Coherent states over symplectic homogeneous spaces

S. De Bièvre
Department of Mathematics, University of Toronto, