101**, 1730 (1956).
[25]E. Filter and E. O. Steinborn, J. Math. Phys. **21**, 2725 (1980).
[26]E. A. Hylleraas, Z. Phys. **48**, 469 (1928).
[27]M. Rotenberg, Adv. At. Mol. Phys. **6**, 233 (1970).
[28]H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Systems*, in *Handbuch der Physik* (Springer-Verlag, Berlin, 1957), Vol. 35, p. 101, Eq. (3.17).
[29]See Ref. 28, p. 110, Eq. (4.23).
[30]J. Čížek and E. R. Vrscay, Int. J. Quantum Chem. **21**, 27 (1982); and references therein.
[31]B. G. Adams, J. Čížek, and J. Paldus, Int. J. Quantum Chem. **21**, 153 (1982); and references therein.
[32]B. Podolsky and L. Pauling, Phys. Rev. **34**, 109 (1929).
[33]P. Kaijser and V. Smith, Jr., Adv. Quantum. Chem. **10**, 37 (1977), see pp. 50–52.
[34]E. J. Weniger, Ph.D. thesis, Universität Regensburg, 1982, p. 45, Eq. (3.3-35). See also Ref. 25, p. 2728, Eq. (3.17).
[35]E. J. Weniger and E. O. Steinborn, J. Chem. Phys. **78**, 6121 (1983), see p. 6123, Eq. (3.7).
[36]H. P. Trivedi and E. O. Steinborn, Phys. Rev. A **27**, 670 (1983).
[37]E. J. Weniger and E. O. Steinborn, J. Math. Phys. **24**, 2553 (1983).
[38]See Ref. 25, p. 2728, Eq. (3.15).
[39]See Ref. 27, p. 241, Eq. (26).
[40]See Ref. 28, p. 125, Eq. (8.8).
[41]M. J. Englefield, *Group Theory and the Coulomb Problem* (Wiley, New York, 1972), p. 69, Eqs. (5.5) and (5.6).
[42]See Ref. 9, p. 351, Eq. (7.4.69).
[43]M. Moshinsky, *The Harmonic Oscillator in Modern Physics: From Atoms to Quarks* (Gordon and Breach, New York, 1969).
[44]N. Ja. Vilenkin, *Special Functions and the Theory of Group Representations* (American Mathematical Society, Providence, RI, 1968), Chap. IX.
[45]J. Avery and Z-Y. Wen, Int. J. Quantum Chem. **22**, 717 (1982).
[46]P. A. Stone, Proc. Cambridge Philos. Soc. **52**, 424 (1956), see p. 427.
[47]L. C. Biedenharn, J. Math. Phys. **2**, 433 (1961), see p. 438, Eqs. (37) and (45).
[48]R. T. Sharp, J. Math. Phys. (Cambridge, MA) **47**, 359 (1968), see p. 362, Eq. (2.6).
[49]See Ref. 41, p. 5, Eq. (1.15).
[50]B. R. Judd, *Angular Momentum Theory for Diatomic Molecules* (Academic, New York, 1975), p. 32, Eq. (2.19).
[51]See Ref. 1, p. 147, Eq. (7*).
[52]T-I. Shibuya and C. Wulfman, Proc. R. Soc. London Ser. A **286**, 376 (1965), see p. 380, Eq. (21).
[53]See Ref. 52, p. 378, Eq. (12).
[54]C. E. Wulfman in *Group Theory and Its Applications II*, edited by E. M. Loebl (Academic Press, New York, 1971), p. 156, Eq. (3.40).
[55]See Ref. 1, p. 146, Eqs. (4) and (5).
[56]K. Ruedenberg, Theoret. Chim. Acta **7**, 359 (1967).
[57]H. J. Silverstone, J. Chem. Phys. **47**, 537 (1967).
[58]E. J. Weniger and E. O. Steinborn, Comput. Phys. Commun. **25**, 149 (1982), see p. 151, Eq. (3.1).
[59]See Ref. 25, p. 2729, Eqs. (4.4) and (4.7); p. 2733, Eq. (5.11).
[60]H. P. Trivedi and E. O. Steinborn, Phys. Rev. A **25**, 113 (1982).
[61]B. K. Novosadov, Int. J. Quantum Chem. **24**, 1 (1983).

Structural invariance of the Schrödinger equation and chronoprojective geometry

G. Burdet and M. Perrin
Centre de Physique Théorique, Centre National de la Recherche Scientifique-Luminy, Case 907-F, 13288 Marseille Cedex 9, France

(Received 11 January 1984; accepted for publication 22 June 1984)

We describe an extension of the chronoprojective geometry and show how its automorphisms are related to the invariance properties of the Schrödinger equation describing a quantum test particle in any Newton–Cartan structure.

I. INTRODUCTION

The chronoprojective geometry over four-dimensional manifolds has been described in Ref. 1(a). This geometry is well adapted to Galilean manifolds, it is a kind of nonrelativistic Weyl's geometry in the sense that it reconciles the notions of conformal equivalence over a Galilean manifold and projective equivalence between Newtonian connections. It has been shown in Ref. 1(b) that the chronoprojective geometry is the very geometry of the Newtonian cosmology since (i) the uniqueness condition of the chronoprojective Cartan connection coincides with the source equations (Poisson's equations) of the Newtonian potential; (ii) the form of the Ricci curvature tensor of an admissible Galilean connection is compatible with the one coming from the absence of rotational curvature; and (iii) the isotropy hypothesis of the Newtonian cosmology is expressed by the notion of chronoprojective flatness.

Moreover, the chronoprojective geometry is also relevant in classical mechanics since it explains various "accidental symmetries," for instance, the Kepler similitudes, the kinematical symmetries of the system of a charged particle in a Dirac magnetic monopole field, etc.

The chronoprojective geometry makes use of the so-called chronoprojective group which contains as a subgroup the Schrödinger group which arose by studying the invariance properties of the Schrödinger equation.[2] So, through the chronoprojective geometry, a geometrical status has been given to the Schrödinger group quite independently of its quantal origin.

It is known that only projective representations of the Schrödinger group are of physical interest or, what comes to the same, the true representations of an extended group which is the central extension of the Schrödinger group by an abelian phase group responsible for the emergence of the nonrelativistic mass. This extended Schrödinger group is contained in a noncentral extension of the chronoprojective group. By using this extended chronoprojective group an extended version of the chronoprojective geometry can be constructed which is described in this paper.

Moreover, we want to show that the extended chronoprojective geometry gives an explicit example of the structural invariance of the Schrödinger equation written upon any Galilean manifold,[3] and how the symmetry properties of such an equation are related to the automorphisms of the extended chronoprojective structure.

To carry out this program the paper is organized as follows: In Sec. II Newton–Cartan structures[4] are defined. Extended Galilean connections are defined in Sec. III, and the notion of extended chronoprojective equivalence of two extended Galilean connections is given in Sec. IV. Section V is devoted to the description of the structural invariance of the Schrödinger equation on a Newtonian space-time and the symmetry properties of this equation with respect to the chronoprojective equivalence notions are examined in Sec. VI. The technical points are treated in two appendices: the extended chronoprojective group and its relevant subgroups are described in Appendix A and the construction of extended chronoprojective Cartan connections is carried out in Appendix B.

II. NEWTON–CARTAN STRUCTURES

Definition 2.1: A Newton-Cartan space-time is a five-tuple $(V_4, \psi, \gamma, U, V)$ where the following hold.

(i) (V_4, ψ, γ) is a Galilean manifold, i.e., a four-dimensional C_∞-manifold endowed with a differential one-form ψ of class one and a positive semidefinite symmetric contravariant tensor field γ of degree 2, such that ker γ is generated by ψ.

(ii) U is an observer, i.e., a timelike unit (local) vector field

$$U \lrcorner \psi = 1. \tag{2.1}$$

(iii) V is the (gravitational) potential, i.e., a suitably differentiable function on V_4.

Let H denote the connected component of the full homogeneous Galilei group, i.e., the group of matrices

$$\begin{pmatrix} A & \overline{B} \\ 0 & 1 \end{pmatrix}, \quad \text{with} A \in O(3), \quad \overline{B} \in \mathbb{R}^3.$$

Definition 2.2: The bundle of Galilean frames $H(V_4)$ over a Galilean manifold (V_4, ψ, γ) is an H-structure of degree 1, i.e., a subbundle of the bundle of linear frames $P^1(V_4)$ corresponding to a reduction of GL(4,**R**) to H.

Definition 2.3: A Galilean connection is a linear connection reducible to a connection in $H(V_4)$ with respect to which ψ and γ are parallel, i.e.,

$$\nabla\psi = 0, \quad \nabla\gamma = 0, \tag{2.2}$$

∇ denoting the covariant derivative with respect to the Galilean connection (see Sec. IV).

 0022-2488/85/020292-08$02.50

Let us denote by Φ_H the curvature form of a Galilean manifold which can be written as $\Phi_H = \{\Phi^j_k, \Phi^j_0, j, k\in[1,3]\}$, where $\Phi := \{\Phi^j_k\}$ is $o(3)$-valued and $\overline{\Phi}_0 := \{\Phi^j_0\}$ is $\mathbf{R}^3$-valued, $\vartheta = \{\theta^0, \overline{\theta}\}$ the $\mathbf{R}^4$-valued canonical form of $P^1(V_4)$ restricted to $H(V_4)$.

Definition 2.4: A Newtonian connection is a torsionless Galilean connection which is such that ${}^t\overline{\theta}\wedge\overline{\Phi}_0 = 0$, or equivalently (and locally), such that the corresponding curvature tensor satisfies

$$R^{\rho}_{\ \lambda\mu\alpha}\gamma^{\alpha\nu} = R^{\ \nu}_{\mu\lambda\alpha}\gamma^{\alpha\rho}. \tag{2.3}$$

Proposition 2.5[4(b)]: Being given (V_4, ψ, γ, U) there is a unique torsionless Newtonian connection $\overset{U}{\Gamma}$ called the special Galilean connection associated with U, which is such that the observer is geodesic

$$\nabla_U U = 0, \tag{2.4}$$

and nonrotating

$$\gamma^{\rho[\alpha}\nabla_\rho U^{\beta]} = 0, \quad \alpha,\beta,\rho\in[0,3]. \tag{2.5}$$

Conversely,[5] for each Newtonian connection there exists (at least locally) such a geodesic and nonrotating observer.

To express $\overset{U}{\Gamma}$ it is convenient to introduce the so-called "associated covariant space metric" $\overset{U}{\gamma}$ which is defined by

$$\overset{U}{\gamma}_{\alpha\rho}\, U^\rho = 0 \tag{2.6a}$$

and

$$\overset{U}{\gamma}_{\alpha\rho}\,\gamma^{\rho\beta} = \delta^\beta_\alpha - \psi_\alpha U^\beta. \tag{2.6b}$$

Then $\overset{U}{\Gamma}$ is given by

$$\overset{U}{\Gamma}{}^{\alpha}_{\beta\gamma} = \overset{U}{\gamma}{}^{\alpha\rho}\{\partial_{(\beta}\gamma_{\gamma)\rho} - \tfrac{1}{2}\partial_\rho\gamma_{\beta\gamma}\} + U^\alpha\,\partial_{(\beta}\psi_{\gamma)}. \tag{2.7}$$

This special Galilean connection associated with the observer U supplies us with a reference for defining a Newtonian connection corresponding to a potential V.

Definition 2.6: The Newtonian connection $\overset{U,V}{\Gamma}$ deriving from the gravitational potential V and associated with the observer U is defined by its components with respect to a natural basis which are given by

$$\overset{U,V}{\Gamma}{}^{\alpha}_{\beta\gamma} = \overset{U}{\Gamma}{}^{\alpha}_{\beta\gamma} + \gamma^{\alpha\rho}\psi_{(\beta}K_{\gamma)\rho}, \tag{2.8}$$

where K is the two-form $\psi\wedge dV$.

Let us note that, K being a closed two-form, the above definition ensures that $\overset{U,V}{\Gamma}$ is a Newtonian connection.[4] Moreover, it is clear that $\overset{U,V}{\Gamma}$ accounts for the axiom according to which the Newtonian gravity comes from a potential,[6] since in a special adapted coordinate system ($U^\alpha = \delta^\alpha_0$ see Ref. 4(b)) the only nonzero components of $\overset{U,V}{\Gamma}$ are the $\overset{U,V}{\Gamma}{}^j_{00}$'s given by $\overset{U,V}{\Gamma}{}^j_{00} = \partial_j V$.

In the following we shall speak of a Newton–Cartan structure for a Newton–Cartan space-time endowed with the above defined gravitational connection $\overset{U,V}{\Gamma}$ referred to the observer U.

III. EXTENDED GALILEAN CONNECTIONS

Let G denote the Galilei group $G\approx\mathbf{R}^4 \circledS (O(3)\otimes\mathbf{R})$ and $\widetilde{G}$ its one-parameter central extension. A realization of $\widetilde{G}$ is given in Appendix A (Remark A.1) where an element of $\widetilde{G}$ is parametrized by the set

$$\{A\in O(3), M = (\overline{B},\overline{C}), \overline{B},\overline{C}\in\mathbf{R}^3, b\in\mathbf{R},$$
$$X = e\mathbb{1}_2 - \tfrac{1}{2}J\,{}^tMM, e\in\mathbf{R}\}.$$

Here $\widetilde{G}$ contains as a subgroup the direct product $\widetilde{H} = \mathbf{R}_e\otimes H$, where the homogeneous Galilei group is parametrized by $\{A\,\overline{B}\}$ as in Sec. II.

Let $\tilde{g}$ and $\tilde{h}$ denote the Lie algebras of $\widetilde{G}$ and $\widetilde{H}$, respectively, and let us denote by a a complementary subspace of $\tilde{h}$ with respect to $\tilde{g}$ such that $\tilde{g} = \tilde{h} + a$ as a vector space. In fact, a is isomorphic to $\mathbf{R}^4$, the four-dimensional abelian algebra.

The linear isotropy representation ρ of $\widetilde{H}$ defined by

$$\rho(g)X = \mathrm{Ad}(g)X\ (\mathrm{mod}\ \tilde{h}), \quad \text{for } g\in\widetilde{H}, \quad X\in a \tag{3.1}$$

is not faithful. Its kernel is isomorphic to $\mathbf{R}_e$; explicitly one gets

$$\rho(g) = \begin{pmatrix} A & \overline{B} \\ 0 & 1 \end{pmatrix},$$

where $g\in\widetilde{H}$ is parametrized by $\{A,\overline{B},e\}$ and its image is isomorphic to the homogeneous Galilei group H.

Let us now consider a principal $\widetilde{H}$-bundle $\widetilde{H}(V_4)$ over a four-dimensional manifold V_4, $\widetilde{H}(V_4)$ is not a subbundle of $P^1(V_4)$.

Definition 3.1: An extended Galilean connection is a Cartan connection in $\widetilde{H}(V_4)$ with respect to the extended Galilei group $\widetilde{G}$, that is to say, an extended Galilean connection is given by a $\tilde{g}$-valued one-form $\widetilde{\varphi}$ on $\widetilde{H}(V_4)$ which satisfies the following conditions.

(i) $\widetilde{\varphi}(X^*) = X$ for every $X\in\tilde{h}$, X^* denoting the fundamental vector field corresponding to X.

(ii) $(R_g)^*\widetilde{\varphi} = \mathrm{ad}(g^{-1})\widetilde{\varphi}$ for every $g\in\widetilde{H}$.

(iii) $\widetilde{\varphi}(Y)\neq 0$ for every nonzero vector Y of $\widetilde{H}(V_4)$.

Definition 3.2: The curvature two-form $\widetilde{\Phi}$ of an extended Galilean connection is defined by the following structure equation:

$$\widetilde{\Phi} = d\widetilde{\varphi} + \tfrac{1}{2}[\widetilde{\varphi},\widetilde{\varphi}]. \tag{3.2}$$

By using standard techniques[7] it can be shown that there does not exist a uniquely defined extended Galilean connection. Given $\widetilde{H}(V_4)$ there is an obvious surjective principal bundle homomorphism $\widetilde{H}(V_4)\rightarrow H(V_4)$. Let us then denote by φ_H the pullback to $\widetilde{H}(V_4)$ of a Galilean connection over $H(V_4)$ through this homomorphism and by ϕ_a the pullback of the canonical form ϑ. As φ_H is h-valued we can set

$$\varphi_H = \begin{pmatrix} \phi & \overline{\phi}_0 \\ 0 & 0 \end{pmatrix},$$

where ϕ is an $o(3)$-valued one-form and

$\bar{\phi}_0 = \{\phi^j_{0'}, j = 1,2,3\}$, a $\mathbb{R}^3$-valued one. Then an extended Galilean connection can be constructed by supplementing the lift φ_H of a Galilean connection with a $\mathbb{R}$-valued one-form ϕ_e and the affine components $\phi_a = \{\phi^\mu_{0'}, \mu = 0,1,2,3\}$ such that properties (i), (ii), (iii) of Def. 3.1 are satisfied. By taking into account the decomposition $\tilde{\varphi} = \{\phi_a, \varphi_H, \phi_e\} = \{\phi_a, \varphi_{\tilde{H}}\}$, property (ii) of Def. 3.1 can be written

$$(R_g)^*\phi_a = \rho(g^{-1})\phi_a, \tag{3.3a}$$

$$(R_g)^*\varphi_H = \rho(g^{-1})\varphi_H\,\rho(g), \tag{3.3b}$$

$$(R_g)^*\phi_e = \phi_e - {}^t\bar{B}\bar{\phi}_{0'} + \tfrac{1}{2}|\bar{B}|^2\phi^0_{0'}. \tag{3.3c}$$

An analogous decomposition is used for the two-form $\tilde{\Phi} = \{\Phi_a, \Phi_H, \Phi_e\}$, where Φ_H denotes the lift to $\tilde{H}(V_4)$ of the curvature form of a Galilean connection. Then $\Phi_a = \{\Phi^\mu_{0'}, \mu = 0, 1, 2, 3\}$ is called the torsion form and $\Phi_{\tilde{H}} := \{\Phi_H, \Phi_e\}$ the curvature form of the extended Galilean connection.

From Eq. (3.2) one gets explicitly

$$\Phi_a = d\phi_a + \varphi_H \wedge \phi_a, \tag{3.4a}$$

$$\Phi_H = d\varphi_H + \varphi_H \wedge \varphi_H, \tag{3.4b}$$

$$\Phi_e = d\phi_e + {}^t\bar{\phi}_0 \wedge \bar{\phi}_{0'}. \tag{3.4c}$$

Proposition 3.3: For a torsionless extended Galilean connection, Φ_e is basic.

Proof: If the connection is torsionless, i.e., if $\Phi_a = 0$, the exterior derivative of Φ_e becomes

$$d\Phi_e = {}^t\bar{\Phi}_0 \wedge \bar{\phi}_{0'}, \tag{3.5}$$

hence $d\Phi_e$ is horizontal. Q.E.D.

Being basic Φ_e can be written as the pullback to $\tilde{H}(V_4)$ of a two-form F on V_4: $\Phi_e = \pi^*F$. Then it is worth noticing that if the two-form F is closed one gets ${}^t\bar{\Phi}_0 \wedge \bar{\phi}_{0'} = 0$, which is the lift to $\tilde{H}(V_4)$ of the Newtonian condition on $H(V_4)$ given in Def. 2.4. Note also that Φ_e is kept invariant under the right action of $\tilde{H}$. Here ϕ_e is not basic but can be constructed over any coordinate neighborhood $\mathscr{U}$ in V_4 from a given one-form Γ_e on $\mathscr{U}$ by using standard techniques.

Let us denote by $\tilde{\Gamma}$ the local connection one-form on $\mathscr{U}$, with values in $\tilde{\mathfrak{g}}$, which corresponds to $\tilde{\varphi}$, then $\tilde{\Gamma} = \{\Gamma_a, \Gamma_H, \Gamma_e\}$. Here Γ_e is related to F through the local version of the structure equation (3.4c):

$$F = d\Gamma_e + {}^t\bar{\Gamma}_0 \wedge \bar{\Gamma}_{0'}. \tag{3.6}$$

Using the arbitrariness of Γ_e we shall impose the following constraint:

$$\Gamma_e(U) = V. \tag{3.7}$$

By taking into account Def. 2.1 and Eq. (2.6a) this condition leads one to look for Γ_e in the form

$$\Gamma_e = V\psi + \overset{U}{\gamma}(Y), \tag{3.8}$$

where Y is a spacelike vector field ($Y \lrcorner \psi = 0$). Hence the local one-form Γ_e, which is necessary to construct an extended Galilean connection from a Galilean one, is in one-to-one correspondence with the spacelike vector fields on V_4 through expression (3.8).

Covariant derivatives with respect to the connection $\tilde{\varphi}$ will be used for writing the Schrödinger operator over a Newton–Cartan structure (cf. Sec. V) and owing to condition (3.7) its expression does not depend on the choice of the vector field Y.

IV. EXTENDED CHRONOPROJECTIVE EQUIVALENCE

The chronoprojective equivalence of two extended Galilean connections corresponds to the following scheme: let $(\tilde{L}^0(V_4), \tilde{\omega})$ be an extended chronoprojective structure (cf. Appendix B) and $\tilde{L}^I(V_4)$ an extended conformal Galilean bundle such that $\tilde{L}^0(V_4)$ can be identified with the k-extension of $\tilde{L}^I(V_4) \overset{\iota,\iota'}{\hookrightarrow} \tilde{L}^0(V_4)$ such that $\iota^*\tilde{\omega} = \tilde{\varphi}$ and $\iota'^*\tilde{\omega} = \tilde{\varphi}'$ define two extended conformal Galilean connections. Such φ and φ' are said to belong to the same chronoprojective structure or to be chronoprojectively equivalent. This notion applies to extended Galilean connections if we suppose that there are also two isomorphic embeddings of $\tilde{H}(V_4)$ into $\tilde{L}^I(V_4)$ such that $\tilde{\varphi}$ and $\tilde{\varphi}'$ can be restricted to extended Galilean connection on each image, respectively. The set of extended Galilean connections arising in this way forms an equivalence class and two connections in this class are said to be chronoprojectively equivalent. Two such connections can be compared at the same point of $\tilde{L}^I(V_4)$. By using the right actions in the different bundles one gets

$$\phi'_e - \phi_e = 0, \tag{4.1a}$$

$$\phi'_H - \phi_H = \eta\mathbb{B}, \tag{4.1b}$$

where η is a $\mathbb{R}$-valued function on V_4 and $\mathbb{B}$ denotes the one-form matrix

$$\mathbb{B} = \begin{pmatrix} \phi^0_{0'}\mathbb{1}_3 & \bar{\phi}_0 \\ 0 & 2\phi^0_{0'} \end{pmatrix}. \tag{4.1c}$$

Obviously the extended chronoprojective geometry contains all the results of the chronoprojective geometry. Let us consider two Galilean manifolds (V_4, ψ, γ) and (V_4, ψ', γ'). They are said to be conformally equivalent iff

$$\psi' = \rho_t\psi \tag{4.2a}$$

and

$$\gamma' = \rho_s\gamma, \tag{4.2b}$$

where ρ_t and ρ_s are positive suitably differentiable functions on V_4. We recall that the most general equivalence relation between two torsionless Galilean connections, respectively, associated to two conformally equivalent Galilean manifolds involves 11 arbitrary functions,[1(a)] while only one function is necessary in the Riemannian case (owing to the presence of the Levi–Cività connection). But, by fixing $\rho_s\rho_t$ = constant function on V_4, these 11 functions can be reduced to only one function: this case corresponds to the chronoprojective equivalence which has been described in Ref. 1. Then one verifies that the function η is no more arbitrary but it is related to ρ_s and ρ_t through the following relation:

$$\eta\psi = \pi^*(d(\log\rho_s)) = \pi^*(d(-\log\rho_t)). \tag{4.3}$$

Proposition 4.1: At each point of $\tilde{L}^I(V_4)$, the one-forms of two chronoprojectively equivalent extended Galilean connections over conformally equivalent Galilean manifolds satisfy Eq. (4.1), where the function η is given by (4.3).

On each open set $\mathscr{U}$ of V_4 the expressions of the local connection one-form are related by

$$\Gamma'_e = (1/\rho_s\rho_t)\Gamma_e, \tag{4.4a}$$

$$\Gamma'^{\alpha}_{\beta\gamma} = \Gamma^{\alpha}_{\beta\gamma} - 2\eta\delta^{\alpha}_{(\beta}\psi_{\gamma)}, \tag{4.4b}$$

where the $\Gamma^{\alpha}_{\beta\gamma}$'s are the components of Γ_H, the local one-form which corresponds to φ_H.

Let us now consider two conformally equivalent Galilean manifolds (V_4, ψ, γ) and (V_4, ψ', γ') together with two corresponding observers U and U', respectively.

Proposition 4.2: The special Galilean connection $\overset{U'}{\Gamma}$ associated to U' will be said to be chronoprojectively equivalent to the special Galilean connection $\overset{U}{\Gamma}$ associated to U, that is to say, in components

$$\overset{U'}{\Gamma}{}^{\alpha}_{\beta\gamma} = \overset{U}{\Gamma}{}^{\alpha}_{\beta\gamma} + (\dot\rho_t/\rho_t)\delta^{\alpha}_{(\beta}\psi_{\gamma)}, \tag{4.5a}$$

iff

$$U' = (1/\rho_t)(U + \gamma(W)), \tag{4.5b}$$

where W is a closed one-form such that

$$U \lrcorner\, W + \tfrac{1}{2}\gamma(W,W) = c \quad \text{constant function on } V_4, \tag{4.5c}$$

$\gamma(W)$ denotes the contraction of $\gamma \otimes W$, locally $\gamma(W)^{\alpha} = \gamma^{\alpha\beta} W_{\beta}$, and it has been set $d\rho_t = \dot\rho_t \psi$.

Then U' is said to be an observer chronoprojectively equivalent to U.

Proof: In Ref. 1 it has been proved that $d\rho_t = \dot\rho_t\psi$ and $\rho_s\rho_t =$ constant function on V_4, which leads to $\dot\rho_t/\rho_t = \dot\rho_s/\rho_s$. The form (4.5b) of U' ensures its unitarity with respect to ψ', i.e., $U' \lrcorner\, \psi' = 1$, by taking into account the conformal equivalence of the two Galilean manifolds through $\psi' = \rho_t\psi$. The closure condition of W corresponds to the fact that U' is nonrotating [Eq. (2.5)] with respect to $\overset{U'}{\Gamma}$.

The condition (4.5c) comes from the fact that U' must be geodesic [Eq. (2.4)] with respect to $\overset{U'}{\Gamma}$.

The proof of the converse is obvious.

Corollary 4.3: Let U' be an observer chronoprojectively equivalent to U, then $\overset{U'}{\gamma}$ is related to $\overset{U}{\gamma}$ as follows:

$$\overset{U'}{\gamma} = (1/\rho_s)\Big(\overset{U}{\gamma} + 2c\psi\otimes\psi - 2W\otimes\psi\Big). \tag{4.6}$$

Proof: This relation directly derives from the definition of $\overset{U}{\gamma}$ given by Eq. (2.6) and Proposition 4.2.

Proposition 4.4: The Newtonian connection $\overset{U',V'}{\Gamma}$, deriving from the gravitational potential V' and associated with the observer U', will be said chronoprojectively equivalent to $\overset{U,V}{\Gamma}$, where U is chronoprojectively equivalent to U', iff

$$V' = (1/\rho_s\rho_t^2)(V + \mathscr{v}), \tag{4.7}$$

where $\mathscr{v}$ is a function on V_4 such that $d\mathscr{v} \wedge \psi = 0$.

Proof: It is a direct consequence of Eq. (2.8) which, by using the relations $d\rho_t = \dot\rho_t\psi$ and $d\rho_s = \dot\rho_s\psi$, implies

$$\overset{U',V'}{\Gamma} - \overset{U,V}{\Gamma} = \overset{U'}{\Gamma} - \overset{U}{\Gamma}. \tag{4.8}$$

Now let us examine how the chronoprojective equivalence interferes with condition (3.7). By using the expression (3.8) of Γ_e and by taking (4.7), (4.1a), and (4.6) into account, the equivalence relation (4.4a) on Γ_e leads to the following relation between the spacelike vector space Y associated to Γ_e and the one Y' associated to Γ'_e,

$$Y' = (1/\rho_t)Y, \tag{4.9}$$

together with the supplementary constraint

$$\mathscr{v} = W(Y). \tag{4.10}$$

Proposition 4.5: The compatibility between the chronoprojective equivalence and the condition (3.7) manifests itself on the spacelike vector field associated to Γ_e, through (3.8), by (4.9) and imposes condition (4.10) on the gravitational potential.

Corollary 4.6:

$$L_Y(W) = d\mathscr{v}.$$

The proof is a direct consequence of Proposition 4.5 and of the fact that W is a closed one-form.

V. STRUCTURAL INVARIANCE OF THE SCHRÖDINGER EQUATION

Lef F be a vector space on which $\widetilde{H}$ acts differentiably on the left by a representation ν and E the fiber bundle over the basis V_4 with standard fiber F associated to $\widetilde{H}(V_4)$ by ν.

Let us denote by $\Sigma(E)$ and $\Sigma(T^*(V_4)\otimes E)$ the spaces of sections of E and of the tensor product $T^*(V_4)\otimes E$, respectively. We recall that a connection on E is an operator $D{:}\Sigma(E)\to\Sigma(E\otimes T^*(V_4))$. Next we introduce the covariant differential with respect to a given vector field U, it is given by $\nabla_U\sigma = (D\sigma)(U)$, where $\sigma\in\Sigma(E)$, and is also called the covariant derivative of σ in the U direction. It is linear in U and σ, and satisfies for an arbitrary function f the following conditions:

$$\nabla_{fU}\sigma = f\nabla_U\sigma, \tag{5.1}$$

$$\nabla_U(f\sigma) = f\nabla_U\sigma + U(f)\cdot\sigma. \tag{5.2}$$

Let $\widetilde{\Gamma}$ denote the local connection one-form associated to $\widetilde{\varphi}$ for each differentiable local section in $\widetilde{H}(V_4)$; $\widetilde{\Gamma}$ is $\widetilde{\mathscr{g}}$-valued and let us denote by $\Gamma_{\widetilde{H}}$ the $\widetilde{\mathscr{h}}$-valued part of $\widetilde{\Gamma}$.

Let us consider U belonging to $T_x(V_4)$, a cross section σ of E defined in a neighborhood of x, and a curve $c(\tau)$ on V_4 such that $c(0) = x_0$ and tangent to U at x_0. Then the covariant derivative of σ at x_0 in the U direction is given by

$$(\nabla_U\sigma)(x_0) = \frac{d\sigma(x)}{d\tau}\bigg|_{\tau=0} + \nu_*(\Gamma_{\widetilde{H}}(U))\sigma(x_0), \tag{5.3}$$

where ν_* denotes the representation of $\widetilde{\mathscr{h}}$ deriving from ν. In local coordinates one sets

$$(\nabla_U\sigma)(x_0) = U^{\mu}(\nabla_{\mu}\sigma)(x_0). \tag{5.4}$$

We shall also need the local coordinates expression of the double covariant derivative

$$(\nabla_U(\nabla_U\sigma))(u_0) = U^{\mu}U^{\rho}(\nabla_{\mu\rho}\sigma)(u_0). \tag{5.5}$$

To describe the Schrödinger equation on a Newtonian space-time and with respect to a given extended Galilean connection we have to consider miscellaneous types of bundles associated to $\tilde{H}(V_4)$.

(i) From now on we denote by E a complex line bundle with a one-dimensional complex vector space as fiber F over V_4 ($F \equiv \mathbb{C}$). Here E becomes a bundle associated to $\tilde{H}(V_4)$ by making $\tilde{H}$ act on F via $\mathbb{R}_e$ only, and by choosing the representation ν in such a way that

$$\nu_*(\tilde{\Gamma}(U)) = (im/\hbar)\Gamma_e(U) \quad (m \in \mathbb{R}^+), \tag{5.6a}$$

which by using (3.7) becomes

$$\nu_*(\tilde{\Gamma}(U)) = (im/\hbar)V. \tag{5.6b}$$

Then for a cross section $\Psi \in \Sigma(E)$ one gets in local coordinates

$$(\nabla_\mu \Psi)(x) = ((\partial_\mu + (im/\hbar)V\psi_\mu)\Psi)(x). \tag{5.7}$$

(ii) The tangent bundle $T(V_4)$ over V_4 is naturally a bundle associated to $H(V_4)$ with standard fiber $\mathbb{R}^4$. We make it an associated fiber bundle to $\tilde{H}(V_4)$ by using the action of $\tilde{H}$ on $\mathbb{R}^4$ given by the linear isotropy representation ρ of $\tilde{H}$ defined in (2.1), so that

$$\rho_*(\tilde{\Gamma}(U)) = \Gamma_H(U). \tag{5.8}$$

Then for a cross section $X \in \Sigma(T(V_4))$ one gets in local coordinates

$$\nabla_\mu X^\nu = \partial_\mu X^\nu + \overset{U,V\,\nu}{\Gamma}_{\mu\lambda} X^\lambda. \tag{5.9}$$

In the same manner an element $g \in \tilde{H}$ acts on $T^*(V_4)$ through ${}^t\rho(g^{-1})$ so that for a cross section $\alpha \in \Sigma(T^*(V_4))$ one gets in local coordinates

$$\nabla_\mu \alpha_\nu = \partial_\mu \alpha_\nu - \overset{U,V\,\lambda}{\Gamma}_{\mu\nu} \alpha_\lambda. \tag{5.10}$$

Over any open subset of a Newtonian space-time the quantum state of a particle with mass m is described with respect to an observer U by a section Ψ of E [$\Psi \in \Sigma(E)$], the "wave function," which is supposed to satisfy the Schrödinger equation described in Ref. 3:

$$-(\hbar^2/2m)\gamma^{\mu\nu}(\nabla_\mu(\nabla_\nu \Psi))(x) = i\hbar(\tfrac{1}{2}\nabla_\lambda U^\lambda + \nabla_U)(\Psi)(x). \tag{5.11}$$

Let us note that the left-hand side of this equation takes a very simple form when covariant derivatives are expressed according to the above definitions, and by taking into account the parallel displacement of structures by the connection $\nabla\psi = 0$, $\nabla\gamma = 0$, and the property according to ψ generates the kernel of γ; one gets

$$(S\Psi)(x) := \left(\left((\hbar^2/2m)\gamma^{\mu\nu}(\partial_\mu\partial_\nu - \overset{U,V\,\lambda}{\Gamma}_{\mu\nu}\partial_\lambda) + i\hbar\left(\frac{1}{2}\nabla_\lambda U^\lambda + U^\lambda\partial_\lambda + \frac{im}{\hbar}V\right)\right)\Psi\right)(x). \tag{5.12}$$

Hence this equation has been explicitly written by using covariant derivatives relative to various relevant fiber bundles associated to $\tilde{H}(V_4)$. Now let us consider the following situation: let us suppose there exists a bundle $\tilde{L}^\circ(V_4)$ with a connection $\tilde{\omega}$ such that $\tilde{H}(V_4)$ is a subbundle of $\tilde{L}^\circ(V_4)$ and $\tilde{\omega}$ is reducible to $\tilde{\varphi}$. Moreover, let us suppose there exist two representations ρ' and ν' of $\tilde{L}^\circ$ into GL(4,$\mathbb{R}$) and $\mathbb{C}$ which coincide with the representations ρ and ν of $\tilde{H}(V_4)$, respectively, when they are restricted to $\tilde{H}$:

$$\rho'(g) = \rho(g|_{\tilde{H}}), \quad \nu'(g) = \nu(g|_{\tilde{H}}), \quad \forall g \in \tilde{L}^\circ.$$

Under these assumptions the complex vector bundle E can be considered as associated to $\tilde{L}^\circ(V_4)$ as well as the tangent bundle $T(V_4)$ and the miscellaneous covariant derivatives considered above remain unchanged. Consequently and under the above assumptions, the Schrödinger equation also remains unchanged and can be considered as associated to $\tilde{H}(V_4)$ as well as to any bigger bundle containing $\tilde{H}(V_4)$ and satisfying all the above convenient assumptions. With respect to this property we can speak of structural invariance of the Schrödinger equation.

An explicit example is provided by the extended chronoprojective geometry which is described in Appendix B. In Sec. IV we have seen that an equivalence notion is associated to this geometry, it is then interesting to study the behavior of the Schrödinger equation with respect to this equivalence notion. This is done in the following section.

VI. AUTOMORPHISMS OF AN EXTENDED CHRONOPROJECTIVE STRUCTURE AND INVARIANCE OF THE SCHRÖDINGER EQUATION

In the previous section the Schrödinger operator relative to a Newton–Cartan structure has been described. In Sec. IV a notion of chronoprojective equivalence for two Newton–Cartan structures has been defined. So it is relevant to examine the relationship which exists between the Schrödinger equations relative to two chronoprojectively equivalent Newton–Cartan structures.

Let $(V_4, \psi, \gamma, U, V, \overset{U,V}{\Gamma})$ and $(V_4, \psi', \gamma', U', V', \overset{U',V'}{\Gamma})$ be two chronoprojectively equivalent Newton-Cartan structures. By using Eqs. (2.2), (4.1a), (4.1b), (4.4a), (4.4b), (4.6), and (4.7), the Schrödinger operator on $(V_4, \psi', \gamma', U', V')$ can be written

$$\begin{aligned} S' = \rho_s(\hbar^2/2m)\gamma^{\alpha\beta}&(\partial_\alpha\partial_\beta - \overset{U,V\,\gamma}{\Gamma}_{\alpha\beta}\partial_\gamma) \\ &+ (i\hbar/\rho_t)\{\tfrac{1}{2}(\partial_\lambda U^\lambda + \overset{U,V}{\Gamma}{}^\lambda_{\lambda\rho} U^\rho) + U^\lambda\partial_\lambda \\ &+ (im/\hbar)(1/\rho_s\rho_t)(V + \omega) + \tfrac{3}{4}(\dot\rho_t/\rho_t) \\ &+ \tfrac{1}{2}\nabla_\lambda(\gamma(W)^\lambda) + (\gamma(W))^\lambda\partial_\lambda\}. \end{aligned} \tag{6.1}$$

From (6.1), it is clear that S' can be expressed in terms of the Schrödinger operator over $(V_4, \psi, \gamma, U, V)$ defined in (5.12) if the condition $\rho_s\rho_t = 1$ is fulfilled; then

$$(S'\Psi')(x) = \left\{\rho_s\left[S + i\hbar\left(\frac{3}{4}\frac{\dot\rho_t}{\rho_t} + \frac{im}{\hbar}\omega + \frac{1}{2}\nabla_\lambda(\gamma(W))^\lambda + (\gamma(W))^\lambda\partial_\lambda\right)\right]\Psi'\right\}(x). \tag{6.2}$$

Let us set

$$\Psi' = (\rho_t)^{-3/4}\exp(-(im/\hbar)f)\Psi, \tag{6.3}$$

where f is a differentiable function on V_4 such that

$$d_f = W + (\omega - c)\psi, \tag{6.4}$$

W, ω, and c having been defined in Sec. IV, Eqs. (4.5b), (4.5c), and (4.7). One can easily verify that if Ψ is a solution of the

Schrödinger equation (5.12), Ψ' is a solution of the Schrödinger equation (6.2). Thus the following proposition has been shown.

Proposition 6.1: Being given a solution Ψ of the Schrödinger equation $(S\Psi)(x) = 0$, $\forall x \in \mathscr{U}$, open set in V_4, relative to a Newton–Cartan structure $(V_4, \psi, \gamma, U, V, \overset{U,V}{\Gamma})$, one gets a solution of the Schrödinger equation $(S'\Psi')(x) = 0$, $\forall x \in \mathscr{U} \subset V_4$, relative to any chronoprojectively equivalent Newton–Cartan structure $(V_4, \psi', \gamma', U', V', \overset{U',V'}{\Gamma})$ such that $\rho_s \rho_t = 1$ by setting $\Psi' = \rho_t^{-3/4} \exp(-(im/\hbar)\mathscr{f})\Psi$, where $\mathscr{f}$ is a differentiable function on V_4 such that $d\mathscr{f} = W + (\mathscr{o} - c)\psi$.

Concerning the probabilistic interpretation of the Schrödinger equation it is worth noticing that the normalization of the "wave function" is conserved under the chronoprojective equivalence owing to the presence of the factor $\rho_t^{-3/4}$.

Now let us consider the automorphisms aut $(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ of the extended chronoprojective structure, i.e., the automorphisms of the bundle $\widetilde{L}^\circ(V_4)$ which map the extended chronoprojective connection on itself. Since $\widetilde{L}^\circ(V_4)$ is parallelizable owing to the existence of a Cartan connection, aut $(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ is a Lie group such that dim aut $(\widetilde{L}^\circ(V_4),\widetilde{\omega}) < \dim(\widetilde{L}^\circ(V_4)) = 14$. These automorphisms are in one-to-one correspondence with the automorphisms of $\widetilde{L}^I(V_4)$ which map an extended Galilean connection onto a chronoprojectively equivalent one. By looking at the projection on the basis, it is ascertained that these automorphisms correspond to chronoprojective Galilean transformations which ensure the chronoprojective equivalence on the local connection one-forms.

Every vector field X'' on $\widetilde{L}^\circ(V_4)$ generates a one-parameter local group of transformations. Let us suppose that such a local one-parameter group generated by X'' corresponds to an automorphism of the extended chronoprojective connection, i.e.,

$$L_X \cdot \widetilde{\omega} = 0. \tag{6.5}$$

According to the above-mentioned property the set of vector fields X'' satisfying (6.5) generates a Lie algebra $\mathscr{aut}(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ of dimension at most equal to 14. If the maximal dimension is reached $\mathscr{aut}(\widetilde{L}^\circ(V_4), \widetilde{\omega})$ is the extended chronoprojective algebra.

Another realization of $\mathscr{aut}(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ can be also obtained by considering the set of vector fields X' on $\widetilde{L}^I(V_4)$, which are such that

$$L_{X'} \bar{\phi}_{0'} = -\tfrac{1}{2}\epsilon_s \bar{\phi}_{0'}, \tag{6.6a}$$

$$L_{X'} \phi^{0}_{0'} = \epsilon_t \phi^{0}_{0'}, \tag{6.6b}$$

$$L_{X'} \phi = 0, \tag{6.6c}$$

$$L_{X'} \bar{\phi}_0 = -(\tfrac{1}{2}\epsilon_s + \epsilon_t)\bar{\phi}_0 + \eta(1 - (\tfrac{1}{2}\epsilon_s + \epsilon_t))\bar{\phi}_{0'}, \tag{6.6d}$$

$$L_{X'} \phi_e = (\epsilon_s + \epsilon_t)\phi_e, \tag{6.6e}$$

where ϵ_s and ϵ_t are two constant functions on the fibers of $\widetilde{L}^I(V_4)$. These expressions are a direct consequence of (4.1a) and (4.1b). By inspecting Eq. (6.6) one can verify that the component of X' which corresponds to the parameter of the extension arises only in (6.6e). Consequently, the projection X of X' onto V_4 will depend on one parameter less than X', and the set of X obtained by projecting the set of X' satisfying (6.6) will generate at most a 13–dimensional Lie algebra.

By looking at the examples one sees that it is the "extension" component of the algebra $\mathscr{aut}(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ which is not realized. The action of diffeomorphisms being natural, it is clear from Proposition 6.1 that the structural invariance of the Schrödinger equation with respect to the chronoprojective geometry leads to the following.

Proposition 6.2: The automorphism group of an extended chronoprojective structure restricted by the condition $\rho_s \rho_t = 1$ is an invariance group for any Schrödinger equation relative to a Newton–Cartan structure subordinate to this extended chronoprojective structure.

Let us remark that the condition $\rho_s \rho_t = 1$ excludes from the automorphisms group a dilation, so that the invariance group of the Schrödinger equation will be at most 13-dimensional. When the maximal dimension is reached one gets the subgroup of the extended chronoprojective group which is known in the literature as the Schrödinger group,[2] to which the above work furnishes a geometrical support.

We have seen that $\mathscr{aut}(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ cannot be realized by vector fields over V_4. Let us introduce the natural extension of the Lie algebra of vector fields by suitably differentiable real functions over V_4 with the Lie bracket

$$[(X_1,\mathscr{f}_1),(X_2,\mathscr{f}_2)] = [X_1,X_2] + X_1(\mathscr{f}_2) - X_2(\mathscr{f}_1).$$

Then the symmetry algebra of the Schrödinger operator S, subalgebra of $\mathscr{aut}(\widetilde{L}^\circ(V_4),\widetilde{\omega})$ can be realized by looking for the family of functions $\mathscr{f}$ which satisfy the condition

$$[S,(X,\mathscr{f})] = \epsilon_t S, \tag{6.7}$$

where X is the chronoprojective vector field of the subjacent Newton–Cartan structure restricted by the condition $\epsilon_s + \epsilon_t = 0$ which excludes a "nonphysical" dilation.

By way of illustration let us consider the two following cases.

(i) The isotropic empty space-time with a cosmological constant is a flat Newton–Cartan structure where the automorphism group reaches its maximal dimension. The corresponding chronoprojective vector field has been given in Ref. 1(b).

A quantum test particle in such a space-time obeys the Schrödinger equation with a (anti)harmonic potential. Then it is easy to determine the family of functions $\mathscr{f}$ satisfying (6.7) which, together with X, gives a realization of the extended Schrödinger algebra as a symmetry algebra for the corresponding Schrödinger operator.

(ii) The Newtonian field of a massive point particle is a Newton–Cartan structure for which the automorphism group reduces to $O(3) \otimes \mathbb{R}^2$ (see Ref. 8). The related Schrödinger equation describes the quantum Kepler problem and the solutions of (6.7) are given by $\mathscr{f} =$ const, so the symmetry algebra is just a trivial extension of $\mathscr{o}(3) \oplus \mathbb{R}$.

ACKNOWLEDGMENTS

We thank C. Duval and H. P. Kunzle for having communicated their results prior to publication.

APPENDIX A: THE EXTENDED CHRONOPROJECTIVE GROUP $\widetilde{\mathrm{Chr}}_3$

The chronoprojective group Chr_3 has been defined in Ref. 1 [where it is denoted by $O^2(3)$]. It possesses a one-parameter noncentral extension denoted $\widetilde{\mathrm{Chr}}_3$ which can be written as a semidirect product of $O(3)\otimes \mathrm{GL}(2,\mathbb{R})$ acting on a covering of the Weyl group. Here $\widetilde{\mathrm{Chr}}_3$ can be defined as the subgroup of $\mathrm{Gl}(7,\mathbb{R})$ generated by the following matrix product which makes clear the semidirect product structure:

$$g=\begin{pmatrix}\mathbf{1}_2 & -J^tM & X\\ \cdot & \mathbf{1}_3 & M\\ \cdot & \cdot & \mathbf{1}_2\end{pmatrix}\begin{pmatrix}-J^tL^{-1}J & \cdot & \cdot\\ \cdot & A & \cdot\\ \cdot & \cdot & L\end{pmatrix}. \tag{A1}$$

where

$M=(\overline{B}\,\overline{C})$ with $\overline{B},\overline{C}\in\mathbb{R}^3$,

$X=e\mathbf{1}_2-\frac{1}{2}J^tMM,\quad e\in\mathbb{R}$,

$A\in O(3)$,

$$J=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\text{ and }\quad L=\begin{pmatrix}a & b\\ c & d\end{pmatrix}\in \mathrm{GL}(2,\mathbb{R}),$$

t as a front superscript denotes the transposition.
The group law $g''=gg'$ is given by

$$A''=AA', \tag{A2a}$$

$$L''=LL', \tag{A2b}$$

$$M''=M+AM'L^{-1}, \tag{A2c}$$

$$e''\mathbf{1}_2=(e+e'(\det L)^{-1})\mathbf{1}_2 + \tfrac{1}{2}J({}^tL^{-1}{}^tM'{}^tAM-{}^tMAM'L^{-1}), \tag{A2d}$$

and (A2d) makes appear clearly the noncentral character of the extension. A central extension is obtained if $\det L=1$ and the corresponding subgroup is known in the literature[2] as the central extension $\widetilde{\mathrm{Sch}}_3$ of the so-called Schrödinger group.

Remark A.1: Let us note that by setting $a=d=1$, $c=0$ into $L\in\mathrm{GL}(2,\mathbb{R})$ another interesting subgroup of $\widetilde{\mathrm{Chr}}_3$ is obtained, namely the central extension $\widetilde{G}$ of the Galilei group G [$\widetilde{G}$ is called the Bargmann group and is denoted by B in Ref. 3(b)].

Now let us consider the subgroup of $\widetilde{\mathrm{Chr}}_3$ obtained by setting $\overline{C}=0$, $b=0$ $(ad\neq 0)$ and denoted by $\widetilde{L}^\circ$. This subgroup can be written as a semidirect product $\widetilde{L}^\circ=\mathbb{R}_e\circledS L^\circ$, where $\mathbb{R}_e$ corresponds to the one-parameter extension in the above notations. Here L° is defined as the group of matrices of $\mathrm{GL}(5,\mathbb{R})$ of the form

$$\begin{pmatrix}A & \overline{B} & \overline{0}\\ \cdot & a & \cdot\\ \cdot & c & d\end{pmatrix},$$

where $A\in O(3)$, $\overline{B}\in\mathbb{R}^3$, $a,c,d\in\mathbb{R}(ad\neq 0)$, and can be written as $L^\circ=\mathbb{R}^3\circledS(O(3)\otimes\mathbb{R}\otimes S_2)$, S_2 denoting the two-dimensional solvable group [see Ref. 1(b)]. Let $\widetilde{\mathscr{chr}}_3$ and $\widetilde{\ell}^\circ$ be the Lie algebras of $\widetilde{\mathrm{Chr}}_3$ and $\widetilde{L}^\circ$, respectively, and let us denote by α a complementary subspace of $\widetilde{\ell}^\circ$ with respect to $\widetilde{\mathscr{chr}}_3$ such that, as a vector space

$$\widetilde{\mathscr{chr}}_3=\widetilde{\ell}^\circ+\alpha. \tag{A3}$$

In fact $\alpha=\mathbb{R}^4$ is a four-dimensional abelian algebra. The linear isotropy representation ρ of $\widetilde{L}^\circ$, defined by

$$\rho(m)X=\mathrm{Ad}(m)X\,(\mathrm{mod}\ \widetilde{\ell}^\circ)\text{ for } m\in\widetilde{L}^\circ\text{ and } X\in\alpha$$

is not faithful. Its kernel N is isomorphic to $\mathbb{R}^2$; explicitly one gets

$$\rho(m)=\rho\left(A,\begin{pmatrix}a & \cdot\\ c & d\end{pmatrix},\overline{B},e\right)=\frac{1}{d}\begin{pmatrix}A & a\overline{B}\\ \cdot & a\end{pmatrix}. \tag{A4}$$

The image $L^I=\rho(\widetilde{L}^\circ)\subset\mathrm{GL}(4,\mathbb{R})$ is isomorphic to $\mathbb{R}^3\circledS(\mathrm{CO}(3)\otimes\mathbb{R})$. In Ref. 1(b) L^I has been called the conformal homogeneous Galilei group and it can also be written as the semidirect product $H\circledS(\dot{\mathbb{R}}_s\otimes\dot{\mathbb{R}}_t)$, where $\dot{\mathbb{R}}_s$ and $\dot{\mathbb{R}}_t$ denote two distinct dilatation subgroups defined as follows: let

$$s=\begin{pmatrix}\widetilde{A} & \overline{\widetilde{B}}\\ \cdot & \widetilde{d}\end{pmatrix}\in L^I,$$

where $\widetilde{A}\in\mathrm{CO}(3)$, $\overline{\widetilde{B}}\in\mathbb{R}^3$, $\widetilde{d}\in\dot{R}$, then $\dot{\mathbb{R}}_s$ is parametrized by $(\det\widetilde{A})^{-1/3}$ and $\dot{\mathbb{R}}_t$ is parametrized by $\widetilde{d}$. It is worth noticing that $\widetilde{L}^\circ$ can also be written as the semidirect product $N\circledS L^I$ with the group law

$$(s,c,e)(s',c',e')=(ss',c\widetilde{d}'+c',e\widetilde{d}'\,|\det\widetilde{A}'|^{-2/3}+e') \tag{A5}$$

where c and e parametrize $N=\mathbb{R}^2$. This group law corresponds to the following choice of the injective homomorphism $\mathscr{K}:L^I\to\widetilde{L}^\circ$:

$$\mathscr{K}\left(\begin{pmatrix}\widetilde{A} & \overline{\widetilde{B}}\\ \cdot & \widetilde{d}\end{pmatrix}\right)=\left\{A=\frac{\widetilde{A}}{|\det\widetilde{A}|^{1/3}},\ L=\frac{1}{|\det\widetilde{A}|^{1/3}}\begin{pmatrix}\widetilde{d} & \cdot\\ \cdot & 1\end{pmatrix},\ M=\left(\frac{\overline{\widetilde{B}}}{\widetilde{d}}\overline{0}\right),X=\begin{pmatrix}\cdot & \cdot\\ \frac{1}{2}(|\overline{\widetilde{B}}|^2/\widetilde{d}^2) & \cdot\end{pmatrix}\right\}, \tag{A6}$$

with the notations introduced in (A1).

Correspondingly the Lie algebra $\widetilde{\ell}^\circ$ can be written as a semidirect sum

$$\widetilde{\ell}^\circ=n\,\square\,\ell^I, \tag{A7}$$

where ℓ^I denotes the Lie algebra of L^I and n is the two-dimensional abelian algebra.

We shall also introduce a group $\widetilde{L}^I=\mathbb{R}_e\circledS L^I$, the extended conformal homogeneous Galilei group. Here, $\widetilde{L}^I$ is a subgroup of $\widetilde{L}^\circ$ and it is defined by $\widetilde{L}^I=\{g\in\widetilde{L}^\circ|c=0\}$. Note also that $\widetilde{L}^I=\widetilde{H}\circledS(\dot{\mathbb{R}}_s\otimes\dot{\mathbb{R}}_t)$. Also $\widetilde{L}^\circ$ can be written as a semidirect product $\mathbb{R}\circledS\widetilde{L}^I$ corresponding to an injective homomorphism $\widetilde{\mathscr{K}}$ which results from $\mathscr{K}$ in an obvious way.

APPENDIX B: EXTENDED CHRONOPROJECTIVE CARTAN CONNECTIONS

We refer to Ref. 1(a) for the general definition of a Cartan connection, classical references about the subject are also given in Ref. 1(b). Let $\widetilde{L}^\circ$ be the subgroup of the extended chronoprojective group defined in Appendix A.

Definition B.1: Let $\widetilde{L}^\circ(V_4)$ be a principal $\widetilde{L}^\circ$-bundle over a four-dimensional manifold V_4. An extended chronoprojective connection is a Cartan connection in $\widetilde{L}^\circ(V_4)$ with respect to the extended chronoprojective group.

Hence an extended chronoprojective connection form $\widetilde{\omega}$ is $c\mathscr{h}\mathscr{r}_3$-valued and can be decomposed as $\widetilde{\omega} = \{ W_\alpha, W_I, W_0^{0'}, W_e \}$, where (i) W_α is α-valued according to the decomposition (A3), i.e., $W_\alpha = \{ W_{0'}^{\mu}, \mu \in [0.3] \}$, (ii) W_I is ℓ^I-valued and can be written as

$$W_I = \begin{pmatrix} \widetilde{W} & \overline{W}_0 \\ \cdot & W_D \end{pmatrix}, \tag{B1}$$

where $\widetilde{W} = \{ W, W_s \}$ is $co(3)$-valued [W being $o(3)$-valued], $\overline{W}_0$ is $\mathbb{R}^3$-valued, W_D is $\mathbb{R}$-valued, and (iii) $\{ W_0^{0'}, W_e \}$ is n-valued (n denoting the Lie algebra of the kernel N of the linear isotropy representation of $\widetilde{L}^\circ$).

The set $\{ W_\alpha, W_I, W_0^{0'} \}$ is $c\mathscr{h}\mathscr{r}_3$-valued and can be identified (see below) with the chronoprojective connection ω studied in Ref. 1(a). So one can write $\widetilde{\omega} = \{ \omega, \omega_e \}$.

Proposition B.2: Under the right action of $m \in \widetilde{L}^\circ$ on $\widetilde{L}^\circ(V_4)$, the extended chronoprojective connection transforms according to

$$R_m^*(W_\alpha) = \rho(m^{-1}) W_\alpha, \tag{B2a}$$

$$R_m^*(W_I) = \rho(m^{-1})(W_I + (c/a)\mathbb{B})\rho(m), \tag{B2b}$$

$$R_m^*(W_0^{0'}) = \frac{a}{d} W_0^{0'} - \frac{c^2}{ad} W_{0'}^0 - \frac{c}{d} W_D, \tag{B2c}$$

$$R_m^*(W_e) = ad\,(W_e - {}^t\overline{B}\,\overline{W}_{0'} + \tfrac{1}{2}|\overline{B}|^2 W_{0'}^0 - e W_s), \tag{B2d}$$

where $\mathbb{B}$ denotes the following one-form matrix:

$$\mathbb{B} = \begin{pmatrix} W_{0'}^0 \, 1_3 & \overline{W}_{0'} \\ \cdot & 2W_{0'}^0 \end{pmatrix}. \tag{B3}$$

Proposition B.3: The components of the two-forms Ω of the extended chronoprojective connection are given by

$$\Omega_\alpha = dW_\alpha + W_I \wedge W_\alpha, \tag{B4a}$$

$$\Omega_I = dW_I + W_I \wedge W_I - W_0^{0'} \wedge \mathbb{B}, \tag{B4b}$$

$$\Omega_0^{0'} = dW_0^{0'} + W_0^{0'} \wedge W_D, \tag{B4c}$$

$$\Omega_e = dW_e + {}^t\overline{W}_0 \wedge \overline{W}_{0'} - W_s \wedge W_e. \tag{B4d}$$

Here Ω_α is called the torsion form and $\{ \Omega_I, \Omega_0^{0'}, \Omega_e \}$ the curvature form of the extended chronoprojective connection.

By using standard techniques it can be shown that a uniquely defined extended chronoprojective connection can be constructed from a given set $\mathfrak{z} = \{ W_\alpha, W_I, W_e \}$ of 13 differential one-forms whose values in each point are linearly independent. The properties of the curvature of the uniquely defined extended chronoprojective connection are identical to the ones of the chronoprojective connection described in Ref. 1(a) (Property 2.3 of this reference).

But the difference between the extended chronoprojective geometry and the chronoprojective one lies in the following fact: $\widetilde{L}^\circ$ cannot be realized as a subgroup of $G^2(4)$, the structure group of the bundle of second-order frames $P^2(V_4)$ over V_4.

As a consequence it does not exist extended chronoprojective Cartan structures, i.e., $\widetilde{L}^\circ(V_4)$ cannot be realized as a subbundle of $P^2(V_4)$ and there is no canonical realization of the set $\mathfrak{z}$.

Finally it is worth noticing that Eqs. (B2) and (B4) can be restricted to extended Galilean connections, this is done in Sec. III. However we shall use the natural chronoprojective structure $(L^\circ(V_4), \omega)$ over V_4 [cf. Sec. 3 §A in Ref. 1(a)] to speak of extended chronoprojective structure in the following sense.

As it has been noted in Appendix A, $\widetilde{L}^\circ$ can be written as a semidirect product $\widetilde{L}^\circ = R_e \circledS L^\circ$ which corresponds to the following exact sequence:

$$1 \to \mathbb{R} \xrightarrow{i} \widetilde{L}^\circ \xrightarrow{\mathscr{h}} L^\circ \to 1. \tag{B5}$$

We can then introduce a $\mathscr{h}$-lifting of $L^\circ(V_4)$, which is a principal $\widetilde{L}^\circ$ bundle over V_4 together with a $\mathscr{h}$-equivariant principal bundle morphism $\widetilde{\mathscr{h}}: \widetilde{L}^\circ(V_4) \to L^\circ(V_4)$.

The natural chronoprojective connection ω over $L^\circ(V_4)$ can then be lifted to $\widetilde{L}^\circ(V_4)$ and is also denoted by ω. Then we have just to choose a one-form W_e *a priori* in order to define a Cartan connection $\widetilde{\omega} = \{ \omega, W_e \}$ over $\widetilde{L}^\circ(V_4)$, $\widetilde{\omega}$ will be called an extended chronoprojective connection and $(\widetilde{L}^\circ(V_4), \widetilde{\omega})$ an extended chronoprojective structure over V_4 (*"par abus de langage"*).

[1](a) G. Burdet, C. Duval, and M. Perrin, Publ. RIMS **19**, 813 (1983); and (b) G. Burdet, C. Duval, and M. Perrin, J. Math. Phys. **24**, 1752 (1983).

[2]U. Niederer, Helv. Phys. Acta **45**, 802 (1972).

[3](a) K. Kuchar, Phys. Rev. D **22**, 1285 (1980); and (b) C. Duval and H. P. Kunzle, Gen. Relativ. Gravit. **16**, 333 (1984).

[4](a) H. P. Kunzle, Ann. Inst. H. Poincarè A **17**, 337 (1972); and (b) H. P. Kunzle, Gen. Relativ. Gravit. **7**, 445 (1976).

[5]H. D. Dombrowski and K. Horneffer, Math. Z. **86**, 291 (1964).

[6]E. Cartan, Ann. Sci. Ecole Norm. Sup. **3**, 1 (1924).

[7](a) S. Kobayashi, Can. J. Math. **8**, 145 (1956); and (b) S. Kobayashi, Ann. Mat. Pura Appl. **43**, 119 (1957).

[8]C. Duval, Thèse de Doctorat d'Etat, Marseille, 1982.

Maxwell's equations and the bundle of null directions on Minkowski space

Steven L. Kent
Department of Mathematics and Computer Science, Youngstown State University, Youngstown, Ohio 44555
Carlos N. Kozameh and E. T. Newman
Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

(Received 8 May 1984; accepted for publication 20 July 1984)

We reexpress the Maxwell field as a cross section of a line bundle over $M\times S^2$, the six-dimensional space of null directions on Minkowski space. Maxwell's equations then become a pair of linear equations for a Herz-like scalar on $M\times S^2$. We obtain a deeper understanding of the simple, yet nontrivial relationship between the self-dual and the anti-self-dual parts of a real Maxwell field. Our results are then applied to study solutions which are globally regular (on $M\times S^2$) namely, the pure radiation solutions, as well as solutions associated with discrete sources (the Lienard–Wiechert fields).

I. INTRODUCTION

The present work is part of a long-range program[1,2] to study the classical Maxwell, Yang–Mills, and gravitational fields from what we believe is a novel point of view. In this paper we will confine the discussion to the empty-space Maxwell equations with point sources.

The basic idea is to reexpress the Maxwell field not as a tensor field on Minkowski space M, but as a cross section of a *line* bundle over the six-dimensional space of null directions on Minkowski space ($M\times S^2$), i.e., the Maxwell field will be expressed as a single scalar function of two angles and points of Minkowski space. The Maxwell equations will become simply a pair of linear equations for this scalar on $M\times S^2$. We will refer to this scalar as the Maxwell scalar.

We will show that the Maxwell scalar can be expressed in terms of a "Herz-like" scalar with a remarkable simplification in the field equations. One of the results of this work is a deeper understanding of the dynamical relationship of the self-dual with the anti-self-dual parts of a real Maxwell field. Though this relationship is simple here, due to the linearity of the field equations, it is not trivial. The interesting cases, however, are the nonlinear Yang-Mills and Einstein theories, where one can see, in this formulation, the interactions of the self- and anti-self-dual parts of the field. These cases will be discussed in future papers.

In Sec. II we will describe our notation which is then used in Sec. III for the reformulation of the Maxwell equations. In Sec. IV we will discuss the subclass of solutions which are globally (on $M\times S^2$) regular, i.e., the retarded minus advanced fields, while in Sec. V we will discuss the solutions associated with point sources, namely the Lienard–Wiechert fields.

II. NOTATION

In Minkowski space M with coordinates x^a we introduce a unit timelike vector t^a which is parallelly propagated throughout M and a null vector field l_a which is normalized by $l_a t^a = 1/\sqrt{2}$ and is parametrized by two coordinates on the sphere most conveniently chosen as complex stereographic coordinates $(\zeta,\bar\zeta)$. At a given point x^a, as $(\zeta,\bar\zeta)$ move over the sphere, $l_a(\zeta,\bar\zeta)$ moves over the light-cone. For fixed $(\zeta,\bar\zeta)$, l_a is parallelly propagated through M. A useful representation of l_a is

$$l_a = [1/\sqrt{2}(1+\zeta\bar\zeta)](1+\zeta\bar\zeta, \zeta+\bar\zeta, i(\bar\zeta-\zeta), -1+\zeta\bar\zeta). \tag{2.1}$$

In addition to l_a we will need the following fields [also parametrized by $(\zeta,\bar\zeta)$]:

$$m_a = \eth l_a, \tag{2.2}$$

$$\bar m_a = \bar\eth l_a, \tag{2.3}$$

$$n_a = l_a + \eth\bar\eth l_a, \tag{2.4}$$

where $\eth(\bar\eth)$ is[2] essentially $\partial/\partial\zeta$ $(\partial/\partial\bar\zeta)$. The set $l_a, n_a, m_a, \bar m_a$ is closed under the $\eth$ and $\bar\eth$ action since

$$\eth m_a = \bar\eth \bar m_a = 0, \quad \eth n_a = -m_a, \quad \text{and} \quad \bar\eth n_a = -\bar m_a. \tag{2.5}$$

For arbitrary, but fixed $(\zeta,\bar\zeta)$ $l_a, m_a, \bar m_a, n_a$ form a null tetrad system, with all scalar products vanishing except

$$l_a n^a = -m_a \bar m^a = 1. \tag{2.6}$$

One easily sees that the Minkowski metric is

$$\eta_{ab} = 2l_{(a}n_{b)} - 2m_{(a}\bar m_{b)} \tag{2.7}$$

for any value of $(\zeta,\bar\zeta)$.

We will need the following $(\zeta,\bar\zeta)$-dependent directional derivatives:

$$D \equiv l^a\nabla_a, \quad \Delta = n^a\nabla_a, \quad \delta = m^a\nabla_a, \quad \bar\delta = \bar m^a\nabla_a. \tag{2.8}$$

Note that if (2.7) is multiplied by x^b we have

$$x^a = l^a n + n^a l - m^a\bar m - \bar m^a m, \tag{2.9}$$

where

$$l = x^a l_a, \quad n = x^a n_a, \quad m = x^a m_a, \quad \bar m = x^a \bar m_a. \tag{2.10}$$

From (2.9) and (2.10) we see that functions of x^a and $(\zeta,\bar\zeta)$ can also be thought of as functions as $l, n, m, \bar m$, and $\zeta,\bar\zeta$. [This can be thought of as a $(\zeta,\bar\zeta)$-dependent coordinate transformation from x_a to $l, n, m, \bar m$.] From this point of view we have

$$D = \frac{\partial}{\partial n}, \quad \Delta = \frac{\partial}{\partial l}, \quad \delta = -\frac{\partial}{\partial \bar m}, \quad \bar\delta = -\frac{\partial}{\partial m}. \tag{2.11}$$

 0022-2488/85/020300-06$02.50

Though ∇_a and ð (and $\bar{ð}$) commute, it is easy to see that ð(and $\bar{ð}$) do not commute with D, δ, $\bar{\delta}$, Δ. In fact, we have, from (2.1), (2.5), and (2.8), that

$$\eth D - D\eth = \delta, \quad \eth\Delta - \Delta\eth = -\delta,$$
$$\eth\delta - \delta\eth = 0, \quad \eth\bar{\delta} - \bar{\delta}\eth = \Delta - D, \tag{2.12}$$

and the conjugates.

Using (2.1) we can construct the useful $(\zeta,\bar{\zeta})$-dependent bivector fields,[3]

anti-self-dual	self-dual	
$l_{[a}m_{b]}$	$l_{[a}\bar{m}_{b]}$	(2.13)
$l_{[a}n_{b]} + \bar{m}_{[a}m_{b]}$ $= \bar{\eth} l_{[a}m_{b]}$	$l_{[a}n_{b]} - \bar{m}_{[a}m_{b]}$ $= \eth l_{[a}\bar{m}_{b]}$	(2.14)
$\bar{m}_{[a}n_{b]} = \frac{1}{2}\bar{\eth}^2 l_{[a}m_{b]}$	$m_{[a}n_{b]} = \frac{1}{2}\eth^2 l_{[a}\bar{m}_{b]}$	(2.15)

with the important identities

$$\eth l_{[a}m_{b]} = \bar{\eth} l_{[a}\bar{m}_{b]} = 0. \tag{2.16}$$

Note that for fixed $(\zeta,\bar{\zeta})$, these anti-self-dual (self-dual) bivectors form a basis set for anti-self-dual (self-dual) bivectors, i.e., for an arbitrary bivector field F_{ab} (x^a) we can define its (anti-self-dual) components by

$$\phi_0(x^a,\zeta,\bar{\zeta}) = F_{ab}l^a m^b, \tag{2.17}$$

$$\phi_1 = \tfrac{1}{2}F_{ab}(l^a n^b + \bar{m}^a m^b) = \tfrac{1}{2}\bar{\eth}\phi_0, \tag{2.18}$$

$$\phi_2 = F_{ab}\bar{m}^a n^b = \tfrac{1}{2}\bar{\eth}^2\phi_0. \tag{2.19}$$

We also have, from (2.16), that

$$\eth\phi_0 = 0. \tag{2.20}$$

We thus see that the complex function $\phi_0(x^a,\zeta,\bar{\zeta})$ with $(\zeta,\bar{\zeta})$ behavior given by (2.20) on the six-dimensional space $M \times S^2$ carries the full information of the bivector field F_{ab}. In the next section we will translate the Maxwell equations for F_{ab} to a pair of differential equations for ϕ_0.

III. THE VACUUM MAXWELL EQUATIONS

One can now easily derive the equation for ϕ_0, equivalent to the Maxwell equations, by either beginning with the Maxwell equations in spin-coefficient notation[4] or from first principles. For completeness we will do the latter.

We write the two sets of equations $\nabla_a F^{ab} = \nabla_a F^{*ab} = 0$ as

$$\nabla_a(F^{ab} + iF^{*ab}) = 0, \tag{3.1}$$

where $*$ denotes a dual. ($F^*_{ab} = \frac{1}{2}\epsilon_{abcd}F^{cd}$, $\epsilon_{0123} = 1$, $\epsilon^{0123} = -1$, and $F^{ab} + iF^{*ab}$ is anti-self-dual.) By contracting (3.1) with l_b we obtain a single equation

$$l_b\nabla_a(F^{ab} + iF^{*ab}) = 0 \tag{3.2}$$

that is equivalent to (3.1). This is easily seen by applying ð,$\bar{ð}$, and ð$\bar{ð}$ to (3.2) and using (2.1) to obtain the tetrad components of (3.1). If (3.2) is rewritten as

$$l^b\nabla_a\eta^{ac}(F_{bc} + iF^*_{bc}) = 0$$

and η^{ac} is expressed by (2.7), we have

$$D\phi_1 - \delta\phi_0 = 0, \tag{3.3}$$

where we have used the anti-self-dual nature of $F_{ab} + iF^*_{ab}$ and the definitions (2.17). Note that (3.3) is one of the Maxwell equations in the spin-coefficient formalism using a parallelly propagated null tetrad. The other equations could be obtained by applying ð, $\bar{ð}$, and ð$\bar{ð}$ to (3.3).

If we now use, from (2.18), that $\phi_1 = \frac{1}{2}\bar{\eth}\phi_0$, we obtain a single equation for ϕ_0, namely,

$$D\bar{\eth}\phi_0 - 2\delta\phi_0 = 0. \tag{3.4}$$

To Eq. (3.4), which determines the spatial behavior of Φ, we must add the angular equation (2.20), i.e.,

$$\eth\phi_0 = 0. \tag{3.5}$$

Equations (3.4) and (3.5) constitute the equations for the local cross sections of a line bundle over the null-cone bundle on Minkowski space which are equivalent to the vacuum Maxwell equations.

The remainder of this section will be devoted to simplifying (3.4) by introducing an alternative variable for the ϕ_0.

If we return to the relationship of the Maxwell field to the vector potential γ_a, i.e.,

$$F_{ab} = 2\nabla_{[a}\gamma_{b]},$$

we have

$$\phi_0 = l^a\nabla_a(m^b\gamma_b) - m^a\nabla_a(l^b\gamma_b) \quad \text{or} \quad \phi_0 = D\eth\gamma - \delta\gamma, \tag{3.6}$$

with

$$\gamma(x^a,\zeta,\bar{\zeta}) = l^a\gamma_a, \quad \eth\gamma = m^a\gamma_a. \tag{3.7}$$

We now wish to introduce, in the following way, a "superpotential" for $\gamma(x^a,\zeta,\bar{\zeta})$: we consider a gauge transformation

$$\gamma'_a = \gamma_a + \nabla_a F,$$

with F depending on $(\zeta,\bar{\zeta})$ as well as x^a [the new potential γ'_a thus also depends on x^a and $(\zeta,\bar{\zeta})$], with the condition that $\gamma' \equiv l^a\gamma'_a = 0$. This can always be done, at least locally. We thus have F defined by

$$DF = -\gamma_a(x^b)l^a = -\gamma. \tag{3.8}$$

Here, $F(x^a,\zeta,\bar{\zeta})$ will be our "superpotential." Using (3.8) in (3.6) we have

$$\phi_0 = -D\eth DF + \delta DF = -D^2\eth F, \tag{3.9}$$

where we have used (2.12) and the fact that D, δ, $\bar{\delta}$, and Δ all commute. Defining

$$\eth F = -\Lambda, \tag{3.10}$$

we have the very simple relationship between the field and Λ, namely,

$$\phi_0 = D^2\Lambda. \tag{3.11}$$

If we substitute (3.11) into the Maxwell equations (3.4) and (3.5), we obtain, after some simplification, using (2.12),

$$D^3\bar{\eth}\Lambda = 0, \tag{3.12}$$

and

$$D(D(\bar{\eth}\Lambda) + 2\delta\Lambda) = 0. \tag{3.13}$$

Equation (3.13) can be replaced by the stronger equation

$$D(\bar{\eth}\Lambda) + 2\delta\Lambda = 0, \tag{3.14}$$

by the following argument: since $\gamma_a(x^a)$ is independent of $(\zeta,\bar{\zeta})$ we have, from (2.1) and (2.5), that

$$\eth^2\gamma = 0, \tag{3.15}$$

from which, using (3.10) and (3.8), we have (3.13).

We will consider (3.12) and (3.14) as Maxwell's equations.

It is important to mention here that there is considerable gauge freedom in the choice of F and Λ. In fact there are two types of freedom in F; the first is the usual gauge transformation of $\gamma_a \to \gamma_a + \nabla_a \lambda$ with λ a function of x_a alone, which induces

$$F' = F + \lambda, \tag{3.16}$$

while the second is

$$F' = F + F_0, \tag{3.17}$$

with $DF_0 = 0$. The first induces no change in the Λ and is uninteresting for us, while the second does, namely,

$$\Lambda' = \Lambda - \eth F_0. \tag{3.18}$$

We will eventually (Secs. IV and V) consider two classes of solutions to the Maxwell equations: (1) the globally regular solutions, such as the half-retarded minus the half-advanced, with "nice" behavior in the infinite future and past, and (2) solutions arising from a finite number of point sources as, for example, the Coulomb field or a superposition of a finite number of advanced or retarded Lienard–Wiechert fields. In class (1) we will be able to find a global solution (on $M \times S^2$) for both F and Λ, i.e., F and Λ will be regular functions of x^a and $(\zeta,\bar{\zeta})$. For class (2), this will not be the case. Both F and Λ will have singularities where the usual Maxwell fields do; however, in addition, they will also have angular singularities at regular points of the Maxwell fields. There will however, always be an F_0 which acts as an overlap function, giving a different cross section F' which will be singular in different directions.

We will show later that, in both of the above mentioned classes, Λ satisfies the stronger set of equations

$$a \equiv D\bar{\eth}\Lambda = 0, \quad b \equiv D\eth\Lambda + 2\delta\Lambda = 0, \tag{3.19}$$

rather than (3.12) and (3.14). [It seems likely that by some appropriate gauge choice, any Maxwell field will satisfy (3.19).] We now find some consequences of (3.19).

By applying $\bar{\eth}$ to b and $\eth$ to a, using (2.12) frequently and $\bar{\eth}\eth\Lambda - \eth\bar{\eth}\Lambda = 2\Lambda$ we obtain

$$\bar{\eth}b - \eth a = \bar{\delta}\delta\Lambda + \delta\bar{\delta}\Lambda + 2\Delta\Lambda = 0. \tag{3.20}$$

Now applying D to (3.20) and using (3.19) we have

$$\Box\Lambda = (D\Delta - \delta\bar{\delta})\Lambda = 0, \tag{3.21}$$

i.e., Λ satisfies the wave equation. From (3.21) it immediately follows that

$$\Box\bar{\eth}\Lambda = 0, \quad \Box\eth\Lambda = 0, \tag{3.22}$$

and hence, using (3.19), we have that

$$\delta\bar{\delta}\bar{\eth}\Lambda = 0. \tag{3.23}$$

If we return to (3.20) and apply δ, we have, using (3.19) and (3.22),

$$\delta^2\bar{\eth}\Lambda + \delta\bar{\delta}\eth\Lambda + 2\Delta\delta\Lambda = 0,$$

$$\delta^2\bar{\eth}\Lambda + D\Delta\eth\Lambda + 2\Delta\delta\Lambda = 0,$$

$$\delta^2\bar{\eth}\Lambda + \Delta(D\eth\Lambda + 2\delta\Lambda) = 0, \quad \delta^2\bar{\eth}\Lambda = 0. \tag{3.24}$$

Finally, by applying $\bar{\eth}\bar{\delta}$ to (3.20) we have, after a brief calculation,

$$\eth(\bar{\delta}^2\bar{\eth}\Lambda) = 0. \tag{3.25}$$

Summarizing, our relevant equations are

$$D\bar{\eth}\Lambda = 0, \quad D\eth\Lambda + 2\delta\Lambda = 0, \quad \Box\Lambda = 0,$$
$$\delta\bar{\delta}\bar{\eth}\Lambda = 0, \quad \delta^2\bar{\eth}\Lambda = 0, \quad \eth\bar{\delta}^2\bar{\eth}\Lambda = 0. \tag{3.26}$$

IV. GLOBALLY REGULAR SOLUTIONS

In this section we will consider solutions to (3.26) which are globally regular on the null-cone bundle.

Due to the assumed regularity, the last of (3.26) integrates to

$$\bar{\delta}^2\bar{\eth}\Lambda = 0. \tag{4.1}$$

Also from regularity and the fact that $\bar{\eth}\Lambda$ is a real spin-weight zero function, we have

$$\bar{\eth}\Lambda = \bar{\eth}A + \eth\bar{A}. \tag{4.2}$$

(Note that though here we are considering real Maxwell fields, if we wanted to consider complex ones, e.g., arbitrary combinations of self- and anti-self-dual fields, then A and $\bar{A}$ would not be complex conjugates of each other. One denotes in that case $\bar{A}$ by $\tilde{A}$.) We now try to determine A and $\bar{A}$. The first of (3.26) forces A to be independent of n, i.e.,

$$DA = 0. \tag{4.3}$$

Thus, A and $\bar{A}$ are, at this point, functions of l, m, $\bar{m}$, $\zeta,\bar{\zeta}$. Now, since

$$\eth\bar{A}(x^a,\zeta,\bar{\zeta}) = \eth'\bar{A} + \frac{\partial\bar{A}}{\partial l}m + \frac{\partial\bar{A}}{\partial\bar{m}}(n - l), \tag{4.4}$$

and

$$\bar{\eth}A = \bar{\eth}'A + \frac{\partial A}{\partial l}\bar{m} + \frac{\partial A}{\partial m}(n - l), \tag{4.5}$$

where $\bar{\eth}'$ and $\eth'$ refer to differentiation holding l, n, m, and $\bar{m}$ constant [see (2.8)–(2.11) and 2.1], we have, from the first of (3.26), that $\bar{A}$ is independent of $\bar{m}$ and A is independent of m. Finally, from (4.4) and (4.1) we have the result that

$$A = A(l,\zeta,\bar{\zeta}), \quad \bar{A} = \bar{A}(l,\zeta,\bar{\zeta}), \tag{4.6}$$

where A is an arbitrary regular spin-weight (s.w.) 1 function of $l,\zeta,\bar{\zeta}$. It constitutes the data for a solution. If we define J by

$$\Lambda = J + A, \tag{4.7}$$

(4.2) becomes

$$\bar{\eth}J = \eth\bar{A}. \tag{4.8}$$

Equation (4.8) (and its complex conjugate) can be easily integrated [by means of a Green's function for $\bar{\eth}$ (and $\eth$)] yielding J (and $\bar{J}$).

Notice that the Maxwell fields

$$\phi_0 = D^2\Lambda = D^2 J \quad \text{and} \quad \bar{\phi}_0 = D^2\bar{\Lambda} = D^2\bar{J} \tag{4.9}$$

are determined by J (and $\bar{J}$), since $DA = D\bar{A} = 0$, and not the full Λ. It is a surprising fact that the J from Λ is determined by the $\bar{A}$ from $\bar{\Lambda}$ (and $\bar{J}$ by the A from Λ).

(A similar type of situation[5] occurs in Yang–Mills theory where J is determined by $\bar{A}$ but the nonlinearity shows itself by an interaction term of J with $\bar{J}$.)

The field ϕ_0 in (4.9) can be easily seen to be identical with the usual Kirchoff integral formula for radiation fields with $\ddot{\bar{A}}$ as characteristic data.

To compute ϕ_0 we will make use of a Green function of the $\bar{\eth}$ operator that can be found in the literature.[6]

If f is a s.w. zero function (not necessarily real) then

$$\bar{\eth} f = \bar{A} \tag{4.10}$$

can be integrated with the use of a function G of s.w. zero in $(\zeta, \bar{\zeta})$ and s.w. 1 in $(\eta, \bar{\eta})$ that satisfies

$$\bar{\eth} G_{0,1} = \delta(\bar{\zeta} - \bar{\eta}).$$

It can be shown[6] that $G_{0,1}$ is given by

$$G_{0,1} = (1 + \eta\bar{\eta})/(\zeta - \eta), \tag{4.11}$$

and hence f can be written as

$$f = \bar{\eth}^{-1}\bar{A} = \iint G\bar{A}\ dS', \tag{4.12}$$

where $dS' = d\eta\, d\bar{\eta}/(1 + \eta\bar{\eta})^2$ is the volume element of S^2 and the indices 0, 1 have been suppressed for simplicity.

If one takes $\eth$ on (4.10) one finds

$$\eth\bar{\eth} f = \bar{\eth}\eth f = \eth\bar{A} = \bar{\eth} J, \tag{4.13}$$

therefore,

$$J = \eth f = \eth(\bar{\eth}^{-1}\bar{A}) = \iint \eth G\bar{A}(x^a l'_a, \eta, \bar{\eta}) dS', \tag{4.14}$$

where l'_a is the null vector of (2.2) expressed in the $(\eta, \bar{\eta})$ system. To obtain (4.9) one computes

$$D^2 J = l^a l^b J_{,ab} = \iint \eth G (l^a l'_a)^2 \ddot{\bar{A}}\ dS'. \tag{4.15}$$

Finally, inserting

$$\eth G = -(1 + \zeta\bar{\zeta})(1 + \eta\bar{\eta})/(\zeta - \eta)^2$$

and

$$l^a l'_a = (\zeta - \eta)(\bar{\zeta} - \bar{\eta})/(1 + \eta\bar{\eta})(1 + \zeta\bar{\zeta})$$

in (4.15), one obtains

$$\phi_0 = -\iint \frac{(\bar{\zeta} - \bar{\eta})^2 \ddot{\bar{A}}\ dS'}{(1 + \eta\bar{\eta})(1 + \zeta\bar{\zeta})}, \tag{4.16}$$

which is the same expression as the projection of the Kirchoff integral for radiation fields in the $l_{[a} m_{b]}$ direction.[7]

It is from this fact that we justify our claim that the globally regular [class (1) fields] satisfy $D\bar{\eth}\Lambda = 0$ rather than $D^3\bar{\eth}\Lambda = 0$.

As a final comment we point out that self-dual (or anti-self-dual) fields arise easily as a specialization of (4.8), (4.9), and their conjugates. We must first consider A and $\bar{A}$ (now denoted by $\tilde{A}$) as independent parts of the data, no longer complex conjugates of each other. If we then take $\tilde{A} = 0$, we have $J = 0$ and hence $\phi_0 = 0$. On the other hand $\tilde{J}$ is determined by A, thus yielding a nonvanishing $\tilde{\phi}_0$, the self-dual part of the Maxwell field. Giving $\tilde{A} \neq 0$ and $A = 0$ yields in the same way the anti-self-dual field. In the former case we have that

$$\bar{\eth} F = -\Lambda = -A, \tag{4.17}$$

the Sparling equation for self-dual Maxwell fields.[8,9]

V. THE LIENARD–WIECHERT SOLUTIONS

In the previous sections we obtained all pure radiation solutions of the Maxwell equations by demanding global regularity for our fields on $M \times S^2$ and using the strengthened equation $D\bar{\eth}\Lambda = 0$ rather than $D^3\bar{\eth}\Lambda = 0$. In the present section we will work backwards, taking the known tensor version of the Lienerd–Wiechert solution and constructing the associates F, Λ, $\bar{\eth}\Lambda$ fields on $M \times S^2$.

The two main points to be made are that (1) again we will have $D\bar{\eth}\Lambda = 0$ (in fact $\bar{\eth}\Lambda =$ const) and (2) the solutions will have two types of singularities; the first occur at the source points and are essential (i.e., cannot be transformed away by a gauge transformation, $F \rightarrow F + F_0$), the second occur as angular singularities, which, however, can be shifted to other directions by a gauge transformation $F \rightarrow F + F_0$. This means that we can think of the Lienard–Wiechert (LW) solutions (advanced and retarded) as well as linear combinations of LW with different source world-lines, as giving rise to a nontrivial line-bundle over $(M - \{\text{sources}\}) \times S^2$ with the F_0 as the overlap function.

In order to prove this we begin with the single LW field described as follows. Assume that the world-line L of a charged particle in M is given by

$$x^a = \xi^a(\tau), \tag{5.1}$$

where τ is the proper time along L. The velocity is then given by

$$v^a = \dot{\xi}^a(\tau), \tag{5.2}$$

with $v^a v_a = 1$.

If x^a is any point on M, then there will be two values of τ (one advanced and one retarded) associated with the intersections of the future and past null-cones of x^a with L. These two values of τ are obtained from

$$(x^a - \xi^a(\tau))(x_a - \xi_a(\tau)) = 0, \tag{5.3}$$

[i.e., $x^a - \xi^a(\tau)$ must be a null vector]. Thus one obtains, at least implicitly, two functions $\tau \equiv \tau_R = g_R(x^a)$ for the retarded proper time, and $\tau \equiv \tau_A = g_A(x^a)$, for the advanced time, of the space-time point relative to L.

For definiteness we will now choose the retarded τ and discuss the retarded Lienard–Wiechert field associated with L.

For a given x^a, one defines the "radial" distance r, from L to x^a, by

$$r = (x^a - \xi^a(\tau))v_a(\tau), \tag{5.4}$$

which is positive. [Note that if

$$\hat{x}^a \equiv (x^a - \xi^a) - v^a(x^b - 1\xi^b)v_b$$

we have $r^2 = -\hat{x}^2$.] The Lienard–Wiechert potential is given by

$$\gamma_a(x^b) = q[v_a(\tau)/r(\tau)]. \tag{5.5}$$

We now proceed to find F from (5.5) via (3.7), i.e., we must evaluate the integral

$$F = -q\int n\frac{v_a l^a}{r} dn. \tag{5.6}$$

Since τ is a function of x^a, which in turn, from (2.6), is a function of n, one can replace n by τ as the integration variable in (5.6). This requires knowledge of $D\tau \equiv \partial\tau/\partial n$. This can be calculated by observing that $\nabla_a \tau$ can be obtained from (5.3) to be

$$\nabla_a \tau = [x_b - \xi_b(\tau)]/r(\tau). \qquad (5.7)$$

[Note that $\nabla_a \tau$ is a null vector pointing from $\xi_a(\tau)$ to the field point x^a and furthermore $v^a \nabla_a \tau = 1$.]

We now have

$$l^a \nabla_a \tau \equiv \frac{\partial}{\partial n}\tau \equiv D\tau = \frac{(x_a - \xi_a)}{r(\tau)} l^a, \qquad (5.8)$$

which, when substituted into (5.6) yields

$$F \equiv F_R = -q \int \tau \frac{(d/d\tau)(\xi_a l^a) d\tau}{(x_b - \xi_b) l^b} \qquad (5.9)$$

$$= -q \int \frac{d(\xi_a l^a)}{l - \xi_a l^a},$$

which integrates immediately to

$$F_R = q \log[(x^a - \xi^a) l_a]. \qquad (5.10)$$

For the advanced fields we replace (5.4) by

$$r = (\xi^a - x^a) v_a, \qquad (5.11)$$

so that r is still positive, and use the advanced τ to obtain in a similar fashion

$$F_A = -q \log[(\xi_a - x_a) l^a]. \qquad (5.12)$$

An important point to note is that $x^a - \xi^a$ is a null vector pointing from L to the field point x^a and hence when l^a points in the same "outward" direction the argument of the log in (5.10) vanishes, yielding a singularity for F_R. Likewise $\xi_a - x_a$ points from the field point towards L and hence F_A has a singularity when l_a points "inward."

We will now argue that an F_{0R} and F_{0A} can be found so that F'_R has a singularity in the "inward" direction while F'_A has a singularity in the "outward" direction.

The pure radiation solution of the Maxwell equations given by the half-retarded minus half-advanced Lienard–Wiechert fields is a globally defined field and hence from the previous section there must exist a regular (on $M \times S^2$) $F(x, \zeta, \bar{\zeta})$ associated with it. Since $F_R - F_A$ yields the same radiation field we must have the patching function $F_0(l, m, \bar{m}, \zeta, \bar{\zeta})$ so that

$$F(x^a, \zeta, \bar{\zeta}) = F_R - F_A + F_0. \qquad (5.13)$$

Since F is regular, F_0 is regular in all directions except the singular directions of F_R and F_A. If we define F'_R and F'_A by

$$F'_R = F_R + F_0 = F + F_A, \qquad (5.14)$$

and

$$F'_A = F_A - F_0 = F_R - F, \qquad (5.15)$$

we see that F'_R (F'_A) yields the same fields as F_R (F_A) but that the directions of the singularities have been switched, i.e., the F_0 defined in (5.13) acts as the patching function in (5.14) and (5.15).

We now examine the special case of the Coulomb solution. In this case the world-line L is given by

$$\xi^a(\tau) = t^a \tau, \qquad (5.16)$$

with $t^a = v^a = (l^a + n^a)/\sqrt{2}$. We find from (5.3) that

$$\tau = x^a t_a \pm \sqrt{(x^a t_a)^2 - x^a x_a}, \qquad (5.17)$$

with the plus and minus signs corresponding to advanced and retarded times. This reduces to the usual definitions

$$\tau = t \pm r, \quad r = \sqrt{x^2 + y^2 + z^2}. \qquad (5.18)$$

In the retarded case we find that

$$(x^a - \xi^a) l_a = l - \tau/\sqrt{2},$$

which with (5.18) and $t = (l + n)/\sqrt{2}$ becomes

$$(x^a - \xi^a) l_a = (1/\sqrt{2})(r + (l - n)/\sqrt{2}).$$

Thus

$$F_R = q \log(r + (l - n)/\sqrt{2}), \qquad (5.19)$$

and similarly

$$F_A = -q \log(r - (l - n)/\sqrt{2}). \qquad (5.20)$$

Both (5.19) and (5.20) yield the same Coulomb solution and hence their difference is the patching function F_0.

It is a simple matter to construct, from the Coulomb F, the F associated with all static multipole fields. For example by taking the gradient of the Coulomb F in some fixed constant direction, we obtain the F of the dipole field, i.e.,

$$F_D = d^a \nabla_a F,$$

which yields a dipole field with dipole moment qd^a. Higher-moment fields are found by further differentiation.

To conclude this section we calculate the A and $\bar{\eth} A$ associated with the Lienard–Wiechert fields. From (5.10) we have

$$A_R = -\eth F_R = -q[(x^a - \xi^a) m_a / (x^a - \xi^a) l_a] \qquad (5.21)$$

and

$$\bar{\eth} A_R = -q[(x^a - \xi^a)(n_a - l_a)(x^b - \xi^b) l_b - (x^a - \xi^a) m_a (x^b - \xi^b) \bar{m}_b] / [(x^a - \xi^a) l_a]^2. \qquad (5.22)$$

It is not difficult to see from (2.4) and the null character of $x^a - \xi^a$ that

$$\bar{\eth} A_R = -q, \qquad (5.23)$$

which proves an earlier contention that $D\bar{\eth} A = 0$ for Lienard–Wiechert.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation and by a University Research Council Grant No. 501.

APPENDIX: AN EXAMPLE

We wish to present here an example of the integration of the equation

$$\bar{\eth} J = \eth \bar{A} \qquad (A1)$$

for a particular, interesting choice of $\bar{A}$, namely,

$$\bar{A} = \bar{\eth} d / L^2, \qquad (A2)$$

with

$$d = d_a l^a, \quad d_a \text{ a constant spacelike vector,}$$

$$L = (x^a - z^a) l_a = l - l_0,$$

$z^a = \mathring{x}^a + i\mathring{y}^a$ being a fixed point in *complex* Minkowski space.

It turns out to be simpler to solve

$$\bar{\eth} f = \bar{A} \qquad (A3)$$

for f and then construct J by

$$J = \eth f. \tag{A4}$$

It is easy to check from (A2) and (A3) that $\bar{\eth}^2(Lf) = 0$ so that

$$f = \alpha_a l^a / L, \tag{A5}$$

with $\alpha_a\ (x^b, d^b)$. Substitution of (A5) into (A3) yields the following algebraic equation for the determination of α_a:

$$2\alpha_{[a} x'_{b]} \bar{m}^a l^b = d_a \bar{m}^a = 2\sqrt{2} d_{[a} t_{b]} \bar{m}^a l^b,$$

where $x'^a = x^a - z^a$ and t^a is the unit timelike vector such that $t_a l^a = 1/\sqrt{2}$. Since $l^{[a}\bar{m}^{b]}$ is the self-dual we have

$$\alpha_{[a} x'_{b]+} = \sqrt{2} d_{[a} t_{b]+}, \tag{A6}$$

where $+$ denotes the self-dual part of the bivector. The solution to (A6) is obtained by multiplying (A6) by x'^b, yielding

$$\alpha_a = 2\sqrt{2}(d_{[a} t_{b]+}) x'^b / x'^2 + x_a \phi(x), \tag{A7}$$

with $\phi(x)$ an arbitrary function of x^b and $x'^2 = x'^a x'_a$. After some computation we have

$$\alpha_a l^a = (2\eth L \bar{\eth} d - L \eth \bar{\eth} d)/x'^2 + l\phi(x) \tag{A8}$$

and hence

$$f = (2\eth L \bar{\eth} d - L \eth \bar{\eth} d)/x'^2 L + \phi(x). \tag{A9}$$

The freedom in f is the arbitrary, inessential function ϕ. One can easily check that (A9) satisfies (A3). We finally have, from (A9) and (A4),

$$J = \eth f = (2/x'^2 L^2)[\eth L \cdot L \eth \bar{\eth} d + L^2 \eth d - (\eth L)^2 \bar{\eth} d]. \tag{A10}$$

Note that in J, the only place that the variable $n = x^a n_a$ appears is in x'^2 and thus the Maxwell field becomes

$$\phi = D^2 J = 2D^2(1/x'^2) L^{-2}[L \eth L \eth \bar{\eth} d + L^2 \eth d - (\eth L)^2 \bar{\eth} d]. \tag{A11}$$

Using $x^a x_a = 2(ln - m\bar{m})$, one find that

$$D^2(1/x'^2) = 8L^2/(x'^a x'_a)^3 \tag{A12}$$

and hence

$$\begin{aligned}\phi &= F_{ab} l^a m^b \\ &= [(x^a - z^a)(x_a - z_a)]^3 \\ &\quad \times [L \eth L \eth \bar{\eth} d + L^2 \eth d - (\eth L)^2 \bar{\eth} d].\end{aligned} \tag{A13}$$

The F_{ab} which can be easily reconstructed from (A13) is a pure (dipole) radiation field.

In a future paper we will show that in a fashion similar to the integration process we have used here, with essentially the same data, we can, for the Yang–Mills equations, produce the single instanton solution, and hence the solution given here can be thought of as the Maxwell "instanton" solution with z^a the instanton "position."

[1]M. Ko, E. T. Newman, and K. P. Tod, in *Asymptotic Structure of Space-Time*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977).

[2]S. L. Kent and E. T. Newman, J. Math. Phys. **24**, 949 (1983).

[3]I. Bialyncki-Birula, E. T. Newman, J. Porter, J. Winicour, B. Lukacs, Z. Perjes, and A. Sebesteyn, J. Math. Phys. **22**, 2530 (1981).

[4]E. T. Newman and K. P. Tod, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 2.

[5]C. N. Kozameh and E. T. Newman, Phys. Rev. D (to be published).

[6]J. Porter, Gen. Relativ. Gravit. **13**, 531 (1981).

[7]R. Penrose, in *Relativity Groups and Topology* (Gordon and Breach, New York, 1964).

[8]G. Sparling, University of Pittsburgh report (unpublished).

[9]E. T. Newman, Phys. Rev. D **22**, 3023 (1980).

The dequantization programme for stochastic quantum mechanics

F. E. Schroeck, Jr.
Department of Mathematics, Florida Atlantic University, Boca Raton, Florida 33431

(Received 28 October 1983; accepted for publication 8 August 1984)

The classical limit $\hbar \to 0$ of stochastic quantum mechanics is investigated. An algebra of stochastic classical observables is canonically associated to the algebra of stochastic quantum observables with a product inherited from sequential instruments and also with a Lie commutator. It is shown that the sharp-point limit (limit of no stochasticity), which implies $\hbar \to 0$, yields the usual algebra of classical observables.

I. INTRODUCTION

The classical limit $\hbar \to 0$ of the algebra of observables obtained from quantum mechanics via the Weyl transformation has been recently shown by Emch[1] to give the usual algebra of classical observables and the Wigner–Weyl–Moyal transform of the quantum mechanical density operator was shown to become a true probability density. To achieve this, Emch took the ordinary product of (quantum) operators as the given product in the algebra of quantum observables. In the stochastic formalism,[2] the observables are defined via instruments in the Davies–Lewis scheme.[3] Thus we take as the product of observables that which follows from sequential measurement. The dequantization program is described generally for any stochastic formalism and detailed calculation in the minimum uncertainty formalism is given. In the latter model it is shown that it is the sharp-point limit rather than the limit $\hbar \to 0$ that gives the correct classical limit. The sharp-point limit, due to the uncertainty relation, nevertheless implies $\hbar \to 0$.

II. GENERAL STOCHASTIC DEQUANTIZATION

For fixed value of $\hbar$, let $\mathscr{W}_\hbar = \{ W_\hbar(z) | z \in \mathbb{R}^{2n}\}$ denote the Weyl algebra in which

$$\begin{aligned} &W_\hbar(z) W_\hbar(s) = \chi_\hbar(z, s) W_\hbar(z+s), \\ &\chi_\hbar(z, s) = \exp\{ i\hbar(z_1 \cdot s_2 - z_2 s_1)/2\}, \\ &z = (z_1, z_2), \quad s = (s_1, s_2), \quad z_i, s_i \in \mathbb{R}^n. \end{aligned} \tag{1}$$

We work in a Hilbert space H in which $\mathscr{W}_\hbar$ is unitarily, irreducibly represented.

Let T_0 be the projection onto some one-dimensional subspace of H. (Usually we take T_0 to project onto vectors in which the position and momentum operators Q and P have expectation = zero.) Let

$$T_z = W_\hbar(z) T_0 W_\hbar^\dagger(z). \tag{2}$$

Then, as a strong integral,

$$\int T_z \, \mathrm{d}z = \lambda \mathbb{1}, \quad \lambda > 0, \tag{3}$$

which follows from Schur's lemma or by the spectral theorem.

A stochastic instrument is given by[4]

$$\rho \mapsto \mathscr{E}(f, \Delta; \rho) \equiv \lambda^{-1} \int_\Delta dz\, f(z) T_z \rho T_z, \tag{4}$$

where ρ is a density operator, Δ is a Borel subset of phase space, and f is a non-negative bounded Lebesgue measurable function. For an ideal (uniform) stochastic instrument $f(z) \equiv 1$. $f(z)$ may be interpreted as the relative probability of the instrument to convert ρ to the state T_z, in which the expected value of (Q, P) is z. Since

$$\mathrm{Tr}(\mathscr{E}(f, \Delta; \rho)) = \mathrm{Tr}\Big(\rho \lambda^{-1} \int_\Delta dz\, f(z) T_z\Big), \tag{5}$$

we identify

$$A(f, \Delta) = \lambda^{-1} \int_\Delta dz\, f(z) T_z \tag{6}$$

as the observables canonically associated with the instrument $\mathscr{E}$. For simplicity, we shall simply write observables as

$$a(g) = \lambda^{-1} \int dz\, g(z) T_z, \tag{7}$$

restricting g to have support in supp(f), g = real-valued measurable function of phase space (classical observable). We shall similarly absorb f into restrictions on the support of g in

$$\mathscr{E}(g, \rho) = \int dz\, g(z) T_z \rho T_z. \tag{8}$$

We next define

$$\rho(z) \equiv \lambda^{-1}\, \mathrm{Tr}(\rho T_z) \tag{9}$$

which is positive, and from (3), and Appendix A,

$$\int dz\, \rho(z) = 1. \tag{10}$$

Thus $\rho(z)$ is a classical probability density which we shall call a stochastic classical density. We now have (see Appendix A)

$$\mathrm{Tr}(\rho a(g)) = \int dz\, g(z)\, \rho(z) \tag{11}$$

so that we may identify $a(g)$ as the stochastic quantum observable canonically associated to the (stochastic) classical observable g. $a(g)$ is a bounded self-adjoint operator in H for g real-valued and in $L_1 \cup L_\infty$; in particular, both the operator norm and trace norm of $a(g)$ are bounded by min $\{ \|g\|_\infty, \lambda^{-1} \|g\|_1 \}$.

We remark that if g, h are such classical observables and σ is a classical density (state), then

$$\sigma \mapsto gh\sigma \tag{12}$$

represents measurement "of h" followed by immediate measurement "of g." In contrast, if A, B are self-adjoint operators (quantum observables) on H and ρ is a quantum state then $\rho \rightarrow AB\rho$ cannot represent a sequential measurement generally since generally AB is not self-adjoint. Even for P^A, P^B spectral projectors for A, B, $P^A P^B$ cannot generally represent an observable. With this well-known difficulty in mind, one next takes the Jordan product, $\frac{1}{2}(AB + BA)$, for the product of observables. A physical motivation for this product may be developed from the assumption that there are sufficiently many dispersion-free states for A, B, $A + B$ so as to define powers of these operators.[5] We do not assume this here. Furthermore, if the map $g \rightarrow a(g)$ is a Jordan homomorphism, then the resulting observables $a(g)$ form a commutative C^*algebra[6] and the positive operator-valued measure associated with a is projection valued only.[6] This in turn is related to measurements, which if repeated "immediately after" an initial measurement, yield the same results.[7] These consequences are known to be too restrictive for quantum measurement and it is precisely their generalization which gives the Davies and Lewis measurement scheme.[3] We take the view that the ordinary operator product and Jordan product are not the appropriate ones for sequential measurement of observables (but see Ref. 8 for further discussion), although the pointwise product of functions is the sequential measurement product for classical observables. We do, however, have, using the technical result in Appendix A,

$$
\begin{aligned}
&\mathrm{Tr}[\mathscr{E}(g, \mathscr{E}(h,\rho))] \\
&\quad = \mathrm{Tr}\left[\lambda^{-2}\int dz\, ds\, g(z)h(s)T_z T_s \rho T_s T_z\right] \\
&\quad = \lambda^{-2}\int dz\, ds\, g(z)h(s)\mathrm{Tr}(\rho T_s T_z T_s) \\
&\quad = \lambda^{-2}\int dz\, ds\, g(z)h(s)\beta(z,s)\mathrm{Tr}(\rho T_s) \\
&\quad = \int dz\, ds\, g(z)h(s)[\lambda^{-1}\beta(z,s)]\,\rho(s),
\end{aligned}
$$

which we define as

$$\equiv \mathrm{Tr}(\rho a(g \circ_\hbar h)), \tag{13a}$$

where

$$\beta(z,s) \equiv \mathrm{Tr}(T_z T_s), \tag{13b}$$

the transition probability from the stochastic point z to the stochastic point s, and where we define

$$(g \circ_\hbar h)(s) = h(s)\int dz\, g(z)\lambda^{-1}\beta(z,s) \tag{13c}$$

the sequential instrument product which reflects the disturbance of the first measurement on the second. From (3) we see that $\lambda^{-1}\beta(z,\cdot)$ and $\lambda^{-1}\beta(\cdot,s)$ are classical probability densities. Thus,

$$|g \circ_\hbar h)(s)| < |h(s)|\,\|g\|_\infty, \tag{13d}$$

so (13c) defines a bilinear map on L_∞.

We also define a sequential product of quantized observables $*_\hbar$ by

$$a(g) *_\hbar a(h) = a(g \circ_\hbar h). \tag{14}$$

[We parenthetically remark that, through $\lambda^{-1}\beta(z,\cdot)$, we associate to each z in phase space a random variable (probability density, in fact). We may therefore consider $\lambda^{-1}\beta$ as a stochastic variable (random function). This justifies the terms "stochastic point," "stochastic geometry."]

In this stochastic geometry, the sharp-point limit is precisely

$$\lambda^{-1}\beta(z,s) \rightarrow \delta(z-s) \tag{15}$$

(as the kernel for an integral operator on the bounded continuous functions), so that in general, the sharp-point limit of the sequential instrument product (13c) for (stochastic) classical observables is well defined in the L_∞^c topology, and yields precisely the pointwise product of classical observables which are bounded, continuous. Similarly, from the L_∞ bound for $a(g)$ we conclude from (14) that the $*_\hbar$ product has a well-defined sharp-point limit which is commutative in that limit.

In the sharp-point limit, the $\lambda^{-1}\beta$ variances of the position and momentum coordinates both vanish. This suggests, by the uncertainty principle, that then $\hbar \rightarrow 0$, the usual classical limit. In the model to be treated in the next section, we show this, and show that we may not replace the sharp-point limit with the (weaker) condition $\hbar \rightarrow 0$.

The other algebraic structures present in the algebras of quantum and classical observables are the Lie commutator and the Poisson bracket.

From (2) and (7) we have

$$
\begin{aligned}
W_\hbar(s)a(g)W_\hbar^\dagger(s) &= \lambda^{-1}\int dz\, g(z)T_{z+s} \\
&= a(s[g]),
\end{aligned}
$$

where

$$(s[g])(z) = g(z-s). \tag{16}$$

More generally, the observables form a system of covariance for some space-time symmetry group G, unitarily and irreducibly represented by U in H and by V in phase space,[9] i.e.,

$$
\begin{aligned}
&U_s a(g)U_s^\dagger = a(s[g]), \\
&(s[g])(z) = g(V_s z), \quad s \in G.
\end{aligned}
\tag{17}
$$

(G may be taken to be either the Galilea group or the Poincaré group, for example.) Thus to any element of the Lie algebra of the Lie group G we canonically associate a self-adjoint operator in H for the corresponding generator in the U representation, which in turn is associated to a generator (vector field) in the V representation, which in turn is connected to a classical observable in the usual manner by

$$g(+\,\mathrm{const}) \leftrightarrow v_g = \sum_i \left[\frac{\partial g}{\partial p_i}\frac{\partial}{\partial q_i} - \frac{\partial g}{\partial q_i}\frac{\partial}{\partial p_i}\right]. \tag{18}$$

[We remark that the self-adjoint operator in H representing the Lie algebra element we assume to be representable in the form $a(g)$ for some g. For these $a(g)$ it then is sensible to take the Lie bracket.]

Since

$$[v_g, v_h] = v_{\{g,h\}}, \tag{19}$$

the Lie commutator of the generators of G in H is naturally connected, via the covariance condition, to the Poisson bracket for the corresponding (stochastic) classical observables. We emphasize that this connection is completely independent of any sharp-point or $\hbar \to 0$ limit, and thus is valid at the stochastic classical level. (We remark that this is unlike the result of Emch in the analysis of the Wigner–Weyl–Moyal dequantization problem[1] in which the connection holds only in the $\hbar \to 0$ limit.)

For more general transformations, we first recall that in the minimum uncertainty model and ideal spin model, almost any operator B may be written in the form

$$B = \int d\,\mu(z) T_z \tag{20}$$

for some measure μ on phase space,[8] respectively, spin space.[10] We consider here a continuous unitary ray representation $\{U(t), t \in G\}$ of some group in H with the property

$$U(t) T_z U^{-1}(t) = \int ds\, \gamma_t(z, s) T_s \tag{21}$$

for some function γ_t. Taking trace of (21) yields

$$\int ds\, \gamma_t(z, s) = 1 \tag{22}$$

for all $t \in G$. Thus,

$$\begin{aligned} U(t) \alpha(g) U^{-1}(t) &= \int d\xi \int ds\, g(\xi) \gamma_t(\xi, s) T_s, \\ &= \alpha(\gamma_t * g), \end{aligned} \tag{23}$$

where

$$(\gamma_t * g)(s) = \int d\xi\, g(\xi) \gamma_t(\xi, s). \tag{24}$$

If g is taken to be the characteristic function for some Borel set, then (22) shows that (24) defines a measure preserving map. Invoking continuity in t, we obtain a flow, and we can again trace the route from commutator to Poisson bracket without invoking the sharp-point limit.

In the other direction, if we begin with a real $(L_1 \cup L_\infty)$ function g, use (18), and then exponentiate to define a unitary group on the algebra of (stochastic) classical observables, and then quantize via (7), we obtain a strongly continuous automorphism of the algebra of stochastic quantum observables which is accordingly unitarily implemented. Again the connection between Lie commutator for the generators and Poisson bracket may be made.

In summary we have shown the following relations among algebras of observables: $\{\alpha(g), *_\hbar, [\,,\,]\}$ = stochastic quantum observable algebra is identified with $\{g, \circ_\hbar, [\,,]\}$ = stochastic classical observable algebra which, in the sharp-point limit, becomes $\{g, \cdot, [\,,\,]\}$ = usual classical observable algebra, where in all three cases, the commutator is to have physical meaning at least for those observables which are associated to elements of the Lie algebra of the space-time symmetry group. In the last two stages, the relation (19) makes the connection between commutator and Poisson bracket, so that the commutator between the v_g's may be replaced with the Poisson bracket on the g's.

We close this section by investigating the states on these algebras.

Let us write the quantum state ρ, in view of (20), as

$$\rho = \int dz\, \mu(z) T_z. \tag{25}$$

Since ρ is positive, μ may be taken positive, and since ρ is of trace one,

$$\int dz\, \mu(z) = 1, \tag{26}$$

so μ is an ordinary probability density. It has previously been thought of as the classical density corresponding to ρ, which, however, is quite different from the classical density defined in (9). From (9) and Appendix A,

$$\begin{aligned} \rho(s) &= \lambda^{-1} \operatorname{Tr}(\rho T_s) \\ &= \int dz\, \mu(z) \lambda^{-1} \operatorname{Tr}(T_z T_s) \\ &= \int dz\, \mu(z) \lambda^{-1} \beta(z, s). \end{aligned} \tag{27}$$

In the sharp-point limit we then have

$$\rho(s) \to \mu(s) \ \text{a.e.},$$

so μ is the classical limit of the stochastic classical density. μ is indeed the classical density corresponding to ρ.

We remark that the procedure

$$\rho \to \lambda^{-1} \operatorname{Tr}(\rho T_s) \underset{\text{sharp point}}{\to} \mu(s)$$

is a program for finding the representation (25). This procedure extends to any operator that is trace class and of form (25).

Finally, we may define the stochastic classical Shannon entropy by

$$\int \rho(z) \ln \rho(z) dz,$$

which in the sharp-point limit, and using the continuity property for entropy,[4] becomes the classical Shannon entropy

$$\int \mu(z) \ln \mu(z) dz.$$

This clarifies the distinction between these two quantities.[4]

III. STOCHASTIC DEQUANTIZATION IN THE MINIMUM UNCERTAINTY MODEL

Lemma: $\{(\pi\alpha)^{-1/2} \exp\{-x^2/\alpha\}\}, \alpha > 0, \alpha \to 0^+$ forms a δ-sequence on the space of bounded continuous functions over the reals.

Proof: Let $K_\alpha(x) \equiv (\pi\alpha)^{-1/2} \exp\{-x^2/\alpha\}$. Then (a) $\int_{-\infty}^{\infty} K_\alpha(x) dx = 1$ for all $\alpha > 0$; (b) $K_\alpha(x)$ is continuous in x on all of $\mathbf{R}$, for all $\alpha > 0$; (c) $K_\alpha(x) > 0$ for all $x \in \mathbf{R}$, $\alpha > 0$; (d) Let $\delta > 0$. Then for all $|x| > \delta$, $0 < K_\alpha(x) < K_\alpha(\delta)$ and $K_\alpha(\delta) \to 0$ as $\alpha \to 0^+$ by L'Hôpital's rule. Hence, $K_\alpha(x) \to 0$ uniformly on $|x| > \delta$. The rest follows from standard analysis. (See Appendix B.)

For the minimum uncertainty case, from Ref. 8, Eq. 6,

$$\begin{aligned} \beta(z, s) &= \operatorname{Tr}(T_z T_s T_z) \\ &= \exp\left\{-\frac{c}{2\hbar}\left[(z_2 - s_2)^2 + \left(\frac{z_1 - s_1}{c^2}\right)^2\right]\right\}. \end{aligned}$$

But, by the Lemma

$$\left(\frac{2\pi\hbar}{c}\right)^{-1/2}\exp\left\{-\frac{c}{2\hbar}z_2^2\right\}(2\pi\hbar c)^{-1/2}\exp\left\{-\frac{1}{2\hbar c}(z_1)^2\right\}$$
$$=\frac{1}{2\pi\hbar}\beta(z,0)=\frac{1}{\lambda}\beta(z,0)$$

is a δ-sequence as $\hbar/c$, $\hbar c\to 0$. We shall call this "$\hbar\to 0$." But for the minimum uncertainty model, $\hbar/c = 2\operatorname{Var} Q$, $\hbar c = 2\times\operatorname{Var} P$ so the limit of interest is in fact when $\operatorname{Var} Q$, $\operatorname{Var} P\to 0$, i.e., the sharp-point limit (limit of precision in the stochasticity). Since $\operatorname{Var} Q \operatorname{Var} P = \hbar^2/4$, the limit $\hbar\to 0$ already requires some sort of loss of stochasticity. For the general stochastic case, we expect that the sharp-point limit (which implies $\hbar\to 0$ by the uncertainty inequality) is the proper limit for dequantization.

The explicit dependence of the stochastic quantum observables and the stochastic classical states is obtained as follows: Let

$$P'' = (\operatorname{Var} P)^{-1/2}P = \sqrt{(2/\hbar c)}\,P,$$
$$Q'' = (\operatorname{Var} Q)^{-1/2}Q = \sqrt{(2c/\hbar)}Q.$$

Then $[P'', Q''] = -(i/2)\mathbf{1}$ which is $\hbar$ invariant. Also $W_\hbar(x,y) = \exp\{ixP + iyQ\} = \exp\{ix''P'' + iy''Q''\}$, where $x'' = \sqrt{(\hbar c/2)}x$, $y'' = \sqrt{(\hbar/2c)}y$.

Now from Ref. 8, Eq. (2),

$$T_{a,b} = \frac{\hbar}{2\pi}\int_{\mathbf{R}^2} dx\,dy$$
$$\times\exp\left\{-\frac{1}{4}\hbar c\left(x^2+\frac{y^2}{c^2}\right) - ibx - i\frac{a}{c}y\right\}W_\hbar(x,y)$$
$$=\frac{\hbar}{2\pi}\int_{\mathbf{R}^2} dx''\,dy''\left(\frac{2}{\hbar}\right)$$
$$\times\exp\left\{-\frac{1}{2}(x''^2+y''^2)\right.$$
$$\left.-ib\sqrt{\frac{2}{\hbar c}}x'' - i\frac{a}{c}\sqrt{\frac{2c}{\hbar}}y''\right\}$$
$$\times\exp\{ix''P''+iy''Q''\}.$$

Since $b\sqrt{2/\hbar c} = b(\operatorname{Var} P)^{-1/2}$ = expected value of P'' in state $T_{a,b}$ and $(a/c)\sqrt{2c/\hbar} = (a/c)(\operatorname{Var} Q)^{-1/2}$ = expected value of Q'' in state $T_{a,b}$, then, $T_{a,b}$ is $\hbar$ invariant for momentum and position expressed in normalized units. Consequently, the $\lambda\,\rho(z)$, $a(f)$ are also $\hbar$ invariant. Because dz/λ is an $\hbar$ invariant measure on $\mathbf{R}^2$, we now have $1 = \int(dz/\lambda)[\lambda\,\rho(z)]$ as an $\hbar$ invariant expression of the fact that $\lambda\,\rho(z)$ is the classical density. Furthermore, as $\hbar\to 0$ the $\mathscr{W}_\hbar$ becomes commutative ($\chi_\hbar(z,s)\to 1$) so the Weyl algebra becomes "classical" also.

The connection between commutator and Poisson bracket may be verified as follows.

The position operator Q_i generates boosts so that in the passage $U_g\to V_g$ we have

$$Q_i\to -\frac{\partial}{\partial p_i} = \sum_j\left|\frac{\partial g}{\partial p_i}\frac{\partial}{\partial q_i} - \frac{\partial g}{\partial q_i}\frac{\partial}{\partial p_i}\right|$$

for $g(z) = q_i + c$. Then we have, by a known result,[8]

$$a(g) = \int dz(q_i + c)T_z = Q_i + c'\mathbf{1}$$

which is entirely consistent. Similarly for the momentum operators. Now $\{q_i + c_1, p_j + c_2\} = \delta_{ij} = [Q_i, P_j] = [a(q_i + c_1), a(p_j + c_2)]$, and we are done.

Note added in proof: In (21) since the left-hand side is again a projection, $\gamma_t(\xi,s)$ must be a delta function at some point s_t. We then obtain "stable dynamics of generalized coherent states."[12]

ACKNOWLEDGMENTS

I would like to thank Dr. J. Brooke and Dr. G. G. Emch for helpful discussions in the development of this work.

APPENDIX A: INTERCHANGE OF TRACE AND INTEGRAL

We justify here the interchange of trace and integral occurring in formulas (10), (11), (13), (26), and (27). It is sufficient to prove that

$$\operatorname{Tr}\int h(z)\rho T_z\,dz = \int h(z)\operatorname{Tr}(\rho T_z)dz$$

for h a non-negative Lebesgue measurable function. Since ρ and T_z are both positive, bounded, traceclass, then in any basis $\{\psi_n\}$ for the representing Hilbert space, H, and η_z a unit vector in T_zH, we have

$$f(z)\equiv\operatorname{Tr}(\rho T_z) = \operatorname{Tr}(\rho^{1/2}T_z\rho^{1/2})$$
$$=\sum_{n=1}^{\infty}\langle\psi_n, \rho^{1/2}T_z\rho^{1/2}\psi_n\rangle$$
$$=\sum_{n=1}^{\infty}\langle\psi_n,\rho^{1/2}\eta_z\rangle\langle\rho^{1/2}\eta_z,\psi_n\rangle.$$

Setting $f_n(z) = \langle\psi_n,\rho^{1/2}\eta_z\rangle\langle\rho^{1/2}\eta_z,\psi_n\rangle$, then $f_n(z)$ is a non-negative continuous (hence, measurable) function of z and $f(z)\equiv\sum_n f_n(z)$ converges for all z; similarly for hf_n. But by a standard corollary of Lebesgue's monotone convergence theorem[11]

$$\int h(z)\operatorname{Tr}(\rho T_z)dz = \int h(z)f(z)dz$$
$$=\sum_{n=1}^{\infty}\int h(z)f_n(z)dz$$
$$=\sum_{n=1}^{\infty}\langle\psi_n, \int h(z)\rho^{1/2}T_z\rho^{1/2}\,dz\,\psi_n\rangle$$
$$=\operatorname{Tr}\int h(z)\rho^{1/2}T_z\rho^{1/2}\,dz$$
$$=\operatorname{Tr}\int h(z)\rho T_z\,dz.$$

APPENDIX B: DELTA NETS

Let $\{K_\alpha(x)\,|\,x\in\mathbf{R},\ \alpha\in$ subset of $(0,\infty)$ clustering at $0\}$ satisfy (a) $\int_{\mathbf{R}}K_\alpha = 1$, for all α; (b) $K_\alpha(x) > 0$ for all α, all $x\in\mathbf{R}$; (c) for any $\delta > 0$, then on $|x|\geqslant\delta$, $K_\alpha(x)\to 0$ uniformly as $\alpha\to 0^+$.

Then for any bounded, measurable function f on $\mathbf{R}$, continuous at zero,

$$\lim_{\alpha\to 0^+}\int K_\alpha f = f(0).$$

Proof: Since K_α, f are measurable, $K_\alpha f$ is measurable. Since f is bounded, K_α integrable, then the integral $\int K_\alpha f$ exists.

Let $\epsilon > 0$. f continuous at zero implies there exists $\delta > 0$ such that $|f(x) - f(0)| < \epsilon/3$ for $|x| \leqslant \delta$. f is bounded implies there exists $M > 0$ such that $|f(x) - f(0)| < M$ for all $x \in \mathbb{R}$. (a) implies there exists $d > 0$ such that $\int_{|x|>d} K_\alpha < \epsilon/3M$. Without loss of generality we may choose $d > \delta$. By (d), given d, δ, M there exists N such that $\alpha \leqslant N$ implies $|K_\alpha(x)| < \epsilon/6(d-\delta)M$ for $|x| \geqslant \delta$. Thus

$$\begin{aligned}\left|\int_{\mathbb{R}} K_\alpha f - f(0)\right| &= \left|\int_{\mathbb{R}} K_\alpha (f - f(0))\right| \\ &\leqslant \int_{|x|>d} + \int_{d>|x|>\delta} + \int_{|x|<\delta} K_\alpha |f - f(0)| \\ &\leqslant M\frac{\epsilon}{3M} + M\frac{\epsilon}{6(d-\delta)M}\int_{d>|x|>\delta} 1 \\ &\quad + \frac{\epsilon}{3}\int_{|x|<\delta} K_\alpha \leqslant \epsilon\end{aligned}$$

using

$$\int_{|x|<\delta} K_\alpha \leqslant \int_{\mathbb{R}} K_\alpha = 1.$$

[1]G. G. Emch, Int. J. Theor. Phys. **22**, 397–420 (1983).
[2]F. E. Schroeck, Jr., Found. Phys. **12**, 825–841 (1982).
[3]E. B. Davies, J. Funct. Anal. **6**, 318–346 (1970); E. B. Davies and J. T. Lewis, Commun. Math. Phys. **17**, 239–260 (1970).
[4]F. E. Schroeck, Jr., "On the Non-Occurrence of Two Paradoxes in the Measurement Scheme of Stochastic Quantum Mechanics," Found. Phys. (in press, 1985).
[5]G. G. Emch, *Algebraic Methods in Statistical Mechanics and Quantum Field Theory* (Wiley-Interscience, New York, 1972).
[6]W. Guz, "On the Non-Classical Character of the Phase-Space Representations of Quantum Mechanics," Found. Phys. (in press, Feb. 1985).
[7]J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton U. P., Princeton, NJ, 1955).
[8]F. E. Schroeck, Jr., J. Math. Phys. **22**, 2562–2572 (1981).
[9]E. Prugovecki, *Stochastic Quantum Mechanics and Quantum Spacetime* (Reidel, Dordrecht, 1984).
[10]F. E. Schroeck, Jr., Found. Phys. **12**, 479–497 (1982).
[11]W. Rudin, *Real and Complex Analysis* (McGraw-Hill, New York, 1974), p. 23.
[12]B. A. Nikolov and D. A. Trifonov, Comm. JINR, E2-81-797 and E2-81-798, 1981.

Inverse scattering by a local impurity in a periodic potential in one dimension. II

Roger G. Newton
Physics Department, Indiana University, Bloomington, Indiana 47405

(Received 25 April 1984; accepted for publication 17 August 1984)

This paper continues and completes the solution to the inverse scattering problem initiated in a recent paper. It allows for the existence of bound states in the band gaps and corrects a number of errors in the first paper.

I. INTRODUCTION

In a recent paper[1] this author studied the inverse scattering problem in one dimension for Hill's equation modified by a nonperiodic potential that tends to zero as $|x|\to\infty$. The nonperiodic potential U was reconstructed from a knowledge of the periodic one, and of the transmission and reflection amplitudes for Bloch waves scattered by U; also sufficient conditions for the existence of a potential U were presented. However, both results required the assumption that there be no eigenvalues (bound states) in the band gaps, i.e., that all eigenvalues (if any) be situated below the continuous spectrum. Since under rather general conditions on U [such as $\int dx\, |U|(1+x^2)<\infty$ together with sgn $U=$ const] there necessarily are eigenvalues in the gaps, in fact infinitely many of them,[2] this restriction severely limits the applicability of the results of Ref. 1. We remove the restriction in the present paper.

In addition to making the restrictive assumption that there be no bound states in the gaps, Ref. 1 did not handle the periodic spectrum correctly. Contrary to the assertion of Lemma 2, the Jost function J is not continuous at the periodic spectrum, i.e., at the points where $\sin k=0$; it behaves as $\csc k$ there, i.e., as $(\lambda-\lambda_n)^{-1/2}$ near $\lambda=\lambda_n$. The resulting zeros of $\det J^{-1}$ at these points have to be removed, just as the poles of J^{-1} at the bound states must, before the Riemann–Hilbert problem can be solved. Since the number of singularities on the real axis that must thus be removed is generally infinite two additional problems arise: The needed factor function which is defined by an infinite product, must be proved to converge, and the reduced S matrix must be shown to differ from 1 by a function in $L^2(\mathbb{R})$. This we also do in the present paper. Finally, we correct some other errors and misprints of Ref. 1, some of which also owe their origin to an incorrect handling of the singularities at the periodic spectrum, including the statement of Levinson's theorem.

In Sec. III we reformulate the Riemann–Hilbert problem that arises here in a more suitable form. In Sec. IV the reducing product that isolates the singularities is defined and proved to converge. Section V deals with the needed asymptotics, both in $\mathbb{C}^+$ and on the real axis. Section VI solves the reduced problem and relates the solution to the potential. The corrections to Ref. 1 are contained in the Appendix. The equation numbers in this paper carry no prefixes; all references to equation numbers with prefixes are to Ref. 1. We shall freely use the notation of Ref. 1.

II. THE RIEMANN–HILBERT PROBLEM

The Jost function defined by (4.24) satisfies (4.27) and (4.27′), which we shall write as[3]

$$J^{\#-1}=\omega Q J^{-1} Q, \tag{1}$$

where

$$\omega=\begin{cases} Q\hat{S}^* Q, & \text{when } \lambda\in\mathbb{R}'_a,\\ Q S_0^* Q, & \text{when } \lambda\in\mathbb{R}\setminus\mathbb{R}'_a.\end{cases} \tag{2}$$

If J_x is the Jost function for a "comprehensively shifted" problem, for which $\hat{S}_x$ and S_{0x} are defined by (4.19) and (4.19′), then the solution $\hat{\psi}$ of (4.1) is given by $\hat{\psi}=X\Psi$ and $\Psi=J_x^{-1}\hat{1}$.[4] The task of removing the bound-state poles from J_x^{-1} was performed in Sec. 5 A of Ref. 1. We now proceed differently, that is, we replace the procedure given in the second half of Sec. 5 A by the following.

As a first step, let J_0 be the Jost function of the one-cell potential $\hat{V}$, so that by (2.15)

$$J_0^{\#-1}=Q S_0 J_0^{-1} Q. \tag{3}$$

Since $V\equiv 0$ for $x<0$, J_0 can be explicitly constructed,

$$J_0=\frac{1}{1+\eta_3}\begin{pmatrix}1+\eta_3 & -\eta_1\\ 0 & 1\end{pmatrix},$$

and for $x<0$ we have

$$J_{0x}=\frac{1}{1+\eta_3}\begin{pmatrix}1+\eta_3 & -\eta_1 e^{-2i\lambda x}\\ 0 & 1\end{pmatrix}.$$

But for $x>0$, J_{0x} cannot in general be explicitly given.

Then define

$$F=J_{0x}J_x^{-1}. \tag{4}$$

By (1), (2), and (3), for $\lambda\in\mathbb{R}$, F satisfies the relation

$$F^{\#}=\Omega Q F Q, \tag{5}$$

where by (2.15) and (4.18)

$$\Omega=\begin{cases} QJ_{0x}M_x^{-1}S_x^* M_x J_{0x}^{-1}Q, & \lambda\in\mathbb{R}'_a,\\ 1, & \lambda\notin\mathbb{R}'_a,\end{cases} \tag{6}$$

and M_x denotes M corresponding to a comprehensively shifted problem. We assume that J_{0x} is known; therefore a determination of F leads to $J_x=F^{-1}J_{0x}$ and $\hat{\psi}=XJ_{0x}^{-1}F\,\hat{1}$.

The asymptotic and analytic properties of F are those it inherits from J_{ox} and J_x^{-1}. It follows from Lemmas 1 and 2 of Ref. 1, as corrected in the Appendix, that $F(\lambda)$ has an analytic continuation into $\mathbb{C}^+\cup\mathbb{R}\setminus\mathbb{R}'_a$ that is meromorphic there, with simple poles at those points $\lambda=\lambda_b$, $b=1,\dots,N$, that are the square roots of the eigenvalues $E_b=\lambda_b^2$ of $-d^2/dx^2+V+U$, and that its asymptotic form for $|\lambda|\to\infty$ in $\mathbb{C}^+$ is given by $F(\lambda)=1+O(1/\lambda)$.

Since $\det F=\det J_{ox}/\det J_x=T$ by (4.26), the function F has zeros nowhere except at the periodic spectrum. (We shall refer to a point at which the determinant of a matrix-valued function vanishes as a zero of the function.) In

 0022-2488/85/020311-06$02.50

particular, it does not have zeros at the bound states of $\hat{V}$, where J_{0x} has zeros. However, one of the difficulties that we have to deal with is the fact that on the real λ axis F has poles in the gaps at the bound states, and in addition it has zeros at all the gap end points. Therefore, certainly along the real axis $F \not\to \mathbb{1}$.

The following expression is easily derived from (4.24), (4.16), (4.11), and (4.5):

$$J^{-1} = (T/2i\lambda)M^{-1}W(\chi,\tilde{\phi})P. \tag{7}$$

Since generically TM^{-1} is bounded at the gap ends, it follows from (3.9) that at the gap ends, where $\sin k = 0$ and $T = 0$, the ranges of the limits of J_x^{-1}, both from above and below, are orthogonal to the vectors $(\Delta - l_x^* e^{i\lambda}, l_x \Delta - e^{i\lambda})$. Therefore, the ranges of F at its zeros are known and so are the ranges of its residues at the bound-state poles. Let the ray s_b be the character of the bound state at E_b, as defined in Sec. 5 of Ref. 1. Then the ray $\mathscr{H}_b := J_{0x}(\kappa_b)X(-\kappa_b)M_x^{-1}(\kappa_b)s_b$ is the range of the residue of F at $\lambda = \kappa_b$. It follows from (5) and (6) that if $\kappa_b > 0$, $\kappa_b \in \mathbb{R}\setminus\mathbb{R}'_a$, then F has a pole also at $\lambda = -\kappa_b$, and its residue there has the range $\bar{\mathscr{H}}_b = Q\mathscr{H}_b$. An analogous relation holds for the positive and negative zeros of F.

Our aim now is to formulate the Riemann–Hilbert problem in such a way that its solution, if it exists, is unique, in spite of the fact that the solution F will not approach $\mathbb{1}$ as $\lambda \to \pm\infty$, nor will $F - \mathbb{1}$ be in L^2 if there are bound states in the gaps. We do that as follows.

Problem $\mathfrak{H}$: Given a sequence $0 < \lambda_1 \leqslant \lambda_2 < \lambda_3 \leqslant \lambda_4 < \cdots$ and a countable set κ_m such that either $\kappa_m = i|\kappa_m|$ (for a finite number of κ_m) or $\lambda_{2n-1} < \kappa_m < \lambda_{2n}$; for each λ_n and κ_m there is a given one-dimensional space $\mathscr{H}_n$ or $\mathscr{H}'_m$, respectively. Also given is a 2×2-matrix-valued function $\Omega(\lambda)$, $\lambda \in \mathbb{R}$, such that $\Omega(-\lambda) = \Omega^*(\lambda)$, $\Omega(-\lambda)^{-1} = Q\Omega(\lambda)Q$, $\Omega - \mathbb{1} \in L^2(\mathbb{R})$, and Ω is analytic in the intervals $\lambda_{2n-1} < \lambda < \lambda_{2n}$. Find a 2×2-matrix-valued function $F(\lambda)$ that is meromorphic and free of zeros in $\mathbb{C}^+$ and in the intervals $(\lambda_{2n-1}, \lambda_{2n})$ with simple poles at the points κ_m and $-\kappa_m^*$ such that the range of the residue at κ_m equals $\mathscr{H}'_m$, the range of the residue of $-\kappa_m$, for real κ_m, equals $Q\mathscr{H}'_m$, and whose limit on the real axis satisfies (5). Furthermore, at the points λ_n, $n = 1,2,\ldots$, $\det F(\lambda_n) = 0$ and $\operatorname{ran} F(\lambda_n) = \mathscr{H}_n$ in such a way that $F^{-1}(\lambda)(\lambda - \lambda_n)^{1/2}$ remains bounded and $\neq 0$ as $\lambda \to \lambda_n$. Elsewhere in each of the intervals $(\lambda_{2n}, \lambda_{2n+1})$ we require that both F and F^{-1} be in L^2. Finally, it is required that

$$\lim_{|\lambda|\to\infty} F(|\lambda|e^{i\theta}) = \mathbb{1} \tag{8}$$

for every θ in the open interval $(0,\pi)$, and that there exists a real sequence $\lambda^{(n)} \to \infty$ such that

$$\lim_{n\to\infty} F(\lambda^{(n)}) = \lim_{n\to\infty} F(-\lambda^{(n)}) = \mathbb{1}. \tag{9}$$

Lemma: If $F(\lambda)$ solves the problem $\mathfrak{H}$ then it is the only solution.

Proof: Suppose F_1 and F_2 both solve the problem $\mathfrak{H}$. Then on $\mathbb{R}$

$$\Omega(\lambda) = F_1(-\lambda)QF_1(\lambda)^{-1}Q = F_2(-\lambda)QF_2(\lambda)^{-1}Q$$

and hence,

$$G(\lambda) := F_2(-\lambda)^{-1}F_1(-\lambda) = QF_2(\lambda)^{-1}F_1(\lambda)Q.$$

The first expression has an analytic continuation into $\mathbb{C}^-$ that is holomorphic there, and the second expression is holomorphic in $\mathbb{C}^+$. (The poles of F_1 at $\lambda = \pm\kappa_m$ there are annihilated by the zeros of F_2^{-1}.) So G is meromorphic in $\mathbb{C}$, with poles at most on the real axis. But again, the poles of F_1 are annihilated by the zeros of F_2^{-1} and the poles of F_2^{-1} are annihilated by the zeros of F_1. Hence, G is entire. Furthermore,

$$\lim_{|\lambda|\to\infty} [G(|\lambda|e^{i\theta}) - \mathbb{1}] = 0$$

for all θ in $0 < \theta < \pi$ and $\pi < \theta < 2\pi$. We now use the Phragmén–Lindelöf theorem[5] to conclude that $G - \mathbb{1}$ must be uniformly bounded for all λ. Consequently, by Liouville's theorem $G = \mathbb{1}$. Q.E.D.

Our task now is to remove the zeros and poles from F and solve a reduced Riemann–Hilbert problem whose solution is to have neither poles nor zeros. In view of Lemma 1 it will not matter by what technique this is accomplished, so long as the final result is a solution of $\mathfrak{H}$. Specifically, it will be of no consequence if we add requirements for the solution of the reduced problem whose necessity we do not prove; if such a solution exists and it leads to a solution of $\mathfrak{H}$ then it must be the only one.

The technique for removing the zeros and poles from F is essentially that of Ref. 1. However, because there are infinitely many zeros and there may be infinitely many poles, we are now confronted with two additional problems. The infinite product of matrices that has to be formed must be shown (a) to converge, and (b) to approach $\mathbb{1}$ at infinity in a suitable sense. The second of these problems is the less trivial one. We first turn to the formation of the product.

III. THE REDUCING PRODUCT

If there are N bound states of negative energy $E_b = \kappa_b^2$, $\kappa_b = i|\kappa_b|$, $b = 1,\ldots,N$, we define N orthogonal projections $B_b = B_b^\dagger = B_b^2$ successively as follows:

$$\Gamma_b(\lambda) := \mathbb{1} - B_b + B_b(\lambda + \kappa_b)/(\lambda - \kappa_b),$$

$$C_b := \Gamma_1(\kappa_b)\cdots\Gamma_{b-1}(\kappa_b), \quad C_1 = \mathbb{1},$$

$$(\mathbb{1} - B_b)C_b^{-1}\mathscr{H}_b = 0.$$

Also define

$$\Pi_B := \Gamma_1\cdots\Gamma_N.$$

The next step is to form an analogous product that isolates the bound-state poles in each band gap and the zeros at the gap ends (the periodic spectrum). We do that simultaneously for the positive and negative gaps. Assuming that the nth gap, which stretches from λ_{2n-1} to λ_{2n}, contains N_n bound states, we define

$$\Gamma_n^{(1)}(\lambda) := \mathbb{1} - B_n^{(1)} + B_n^{(1)}\left(\frac{\lambda - \lambda_{2n-1}}{\lambda - \lambda_{2n-1} + i\epsilon_n}\right)^{1/2},$$

$$\Gamma_n^{(2)}(\lambda) := \mathbb{1} - B_n^{(2)} + B_n^{(2)}\left(\frac{\lambda + \lambda_{2n-1}}{\lambda + \lambda_{2n-1} + i\epsilon_n}\right)^{1/2},$$

$$\Gamma_n^{(3)}(\lambda) := \mathbb{1} - B_n^{(3)} + B_n^{(3)}\left(\frac{\lambda - \lambda_{2n}}{\lambda - \lambda_{2n} + i\epsilon_n}\right)^{1/2},$$

$$\Gamma_n^{(4)}(\lambda) := 1 - B_n^{(4)} + B_n^{(4)}\left(\frac{\lambda + \lambda_{2n}}{\lambda + \lambda_{2n} + i\epsilon_n}\right)^{1/2},$$

$$\Gamma_n^{(i)}(\lambda) := 1 - B_n^{(i)} + B_n^{(i)}\left(\frac{\lambda - \kappa_n^{(i)} + i\epsilon_n}{\lambda - \kappa_n^{(i)}}\right),$$

$$i = 5,\ldots,N_n + 4,$$

$$\Gamma_n^{(i)}(\lambda) := 1 - B_n^{(i)} + B_n^{(i)}\left(\frac{\lambda + \kappa_n^{(i)} + i\epsilon_n}{\lambda + \kappa_n^{(i)}}\right),$$

$$i = N_n + 5,\ldots,2N_n + 4.$$

The orthogonal projections $B_n^{(i)} = B_n^{(i)\dagger} = B_n^{(i)2}$ are defined by the equations

$$\Gamma_n(\lambda) := \Gamma_n^{(1)}(\lambda)\cdots\Gamma_n^{(2N_n+4)}(\lambda),$$

$$\Pi_n(\lambda) := \Pi_B(\lambda)\Gamma_1(\lambda)\cdots\Gamma_n(\lambda),$$

$$C_n^{(1)} := \Pi_{n-1}(\lambda_{2n-1}),$$

$$C_n^{(2)} := \Pi_{n-1}(-\lambda_{2n-1})\Gamma_n^{(1)}(-\lambda_{2n-1}),$$

$$C_n^{(3)} := \Pi_{n-1}(\lambda_{2n})\Gamma_n^{(1)}(\lambda_{2n})\Gamma_n^{(2)}(\lambda_{2n}),$$

$$C_n^{(4)} := \Pi_{n-1}(-\lambda_{2n})\Gamma_n^{(1)}(-\lambda_{2n})\Gamma_n^{(2)}(-\lambda_{2n})\Gamma_n^{(3)}(-\lambda_{2n}),$$

$$C_n^{(5)} := \Pi_{n-1}(\kappa_n^{(1)})\Gamma_n^{(1)}(\kappa_n^{(1)})\cdots\Gamma_n^{(4)}(\kappa_n^{(1)}),$$

$$C_n^{(N_n+5)} := \Pi_{n-1}(-\kappa_n^{(1)})\Gamma_n^{(1)}(-\kappa_n^{(1)})\cdots\Gamma_n^{(N_n+4)}(-\kappa_n^{(1)}),$$

$$B_n^{(1)}C_n^{(1)-1}\mathscr{H}'_{2n-1} = B_n^{(2)}C_n^{(2)-1}Q\mathscr{H}'_{2n-1} = 0,$$

$$B_n^{(3)}C_n^{(3)-1}\mathscr{H}'_{2n} = B_n^{(4)}C_n^{(4)-1}Q\mathscr{H}'_{2n} = 0,$$

$$(1 - B_n^{(5)})C_n^{(5)-1}\mathscr{H}_n^{(1)} = 0,\ldots,$$

$$(1 - B_n^{(N_n+5)})C_n^{(N_n+5)-1}Q\mathscr{H}_n^{(1)} = 0,\ldots .$$

Also, $\epsilon_n > 0$, to be specified below such that $\epsilon_n \to 0$. These equations recursively define the product $\Pi_n(\lambda)$ for all n.

In order to prove that it converges in the operator norm $\|\cdot\|$ for each fixed value of λ we need only prove that the series

$$\sum_{i,n}\|\Gamma_n^{(i)}(\lambda) - 1\|$$

converges. But

$$\Gamma_n^{(1)}(\lambda) - 1 = B_n^{(1)}\left[\left(1 - \frac{i\epsilon_n}{\lambda - \lambda_{2n-1} + i\epsilon_n}\right)^{1/2} - 1\right].$$

Since $B_n^{(1)}$ is an orthogonal projection, $\|B_n^{(1)}\| = 1$, and it is known[6] that $\lambda_{2n-1} \sim \lambda_{2n} \sim \pi n$, therefore asymptotically for large n

$$\|\Gamma_n^{(1)}(\lambda) - 1\| \sim \epsilon_n/\pi n$$

and similarly for $\Gamma_n^{(i)}$, $i = 2,\ldots, N_n + 5$. If $(1 + |x|)U \in L^1$ it is known[7] that for sufficiently large n, $N_n < 2$. Consequently, if the ϵ_n are chosen sufficiently small, for example, $\epsilon_n = 1/n^\delta$, $\delta > 0$, then the product $\Pi_n(\lambda)$ converges pointwise in the operator norm and we may define

$$\Pi(\lambda) = \lim_{n\to\infty}\prod_n(\lambda) \tag{10}$$

in that sense.

We now form

$$F^{\text{red}}(\lambda) = \Pi^{-1}(\lambda)F(\lambda). \tag{11}$$

It is easily checked that Π has been so defined that $F^{\text{red}}(\lambda)$ is free of poles and zeros in $\mathbb{C}^+ \cup \mathbf{R}\setminus\mathbf{R}'_a$ and that in the allowed bands, including their end points, both F^{red} and $(F^{\text{red}})^{-1}$ are bounded, provided that F is differentiable as a function of λ. By (7) this will follow if χ is differentiable, which in turn follows[8] if $|x|^2U \in L^1(\mathbf{R})$. We shall therefore assume that this is the case.

IV. ASYMPTOTICS

A. Complex λ

As $|\lambda| \to \infty$ in $\mathbb{C}^+$, in the sense that $\lambda = |\lambda|e^{i\varphi}$, $0 < \varphi < \pi$, we have

$$\lim \Pi(\lambda) = 1. \tag{12}$$

This is proved by noting that

$$\|\Gamma_n^{(1)}(\lambda) - 1\| - \left\|B_n^{(1)}\left[\left(1 - \frac{i\epsilon_n}{\lambda - \lambda_n + i\epsilon_n}\right)^{1/2} - 1\right]\right\|$$

$$< \frac{\epsilon_n}{2|\lambda|\sin\varphi}$$

and similarly for $\Gamma_n^{(i)}$, $i = 2,\ldots, 2N_n + 4$. Therefore, for sufficiently large $|\lambda|$

$$\Pi(\lambda) - 1 = \exp[\ln \Pi(\lambda)] - 1 = \exp\sum_{i,n}\ln\Gamma_n^{(i)} - 1,$$

which converges and approaches zero as $1/|\lambda|\sin\varphi$ if $\epsilon_n < n^{-1-\epsilon}$, $\epsilon > 0$. [Since the matrices $\Gamma_n^{(i)}$ are close to 1, their logarithms are well defined.] Thus for $0 < \varphi < \pi$,

$$\lim_{|\lambda|\to\infty}\Pi(|\lambda|e^{i\varphi}) = \lim_{|\lambda|\to\infty}[\Pi(|\lambda|e^{i\varphi})]^{-1} = 1.$$

Consequently, $F^{\text{red}}(\lambda)$ has the same property because $F(\lambda)$ has it.

B. Points in the allowed bands

By the same argument as above, because for every $0 < l < 1$ $\|\Gamma_n^{(1)}(\pi(n + l)) - 1\| < \epsilon_n/\pi l$, etc. for sufficiently large n, it follows that $\Pi(\pi(n + l))$, $\Pi(-\pi(n + l))$, $[\Pi(\pi(n + l))]^{-1}$, $[\Pi(-\pi(n + l))]^{-1} = 1 + O(1/n)$.

C. In the band gaps

If follows from (5) and (6) that for $\lambda \in \mathbf{R}$ the function $F^{\text{red}}(\lambda)$ satisfies the equation

$$F^{\text{red}\#} = \Omega^{\text{red}}QF^{\text{red}}Q, \tag{13}$$

where

$$\Omega^{\text{red}}(\lambda) = \Pi^{-1}(-\lambda)\Omega(\lambda)Q\Pi(\lambda)Q. \tag{14}$$

We must now study the behavior of Ω^{red} for large $|\lambda|$ along the real axis.

For a given λ in the nth gap, $\lambda_{2n-1} < \lambda < \lambda_{2n}$, $n \gg 1$, let us divide

$$\Omega^{\text{red}}(\lambda) = \Pi^{-1}(-\lambda)Q\Pi(\lambda)Q$$

into five factors:

$$\Omega^{\text{red}} = [\Pi'_n(-\lambda)]^{-1}\Gamma_n^{-1}(-\lambda)\Omega_n(\lambda)Q\Gamma_n(\lambda)\Pi'_n(\lambda)Q,$$

where

$$\Pi'_n = \Pi_n^{-1}\Pi,$$

$$\Omega_n(\lambda) = \Pi_{n-1}^{-1}(-\lambda)Q\Pi_{n-1}(\lambda)Q.$$

Our first observation is that for λ in the nth gap or beyond, $\Omega_n(\lambda)$ approaches 1 as $n\to\infty$. This is proved as follows. We have

$$\|\Gamma_m^{(i)}(\lambda) - \mathbf{1}\| \leqslant \frac{\epsilon_m}{\lambda - \lambda_{2m}}, \quad m \leqslant n-1.$$

The norm of $\Pi_{n-1}(\lambda) - \mathbf{1}$ is determined by that of the sum

$$\sum_{m=1}^{n-1} [\Gamma_m^{(i)}(\lambda) - \mathbf{1}] = \sum_{m=1}^{[n/2]} + \sum_{[n/2]+1}^{n-1}.$$

In the first sum on the right-hand side we replace $\lambda - \lambda_{2n}$ by its smallest value $\lambda_{2n} - \lambda_n \simeq \pi n$ for $n \gg 1$, and in the second sum we replace ϵ_m by its largest value $\epsilon_{n/2}$

$$\left\| \sum_{m=1}^{n-1} (\Gamma_m^{(i)}(\lambda) - \mathbf{1}) \right\| \leqslant \frac{1}{\pi n} \sum_{1}^{[n/2]} \epsilon_m + \epsilon_{n/2} \sum_{[n/2]+1}^{n-1} \frac{1}{2\pi(n-m)}.$$

If we choose $\epsilon_n = n^{-1-\delta}$, $\delta > 0$, then the first term is $O(1/n)$ and the second $O(n^{-1-\delta} \ln n)$. The same holds for each i, where for large enough n, $i \leqslant 8$, and similarly for $\Gamma_m^{(i)-1}$. Consequently we find that for $\lambda > \lambda_{2n+1}$ or $\lambda < -\lambda_{2n+1}$,

$$\Omega_n(\lambda) = \mathbf{1} + O(1/n). \tag{15}$$

A similar argument shows that its derivative decreases at least equally fast

$$\frac{d\Omega_n(\lambda)}{d\lambda} = O\left(\frac{1}{n}\right). \tag{16}$$

Next let us consider

$$\Pi_n'(\lambda) = \Gamma_{n+1}\Gamma_{n+2}\cdots$$

for λ in the nth gap. Again we use, for $m > n$

$$\|\Gamma_m^{(i)}(\lambda) - \mathbf{1}\| \leqslant \frac{\epsilon_m}{\lambda_{2m} - \lambda} \leqslant \frac{\epsilon_m}{2\pi(m-n)} \leqslant \frac{1}{2\pi}\epsilon_m,$$

$$\left\| \sum_{m=n+1}^{\infty} (\Gamma_m^{(i)}(\lambda) - \mathbf{1}) \right\| < \frac{1}{2\pi} \sum_{n+1}^{\infty} \epsilon_m = O(n^{-\epsilon}),$$

if $\epsilon_n = n^{-1-\delta}$. Therefore, for $\lambda_{2n-1} \leqslant \lambda \leqslant \lambda_{2n}$

$$\Pi_n'(\lambda) = \mathbf{1} + O(n^{-\delta}),$$
$$[\Pi_n'(-\lambda)]^{-1} = \mathbf{1} + O(n^{-\delta}). \tag{17}$$

Similarly we find for the derivative

$$\frac{d\Pi_n'(\lambda)}{d\lambda} = O(n^{-\delta}). \tag{18}$$

Next we examine $\Gamma_n(\lambda)$ for λ in the nth gap. Since the gap length decreases and is $O(1/n^2)$ for large n it will suffice to prove that Γ_n is uniformly bounded with respect to n. We first note that because by (13)

$$\Omega^{\text{red}} = F^{\#\text{red}} Q F^{\text{red}-1} Q, \tag{19}$$

and F^{red} and $F^{\text{red}-1}$ are bounded at the bound states and at the gap ends, Ω^{red} must also be bounded there. Therefore, the poles and $(\lambda - \lambda_{2n})^{-1/2}$ terms that appear to be present in $\Pi^{\#-1} Q\Pi Q$ must cancel out. Such cancellation of the pole of

$$\Gamma_n^{(5)}(\lambda) = \mathbf{1} + B_n^{(5)} i\epsilon_n/(\lambda - \kappa_n^{(1)}),$$

for example, can occur by multiplication either by a singular matrix whose product with $B_n^{(5)}$ vanishes, or by a term that is proportional to $(\lambda - \kappa_n^{(1)})$. The first leaves nothing in Ω_{n+1}, and the second leaves the derivative of the multiplying function. Because of (15) to (18) we need to pay attention only to $\Gamma_n^{\#-1} Q \Gamma_n Q$ when $\lambda_{2n-1} \leqslant \lambda \leqslant \lambda_{2n}$ or $-\lambda_{2n} \leqslant \lambda \leqslant \lambda_{2n-1}$. In fact, for positive λ we may for the same reason also replace $\Gamma_n^{(2)}, \Gamma_n^{(1)\#}, \Gamma_n^{(4)}, \Gamma_n^{(3)\#}, \Gamma_n^{(7)}, \Gamma_n^{(8)}, \Gamma_n^{(5)\#}$, and $\Gamma_n^{(6)\#}$ by $\mathbf{1}$, and we need to consider only

$$A_n := \Gamma_n^{(8)\#-1} \Gamma_n^{(7)\#-1} \Gamma_n^{(4)\#-1} \Gamma_n^{(2)\#-1} \times Q\Gamma_n^{(1)} \Gamma_n^{(3)} \Gamma_n^{(5)} \Gamma_n^{(6)} Q.$$

Assume now that

$$\epsilon_m \leqslant \min\{ |\kappa_m^{(1)} - \kappa_n^{(2)}|^4, |\kappa_n^{(1)} - \lambda_{2n}|^4, |\kappa_n^{(2)} - \lambda_{2n}|^4, |\kappa_n^{(1)} - \lambda_{2n-1}|^4, |\kappa_n^{(2)} - \lambda_{2n-1}|^4, (\lambda_{2n} - \lambda_{2n-1})^4 \} \tag{20}$$

(with obvious modifications if the nth gap contains only one bound state or none). The innermost product is

$$\begin{aligned} A_n^{(1)} := \Gamma_n^{(2)\#-1} Q \Gamma_n^{(1)} &= (\mathbf{1} - B_n^{(2)})(\mathbf{1} - B_n^{(1)\cdot}) \\ &+ B_n^{(2)}(\mathbf{1} - B_n^{(1)\cdot}) \left(\frac{\lambda - \lambda_{2n-1} - i\epsilon_n}{\lambda - \lambda_{2n-1}} \right)^{1/2} \\ &+ (\mathbf{1} - B_n^{(2)}) B_n^{(1)\cdot} \left(\frac{\lambda - \lambda_{2n-1}}{\lambda - \lambda_{2n-1} + i\epsilon_n} \right)^{1/2} \\ &+ B_n^{(2)} B_n^{(1)\cdot} \left(\frac{\lambda - \lambda_{2n-1} - i\epsilon_n}{\lambda - \lambda_{2n-1} + i\epsilon_n} \right)^{1/2}, \end{aligned}$$

where $B_n^{(i)\cdot} := Q B_n^{(i)} Q$. The second term on the right must vanish because there can be no $(\lambda - \lambda_{2n-1})^{1/2}$ term. (Note that none of the other factors in A_n has a zero at λ_{2n-1}.) The other three terms are bounded uniformly with respect to n for all λ in the closed nth gap. Furthermore, if ϵ_n is chosen as in (20), the first three derivatives of $A_n^{(1)}$, evaluated at λ_{2n}, $\kappa_n^{(1)}$, or $\kappa_n^{(2)}$, are small compared to one. The next product,

$$A_n^{(2)} = \Gamma_n^{(4)\#-1} A_n^{(1)} Q \Gamma_n^{(3)} Q$$

is of the same structure as $A_n^{(1)}$, except that between any product of projections there appears a factor of $A_n^{(1)}$. This time the second term vanishes, not identically, but only with $A_n^{(1)}$ evaluated at $\lambda = \lambda_{2n}$. Since $A_n^{(1)}$ is bounded uniformly with respect to n, and its first three derivatives at λ_{2n}, $\kappa_n^{(1)}$, and $\kappa_n^{(2)}$ are small, it follows that $A_n^{(2)}$ too is bounded uniformly with respect to n for all λ in the closed nth gap, and its first two derivatives at $\kappa_n^{(1)}$ and $\kappa_n^{(2)}$ are small compared to one. This reasoning is repeated twice more and we finally conclude that A_n is uniformly bounded with respect to n for all λ in the nth gap $[\lambda_{2n-1}, \lambda_{2n}]$. The same, of course, holds in $[-\lambda_{2n}, -\lambda_{2n-1}]$. Consequently $\Omega^{\text{red}}(\lambda)$ is uniformly bounded for all λ in gaps, including their endpoints. As a result

$$\int_{\mathbf{R}\setminus\mathbf{R}_a'} d\lambda \, \|\Omega^{\text{red}}(\lambda) - \mathbf{1}\|^2 = \Sigma_n \left[\int_{\lambda_{2n-1}}^{\lambda_{2n}} d\lambda \, \|\Omega^{\text{red}}(\lambda) - \mathbf{1}\|^2 + \int_{-\lambda_{2n}}^{-\lambda_{2n-1}} d\lambda \, \|\Omega^{\text{red}}(\lambda) - \mathbf{1}\|^2 \right] < \infty, \tag{21}$$

because the gap lengths are $O(1/n^2)$.

D. In the allowed bands

There is a further problem to be considered before we have a proof that $(\Omega^{\text{red}} - \mathbf{1}) \in L^2(\mathbf{R})$: the periodic spectrum, when approached from the allowed bands. Equation (19) shows that, as λ approaches a given gap endpoint from the allowed side, $\Omega^{\rm red}$ remains bounded; from the gap side, the inverse square roots in $\Pi^{\#-1}\Omega Q\Pi Q=\Omega^{\rm red}$ must cancel. The question, as from inside the gaps, is whether these cancellations produce unbounded growth as $n\to\infty$.

The generic behavior of S for large λ and near $\sin k=0$ is obtained from the equations below (2.3), which leads to

$$\eta_3=-ic/\lambda+o(1/\lambda),\quad \eta_1=o(1/\lambda),$$

$$c=\frac{1}{2}\int_0^1 dx\, V(x),$$

and from (2.2), (3.7), (3.8), (3.4), (3.11), (3.18), and (4.2). One finds that

$$\beta_1=(1/\lambda)\sin k\,\gamma_1-(c/\lambda^2)\cos k\,(\gamma_1+\gamma_2)+o(1/\lambda^2),$$

$$\beta_2=(1/\lambda)\sin k\,\gamma_2-(c/\lambda^2)\cos k\,(\gamma_1+\gamma_2)+o(1/\lambda^2),$$

$$\epsilon=(1/\lambda)\sin k-(2c/\lambda^2)\cos k+o(1/\lambda^2).$$

As a result one finds that for large λ

$$S^*\simeq a\mathbf{1}+bQ+\sin k\,O(1/\lambda),$$

where

$$a=\left[\frac{2i\lambda^2}{2i\lambda^2\sin k-c'\cos k}+O\left(\frac{1}{\lambda}\right)\right]\sin k,$$

$$b=\frac{2cc'}{(\lambda\sin k-2c\cos k)(c'\cos k-2i\lambda^2\sin k)}+O\left(\frac{1}{\lambda}\right),$$

$$c'=c\int_{-\infty}^{\infty}dx\,U.$$

Furthermore,

$$M=\frac{1}{2i\lambda}(2i\sin k\,\mathbf{1}+\sigma),$$

$$M^{-1}=\frac{-i(1+\eta_3)}{2\epsilon\sin k}\left[2i\sin k\,\mathbf{1}+\sigma'+\sin k\,O\left(\frac{1}{\lambda}\right)\right],$$

where

$$\sigma=\begin{pmatrix}\sigma_1 & \sigma_2\\ \sigma_1 & \sigma_2\end{pmatrix},\quad \sigma'=\begin{pmatrix}\sigma_1 & -\sigma_2\\ -\sigma_1 & \sigma_2\end{pmatrix},$$

$$\sigma_1:=e^{i\lambda}l^*-e^{ik}=-(2ic/\lambda)e^{ik}+o(l/\lambda),$$

$$\sigma_2'=e^{i\lambda}-le^{ik}=-(2ic/\lambda)e^{ik}+o(1/\lambda).$$

Consequently

$$M^{-1}S^*M=\mathbf{1}a+Qb+O(1/\lambda).$$

Now examine a and b; they are bounded with respect to λ. (The real denominator of b, which is ϵ, will not vanish in an allowed band.) When $\sin k>1/\lambda$, then $a\simeq1$, and $b\simeq0$, so that $M^{-1}S^*M\simeq\mathbf{1}$. On the other hand, when $\sin k<1/\lambda^2$, then a gets small and b approaches -1. For $\sin k\ll 1/\lambda^2$,

$$a=O(\lambda^2\sin k),\quad b=-1+O(\lambda^2\sin k),$$

so that

$$\Omega=-Q+O(\lambda^2\sin k).$$

When $\sin k$ approaches zero to within $\epsilon_n^{1/2}$ then $\Gamma^{(4)\#-1}\Gamma^{(2)\#-1}$ in $\Omega^{\rm red}$ will lead to

$$B_n^{(4)}\left(1-\frac{i\epsilon_n}{\lambda-\lambda_{2n}}\right)^{1/2}\lambda^2\sin k\to B_n^{(4)}\epsilon_n^{1/2}\lambda_{2n}^2,\quad \text{as }\lambda\to\lambda_{2n},$$

$$B_n^{(2)}\left(1-\frac{i\epsilon_n}{\lambda-\lambda_{2n-1}}\right)^{1/2}\lambda^2\sin k\to B_n^{(2)}\epsilon_n^{1/2}\lambda_{2n-1}^2,$$

$$\text{as }\lambda\to\lambda_{2n-1}.$$

By (20), and because $\lambda_{2n}=\pi n+O(1/n)$ and $\lambda_{2n}-\lambda_{2n-1}=O(1/n^2)$, these remain bounded as $n\to\infty$. As a result $\Omega^{\rm red}$ does not grow with n at the periodic spectrum. When $\lambda-\lambda_{2n}>\epsilon_n$ or $\lambda_{2n-1}-\lambda>\epsilon_n$ the factors $\Pi^{\#-1}$ and Π may be replaced by $\mathbf{1}$.

E. The whole real line

The conclusion of these considerations is that there is an interval of length $\sim1/n^2$ (i.e., $\sin k\simeq1/n$) on both sides of the nth gap in which $\Omega^{\rm red}-\mathbf{1}$ is bounded uniformly in n, but not necessarily small. The sum of the integrals of $\|\Omega^{\rm red}-\mathbf{1}\|^2$ over these intervals converges; outside of them, $\Omega^{\rm red}-\mathbf{1}=O(1/\lambda)$, and hence, the sum of the integrals of $\|\Omega^{\rm red}-\mathbf{1}\|^2$ over the allowed bands converges. Together with (21), this proves that if the ϵ_n are chosen as in (20) then

$$\int_{-\infty}^{\infty}d\lambda\,\|\Omega^{\rm red}-\mathbf{1}\|^2<\infty. \tag{22}$$

V. THE REDUCED PROBLEM AND THE POTENTIAL

The reduced problem $\mathfrak{H}$ is now the standard Riemann–Hilbert problem based on (13) and (22). The requirements on $F^{\rm red}$ are (a) that in $\mathbb{C}^+$ it be analytic, zero-free, and aymptotic to $\mathbf{1}$; and (b) that $(F^{\rm red}-\mathbf{1})\in L^2(\mathbb{R})$. If this problem has a solution Φ, and if this solution is such that for some unbounded sequence ι_n, $\lim\Phi(\iota_n)=\mathbf{1}$, then, by the result of Sec. IV B, $F=\Pi\Phi$ solves the original problem $\mathfrak{H}$, and the Lemma of Sec. II ensures that it is the only solution of $\mathfrak{H}$.

The solution of the reduced problem, of course, proceeds by means of the Marchenko procedure, as in Sec. 6 of Ref. 1, and the starting point is (6), with the S matrix of a "comprehensively shifted" problem. The function Ψ of (4.20) is then given by

$$\Psi=\Sigma\,\Phi\,\hat{1}=\Sigma F^{\rm red}\hat{1},$$

where $\Sigma:=J_{0x}^{-1}\Pi$. Definition of the Fourier transforms (where the x-dependence is not explicitly shown)

$$\xi(\alpha)=\left(\frac{1}{2\pi}\right)\int_{-\infty}^{\infty}d\lambda\,e^{-i\lambda\alpha}[\Psi(\lambda)-\hat{1}],$$

$$A(\alpha)=\left(\frac{1}{2\pi}\right)\int_{-\infty}^{\infty}d\lambda\,e^{-i\lambda\alpha}[\Sigma(\lambda)-\mathbf{1}], \tag{23}$$

$$\eta(\alpha)=\left(\frac{1}{2\pi}\right)\int_{-\infty}^{\infty}d\lambda\,e^{-i\lambda\alpha}[F^{\rm red}(\lambda)-\mathbf{1}],$$

then leads to the equation

$$\xi(\alpha)=A(\alpha)\hat{1}+\eta(\alpha)\hat{1}+\int_0^{\infty}d\beta\,A(\alpha-\beta)\eta(\beta). \tag{23'}$$

The differential equation

$$\left(\frac{d^2}{dx^2}+2i\lambda I\frac{d}{dx}\right)\Psi=(V+U)\Psi \tag{24}$$

now leads to the condition

$$2I\frac{d}{dx}[\xi(0-)-\xi(0+)]=(V+U)\hat{1}, \tag{25}$$

and therefore from (23),

$$(V+U)\hat{1}=-2I\frac{d}{dx}\eta(0+)\hat{1}$$
$$+2I\frac{d}{dx}[\Lambda(0-)-\Lambda(0+)]\hat{1}. \tag{26}$$

The first term on the right-hand side is the end-point limit of the solution of the matrix Marchenko equation; the second arises from the reducing product Π and J_{ox}, in as (6.11). That the right-hand side of (26) must be a multiple of $\hat{1}$, because the left-hand side is, is the *miracle*.

Since $\Sigma(\lambda)$ is neither integrable nor square integrable, the question is how to handle the poles on the real axis in it when defining its Fourier transform $\Lambda(\alpha)$. One easily sees that the discontinuity $\Lambda(0-)-\Lambda(0+)$ at $\alpha=0$ is independent of whether the Fourier transform is defined by going around the poles in the upper or lower half-plane, or whether it is defined as Cauchy's principal value. Therefore, any of these prescriptions will do in (26).

Finally there is a problem that arises from the fact that, contrary to the statement in Ref. 1, Eq. (3.26) holds only for $x\in[0,1]$ (see the Appendix), and hence (26) gives us U only in that interval. We therefore define the solution $\psi_n(x)$ for a shifted $U(x)$, $U_n(x):=U(x+n)$

$$\psi_n(x)=\beta(x)+\int_{-\infty}^{\infty}dy\,g^+(x,y)U_n(y)\psi_n(y), \tag{27}$$

in which n is a positive or negative integer. Then,

$$\psi_n(x)=e^{iknI}\psi(x+n)$$

and hence, if we define

$$S^{(n)}=e^{iknI}Se^{-iknI}, \tag{28}$$

then in allowed bands

$$\psi_n^*(x)=QS^{(n)*}\psi_n(x).$$

The function $\Psi_n=X^{-1}\hat{\psi}_n=X^{-1}M^{-1}\psi_n$ then satisfies

$$\Psi_n^{\#}=Q\hat{S}_x^{(n)*}\Psi_n \tag{29}$$

in allowed bands, and

$$\Psi_n^{\#}=QS_{ox}^*\Psi_n \tag{29'}$$

in the gaps, where

$$\hat{S}_x^{(n)}=XQM^{-1}QS^{(n)}M^*X^{-1}.$$

Now since (3.26) holds for $0\leqslant x\leqslant 1$, it follows from (27) that for $0\leqslant x\leqslant 1$ and $n=\pm 1,\pm 2,\ldots$

$$\Psi_n=\hat{1}+O(1/\lambda)$$

as $\lambda\to\pm\infty$ or $|\lambda|\to\infty$ in $\mathbb{C}^+$. Therefore, for $n\leqslant x\leqslant n+1$ we solve the same problem as for $0\leqslant x\leqslant 1$, except that S is replaced by $S^{(n)}$ and x by $x-n$. Equation (26) then gives us $U_n(x-n)=U(x)$ for $n\leqslant x\leqslant n+1$.

Note added in proof: The following two papers by N. E. Firsova have come to my attention after this work was completed: Mat. Zametki **18**, 831 (1975) [Math. Notes **18**, 1085 (1975)]; Zap. Nauch. Sem. Leningrad. Otdel. Mat. Inst. Steklov. **51**, 183 (1975) [J. Sov. Math. **11**, 487 (1979)]. They solve the same problem as Ref. 1 and this paper, but by an extension of Faddeev's method.

ACKNOWLEDGMENT

This work was supported in part by grant PHY-80-20457 from the National Science Foundation.

APPENDIX: CORRECTIONS TO REFERENCE 1

The following statements in Ref. 1 pertaining to the periodic spectrum require correction:

p. 2155, two lines above Eq. (3.25): J_0M^{-1} is generally not continuous at $k=0(\mathrm{mod}\ \pi)$;

p. 2156, two lines above Eq. (3.27): g^+ is not continuous at $k=0(\mathrm{mod}\ \pi)$, and (3.27) holds only away from these points;

p. 2158, two lines above Eq. (4.31): J is generally not continuous at $k=0(\mathrm{mod}\ \pi)$, and (4.31) does not hold at these points and in the gaps;

p. 2158, right-hand column, line 26: $1/T$ is generally not continuous at $k=0(\mathrm{mod}\ \pi)$; line 32: This holds only away from these points;

p. 2160, lines 7–13 below Eq. (5.9): in the generic case T has simple zeros at the periodic spectrum as a function of k, which implies that as $\lambda\to\lambda_n$, T goes as $(\lambda-\lambda_n)^{1/2}$. As we circumscribe such a point clockwise in $\mathbb{C}^+$, its phase decreases by $\pi/2$. Therefore, we must define σ so that at each band gap the difference between its left-hand limit at the left gap end and its right-hand limit at the right gap end is π. At $\lambda=0$, the generic zero of T is simple as a function λ, because there $k\sim\lambda$. With the new definition of σ, the statement of Levinson's theorem remains correct. Its proof requires some simple changes of wording on the right-hand column of p. 2180: line 1, read $\pi/2$ for π; line 2, read π for 2π; line 5, read $-\pi$ for π.

In addition, there should be the following corrections: p. 2155, line 9 below Eq. (3.16): read β_1 for φ_1; p. 2156, Eq. (3.26) is valid only for $0\leqslant x\leqslant 1$; first line of (4.2): $-$ should read $+$; one line below (4.2), u should read U; p. 2157, Eq. (4.22′): $-$ should read $+$; line 1 below Eq. (4.26): (4.18) should read (4.17); p. 2158, right-hand column: delete line 14; line 17: $T=1$ should read $T\neq 0$; line 29: (3.28) should read (3.27); line 4 from bottom: $S\to 1$ should read $S\not\to -Q$; p. 2160, two lines below Eq. (6.1): $\mathbf{R}_a^*$ should read $\mathbf{R}_a'$; p. 2161, right-hand column, line 12: $\sigma(x)$ should read $\bar{\sigma}(x)$.

[1]R. G. Newton, J. Math. Phys. **24**, 2152–2162 (1983).
[2]V. A. Zheludev, in *Topics in Mathematical Physics*, edited by M. Sh. Birman (Consultants Bureau, New York, 1968), Vol. 2, p. 87.
[3]Recall that $J^{\#}(\lambda):=J(-\lambda)$.
[4]$\hat{1}:=\begin{pmatrix}1\\1\end{pmatrix}$.
[5]R. P. Boas, *Entire Functions* (Academic, New York, 1954), p. 4.
[6]W. Magnus and S. Winkler, *Hill's Equation* (Interscience, New York, 1966).
[7]S. Rofe-Beketov, Dokl. Akad. Nauk SSSR **156**, 515 (1964) [Sov. Phys. Dokl. **5**, 689 (1964)].
[8]This is easily proved in the standard way be means of (4.2).

An electromagnetic inverse problem for dispersive media

R. S. Beezley
Department of Mathematics, University of Nebraska, Lincoln, Nebraska 68588

R. J. Krueger
Applied Mathematical Sciences, Ames Laboratory–United States Department of Energy, Iowa State University, Ames, Iowa 50011

(Received 17 July 1984; accepted for publication 24 August 1984)

The dispersion of transient electromagnetic waves in a homogeneous medium can be characterized by expressing either the complex permittivity as a function of frequency or the susceptibility kernel as a function of time. In this paper, a time domain technique is used to derive a nonlinear integrodifferential equation which relates the susceptibility kernel for a one-dimensional homogeneous slab to the reflection operator for the medium. Thus, the susceptibility kernel (which is a function of time) can be determined from reflection data. A numerical implementation of this technique is shown. The more general case of a medium consisting of a stack of homogeneous dispersive layers is also addressed.

I. INTRODUCTION

Linear wave propagation in a dispersive medium is characterized by the fact that the phase and group velocities are functions of frequency. Thus, a transient pulse in a dispersive medium will tend to spread and change shape, even if the medium is homogeneous.

In the case of electromagnetic wave propagation, the subject of this paper, the physical basis for this dispersive phenomenon lies in the constitutive relation between the displacement field $\mathbf{D}(\mathbf{x},t)$ and the electric field $\mathbf{E}(\mathbf{x},t)$. In the time domain this can be expressed in the simplest case as[1,2]

$$\mathbf{D}(\mathbf{x},t) = \epsilon_0\left[\mathbf{E}(\mathbf{x},t) + \int_0^\infty G(s)\mathbf{E}(\mathbf{x},t-s)ds\right], \tag{1.1}$$

where ϵ_0 is the permittivity of free space. This relation says that the displacement field at a point in a homogeneous medium depends on the properties of the medium (as expressed in the susceptibility kernel G) and the past history of the electromagnetic field at that point [$\mathbf{E}(\mathbf{x},s)$ for $-\infty < s < t$]. Equation (1.1) can be shown to be equivalent to the frequency domain Kramers–Kronig dispersion relations, which relate the real and imaginary parts of the complex permittivity $\epsilon(\omega)$. The connection between time domain and frequency domain results is provided by the Fourier transform

$$\frac{\epsilon(\omega) - \epsilon_0}{\epsilon_0} = \int_0^\infty G(t)e^{i\omega t}\,dt. \tag{1.2}$$

The inverse problem considered in this paper involves determining the dispersive properties of a homogeneous medium (i.e., the susceptibility kernel G) by means of scattering experiments. The precise formulation of this problem will be given in Sec. II. Notice that while this problem is equivalent to determining the complex permittivity of the medium, the approach here will be entirely in the time domain and will not depend on Fourier transforming back to the frequency domain. Most previous work on inverse problems for dispersive electric media is carried out in the frequency domain, with measurements being made at a fixed frequency but varying the angle of incidence.[3,4] Thus, the dispersive character of the problem is in fact not a central issue in the solution technique. On the other hand, such techniques do apply to inhomogeneous media, modulo problems with measuring evanescent waves. Although dispersive, dissipative inverse problems have been studied in the time domain,[5–9] the models used are not applicable to the physics expressed in Eq. (1.1). Rather, they apply in the case in which the permittivity and conductivity vary spatially but are independent of ω. Thus, this paper should be viewed as a first step in the study of dispersive inverse problems which uses a physically motivated model of dispersion and exploits causality.

In Sec. II the precise form of the inverse problem considered herein is given. Additionally, splitting and reflection operators are introduced which form the framework for this time domain approach. In Sec. III an integrodifferential equation is derived for the reflection operator for a finite slab. This equation relates the reflective behavior of a medium to the susceptibility kernel G and therefore is useful both for direct and inverse scattering studies. A semi-infinite medium is considered in Sec. IV by suitably modifying the analysis of Sec. III. In this case, the integrodifferential equation reduces to a Volterra equation of the second kind for G. Section V shows a comparison of classical frequency domain results and the time domain results given herein. The more general problem of a layered medium consisting of a stack of homogeneous dispersive slabs is addressed in Sec. VI. Section VII presents an outline of a numerical implementation of the equations developed in Secs. III and IV. In Sec. VIII numerical examples of inversions are given. A summary follows in Sec. IX, which points out what is done and not done in this paper. Finally, three appendices present some more detail regarding the analysis.

II. PROBLEM FORMULATION

The scattering model considered in this section consists of a homogeneous, isotropic, dispersive medium bounded by the planes $z = 0$ and $z = L > 0$. The magnetic permeability is assumed to be constant (μ_0) and Eq. (1.1) is assumed to hold. Free space occupies the regions on either side of this medium. A right-moving electromagnetic plane wave in the region $z<0$ impinges on the medium at normal incidence, producing a left-moving reflected wave as well as establishing a transient field within the medium. Letting $E(z,t)$ denote a transverse component of the electric field, it follows from Maxwell's equations that

$$E_{zz} - (1/c^2)E_{tt} = 0, \quad z<0 \quad \text{or} \quad z>L, \tag{2.1}$$

and, using (1.1),

$$E_{zz} - \frac{1}{c^2}\left[E_{tt} + \partial_t^2 \int_0^\infty G(s)E(z,t-s)ds\right] = 0, \quad 0<z<L, \tag{2.2}$$

where $c^2 = (\epsilon_0\mu_0)^{-1}$, ϵ_0 and μ_0 being the permittivity and permeability of free space, respectively. Notice that the velocity c is assumed to be the same in the dispersive medium (2.2) as in the host medium (2.1). This assumption will be relaxed in Sec. VI.

In the region $z<0$, the field $E(z,t)$ can be split into a sum of right- and left-moving components,

$$E(z,t) = E^+(z,t) + E^-(z,t), \quad z<0,$$

where

$$E^+(z,t) = f(t-z/c) \quad \text{(incident field)}, \tag{2.3a}$$

$$E^-(z,t) = g(t+z/c) \quad \text{(reflected field)}. \tag{2.3b}$$

Using a variation of Duhamel's integral[10] it can be shown that these fields are related via a reflection operator

$$E^-(0,t) = [\tilde{R}E^+(0,\cdot)](t) = \int_{-\infty}^t R(t-s)E^+(0,s)ds. \tag{2.4}$$

The kernel R is the impulse response function for the dispersive medium. It is a difference kernel because Eq. (2.2) is invariant under time translation. Furthermore, it does not depend on the field E, but rather depends only on the properties of the medium. Notice also that $\tilde{R}$ is a causal operator, since the reflected field at time t depends only on the incident field at earlier times.

At this point the inverse problem for Eq. (2.2) can be stated precisely: given $R(t)$ for $0\leqslant t\leqslant T$ (for some T), find $G(t)$ for $0\leqslant t\leqslant T$. Notice that unlike other one-dimensional inverse problems, which seek to reconstruct some function of the spatial variable z, this problem involves reconstructing a function of the time variable t.

The data for the inverse problem, $R(t)$, are the result of a deconvolution of Eq. (2.4). The deconvolution problem itself will not be discussed here. Rather, it is assumed that $R(t)$ has been accurately obtained by some means. However, the equations derived in this paper appear to also be suitable for studying the effects of deconvolution on the solution of the inverse problem.

In order to solve the inverse problem, a relation between $R(t)$ and $G(t)$ will be established via a wave-splitting approach to Eq. (2.2), coupled with an invariant imbedding technique. These ideas have been utilized in other types of direct and inverse problems[5,11-14] but they manifest themselves somewhat differently for the dispersive problem now under consideration. Consequently, the machinery behind this approach will be shown in some detail.

For any z in $[0,L]$ define functions $E^\pm$ by

$$E^\pm(z,t) = \tfrac{1}{2}[E(z,t) \mp c\,\partial_t^{-1}E_z(z,t)], \tag{2.5}$$

where E is a solution of (2.2) and

$$\partial_t^{-1}E_z(z,t) = \int_{-\infty}^t E_z(z,s)ds.$$

In the region $z<0$ the definition (2.5) with E a solution of (2.1) results in Eqs. (2.3a) and (2.3b). Thus, E^+ and E^- in (2.5) can be thought of as approximate right- and left-moving waves in the medium. Although this is not a physically well-defined concept for a transient field in a dispersive medium, the analysis that stems from definition (2.5) is in fact precise. Using the time translation invariance of (2.2), the existence of a "reflection" operator $\tilde{R}(z)$ given by

$$E^-(z,t) = [\tilde{R}(z)E^+(z,\cdot)](t) = \int_{-\infty}^t R(z,t-s)E^+(z,s)ds \tag{2.6}$$

can be proved. This operator can be thought of as the reflection operator for the portion of the dispersive medium occupying the region $[z,L]$, with free space everywhere else. With this notation, the kernel $R(t)$ given in (2.4) should now be written $R(0,t)$.

In the next section, the behavior of the kernel $R(z,t)$ will be examined. This will provide the link between the impulse response $R(0,t)$ and the susceptibility kernel $G(t)$.

III. FINITE SLAB

In order to simplify the derivation which follows, some preliminary observations are in order. First, in verifying the existence of the reflection operator $\tilde{R}(z)$ given in Eq. (2.6), it becomes clear that the kernel $R(z,t)$ is independent of the fields in and around the medium; and depends only on the properties of the medium itself. Second, Eq. (2.6) is valid for arbitrary fields $E^+(z,t)$. However, the analysis which follows is greatly simplified by assuming that $E(z,t)$ is twice continuously differentiable for all (z,t). The resulting equations derived for $R(z,t)$ are not altered by requiring E to be C^2 since R is independent of E. Third, in addition to assuming E is C^2, the initial conditions

$$E(z,0) = E_t(z,0) = 0, \quad z>0$$

will be imposed with Eqs. (2.1) and (2.2). This implies that the incident field $E^+(z,t)$ (for $z<0$) does not impinge on the medium prior to $t=0$. In other words, $f(t)=0$ for $t<0$ in Eq. (2.3a). Fourth, the susceptibility kernel $G(t)$ will be assumed to be differentiable for $t>0$.

The derivation begins by rewriting Eq. (2.2) in terms of $E^\pm(z,t)$. This is done by first writing

$$\partial_z \begin{pmatrix} E \\ E_z \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ (\partial_t^2 + G * \partial_t^2)/c^2 & 0 \end{pmatrix} \begin{pmatrix} E \\ E_z \end{pmatrix} \equiv D \begin{pmatrix} E \\ E_z \end{pmatrix}, \tag{3.1}$$

where the $*$ operation denotes convolution in time,

$$G * \partial_t^2 E(z,t) = \int_0^t G(s)E_{tt}(z,t-s)ds.$$

Now set

$$\begin{pmatrix} E^+ \\ E^- \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & -c\,\partial_t^{-1} \\ 1 & c\,\partial_t^{-1} \end{pmatrix}\begin{pmatrix} E \\ E_z \end{pmatrix} \equiv T \begin{pmatrix} E \\ E_z \end{pmatrix}, \tag{3.2}$$

where now

$$\partial_t^{-1} E_z(z,t) = \int_0^t E_z(z,s) ds.$$

The operator matrix T is called a splitting matrix[15] since it splits the field into $\pm$ components. From Eqs. (3.1) and (3.2),

$$\partial_z \begin{pmatrix} E^+ \\ E^- \end{pmatrix} = TDT^{-1} \begin{pmatrix} E^+ \\ E^- \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} E^+ \\ E^- \end{pmatrix}, \qquad (3.3)$$

where

$$\delta = -\alpha = [\partial_t + \tfrac{1}{2} G * \partial_t]/c,$$

$$\gamma = -\beta = [\tfrac{1}{2} G * \partial_t]/c.$$

Observe that Eq. (3.3) is equivalent to Eq. (2.2). Thus, the degree to which $E^\pm$ represents right- and left-moving waves is immaterial.

Next, the equation

$$E^-(z,t) = [\widetilde{R}(z) E^+(z,\cdot)](t) = \int_0^t R(z,t-s) E^+(z,s) ds \qquad (3.4)$$

[which follows from (2.6) since $E^+(z,t) = 0$ for $t < 0$] implies that

$$E_z^- = \widetilde{R}_z E^+ + \widetilde{R} E_z^+. \qquad (3.5)$$

Using Eqs. (3.3)–(3.5) and the fact that $E^+(z,t)$ can be considered to be arbitrary yields

$$\widetilde{R}_z = \gamma + \delta \widetilde{R} - \widetilde{R} \alpha - \widetilde{R} \beta \widetilde{R}. \qquad (3.6)$$

In terms of the kernel $R(z,t)$, Eq. (3.6) implies that

$$2cR_z = 4R_t + G'(t) + G(0)[2R + R*R] + G'*[2R + R*R], \quad 0<z<L, \quad t>0 \qquad (3.7)$$

(where $' = d/dt$) and

$$R(z,0^+) = -\tfrac{1}{4} G(0^+), \quad 0<z<L. \qquad (3.8)$$

It is also clear that

$$R(L,t) = 0, \quad t>0. \qquad (3.9)$$

Finally, using standard propagation of singularities arguments[16] it follows from Eq. (3.6) that discontinuities in $R(z,t)$ can propagate only along lines of the form

$$t = -2(z - z_0)/c \quad (z_0 \text{ constant})$$

in the (z,t) plane. Consequently from Eq. (3.7)–(3.9) the only possible discontinuity in R is along the line

$$t = 2(L - z)/c,$$

where

$$R(z,(2/c)(L-z)^+) - R(z,(2/c)(L-z)^-) = \tfrac{1}{4} G(0^+) \exp[G(0^+)(z-L)/c]. \qquad (3.10)$$

Thus, if $G(0^+) = 0$ then R is continuous everywhere.

The integrodifferential equation (3.7) with initial condition (3.8), boundary condition (3.9), and jump condition (3.10) provides a link between the impulse response $R(0,t)$ and $G(t)$. Hence, the system (3.7)–(3.10) can be used to study the inverse problem, as is demonstrated in Sec. VIII. This system can also be used to study the direct scattering problem in which $G(t)$ is given and $R(0,t)$ is to be determined. (Again, see Sec. VIII.) Figure 1 shows the domain for the system (3.7)–(3.10). The solution of the inverse problem involves all three regions of Fig. 1. This is in contrast to earlier splitting/invariant imbedding approaches to inverse problems,[5,11–14] in which data from only one round trip through the medium is used, which corresponds to region 1.

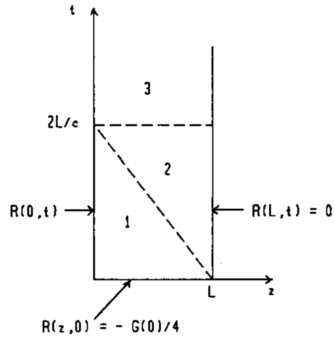


FIG. 1. Domain for the system of equations (3.7)–(3.10).

IV. SEMI-INFINITE MEDIUM

Some insight into Eq. (3.7) can be gained by considering the special case of a semi-infinite medium, $0<z<\infty$. Notice that with $L = \infty$, the scattering produced by the portion of the medium (z,∞) is independent of z. In other words, the operator $\widetilde{R}$ and kernel R are independent of z. Thus, integrating Eq. (3.7) with respect to t and using Eq. (3.8) results in

$$4R(t) + G(t) + [G*(2R + R*R)](t) = 0, \quad t>0. \qquad (4.1)$$

Now for the direct scattering problem [$G(t)$ given], Eq. (4.1) is a nonlinear integral equation for $R(t)$. Under the mild assumption that $G(t)$ is bounded, this problem is well posed in the sense that a unique solution $R(t)$ exists and that solution depends continuously on the "data," $G(t)$. This is demonstrated in Appendix A. The existence proof in Appendix A also provides an iterative approach to the solution of Eq. (4.1).

For the inverse scattering problem [$R(t)$ given], Eq. (4.1) is simply a Volterra equation of the second kind for $G(t)$. Hence, this problem is also well posed[17] and in particular, small changes in the measured data $R(t)$ produce only small changes in the reconstructed susceptibility kernel $G(t)$.

Equation (4.1) can be solved exactly for the special case in which

$$G(t) = \alpha e^{\beta t} \quad (\alpha, \beta \text{ constant}), \quad t>0.$$

Upon setting

$$R(t) = f(t) e^{\beta t},$$

Eq. (4.1) becomes

$$4f + \alpha + \alpha*[2f + f*f] = 0.$$

Now f can be easily found using Laplace transforms, and so finally

$$R(t) = -\exp[(\beta - \alpha/2)t] \cdot I_1(\alpha t/2)/t,$$

where I_1 is the modified Bessel function of the first kind.

In the same manner, it can be shown that if

$$G(t) = \alpha t e^{\beta t}, \quad t>0,$$

then

$$R(t) = -2e^{\beta t}J_2(\alpha^{1/2}t/2)/(\alpha^{1/2}t),$$
where J_2 is the Bessel function of the first kind.

V. COMPARISON OF FREQUENCY DOMAIN AND TIME DOMAIN RESULTS

Linear wave propagation in dispersive media is more commonly considered in the frequency domain than in the time domain. For example, the causal, nonlocal relation (1.1) between **D** and **E** is shown by Jackson[1] to be equivalent to the usual frequency domain result

$$\widehat{\mathbf{D}}(\mathbf{x},\omega) = \epsilon(\omega)\widehat{\mathbf{E}}(\mathbf{x},\omega), \tag{5.1}$$

where $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{E}}$ are the Fourier transforms of **D** and **E**, respectively. Equations (5.1) and (1.1) provide the link between the susceptibility kernel G and the complex, frequency dependent permittivity ϵ as expressed in Eq. (1.2) or equivalently,

$$G(t) = \frac{1}{2\pi\epsilon_0}\int_{-\infty}^{\infty} [\epsilon(\omega) - \epsilon_0]e^{-i\omega t}\,d\omega.$$

Some specific examples are as follows.

(1) For a nonmagnetic medium of relatively low density a simple resonance model of the electron contribution to the permittivity yields[1]

$$\epsilon(\omega) = \epsilon_0[1 + \omega_p^2(\omega_0^2 - \omega^2 - i\gamma\omega)^{-1}],$$

where ω_p is the plasma frequency, ω_0 the resonant frequency, and γ is a damping constant. Jackson shows that the corresponding susceptibility kernel is

$$G(t) = H(t)\omega_p^2 e^{-\gamma t/2}\sin(\nu_0 t)/\nu_0,$$

where $\nu_0^2 = \omega_0^2 - \gamma^2/4$ and $H(t)$ is the Heaviside function, vanishing for $t<0$.

(2) A Debye model of dispersion[18] results in

$$\epsilon(\omega) = \epsilon_\infty + (\epsilon_s - \epsilon_\infty)(1 + i\omega\tau)(1 + \omega^2\tau^2)^{-1},$$

where ϵ_∞, ϵ_s, and τ are given parameters. In this case, Eq. (1.1) is replaced with

$$\mathbf{D}(\mathbf{x},t) = \epsilon_\infty\left[\mathbf{E}(\mathbf{x},t) + \int_0^\infty G(s)\mathbf{E}(\mathbf{x},t-s)ds\right], \tag{5.2}$$

where

$$G(t) = H(t)e^{-t/\tau}(\epsilon_s - \epsilon_\infty)/(\epsilon_\infty\tau).$$

The factor ϵ_∞ in place of ϵ_0 in Eq. (5.2) causes some complications in the analysis. This will be examined in Sec. VI.

(3) Notice from Eq. (1.2) that if G is a multiple of $\delta(t)$, the Dirac delta function, then ϵ is independent of ω and the medium is nondispersive.

At this point it is convenient to discuss assumptions regarding the behavior of $G(t)$ for $t\to 0^+$. While Jackson argues that "it is unphysical to have $G(0)\neq 0$,"[1] Chelkowski[18] shows that under a more macroscopic point of view, it is reasonable to consider cases in which $G(0)\neq 0$ as in the Debye model above. In this paper the more general situation [$G(0)$ not necessarily 0] is considered.

The reflection kernel $R(z,t)$ is related to the reflection coefficient $\rho(z,\omega)$ via the Fourier transform. Upon transforming Eq. (2.6) it follows that

$$\rho(z,\omega) = \int_0^\infty R(z,t)e^{i\omega t}\,dt.$$

Thus, Eq. (3.7) can be rewritten in the frequency domain as

$$2c\epsilon_0\rho_z = i\omega[\epsilon_0(1-\rho)^2 - \epsilon(1+\rho)^2], \tag{5.3}$$

with $\rho(L,\omega) = 0$. Notice that from (5.3) it follows that

$$\rho(0,\omega) = r(1 - e^{-aL})/(r^2 - e^{-aL}),$$

where

$$r = (\sqrt{\epsilon_0} - \sqrt{\epsilon})/(\sqrt{\epsilon_0} + \sqrt{\epsilon}),$$
$$a = 2i\omega c^{-1}\sqrt{\epsilon/\epsilon_0},$$

which agrees with standard results.[19]

VI. LAYERED MEDIA

The problem formulated in Sec. II is generalized in this section by considering a layered medium consisting of a stack of homogeneous dispersive slabs. As a further generalization, the velocity c can differ from slab to slab. The boundaries of the individual layers are at $0 = z_0 < z_1 < z_2 < \cdots$. In the nth slab the constitutive relation between D and E is

$$D(z,t) = \epsilon_n\left[E(z,t) + \int_0^\infty G_n(s)E(z,t-s)ds\right],$$

where ϵ_n is a constant. Now E satisfies

$$E_{zz} - \frac{1}{c_n^2}\left[E_{tt} + \partial_t^2\int_0^\infty G_n(s)E(z,t-s)ds\right] = 0,$$
$$z_{n-1} < z < z_n,$$

where $c_n^2 = (\epsilon_n\mu_0)^{-1}$. For this situation a reflection operator $\widetilde{R}$ again transforms the incident to the reflected field. The inverse problem is to determine the functions $G_1, G_2, \ldots$ and velocities $c_1, c_2, \ldots$ from knowledge of the reflection kernel $R(0,t)$.

Intuitively, this problem has nonunique solutions without some further assumptions. This is because the dispersive characteristics of the deeper portions of the medium can be erroneously attributed to the large time behavior of shallow portions of the medium.

The inverse scattering problem can be made tractable by assuming that in each layer of the medium the form of the function G_n is known. For example, it might be assumed that each layer is a Debye medium, in which case

$$G_n(t) = \alpha_n e^{-\beta_n t}.$$

The inverse problem now reduces to finding the velocity c_n and parameters α_n, β_n for each layer, $n = 1,2,\ldots$. The solution procedure for a layered medium is now a recursive process. Given the kernel $R(0^-,t)$, the first step is to determine the velocity c_1 and then determine $R(0^+,t)$; i.e., step the data across the discontinuity at $z=0$. The function G_1 can then be determined. Finally, the data $R(0^+,t)$ is propagated through the first layer, producing the reflection kernel $R(z_1^-,t)$. At this stage the solution process commences in the second layer in the same manner as in the first.

For cases in which the velocity is discontinuous across the interfaces $z_0, z_1, \ldots$, the reflection kernel contains δ function singularities. In particular, if $c_0 \neq c_1$, then for t sufficiently small the reflection operator $\widetilde{R}(0^-)$ has the form

$$[\widetilde{R}(0^-)f](t) = r^+ f(t) + [R(0^-,\cdot)*f](t), \tag{6.1}$$

where

$$r^+ = (c_1 - c_0)/(c_1 + c_0),$$

and $R(0^-,t)$ is the nonsingular portion of the reflection kernel. Thus, the velocity c_1 is determined from Eq. (6.1) via the strength of the singularity in $\widetilde{R}(0^-)$. Using the star product of operators[13] it now follows that the reflection operator at $z = 0^+$ can be determined from

$$\widetilde{R}(0^-) = r^+ I + t^-[I - \widetilde{R}(0^+)r^-]^{-1}\widetilde{R}(0^+)t^+, \quad (6.2)$$

where

$$r^- = -r^+, \quad t^+ = 2c_1/(c_1 + c_0), \quad t^- = 2c_0/(c_1 + c_0),$$

and I is the identity operator. Equation (6.2) is written in terms of reflection kernels in Appendix B.

Once $\widetilde{R}(0^+)$ is known, the function G_1 can be found for $0 \leqslant t < \infty$. This is because of the fact that for $0 \leqslant t < 2(z_1 - z)/c_1$ the kernel $R(z,t)$ is independent of z and, consequently, the analysis of Sec. IV applies. Thus, G_1 can be found for $0 \leqslant t < 2z_1/c_1$ upon solving Eq. (4.1). Because the functional form of G_1 is assumed to be known, it follows that $G_1(t)$ is known for all t.

Finally, in order to determine $\widetilde{R}(z_1^-)$ the analysis of Sec. III applies and, consequently,

$$\widetilde{R}_z = \gamma_1 + \delta_1\widetilde{R} - \widetilde{R}\alpha_1 - \widetilde{R}\beta_1\widetilde{R} \quad (6.3)$$

follows directly from Eq. (3.6), with

$$\delta_1 = -\alpha_1 = [\partial_t + \tfrac{1}{2}G_1 * \partial_t]/c_1,$$
$$\gamma_1 = -\beta_1 = [\tfrac{1}{2}G_1 * \partial_t]/c_1.$$

An alternate approach to determining $\widetilde{R}(z_1^-)$ is available for the case in which

$$G_1(t) = \alpha_1 e^{-\beta_1 t}$$

or

$$G_1(t) = \alpha_1 t e^{-\beta_1 t}.$$

Set

$$\widetilde{R}(z) = \widetilde{R}_1 + \widetilde{R}_2(z), \quad (6.4)$$

where $\widetilde{R}_1$ is the corresponding reflection operator found in Sec. IV. Substituting (6.4) into (6.3) yields

$$\partial_z\widetilde{R}_2 = \delta_1\widetilde{R}_2 - \widetilde{R}_2\alpha_1 - 2\widetilde{R}_1\beta_1\widetilde{R}_2 - \widetilde{R}_2\beta_1\widetilde{R}_2.$$

It is not necessary to assume a functional form for $G(t)$ in the deepest layer of the medium, as this can be discerned in the same manner as for a finite slab. In particular, the problem formulated in Sec. II can now be generalized to the case in which a dispersive slab with unknown characteristic velocity c_1 is situated in a nondispersive host medium with velocity c_0 for $z < 0$ and unknown velocity c_2 for $z > L$. The solution technique for this problem consists of first determining c_1 as outlined above, then stepping the data across the discontinuity at $z = 0$. For t sufficiently small, these new data take the form

$$[\widetilde{R}(0^+)f](t) = t_0^2 r_1^+ f(t - 2L/c_1) + [R(0^+,\cdot) * f](t),$$

where

$$t_0 = \exp[-G(0^+)L/(2c_1)],$$
$$r_1^+ = (c_2 - c_1)/(c_2 + c_1).$$

Consequently, c_2 can be determined from the strength of the first singularity in $\widetilde{R}(0^+)$. Finally, G is determined via Eq. (3.6).

VII. A NUMERICAL SCHEME FOR THE DIRECT AND INVERSE PROBLEM

In this section a numerical scheme is presented for an approximate solution of Eq. (3.7) on the rectangle $0 \leqslant z \leqslant L$, $0 \leqslant t \leqslant T$ for some T. For the direct problem the susceptibility kernel $G(t)$ is known and the problem is to calculate $R(0,t)$ for $0 \leqslant t \leqslant T$. In the inverse problem, $R(0,t)$ is specified for $0 \leqslant t \leqslant T$ and $G(t)$ is to be determined.

The following observations apply to both the direct and inverse problem. In the lower triangular region (1) of Fig. 1, R is independent of z because reflections off the back wall of the slab have not yet returned to make a contribution to R. Consequently, R satisfies Eq. (4.1) in region 1, being constant on horizontal lines $t = t_0$, $0 \leqslant t_0 < 2L/c$. In regions 2 and 3, R is a function of both z and t, so Eq. (3.7) applies with the jump condition (3.10) holding along the line $t = 2(L - z)/c$.

Now consider a discretization of Eq. (3.7). Begin by writing (3.7) as

$$\begin{aligned}2c\frac{d}{dz}R(z,t - 2z/c) &= G'(t - 2z/c) + G(0)[2R + R * R](z,t - 2z/c)\\ &\quad + \{G' * [2R + R * R]\}(z,t - 2z/c). \qquad (7.1)\end{aligned}$$

Integrate with respect to z from $z = z_0$ to $z = z_0 + h$ using the trapezoidal rule to approximate the integral of the right-hand side of (7.1). With $t_0 = t - 2z_0/c$ this results in

$$\begin{aligned}2c[R(z_0 + h,t_0 - 2h/c) - R(z_0,t_0)] &= \tfrac{1}{2}h[G'(t_0 - 2h/c) + G'(t_0)]\\ &\quad + hG(0)[R(z_0 + h,t_0 - 2h/c) + R(z_0,t_0)] \qquad (7.2)\\ &\quad + h(S * R)(z_0 + h,t_0 - 2h/c)\\ &\quad + h(S * R)(z_0,t_0) + O(h^2),\end{aligned}$$

where

$$S(z,t) = G'(t) + \tfrac{1}{2}[G(0)R + G' * R](z,t).$$

Introduce a rectangular grid of points in the region $0 \leqslant z \leqslant L$, $0 \leqslant t \leqslant T$ with $L/N = h$ being the spacing in the z direction, and $2h/c$ being the spacing in the t direction. Let

$$R_{ij} \doteq R(ih,2jh/c),$$
$$G_j \doteq G(2jh/c), \quad G'_j \doteq G'(2jh/c)$$

denote approximations to R,G,G', where $i = 0,1,\ldots,N$, $j = 0,1,\ldots,J$ and $J = [cT/2h]$. Using the trapezoidal rule to approximate the convolutions in (7.2) results in the discretization

$$\begin{aligned}2c[R_{i+1,j-1} - R_{i,j}] &= \tfrac{1}{2}h[G'_{j-1} + G'_j] + hG_0[R_{i+1,j-1} + R_{i,j}]\\ &\quad + h^2G_0[A_{i+1,j-1} + A_{i,j}]/c \qquad (7.3)\\ &\quad + 2h^2[B_{i+1,j-1} + B_{i,j}]/c\\ &\quad + 2h^3[C_{i+1,j-1} + C_{i,j}]/c^2,\end{aligned}$$

where

$$A_{i,j} = \sum_{k=1}^{j} R_{i,j-k}R_{i,k},$$

$$B_{i,j} = \tfrac{1}{2}(G'_jR_{i,0} + G'_0R_{i,j}) + \sum_{k=1}^{j} G'_{j-k}R_{i,k},$$

$$C_{i,j} = \tfrac{1}{2}G'_0 A_{i,j} + \sum_{k=1}^{j-1} G'_{j-k} A_{i,k}.$$

In the lower triangular region (region 1) of Fig. 1, Eq. (4.1) can be discretized in a manner similar to that above yielding

$$4R_{0,j} + G_j + 4hE_j/c + 4h^2F_j/c^2 = 0, \qquad (7.4)$$

where

$$E_j = \tfrac{1}{2}(G_j R_{0,0} + G_0 R_{0,j}) + \sum_{k=1}^{j-1} G_{j-k} R_{0,k},$$

$$F_j = \tfrac{1}{2}G_0 A_{0,j} + \sum_{k=1}^{j-1} G_{j-k} A_{0,k}.$$

For the direct problem, Eq. (7.4) is solved for $R_{0,j}, j = 0,1,...,N$ and then $R_{i,j}$ in region 1 is given by

$$R_{i,j} = R_{0,j}, \quad j = 0,1,...,N, \quad i = 0,1,...,N-j. \qquad (7.5)$$

This gives $R_{i,N-i}$ as an approximation to $R\,[ih,2(N-i)h/c^-]$. Condition (3.10) then computes an approximation to $R\,[ih,2(N-i)h/c^+]$. Finally, Eq. (7.3) is solved for $R_{i,j}$ in regions 2 and 3. The calculation proceeds from right to left, bottom to top, using the previously determined values of $R_{i,j}$ and condition (3.9),

$$R_{N,j} = 0, \quad j = 0,1,...,J.$$

Since the trapezoidal rule was used to derive Eq. (7.3), the value of $R_{i,N-i}$ in that equation can be taken to be the average of the R values at $j = N - i$, yielding

$$R_{i,N-i} = R_{0,N-i} + \tfrac{1}{8}G_0 \exp[-G_0(N-i)h/c]. \qquad (7.6)$$

For the inverse problem, Eq. (7.4) is first solved for G_j, $j = 0,1,...,N$, and $R_{i,j}$ is determined in region 1 via Eq. (7.5). A difference formula is used to compute $G'_j, j = 0,1,...,N$, and then $R_{i,j}$ is determined in region 2 via Eq. (7.3), again proceeding from right to left, bottom to top and using (7.6). Finally, assuming G'_j and $R_{i,j}$ are known for $j = 0,1,...,k$ and $i = 0,1,...,N$, G'_{k+1} is determined by solving (7.3) with $i = 0, j = k + 1$, since $R_{0,k+1}$ is known. Then $R_{i,k+1}$ is obtained from (7.3) for $i = 1,...,N-1$ and the procedure continues, sweeping left to right, bottom to top across region 3.

VIII. EXAMPLES

In this section two numerical examples are given which illustrate the use of both the forward and inverse algorithms presented in this paper. The approach to each example is similar: a kernel $G(t)$ is selected, scattering data are generated, and a form of the data is then used in the inversion algorithm.

The depth of the medium was chosen to be $L = 0.8$. The time variable was scaled by c, and in these scaled units T was chosen to be 6.0, corresponding to 3.75 round trips through the medium. The scattering data were produced using the numerical scheme of the previous section, first with a step size $h = \frac{1}{80}$, then with a step size $h = \frac{1}{40}$, and finally passively extrapolating to determine $R(0,t)$. The extrapolated scattering data were then used in the inversion algorithm. In the inverse problem, G' was obtained from G in region 1 by means of a fourth-order difference formula. In region 3, G was obtained from G' using a second-order quadrature formula.

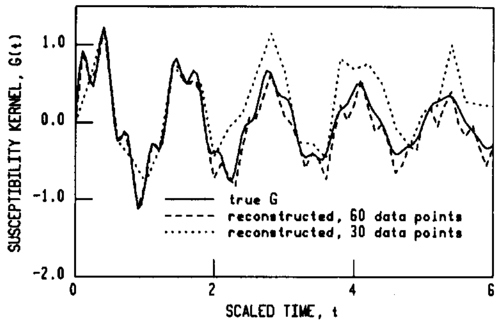


FIG. 2. Reconstruction of the susceptibility kernel for the two-resonance model given in Sec. VIII, example 1. With 120 data points the reconstructed G is indistinguishable from the true G.

Example 1: A two-resonance model for the electron contribution to the permittivity is used in this example. Thus, G is given by

$$G(t) = e^{-0.2t}\sin(1.6\pi t) + 0.5e^{-0.5t}\sin(6\pi t),$$
$$0 < t < 6.$$

The constants in the above formula were chosen to provide a severe test of the inversion algorithm. Figure 2 shows the results of two such tests. If 120 values of $R(0,t)$ are used with equally spaced values of t, the reconstruction is virtually indistinguishable from the true G. Figure 2 shows the reconstructions using every second and every fourth data value for R.

Example 2: This example tests the performance of the inversion algorithm in the presence of noise. Again, G was chosen to severely test the inversion procedure, being given by

$$G(t) = (1 + 3t + t^2)e^{-t}, \quad 0 < t < 6.$$

Such a G can be thought of as representing a modified Debye medium. Gaussian noise with zero mean and 0.001 variance was added to the reflection kernel, yielding a data set with signal to noise ratio of 7.8. This noisy kernel was then smoothed three times using a five point linear least squares smoother. This smooth data was then used twice in the inversion algorithm. First, the full set of data was employed, then every other data point was employed, and finally G was

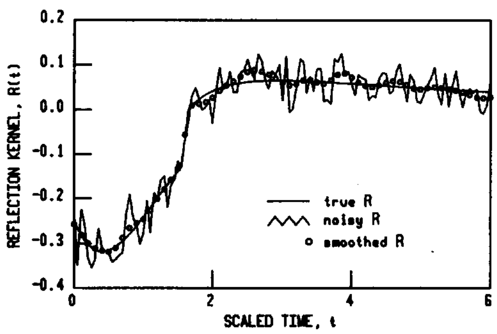


FIG. 3. Reflection kernel for the modified Debye medium given in Sec. VIII, example 2.

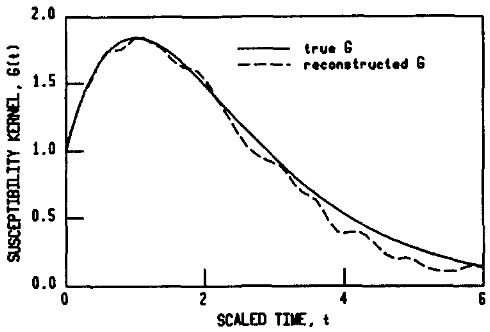


FIG. 4. Reconstruction of the susceptibility kernel for example 2 using noisy data.

determined by passively extrapolating these two results.

Figure 3 shows the true R produced by the forward algorithm, along with the noisy and smoothed data. For graphical clarity, every other smoothed R data point is shown. The jump in the kernel at $t = 1.6$ corresponds to the completion of the first round trip through the medium. Figure 4 shows the result of the reconstruction using the smoothed R. Notice the effects of accumulating error on the quality of the reconstruction. In the absence of noise, the reconstructed G is indistinguishable from the true G.

IX. SUMMARY

A method for solving one-dimensional electromagnetic scattering and inverse scattering problems for homogeneous dispersive media has been presented. The method is based on an integrodifferential equation which relates the susceptibility kernel and reflection kernel for the medium. A numerical implementation of these techniques has been demonstrated, along with an example of the effects of noise on the reconstruction. Under suitable assumptions the inverse problem for a stack of homogeneous dispersive layers has also been considered.

The constitutive relation which models the dispersive behavior in this problem is quite limited. However, it is felt that the technique presented in this paper can be considerably expanded so that more general problems can be studied. For example, the problem of inhomogeneities in the medium is not addressed in this paper. However, work is currently underway to solve this problem using techniques similar to those used in Secs. II and III. Another shortcoming of the model used in this paper is that it is not suitable for good conductors. Nonlocal spatial effects must be added to the model in such a case.

ACKNOWLEDGMENTS

This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under Contract No. W-7405-ENG-82, and in part by the Office of Naval Research, Contract No. N0014-83-K-0038.

APPENDIX A: EXISTENCE, UNIQUENESS, AND CONTINUOUS DEPENDENCE ON DATA FOR THE DIRECT SCATTERING PROBLEM FOR A SEMI-INFINITE MEDIUM

Set $H(t) = -4G(t)$ in Eq. (4.1) to obtain

$$R(t) = H(t) + [H*(2R + R*R)](t). \tag{A1}$$

This equation will be considered for $0 \leqslant t \leqslant T$, where T is an arbitrary positive number. Let

$$B = \{H \in L^2[0,T]: H \text{ bounded}\}.$$

Theorem 1: Assume $H \in B$. Then Eq. (A1) has a solution R, with $R \in B$.

Proof: Let H_B be a constant such that $|H(t)| \leqslant H_B$ for $0 \leqslant t \leqslant T$, and let S be the solution of

$$S(t) = H_B + [H_B*(2S + S*S)](t), \quad 0 \leqslant t \leqslant T.$$

Using Laplace transforms, it is easily verified that

$$S(t) = \exp(2H_B t) \cdot I_1(2H_B t)/t,$$

where I_1 is the modified Bessel function of the first kind. Thus, S is positive and bounded on $[0,T]$.

Now define a sequence of iterates,

$$R_0(t) = S(t),$$

$$R_{n+1}(t) = H(t) + [H*(2R_n + R_n*R_n)](t), \quad n \geqslant 0$$

for $0 \leqslant t \leqslant T$. It is now shown that each iterate satisfies

$$|R_n(t)| \leqslant S(t) \tag{A2}$$

on $[0,T]$. Clearly this is true for $n = 0$. Proceeding by induction, assume (A2) is true for $n = k$. Then

$$\begin{aligned} |R_{k+1}(t)| &\leqslant H_B + [H_B*(2|R_k| + |R_k|*|R_k|)](t) \\ &\leqslant H_B + [H_B*(2S + S*S)](t) \\ &= S(t), \end{aligned}$$

and the induction is complete.

Finally, it is shown that the sequence of iterates, $\{R_n\}$, converges in $L^2[0,T]$ to a function R which is a solution of (A1). Note that

$$R_1(t) - R_0(t) = H(t) - H_B + [(H - H_B)*(2S + S*S)](t). \tag{A3}$$

Upon setting

$$S_B = \max_{0 \leqslant t \leqslant T} S(t),$$

it follows from (A3) that

$$|R_1(t) - R_0(t)| \leqslant 2H_B[1 + S_B T(2 + S_B T)] \equiv D_B.$$

Now for any $n > 1$,

$$\begin{aligned} &|R_{n+1}(t) - R_n(t)| \\ &= |[(R_n - R_{n-1})*(2H + H*(R_n + R_{n-1}))](t)| \\ &\leqslant 2H_B(1 + S_B T)\int_0^t |R_n(s) - R_{n-1}(s)|ds. \end{aligned} \tag{A4}$$

It follows from (A4) by induction that

$$|R_{n+1}(t) - R_n(t)| \leqslant D_B[2H_B(1 + S_B T)]^n t^n/n!.$$

By comparison with the exponential series, it can now be seen that the series

$$R(t)=R_0(t)+\sum_{n=0}^{\infty}[R_{n+1}(t)-R_n(t)]$$
$$=\lim_{n\to\infty}R_n(t) \tag{A5}$$

converges pointwise almost everywhere for $0<t<T$. But since $H, S\in L^2[0,T]$ it follows that $R_n\in L^2[0,T]$ for all n. Hence, using (A2), (A5) and the dominated convergence theorem, R is a bounded square-integrable solution of (A1). This completes the proof.

Corollary: If H is continuous on $[0,T]$, then the solution R of Eq. (A1) is continuous on $[0,T]$.

Proof: If H is continuous, then clearly all the iterates R_n are continuous. But now the sequence (A5) converges uniformly on $[0,T]$, and the result follows.

Having established that Eq. (A1) has a solution, it is now shown the solution is unique.

Theorem 2: If $H\in B$, then the solution of Eq. (A1) is uniuqe in $L^2[0,T]$.

Proof: Suppose R and U are L^2 solutions of Eq. (A1). Then

$$R(t)-U(t)=[(R-U)*K](t),\quad 0<t<T, \tag{A6}$$

where

$$K(t)=2H(t)+[H*(R+U)](t).$$

Thus, $R-U$ must be an eigenfunction of a Volterra equation of the second kind with square integrable kernel K. It follows that the only solution of (A6) is the zero solution; i.e.,

$$R(t)=U(t)\quad \text{a.e.}$$

Next, the question of continuous dependence on data is addressed. Let $F:B\to B$ denote the mapping from H to R given by Eq. (A1).

Theorem 3: The mapping F is continuous; i.e., small changes in the L^2 norm of H produce small changes in the L^2 norm of R.

Proof: Let R_1,R_2 be the solutions of (A1) corresponding to H_1,H_2. Then

$$R_1(t)-R_2(t)=A(t)+[K*(R_1-R_2)](t), \tag{A7}$$

where

$$A(t)=H_1(t)-H_2(t)+[(H_1-H_2)*(2R_1+R_1*R_1](t),$$
$$K(t)=2H_2(t)+[H_2*(R_1+R_2)](t).$$

For H_1 and R_1 fixed, A can be made arbitrarily small in $L^2[0,T]$ by closing H_2 sufficiently close to H_1 in $L^2[0,T]$. This implies that R_1-R_2 can be made arbitrarily small in $L^2[0,T]$, since (A7) is a Volterra equation of the second kind. This completes the prooof.

Finally, note that the above proof can be easily modified to show that small changes in the sup norm of H produce small changes in the sup norm of R.

APPENDIX B: PROPAGATING REFLECTION DATA ACROSS DISCONTINUITIES

The following discussion refers to the problem considered in Sec. VI. In general, the reflection operator for a dispersive medium with piecewise constant characteristic velocity is given by

$$[\widetilde{R}(0^-)f](t)=[R_s(0^-,\cdot)*f](t)$$
$$=\left[\sum_{n=0}a_n\delta(t-s_n)+R(0^-,t)\right]*f(t),$$
$$=\sum_{n=0}a_nf(t-s_n)+[R(0^-,\cdot)*f](t). \tag{B1}$$

Here, R_s is the reflection kernel, including δ function singularities, and R denotes the classical (or nonsingular) portion of the reflection kernel. In Eq. (B1),

$$a_0=r^+=(c_1-c_0)/(c_1+c_0),\quad s_0=0,$$

and $s_n>0$ for $n>0$.

Substituting Eq. (B1) into Eq. (6.2) and rearranging yields

$$t^+t^-R_s(0^+,t)=\sum_{n>0}a_n[\delta(t-s_n)+r^+R_s(0^+,t-s_n)]$$
$$+R(0^-,t)+r^+R(0^-,t)*R_s(0^+,t), \tag{B2}$$

which is a delay Volterra equation for $R_s(0^+,t)$. This equation can first be used to separate out the δ function terms from $R_s(0^+,t)$, after which the kernel $R(0^+,t)$ can be uniquely determined.[20]

As an example of the form of the a_n's, s_n's, and $R_s(0^+,t)$ consider the situation of a single dispersive layer situated between $z=0$ and $z=L$. Then in Eq. (B1),

$$s_n=nT_1,$$

where $T_1=2L/c_1$. Also,

$$a_1=t^+t^-t_0^2r_1^+,$$

and

$$a_n=-t_0^2r_1^+r^+a_{n-1},\quad n\geqslant 2.$$

It follows from (B2) that

$$R_s(0^+,t)=t_0^2r_1^+\delta(t-T)+R(0^+,t),$$

where R satisfies

$$t^+t^-R(0^+,t)=r^+\sum_{n=1}^{\infty}a_nR(0^+,t-nT)$$
$$+R(0^-,t)+t_0^2r_1^+r^+R(0^-,t-T)$$
$$+r^+[R(0^-,\cdot)*R(0^+,\cdot)](t).$$

APPENDIX C: TRANSMISSION OPERATORS

For the finite slab model considered in Sec. III it is possible to consider a transmission operator as well as a reflection operator. Such an operator maps incident fields form one side of the slab into transmitted fields emerging from the other side of the slab. Thus, let $\widetilde{T}(z)$ denote the transmission operator which maps right-moving incident fields through the portion of the slab occupying $[z,L]$ into right-moving fields in the region $z>L$. Notice that since the medium is homogeneous, it is in fact not necessary to distinguish between incidence from the right and incidence from the left for either reflection or transmission.

The representation for the transmission operator is

$$[\widetilde{T}(z)f](t) = af(t-l) + \int_{-\infty}^{t-l} T(z,t-s)f(s)ds,$$

where

$$l = (L-z)/c, \quad a = \exp[-\tfrac{1}{2}lG(0^+)].$$

In a manner similar to that in Sec. III (or, see Ref. 12) it can be shown that

$$\widetilde{T}_z = -\widetilde{T}(\alpha + \beta\widetilde{R})$$

and

$$\widetilde{R}_z = -\widetilde{T}\beta\widetilde{T}.$$

In terms of kernels, this translates to

$$\begin{aligned} 2cT_z = 2T_t + aG'(t-l) + G(0) \\ \times [T + aR(z,t-l) + T*R] \\ + G'*[T + aR(z,t-l) + T*R] \end{aligned}$$

and

$$\begin{aligned} 2cR_z = a^2G'(t-2l) + G(0)[2aT(z,t-l) + T*T] \\ + G'*[2aT(z,t-l) + T*T], \end{aligned}$$

with

$$R(L,t) = 0,$$
$$[R(z,t)]_{t=2l^-}^{t=2l^+} = \tfrac{1}{4}G(0^+)e^{-lG(0^+)},$$
$$R(z,t) = R(L-ct/2,t), \quad 0 < z < L - ct/2.$$

[1] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), pp. 306–310.

[2] F. Bloom, *Ill-Posed Problems for Integrodifferential Equations in Mechanics and Electromagnetic Theory* (SIAM, Philadelphia, 1981), p. 101.

[3] H. E. Moses and C. M. deRidder, "Properties of dielectrics from reflection coefficients in one dimension," MIT Lincoln Lab. Tech. Report 322, 1963.

[4] E. Vogelzang, D. Yevick, and H. A. Ferwerda, Opt. Comm. **45**, 376 (1983).

[5] J. P. Corones, M. E. Davison, and R. J. Krueger, "Dissipative inverse problems in the time domain" to appear in *Proceedings NATO-ARW-IMEI*, edited by W-M. Boerner.

[6] R. J. Krueger, Q. Appl. Math. **36**, 235 (1978).

[7] R. J. Krueger, IEEE Trans. Antennas Propag. **AP-29**, 253 (1981).

[8] A. Tijhuis, IEEE Trans. Antennas Propag. **AP-29**, 239 (1981).

[9] V. Weston, J. Math. Phys. **13**, 1952 (1972).

[10] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1962), Vol. 2., p. 512.

[11] J. P. Corones, M. E. Davison, and R. J. Krueger, J. Acoust. Soc. Am. **74**, 1535 (1983).

[12] J. P. Corones, M. E. Davison, and R. J. Krueger, "Wave splittings, invariant imbedding and inverse scattering," in *Inverse Optics*, *Proceedings of SPIE* 413, edited by A. J. Devaney (SPIE, Bellingham, WA, 1983), p. 102.

[13] J. P. Corones and R. J. Krueger, J. Math. Anal. Appl. **95**, 393 (1983).

[14] M. Davison, "A general approach to splitting and invariant imbedding techniques for linear wave equations," to appear in J. Math. Anal. Appl.

[15] J. P. Corones, J. Math. Anal. Appl. **50**, 361 (1975).

[16] I. Stakgold, *Boundary Value Problems of Mathematical Physics* (Macmillan, New York, 1968), Vol. 2., pp. 44–46.

[17] H. Hochstadt, *Integral Equations* (Wiley, New York, 1975), pp. 31–34.

[18] A. Chelkowski, *Dielectric Physics* (Elsevier, Amsterdam, 1980), p. 97.

[19] M. Born and E. Wolf, *Principles of Optics* (Pergamon, New York, 1975), p. 61.

[20] R. J. Krueger, J. Math. Anal. Appl. **67**, 232 (1979).

Solitary wave solutions to the Einstein equations

P. S. Letelier
Departamento de Física, Universidade de Brasília, 70 910-Brasília, Brazil

(Received 25 May 1984; accepted for publication 9 August 1984)

The soliton solutions to the vacuum Einstein equations generated by the special class of Einstein–Rosen metrics described by linear combinations of homogeneous solutions to the usual cylindrically symmetric wave equation are studied.

I. INTRODUCTION

Recently we studied the problem of solving the vacuum Einstein equations for cylindrically symmetric solitary waves using the inverse scattering method.[1] We found that for the particular class of "seed" solutions known as Einstein–Rosen waves the inverse scattering method reduces to the problem of finding exact solutions to the system of equations[1]

$$(t\,\partial_t - \lambda\,\partial_r + 2\lambda\,\partial_\lambda)F = t\phi_{,t}, \tag{1.1a}$$

$$(t\,\partial_r - \lambda\,\partial_t)F = t\phi_{,r}, \tag{1.1b}$$

with the boundary condition

$$F_{|\lambda=0} = \phi, \tag{1.1c}$$

where $(\ )_{,t} \equiv \partial_t$ and $(\ )_{,r} \equiv \partial_r$. The function ϕ depends on the variables t and r only and it is the "gravitational potential" that appears in the Einstein–Rosen metric,[2] i.e.,

$$ds^2 = (e^{\sigma_{(0)}}/\pm\sqrt{t})(dt^2 - dr^2) - t(e^{\phi}\,d\theta^2 + e^{-\phi}\,dz^2), \tag{1.2}$$

where

$$2\sigma_{(0)} \equiv \int t\,[(\phi_{,t}^2 + \phi_{,r}^2)dt + 2\phi_{,t}\phi_{,r}\,dr]. \tag{1.3}$$

The integrability condition of (1.3) is

$$\phi_{,tt} + \phi_{,t}/t - \phi_{,rr} = 0. \tag{1.4}$$

The vacuum Einstein equations for the metric (1.2) are equivalent to (1.3) and (1.4). The function F depends on the variables t and r, and on the spectral parameter λ that, in general, is a complex parameter. The function F is related to the function Λ by

$$F = \ln\Lambda, \tag{1.5}$$

and Λ is closely related to the wave function Ψ_0, solution to the "Schrödinger equations" used in the inverse scattering method.[3]

We also found[1] the explicit form of the Λ functions associated to the solutions to (1.4) given by $\phi = 1$, $\phi = r$, and $\phi = r^2 + \frac{1}{2}t^2$. These three solutions are homogeneous functions of degree 0, 1, and 2, respectively.

The purpose of this paper is to find a class of homogeneous functions of degree n that are solutions to (1.4) and their corresponding Λ functions, and to use these functions to construct soliton solutions to the Einstein equations.

In Sec. II we study particular cases of Einstein–Rosen waves that can be constructed using a class of homogeneous solutions to (1.4). Their corresponding Λ functions are presented in Sec. III. In Sec. IV we give the metric for one-, two-, and N-soliton solutions to the vacuum Einstein equations. Finally, in Sec. V we study some of the results.

II. A CLASS OF EINSTEIN–ROSEN METRICS

A direct verification shows that the polynomials

$$L_n(t,r) = \sum_{k=0}^{[n/2]} \frac{n!}{(n-2k)!(k!)^2}\, r^{n-2k}t^{2k} \tag{2.1}$$

are homogeneous solutions of degree n to Eq. (1.4), where $[\cdots]$ means the integer part of the enclosed number. The first five polynomials are

$$L_0 = 1, \tag{2.2}$$

$$L_1 = r, \tag{2.3}$$

$$L_2 = r^2 + \tfrac{1}{2}t^2, \tag{2.4}$$

$$L_3 = r^3 + \tfrac{3}{2}t^2 r, \tag{2.5}$$

$$L_4 = r^4 + 3r^2t^2 + \tfrac{3}{8}t^4. \tag{2.6}$$

The polynomials L_n are related to the zonal harmonics by

$$L_n(t,r) = \rho^n P_n(\cos\theta), \tag{2.7}$$

where

$$t = i\rho\sin\theta, \quad r = \rho\cos\theta. \tag{2.8}$$

Using the definition (2.1) one can show the following useful identities:

$$L_{n,r} = nL_{n-1}, \tag{2.9}$$

$$L_n(t, r+a) = \sum_{i=0}^{n} \binom{n}{i} L_i a^{n-1}, \tag{2.10}$$

$$L_n(-t,r) = L_n(t,r), \tag{2.11}$$

$$L_n(t,-r) = (-1)^n L_n(t,r), \tag{2.12}$$

$$L_n(0,r) = r^n, \tag{2.13}$$

$$L_{2n}(t,0) = [(2n)!/(n!)^2 2^{2n}]\,t^{2n}, \tag{2.14}$$

where, as usual $\binom{n}{i} = n!/(n-i)!i!$.

An integral representation of the functions (2.1) is given by

$$L_n = \frac{1}{\pi}\int_0^{\pi} (r + t\cos\theta)^n\, d\theta. \tag{2.15}$$

Due to the linearity of (1.4), a linear combination of functions L_n,

$$\phi = \sum_{i=0}^{n} a_i L_i, \tag{2.16}$$

is also a solution to (1.4). A particularly interesting linear

 0022-2488/85/020326-03$02.50

combination is obtained by taking the constant coefficients a_n as

$$a_n = \frac{\pi}{n!}\frac{d^n}{dr^n} f(r)|_{r=0}, \tag{2.17}$$

where the function f is analytic. From (2.15)–(2.17) we find that the limit

$$\phi' = \lim_{n\to\infty} \sum_{i=0}^{n} a_i L_i, \tag{2.18}$$

is

$$\phi' = \int_0^{\pi} f(r + t\cos\theta)\, d\theta. \tag{2.19}$$

Thus the polynomials (2.1) are the "Taylor base" of the solutions to (1.4) that can be written[4] as (2.19).

The integral $\sigma_{(0)}$ that appears in the Einstein–Rosen metric can be easily computed when ϕ is taken as a polynomial L_n. We get

$$\sigma_{(0)n} \equiv \sigma_{(0)}[L_n] = \sum_{l,k=0}^{[n/2]} \frac{(n!)^2}{(n-2k)!(n-2l)!(k!l!)^2 2^{2(k+l)}} \times\left[\frac{kl}{k+l} + \frac{(n-2k)(n-2l)}{4(k+l+1)}\frac{t^2}{r^2}\right] r^{2(n-k-l)} t^{2(k+l)}. \tag{2.20}$$

For $k = l = 0$, one replaces $kl/(k+l)$ by its limit value. For the linear combination (2.16) we get

$$\sigma_{(0)} = \sum_{i=1}^{n} a^2{}_i \sigma_{(0)i} + \sum_{\substack{i=0\ j=0\\ i\neq j}}^{n\ \ n} a_i a_j \sigma_{(0)ij}, \tag{2.21}$$

where

$$\sigma_{(0)nm} = \int t\,[(L_{n,t}L_{m,t} + L_{n,r}L_{m,r})dt + (L_{n,t}L_{m,r} + L_{n,r}L_{m,t})dr]. \tag{2.22}$$

The existence of $\sigma_{(0)nm}$ is guaranteed by the fact that L_n satisfies (1.4). From (2.1) we get

$$\sigma_{(0)nm} = \sum_{k=0}^{[n/2]}\sum_{l=0}^{[m/2]} \frac{n!m!}{(n-2k)!(m-2l)!(k!l!)^2 2^{2(k+l)}} \times\left(\frac{2kl}{k+l} + \frac{(n-2k)(m-2l)}{2(n+l+1)} \times\frac{t^2}{r^2}\right) r^{n+m-2k-2l} t^{2(k+l)}. \tag{2.23}$$

Note that

$$\sigma_{(0)mm} = 2\sigma_{(0)m}. \tag{2.24}$$

Thus, in principle we can compute the function $\sigma_{(0)}$ associated to (2.19) by taking the coefficients a_i given by Eq. (2.17) and doing $n\to\infty$ in (2.21).

III. THE FUNCTION $F = \ln A$

Equations (1.1a) and (1.1b) with the boundary condition (1.1c) for the functions L_n are

$$(t\,\partial_t - \lambda\,\partial_r + 2\lambda\,\partial_\lambda)F_n = tL_{n,t}, \tag{3.1a}$$

$$(t\,\partial_r - \lambda\,\partial_t)F_n = tL_{n,r}, \tag{3.1b}$$

$$F_{n|\lambda=0} = L_n. \tag{3.1c}$$

From (1.4) and (2.9) we find

$$F_{n,r} = nF_{n-1}. \tag{3.2}$$

One can prove by induction that

$$F_n = \frac{n}{\lambda}\int t(F_{n-1} - L_{n-1})dt + \left(r + \frac{1}{2}\lambda\right)^n \tag{3.3}$$

satisfies Eq. (3.1). The expression (3.3) can be used to compute the function F_n associated to the particular cases of L_n given by (2.2)–(2.8); we find

$$F_0 = 1, \tag{3.4}$$

$$F_1 = r + \tfrac{1}{2}\lambda, \tag{3.5}$$

$$F_2 = (r + \tfrac{1}{2}\lambda)^2 + \tfrac{1}{2}t^2, \tag{3.6}$$

$$F_3 = (r + \tfrac{1}{2}\lambda)^3 + \tfrac{3}{2}t^2(r + \lambda/4), \tag{3.7}$$

$$F_4 = (r + \tfrac{1}{2}\lambda)^4 + 3(r^2 + \tfrac{1}{2}\lambda r + \tfrac{1}{12}\lambda^2)t^2 + \tfrac{3}{8}t^4. \tag{3.8}$$

The same expression (3.3) can be used for the generic case; we get

$$F_n = \sum_{l=0}^{n}\binom{n}{l}2^{-l}\lambda^{-l}\left(r + \frac{\lambda}{2}\right)^{n-l} t^{2l} - \sum_{l=1}^{n}\sum_{k=0}^{[(n-l)/2)]} \frac{n!r^{n-l-2k}t^{2k+2l}}{(n-l-2k)!(k+l)!k!2^{2k+l}\lambda^l}. \tag{3.9}$$

The functions A associated to the functions ϕ and ϕ' given by (2.16) and (2.19) are, respectively,

$$A = \exp\left(\sum_{i=0}^{n} a_i F_i\right), \tag{3.10}$$

$$A' = \exp\left(\sum_{i=0}^{n} a_i F_i\right)\Bigg|_{n\to\infty}; \tag{3.11}$$

the coefficients a_i in the last case are given by (2.17).

IV. SOLITON SOLUTIONS

For a digaonal "seed" solution like (1.2)–(1.4) the corresponding soliton solutions can be cast as[1]

$$ds^2 = (e^\sigma / \pm\sqrt{t})(dt^2 - dr^2) - \gamma_{11}\,d\theta^2 - 2\gamma_{12}\,d\theta\,dz - \gamma_{22}\,dz^2. \tag{4.1}$$

The one-soliton solution is characterized by

$$\gamma_{11} = t\,[\cosh(p+\delta)/\cosh(x+\delta)]\,e^{\phi}, \tag{4.2a}$$

$$\gamma_{12} = -\eta t\sinh y/\cosh(x+\delta), \tag{4.2b}$$

$$\gamma_{22} = t\,[\cosh(q+\delta)/\cosh(x+\delta)]\,e^{-\phi}, \tag{4.2c}$$

$$\sigma = \sigma_{(0)} + \ln[t^{-1/2}\cosh(x+\delta)/\sinh y] + \ln C_1, \tag{4.3}$$

where

$$x = \phi - 2\ln A_1, \tag{4.4a}$$

$$p = \phi - \ln(\mu_1 A_1^2/t), \tag{4.4b}$$

$$q = \phi - \ln(t A_1^2/\mu_1), \tag{4.4c}$$

$$2y = q - p, \tag{4.4d}$$

$$\eta = m_1 m_2/|m_1 m_2|, \tag{4.5a}$$

$$\tanh\delta = [(m_1)^2 - (m_2)^2]/[(m_1)^2 + (m_2)^2]. \tag{4.5b}$$

The function μ_k is defined as

$$\mu_k = \alpha_k - r \pm [(\alpha_k - r)^2 - t^2]^{1/2}, \tag{4.6}$$

and A_k as

$$A_k \equiv A_{|\lambda = \mu_k}. \tag{4.7}$$

The index k runs from 1 to N, and α_k, m_1, and m_2 are arbitrary constants.

The two-soliton solution is characterized by

$$\gamma_{11} = t\frac{[t(\mu_2-\mu_1)P_1]^2 + [(\mu_1\mu_2 - t^2)P_2]^2}{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_1\mu_2 - t^2)S_2]^2}e^{\phi}, \tag{4.8a}$$

$$\gamma_{12} = t\frac{(\mu_2-\mu_1)(\mu_1\mu_2 - t^2)}{\mu_1\mu_2}\,\frac{\mu_2(\mu_1^2 - t^2)m_{01}^{(1)}m_{02}^{(1)}T_1 - \mu_1(\mu_2^2 - t^2)m_{01}^{(2)}m_{02}^{(2)}T_2}{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_1\mu_2 - t^2)S_2]^2}, \tag{4.8b}$$

$$\gamma_{22} = t\frac{[t(\mu_2-\mu_1)Q_1]^2 + [\mu_1\mu_2 - t^2)Q_2]^2}{[t(\mu_2-\mu_1)S_1]^2 + [\mu_1\mu_2 - t^2)S_2]^2}e^{-\phi}, \tag{4.8c}$$

$$\sigma = \sigma_{(0)} - \ln[(\mu_1^2 - t^2)(\mu_2^2 - t^2)(\mu_1\mu_2 - t^2)^2(1/\mu_1 - 1/\mu_2)^2] + \sigma' + \ln C_2, \tag{4.9a}$$

$$\sigma' = \ln\{[t(\mu_2-\mu_1)S_1]^2 + [(\mu_1\mu_2 - t^2)S_2]^2\}. \tag{4.9b}$$

The functions $S_1, S_2, P_1, P_2, Q_1, Q_2, T_1$, and T_2 depend on μ_1, μ_2, A_1, A_2, and ϕ, and on the set of four constants $\{m_{02}^{(1)}, m_{01}^{(1)}, m_{01}^{(2)}, m_{02}^{(2)}\}$. The explicit form of these functions can be found in Ref. 1. Depending on the value of the above-mentioned constants one can cast (4.8) in three different forms, in terms of hyperbolic and circular functions.[1]

The solutions (4.2) and (4.3) and (4.8) and (4.9) are the most general solutions that can be obtained using the inverse scattering method with (1.2) as a seed solution.

A particular case of an N-soliton[1] is characterized by

$$\gamma_{22} = t^2/\gamma_{11}, \tag{4.10a}$$

$$\gamma_{12} = 0, \tag{4.10b}$$

$$\gamma_{11} = \left[\prod_{l=1}^{N}\left(\frac{\mu_l}{t}\right)\epsilon_l\right]te^{\phi}, \tag{4.10c}$$

$$\begin{aligned}\sigma = \sigma_0 &- \phi\sum_{l=1}^{N}\epsilon_l + 2\sum_{l=1}\epsilon_l \ln A_l \\ &+ \frac{1}{2}\ln t\left(\sum_{l=1}^{N}\epsilon_l\right)^2 - \sum_{l,k=1}^{N}\epsilon_l\,\epsilon_k \ln\mu_l \\ &+ \sum_{l}^{N}\epsilon_l^2 \ln\frac{\mu_l^2}{\mu_l^2 - t^2} \\ &+ 2\sum_{l>k=1}^{N}\epsilon_l \ln(\mu_l - \mu_k) + \ln C_n,\end{aligned} \tag{4.11}$$

where $\epsilon_l = \pm 1$. Note that if one takes ϵ_l as a set of arbitrary constants, (4.10) and (4.11) also give a solution to the Einstein equations.

V. DISCUSSION

First we note that depending on the sign of the square root that appears in (1.2) and (4.1) the roles of t and r can be interchanged. The metric (1.2), as well as the field equation (1.4), have a singularity at $t = 0$, nevertheless the polynomials L_n are not singular at the above-mentioned instant. The metric (1.2), the field equation (1.4), and the polynomials L_n are well behaved at $r = 0$.

When t is a timelike variable the metric (1.2) describes a cosmological model with a big-band type of singularity, and in the complementary case, i.e., when t is a spacelike variable, this metric represents a cylindrically symmetric gravitational wave with a singularity on the axis of symmetry $t = 0$. The metric (1.2) can also be interpreted as an exact perturbation of the plane-symmetric solution to the vacuum Einstein equations, since letting $\phi = 0$ in (1.2), we end up with the well-known Taub metric. Furthermore, taking $\phi = a_0$, $L_0 = a_0$, redefining θ and z in (1.2), the Taub metric is also obtained. When t represents a timelike variable the Taub metric is a special case of the Kasner metric, i.e., a special type of Bianchi I cosmological model.[5]

The soliton solutions constructed using L_n as seed solutions describe exact perturbations of either a cosmological model or a cylindrical wave depending on the timelike or spacelike character of the t coordinate as Eq. (4.1) indicates.[1] In the case of a "perturbed" cosmological model we can say that the solitons are created near the big band. And in the case of "perturbed" cylindrical waves the solitons are incident and reflected from the axis of symmetry.[3,6]

The solitons' velocity of propagation, as well as their other properties like the position of the "bumps," shape, etc., depend on the particular form of the function ϕ and on its functionally related function A, and on the value of the constants $m_{0k}^{(l)}$ and α_k. Special cases of one- and two-soliton solutions are studied in Refs. 1 and 7.

In general the Einstein–Rosen solutions will diverge at $r, t \to \infty$, and in consequence, their associated soliton solutions will have the same singular behavior. A discussion of this point for the elliptic case can be found in Ref. 8.

[1] P. S. Letelier, J. Math. Phys. **25**, 2675 (1984).

[2] See, for instance, J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1966), p. 352. The equivalence of (1.2) with the Einstein–Rosen metric can be proved performing the transformation $\phi \to -\phi + \ln t$.

[3] V. A. Belinsky and V. E. Zakharov. Zh. Eksp. Teor. Fiz. **75**, 1955 (1978) [Sov. Phys. JETP **48**, 985 (1976)].

[4] Similar expressions like (2.19) can be found in H. Lamb, *Hydrodynamics* (Dover, New York, 1945), p. 298; E. T. Whittaker and G. N. Watson, *A Course in Modern Analysis* (Cambridge U. P., Cambridge, 1962), p. 399. Their use in the context of general relativity can be found in P. S. Letelier and R. Tabensky, J. Math. Phys. **16**, 8 (1975); P. C. Waylen. Proc. R. Soc. London Ser. A **382**, 467 (1982).

[5] See, for instance, M. P. Ryan, Jr. and L. C. Shepley, *Homogeneous Relativistic Cosmologies* (Princeton U. P., Princeton, 1975), p. 133.

[6] P. S. Letelier, Phys. Rev. D **26**, 2623 (1982).

[7] V. A. Belinsky and D. Fargion, Nuovo Cimento B **59**, 143 (1980).

[8] P. S. Letelier, Phys. Rev. D **26**, 3728 (1982), and references therein.

The well-posedness of ($N = 1$) classical supergravity

David Bao
School of Mathematics, The Institute for Advanced Study, Princeton, New Jersey 08540, and Department of Mathematics, University of Houston–University Park, Houston, Texas 77004

Yvonne Choquet-Bruhat
IMTA, Université Paris VI, Paris, France

James Isenberg
Department of Mathematics, University of Oregon, Eugene, Oregon 97403-1222

Philip B. Yasskin
Department of Mathematics, Texas A & M University, College Station, Texas 77843

(Received 17 July 1984; accepted for publication 24 August 1984)

In this paper we investigate whether classical ($N = 1$) supergravity has a well-posed locally causal Cauchy problem. We define well-posedness to mean that any choice of initial data (from an appropriate function space) which satisfies the supergravity constraint equations and a set of gauge conditions can be continuously developed into a space-time solution of the supergravity field equations around the initial surface. Local causality means that the domains of dependence of the evolution equations coincide with those determined by the light cones. We show that when the fields of classical supergravity are treated as formal objects, the field equations are (under certain gauge conditions) equivalent to a coupled system of quasilinear nondiagonal second-order partial differential equations which is formally nonstrictly hyperbolic (in the sense of Leray–Ohya). Hence, if the fields were numerical valued, there would be an applicable existence theorem leading to well-posedness. We shall observe that well-posedness is assured if the fields are taken to be Grassmann (i.e., exterior algebra) valued, for then the second-order system decouples into the vacuum Einstein equation and a sequence of numerical valued linear diagonal strictly hyperbolic partial differential equations which can be solved successively.

I. INTRODUCTION

Soon after the equations of supergravity were first proposed,[1,2] the system was shown to have a consistent initial value formulation.[3–5] That is, one can write the field equations as (i) a set of constraints on the choice of initial data (on some spacelike Cauchy three-surface) and (ii) a set of evolution equations for the time development of the initial data, with the evolution equations preserving the constraints.

However, the existence of a consistent initial value formulation does not guarantee that one can (even in principle) use it to find solutions, except in the analytic case (but then causality may be violated). One says that a system of field equations with a consistent initial value formulation is well-posed if any choice of initial data which satisfies the constraints can be evolved into a nonsingular space-time solution which depends continuously (with respect to some appropriate function space topology) upon the choice of initial data. The solution may not necessarily extend for infinite time, but it is guaranteed that the solution exists in an open space-time neighborhood of the initial surface. A well-posed system of field equations is said to be locally causal if its domains of dependence coincide with those determined by the light cones; that is, the initial data on a compact subset of the initial three-surface uniquely determines (up to gauge transformations) the solution at points whose causal past intersects the initial surface within the compact set.[6]

Most of the important field theories of theoretical physics are known to be well-posed and locally causal. This includes Maxwell, Dirac, Klein–Gordon, Yang–Mills, Higgs, Einstein, and various coupled combinations of the above. Recently, Choquet-Bruhat[7] has shown that when Grassmann (exterior algebra) valued fields are used, classical ($N = 1$) supergravity is also well-posed and locally causal. For the Grassmann formulation of supergravity, we give a similar proof which differs from the one in Ref. 7 in three ways.

(1) We impose a different set of gauge conditions.

(2) We assume that the spin-$\frac{3}{2}$ gravitino is a Majorana spinor, whereas that in Ref. 7 is a Weyl spinor.

(3) In Ref. 7 one works with the original supergravity field equations; here we work with an equivalent (under gauge conditions) second-order system. As will be discussed, the Grassmann formulation leads to a decoupling (noted already in Ref. 8) of the field equations, and the basic idea in the aforementioned proofs is a systematic exploitation of this decoupling.

There are mixed reactions among researchers towards this decoupling.[9] To some, the decoupling seems to be unphysical in that the gravitino does not affect the rank zero part of the tetrad which completely controls the characteristics of the wave operators in all of the decoupled field equations (see Sec. IV of this paper). To others, such decoupling is perfectly consistent with their experience from quantum field theories. Indeed, it must be emphasized that there is as yet no concrete physical justification for assuming that the supergravity fields are Grassmann-valued or otherwise. We therefore find it instructive to examine the well-posedness problem within the realm of formal supergravity, in which the fomal rules for manipulating the fields are satisfied by letting the latter take values in some hypothetical $\mathbf{Z}_2$-graded

 0022-2488/85/020329-05$02.50

algebra $\mathfrak{A} = \mathfrak{A}^+ \oplus \mathfrak{A}^-$, where elements of $\mathfrak{A}^+$ are commuting (bosonic) and those of $\mathfrak{A}^-$ are anticommuting (fermionic). A detailed description of the requisite properties of $\mathfrak{A}$ is given in Refs. 9–11. The algebra is not *a priori* Grassmann nor graded by the non-negative integers. Thus, in general, the field equations do not decouple. Nevertheless, the field equations are found to be equivalent (under certain gauge conditions) to a quasilinear nondiagonal second-order system of coupled partial differential equations which is nonstrictly hyperbolic (*à la* Leray–Ohya[12]) in a formal sense. It follows that if such a system were numerical valued, it would be well-posed in Gevrey classes of C^∞ functions.

II. NOTATIONS AND PRELIMINARIES

The basic fields of ($N = 1$) supergravity are the spin-$\frac{3}{2}$ fermion field $\psi = \psi_\mu dx^\mu$ (an anticommuting, i.e., components in $\mathfrak{A}^-$, Majorana spinor valued one-form) and the spin-2 boson field $e^{\hat{\alpha}} = e^{\hat{\alpha}}{}_\mu\, dx^\mu$ (a commuting, i.e., components in $\mathfrak{A}^+$, Lorentz vector valued one-form). The field equations for supergravity are

$$G^{\mu\xi} = \epsilon'^{\lambda\mu\nu\rho}\bar{\psi}_\lambda \gamma_5 \gamma^\xi D_\nu \psi_\rho, \tag{1}$$

$$(i_\lambda D\psi)_\lambda := \gamma^\xi (D_\xi \psi_\lambda - D_\lambda \psi_\xi) = 0, \tag{2}$$

with torsion

$$Q^\xi{}_{\mu\nu} = -\tfrac{1}{4}\bar{\psi}_\mu \gamma^\xi \psi_\nu. \tag{3}$$

Here, $\gamma_{\hat{\alpha}}$ are the standard Dirac matrices with $\gamma_\mu := \gamma_{\hat{\alpha}} e^{\hat{\alpha}}{}_\mu$ and $\gamma_5 := \gamma^{\hat{0}}\gamma^{\hat{1}}\gamma^{\hat{2}}\gamma^{\hat{3}}$, $\epsilon'^{\gamma\mu\nu\rho}$ is the Levi–Civita tensor (not density), $G^{\mu\xi}$ is the Einstein tensor of the metric-compatible connection $\Gamma^\xi{}_{\nu\mu}$ with torsion $Q^\xi{}_{\mu\nu} := \frac{1}{2}(\Gamma^\xi{}_{\mu\nu} - \Gamma^\xi{}_{\nu\mu})$, and our metric has signature $(-+++)$. The torsion equation (3) will be used as an identity throughout the paper. Note our index conventions: $\hat{\alpha}$ are $O(3, 1)$ frame indices, while μ are space-time coordinate indices. On scalar spinors ϕ we use the notation $D\phi := (D_\nu \phi)dx^\nu$; similarly, if ϕ is a spinor valued one-form, we define $i_\gamma \phi := \gamma^\nu \phi_\nu$ and $D\phi := (D_\mu \phi_\nu - D_\nu \phi_\mu) dx^\mu \wedge dx^\nu$.

We remind the reader that here

$$D_\nu \psi_\rho = \partial_\nu \psi_\rho + \tfrac{1}{2}\Gamma^{\hat{\alpha}\hat{\beta}}{}_\nu \sigma_{\hat{\alpha}\hat{\beta}} \psi_\rho, \tag{4}$$

where $\Gamma^{\hat{\alpha}\hat{\beta}}{}_\nu$ is related to $\Gamma^\xi{}_{\mu\nu}$ via the change-of-basis formula

$$\partial_\nu e^{\hat{\alpha}}{}_\mu + e^{\hat{\beta}}{}_\mu \Gamma^{\hat{\alpha}}{}_{\hat{\beta}\nu} - e^{\hat{\alpha}}{}_\xi \Gamma^\xi{}_{\mu\nu} = 0. \tag{5}$$

The presence of $\epsilon'^{\lambda\mu\nu\rho}$ in (2) antisymmetrizes the ν, ρ indices of $D_\nu \psi_\rho$; therefore, $D_\nu \psi_\rho$ can be replaced by

$$\overline{D_\nu} \psi_\rho = \partial_\nu \psi_\rho + \tfrac{1}{2}\Gamma^{\hat{\alpha}\hat{\beta}}{}_\nu \sigma_{\hat{\alpha}\hat{\beta}} \psi_\rho - \{{}^\xi{}_{\rho\nu}\}\psi_\xi$$

(i.e., correcting the space-time coordinate index ρ with the torsion-free Christoffel connection) without affecting (2). Since $\overline{D_\nu}\psi_\rho$ is covariant, it is preferred by some authors, for example, Choquet-Bruhat[7] and Yasskin.[13] We will use $D_\nu \psi_\rho$ in this paper in order to facilitate comparison with most of the articles listed in the bibliography.

Equation (2) is the Rarita–Schwinger field equation and is frequently written as

$$(\mathrm{RS})^\lambda := \epsilon'^{\lambda\mu\nu\rho}\gamma_5 \gamma_\mu D_\nu \psi_\rho = 0. \tag{6}$$

Equations (2) and (6) are algebraically equivalent through the use of the identities[10,14]

$$(i_\gamma D\psi)^\lambda = (\mathrm{RS})^\lambda - \tfrac{1}{2}\gamma^\lambda i_\gamma (\mathrm{RS}) \tag{7}$$

$$(\mathrm{RS})^\lambda = (i_\gamma D\psi)^\lambda - \tfrac{1}{2}\gamma^\lambda i_\gamma i_\gamma D\psi. \tag{8}$$

Clearly, (7) can be obtained from (8) (and vice versa) by "trace-reversing": contract with γ^λ and use $\gamma^\lambda \gamma_\lambda = 4$. We note for later purposes that *if* a configuration (e, ψ) satisfies (1) and (3), then

$$D_\lambda (\mathrm{RS})^\lambda = 0. \tag{9}$$

Equation (9) is the result which guarantees that supergravity is consistent in the sense of Buchdal: taking the divergence (with D) of the Rarita–Schwinger field equation does not introduce new conditions "on shell." A detailed proof of (9) may be found in Zumino[15]; an outline of the strategy of that proof, using our notation and conventions, is given in Ref. 10.

We now discuss the gauge conditions which we will be using later in the paper:

$$g^{\mu\nu}\{{}^\lambda{}_{\mu\nu}\} = 0 \quad \text{(harmonic gauge)}, \tag{10}$$

$$\partial^\mu \Gamma^{\hat{\alpha}}{}_{\hat{\beta}\mu} = 0 \quad [O(3, 1)\text{gauge}^{9,10,16-18}], \tag{11}$$

and

$$i_\gamma \psi = 0 \quad \text{(Rarita–Schwinger gauge)}. \tag{12}$$

A well-known consequence of the harmonic gauge is that[19]

$$\tilde{R}_{\mu\nu} = -\tfrac{1}{2}\Box g_{\mu\nu} + O(g, \partial g), \tag{13}$$

where $\tilde{R}_{\mu\nu}$ is the Ricci curvature tensor of the torsion-free Christoffel connection $\{\ \}$, $\Box := g^{\mu\nu}\partial_\mu \partial_\nu$ is the d'Alembertian operator, and $O(g, \partial g)$ is a functional of g and its first derivatives. If one imposes *both* the harmonic and $O(3, 1)$ gauges, then a noteworthy result[10,16,17] is the reduction of the statement $e_{\hat{\alpha}\nu}\Box e^{\hat{\alpha}}{}_\mu = e_{\hat{\alpha}(\nu}\Box e^{\hat{\alpha}}{}_{\mu)} + e_{\hat{\alpha}[\nu}\Box e^{\hat{\alpha}}{}_{\mu]}$ (parentheses denote symmetrization and brackets denote antisymmetrization) to

$$e_{\hat{\alpha}(\nu}\Box e^{\hat{\alpha}}{}_{\mu)} = e_{\hat{\alpha}\nu}\Box e^{\hat{\alpha}}{}_\mu + O(1). \tag{14}$$

Throughout this paper, $O(n)$ will abbreviate terms containing at most the first n derivatives of (e, ψ).

As a consequence of Eqs. (13) and (14), we obtain a result which is important in our well-posedness proofs—that the Einstein field equations (1) can be rewritten in the reduced form

$$\Box e^{\hat{\alpha}}{}_\mu = O(1). \tag{15}$$

To derive (15), one first writes (1) in the trace-reversed form $R_{\mu\nu} = T_{\mu\nu} - \frac{1}{2}g_{\mu\nu}T^\lambda{}_\lambda$, where T is the right-hand side of (1). (Incidentally, note that due to the presence of torsion, neither $R_{\mu\nu}$ nor $G_{\mu\nu}$ is symmetric.) Using the torsion equation (3) and the explicit form of the stress energy T [see (1)], we can rewrite the above as $\tilde{R}_{\mu\nu} = O(1)$, which, upon the application of (13) and (14), gives (15).

Another result we shall need later is that, if (e, ψ) satisfies the Einstein field equations (1), as well as the harmonic and $O(3,1)$ gauges, then, for any Majorana spinor valued one-form $\phi = \phi_\lambda\, dx^\lambda$, we have

$$[(i_\gamma D + Di_\gamma)^2 \phi]_\lambda = \Box\phi_\lambda - R^\xi{}_{\lambda\mu\nu}\sigma^{\mu\nu}\phi_\xi + O(1), \tag{16}$$

where $O(1)$ is a functional of (e, ψ, ϕ) and their first derivatives. Note that the operator D depends on (e, ψ). The verification of (16) is tedious and is summarized in Ref. 10.

III. THE SECOND-ORDER SYSTEM

To show that the Cauchy problem for the supergravity field equations (1) and (2) is (formally) well-posed, we shall work with an auxiliary system

$$G^{\mu\xi} = \epsilon^{\lambda\mu\nu\rho}\, \overline{\psi_\lambda}\gamma_5\gamma^\xi D_\nu \psi_\rho, \tag{17}$$

$$(i_\gamma D + D i_\gamma)^2 \psi = 0. \tag{18}$$

[Note that (17) is identical to (1), and is merely reproduced here for convenience.] Unlike Eqs. (1) and (2), which are mixed in order (second derivatives on e, first derivatives on ψ), this auxiliary system is purely second order. In the next section, we shall discuss the well-posedness of the Cauchy problem for Eqs. (17) and (18). Here, we show that when the Rarita–Schwinger gauge condition (12) is imposed, the system (17) and (18) is, in a certain sense, equivalent to the system (1) and (2). Specifically, we show that if initial data for the mixed system (1) and (2) is chosen so that it satisfies the usual supergravity constraints[3] (obtained through a canonical analysis)

$$\Phi(0,x) = 0, \tag{19}$$

then one can always extend it to data for the second-order system (17) and (18), which also satisfies

$$(i_\gamma \psi)(0,x) = 0, \tag{20a}$$

$$[\partial_t (i_\gamma \psi)](0,x) = 0, \tag{20b}$$

and

$$(i_\gamma D\psi)(0,x) = 0. \tag{20c}$$

Then (most importantly) the space-time fields (e, ψ) obtained by evolving this data using the second-order system (17) and (18) will always satisfy the supergravity field equations (1) and (2), together with the gauge condition $i_\gamma \psi = 0$, throughout space-time.

The extension of the initial data can always be done because (20) can be satisfied by fixing $\psi_0(0,x)$, $(\partial_t \psi_0)(0,x)$, and $(\partial_t \psi_i)(0,x)$, and these quantities are left unspecified by the supergravity Cauchy data.

Now let us be *given* a solution (e, ψ) of (17) and (18), satisfying the initial conditions (19) and (20). Define the spinor valued one-form

$$\phi := (i_\gamma D + D i_\gamma)\psi,$$

which is a functional of (e, ψ). Equation (18) then says that

$$(i_\gamma D + D i_\gamma)\phi = 0. \tag{21}$$

If we now apply the operator $(i_\gamma D + D i_\gamma)$ to (21) and use coordinate changes and Lorentz rotations to impose the harmonic and $O(3,1)$ gauge conditions on the fields (e, ψ), then it follows from (16) that

$$\Box\phi + \text{l.o.t.} = 0. \tag{22}$$

In (22), "l.o.t." is a functional which contains no higher than first derivatives of ϕ, and which is homogeneous in ϕ in the sense that l.o.t. $= 0$ if ϕ (and therefore all its derivatives) is zero.

Since $\phi(t,x) = 0$ is clearly a solution of the strictly hyperbolic system (22), it is the only solution if (i) $\phi(0,x) = 0$, and (ii) $(\partial_t \phi_\lambda)(0,x) = 0$. Condition (i) is guaranteed by the initial conditions (20); consequently one also has (iii) $(\partial_k \phi_\lambda)(0,x) = 0$, where k is a coordinate index on the initial three-surface. To verify (ii), we write out (20) in components, evaluate at $(0,x)$, and then use (i) and (iii).

We now know that

$$0 = \phi(t,x) = (i_\gamma D\psi + D i_\gamma \psi)(t,x). \tag{23}$$

It follows that if $(i_\gamma \psi)(t,x) = 0$, then the fields (e, ψ) will satisfy (1) and (2). In view of the initial conditions (20a) and (20b), it suffices to check that the propagation of the scalar spinor $i_\gamma \psi$ is governed by a homogeneous hyperbolic equation. This would at the same time tell us that the Rarita–Schwinger gauge condition $i_\gamma \psi = 0$ is compatible with the use of (17) and (18) to evolve initial data. The equation governing the propagation of $i_\gamma \psi$ is not hard to find. In fact, taking the divergence of (8) with D and using the conservation law (9) [which is applicable because the fields (e, ψ) satisfy the Einstein equation (17) and the torsion equation (3)], we get

$$D_\lambda (i_\gamma D\psi)^\lambda - \tfrac{1}{2} D_\lambda (\gamma^\lambda i_\gamma i_\gamma D\psi) = 0. \tag{24}$$

Now the (e, ψ) at hand satisfies (23), so (24) becomes

$$-D_\lambda (D i_\gamma \psi)^\lambda + \tfrac{1}{2} D_\lambda (\gamma^\lambda i_\gamma D i_\gamma \psi) = 0. \tag{25}$$

A routine computation shows that the left-hand side of (25) is $-\tfrac{1}{2}\Box(i_\gamma \psi) + \text{l.o.t.}$, where now l.o.t. is a functional containing no higher than first derivatives of $i_\gamma \psi$, with l.o.t. $= 0$ if $i_\gamma \psi$ (and hence its derivatives) is zero. Equation (25) is the homogeneous hyperbolic equation we need.

Note that since we do not have the closed form of finite supersymmetry transformations (for "rigid" ones in a special case, see Ref. 20) at our disposal, we are unable to show that every configuration (e, ψ) can be transformed into one which satisfies the Rarita–Schwinger gauge condition (12). As a result, we do not know whether solving the second-order system (17) and (18) with initial data which satisfies (19) and (20) will actually generate all solutions of the supergravity field equations (1) and (2) satisfying the usual supergravity constraints, nor do we know whether the solutions generated this way are unique up to gauge transformations. Obviously, settling the latter question in the affirmative will imply the same for the former.

IV. WELL-POSEDNESS OF THE SECOND-ORDER SYSTEM

In the harmonic and $O(3,1)$ gauges (10) and (11), it follows from (15) and (16) that our second-order system (17) and (18) can be rewritten in the reduced form

$$A \begin{pmatrix} e \\ \psi \end{pmatrix} = O(1), \tag{26}$$

where

$$A = \left(\begin{array}{c|c} \Box & 0 \\ \hline \text{some 2nd-order} & \\ \text{curvature operator} & \Box \end{array}\right)$$

consists of four 16×16 blocks. As usual, the harmonic and $O(3,1)$ gauge choices must be shown to be consistent with the system (26), which is invariant neither under change of coordinates nor under local Lorentz rotations. Consistency means that if the initial data obeys these gauge conditions and the usual constraints, then the (e, ψ) obtained by evolving such initial data with (26) also satisfies the gauge conditions.

This consistency can be proved by standard methods parallel in part to those give in Ref. 21 for the vacuum Einstein system, by using the various Noether identities which correspond to the invariances of the theory.

We now show that the second-order quasilinear nondiagonal system (26) is nonstrictly hyperbolic in a formal sense. The symbol of A is the 32×32 matrix.

$$\sigma A = \left(\begin{array}{c|c} p^\xi p_\xi \cdot 1_{16\times 16} & 0 \\ \hline \text{quadratic in } p_\xi & p^\xi p_\xi \cdot 1_{16\times 16} \end{array}\right),$$

where p_ξ is an element of the cotangent bundle. The highest-order homogeneous part of $\det(\sigma A)$ is the characteristic polynomial h. Here

$$h = (p^\xi p_\xi)^{32} \tag{27a}$$

is homogeneous with

$$\deg(h) = 64. \tag{27b}$$

We note for later purposes that $h = h_1 \cdots h_{32}$, where

$$h_i = p^\xi p_\xi = \sigma(\Box), \tag{28a}$$

and

$$\deg(h_i) = 2. \tag{28b}$$

Let $m = (m_1,\ldots,m_{32})$, $n = (n_1,\ldots,n_{32})$. We say that A is of type (n, m) if its (j, k)th entry has order $\leqslant m_k - n_j$. For the case at hand, A is of type (n, m), with

$$m_k = 2,\ \forall k,\ n_j = 0,\ \forall j. \tag{29}$$

One of the tests for nonstrict hyperbolicity, as given in Ref. 22, is applicable here. It consists of verifying that

$$\deg(h) \text{ and } \deg(h_i) \text{ are both } \geqslant \sup m_k - \inf n_j. \tag{30}$$

In view of (27)–(29), the inequalities in (30) are clearly satisfied. Thus, if (26) were a numerical valued system, it would be nonstrictly hyperbolic in the sense of Leray–Ohya,[12] and would therefore be well-posed in Gevrey classes of C^∞ functions. Whether or not such existence theorems generalize to systems with $\mathfrak{A}$-valued fields is a matter of ongoing research.

For the special case in which $\mathfrak{A}$ is an exterior algebra, one does not have to worry about whether or not the Leray–Ohya results apply to $\mathfrak{A}$-valued fields. For in the exterior algebra case, one finds that the $\mathfrak{A}$-valued field equations (26) for the $\mathfrak{A}$-valued fields (e, ψ) break down into a sequence of ordinary field equations for ordinary C^∞ functions. Standard theory then applies. To see how this occurs, let $\mathfrak{A} = \mathfrak{A}^+ \oplus \mathfrak{A}^-$, and choose a basis so that $\mathfrak{A}^+$ is spanned by $\{1, v_M v_N, v_M v_N v_P v_Q, \ldots\}$ (elements of even rank) and $\mathfrak{A}^-$ is spanned by $\{v_M, v_M v_N v_P, \ldots\}$ (elements of odd rank). Then, expanding the fields (e, ψ) in terms of this basis, we have

$$e = \underset{(0)}{e} + \underset{(2)}{e}vv + \underset{(4)}{e}vvvv + \cdots \tag{31a}$$

and

$$\psi = \underset{(1)}{\psi}v + \underset{(3)}{\psi}vvv + \cdots; \tag{31b}$$

while expanding the field equations (26) in terms of the basis and reading off coefficient equations, we obtain the sequence

$$\text{rank 0:}\quad \Box_{(0)}\underset{(0)}{e} = O(\underset{(0)}{e}, \partial \underset{(0)}{e}) \quad (\text{where } \underset{(0)}{\Box} := \underset{(0)}{g}^{\alpha\beta}\partial_\alpha\partial_\beta), \tag{32a}$$

$$\text{rank 1:}\quad \underset{(0)}{\Box}\underset{(1)}{\psi} = \underset{(1)}{\psi}O(\underset{(0)}{e}, \partial\underset{(0)}{e}, \partial\partial\underset{(0)}{e}), \tag{32b}$$

$$\text{rank 2:}\quad \underset{(0)}{\Box}\underset{(2)}{e} = -\underset{(2)}{\Box}\underset{(0)}{e} + O(\underset{(0)}{e}, \partial\underset{(0)}{e}, \underset{(1)}{\psi}, \partial\underset{(1)}{\psi}) \tag{32c}$$

(note that $\underset{(2)}{\Box} = \underset{(2)}{g}^{\mu\nu}\partial_\mu\partial_\nu = \underset{(2)}{e}^{(\mu}\underset{(0)}{e}^{\nu)}\partial_\mu\partial_\nu$ is linear in $\underset{(2)}{e}$), etc. One can ascertain by induction that the above sequence of equations, which we shall label as $\{\underset{(n)}{F} = 0:\ n = 0, 1, 2, \ldots\}$, has the following properties.

(i) $\underset{(0)}{F} = 0$ is the standard (numerical valued) nonlinear vacuum Einstein equation with the harmonic and $O(3, 1)$ gauge conditions imposed. It is in strictly hyperbolic form, hence solvable[23,24] for the sole unknown $\underset{(0)}{e}$.

(ii) For $n > 0$, $\underset{(n)}{F} = 0$ takes the form $\underset{(0)}{\Box}\underset{(n)}{e} = \underset{(n)}{e}a + b$ for n even and $\underset{(0)}{\Box}\underset{(n)}{\psi} = \underset{(n)}{\psi}c + d$ for n odd. Here, a, b, c, d are functionals of $\underset{(0)}{e}, \underset{(1)}{\psi}, \underset{(2)}{e}, \underset{(3)}{\psi}, \ldots, \underset{(n-1)}{e}$ or $\underset{(n-1)}{\psi}$, and their derivatives. Thus these can be solved successively as linear equations for the unknowns $\underset{(1)}{\psi}, \underset{(2)}{e}, \underset{(3)}{\psi}, \underset{(4)}{e}, \ldots$, etc. Note that since the top-order operator of each $\underset{(n)}{F}$ is always $\underset{(0)}{\Box}$, local causality is immediate with the light cones being those of $\underset{(0)}{g}$. Compare also with Ref. 7.

V. CONCLUSION

We have shown that the field equations of $(N = 1)$ supergravity, when treated "classically" as a set of $\mathfrak{A}$-valued partial differential equations, are equivalent (under some gauge conditions) to a certain formally hyperbolic system. In the event that $\mathfrak{A}$ is an exterior algebra (or more generally, if $\mathfrak{A}$ should admit a grading by the non-negative integers), we have seen that this formally hyperbolic system decouples and can be solved in an iterative manner. It follows from the aforementioned equivalence that the Cauchy problem for the Grassmann formulation of supergravity is well-posed, and thus in principle (for example, generalizing ideas in Ref. 25) any properly constrained initial data generates a spacetime solution of the supergravity field equations.

Why should this matter to physicists? If supergravity were, like Maxwell's theory and Einstein's theory, a field theory with obvious classically observable manifestations, then the answer would be clear: One cannot expect signals to propagate causally in a universe governed by an ill-posed theory, and so the theory would be suspect.

But the physical meaning of classical supergravity is far from clear. Its fields seem to be inherently quantum field operators rather than classical observables. In the Grassmann formulation of supergravity, one of the symptoms of this problem of classical interpretation is the controversy over how one should think about the decoupling of the field equations which results from the presumed Grassmann algebraic structure of the fields [see (31) and (32)]. On the one hand, this decoupling seems to be somewhat unphysical—

the effects of fermions and higher-rank components of bosons become linearized, and the entire causal structure is determined by $\underset{(0)}{e}$, which is unaffected by the ψ source terms. On the other hand, the decoupling seems to be an unavoidable consequence of using the only currently known $\mathfrak{A}$ to mathematically formulate classical supergravity.

In spite of the lack of a clear physical interpretation for classical supergravity, we believe that it is important to know that the theory is well-posed. We note, for example, that many formulations of quantum field theory rely upon the space of classical solutions, and this space would most likely be quite strange if the field equations were ill-posed. In particular, the perturbations relied upon by the Feynman approach might be nonsensical.

Of course, it is experiment which must ultimately determine whether or not the theory of supergravity is a useful tool for describing the physics in our universe. But the fact that the theory has a well-posed Cauchy problem does, we think, increase the possibility that this is the case.

ACKNOWLEDGMENTS

D. Bao's research was partially supported by National Science Foundation (NSF) Grants No. MCS 81-07086 and No. 81-08814 (A02); that of J. Isenberg was partially supported by NSF Grant No. MCS 83-03998. J. Isenberg would also like to acknowledge the hospitality of the Department of Mathematics, Rice University, Houston, Texas 77251.

[1]D. Freedman, P. van Nieuwenhuizen, and S. Ferrara, Phys. Rev. D **13**, 3214 (1976).

[2]S. Deser and B. Zumino, Phys. Lett. B **62**, 335 (1976).

[3]M. Pilati, Nucl. Phys. B **132**, 138 (1978).

[4]E. Fradkin and M. Vasiliev, Phys. Lett. B. **72**, 70 (1977).

[5]S. Deser, J. Kay, and K. Stelle, Phys. Rev. D **16**, 2448 (1977).

[6]S. Hawking and G. Ellis, *The Large Scale Structure of Spacetime* (Cambridge U. P., Cambridge, 1973).

[7]Y. Choquet-Bruhat, Lett. Math. Phys. **7**, 459 (1983).

[8]R. Finkelstein and J. Kim, J. Math. Phys. **22**, 2228 (1981).

[9]J. Isenberg, D. Bao, and P. Yasskin, *Classical Supergravity* (to appear).

[10]D. Bao, Ann. Phys. **158**, 211 (1984).

[11]The algebra $\mathfrak{A}$ is a "coefficient" algebra. Some of its properties, however, coincide with that of a superalgebra in supermanifold theory, as described by D. Leites, Russ. Math. Surveys **35** (**1**), 1–64 (1980). The conceptual difference between coefficient algebras (in the space-time component formulation of supergravity as presented here, and in Wess-Zumino superspace) and superalgebras (in supermanifold theory) is explained in Ref. 10 and in L. Smolin, Ann. Phys. **131**, 398 (1981), and by references therein.

[12]J. Leray and Y. Ohya, Math. Annalen. **170**, 167 (1967).

[13]P. Yasskin, Ph. D. thesis, University of Maryland at College Park, 1979.

[14]P. van Nieuwenhuizen, Phys. Rep. **68:4** (1981).

[15]B. Zumino, Lecture Notes on Supersymmetry and Supergravity, University of California at Berkeley, 1981.

[16]J. Isenberg, D. Bao, and P. Yasskin, "Is Supergravity Well-Posed?," in *Proceedings of the Third Marcel Grossmann Meeting on General Relativity*, edited by H. Ning (North–Holland, Amsterdam, 1983).

[17]D. Bao, J. Isenberg, and P. Yasskin, "The Dynamics of the Einstein-Dirac System, Part II," in preparation.

[18]P. van Nieuwenhuizen (private communication).

[19]V. Fock, *The Theory of Spacetime and Gravitation* (Pergamon, New York, 1964), 2nd ed., pp. 422–430.

[20]N. Baaklini, S. Ferrara, and P. van Nieuwenhuizen, Lett. Nuovo Cimento **20**, 113 (1977).

[21]Y. Choquet-Bruhat, in *Gravitation, an Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), pp. 130–168.

[22]Y. Choquet-Bruhat, J. Math. Pures Appl. **45**, 371 (1966).

[23]J. Leray *Hyperbolic Differential Equations* (The Institute for Advanced Study, Princeton, NJ, 1952).

[24]Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics*, (North-Holland, Amsterdam, 1981), 2nd ed.

[25]J. Isenberg, Ann. Phys. **129**, 223 (1980).

On the Cauchy problem for the nonlinear Boltzmann equation global existence uniqueness and asymptotic stability

N. Bellomo
Dipartimento Matematico, Politecnico di Torino, Torino, Italy
G. Toscani
Dipartimento Matematico, Univarsità di Pavia, Pavia, Italy

(Received 25 May 1984; accepted for publication 17 August 1984)

The analysis of the initial value problem for the nonlinear Boltzmann equation is considered in this paper. A theorem defining global existence and uniqueness for initial data which decay at infinity with an inverse power law is the main result of this work and is obtained by suitable application of fixed point theorems in Banach spaces. The theorem also defines the asymptotic stability of the solutions.

I. INTRODUCTION

The analysis of the initial value problem for the full nonlinear Boltzmann equation, as carefully reviewed in Refs. 1 and 2, has been developed in the last 40 years with the final objective of supplying a proof for global existence under reasonable assumptions on the initial data. Such a proof does not involve a purely mathematical interest. In fact, it can also be regarded as an indirect validation of the mathematical model defined by the Boltzmann equation and hopefully can shed some light on the rigorous derivation of this equation.

In spite of the relevance of the subject and of several previous attempts, as documented in the large bibliography quoted in Refs. 1 and 2, a proof for global existence for the nonlinear unmodified Boltzmann equation has been only very recently given by Illner and Shinbrot[3] for initial conditions which decay exponentially to zero at infinity. Their paper shows how the solution, which exists locally,[4] is globally bounded and stays in the considered function space if the initial conditions are bounded in the aforementioned fashion. The proof, which refers to the equation in absence of an external field, has been supplied for hard spheres, but it can be extended to other classes of interaction potentials.

A similar assumption on the initial data has supplied global existence and uniqueness for a class of discrete velocity models.[5] There the proof was obtained by application of fixed point theorems.[6,7]

This paper considers the full nonlinear Boltzmann equation,[8] in the absence of a force field, and applies a method close to the one of Ref. 5 in order to prove global existence and uniqueness of the solution of the initial value problem. The proof, supplied in the third section, after the mathematical formulation of the problem proposed in the second section, holds for inverse power gas–particle interaction potential with cutoff characterized by "hard' interactions. More in detail, the result is proved for initial data which go to zero in terms of $1/|\mathbf{x}|^p, p>1$, where $\mathbf{x}$ is the space coordinate, and which are bounded by a velocity distribution which tends to zero at infinity. Therefore some analogy can be found about the assumption on the initial data with respect to the ones of Refs. 4 and 5. In fact, as in the quoted papers, the gas is assumed to be confined in a central region. However, the decay is here assumed to be very smooth. In fact some simple calculations realized in the last section show that the mean free path of the gas molecules can be sufficiently small.

II. PRELIMINARIES

The Boltzmann equation, which defines the time–space evolution of the one-particle distribution function f of a dilute monoatomic gas, can be written, in absence of an external field, in the following form:

$$f=f(t,\mathbf{x},\mathbf{V}),\quad \frac{\partial f}{\partial t}+\mathbf{V}\cdot\nabla f=J(f,f), \tag{1}$$

where f is a non-negative real-valued function of the time $t\in T$, of the space $\mathbf{x}\in \mathbf{R}^3$, and of the velocity $\mathbf{v}\in\mathbf{R}^3$, and where J is the collisional operator defining a bilinear map from two copies of the same function space into another. The operator J, for interaction potentials with "cutoff," can be split into two terms

$$J(t,\mathbf{x},\mathbf{V})=J_1(t,\mathbf{x},\mathbf{V})-f(t,\mathbf{x},\mathbf{V})J_2(t,\mathbf{x},\mathbf{V}), \tag{2}$$

namely into the "gain" and "loss" operators which can be defined as follows:

$$J_1=\int_D B(\theta,\mathbf{q})f(t,\mathbf{x},\mathbf{V}')f(t,\mathbf{x},\mathbf{V}_1')d\epsilon\, d\theta\, d\mathbf{V}_1, \tag{3}$$

$$J_2=\int_D B(\theta,\mathbf{q})f(t,\mathbf{x},\mathbf{V}_1)d\epsilon\, d\theta\, d\mathbf{V}_1, \tag{4}$$

where $(\mathbf{V},\mathbf{V}_1)$ are the precollisional velocities, $(\mathbf{V}',\mathbf{V}_1')$ the postcollisional velocities, $\mathbf{q}=(\mathbf{V}_1-\mathbf{V})$, and the angles ϵ and θ are, respectively, the polar and azimuthal angles of $\mathbf{V}'$ in a spherical coordinate system attached to $\mathbf{V}$ with the z axis oriented in the direction of $\mathbf{q}$. Consequently, $D=[0,2\pi]\cdot[0,\frac{1}{2}\pi]\cdot\mathbf{R}^3$. The precollisional and postcollisional velocities are related by the conservation equations for the momentum and energy. The structure of B depends upon the physical assumptions on the interaction potentials. The term B can be, for inverse power interaction potentials, written as follows:

$$B=B(\theta,q;s)=\beta_s(\theta)q^{(s-4)/s} \tag{5}$$

(where $s=4$ means Maxwellian molecules, where $s<4$ means soft interaction, and where $s>4$ means hard interaction). As a limiting case one gets the hard sphere model: $B=a^2q\sin\theta\cos\theta$, where a is the radius of the hard sphere, $s\to\infty$, and $\beta_s=a^2\sin\theta\cos\theta$.

In order to deal with the Cauchy problem related to Eq. (1), some function space has to be introduced with a defini-

 0022-2488/85/020334-05$02.50

tion of solution. Keeping this in mind, Eq. (1), with given initial conditions f_0, is now rewritten in integral form

$$f(t,\mathbf{x},\mathbf{V}) = f_0(\mathbf{x} - \mathbf{V}t,\mathbf{V}) + \int_0^t J_1(s,\mathbf{x} - \mathbf{V}(t-s),\mathbf{V})ds - \int_0^t f(s,\mathbf{x} - \mathbf{V}(t-s),\mathbf{V})J_2(s,\mathbf{x} - \mathbf{V}(t-s),\mathbf{V})ds. \tag{6}$$

Let now $\widetilde{C}_b(\mathbf{R}^3 \oplus \mathbf{R}^3)$ be the space of all continuous functions on $\mathbf{R}^3 \oplus \mathbf{R}^3$ which go to zero at infinity and let $\mathbf{B}_T$ be the space

$$\mathbf{B}_T = C^0([0,t];\widetilde{C}_b(\mathbf{R}^3 \oplus \mathbf{R}^3)) \tag{7}$$

of all continuous functions mapping $(t,\mathbf{x},\mathbf{V})$ into $\mathbf{R}$, which go to zero as $\mathbf{x}$ and $\mathbf{V}$ tend to infinity, equipped with the norms

$$\|f\|_\infty = \sup_{(\mathbf{x},\mathbf{V})\in\mathbf{R}^3\oplus\mathbf{R}^3} |f|, \quad \|f\| = \sup_{t\in T} \|f\|_\infty. \tag{8}$$

The following definition can now be supplied.

Definition: A function $f = f(t,\mathbf{x},\mathbf{V})$ is defined as a "mild solution" of Eq. (1) if $f \geqslant 0$ and Eq. (6) is satisfied, with $f \in \mathbf{B}_T$.

III. THE INITIAL VALUE PROBLEM

Consider now the expression of B and assume the following.

Hypothesis: The function $\beta(\theta)$ is regular in its arguments and the integral

$$\int_0^{(1/2)\pi} \left(\frac{\beta(\theta)}{\sin\theta\cos\theta}\right)d\theta = F < \infty \tag{9}$$

is bounded. Moreover, in Eq. (5), $s > 4$.

This hypothesis certainly holds for several cutoff potentials as also discussed in Ref. 2. In particular, for the hard spheres model, $F = a^2/2$.

Before entering into the main problem of this section, the following lemma needs to be proved.

Lemma 1: Let $\mathbf{u},\mathbf{v}\in\mathbb{R}^3$ be two orthogonal vectors $\mathbf{u}\cdot\mathbf{v} = 0$, then the integral

$$I_p = I_p(t,\mathbf{x};\mathbf{u},\mathbf{v},p) = \int_0^t g(s,\mathbf{x};p)ds, \tag{10}$$

where

$$g = \frac{1}{(1 + |\mathbf{x} + \mathbf{u}s|^2)^{p/2}} \frac{1}{(1 + |\mathbf{x} + \mathbf{v}s|^2)^{p/2}} \tag{11}$$

is bounded as follows:

$$I_p \leqslant \frac{1}{(1+x^2)^{p/2}} \frac{4}{(p-1)} \left\{\frac{1}{(u^2+v^2)^{1/2}} + \frac{1}{\inf(u,v)}\right\}. \tag{12}$$

Proof: Consider the integral defined in Eqs. (10) and (11) and that the following inequalities hold:

$$s \geqslant -2\mathbf{x}\cdot\mathbf{u}/u^2 \Rightarrow |\mathbf{x} + \mathbf{u}s| \geqslant |\mathbf{x}|;$$
$$s \geqslant -2\mathbf{x}\cdot\mathbf{v}/v^2 \Rightarrow |\mathbf{x} + \mathbf{v}s| \geqslant |\mathbf{x}|. \tag{13}$$

Therefore if $H = H(\mathbf{x};\mathbf{u},\mathbf{v}) = \inf\{(-2\mathbf{x}\cdot\mathbf{u}/u^2),(-2\mathbf{x}\cdot\mathbf{v}/v^2)\}$ then I_p can be decomposed into $I_p = I_{p1} + I_{p2}$, where

$$I_{p1} = \int_0^H g(s,\mathbf{x};p)ds, \quad \text{with } I_{p1} = 0 \text{ when } H \leqslant 0, \tag{14}$$

and

$$I_{p2} = \int_H^t g(s,\mathbf{x};p)ds. \tag{15}$$

Assume now $0 < H < t$ and note that, according also to the inequality (13), the following further inequalities hold:

$$s \leqslant H: (1 + |\mathbf{x} + \mathbf{u}s|^2)(1 + |\mathbf{x} + \mathbf{v}s|^2) \geqslant (1 + x^2)(1 + x^2 + (u^2+v^2)s^2 + 2\mathbf{x}\cdot(\mathbf{u}+\mathbf{v})s), \tag{16}$$

where $|\mathbf{u}+\mathbf{v}| = (u^2+v^2)^{1/2}$ and

$$(x^2 + (u^2+v^2)s^2 + 2\mathbf{x}\cdot(\mathbf{u}+\mathbf{v})s) \geqslant [(s(u^2+v^2)^{1/2} + \mathbf{x}\cdot(\mathbf{u}+\mathbf{v}))/(u^2+v^2)^{1/2}]^2. \tag{17}$$

Consequently

$$I_{p1} \leqslant \int_0^H \frac{1}{(1+x^2)^{p/2}} \frac{ds}{(1 + (s(u^2+v^2)^{1/2} + \mathbf{x}\cdot(\mathbf{u}+\mathbf{v})/(u^2+v^2)^{1/2})^2)^{p/2}} \leqslant \frac{1}{(1+x^2)^{p/2}} \int_0^H \frac{ds}{(1 + (u^2+v^2)s^2 + 2s\mathbf{x}\cdot(\mathbf{u}+\mathbf{v}) + (\mathbf{x}\cdot(\mathbf{u}+\mathbf{v}))^2/(u^2+v^2))}. \tag{18}$$

Then,

$$I_{p1} \leqslant \frac{1}{(1+x^2)^{p/2}} \int_{-\infty}^{\infty} \frac{1}{(1+(u^2+v^2)s^2)^{p/2}} ds \leqslant (4/((p-1)(u^2+v^2)^{1/2}))(1/(1+x^2)^{p/2}). \tag{19}$$

Consider now the second integral

$$H = -2\mathbf{x}\cdot\mathbf{u}/u^2, \quad s \geqslant H \Rightarrow |\mathbf{x} + \mathbf{u}s| \geqslant |\mathbf{x}|,$$

consequently

$$I_{p2} \leqslant \int_H^t (1+x^2)^{-p/2}(1 + |\mathbf{x} + \mathbf{v}s|^2)^{-p/2} ds \leqslant (1+x^2)^{-p/2} \int_{-\infty}^{\infty} (1 + (vs - |\mathbf{x}|)^2)^{-p/2} ds \leqslant (1+x^2)^{-p/2}(4/\mathbf{v}(p-1)). \tag{20}$$

Analogously,

$$H = -2\mathbf{x}\cdot\mathbf{v}/v^2, \quad s \geqslant H \Rightarrow I_{p2} \leqslant (1+x^2)^{-p/2}(4/u(p-1)).$$

Finally considering the sum of both integrals,

$$I_p \leqslant \frac{1}{(1+x^2)^{p/2}} \frac{4}{(p-1)} \left(\frac{1}{(u^2+v^2)^{1/2}} + \frac{1}{\inf\{u,v\}}\right). \tag{21}$$

Inequality (21) holds both for $H \leqslant 0$, as in the case $I_p = I_{p2}$, and for $H > T$, as in the case $I_p \leqslant I_{p1}$. Lemma 1 is then proved.

Lemma 2: Let

$$B_{pr}(t,\mathbf{x},\mathbf{V}) = (1 + |\mathbf{x} - \mathbf{V}t|^2)^{-p/2}\exp(-r^2V^2),$$

then if the hypothesis holds, the integral

$$H_{pr} = \int_0^t J_1(B_{pr},B_{pr})(s,\mathbf{x} - \mathbf{V}(t-s),\mathbf{V})ds \tag{22}$$

is bounded as follows:

$$H_{pr} < \frac{16(F/r^2)\pi^3(1/2r^2)^{(s-4)/4s}}{(p-1)(1+|\mathbf{x} - \mathbf{V}t|^2)^{p/2}} \exp(-r^2V^2). \tag{23}$$

Proof: The proof of Lemma 2 is founded on the observation that the vectors $(\mathbf{V} - \mathbf{V}')$ and $(\mathbf{V} - \mathbf{V}_1')$ are orthogonal, i.e.,

$$(\mathbf{V} - \mathbf{V}')\cdot(\mathbf{V} - \mathbf{V}_1') = 0, \tag{24}$$

and that their vector sum is

$$(\mathbf{V} - \mathbf{V}') + (\mathbf{V} - \mathbf{V}_1') = \mathbf{q} = (\mathbf{V} - \mathbf{V}_1). \tag{25}$$

Now replacing the expression B_{pr} into (22) yields

$$\begin{aligned} H_{pr} = \int_0^t \int_D & B(\theta,\mathbf{q})(1 + |\mathbf{x} - \mathbf{V}(t-s) - \mathbf{V}'s|^2)^{-p/2} \\ & \times(1 + |\mathbf{x} - \mathbf{V}(t-s) - \mathbf{V}_1's|^2)^{-p/2} \\ & \times\exp(-r^2(V'^2 + V_1'^2))d\epsilon\, d\theta\, d\mathbf{V}_1\, ds. \end{aligned} \tag{26}$$

Setting $\mathbf{y} = \mathbf{x} - \mathbf{V}t$ and considering that conservation of energy in the collision process implies $(V'^2 + V_1'^2) = (V^2 + V_1^2)$, then Eq. (26) (recalling also the statements at the beginning of the proof as well as the result of Lemma 1) can be rewritten, after integration over s and ϵ, as follows:

$$\begin{aligned} H_{pr} = & \frac{8}{(p-1)}(1+y^2)^{-p/2}\exp(-r^2V^2) \\ & \times\int_{[0,(1/2)\pi]\cdot\mathbf{R}^3} B(\theta,\mathbf{q}) \\ & \times\left(\frac{1}{\inf\{|\mathbf{V} - \mathbf{V}_1'|\}} + \frac{1}{|\mathbf{V} - \mathbf{V}_1|}\right) \\ & \times\exp(-r^2V_1^2)d\theta\, d\mathbf{V}_1. \end{aligned} \tag{27}$$

Recalling now that $|\mathbf{V} - \mathbf{V}'| = q\cos\theta$ and $|\mathbf{V} - \mathbf{V}_1'| = q\sin\theta$, we have

$$(1/\inf\{|\mathbf{V} - \mathbf{V}_1'|,|\mathbf{V} - \mathbf{V}'|\} + 1/|\mathbf{V} - \mathbf{V}_1|) \leqslant 2/q\sin\theta\cos\theta.$$

Then replacing the actual expression of $B(\theta,\mathbf{q})$ into Eq. (27) the following result is obtained:

$$\begin{aligned} H_{pr} < & [8/(p-1)](1+y^2)^{-p/2}\exp(-r^2V^2) \\ & \times 2F\int_{\mathbf{R}^3}\exp(-r^2V_1^2)q^{-4/s}\, d\mathbf{V}_1, \end{aligned} \tag{28}$$

where standard calculations give

$$\begin{aligned} & \int_{\mathbf{R}^3}\exp(-r^2V_1^2)|\mathbf{V} - \mathbf{V}_1|^{-4/s}\, dV_1 \\ & \quad \leqslant (\pi^2/r^2)(1/2r^2)^{(s-4)/4s}. \end{aligned} \tag{29}$$

Consequently,

$$H_{pr} < \frac{16\pi^3(F/r^2)}{(p-1)}\left(\frac{1}{2r^2}\right)^{(s-4)/4s}(1+y^2)^{-p/2}\exp(-r^2V^2). \tag{30}$$

Now resetting $\mathbf{x} = \mathbf{y} + \mathbf{V}t$ proves the lemma.

Remark 1: Details on the inequalities (16), (17) and (29) are supplied in the Appendix.

The initial value problem can now be considered and the following result for global existence and uniqueness can be proposed.

Theorem: If the hypothesis holds and the initial conditions are such that

$$\begin{aligned} & f_0(\mathbf{x},\mathbf{V})\in\widetilde{C}_b(\mathbf{R}^3\oplus\mathbf{R}^3), \\ & 0 \leqslant f_0 \leqslant \alpha_{pr}\exp(-r^2V^2)[1/(1+x^2)^{p/2}], \end{aligned} \tag{31}$$

where $p > 1$ and

$$\alpha_{pr} = [(p-1)/64\,\pi^3\,(F/r^2)](2r^2)^{(s-4)/4s}, \tag{32}$$

then the Cauchy problem has a unique global "mild solution"

$$\forall T > 0: f(t,\mathbf{x},\mathbf{V})\in\mathbf{B}_T.$$

Moreover

$$f(t,\mathbf{x},\mathbf{V}) \leqslant 2\alpha_{pr}\exp(-r^2V^2)/(1 + |\mathbf{x} - \mathbf{V}t|^2)^{p/2}. \tag{33}$$

Proof: The local uniqueness and positivity of solutions has been proved by Glickson[6] in the presence of a force field and by Kaniel and Shinbrot[4] in a bounded domain. Also, the method proposed in Ref. 5 for the discrete Boltzmann equation works for the full equation. Then, one has to prove that the solution, which exists uniquely and positively locally, is globally bounded and stays in the considered function space. Keeping this in mind, consider the following space:

$$\mathscr{B}_T^r = \{f(t,\mathbf{x},\mathbf{V}) = f^*(t,\mathbf{x},\mathbf{V})\exp(-r^2V^2): f^*\in\mathbf{B}_T\}, \tag{34}$$

where the norm is defined as

$$\|f\|^r = \sup_{t\in T}\|f^*(t,\mathbf{x},\mathbf{V})\|_\infty. \tag{35}$$

Obviously, $\mathscr{B}_T$ with the norm (35) is a complete Banach space.[7] Moreover, one can note that if there exists a time interval T such that the mild solution exists uniquely and positively in the said interval, then the solution is bounded by the solution of the truncated equation

$$f = Uf{:}f(t,\mathbf{x},\mathbf{V}) = f_0(\mathbf{x} - \mathbf{V}t,\mathbf{V}) + \int_0^t J_1(s,\mathbf{x} - \mathbf{V}(t-s),\mathbf{V})ds. \tag{36}$$

It will be proved that if the initial conditions are as in the theorem, then U defines a contractive mapping from a close convex subset of $\mathscr{B}_T^r$ into itself and that, as a consequence, U has a unique fixed point in $\mathscr{B}_T^r$ for every $t > 0$. The proof is in two steps: (a) $f_0\in\mathscr{B}_T^r \Rightarrow Uf\in\mathscr{B}_T^r$, $U{:}\mathscr{B}_T^r\to\mathscr{B}_T^r$; and (b) U is a contractive operator on

$$\mathscr{A}_T^r = \{f\in\mathscr{B}_T^r: 0 \leqslant f \leqslant 2\alpha_{pr}\exp(-r^2V^2)/(1 + |\mathbf{x} - \mathbf{V}t|^2)^{p/2}\}.$$

The first step is easily proved. In fact according to the statements of the theorem $f_0\in\mathbf{B}_T$. Moreover

$$\|f_0(\mathbf{x} - \mathbf{V}t,\mathbf{V}\|^r \leqslant \alpha_{pr} < \infty \Rightarrow f_0(\mathbf{x} - \mathbf{V}t,\mathbf{V})\in\mathscr{B}_T^r.$$

Consider now the second term of the operator U with $f\in\mathscr{B}_T^r$:

$$\begin{aligned} J_1 = \int_D & B(\theta,\mathbf{q})f^*(t,\mathbf{x},\mathbf{V}')f^*(t,\mathbf{x},\mathbf{V}_1') \\ & \times\exp(-r^2(V'^2 + V_1'^2))d\epsilon\, d\theta\, d\mathbf{V}_1. \end{aligned} \tag{37}$$

This implies, after simple calculations and considering both that B is continuous and bounded in all its arguments and the conservation of energy in the collision already recalled in the proof of Lemma 2, the following:

$$Uf \in \mathscr{B}^r_T, \quad U:\mathscr{B}^r_T \to \mathscr{B}^r_T. \tag{38}$$

Consider now the second step of the proof. The result of the previously proved Lemma 2 implies

$$\int_0^t J_1(B_{pr},B_{pr})(s,\mathbf{x} - \mathbf{V}(t-s),\mathbf{V})ds$$
$$< \frac{1}{4\alpha_{pr}} B_{pr}(t,\mathbf{x},\mathbf{V}). \tag{39}$$

Let $f \in \mathscr{A}^r_T$ then $Uf \in \mathscr{A}^r_T$ in fact easily verifies the following:

$$0 < Uf(t,\mathbf{x},\mathbf{V}) < 2\alpha_{pr} B_{pr}(t,\mathbf{x},\mathbf{V}). \tag{40}$$

Let now $f,g \in \mathscr{A}^r_T$, then

$$\begin{aligned} |Uf - Ug| &< \int_0^t \int_D B(\theta,\mathbf{q}) |f(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}')f(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}'_1) \\ &\quad - g(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}')g(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}'_1)| d\epsilon\, d\theta\, d\mathbf{V}_1\, ds \\ &< \int_0^t \int_D B(\theta,\mathbf{q}) \{ |f^*(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}' \\ &\quad - g^*(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}')| g^*(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}'_1) \\ &\quad + |f^*(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}'_1) - g^*(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}'_1)| \\ &\quad \times f^*(s,\mathbf{x} + \mathbf{V}s,\mathbf{V}'_1)\} \\ &\quad \times \exp(-r^2(V^2 + V_1^2)) d\epsilon\, d\theta\, d\mathbf{V}_1\, ds. \end{aligned} \tag{41}$$

Consequently, considering also the result of Lemma 2,

$$|Uf - Ug| \leq \tfrac{1}{2} \|f - g\|^r \exp(-r^2V^2). \tag{42}$$

Then, considering the proposed definition of norm

$$\|Uf - Ug\|^r \leq \tfrac{1}{2} \|f - g\|^r, \tag{43}$$

which states that U is a contractive operator from a closed convex subset of the Banach space $\mathscr{B}^r_T$, the theorem is proved.

Remark 2: The final result of the theorem supplies a solution procedure for the initial value problem in terms of iterations with convergence rate:

$$\|f_n - f\|^r < [c^n/(1-c)] \| f_1 - f_0 \|^r, \text{ with } c \leq \tfrac{1}{2}. \tag{44}$$

Moreover, the theorem supplies a result for asymptotic stability according to the following corollary.

Corollary: There exists a nonempty set of initial conditions f_0 defined as in the theorem such that

$$f_0 \in \Phi \Rightarrow \forall f_0 \in \Phi, \quad \|f_0\| \geq 0, \quad \lim_{t\to\infty} \|f\| \to 0.$$

Proof: The proof is only a direct consequence of the inequality (33) of the theorem.

IV. DISCUSSION

Global existence and uniqueness has been proved, in the preceding section, for initial data which satisfy the conditions of the theorem. The proof holds for a large class of inverse power gas–particle interaction potentials with cutoff, for initial data which decay to zero at infinity in space and velocity. The particular structure of the aforementioned decay and bounds is specified in the theorem.

In order to supply a quantitative estimate on the initial conditions, the particular case of the "hard spheres" interaction potential ($s \to \infty$) can be considered. According to Eqs. (5) and (9), $F = \pi a^2/2$. Then the bound for the initial conditions is detailed by

$$f_0(\mathbf{x},\mathbf{V}) < \frac{(p-1)r^2\sqrt{r}2^{1/4}}{32\,\pi^3\, a^2} \frac{\exp(-r^2V^2)}{(1+x^2)^{p/2}}. \tag{45}$$

After Eq. (45), a bound for the local number density is simply obtained as follows:

$$n_0(\mathbf{x}) = \int f_0(\mathbf{x},\mathbf{V}) d\mathbf{V} < \frac{((p-1)2^{1/4}\sqrt{r}/32\pi\sqrt{\pi}a^2)}{(1+x^2)^{p/2}}. \tag{46}$$

The inverse of the product na^2 can be assumed as a measure of the mean free path:

$$\lambda \propto (32\pi\sqrt{\pi}\,\sqrt{r}/(p-1))(1+x^2)^{p/2} \tag{47}$$

which shows that very small mean free paths can be obtained at $x = 0$. However, the smaller the mean free path is at the origin, the larger is the increase as $\mathbf{x}$ increases.

If now the time evolution of the number density is considered, the result of the theorem and some simple calculations supply the following:

$$n(\mathbf{x} = 0,t) < \int \alpha_{pr} \frac{\exp(-r^2V^2)}{(1+t^2V^2)^{p/2}} d\mathbf{V}, \tag{48}$$

namely,

$$n(0,t) = o(1/t^p) \quad \text{as } t \to \infty. \tag{49}$$

Therefore, the result of the theorem and Eqs. (45)–(49) indicate how this paper develops the basic idea, considered previously in Refs. 3 and 5, of considering the time evolution of a gas confined in a central region and described by the nonlinear Boltzmann equation. The analysis indicated in this work shows that it is possible to obtain global existence and asymptotic behavior of the solution for very general assumptions on the initial data as far as the decay at infinity is assured (even if it is very smooth as stated in the theorem). As a particular final result, Eq. (49) shows how the decay in time of the local number density can be realized in a very smooth fashion.

ACKNOWLEDGMENTS

The authors are obliged to R. Illner for a helpful discussion during the preparation of the manuscript.

This research has been partially supported by the Italian Minister for Education M.P.I.

APPENDIX: SOME PROOFS

1. Proof of inequality (16)

We have

$$\begin{aligned} s < H:\ &(1 + |\mathbf{x} + \mathbf{u}s|^2)(1 + |\mathbf{x} + \mathbf{v}s|^2) \\ &= (1 + x^2 + (u^2s^2 + 2s\mathbf{x}\cdot\mathbf{u})) \\ &\quad \times (1 + x^2 + (v^2s^2 + 2s\mathbf{x}\cdot\mathbf{v})) \\ &= (1 + x^2)((1 + x^2) + s^2(u^2 + v^2) + 2s\mathbf{x}\cdot(\mathbf{u} + \mathbf{v})) \\ &\quad + (u^2s^2 + 2s\mathbf{x}\cdot\mathbf{u})(v^2s^2 + 2s\mathbf{x}\cdot\mathbf{v}). \end{aligned}$$

Then, since for $s \leqslant H$: $2s\mathbf{x}\cdot\mathbf{u} \leqslant -s^2u^2$, $2s\mathbf{x}\cdot\mathbf{v} \leqslant -s^2v^2$, the inequality is proved.

2. Proof of inequality (17)

We have

$$s \leqslant H: (x^2 + (u^2 + v^2)s^2 + 2s\mathbf{x}\cdot(\mathbf{u} + \mathbf{v}))$$
$$= \left(x^2 - \left(\mathbf{x}\cdot \frac{(\mathbf{u} + \mathbf{v})}{(u^2 + v^2)}\right)^2 + \left(s(u^2 + v^2)^{1/2} + \mathbf{x}\cdot \frac{(\mathbf{u} + \mathbf{v})}{(u^2 + v^2)}\right)^2\right)$$
$$- (s(u^2 + v^2)^{1/2} + \mathbf{x}(\mathbf{u} + \mathbf{v})/(u^2 + v^2)^{1/2})^2.$$

The inequality is then proved.

3. Proof of equalities (24) and (25)

Let $\mathbf{n}$ be the unit vector, in the direction of the apse line bisecting $-\mathbf{V}$ and $\mathbf{V}'$, then following Ref. 8,

$$\mathbf{V}' = \mathbf{V} - \mathbf{n}(\mathbf{n}\cdot\mathbf{q}) \quad \text{and} \quad \mathbf{V}_1' = \mathbf{V}_1 + \mathbf{n}(\mathbf{n}\cdot\mathbf{q}).$$

Then,

$$(\mathbf{V} - \mathbf{V}')\cdot(\mathbf{V} - \mathbf{V}_1') = \mathbf{n}(\mathbf{n}\cdot\mathbf{q})\cdot\mathbf{q} - \mathbf{n}(\mathbf{n}\cdot\mathbf{q})\cdot\mathbf{n}(\mathbf{n}\cdot\mathbf{q})$$
$$= (\mathbf{n}\cdot\mathbf{q})(\mathbf{n}\cdot\mathbf{q}) - (\mathbf{n}\cdot\mathbf{q})^2 = 0.$$

Moreover,

$$(\mathbf{V} - \mathbf{V}') + (\mathbf{V} - \mathbf{V}_1')$$
$$= 2\mathbf{V} - (\mathbf{V}' + \mathbf{V}_1') = 2\mathbf{V} - (\mathbf{V} + \mathbf{V}_1) = q.$$

The equalities are then proved.

4. Proof of inequality (29)

Consider the integral

$$L = \int_{\mathbf{R}^3} \exp(-r^2V_1^2)\left(\frac{1}{|\mathbf{V} - \mathbf{V}_1|^{4/s}}\right) d\mathbf{V}_1.$$

Setting $e = 4/s$, choosing an orthogonal frame with the z axis directed as $(\mathbf{V} - \mathbf{V}_1)$, and performing the integration with respect to z give the following:

$$L \leqslant \left(\frac{\pi}{r}\right) \int_{\mathbf{R}^2} \exp(-r^2(x^2 + y^2)) \left(\frac{1}{(x^2 + y^2)^{e/2}}\right) dx\, dy.$$

In polar coordinates,

$$L \leqslant \left(\frac{2\pi\sqrt{\pi}}{r}\right) \int_o^\infty R^{(1-e)} \exp(-r^2R^2) dR.$$

Applying the Hölder inequality,[7] namely,

$$E(|xy|) \leqslant E(|x|^p)^{1/p} E(|y|^q)^{1/q}$$

with $(1/p + 1/q) = 1$ and with E denoting the mean expected value, we get

$$L \leqslant (2\pi\sqrt{\pi}/r)(\sqrt{\pi}/2r)(1/2r)^{(1/2)(1-e)}$$
$$= (\pi^2/r^2)(1/2r^2)^{(s-4)/4s},$$

which proves the inequality.

[1] W. Grenberg, J. Palewczak, and P. Zweifel, "Global Existence Proofs for the Boltzmann Equation," in *Nonlinear Equilibrium Phenomena: The Boltzmann Equation*, edited by J. L. Lebowitz and E. W. Montroll (North-Holland, Amsterdam, 1983), pp. 21–49.

[2] W. Fiszdon, M. Lachowicz, and A. Palchewski, "Existence Problems for the Nonlinear Boltzmann Equation," in *Proceedings of the International Conference on Mathematical Methods in Mechanics*, Paris, November 1983.

[3] R. Illner and M. Shinbrot, Commun. Math. Phys. **95**, 217 (1984).

[4] S. Kaniel and M. Shinbrot, Commun. Math. Phys. **58**, 65 (1978).

[5] G. Toscani, "On the Discrete Velocity Models of the Boltzmann Equation in Several Dimensions," Ann. Mat. Pura Appl. (to appear).

[6] A. Glikson, Arch. Rat. Mech. Anal. **47**, 289 (1972).

[7] L. V. Kantorovic and G. P. Akilov, *Funkzionalij Analiz* (Mir, Moscow, 1977).

[8] C. Cercignani, *Theory and Application of the Boltzmann Equation* (Scottish Academic, Edinburgh, (1975).

Representations of the current algebra of a charged massless Dirac field

R. Borek
Institut für Theoretische Physik, Universität Göttingen, Göttingen, West Germany

(Received 28 February 1984; accepted for publication 27 July 1984)

It is shown that the current algebra A_0 of a charged, massless Dirac particle has representations with positive energy of all types I_∞, II_∞, and III in the classification of Araki–Woods.

I. INTRODUCTION

The von Neumann algebra generated by a representation π of the C^* algebra A of quasilocal observables of a local quantum theory is known to be type I if π is covariant under the space-time translation group, the spectrum condition is satisfied and there is a translation invariant vector which is cyclic for the representation π.

In the absence of a vacuum state the von Neumann algebra $\pi(A)''$ is also type I if the spectrum condition is sharpened by requiring the existence of a massive particle isolated from the rest of the spectrum.[1]

In order to consider these questions further within the framework of the algebraic quantum field theory we shall state the Haag–Kastler–Araki axioms in the notation of Ref. 2.

(1) To every bounded open region $O\subset\mathbb{R}^d$, $d\geqslant 2$ one assigns a C^* algebra $A(O)$ such that (a) $O_1\subset O_2$ implies $A(O_1)\subset A(O_2)$; (b) if the regions O_1 and O_2 are spacelike separated then the elements of $A(O_1)$ commute with all elements of $A(O_2)$. The algebra of quasilocal observables will denote the C^* algebra generated by the union of $\{A(O)\}$.

(2) There exists a representation of the vector group $\mathbb{R}^d$ as automorphisms of A, $\alpha:\mathbb{R}^d\to\mathrm{Aut}\,A$, and furthermore, $\alpha_a A(O)=A(O+a)$ for every a in $\mathbb{R}^d$.

(3) A representation π of A on a Hilbert space H is called a representation satisfying the spectrum condition if the following holds: (a) there exists a strongly continuous unitary representation of the vector group $\mathbb{R}^d$ on the Hilbert space H; (b) the representation $\mathrm{U}(a)$ implements the automorphisms α_a, that is

$$\mathrm{U}(a)\pi(x)\mathrm{U}(a)^{-1}=\pi(\alpha_a(x))\quad\text{for every } x\in A;$$

(c) the spectrum of the representation $\mathrm{U}(a)$ is contained in the future light cone. Such representations are said to be representations with positive energy.

We shall consider C^*-dynamical systems which satisfy axioms I and II and possess at least one nontrivial faithful representation which satisfies the spectrum condition. Such a system will be called a theory of local observables and will be denoted by $(A(O),\mathbb{R}^d,\alpha)$.

A state ω on A is called a vacuum state if it is invariant under the automorphisms α_a and the cyclic representation (π_ω,H_ω) induced by ω is a representation satisfying the spectrum condition.

In the presence of massless particles it is experimentally impossible to distinguish a vacuum from an infrared cloud formed by massless particles of very low momenta. Therefore, there exist representations of massless Fermi fields with positive energy which do not possess a vacuum state. We shall show that there are representations of the algebra of a free Fermion field with positive energy which are representations of types II_∞ and III in the sense of von Neumann.

We shall consider the CAR algebra $A(K)$ over K, when K is the direct sum of the Hilbert spaces of the irreducible unitary representations of the covering group of the Poincaré group of zero mass, spin $\frac{1}{2}$, and helicity $\pm$. The creation and annihilation operators $a(f)^*$, $a(g)$, f, $g\in K$, fulfilling the CAR are related in the standard way to the negative and positive frequency parts of the free massless Fermi field Ψ. The local field algebras $F(O)$ are the C^* subalgebras of $A(K)$ generated by Ψ regularized with test functions with support in a given region O. The gauge transformation to the angle π defines an automorphism γ of $A(K)$ such that

$$\gamma(a(f))=-a(f),\quad \gamma(\Psi)=-\Psi.$$

The γ-fixed point subalgebras $A(O)$ of $F(O)$ provide a model of local quantum field theory fulfilling the above axioms where the quasilocal algebra $A=A(K)_e$ is simple, separable, and the action of the Poincaré group is strongly continuous.

Alternatively, for given translation covariant, locally conserved currents $j^\mu(x)$, $\mu=0,1,2,3$, we let K denote the Hilbert space of vector states of a charged massless Dirac field ϕ. We let $A(K)$ denote the CAR algebra over K. A strongly continuous unitary representation of the gauge group U(1) induces automorphisms β of the CAR algebra $A(K)$ such that

$$\beta_\alpha(a(f))=e^{i\alpha}a(f);\quad \beta_\alpha(\phi)=e^{i\alpha}\phi\quad\text{for } f \text{ in K}.$$

The current algebra of a massless Dirac field ϕ is then the β-fixed point subalgebra of the CAR algebra $A(K)$. The β-fixed point subalgebras $A(O)$ of the local field algebras $F(O)$ also provide a model for local quantum field theory which fulfills the above axioms.

We shall construct non-type I representations with spectrum condition of the current algebra of a charged massless Dirac field and quasilocal algebra of a free massless Majorana field in the following way.

We consider subspaces of the Hilbert space K generated by finite particle unit vectors which are created from the vacuum and are localized in a bounded region of the momentum space. If we require that the sum of the energies carried by these finite particle vector states remain finite then the weak limit of these vector states induce representations of the quasilocal algebra which are type I, II, or III and satisfy the spectrum condition. Since every vacuum representation of free massless Fermi fields is equivalent to the Fock repre-

 0022-2488/85/020339-06$02.50

sentation one can see that it is convenient to consider representations of the Fermion algebra.

II. THE SPECTRUM CONDITION AND DIRECT PRODUCTS OF FOCK REPRESENTATIONS

Let A be a C^* algebra and A^{**} the double dual space of A. Then A^{**} becomes a von Neumann algebra, in a natural manner, if it is endowed with the weak topology induced by the topology of A^*. If A is a C^* algebra we shall denote by $S(A)$ the set of states of A. Let V denote the future light cone. Then, $S_0(\hat{V})$ will denote the set of states with the following properties: (a) $\omega(x\alpha_a y)$ is continuous for every $x,y\in A^{**}$; (b) $\omega(x\alpha_a y)=f(a)$ is the boundary value of an analytic function $f(z)$ holomorphic in the tube $\mathbb{R}^4 + iV^0 = T^+$, where V^0 denotes the interior of V; (c) there exists a constant $m>0$ depending on ω such that $|f(z)|\leqslant\|x\|\,\|y\|\exp\{m|\operatorname{Im} z|\}$. $S(\hat{V})$ will denote the norm closure of $S_0(\hat{V})$.

The set $S(\hat{V})$ has the following properties.

(1) If (π,H) is a representation of A there exists a strongly continuous unitary representation $\mathrm{U}(a)$ of the translation group $\mathbb{R}^d$ which implements the automorphisms α_a, that is, $\mathrm{U}(a)\pi(x)\mathrm{U}(a)^{-1} = \pi(\alpha_a(x))$, $x\in A$, and the spectrum of $\mathrm{U}(a)$ is contained in $\hat{V}$ if and only if all normal states of π are in $S(\hat{V})$.

(2) $S(\hat{V})$ is a folium. This means that there exists a projection $E(\hat{V})\in Z(A^{**})$ such that $\omega\in S(A)$ is in $S(\hat{V})$ if and only if $\omega(E(\hat{V}))=1$, where $Z(A^{**})$ denotes the center of the von Neumann algebra A^{**}.

(3) $S(\hat{V})$ is invariant under α_a for every $a\in\mathbb{R}^d$. This implies that $E(\hat{V})$ is invariant under the automorphisms α_a.

Let $(A(O),\mathbb{R}^d,\alpha)$ be a theory of local observables. It follows from Ref. 3 that the automorphisms α_a are spatial in $A^{**}E(\hat{V})$. There exists a strongly continuous unitary representation $\mathrm{U}(a)$ of $\mathbb{R}^d$ which implements the automorphisms α_a and $\mathrm{U}(a)\in A^{**}E(\hat{V})$. Furthermore, the representation $\mathrm{U}(a)$ is minimal in the sense that if $V(a)$ is a strongly continuous unitary representation of $\mathbb{R}^d$ which has the same properties, then $\mathrm{U}(a)V(a)^{-1}\in Z(A^{**}E(\hat{V}))$.

We shall give a construction of representations of the quasilocal algebra with spectrum condition which are representations of types I_∞, II_∞, and III.

Let K be the Hilbert space formed as a direct sum of vector states of a free massless Majorana field, and let $A=A(K)_e$ be the algebra of quasilocal observables described in the Introduction. Let K_i denote the translation invariant subspaces of K defined by $K_i=\{f(x)\in K:\operatorname{supp}\tilde{f}\subset\{\epsilon_{i-1}\leqslant|p|\leqslant\epsilon_i\}\}$ and $\Sigma\epsilon_i<\infty$. Let $A(K_i)_e$ denote the even CAR algebra over K_i. Let $\pi_{i,F}$ denote the Fock representation of $A(K_i)_e$ on a Hilbert space $H_{i,F}$ with a cyclic vector Ψ_i. Then $\pi(A(K)_e)'' = \otimes(\pi_{i,F}(A(K_i)_e), H_{i,F},\Psi_i)$ is a uniformly hyperfinite factor in the classification of Araki and Woods.[4]

Let $\Psi_{0,i}$ denote the Fock vacuum for the representation $\pi_{i,F}$. Let H_i denote the generator of a representation of the time-translation group on the Hilbert space H_i, $i=1,2,\ldots$. Since the vectors Ψ_i are in the spectrum of the energy operator we can assume that they had been chosen such that $\|\mathrm{H}_i\Psi_i\|\leqslant\epsilon_i$ and given $\delta_i>0$, $\|\Psi_i-\Psi_{0,i}\|\geqslant\delta_i$ for all but finitely many i. We form the incomplete infinite tensor product space $H=\otimes(H_i,\Psi_i)$. The rotation group is compact. By specifying spin indices we obtain a representation $V_i(r)$ of the rotation group R on the Hilbert space H_i such that $V_i(r)\Psi_i=\Psi_i, i=1,2,\ldots$. Then the representation π is rotation covariant under the representation $V(r)=\otimes V_i(r)$ on the Hilbert space H. Since $\Sigma_i\|\Psi_i-\Psi_{0,i}\|=\infty$ the product vectors $\Psi=\otimes\Psi_i$ and $\Psi_0=\otimes\Psi_{0,i}$ are not in the same weak equivalence class and therefore H does not contain any vacuum. We define sequences of vectors $\Psi_i^{(n)}$ in the Hilbert spaces H_i as follows.

If $n=1$ we let $\Psi_i^{(1)}=\Psi_i$, $i=1,2,\ldots$.

If $n=2$ we let $\Psi_1^{(2)}=\Psi_1^{(1)}$, and for $i>1$ we choose the $\Psi_i^{(2)}$ such that $\epsilon_i<\|\mathrm{H}_i\Psi_i^{(2)}\|\leqslant\epsilon_{i+1}$.

If $n>2$ we let $\Psi_1^{(n+1)}=\Psi_1^{(n)},\ldots,\Psi_n^{(n+1)}=\Psi_n^{(n)}$, and for $i>n$ we choose the vector $\Psi_i^{(n+1)}$ such that

$$\epsilon_{i+n-2}<\|\mathrm{H}_i\Psi_i^{(n+1)}\|\leqslant\epsilon_{i+n-1}.$$

We let $\Psi^{(n)}=\otimes\Psi_i^{(n)}$. Then $\|\Psi_i^{(n)}-\Psi_i^{(n')}\|\geqslant\delta_i$, $\delta_i>0$ holds for some $\delta_i\not\to 0$ for all but finitely many i if $n\neq n'$. We note that the vectors $\{V_i(r)\Psi_i^{(n)}, n=1,2,\ldots; r\in R\}$ can be chosen to form a basis for the space H_i, $i=1,2,\ldots$. Therefore if Ψ_i is any vector in H_i with $0<\Pi_{i=1}^\infty\|\Psi_i\|<\infty$, then there exists a product vector $\Psi^{(n)}=\otimes\Psi_i^{(n)}$ in $H=\otimes(H_i,\Psi_i)$ with $\Sigma_i\|\mathrm{H}_i\Psi_i^{(n)}\|^2<\infty$. It is shown by Araki and Woods[4] that given any incomplete infinite tensor product space $H=\otimes(H_i,\Psi_i)$ a uniformly hyperfinite factor generated by the type I factors $M_i=B(K_i), i=1,2,\ldots$ is an infinite tensor product $\otimes M_i$. As Ψ_i vary over all basis vectors of H_i, $i=1,2,\ldots$, $\otimes B(K_i)$ vary over all uniformly hyperfinite factors of Araki–Woods.

We shall show that the representation (π,H) satisfies the spectrum condition. Let $(A(O),\mathbb{R}^d,\alpha)$ be a theory of local observables and A the quasilocal algebra of a free massless Majorana particle which we have described in the Introduction. We assume that the representation α_a of the translation group is strongly continuous.

Let ω be a normal state of π. Then ω is the weak* limit of the Fock states $\omega_{i,F}$, where each $\omega_{i,F}$ is the vector state of Ψ_i. Since each $\omega_{i,F}$ is an element of $S(\hat{V})$ and α_a is strongly continuous the conditions (a) and (b) remain valid under limits. It follows that ω satisfies (a) and (b). For x,y in A^{**} let $\omega(x\alpha_a y)=f(a)$ be the boundary value of an analytic function $f(z)$ holomorphic in the tube $\mathbb{R}^4+iV^0$. Since each $\omega_{i,F}$ is an element of $S(\hat{V})$ and the sequence $\|\mathrm{H}_i\Psi_i\|$ is uniformly bounded it follows from Vitali's theorem[5] that there exists a real number $s>0$ such that $|f(z)|\leqslant\|x\|\,\|y\|\exp\{s|\operatorname{Im} z|\}$. The state ω is therefore an element of $S(\hat{V})$. It follows that the representation (π,H) satisfies the spectrum condition.

III. REPRESENTATIONS OF THE CURRENT ALGEBRA

Let K be the Hilbert space of vector states of massless Fermi particles

$$K=\left\{\tilde{f}_\nu(|\mathbf{p}|,\mathbf{p}):\sum_\nu\int|\tilde{f}_\nu(|\mathbf{p}|,\mathbf{p})|^2\frac{d^3\mathbf{p}}{|\mathbf{p}|}<\infty\right\}.$$

It follows from the definition of Fermi fields that there exists a unitary representation $\mathrm{U}(a,\Lambda)$ of the Poincaré group on K. Let $(\pi_F,H_F,V(a,\Lambda))$ be the induced Fock representation, where $V(a,\Lambda)$ is a strongly continuous unitary representation of the Poincaré group and $V(a,1)$ a representation of the translation group which satisfies the spectrum condition.

A representation of the time translation group is then given by

$$V(t,1) = \int_0^\infty \exp(itp) dE(p)$$

and

$$E(\epsilon) = \int_0^\epsilon dE(p).$$

Lemma 3.1: Let $\{\Psi_i\}_{i=1}^\infty \in H_F$ such that (a) $|\,\|\Psi_i\| - 1| \leqslant \delta_i < 1$ and $\Sigma_i \delta_i < \infty$; and (b) for $i > 1$ there exist ϵ_i such that $E(\epsilon_i)\Psi_i = \Psi_i$ and $\Sigma_i \epsilon_i < \infty$. Then

$$\sum_i \|V(a,1)\Psi_i - \Psi_i\| < \infty$$

holds for each a in $\mathbb{R}^4$.

Proof: for each ϕ in H_F $(V(a,1)\Psi_i - \Psi_i, \phi)$ is the boundary value of an analytic function holomorphic in the tube $\mathbb{R}^4 + iV^0$. Therefore, it follows from Lemma 11.4.1 in Ref. 6 that

$$\|V(a,1)\Psi_i - \Psi_i\| \leqslant 2 \sin(\|a\|\epsilon_i/\sqrt{2})\|\Psi_i\|.$$

The lemma follows from this relation. We choose a sequence of real numbers $\{\delta_i\}$ with $\delta_{i+1} < \delta_i$. Let

$$K_0 = \{f_\nu \in K : \operatorname{supp} \tilde{f}_\nu(|\mathbf{p}|,\mathbf{p}) \subset [\delta_1, \infty)\},$$
$$K_i = \{f_\nu \in K : \operatorname{supp} \tilde{f}_\nu(|\mathbf{p}|,\mathbf{p}) \subset [\delta_{i+1}, \delta_i]\}.$$

The subspaces K_i are invariant under translations and rotations. Therefore, the Fock representations $\pi_{i,F}$ of $A(K_i)$ on $H_{i,F}$ are also covariant under translation and rotation groups. The translations in $H_{i,F}$ satisfy the spectrum condition and the spectrum is contained in the set $\{0\} \cup [\delta_{i+1}, \infty)$. We consider the infinite tensor product

$$\bigotimes_{i=1}^\infty \pi_{i,F}(A(K_i))$$

on

$$\bigotimes_{i=1}^\infty (H_{i,F}, \Psi_i),$$

where

$$\sum_i |\,\|\Psi_i\| - 1| < \infty.$$

We choose ϵ_i with

$$\sum_i \epsilon_i < \infty, \quad \delta_i < \epsilon_i$$

and

$$E(\epsilon_i)\Psi_i = \Psi_i.$$

It follows from Lemma 2.1 that the product vectors $\otimes \Psi_i$ and $\otimes V(a,1)\Psi_i$ are weakly equivalent. We define on the incomplete infinite tensor product space $\otimes (H_{i,F}, \Psi_i)$ a representation of the translation group by

$$W(a) \bigotimes_{i=1}^\infty \pi_{i,F}(x_i)\Psi_i = \bigotimes_{i=1}^\infty \pi_{i,F}(\alpha_a x_i) V(a,1)\Psi_i.$$

Lemma 3.2: $W(a)$ is a strongly continuous unitary representation of the translation group which satisfies the spectrum condition.

Proof: Since $V(a,1)$ is a strongly continuous unitary representation and

$$\left(\bigotimes_{i=1}^\infty \pi_{i,F}(\alpha_a x_i) V(a,1)\Psi_i, \phi \right)$$

is a Borel function for all vectors ϕ in

$$\bigotimes_{i=1}^\infty (H_{i,F}, \Psi_i)$$

it follows from Ref. 7 that $W(a)$ is a strongly continuous unitary representation.

Let H_i be the generator of the representation of the time translation group for the Fock representation $\pi_{i,F}$. Since the vectors $\otimes \Psi_i$ and $\otimes V(a,1)\Psi_i$ are in the same weak equivalence class it follows that

$$\sum_i |(\mathrm{H}_i \Psi_i, \Psi_i) - 1| < \infty, \quad \sum_i |(\mathrm{H}_i \Psi_i, \Psi_i)| < \infty.$$

It follows from a theorem by Kraus and Streater[8] that the representation $W(a)$ satisfies the spectrum condition.

Corollary 3.3: The CAR algebra $A(K)$ has covariant representations with positive energy of all types I_∞, II_∞, and III.

Proof: The corollary follows from Lemma 3.2 and the construction of these representations as infinite direct products of Fock representations. If (π, H) is a representation of $A(K)$ with positive energy then it is an infinite direct product of type I representations. Hence, any finite normal trace on $\pi(A(K))''$ is zero and therefore (π, H) cannot be a representation of type II_1.

Let $j^\mu(x), \mu = 0,1,2,3$ be given translation covariant, locally conserved currents and let K denote the Hilbert space of vector states of a charged, massless Dirac particle. Let $A(K)$ be the CAR algebra over K.

It follows from Noether's theorem that the currents $j^\mu(x), \mu = 0,1,2,3$ are induced by a strongly continuous unitary representation of the gauge group U(1). This representation defines a gauge transformation of the first kind on the elements of the CAR algebra $A(K)$. We define a C^* algebra of quasilocal observables associated to a region O in space-time as the C^* algebra which is generated by polynomials of the form $\{\phi(f)\phi(g)^* : \operatorname{supp} f \cup \operatorname{supp} g \subset O\rangle\}$. The quasilocal algebra is the C^* algebra A_0 which is generated by the elements of $A(K)$ which are invariant under the gauge group U(1). The quasilocal algebra A_0 is called the current algebra of a charged massless Dirac particle.

Corollary 3.4: The current algebra A_0 has covariant representations with positive energy of all types I_∞, II_∞, and III of Araki–Woods.

Proof: Let K_i denote the translation invariant subspaces defined above and $A(K_i)$ the CAR algebra over K_i. Let $A(K_i)_0$ denote the gauge invariant C^* subalgebra of $A(K_i)$. Then A_0 is isomorphic to the unique minimal C^* algebra tensor product of the C^* algebras $A(K_i)_0$. If π_i is the Fock representation of $A(K_i)_0$, then, as in the proof of Corollary 2.3, we construct representations with positive energy of all types I_∞, II_∞, and III of the current algebra A_0 on an infinite incomplete tensor product space.

ACKNOWLEDGMENT

The author wishes to thank Professor H. J. Borchers for several discussions.

APPENDIX: A CLASSIFICATION OF THE REPRESENTATIONS OF THE CAR ALGEBRA

The main references for this section are the works of von Neumann,[9] Glimm,[10] and Araki–Woods.[4]

Let $A(K)$ be the CAR algebra over a Hilbert space K. Since the CAR algebra is a uniformly hyperfinite algebra there exists an infinite subset J of positive integers and an increasing sequence $\{M_\nu:\nu\in J\}$ of finite type I factors such that $A(K)=\cup_{\nu\in J} M_\nu^-$, where the closure is in the norm topology. We assume that $M_\nu = B(H_\nu)$ for some finite-dimensional Hilbert space H_ν. By Ref. 11 we let M_ν to be a $(2^{n_\nu}\times 2^{n_\nu})$-matrix algebra and $N_\nu = N_{\nu-1}\otimes M_\nu$. We form the incomplete infinite tensor product space $H=\otimes_{\nu\in J}(H_\nu,\Psi_\nu)$ of the Hilbert spaces H_ν which contains the product vector

$$\Psi=\otimes\Psi_\nu,\quad \Psi_\nu\in H_\nu,\quad 0<\prod_{\nu=1}^{\infty}\|\Psi_\nu\|<\infty.$$

The product vectors $\Psi=\otimes\Psi_\nu$ and $\chi=\otimes\chi_\nu$ belong to the same infinite tensor product space if and only if Ψ is in the weak equivalence class of χ, that is,

$$\sum_\nu |1-(\chi_\nu,\Psi_\nu)|<\infty,\quad \sum_\nu |1-\|\chi_\nu\||<\infty.$$

This implies that $\Sigma_\nu\|\chi_\nu-\Psi_\nu\|<\infty$. One defines a canonical mapping from $B(H_\nu)$ to $B(H)$ by

$$pS=\underset{\mu\neq\nu}{\otimes}(I_\mu)\otimes S,$$

where $S\in B(H_\nu)$ and I_μ is the identity operator on H_μ. Given the incomplete infinite tensor product space

$$H=\underset{\nu\in J}{\otimes}(H_\nu,\Psi_\nu),$$

one defines the infinite tensor product of finite type I factors N_ν by

$$\otimes N_\nu=\{p(N_\nu):\nu\in J\}''.$$

The von Neumann algebra $\otimes N_\nu$ is a uniformly hyperfinite factor. If $\pi(\pi\neq 0)$ is a representation of the CAR algebra $A(K)$ then it is faithful and $\pi(A(K))$ is a uniformly hyperfinite algebra.

A classification of uniformly hyperfinite factors is given by Araki and Woods.[4] We shall show that a classification of quasiequivalence classes of representations of the CAR algebra follows from the classifications given in Ref. 4.

We assume that the CAR algebra is generated by the field operators $\phi(f)$, $\phi(g)^*$, $f, g\in K$.

Let $0\leqslant \mathrm{A}\leqslant I$ be a self-adjoint operator. A state ω_A of the CAR algebra $A(K)$ is said to be a quasi-free state if its n-point functions are of the form

$$\omega_\mathrm{A}(\phi(f_1)\cdots\phi(f_{2n+1}))=0,$$

$$\omega_\mathrm{A}(\phi(f_1)\cdots\phi(f_{2n}))$$
$$=(-1)^{n(n-1)/2}\sum\sigma(s)\prod_{j=1}^{n}(f_{s(j)},\mathrm{A}f_{s(j+n)})$$

with $s(1)<\cdots<s(n)$, $s(j)<s(j+n)$, $j=1,\dots,n$ and $\sigma(s)$ is the signature of s.

If A has a pure point spectrum it follows from p. 4 of Ref. 11 that the n-point functions of the state ω_A can be given by

$$\omega_\mathrm{A}(E(i_1,\dots,i_{n_\nu};j_1,\dots j_{n_\nu}))=\delta^{j_1\cdots j_{n_\nu}}_{i_1\cdots i_{n_\nu}}\lambda_{i_1\ldots}\lambda_{i_{n_\nu}},$$

where

$$\{E(i_1,\dots,i_{n_\nu};j_1,\dots,j_{n_\nu}):i_k,j_k=0,1;k=1,\dots,n_\nu\}$$

are matrix units which span a $(2^{n_\nu}\times 2^{n_\nu})$-matrix algebra and

$$\lambda_{i_s}=\lambda_s\quad\text{if } i_s=0,\quad \lambda_{i_s}=1-\lambda_s\quad\text{if } i_s=1,$$
$$0\leqslant\lambda_s\leqslant 1,$$
$$s=1,\dots,n_\nu.$$

These matrix units are products of (2×2)-matrix units $\{E_{i_r j_r}\}$ and we have

$$E(i_1,\dots,i_{n_\nu};j_1,\dots,j_{n_\nu})=E_{i_1j_1}\cdots E_{i_{n_\nu}j_{n_\nu}};$$
$$E_{i_rj_r}E_{m_rn_r}=E_{m_rn_r}E_{i_rj_r}.$$

By theorem 5.1 of Ref. 11 it follows that two quasi-free states ω_A and $\omega_{\mathrm{A}'}$ are quasiequivalent if and only if the operators $\mathrm{A}^{1/2}-\mathrm{A}'^{1/2}$ and $(I-\mathrm{A})^{1/2}-(I-\mathrm{A}')^{1/2}$ are of the Hilbert–Schmidt class. If A has a continuous spectrum it follows from von Neumann's spectral theorem that one can choose a self-adjoint operator $0\leqslant\mathrm{A}'\leqslant I$ with a pure point spectrum such that the operators $\mathrm{A}^{1/2}-\mathrm{A}'^{1/2}$ and $(I-\mathrm{A})^{1/2}-(I-\mathrm{A}')^{1/2}$ are of Hilbert–Schmidt class. If π_A is a representation of the CAR algebra canonically associated to the quasi-free state ω_A, then the representation π_A can be determined up to quasiequivalence by a state $\omega_{\mathrm{A}'}$, where A' is a self-adjoint operator with a pure point spectrum whose eigenvalues are dense in the spectrum of A.

We let ω_A be a quasi-free state of the CAR algebra $A(K)$ and ω_ν be the restrictions of ω_A to the matrix algebras M_ν which generate $A(K)$. We let π be a representation of the CAR algebra $A(K)$ formed as an infinite tensor product of representations $\pi_\nu(B)$ acting on the incomplete infinite tensor product space $H'=\otimes H'_\nu$ containing a product vector $\Psi'=\otimes\Psi'_\nu$.

We shall show that the von Neumann algebra $\pi(A(K))''$ is a uniformly hyperfinite factor of types I, II, or III given in the classification of Araki and Woods.[4]

The state ω_ν of the algebra M_ν is given by the relations $\omega_\nu(B)=(\Psi_\nu,\pi_\nu(B)\Psi_\nu)$, where π_ν is a representation of M_ν canonically associated to ω_ν. Hence, there exists a trace class operator T_ν such that $\omega_\nu(B)=\mathrm{tr}\,T_\nu\pi(B)$. We let

$$T_\nu=\sum_j\lambda_{\nu j}P_j$$

be a spectral decomposition of the operator T_ν where each P_j is a one-dimensional projection $\lambda_{\nu j}\geqslant 0$ and

$$\sum_j\lambda_{\nu j}=\|\Psi'_\nu\|^2=1.$$

The eigenvalues of the operator T_ν appearing in the list $\{\lambda_{\nu j}:j=1,\dots,n_\nu\}$ are said to be the eigenvalue list of the state ω_ν relative to the type I factor M_ν. In fact, if $0\leqslant\mathrm{A}\leqslant I$ is a self-adjoint operator with a pure point spectrum and ω_A is a quasi-free state of the CAR algebra $A(K)$, whose restrictions to the M_ν are denoted by ω_ν then the eigenvalue lists $\{\lambda_{\nu j}; j=1,\ldots,n_\nu; \nu=1,2,\ldots\}$ of the state ω_ν relative to the type factors M_ν are the pure point spectrum of the operator A.

It follows from Ref. 4 that the quasiequivalence class of π is determined by the eigenvalue list $\{\lambda_{\nu j}; j=1,\ldots,n_\nu; \nu=1,2,\ldots\}$ and does not depend on the factorization of the type I factors M_ν. The eigenvalue list $\{\lambda_{\nu j}; j=1,\ldots,n_\nu; \nu=1,2,\ldots\}$ exists for every quasi-free state $\omega_{\rm A}$ and conversely this eigenvalue list uniquely characterizes this state.

If $\mathrm{A}=0$ then ω_0 is the Fock state which induces a representation of type I.

We shall give a construction of factors of types II and III by means of von Neumann's construction.[9] For $n=1,2,\ldots$ we let X_n be the measure space $\{0,1\}$, B_n the set of subsets of $\{0,1\}$, and μ'_n the measure on X_n defined by $\mu'_n(\{0\})=\lambda_n, \mu'_n(\{1\})=\lambda'_n$, where $\lambda_n+\lambda'_n=1$ and $\{\lambda_n : n=1,2,\ldots\}$ is a remuneration of the real numbers appearing in the set $\{\lambda_{\nu j}; j=1,\ldots,n_\nu, \nu=1,2,\ldots\}$. We let

$$(X,B,\mu')=\left(\prod_{n=1}^{\infty} X_n, \prod_{n=1}^{\infty} B_n, \prod_{n=1}^{\infty} \mu_n\right).$$

(X,B,μ) will denote the measure space formed by the completion of μ'. If x is in X then x is identified with the sequence (x_n), where each $x_n=0$ or 1. If $y=(y_n)$ is in X we define $x+y$ to be the sequence (x_n+y_n) reduced mod 2. Then X is a group and $\Delta=\{(x_n): x_n\neq 0$ for at most finite number of $n\}$ is a countable subgroup of X generated by the elements $\gamma_k=((\gamma_k)_n)$, where $(\gamma_k)_n=\delta^n_k$. For γ in Δ we define a mapping of X onto itself by $x\gamma=x+\gamma$.

Lemma 1: The measure μ is quasi-invariant under the action of Δ on X.

A proof of this lemma can be found in Ref. 19 (p. 179). Let $d\mu_a/d\mu(x)$ denote the Radon–Nikodym derivative of the translated measure with respect to the original measure. We let H_0 be the Hilbert space of functions $F(\gamma,x)$ $(\gamma\in\Delta, x\in X)$ for which

$$\sum_{\gamma\in\Delta}\int_X |F(\gamma,x)|^2\, d\mu(x)<\infty$$

with inner product

$$(F,G)=\sum_{\gamma\in\Delta}\int_X F(\gamma,x)\,\overline{G(\gamma,x)}d\mu(x) \quad \text{for } F \text{ and } G \text{ in } H_0.$$

It follows from Ref. 9 that the ring of operators R generated by the operators U_a and L_ϕ [a in Δ, $\phi(x)$ any bounded measurable function on X] is a factor of type II or type III, where

$$(U_aF)(\gamma,x)=(d\mu_a/d\mu(x))^{1/2}F(\gamma+a,xa),$$

$$(L_\phi F)(\gamma,x)=\phi(x)F(\gamma,x), a,\gamma\in\Delta, x\in X.$$

The group Δ is said to be (1) free, if for $\gamma\in\Delta$, $\gamma\neq 0$ the set of points satisfying the condition $x=x\gamma (x\in X)$ is a set of μ measure zero; (2) ergodic, if $E\gamma\subseteq E$ for $E\in B$ and every $\gamma\in\Delta$ implies either $\mu(E)=0$ or $\mu(X\setminus E)=0$; and (3) nonmeasurable if there exists no σ–finite measure ν on X which is equivalent to μ and invariant under Δ.

If $\lambda_{\nu j}=\frac{1}{2}$ for all ν and j, then $d\mu_a/d\mu(x)\equiv 1$. The group Δ is then measurable and the Haar measure μ is equivalent to the Lebesgue measure on the closed interval [0,1], where the equivalence is given by the map $x\to\Sigma_k x_k 2^{-k}$ except at countably many points. In this case the factor R generated by the operators U_a, L_ϕ is of type II.

Lemma 2: Assume that $\lambda_n\neq\frac{1}{2}$ for infinitely many n. Then the group Δ is free, ergodic, and nonmeasurable.

A proof of Lemma 2 is given in Ref. 12 (p. 180).

By theorem IX Ref. 9, Lemma 2 implies that R is a factor of type III. Every element γ in Δ is of the form $\gamma=\gamma_1+\cdots+\gamma_k$ and it can be shown that the factor R is generated by the operators $U_a, L_{\chi(\gamma_1,\ldots,\gamma_k)}$ (a in Δ, γ_i in X_i; $k=1,2,\ldots$), where $\chi(\gamma_1,\ldots,\gamma_k)(x)$ is the characteristic function of the set $\{(x_n): x_i=\gamma_i, i=1,\ldots,k\}$.

Let C be the algebra of linear combinations of the functions $\chi(\gamma_1,\ldots,\gamma_k)(x)$. Then the strong closure L_C^- of L_C is a subalgebra of $L_{l^\infty(X)}$ which is closed under monotone limits and thus it contains L_χ, where $\chi(x)$ is the characteristic function of an arbitrary measurable set and so $L_C^-=L_{l^\infty(X)}$. We let

$$W(i_1,\ldots,i_{n_\nu};j_1,\ldots,j_{n_\nu})=\pi_\nu(E(i_1,\ldots,i_{n_\nu};j_1,\ldots,j_{n_\nu})E'_\Psi,$$

where E'_Ψ is the projection onto the vector Ψ':

$$\theta(L_{\chi(i_1,\ldots,i_{n_\nu})})=W(i_1,\ldots,i_{n_\nu};i_1,\ldots,i_{n_\nu}),$$

$$\theta(U_{\gamma_{n_\nu}})=\sum_{i_1,\ldots,i_{n_\nu-1}}\{W(i_1,\ldots,i_{n_\nu-1},0;i_1,\ldots,i_{n_\nu-1},1)$$

$$+W(i_1,\ldots,i_{n_\nu-1},1;i_1,\ldots,i_{n_\nu-1},0)\}.$$

By linearity θ extends uniquely to an isomorphism of the C^* algebra N generated by U_Δ and L_C onto $\pi_A(A(K))E'_\Psi$.

It follows from Ref. 10 (p. 587) that the map θ extends to an isomorphism $\bar\theta$ of $\pi_A(A(K))''$ into the factor R. It follows from Ref. 4 that (1) $\pi_A(A(K))''$ is type I if and only if

$$\sum_\nu |1-\lambda_{\nu i}|<\infty,\quad \lambda_{\nu 1}\geqslant\lambda_{\nu 2}\geqslant\cdots\geqslant 0;$$

(2) $\pi_A(A(K))''$ is type II_1 if and only if $n_\nu<\infty$ for all ν and

$$\sum_{\nu,i}|(n_\nu)^{-1/2}-(\lambda_{\nu i})^{1/2}|^2<\infty;$$

(3) $\pi_A(A(K))''$ is a factor of type II_∞ if and only if it is isomorphic to a factor of the form $M_1\otimes M_2$, where M_1 is a factor of type I_∞ and M_2 is a factor of type II_1; and (4) if $\lambda_{\nu 1}\geqslant\delta$ for some $\delta>0$ for all ν, then $\pi_A(A(K))''$ is type III if and only if

$$\sum_{\nu,i}\inf\left\{\left|\frac{\lambda_{\nu 1}}{\lambda_{\nu i}}-1\right|^2, C\right\}=\infty$$

for all positive C.

It can be seen that any uniformly hyperfinite factor in the classification of Araki and Woods[4] can be described by the construction which we have given above. A classification of the quasiequivalence classes of the representations of the CAR algebra therefore follows from the classification of factors given in Ref. 4.

[1]D. Buchholz and K. Fredenhagen, Commun. Math. Phys. **84**, 1 (1982).
[2]H. J. Borchers, "On the Existence of Vacuum States," University of Göttingen (unpublished) (1982).
[3]H. J. Borchers, "Translation Group and the Spectrum Condition," Uni-

versity of Göttingen preprint (1982).
[4]H. Araki and E. J. Woods, Publ. Res. Inst. Math. Sci. **3**, 5 (1968).
[5]E. C. Titchmarsh, *The Theory of Functions* (Oxford U. P., New York, 1936).
[6]R. P. Boas, *Entire Functions* (Academic, New York, 1956).
[7]V. S. Varadarajan, *Geometry of Quantum Theory. II* (Van Nostrand-Rheinhold, New York, 1970).
[8]K. Kraus and R. F. Streater, J. Phys. A: Math. Gen. **14**, 2467 (1981).
[9]J. von Neumann, Ann. Math. **41**, 94 (1940).
[10]J. Glimm, Ann. Math. **73**, 572 (1961).
[11]R. T. Powers and E. Stormer, Commun. Math. Phys. **16**, I (1970).
[12]S. Sakai, *C*-algebras and W*-algebras* (Springer-Verlag, Berlin, 1970).

Superfield actions for $N = 2$ degenerate central charges

J. Hassoun, A. Restuccia, and J. G. Taylor
Department of Mathematics, King's College, London, WC2R 2LS, England

(Received 3 April 1984; accepted for publication 22 June 1984)

We construct the superfield actions for degenerate (spin-reducing) multiplets of $N = 2$ extended supersymmetry as integrals over all superspace. This requires the integration over the two available central charges as well. We evaluate the detailed component contributions to these actions and show they are total derivatives with respect to central charge dimensions. The resulting spectrum of the theories are analyzed in four dimensions in terms of various boundary conditions in the higher dimensions and the nature of the integration domain.

I. INTRODUCTION

The hope of constructing a unified theory of the forces of nature in terms of maximally extended supergravity ($N = 8$ SGR) has proved difficult to justify in the absence of a superfield formulation of the theory. Without such a version the ultraviolet divergence cancellation known to occur to all orders for maximally extended supersymmetric Yang–Mills theory ($N = 4$ SYM)[1,2] cannot also be shown to arise for $N = 8$ SGR. The difficulty in constructing a superfield framework for $N = 8$ SGR is that we encounter a barrier to such efforts at $N = 3$. No N-SGR, for $N > 3$, has a suitable set of auxiliary fields for the theory to be able to be put into superfield form (Refs. 3 and 4a).

Two methods were used for $N = 4$ SYM to broach the similar $N = 3$ barrier for N-SYM (Refs. 5 and 3b), that of light-cone gauge techniques and the use of $N = 2$ superfields, respectively. However, though similar techniques are available for $N = 8$ SGR they do not seem to be as satisfactory. The use of light-cone gauge methods has allowed the first nontrivial order interaction for $N = 8$ SGR in light-cone superspace to be constructed but no clear indication of finiteness for the resulting theory has appeared [Ref. 4(b)]. Similarly, $N = 8$ SGR may be constructed in terms of $N = 4$ superfields[6] but the construction of a counter term at three loops seems possible.[7] Thus the only alternative that seems available is by means of degenerate central charges [Ref. 4(c)].

Central charges z^{ij} may be introduced into the N-extended supersymmetry algebra $\mathscr{S}_N$ as

$$[S^i_{\alpha+}, S^j_{\beta+}]_+ = -2\epsilon_{\alpha+\beta+} Z^{ij} \tag{1.1a}$$

and the complex conjugate (where we use the convention that the complex conjugate of products of fermions is performed without interchange of order of terms)

$$[S_{\alpha-i}, S_{\beta-j}]_+ = -2\epsilon_{\alpha-\beta-} Z^{ij*}. \tag{1.1b}$$

We may regain the usual anticommutator between $S^i_{\alpha+}$, $S_{\beta-j}$

$$[S^i_{\alpha+}, S_{\beta-j}]_+ = -2(\not p C)_{\alpha+\beta-}\, S^i_j \tag{1.2}$$

if $S^i_{\alpha+}$ is determined in terms of $S_{\alpha-i}$ and the Z^{ij} by the Dirac constraint [Ref. 4(d)].

$$S^i_{\alpha+} = (\not p^{-1})_{\alpha+}{}^{\beta-} Z^{ij} S_{\beta-j}, \tag{1.3}$$

with the further constraint

$$Z^{ij} Z^{*jk} = p^2 S^i_k. \tag{1.4}$$

The constraints (1.3) and (1.4) cause a degeneration of the algebra so that only half the number of Fermi generators are required for the algebra, though at the same time the usual symmetries of Lorentz covariance, etc. are explicitly preserved. This is in sharp distinction to the other two methods of penetrating the $N = 3$ barrier, where either explicit Lorentz covariance, in the light-cone analysis, or explicit N-extended supersymmetry, on use of $N/2$ superfields, are lost. However, it is necessary to take the degeneracy constraints (1.3) and (1.4) in producing a radically new framework in which to build the theory.

One of the important features arising from the introduction of the central charges is the possibility of constructing fully geometric superfield actions. This has already been discussed in detail for the $N = 2$ hypermultiplet described by the superfield $\Phi_i(x,z,\bar z,\theta)$, where $1 < i < 2$ is the internal SU(2) label.[8] In this case we have the representation

$$Z^{ij} = Z\epsilon^{ij}, \quad Z^{*ij} = Z^*\epsilon^{ij}, \quad Z = \frac{\partial}{\partial z}, \quad Z^* = \frac{\partial}{\partial \bar z}. \tag{1.5}$$

In terms of $z = x^5 + ix^6$ we may write the action as

$$I = \int_\Gamma d^6x\, d^8\theta\, \Phi_i^+ \Phi_i, \tag{1.6}$$

$$D_{\alpha+(i}\Phi_{j)} = D_{\alpha-(i}\,\Phi_{j)} = 0. \tag{1.7}$$

The presence of the two extra integration variables in (1.6) have allowed us integration with the full measure $d^8\theta$ of the Grassmann variables. Moreover, on dimensional grounds the action (1.6) appears unique. It was shown that an unconstrained action derived from (1.6) leads to the correct equations of motion provided that the region Γ of integration over the central charge variables is limited to a cone and suitable boundary conditions are imposed on Φ_i. The purpose of this paper is to analyze the constrained version of (1.6) more fully in terms of components. Our analysis will allow us to extend our earlier analysis of the equations of motion and so give a more complete account of the relation between the spectrum, the region Γ, and the boundary conditions.

We begin our analysis in the next section by describing the properties satisfied by the covariant derivative D_α and its powers so as to evaluate (1.6) suitably. The Dirac condition for D_α is then solved in terms of the dependence of Φ_i on $\theta_{\alpha+i}$, $\theta^j_{\alpha-}$, and the resulting component expression tentatively evaluated. A similar expression is obtained for other

 0022-2488/85/020345-07$02.50

irreps expressed in superfield form. In the following section a detailed evaluation of the action is then given, leading to the resultant total derivative; the evaluation is aided by product formulas for D_α's given in the Appendix. The remaining section then analyzes the relation between the resulting spectrum, the shape of Γ, and the constraints on Φ_i on the boundary of Γ.

II. THE BASIC EXPRESSIONS

It is known[9] that the constraint (1.7) implies the Dirac constraint for the covariant derivatives D_α ($\alpha = {}^{i}_{\alpha+}$)

$$D^{i}_{\alpha+} = +(\not{p}^{-1})^{\beta-}_{\alpha+} Z^{ij} D_{\beta-j}, \tag{2.1}$$

and the complex conjugate of (2.1). We note that (2.1) is required for all degenerate representations if (1.3) and (1.4) are valid, since otherwise the eigenvalues of the Casimir generalizing the Pauli–Lubanski vector would be infinite.[10]

We wish to evaluate the action (1.6), which we write in the more usual form

$$I = \int_\Gamma d^6x \cdot D^{(0)4}\overline{D}^{(0)4}(\Phi_i^+ \Phi_i), \tag{2.2}$$

where (2.2) is to be evaluated at $\theta_{\alpha+i} = \theta^{i}_{\alpha-} = 0$. We have introduced the notation in (2.2) that $D^{(0)}{}_\alpha$ is obtained from the covariant derivative D_α by removal of the central charge term

$$D_\alpha = D^{(0)}{}_\alpha + Z_\alpha, \tag{2.3}$$

where

$$Z_\alpha = Z^{ij}\theta_{\alpha+j} \tag{2.4}$$

and

$$D^4 = \epsilon^{\alpha\beta\gamma\delta} D_\alpha D_\beta D_\gamma D_\delta, \tag{2.5}$$

with $\epsilon^{\alpha\beta\gamma\delta}$ the SU(2)$\times$SL(2C) alternating symbol

$$\epsilon^{\alpha\beta\gamma\delta} = \epsilon^{\alpha\delta}\epsilon^{\beta\gamma}\epsilon_{ij}\epsilon_{kl} - \epsilon^{\alpha\beta}\epsilon^{\gamma\delta}\epsilon_{il}\epsilon_{jk}. \tag{2.6}$$

Since

$$[D_{\alpha_1}, Z_{\alpha_2}]_+ - [D_{\alpha_2}, Z_{\alpha_1}]_+ = 0 \tag{2.7}$$

we may write

$$D^{(0)4} = D^4 - 4ZD^3 + 6Z^2D^2 - 4Z^3D + Z^4. \tag{2.8}$$

Since we have the representations

$$D_\alpha = \frac{\partial}{\partial\theta^\alpha} - i(\not{\partial}\bar\theta)_\alpha + Z_\alpha,$$
$$\overline{D}_{\dot\alpha} = \frac{\partial}{\partial\bar\theta^{\dot\alpha}} - i(\not{\partial}\theta)_{\dot\alpha} + \overline{Z}_{\dot\alpha}, \tag{2.9}$$

then, when evaluated at $\theta = 0$,

$$D^{(0)4}\overline{D}^{(0)4}(\Phi_i^+\Phi_i) = D^4\overline{D}^4(\Phi_i^+\Phi_i). \tag{2.10}$$

Thus the value of I is to be obtained by obtaining the various terms in the right-hand side of (2.10) on letting the derivatives act separately on the two superfield factors; we then must use the Dirac constraint (2.1) or its conjugate.

In order to appreciate most rapidly which components should be present in (2.10) we will go back to (1.6), and solve the Dirac constraint explicitly in terms of the dependence of Φ_i on θ_α and $\theta_{\dot\alpha}$, with $\theta_\alpha = \epsilon^{ij}\theta_{\alpha+j}$, $\theta_{\dot\alpha} = \epsilon_{ij}\theta^{j}_{\alpha-}$. For that we use (2.9) in (2.1) to rewrite (2.1), using (1.4), as

$$\left(\frac{\partial}{\partial\theta^\alpha} - \not{p}^{-1}Z\frac{\partial}{\partial\theta^{\dot\alpha}}\right)\Phi_i = 0. \tag{2.11}$$

If we introduce the Grassmann-valued operator

$$\psi_{\alpha+i} = \theta_{\alpha+i} + \not{p}^{-1\alpha-}_{\alpha+} Z^*_{ij}\theta^{j}_{\alpha-}, \tag{2.12}$$

then (2.11) indicates that

$$\Phi_i = \Phi_i(\psi, x, z_{ij}). \tag{2.13}$$

We note that from the constraint (1.4),

$$\psi^{i}_{\alpha-} = \not{p}^{-1}{}_{\alpha-}{}^{\alpha+} Z^{ij}\psi_{\alpha+j}, \tag{2.14}$$

so that $\psi^{i}_{\alpha-}$ is a redundant variable when (1.4) and (2.11) are taken into account. We add that we interpret (2.13) as a power series expansion of the right-hand side in the Grassmann variable $\psi_{\alpha+i}$, so that

$$\Phi_i = A_i(x,z) + \psi_\alpha\chi^\alpha_i(x,z) + \psi^2{}_{\alpha\beta}B_i^{\alpha\beta}(x,z)$$
$$+ \psi^{3\alpha}\eta_{i\alpha}(x,z) + \psi^4 C_i(x,z), \tag{2.15}$$

where $\psi^2{}_{\alpha\beta} = \frac{1}{2}\psi_{[\alpha}\psi_{\beta]}$, $\psi^{3\alpha} = \epsilon^{\alpha\beta\gamma\delta}\psi_\beta\psi_\gamma\psi_\delta$, $\psi^4 = \epsilon^{\alpha\beta\gamma\delta} \times \psi_\alpha\psi_\beta\ \psi_\gamma\gamma_\delta$ and we have used ϵ^{ij}, ϵ_{ij} to raise and lower, so $\psi_\alpha = \epsilon^{ij}\psi_{\alpha+j}$. We note that the component content of (2.15) is as expected from $N = 1$ irreps with superspin $Y = 0$ and $\frac{1}{2}$. In particular the $Y = 0$ irrep has no vector component field (the components have been discussed in Ref. 9). We will discuss these components more fully in the next section.

We see immediately from (2.15) that there are at most 4 powers of θ_α and $\theta_{\dot\alpha}$ together in (2.15), and these multiply $C_i(x,z)$ and its various derivatives with respect to Z, Z^* and (possibly inverse) powers of $\not{p}$. The only term that can therefore arise in the action of (1.6) is quadratic in C_i alone, with various powers of Z, Z^*, and $\Box$ acting on it:

$$I = \int_\Gamma d^6x\{\Sigma a_{rstuv} Z^r(Z^*)^s C_i^+ \Box^t Z^u (Z^*)^v C_i\}. \tag{2.16}$$

The numerical coefficients a_{rstuv} have yet to be determined, but only involve values of r,s,u,v which do not allow factors ZZ^* to be replaced by $\Box$ by use of (1.4); by dimensional arguments

$$r + s + u + v + 2t = 0. \tag{2.17}$$

We will calculate the coefficients a_{rstuv} in the next section to obtain a remarkable simplification, but before doing so we will comment on the result (2.16).

We first note that if there are positive values of r, s, u, or v in the sum (2.16) then the associated factor $\Box^t$ will introduce apparent nonlocality. We will show in the next section that there is no nonlocality in (2.16) when expressed in terms of $A_i(x,z)$ and derivatives of it with respect to x and z. This is because A_i and C_i are related for each of the $Y = 0$ and $Y = \frac{1}{2}$ irreps.

The second, and more substantial, point is that (2.16), containing only a scalar term alone, appears to violate supersymmetry. However, that cannot be true, since the original expression (1.6) is invariant under SUSY, as was discussed in detail in this case earlier.[8] We will return to explain the peculiarity of the situation in more detail at the end of the next section.

We now have to obtain the explicit form of the coefficient a_{rstuv} in (2.16). We will do that by using the expression (2.10) and evaluating it in detail; we discuss that in the next section.

III. EVALUATION OF THE COEFFICIENTS

We now wish to construct constrained actions for all degenerate irreducible multiplets. As was shown earlier (Ref. 8) dimensional arguments yield the unique functional I, quadratic in the superfields, defined over the whole superspace including the two central charges available in $N=2$ SUSY theories,

$$I=\int_\Gamma d^4x\, d^2z\, d^4\theta\, d^4\bar\theta\, \bar\phi_{[i]}\, \phi_{[i]}, \tag{3.1}$$

where $\phi_{[i]}$ is a scalar superfield which satisfies the spin reducing condition, and $[i]$ indicates a set of internal indices.

As a consequence of the Dirac condition, the independent components of a spin-reducing multiplet $\phi_{[i]}$ are only $D^m\, \phi_{[i]}|_{\theta=\bar\theta=z=0}$ and $D^m Z\phi_{[i]}|_{\theta=\bar\theta=z=0}$, where D^m are the totally antisymmetric derivatives defined in Appendix A. Therefore to isolate an irreducible representation from $\phi_{[i]}$ it is only necessary to impose a set of constraints on these derivatives. Let us denote these constraints as

$$\Sigma D^m\phi_{[i]}=0. \tag{3.2}$$

We propose as the constrained action for a general degenerate irreducible representation contained in $\phi_{[i]}$, the action (3.1) constrained by (3.2). In the case of the $N=2$ hypermultiplet this corresponds to the action (1.6). For the $Y=\frac{1}{2}$ representation contained in the scalar superfield ϕ which describes the $N=2$ abelian SYM, the constrained action is

$$I=\int_\Gamma d^4x\, d^2z\, d^4\theta\, d^4\bar\theta\bar\phi\phi, \qquad D^{2ij}\phi=0. \tag{3.3}$$

This constraint arises directly from the representation theory of central charges multiplets[10] and additionally it can be shown[11] that it satisfies the gauge covariant constraints associated with $N=2$ SYM. We will first characterize in a more precise way the structure of the constraints (3.2).

Let $\phi_{[i]}$ be a degenerate irreducible multiplet of the $N=2$ SUSY algebra. Then it satisfies the necessary condition

$$\not{D}^4\phi_{[i]}=\pm 4!Z^2\phi_{[i]}, \tag{3.4}$$

where the $+$ sign applies to the $Y=0$ representations while the $-$ sign applies to the $Y=\frac{1}{2}$ representations contained in $\phi_{[i]}$. To prove this relation we notice that $\phi_{[i]}$ being a degenerate irreducible multiplet is defined completely by constraints of the form (3.2). If we now apply $\not{D}^4$ to this constraint and we use (A8)

$$\not{D}^4D^n=(-1)^nD^n\not{D}^4$$

we obtain that $\not{D}^4\phi_{[i]}$ also satisfies the constraints (3.2). Hence, $\not{D}^4\phi_{[i]}$ is also a degenerate irreducible multiplet and it satisfies exactly the same constraints as $\phi_{[i]}$. The zero-order component of each superfield, $\phi_{[i]}|_{\theta=\bar\theta=0}$ and $\not{D}^4\phi_{[i]}|_{\theta=\bar\theta=0}$ must be proportional, because otherwise the hypothesis of irreducibility of $\phi_{[i]}$ would be contradicted. Hence, both superfields are proportional

$$\not{D}^4\phi_{[i]}=\chi\cdot\phi_{[i]}. \tag{3.5a}$$

Moreover from (A7) we know

$$\not{D}^4\not{D}^4=4!4!Z^4,$$

therefore,

$$\chi=\pm 4!Z^2. \tag{3.5b}$$

We are now able to analyze the sign in relation (3.5). From (A7)

$$Z^2\not{D}^{2\alpha\beta}=-\tfrac{1}{12}D^{2\alpha\beta}\not{D}^4,$$

and the decomposition of $D^2_{\alpha\beta}$ into its irreducible parts

$$D^2_{\alpha\beta}=\epsilon_{\alpha\beta}D^{2ij}+\epsilon^{ij}D^2_{\alpha\beta},$$

we finally obtain

$$4!Z^2D^2_{\alpha\beta}=-D^2_{\alpha\beta}\not{D}^4, \tag{3.6a}$$

$$4!Z^2D^2_{ij}=+D^2_{ij}\not{D}^4. \tag{3.6b}$$

If ϕ satisfies (3.4) with the $+$ sign, then (3.6b) is an identity and (3.6a) can be rewritten

$$D^2_{\alpha\beta}\,\phi_{[i]}=0,$$

which means that the spin-one component of $\phi_{[i]}$ is missing. Hence, we obtain the $Y=0$ representations of $\phi_{[i]}$. Analogously the $-$ sign is associated with the $Y=\frac{1}{2}$ representations of $\phi_{[i]}$.

We are now able to decompose the action for a general degenerate irreducible multiplet in terms of the superfield components. We are going to show that the Lagrangian density can be expressed as a total derivative in ZZ^*. The only property of the multiplets we need to use in our evaluation is (3.4). We are going to perform the explicit calculation in a particular frame of reference, one in which the Γ-cone is independent of the θ-coordinates. This allows us first to integrate in the $\theta,\bar\theta$ variables. Finally, in the next section, after defining the integration over a θ-dependent Γ-cone, we extend the above result to all frames of reference. It has been proved in the previous section that

$$I=\int_\Gamma d^4x\, d^2z\, d^4\theta\, d^4\bar\theta\,\bar\phi_{[i]}\,\phi_{[i]} =\int_\Gamma d^4x\, d^2z\,[\,\not{D}^4\bar{\not{D}}^4(\bar\phi_{[i]}\,\phi_{[i]})]_{\theta=\bar\theta=0}. \tag{3.7}$$

We may now expand the last expression in terms of the components of $\phi_{[i]}$ and $\bar\phi_{[i]}$. We obtain

$$\bar{\not{D}}^4[\bar\phi\phi\,]=\not{D}^4\bar\phi\cdot\phi-4\bar{\not{D}}^{3\dot\alpha}\bar\phi\cdot\bar D_{\dot\alpha}\,\phi+6^{2\dot\alpha\dot\beta}\cdot\bar D^2_{\dot\alpha\dot\beta}\,\phi +4\bar D_{\dot\alpha}\,\bar\phi\cdot\bar{\not{D}}^{3\dot\alpha}\,\phi+\bar\phi\cdot\bar{\not{D}}^4\phi. \tag{3.8}$$

We are now able to use explicitly the Dirac condition in each term; we have

$$\bar D_{\dot\alpha}\,\phi=(Z\not{p}_{\dot\alpha}{}^{\alpha}/\Box)\epsilon_{\tau i}D_\alpha\,\phi,$$

$$\bar{\not{D}}^4\phi=\not{D}^4(Z^{*2}/\Box)\,\phi,$$

$$\int d^4x\,\bar{\not{D}}^{3\dot\alpha}\psi\cdot\bar D_{\dot\alpha}\,\phi=-\int d^4x\,\not{D}^{3\alpha}\frac{Z^{*3}}{\Box}\times\psi\cdot D_\alpha\frac{Z^*}{\Box}\phi, \tag{3.9}$$

$$\int d^4x\,\bar{\not{D}}^{2\dot\alpha\dot\beta}\psi\cdot\bar D^2_{\dot\alpha\dot\beta}\,\phi=\int d^4x\times\not{D}^{2\alpha\beta}\frac{Z^{*2}}{\Box}\,\psi\cdot D^2_{\alpha\beta}\,\frac{Z^{*2}}{\Box}\,\phi.$$

After the substitution of (3.9) in (3.8) and the further application of D^4 to the resulting expression we obtain

$$\begin{aligned}\int d^4x\, d^2z\, \not{D}^4\bar{\not{D}}^4(\bar{\phi}\phi) = \int d^4x\, d^2z\Big[& \not{D}^4\not{D}^4\frac{Z^{*4}}{\Box^2}\bar{\phi}\cdot\phi + \not{D}^4\frac{Z^{*4}}{\Box^2}\bar{\phi}\cdot\bar{\phi}\cdot\not{D}^4\phi \\ & + \not{D}^4\bar{\phi}\cdot\not{D}^4\frac{Z^{*4}}{\Box^2}\phi + \bar{\phi}\cdot\not{D}^4\not{D}^4\frac{Z^{*4}}{\Box^2}\phi + 16\not{D}^{3\alpha}\not{D}^{3\beta}\frac{Z^{*3}}{\Box^2}\bar{\phi}\cdot D_\alpha D_\beta Z^*\phi \\ & - 16D_\alpha\not{D}^{3\beta}\frac{Z^{*3}}{\Box^2}\bar{\phi}\cdot\not{D}^{3\alpha}D_\beta Z^*\phi - 16\not{D}^{3\alpha}D_\beta Z^*\bar{\phi}\cdot D_\alpha\not{D}^{3\beta}\frac{Z^{*3}}{\Box^2}\phi \\ & + 16D_\alpha D_\beta Z^*\bar{\phi}\cdot\not{D}^{3\alpha}\not{D}^{3\beta}\frac{Z^{*3}}{\Box^2}\phi \\ & + 36\not{D}^{2\alpha\beta}\not{D}^{2rs}\frac{Z^{*2}}{\Box}\bar{\phi}\cdot D^2_{\alpha\beta}D^2_{rs}\frac{Z^{*2}}{\Box}\phi + 6\not{D}^4\not{D}^{2\alpha\beta}\frac{Z^{*2}}{\Box^2}\bar{\phi}D^2_{\alpha\beta}Z^{*2}\phi \\ & + 6\not{D}^{2\alpha\beta}\frac{Z^{*2}}{\Box^2}\bar{\phi}\cdot\not{D}^4\not{D}^2_{\alpha\beta}Z^{*2}\phi + 6\not{D}^{2\alpha\beta}\bar{\phi}D^2_{\alpha\beta}\not{D}^4\frac{Z^{*4}}{\Box^2}\phi \\ & + 6\not{D}^{2\alpha\beta}\not{D}^4\frac{Z^*}{\Box^2}\bar{\phi}D^2_{\alpha\beta}\,\phi\Big].\end{aligned}$$

We may now use (A7) to get rid of all the dual covariant derivatives, and (A9) and (A10) to express all the terms as the product of two totally antisymmetric expressions. After some manipulation we obtain

$$\begin{aligned}I &= \int d^4x\, d^2z\, \not{D}^4\bar{\not{D}}^4(\bar{\phi}\phi)\Big|_{\theta=\bar{\theta}=0} \\ &= \int d^4x\, d^2z\, 4!4!(ZZ^*)^2(\bar{\phi}\phi)\}_{\theta=\bar{\theta}=0}. \qquad (3.10)\end{aligned}$$

We notice that the action is nontrivial only in our cone formulation.

Finally, let us discuss the supersymmetric properties of our constrained actions. The basic assumption is that ϕ transforms as a scalar superfield under supersymmetric and central charge transformations. Considering such a transformation

$$\begin{aligned}x' &= x + \theta\sigma\bar{\xi} + \xi\sigma\bar{\theta}, \\ z' &= z + w + \epsilon\theta + \bar{\theta}\bar{\epsilon}, \qquad (3.11) \\ \theta' &= \theta + \epsilon,\end{aligned}$$

the transformation law for Φ is

$$\begin{aligned}&\Phi'(x + \theta\sigma\bar{\xi} + \xi\sigma\bar{\theta}, z + w + \epsilon\theta + \bar{\theta}\bar{\epsilon}, \theta + \epsilon, \bar{\theta} + \bar{\epsilon}) \\ &\quad = \Phi(x,z,\theta,\bar{\sigma}). \qquad (3.12)\end{aligned}$$

From this transformation law and the fact that the Jacobian of the supersymmetric transformations is one, the invariance of the action follows directly. This is so by the usual argument that the action is invariant provided that the Lagrangian density transforms as a scalar density under the corresponding coordinate transformation; we have used this argument earlier.[8]

It is interesting to analyze the behavior of the Lagrangian density in the (x,z) space, that is to say after performing the θ-integration.

Let us rewrite (3.10) in the following way:

$$I = \int_{z_0} d^4x\, d^2z (ZZ^*)^2 L(x,z,\theta,\bar{\theta})|_{\theta=\bar{\theta}=0}. \qquad (3.13)$$

After the transformation (3.11) we have

$$I' = \int_{z_0+\epsilon\theta'+\bar{\theta}'\bar{\epsilon}} d^4x'\, d^2z'\, d^4\theta'\, L'(x',z',\theta',\bar{\theta}'), \qquad (3.13')$$

where now Γ' is explicitly θ' dependent and we cannot proceed directly to integrate first the $\theta,\bar{\theta}$ variables. However, we may proceed further by first eliminating the θ' dependence of Γ' with the following change of variables:

$$\begin{aligned}x' &= x, \\ \theta' &= \theta, \\ z' &= z + \epsilon\theta + \bar{\theta}\bar{\epsilon}.\end{aligned}$$

We obtain

$$I' = \int_{z_0} d^4x\, d^2z\, d^4\theta\, d^4\bar{\theta}\, L'(x,z + \epsilon\theta + \bar{\theta}\bar{\epsilon},\theta,\bar{\theta}),$$

and to the first order in ϵ

$$\begin{aligned}I' &= \int_{z_0} d^4x\, d^2z (ZZ^*)^2 L'(x,z,\theta,\bar{\theta})|_{\theta=\bar{\theta}=0} \\ &\quad + \int_{z_0} d^4x\, d^2z\, d^4\theta\, d^4\bar{\theta}\, \epsilon\theta\, \partial_z L'(x,z,\theta,\bar{\theta}). \qquad (3.14)\end{aligned}$$

The first term in (3.14) gives a similar contribution to the one we have in (3.13) but now in terms of the components of L'; the second term in (3.14) is the contribution from the variation of the Γ-cone. It is a pure z-divergence and has a nonzero contribution to the corner of the manifold. In particular we note that this second term in (3.14) involves the fermionic components of ϕ, as is to be expected if we rewrite the scalar contribution (2.16) in the original frame in terms of the components in the primed frame

$$C = C' + \bar{\epsilon}\eta'.$$

We conclude that the action is invariant under SUSY and central charge transformations

$$\begin{aligned}I &= \int_\Gamma d^4x\, d^2z\, d^4\theta\, d^4\bar{\theta}\, \bar{\phi}\phi \\ &= \int_{\Gamma'} d^4x'\, d^2z'\, d^4\theta'\, d^4\bar{\theta}'\, \bar{\phi}'\phi' = I',\end{aligned}$$

and its structure as a full superspace integration is also preserved under both transformations. As a consequence of these two properties the explicit structure of the four-dimensional space-time Lagrange density is not preserved.

IV. NATURE OF THE SPECTRUM

In order to discuss the spectrum of component fields which arises from (3.10) we have to expand a little on our remarks at the end of the last section. In particular we must take account of cones Γ which also have some θ-dependence, as already appeared in (3.13′). We only considered an infinitesimal contribution from such terms there, but note that a general formula can be given for the integration over a cone of form $\Gamma_0 + f(\theta)$, where Γ_0 is θ-independent. This extension is by means of the definition

$$\int_{\Gamma_0 + f} F(z)dz = \int_{\Gamma_0} F(z)dz - f\cdot \int_0^1 F(z_0 + tf)dt, \quad (4.1)$$

when z is a single variable, and the extension of (4.1) for z being two dimensional

$$\begin{aligned}\int_{\Gamma_0 + f} F(z)d^2z = &\int_{\Gamma_0} F(z)d^2z \\ &- f_1 \int_0^1 dt_1 \int_{\Gamma_{02}} dz_2\, F(z_{01} + t_1 f_1, z_2) \\ &- f_2 \int_0^1 dt_2 \int_{\Gamma_{01}} dz_1\, F(z_1, z_{02} + t_2 f_2) \\ &+ f_1 f_2 \int_0^1 dt_1 \int_0^2 dt_2 \\ &\times F(z_{01} + t_1 f_1, z_{02} + t_2 f_2) \qquad (4.2)\end{aligned}$$

with $\Gamma_0 = \Gamma_{01} \times \Gamma_{02}$, Γ_{oi} being intervals with left-hand endpoints z_{oi}. By a similar analysis to that given at the end of the last section we see that we can include all of the components of a given irreducible representation in the final integration over bosonic coordinates upon integration over all of the θ-variables. We may therefore simplify the analysis by restricting ourselves to choosing Γ_0 to be θ-independent, and so concentrating on the purely scalar terms in the action.

The next aspect requiring clarification is concerned with the number of central charges present in the superfield $\phi_{[i]}$. For $N = 2$ we have already indicated in (1.5) that there are two independent real central charges ∂_5, ∂_6, with

$$Z = \partial_5 + i\partial_6$$

and we take $\partial_5 = \partial/\partial x^5$, $\partial_6 = \partial/\partial x^6$. Analysis of the purely scalar part of the constrained Lagrangian (2.16) has already been given when ∂_5 and ∂_6 are independent, and leads to an infinite set of propagating scalar modes.[12] These may be defined as the values of $(\partial_5^2/\Box)^n \phi_{[i]}(x, x^5_{(0)}, x^6_{(0)})$, where $(x^5_{(0)}, x^6_{(0)})$ is the vertex of Γ_0, and $n = 0, 1, \ldots$.

In order to avoid irreducible representations with an infinite number of components we imposed the additional constraint

$$\partial_6 \phi_{[i]} = 0 \qquad (4.3)$$

in our earlier deduction of the equation of motion from the action (1.6).[13] However, the most general irreducible representation is given by the more general constraint

$$Z = e^{2i\alpha} Z^*, \qquad (4.4)$$

where α is an arbitrary real number.[13] We will assume (4.4) as the further constraint instead of (4.3) (when $\alpha = 0$), so have that

$$\partial_5 = \cos\alpha\partial, \quad \partial_6 = \sin\alpha\partial, \quad Z = e^{i\alpha}\partial \qquad (4.5)$$

with (1.4) reducing to

$$\partial^2 = \Box. \qquad (4.6)$$

We may consider $\partial = \partial/\partial y$, where $y = \cos\alpha x^5 + \sin\alpha x^6$, and all superfields depend only on the combination of x^5 and x^6.

We may now evaluate the z-integral in (3.10) as

$$\iint_\Gamma dx^5\, dx^6 (\partial^2)^2 (\bar{\phi}_{[i]} \phi_{[i]})_{\theta = \bar{\theta} = 0}. \qquad (4.7)$$

This may be rewritten, using (4.5), as

$$2(\sin 2\alpha)^{-1} \iint_\Gamma dx^5\, dx^6\, \partial_5 \partial_6 [\partial^2 (\bar{\phi}_{[i]} \phi_{[i]})_{\theta = \bar{\theta} = 0}] \qquad (4.8)$$

(provided $\alpha \neq 0$ or $\pi/2$).

The spectrum of particles described by (4.8) can be most clearly seen if we choose Γ to be the region R_1: $x^5_{(0)} < x^5 < x^5_{(1)}$, $x^6_{(0)} < x^6 < x^6_{(1)}$. We may integrate (4.8) by parts, and if $A_{[i]} = \phi_{[i]}|_{\theta = \bar{\theta} = 0}$, (4.8) becomes

$$4(\sin 2\alpha)^{-1} [\bar{A}_{[i]} \Box A_{[i]} + \partial \bar{A}_{[i]} \partial A_{[i]}]|_{3-2-1+0}, \qquad (4.9)$$

where the points of 0,1,2,3 are the corners of R_1:

$$\begin{aligned} 0 &= (x^5_{(0)}, x^6_{(0)}), \quad 1 = (x^5_{(0)}, x^6_{(1)}), \\ 2 &= (x^5_{(1)}, x^6_{(0)}), \quad 3 = (x^5_{(1)}, x^6_{(1)}). \end{aligned} \qquad (4.10)$$

If we add the further constraint

$$[\bar{A}_{[i]} \Box A_{[i]} + \partial \bar{A}_{[i]} \partial A_{[i]}]|_{3-2-1} = 0 \qquad (4.11)$$

we obtain only the contribution to (4.9) from the corner 0. The modes then present in four dimensions correspond to a propagating scalar $A_{[i]}$ and an auxiliary scalar $\partial A_{[i]}$. However, the constraint (4.11), in combination with (4.6), reduces these two independent modes to one, leaving one propagating scalar. In order to regain the two independent modes present in the usual degenerate central charge representations we have to forego (4.11). The spectrum in (4.9) then appears to have both positive and negative energies, and so seems physically unacceptable. This relationship between the spectrum of the resulting theory and the boundary conditions in central charge space was already recognized in our earlier derivation of the field equations from actions integrated over central charge dimensions,[8] and it is clearly important to explore it further.

The main feature needing clarification is that of the dependence of the spectrum on the general shape of Γ_0. Since $ZZ^* = \partial_5^2 + \partial_6^2 = \nabla_2^2$, we may rewrite (3.13) as

$$\int d^4x \int_{\partial\Gamma_0} dr\, \mathbf{n}\cdot\nabla(\nabla_2^2 L|_{\theta = \bar{\theta} = 0}), \qquad (4.12)$$

where n is the unit outward normal to $\partial\Gamma_0$, which we assume to be piecewise differentiable, and dr is the arc length on $\partial\Gamma_0$. We may wish to impose the constraint (4.4) to reduce the component content of (4.12) to a finite number of degrees of freedom. We may further desire (4.12) to reduce explicitly to an integration over four-dimensional space-time. This appears to have no great urgency since there is in any case boundary control of (3.13). However, we are here trying to specify more precisely how such control can be defined so as to lead to component content of satisfactory sort to be relevant to building supersymmetric theories of physical interest. In particular we may be concerned with such features as positivity of the energy of the theory, where reduction of I to a four-dimensional integral brings us back to familiar ground.

We assume therefore that (4.12) reduces to a four-dimensional integral. For this to occur we must be able to integrate (4.12) one further time. To achieve this it would appear necessary that the constraint (4.4) (giving an irreducible degenerate representation of supersymmetry) leads to

$$\mathbf{n}\cdot\nabla(\nabla_2^2 L) \equiv \hat{\mathbf{t}}\cdot\nabla(\nabla_2^2 L), \tag{4.13}$$

where $\hat{\mathbf{t}}$ is the unit tangent vector along $\partial\Gamma_0$. If (4.13) is satisfied then a nonzero contribution to (4.12) can only arise from corners of $\partial\Gamma_0$, where $\hat{\mathbf{t}}$ is discontinuous.

The region R_1 described earlier is one such case, as is the region

$$R_2 = \{\theta_1 < \theta < \theta_2,\ 0 < r < R\},$$

where (r,θ) are polar coordinates in the (x^5,x^6) plane. For the constraint (4.4) reduces, in polar coordinates, to

$$\frac{\partial}{\partial r} = i r^{-1} g(\theta) \frac{\partial}{\partial \theta}, \tag{4.14}$$

where

$$g(\theta) = \frac{1 + e^{2i(\theta-\alpha)}}{1 - e^{2i(\theta-\alpha)}}. \tag{4.15}$$

Then the contribution to I from R_2 is

$$\begin{aligned}&[\tan(\theta_1-\alpha) - \tan(\theta_2-\alpha)](\nabla_2^2 L)_{r=0} \\ &\quad + i^{-1} g(\theta)(\nabla_2^2 L)\Big|_{r=R,\theta=\theta_1}^{r=R,\theta=\theta_2} \\ &\quad + i\int_{\theta_1}^{\theta_2} d\theta\, g(\theta)\frac{\partial}{\partial\theta}(\nabla_2^2 L)_{(R,\theta)}.\end{aligned} \tag{4.16}$$

We may regain the contribution at the origin in central charge space if we set the second term in (4.16) to zero. Again this constraint removes potentially dangerous negative-energy contributions. The region R_2 can be generalized to a segment of a circle with center at any point.

The above two examples indicate the close relation between the spectrum of I, the shape of Γ_0, and boundary conditions on $\phi_{[i]}$ at the corners of $\partial\Gamma_0$. More generally it is clear from (4.12) that there is no spectrum at all if $\partial\Gamma_0 = \phi$, for example if Γ_0 is a two-dimensional torus. Thus we only get a nontrivial spectrum if

$$\partial\Gamma_0 \neq \phi, \tag{4.17}$$

and I can only be reduced to a four-dimensional integral if

$$\text{corners of } \partial\Gamma_0 \neq \phi. \tag{4.18}$$

We note that the above analysis can be extended to any boundary $\partial\Gamma_0$ with at least one corner, which we identify with the origin. Provided we have the constraint, generalizing (4.11) and that associated with the latter part of (4.16),

$$\int_{\partial\Gamma_0} dr\, \mathbf{n}\cdot\nabla(\nabla_2^2 L|_{\theta=\bar\theta=0}) = (\nabla_2^2 L|_{\theta=\bar\theta=0})_{x^5=x^6=0}, \tag{4.19}$$

we have a satisfactory spectrum.

It may be possible to obtain a satisfactory physical theory even if (4.18) is not true. For example in the presence of interaction we might expect additional terms in I which are not total derivatives of the form (3.10). We must therefore turn to the case of interacting theories, and their quantization, in order to obtain further restrictions on Γ_0. We propose to make such an analysis elsewhere.

ACKNOWLEDGMENTS

J. H. would like to thank the Lebanese University Faculty of Science. A. R. would like to thank CONICIT of Venezuela for grants whilst this research was being completed. We would all like to thank G. Amerighi, L. Hornfeldt, and B. Rands for useful discussions during the investigation.

APPENDIX

We wish to give in this appendix some general results which follow directly from the supersymmetry algebra with central charges. The only assumption is the usual $N = 2$ anticommutation relation

$$\{D_\alpha, D_\beta\} = -2\epsilon_{\alpha\beta} z, \tag{A1}$$

where $\epsilon_{\alpha\beta} = \epsilon_{\alpha\beta}\epsilon^{ij}$. We define the totally antisymmetric objects

$$\begin{aligned} D^2_{\alpha\beta} &\equiv \sum_{\alpha\wedge\beta} D_\alpha D_\beta, \\ D^3_{\alpha\beta n} &\equiv \sum_{\alpha\wedge\beta n} D_\alpha D_\beta D_n, \\ D^4_{\alpha\beta n s} &\equiv \sum_{\alpha\beta n s} D_\alpha D_\beta D_n D_s, \end{aligned} \tag{A2}$$

whose $\Sigma(\cdot)$ means the antisymmetric part of the corresponding geometrical objects.

Applying (A1) several times we obtain

$$\begin{aligned} D_\alpha D_\beta &= D^2_{\alpha\beta} - \epsilon_{\alpha\beta} Z, \\ D_\alpha D^2_{\beta\gamma} &= D^3_{\alpha\beta\gamma} - \epsilon_{\alpha\beta} Z D_\gamma + \epsilon_{\alpha\gamma} Z D_\beta, \end{aligned} \tag{A3}$$

and in general

$$D_\alpha D^n_{\beta_1\beta_2\cdots\beta_i\beta_n} = D^{n+1}_{\alpha\beta_1\cdots\beta_n} + \sum_{i=1}^{n}(-1)^i \epsilon_{\alpha\beta_i} Z D^{n-1}_{\beta_1\cdots\beta_{i-1}\beta_{i+1}\cdots\beta_n}. \tag{A4}$$

Now we may introduce the dual covariant derivatives $\not{D}^{m_{\alpha_1}\cdots\alpha_n}$

$$\not{D}^{m_{\alpha_1}\cdots\alpha_n} = \epsilon^{\alpha_1\cdots\alpha_n\beta_1\cdots\beta_m} D^m_{\beta_1\cdots\beta_m},$$

in particular

$$\begin{aligned} \not{D}^4 &\equiv \epsilon^{\beta_1\beta_2\beta_3\beta_4} D^4_{\beta_1\beta_2\beta_3\beta_4}, \\ \not{D}^{3\alpha} &\equiv \epsilon^{\alpha\beta_1\beta_2\beta_3} D^3_{\beta_1\beta_2\beta_3}, \\ \not{D}^{2\alpha_1\alpha_2} &\equiv \epsilon^{\alpha_1\alpha_2\beta_1\beta_2} D^2_{\beta_1\beta_2}. \end{aligned} \tag{A5}$$

The $\epsilon^{\alpha\beta\gamma\delta}$ is the totally antisymmetric SL(4,C) tensor,

$$\epsilon^{\alpha\beta\gamma\delta} = \epsilon_{ij}\epsilon_{kl}\epsilon^{\alpha\delta}\epsilon^{\beta n} - \epsilon_{il}\epsilon_{jk}\epsilon^{\alpha\beta}\epsilon^{r\delta}, \tag{A6}$$

which satisfies

$$\epsilon^{\alpha\beta\gamma\delta}\epsilon_{\alpha_1\beta_1\gamma_1\delta_1} = \delta^{[\alpha}_{\alpha_1}\delta^{\beta}_{\beta_1}\delta^{\gamma}_{\gamma_1}\delta^{\delta]}_{\delta_1}.$$

It can be shown that

$$\begin{aligned} Z\bar{D}^{3\alpha} &= -\tfrac{1}{4}D^{\alpha}\bar{D}^{4}, \\ Z^2\bar{D}^{2\alpha\beta} &= -\tfrac{1}{12}D^{2\alpha\beta}\bar{D}^{4}, \\ \bar{D}^{4}\bar{D}^{4} &= 4!Z^{4}, \end{aligned} \tag{A7}$$

where $D^{\alpha} \equiv \epsilon^{\alpha\beta}D_{\beta}$, $D^{2\alpha\beta} = \epsilon^{\alpha\gamma}\epsilon^{\beta\delta}D^{2}_{\gamma\delta}$.

In addition,

$$D^{n}_{\alpha_1\cdots\alpha_n}\bar{D}^{4} = (-1)^{n}\bar{D}^{4}D^{n}_{\alpha_1\cdots\alpha_n}, \tag{A8}$$

and, with ψ and ϕ arbitrary superfields,

$$\begin{aligned} &D^{2\alpha_1\alpha_2}D^{2\beta_1\beta_2}\psi\cdot D^{2}_{\alpha_1\alpha_2}D^{2}_{\beta_1\beta_2}\phi \\ &\quad = D^{4\alpha_1\alpha_2\alpha_3\alpha_4}\psi\cdot D^{4}_{\alpha_1\alpha_2\alpha_3\alpha_4}\phi \\ &\qquad + 8ZD^{2\alpha_1\alpha_2}\psi\cdot ZD^{2}_{\alpha_1\alpha_2}\phi + 24z^2\psi\cdot Z^2\phi, \end{aligned} \tag{A9}$$

$$D^{2\alpha_1\alpha_2}D^{\beta}\psi\cdot D^{2}_{\alpha_1\alpha_2}D_{\beta}\phi = D^{\beta}D^{2\alpha_1\alpha_2}\psi\cdot D_{\beta}D^{2}_{\alpha_1\alpha_2}\phi. \tag{A10}$$

[1]Mandelstam, Nucl. Phys. B **213**, 149 (1983).

[2]P. Howe, K. Stelle, and P. Townsend, J. Phys. **43**, Coll. C-3, Suppl. to 12 (1982).

[3]V. O. Rivelles and J. G. Taylor, (a) Phys. Lett. B **104**, 131 (1981); (b) Phys. Lett. B **121**, 37 (1983).

[4]J. G. Taylor, (a) J. Phys. A **15**, 867 (1981); (b) "Extended Supergravity in Light-cone Superspace," in *Proceedings of the Karpacz School on Supersymmetry and Supergravity*, edited by B. Milewski (World Scientific, Singapore, 1983); (c) Future Avenues in Supergravity, in *Quantum Theory of Gravity*, edited by S. M. Christensen (A. Hilger, Bristol, 1984); (d) Phys. Lett. **B94**, 147 (1980).

[5]M. Rocek and W. Siegel, Phys. Lett. B **105**, 278 (1981).

[6]G. R. J. Bufton and J. G. Taylor, KCL preprint (in preparation).

[7]M. Grisaru and W. Siegel, Nucl. Phys. B **201**, 292 (1982).

[8]A. Restuccia and J. G. Taylor, "How to Integrate over Central Charges," J. Phys. A: Math. Gen. 16, 4097–4123 (1983).

[9]M. Sohnius, Nucl. Phys. B **138**, 109 (1978).

[10]B. Rands and J. G. Taylor, J. Phys. A **16**, 1005 (1983).

[11]J. Hassoun, A. Restuccia, and J. G. Taylor, KCL preprint (in preparation).

[12]D. Gorse, A. Restuccia, and J. G. Taylor, J. Phys. A **16**, 3037 (1983).

[13]A. Restuccia and J. G. Taylor, J. Phys. A **16**, 4125 (1983).

Planar factors of proper homogeneous Lorentz transformations

Donald E. Fahnline
The Pennsylvania State University, Altoona, Pennsylvania 16603

(Received 24 May 1984; accepted for publication 24 August 1984)

This article discusses two constructions factoring proper homogeneous Lorentz transformations H into the product of two planar transformations. A planar transformation is a proper homogeneous Lorentz transformation changing vectors in a two-flat through the origin, called the transformation two-flat, into new vectors in the same two-flat and which leaves unchanged vectors in the orthogonal two-flat, called the pointwise invariant two-flat. The first construction provides two planar factors such that a given timelike vector lies in the transformation two-flat of one and in the pointwise invariant two-flat of the other; it leads to several basic conditions on the trace of H and to necessary and sufficient conditions for H to be planar. The second construction yields explicit formulas for the orthogonal factors of H when they exist and are unique, where two planar transformations are orthogonal if the transformation two-flat of one is the pointwise invariant two-flat of the other.

I. INTRODUCTION

This paper generalizes two types of factorization of proper homogeneous Lorentz transformations (HLT) into the product of two planar transformations. A planar transformation is a proper HLT which in the active interpretation changes vectors in a two-flat through the origin, called the transformation two-flat, into new vectors in the same two-flat and leaves vectors in the orthogonal two-flat, called the pointwise invariant two-flat, unchanged.[1,2] The properties of several convenient expressions for planar transformations reviewed in Sec. II are the basis for the constructions.[3,4]

Section III generalizes a well-known method for expressing a given restricted (proper and orthochronous) HLT as the product of a boost and a pure spatial rotation.[5] Given a timelike vector a, the construction expresses a proper HLT as the product of two planar transformations, where a lies in the transformation two-flat of the first and in the pointwise invariant two-flat of the second. (Either or both planar transformations can degenerate to the identity.) This factorization leads to the next section's discussion of some general necessary conditions on the trace of a proper HLT and on the trace of its square.

Section V applies these conditions to provide alternate necessary and sufficient conditions for a proper HLT to be planar. One of these simplifies and generalizes a similar condition given by Rao, Saroja, and Rao.[2]

The final factorization is a modification of one considered previously by several authors. Synge[6] and Schwartz[7] discuss the factorization of a restricted HLT into the product of a timelike transformation and an orthogonal spacelike transformation using infinitesimal transformations. In this context two planar transformations are orthogonal to each other if their transformation two-flats are orthogonal. Schwartz[8] and Wigner[9] provide an alternate approach based on an analysis of the eigenvalue problem for restricted HLT. Rao, Saroja, and Rao base their discussion of this factorization on electromagnetic theory.[2] All of these authors emphasize in particular that null transformations cannot be factored into the product of orthogonal timelike and spacelike transformations. As a replacement for such a factorization Rao, Saroja, and Rao express null transformations as the product of two "exceptional" spacelike transformations.[2] However, this replacement is not appropriate because the factors are not orthogonal and because Wigner has shown that *any* restricted HLT can be expressed as the product of two exceptional transformations, which Wigner calls involutions.[9]

The modification presented here in Sec. VI attempts to express any proper HLT, nonorthochronous as well as orthochronous, as the product of two orthogonal planar transformations, null as well as timelike and spacelike. The previous results on planar transformations yield a simple derivation of explicit formulas for the factors. Applying the Cayley–Hamilton theorem provides a basis for discussing the conditions for the validity and uniqueness of the solution and for constructing projection operators onto the transformation and pointwise invariant two-flats of the factors. The section concludes with the exceptional cases. For example, the formulas fail for null transformations not because there is no solution, but because the solution is not unique; the discussion establishes the family of all such solutions. The negative of a null transformation, on the other hand, is a proper HLT for which no orthogonal planar factors exist.

II. PLANAR TRANSFORMATIONS

This section reviews and generalizes some of the properties of planar homogeneous Lorentz transformations.[3] A four-vector x has components $x^\mu \equiv (x^0; x^i)$ relative to a Lorentz frame. The scalar product of two vectors is $x\cdot y \equiv x^\mu y_\mu \equiv g_{\mu\nu} x^\mu y^\nu$, where $g^{ii} = -g^{00} = 1$ and $g^{\mu\nu} = 0$ for $\mu \neq \nu$.

The identity transformation E is the proper HLT which leaves all vectors unchanged; its elements are $E^\mu{}_\nu = g^\mu{}_\nu = \delta^\mu{}_\nu$ relative to any Lorentz frame. For any vector a such that $a\cdot a \neq 0$, the dyadic

$$I\{a\} \equiv E - 2aa/a\cdot a \tag{1}$$

is an improper HLT reflecting vectors which are multiples of a through the origin and leaving vectors which are orthogonal to a unchanged.[10]

The dyadic

$$P\{a,b\} \equiv E + 2ab/a\cdot a - (a+b)(a+b)/(a\cdot a + a\cdot b), \tag{2}$$

 0022-2488/85/020352-07$02.50

where $a\cdot a = b\cdot b \neq 0$ and $a\cdot a + a\cdot b \neq 0$ is a proper HLT (see Ref. 3). If $a = b$, it reduces to the identity E. If $a \neq b$, it is a planar transformation which changes vectors in the two-flat determined by a and b into new vectors in the same two-flat and which leaves vectors in the orthogonal two-flat invariant. The planar transformation is an orthochronous timelike transformation T for $\omega \equiv a\cdot b / a\cdot a > 1$, a null transformation N for $\omega = 1$, a spacelike transformation S for $-1 < \omega < 1$, and a nonorthochronous timelike transformation T for $\omega < -1$. Some of its other properties are

$$P\{a,b\} = I\{a\}I\{a+b\}, \tag{3}$$

$$[P^2 - 2(a\cdot b/a\cdot a)P + E][P - E] = 0, \tag{4}$$

$$P_1 = 2(1+\omega), \tag{5}$$

and

$$P_2 = (P_1 - 2)^2 = 4\omega, \tag{6}$$

where $P_1 \equiv \mathrm{Tr}\, P \equiv P^{\mu}{}_{\mu}$ and $P_2 \equiv \mathrm{Tr}(P^2)$.

For the case $\omega = -1$, Eq. (2) is indeterminant and must be replaced by the proper HLT transformation

$$\Pi\{c,b\} \equiv E + 2[(b\cdot b)cc - (b\cdot c)(bc + cb) + (c\cdot c)bb]/\delta, \tag{7}$$

where $\delta \equiv (b\cdot c)^2 - (b\cdot b)(c\cdot c) \neq 0$ (see Ref. 3). This "exceptional" planar transformation is a nonorthochronous timelike transformation T for $\delta > 0$, and it is a spacelike transformation S for $\delta < 0$. Other properties are that $\Pi\{c,b\}$ decomposes into the product of two reflections and that

$$\Pi^2 - E = 0, \tag{8}$$

$$\Pi_1 = 0, \tag{9}$$

$$\Pi_2 = 4 = (\Pi_1 - 2)^2. \tag{10}$$

The product of two reflections always yields the identity E or a planar transformation of the form P or Π:

$$B\{a,c\} \equiv I\{a\}I\{c\} \tag{11}$$

$$= E - 2aa/a\cdot a - 2cc/c\cdot c + 4(a\cdot c)ac/(a\cdot a)(c\cdot c) \tag{12}$$

$$= E, \quad \text{for } c = \phi a, \quad \phi \neq 0 \tag{13}$$

$$= P\{a, 2(a\cdot c)c/c\cdot c - a\}, \quad \text{for } a\cdot c \neq 0, \quad c \neq \phi a \tag{14}$$

$$= \Pi\{a,c\}, \quad \text{for } a\cdot c = 0, \tag{15}$$

where $a\cdot a \neq 0$ and $c\cdot c \neq 0$. For this expression one has

$$B_1 \equiv \mathrm{Tr}\, B\{a,c\} \equiv B^{\mu}{}_{\mu} = 4(a\cdot c)^2/(a\cdot a)(c\cdot c). \tag{16}$$

If $0 < (a\cdot a)(c\cdot c) < (a\cdot c)^2$, then B is an orthochronous timelike transformation T with $T_1 > 4$. If $0 < (a\cdot a)(c\cdot c) = (a\cdot c)^2$ and $c \neq \phi a$, then B is a null transformation N with $N_1 = 4$. If $0 < (a\cdot c)^2 < (a\cdot a)(c\cdot c)$, then B is a spacelike transformation S with $0 < S_1 < 4$. If $a\cdot c = 0 < (a\cdot a)(c\cdot c)$, then B is a spacelike exceptional transformation S with $S_1 = 0$. If $(a\cdot a)(c\cdot c) < 0 = a\cdot c$, then B is a nonorthochronous timelike exceptional transformation T with $T_1 = 0$. Finally, if $(a\cdot a)(c\cdot c) < 0$ with $a\cdot c \neq 0$, then B is a nonorthochronous timelike transformation T with $T_1 < 0$.

If c is linearly independent of a, then a and c determine the transformation plane of $B\{a,c\}$. If in addition at least one of these vectors is timelike, then the transformation two-flat is timelike and $B\{a,c\}$ is timelike. If both are timelike, then $B\{a,c\}$ is orthochronous; if one is timelike and the other spacelike, $B\{a,c\}$ is nonorthochronous.

Equation (11) implies

$$B^{-1}\{a,c\} = B\{c,a\} \tag{17}$$

because $I^2 = E$ for all I. The equation

$$[B^2 - (B_1 - 2)B + E][B - E] = 0 \tag{18}$$

follows either from consolidating Eqs. (4) and (5) with Eqs. (8) and (9) or directly from Eqs. (12) and (16). Expanding Eq. (18) and multiplying the result by B^{-1} yields

$$B^2 + B - E - B^{-1} = B_1(B - E). \tag{19}$$

Taking the trace of this gives

$$B_2 = (B_1 - 2)^2, \tag{20}$$

because $\mathrm{Tr}\,(B^{-1}) = \mathrm{Tr}\, B = B_1$ by Eqs. (12), (16), and (17).

If H is a proper HLT possessing a pointwise invariant two-flat through the origin, then H is the identity or is a planar transformation of the form P or Π (see Ref. 3). Hence one can always express such a transformation in the form $B\{a,c\}$.

III. FACTORIZATION OF A PROPER HLT RELATIVE TO A TIMELIKE VECTOR

It is well-known that a restricted (i.e., proper and orthochronous) HLT is a boost, a pure spatial rotation, or the product of a boost and a pure spatial rotation.[5] This section generalizes this factorization for use in deriving several conditions on the trace of a proper HLT.

Let H be a proper HLT, let a be an arbitrary timelike vector, and define a vector b by

$$b \equiv H^{-1}a. \tag{21}$$

Since $b\cdot b = a\cdot a$, either $b = \pm a$ holds or else b is linearly independent of a. If $b = a$, define $c \equiv a$; if $b = -a$, define c as an arbitrary spacelike vector orthogonal to a; if $b \neq \pm a$, define $c \equiv \phi(a + b)$, where $\phi \neq 0$. In this last case one has $c\cdot c = 2\phi^2 a\cdot(a+b) \neq 0$ because a is timelike and $b \neq -a$. Hence, for all three cases one can write

$$b = 2(c\cdot a)c/c\cdot c - a. \tag{22}$$

From a and c construct

$$B \equiv B\{a,c\} \tag{23}$$

using Eq. (12). Since a is timelike, B must be timelike or the identity. The three alternate definitions of c yield $B = E$, $B = \Pi$, and $B = P$, respectively, according to Eqs. (13), (15), and (14).

If H is orthochronous, a and b are either both future pointing or both past pointing; it follows then that c is timelike and that B is orthochronous. If H is nonorthochronous, one of a and b is future pointing and the other is past pointing; in this case it follows that c is spacelike and that B is nonorthochronous.

Next construct the proper HLT

$$C \equiv HB^{-1} = HB\{c,a\} \tag{24}$$

using Eqs. (12) and (17). It follows from Eqs. (12), (22), and (21) that

$$Ca = HB\{c,a\}a = Hb = a. \tag{25}$$

Euler's theorem applied to the restriction of C to the three-dimensional space orthogonal to a implies that C possesses at least one more invariant direction. Let s be a nonzero vector along such a direction; then $s\cdot a = 0$ implies that s must be spacelike. It follows that C has a timelike pointwise invariant plane determined by a and s and hence that C is a spacelike planar transformation S or the identity E. (Note that s exists, is nonzero, and obeys $s\cdot a = 0$ and $Cs = s$ both for $C = S$ and for $C = E$.)

Thus Eq. (24) yields

$$H = CB, \tag{26}$$

where B is the identity E or a timelike transformation T such that its transformation two-flat contains the timelike vector a, and C is the identity E or a spacelike transformation which leaves a invariant. This is called the factorization of H relative to the timelike vector a.

If $Ha \neq -a$, the factorization is unique for the given timelike vector a, because $H = C'B\{a,c'\}$ with $C'a = a$ implies

$$b' \equiv B^{-1}\{a,c'\}a = B^{-1}\{a,c'\}C'^{-1}a = H^{-1}a = b$$

by Eq. (21). Using Eqs. (12) and (17) to express $B^{-1}\{a,c'\}$ in terms of a and c' and applying it to a yield b' in terms of a and c'; comparing the result to Eq. (22) for b gives

$$(a\cdot c')c'/c'\cdot c' = (a\cdot c)c/c\cdot c$$

so that $c' = \phi' c$ for some scalar ϕ'. Equation (12) then yields $B\{a,c'\} = B\{a,c\}$, and it follows that $C' = C$. If $Ha = -a$, on the other hand, the vector c is not unique and neither is the factorization.

Applying the factorization to the inverse transformation H^{-1} yields $H^{-1} = C'B'$; hence $H = B'^{-1}C'^{-1} \equiv B''C''$ is a factorization of H with respect to a in reverse order, and it is unique if $Ha \neq -a$. Since Eqs. (26), (12), and (25) yield

$$H = (CBC^{-1})C = B\{Ca,Cc\}C = B\{a,Cc\}C,$$

the uniqueness implies $C'' = C$ and $B'' = B\{a,Cc\}$.

IV. TRACE CONDITIONS ON PROPER HLT

Equation (26) and the properties of B and C lead to useful expressions for the trace of H and of H^2. By Eqs. (26), (23), (12), and (25) one has

$$H = C[E - 2aa/a\cdot a - 2cc/c\cdot c + 4(a\cdot c)ac/a\cdot ac\cdot c]$$

$$= B + C - E - 2(c' - c)c/c\cdot c, \tag{27}$$

where $c' \equiv Cc$. Taking the trace of this result yields

$$H_1 = B_1 + C_1 - 4 - \psi, \tag{28}$$

where $\psi \equiv 2(c' - c)\cdot c/c\cdot c$. Transposing E in Eq. (27) to the left member, squaring the result, and then taking the trace yield

$$H_2 + 4 = B_2 + C_2 + \psi^2 - \zeta - \xi, \tag{29}$$

where $\zeta \equiv 4[B(c' - c)]\cdot c/c\cdot c$ and $\xi \equiv 4[C(c' - c)]\cdot c/c\cdot c$.

Using Eqs. (23) and (12) to find Bc, taking the scalar product with c, and using Eq. (16) yield

$$(Bc)\cdot c/c\cdot c = B_1/2 - 1. \tag{30}$$

Similarly, using $B^{-1} = B\{c,a\}$ and $c'\cdot a = c\cdot(C^{-1}a) = c\cdot a$ with the same equations yields

$$(Bc')\cdot c/c\cdot c = c'\cdot(B^{-1}c)/c\cdot c$$

$$= (B_1 - 1)c'\cdot c/c\cdot c - B_1/2. \tag{31}$$

Combining Eqs. (30) and (31) gives

$$\zeta = 2(1 - B_1)\psi. \tag{32}$$

To reduce ξ, apply Eq. (19) with C substituted for B to c, take the scalar product of the resulting equation with c, and rearrange to obtain

$$\xi = 2(1 - C_1)\psi. \tag{33}$$

Substituting Eqs. (32) and (33) into Eq. (29), using Eq. (20) to reduce B_2 and C_2, and using Eq. (28) yield

$$H_2 = (H_1 + 2)^2 - 2B_1C_1. \tag{34}$$

It remains to evaluate ψ. Define a vector d by the equations

$$d \equiv \begin{cases} c - (c\cdot a)a/a\cdot a, & \text{if } B \neq E, \\ s, & \text{if } B = E, \end{cases} \tag{35}$$

where s is the nonzero vector defined beneath Eq. (25). It follows that d is always nonzero and that $a\cdot d = 0$. Since a is timelike, one must have

$$d\cdot d > 0; \tag{36}$$

similarly, one has

$$s\cdot s > 0. \tag{37}$$

Next define the vector

$$f \equiv d - (s\cdot d)s/s\cdot s, \tag{38}$$

which obeys $f\cdot s = f\cdot a = 0$. It follows that f lies in the transformation plane of C, that

$$f\cdot f \geqslant 0 \tag{39}$$

and that

$$d\cdot d = f\cdot f + (s\cdot d)^2/s\cdot s. \tag{40}$$

Equations (16) and (35) yield

$$c\cdot d/c\cdot c = 1 - B_1/4, \tag{41}$$

while Eqs. (5) and (9) yield

$$f\cdot(f' - f) = (C_1/2 - 2)f\cdot f, \tag{42}$$

where $f' = Cf$. It follows from $Ca = a$, $Cs = s$, and Eqs. (35) and (38) that

$$c\cdot(c' - c) = c\cdot(f' - f) = f\cdot(f' - f).$$

Consequently, the definition of ψ with Eqs. (40) and (41) yields

$$\psi = \frac{2f\cdot(f' - f)}{c\cdot c} = \frac{2c\cdot d}{c\cdot c}\frac{f\cdot(f' - f)}{d\cdot d}$$

$$= (B_1 - 4)(4 - C_1)f\cdot f/4d\cdot d. \tag{43}$$

[If $B \neq E$, one has $c\cdot d = d\cdot d$ from Eq. (35); if $B = E$, one has $c\cdot d = a\cdot s = 0 = \psi$ and Eq. (43) is still correct.]

It is convenient to also define

$$\eta \equiv (B_1 - 4)(4 - C_1)/4 - \psi \tag{44}$$

$$= (B_1 - 4)(4 - C_1)(s\cdot d)^2/4(d\cdot d)(s\cdot s), \tag{45}$$

where the equality results from Eqs. (40) and (43). Using Eq. (44) in Eq. (28) yields

$$H_1 = B_1 + C_1 - 4 - \psi = B_1C_1/4 + \eta, \tag{46}$$

and combining Eq. (46) with Eq. (34) yields

$$H_2 = (H_1 + 2)^2 - 2B_1C_1 = (H_1 - 2)^2 + 8\eta. \tag{47}$$

Equation (47) implies

$$2H_2 - H_1^2 + 8 = (H_1 + 4)^2 - 4B_1C_1 = (H_1 - 4)^2 + 16\eta. \tag{48}$$

Since C can be spacelike, spacelike exceptional, or the identity, one has $0 \leqslant C_1 \leqslant 4$. If H is orthochronous, B is also orthochronous and one has $B_1 \geqslant 4$. Then Eqs. (36), (37), and (45) yield $0 \leqslant \eta \leqslant (B_1 - 4)(4 - C_1)$, Eq. (46) yields

$$0 \leqslant C_1 \leqslant B_1C_1/4 \leqslant H_1 \leqslant B_1 + C_1 - 4 \leqslant B_1, \tag{49a}$$

and Eq. (48) yields

$$0 \leqslant (H_1 - 4)^2 \leqslant 2H_2 - H_1{}^2 + 8 \leqslant (H_1 + 4)^2. \tag{50a}$$

Similarly, if H is nonorthochronous it follows that B is nonorthochronous, that $B_1 \leqslant 0$, and that $(B_1 - 4)(4 - C_1) \leqslant \eta \leqslant 0$. In this case Eqs. (46) and (47) yield

$$B_1 \leqslant B_1 + C_1 - 4 \leqslant H_1 \leqslant B_1C_1/4 \leqslant 0 \tag{49b}$$

and

$$0 \leqslant (H_1 + 4)^2 \leqslant 2H_2 - H_1^2 + 8 \leqslant (H_1 - 4)^2. \tag{50b}$$

V. CONDITIONS FOR PLANAR TRANSFORMATIONS

According to Sec. I the form $B\{a,c\}$ given in Eqs. (11) and (12) suffices for expressing all planar HLT. Hence all planar HLT must satisfy Eq. (20). The present section shows that this equation is also a sufficient condition for a proper HLT to be a planar transformation or the identity.

If H is a proper HLT and it satisfies

$$H_2 = (H_1 - 2)^2, \tag{51}$$

then for any choice of the timelike vector a the constructions of the previous section lead to $\eta = 0$ via Eq. (47). It then follows from Eq. (45) that at least one of $B_1 = 4$, $C_1 = 4$, or $s\cdot d = 0$ must hold. In the first case one has $B = E$ and hence $H = C$, a planar transformation or E; in the second case one has $C = E$ and hence $H = B$, a planar transformation or E. If $B \neq E$ and $C \neq E$, one must have the third case $s\cdot d = 0$; Eq. (35) then gives $s\cdot c = 0$. This result combined with $s\cdot a = 0$ and Eqs. (23) and (12) yields $Bs = s$. Since s obeys $Cs = s$ by definition, it follows from Eq. (26) that $Hs = s$, where s is a spacelike vector. Although the desired conclusion that H is planar is already evident for the first two of these cases, it is convenient to use the fact that the definition of s is valid for all three cases to obtain the following summary: If H is a proper HLT satisfying Eq. (51) and a is any given timelike vector, then H possesses an invariant spacelike vector s orthogonal to a.

After having found such a vector for a given timelike vector , one can repeat the entire construction starting from a new timelike vector $a' \equiv \alpha a + \sigma s$, where $\sigma \neq 0$. The result is a new spacelike vector s' invariant under H and orthogonal to a'. Since the definition of a' shows that s is not orthogonal to a', it follows that s' and s are linearly independent. Hence a proper HLT satisfying Eq. (51) possesses a pointwise invariant two-flat through the origin and must be a planar transformation or the identity.

Since the equivalent Eqs. (18) and (19) imply Eq. (21), it follows that they are each a necessary and sufficient condition for a proper HLT to be a planar transformation or the identity.

VI. ORTHOGONAL PLANAR FACTORS FOR PROPER HLT

This section examines the factorization $H = BC$ of a proper HLT into the product of two orthogonal planar transformations, where two planar transformations are called orthogonal if their transformation planes are orthogonal. Suppose first that such a factorization is possible for a given H; then Eq. (27) gives

$$H = CB = BC \tag{52}$$

and

$$B + C = H + E, \tag{53}$$

because $c' \equiv Bc = c$. Equation (52) implies $H^{-1} = B^{-1}C^{-1} = C^{-1}B^{-1}$ so that Eq. (27) similarly yields

$$B^{-1} + C^{-1} = H^{-1} + E. \tag{54}$$

Squaring Eq. (53) and using Eq. (52) to eliminate H leads to

$$B^2 + C^2 = H^2 + E. \tag{55}$$

Since B is planar, it must satisfy Eq. (19), which can be written in the form

$$B_1B = B^2 + B + (B_1 - 1)E - B^{-1}. \tag{56}$$

Adding the similar expression for C_1C to Eq. (56) and using Eqs. (53)–(55) yield

$$B_1B + C_1C = H^2 + H + (B_1 + C_1 - 1)E - H^{-1}. \tag{57}$$

Solving Eqs. (53) and (57) simultaneously for B and C yields

$$(B_1 - C_1)B = H^2 - (C_1 - 1)H + (B_1 - 1)E - H^{-1} \tag{58}$$

and

$$(C_1 - B_1)C = H^2 - (B_1 - 1)H + (C_1 - 1)E - H^{-1}. \tag{59}$$

The orthogonality condition on H in the form

$$(H^{-1})^{\mu}{}_{\nu} = H_{\nu}{}^{\mu} \tag{60}$$

implies

$$\mathrm{Tr}\,(H^{-1}) = H_1. \tag{61}$$

Hence, the trace of Eq. (57) yields

$$B_1^2 + C_1^2 = H_2 + 4(B_1 + C_1 - 1). \tag{62}$$

Solving this simultaneously with the trace of Eq. (53)

$$B_1 + C_1 = H_1 + 4 \tag{63}$$

yields

$$B_1 - C_1 = \epsilon(2H_2 - H_1^2 + 8)^{1/2}, \tag{64}$$

$$2B_1 = H_1 + 4 + \epsilon(2H_2 - H_1^2 + 8)^{1/2}, \tag{65}$$

and

$$2C_1 = H_1 + 4 - \epsilon(2H_2 - H_1^2 + 8)^{1/2}, \tag{66}$$

where, for convenience, $\epsilon \equiv 1$ for orthochronous H and $\epsilon \equiv -1$ for nonorthochronous H.

It has been shown so far that if H is the product of two orthogonal planar transformations B and C, then they must satisfy Eqs. (58), (59), (65), and (66). However, it remains to show whether or not such a factorization exists for a given proper HLT; one must check that B_1 and C_1 are real, that B and C exist as real proper HLT's, and that B and C are orthogonal planar factors of H.

The first checks are already easy to perform: Eqs. (50a) and (50b) show that B_1 and C_1 defined by Eqs. (65) and (66) are real given that H is a proper HLT. Further, if H is also orthochronous, Eq. (50a) implies

$$4<B_1<H_1+4, \tag{67a}$$

$$0<C_1<4; \tag{68}$$

if H is proper and nonorthochronous, Eq. (50b) yields Eq. (68) and

$$H_1<B_1<0. \tag{67b}$$

The matrices B and C determined by Eqs. (58) and (59) both exist and are real as long as $B_1 \neq C_1$.

The remaining checks require some of the results of the eigenvalue problem for H. For a general HLT, proper or improper, the eigenvalue equation

$$Hx=\lambda x \tag{69}$$

has the characteristic equation

$$|H-\lambda E| = \mathrm{Det}||H^{\mu}{}_{\nu} - \lambda g^{\mu}{}_{\nu}|| = 0. \tag{70}$$

One may use

$$|A| \equiv \delta^{\mu\nu\rho\sigma}_{\alpha\beta\gamma\delta} A^{\alpha}{}_{\mu} A^{\beta}{}_{\nu} A^{\gamma}{}_{\rho} A^{\delta}{}_{\sigma}/4!$$

to expand Eq. (70), where $\delta^{\mu\nu\rho\sigma}_{\alpha\beta\gamma\delta} = -\epsilon^{\mu\nu\rho\sigma}\epsilon_{\alpha\beta\gamma\delta}$ and $\epsilon_{\alpha\beta\gamma\delta}$ is the completely antisymmetric Levi–Civita tensor with $\epsilon_{0123}=1$ (see Ref. 11). The coefficient of λ^4 in the result is $\delta^{\alpha\beta\gamma\delta}_{\alpha\beta\gamma\delta}/4! = 1$. The coefficient of λ^3 is $-4\delta^{\mu\nu\rho\sigma}_{\alpha\nu\rho\sigma}H^{\alpha}{}_{\mu}/4! = -H^{\mu}{}_{\mu} \equiv -H_1$. The coefficient of λ^2 is $6\delta^{\mu\nu\rho\sigma}_{\alpha\beta\rho\sigma}H^{\alpha}{}_{\mu}H^{\beta}{}_{\nu}/4! = (H_2 - H_1^2)/2$. The coefficient of λ is $-4\delta^{\mu\nu\rho\sigma}_{\alpha\beta\gamma\sigma}H^{\alpha}{}_{\mu}H^{\beta}{}_{\nu}H^{\gamma}{}_{\rho}/4! = -|H|H_1$, where the formula $(A^{-1})^{\mu}{}_{\nu} = \delta^{\mu\nu\rho\sigma}_{\alpha\beta\gamma\delta}A^{\beta}{}_{\nu}A^{\gamma}{}_{\rho}A^{\delta}{}_{\sigma}/3!|A|$ and the orthogonality condition in Eq. (60) have been used.[11] Finally, the coefficient of λ^0 is $\delta^{\mu\nu\rho\sigma}_{\alpha\beta\gamma\delta}H^{\alpha}{}_{\mu}H^{\beta}{}_{\nu}H^{\gamma}{}_{\rho}H^{\delta}{}_{\sigma}/4! = |H|$. Hence the expanded form of Eq. (70) is

$$\lambda^4 - H_1\lambda^3 + \tfrac{1}{2}(H_1^2 - H_2)\lambda^2 - |H|H_1\lambda + |H| = 0. \tag{71}$$

The Cayley–Hamilton theorem[12] applied to Eq. (71) implies that all HLT must satisfy the equation

$$H^4 - H_1H^3 + \tfrac{1}{2}(H_1^2 - H_2)H^2 - |H|H_1H + |H|E = 0. \tag{72}$$

Taking the trace of Eq. (72) and multiplying by 2 yield

$$2H_4 - 2H_1H_3 + H_1^2H_2 - H_2^2 - 2|H|H_1^2 + 8|H| = 0, \tag{73}$$

where $H_3 \equiv \mathrm{Tr}(H^3)$ and $H_4 \equiv \mathrm{Tr}(H^4)$. Multiplying Eq. (72) by $2H^{-1}$, taking the trace, and using Eq. (61) yield

$$2H_3 - 3H_1H_2 + H_1^3 - 6|H|H_1 = 0. \tag{74}$$

Multiplying Eq. (73) by 3 and subtracting the product of Eq. (74) with H_1 yield

$$24|H| = H_1^4 - 6H_1^2H_2 + 3H_2^2 + 8H_1H_3 - 6H_4. \tag{75}$$

Multiplying Eq. (72) by H^{-2} yields

$$H^2 - H_1H + \tfrac{1}{2}(H_1^2 - H_2)E - |H|H_1H^{-1} + |H|H^{-2} = 0, \tag{76}$$

and the trace of this yields

$$(|H|-1)(H_1^2 - H_2) = 0. \tag{77}$$

Assume for the remainder of this paper that H is proper so that $|H|=1$. Then Eqs. (71) and (72) factor to

$$[\lambda^2 - (B_1-2)\lambda + 1][\lambda^2 - (C_1-2)\lambda + 1] = 0 \tag{78}$$

and

$$[H^2 - (B_1-2)H + E][H^2 - (C_1-2)H + E] = 0, \tag{79}$$

where B_1 and C_1 are given by Eqs. (65) and (66) and are real quantities obeying Eqs. (67a), (67b), and (68). Although Eqs. (78), (67a), (67b), and (68) lead immediately to a complete solution for the eigenvalues of a proper HLT, the results are already well-known and are not needed here.[9]

If H is a proper HLT for which $B_1 \neq C_1$, the operators P_B and P_C defined by

$$P_B \equiv [H + H^{-1} - (C_1-2)E]/(B_1 - C_1), \tag{80}$$

$$P_C \equiv [H + H^{-1} - (B_1-2)E]/(C_1 - B_1) \tag{81}$$

exist. Adding Eqs. (80) and (81) yields

$$P_B + P_C = E, \tag{82}$$

while multiplying them and using Eq. (79) yield

$$P_BP_C = P_CP_B = 0. \tag{83}$$

Equation (79) also yields

$$P_B^2 = P_B \tag{84}$$

and

$$P_C^2 = P_C. \tag{85}$$

Equations (82)–(85) show that P_B and P_C are projection operators.[12] The operators B and C defined by Eqs. (58) and (59) for $B_1 \neq C_1$ are the same as

$$B = P_C + HP_B, \tag{86}$$

$$C = P_B + HP_C. \tag{87}$$

Since P_B and P_C commute with H, multiplying Eqs. (86) and (87) and using Eqs. (82)–(85) to simplify yield

$$H = BC = CB. \tag{88}$$

It follows from Eqs. (60), (80), and (81) that $P_{B\nu}{}^{\mu} = P_B{}^{\mu}{}_{\nu}$ and $P_{C\nu}{}^{\mu} = P_C{}^{\mu}{}_{\nu}$; then using Eq. (60) in Eq. (86) yields

$$B_{\nu}{}^{\alpha} = P_C{}^{\alpha}{}_{\nu} + H^{-1\alpha}{}_{\beta}P_B{}^{\beta}{}_{\nu}. \tag{89}$$

Equations (86) and (89) with Eqs. (82)–(85) give

$$B^{\mu}{}_{\alpha}B_{\nu}{}^{\alpha} = g^{\mu}{}_{\nu}, \tag{90}$$

which shows that B is a HLT. From Eq. (88) it then follows that $C = HB^{-1}$ is also a HLT.

To calculate B^2, square Eq. (86) and simplify using Eqs. (82)–(85) to obtain

$$B^2 = P_C + H^2P_B. \tag{91}$$

Substituting Eqs. (80) and (81) into Eq. (91), multiplying out, collecting terms, taking the trace, and using Eq. (74) to eliminate H_3 yield

$$B_2 = (B_1 - 2)^2. \tag{92}$$

Since B is a HLT, it must obey Eq. (77):

$$(|B|-1)(B_1^2 - B_2) = 0. \tag{93}$$

Substituting Eq. (92) into Eq. (93) to eliminate B_2 yields

$$(|B|-1)(B_1-1) = 0.$$

But Eqs. (67a) and (67b) state that $B_1 \neq 1$, so the last equation implies that $|B| = 1$ and that B is proper. It then follows that C is also proper. Equation (92) now proves that B is a planar transformation.

It follows from Eqs. (82)–(86) that

$$(P_C + H^{-1}P_B)B = E,$$

which proves that

$$B^{-1} = P_C + H^{-1}P_B. \tag{94}$$

Assuming that $B_1 \neq 4$ in addition to $B_1 \neq C_1$, one can apply the previous results to $H' \equiv B$. Equations (65) and (66) give $B_1' = B_1$ and $C_1' = 4$: Eqs. (79) and (80) then yield

$$P_B' = P_B. \tag{95}$$

From this and Eq. (82) it follows that

$$P_C' = E - P_B' = P_C. \tag{96}$$

Since B is a planar transformation, one can express it in the form $B\{a,c\}$ using Eq. (12). Substituting this expression into Eq. (80) yields

$$P_B = P_B' = [(a\cdot c)(ac + ca) - (c\cdot c)aa - (a\cdot a)cc] \times [(a\cdot c)^2 - (a\cdot a)(c\cdot c)]^{-1}, \tag{97}$$

which is the projection operator onto the transformation two-flat of B:

$$P_B\cdot a = a,$$

$$P_B\cdot c = c,$$

$$P_B\cdot x = 0, \quad \text{for } x\cdot a = x\cdot c = 0. \tag{98}$$

From Eq. (96) it now follows that P_C projects onto the pointwise invariant plane of B.

Similarly, if $C_1 \neq 4$ in addition to $B_1 \neq C_1$, then P_B projects onto the pointwise invariant two-flat of C, and P_C projects onto the transformation two-flat of C.

For the case $B_1 \neq 4$, $C_1 \neq 4$, and $B_1 \neq C_1$, these results indicate that B and C are orthogonal factors because the transformation two-flat of B coincides with the pointwise invariant two-flat of C and vice versa. From Eqs. (67a), (67b), and (68) it follows that B is timelike and that C is spacelike.

If B is timelike, then P_B exists and $P_B n$, where $n^\mu = (1;0,0,0)$, is a nonzero vector lying in the transformation two-flat of B. Applying H or B to this vector yields a second vector lying in the transformation two-flat of B. The two vectors together determine this transformation two-flat and B.

Combining Eqs. (62) and (63) yields

$$H_2 - (H_1 - 2)^2 = 2(B_1 - 4)(4 - C_1). \tag{99}$$

Hence, if either B_1 or C_1 is equal to 4, then H satisfies Eq. (51) and must be planar. For the case $B_1 \neq 4$ and $C_1 = 4$, Eq. (63) gives $B_1 = H_1$, and Eq. (19) applied to H gives

$$H^2 = (B_1 - 1)(H - E) + H^{-1}.$$

Substituting this expression for H^2 into Eqs. (58) and (59) yields

$$B = H, \quad C = E.$$

In other words, a timelike planar transformation has no orthogonal planar factors. Nevertheless, Eqs. (95)–(98) show that the operators P_B and P_C still exist and are the appropriate projection operators.

Similarly, applying Eq. (99) to the case $B_1 = 4$ and $C_1 \neq 4$ leads to

$$B = E, \quad C = H,$$

which says that a spacelike planar transformation has no orthogonal planar factors.

Now consider the case $B_1 = C_1$, for which Eqs. (58) and (59) fail. According to Eqs. (67a), (67b), and (68) this can occur only for $B_1 = C_1 = 4$ or $B_1 = C_1 = 0$. If $B_1 = C_1 = 4$, Eq. (63) gives $H_1 = 4$ and Eq. (62) gives $H_2 = 4$. It follows that H satisfies Eq. (51) and hence that H is either a null planar transformation N or the identity E. On the other hand, for this case Eq. (57) reduces to Eq. (53) so that these two equations for B and C become indeterminate. For $H = E$, Eq. (52) gives $C = B^{-1}$; in other words, there are no orthogonal planar factors of the identity. The following discussion shows that for $H = N$ there are such factors, and they are not unique.

Any null planar transformation has the form

$$N \equiv N[z,b] \equiv E + (2zb - 2bz - zz)/2b\cdot b, \tag{100}$$

where $z\cdot b = z\cdot z = 0$ and $b\cdot b > 0$ (see Ref. 3). It has the properties

$$Nz = z, \tag{101}$$

$$Nb = b + z, \tag{102}$$

$$Nc = c, \quad \text{for } c\cdot z = c\cdot b = 0, \tag{103}$$

$$N^{-1} = N[z, -b] = N[-z,b], \tag{104}$$

$$N[az,ab] = N[z,b], \quad \text{for } \alpha \neq 0. \tag{105}$$

Let d be any spacelike vector orthogonal to z and let α be a nonzero scalar; then Eqs. (100) and (105) yield

$$N[az,d]N[z,b] = N[z,d/\alpha]N[z,b] = E + zs - sz - (s\cdot s)zz/2, \tag{106}$$

where $s \equiv b/b\cdot b + \alpha d/d\cdot d$. It follows from $b\cdot z = d\cdot z = 0$ that $s\cdot z = 0$; hence one has $s\cdot s \geqslant 0$. If $s\cdot s = 0$, either $s = 0$ holds or else s is a scalar multiple of z; both cases give

$$N[az,d]N[z,b] = E \quad \text{for } s\cdot s = 0. \tag{107}$$

If $s\cdot s > 0$ holds, then $f \equiv s/s\cdot s$ exists and is spacelike, and Eq. (106) gives

$$N[az,d]N[z,b] = N[z,f]. \tag{108}$$

Thus the product of two null planar transformations with a common null invariant direction is either a new null planar transformation or the identity.

Now let H be a null transformation. Then there exist vectors z and b such that $H = N[z,b]$, where $z\cdot z = z\cdot b = 0$ and $b\cdot b > 0$. There also exists a vector c such that $z\cdot c = b\cdot z = 0$ and $c\cdot c > 0$. Choose any nonzero vector $d \equiv \beta b + \gamma c$, where β and γ are arbitrary scalars. Applying Eqs. (104) and (105) to d yields

$$N^{-1}d = N[z, -b]\, d = d - \beta z. \tag{109}$$

Since $d\cdot z = 0$ and $d\cdot d > 0$, the null transformation $N_d \equiv N[\beta z, d]$ exists. Hence the transformation

$$N_e \equiv N N_d^{-1} \tag{110}$$

exists, and it must be a null transformation or E because it is the product of two null transformations with a common null direction. (One has $N_e = E$ if and only if $\gamma = 0$.) The definitions of N_e, N, and N_d yield

$$N_e z = z, \tag{111}$$

$$N_e d = d. \tag{112}$$

Hence N_d and N_e are orthogonal null planar transformations with

$$N = N_e N_d$$

by Eq. (110). Since β and γ in the definition of d are arbitrary scalars, the factorization is not unique. However, since $B_1 = C_1 = 4$ implies that B and C can ony be null planar transformations or the identity, this is the only type of orthogonal planar factorization of a null planar transformation.

The only case remaining has $B_1 = C_1 = 0$, which by Eqs. (63) and (62) requires that $H_1 = -4$ and $H_2 = 4$. Let $G = -H$ so that G is a proper HLT with $G_1 = 4$ and $G_2 = 4$; then G satisfies Eq. (51) and must be a null transformation or the identity. Hence, this case occurs only for $H = -E$ or $H = -N$. One also has from Eq. (19) applied to G that

$$H^2 = -3H - 3E - H^{-1}, \tag{113}$$

while Eq. (57), which must be true if H has orthogonal planar factors, yields

$$H^2 = H^{-1} + E - H \tag{114}$$

for this case. Equating Eqs. (113) and (114) gives

$$(H+E)^2 = 0 \tag{115}$$

as a necessary condition in this case for H to have orthogonal planar factors. However, using $H = -N[z,b]$ with Eq. (100) yields

$$(H+E)^2 = -zz \neq 0,$$

which shows that the case $H = -N$ cannot be factored. It is obvious that the case $H = -E$ satisfies Eq. (115) and that any exceptional transformation $B = \Pi$ and its negative $C = -\Pi$ are orthogonal planar factors. Since $B_1 = C_1 = 0$ requires that B and C be exceptional planar transformations, these are the only solutions.

VII. SUMMARY

For any proper homogeneous Lorentz transformation H and any timelike vector a, the first construction presented here gives a factorization $H = CB$ into the product of two planar transformations, where a lies in the transformation two-flat of B and in the pointwise invariant plane of C. For $a^\mu = n^\mu \equiv (1{:}0,0,0)$, it reduces to the well-known factorization of H into the product of a pure Lorentz transformation and a pure spatial rotation. In combination with the properties of planar transformations, the factorization yields the conditions expressed in Eqs. (49) and (50) on the trace of H and of H^2. The freedom to choose the vector a then leads to a proof that Eqs. (18) and (20) are each a necessary and sufficient condition for H to be a planar transformation or the identity.

Equations (58), (59), (65), and (66) factor H into the product of two orthogonal planar transformations whenever these factors exist and are unique; Eqs. (80) and (81) give projection operators onto the orthogonal transformation two-flats of the factors. This factorization is trivial if H is a timelike or spacelike planar transformation, but the projection operators are still useful for projecting onto the transformation and pointwise invariant two-flats of H. The construction fails if H is the identity or the negative of a null transformation because no orthogonal factorization exists. The construction also fails if H is a null transformation or the negative of the identity because, although orthogonal factorizations exist, they are not unique. In all other cases the construction is valid and yields the unique planar factors.

[1] J. L. Synge, *Relativity: The Special Theory* (North-Holland, Amsterdam, 1972), 2nd ed., pp. 60–1, 79–80.
[2] K. N. S. Rao, D. Saroja, and A. V. G. Rao, J. Math. Phys. **22**, 2167 (1981).
[3] D. E. Fahnline, J. Math. Phys. **24**, 1080 (1983).
[4] H. Bacry, *Lectures on Group Theory and Particle Theory* (Gordon and Breach, London, 1977), p. 335.
[5] See, for example, J. L. Anderson, *Principles of Relativity Physics* (Academic, New York, 1967), p. 147.
[6] Reference 1, J. L. Synge, pp. 85–9.
[7] H. M. Schwartz, Am. J. Phys. **33**, 376 (1965).
[8] H. M. Schwartz, Am. J. Phys. **31**, 864 (1963).
[9] E. Wigner, Ann. Math. **40**, 149 (1939).
[10] A. O. Barut, *Electrodynamics and Classical Theory of Fields and Particles* (Macmillan, New York, 1964), pp. 14 and 15.
[11] See, for example, C. W. Misner, K. S.Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), pp. 160–1.
[12] See, for example, G. Birkhoff and S. MacLane, *A Survey of Modern Algebra* (Macmillan, New York, 1969), 3rd ed., pp. 299 and 300.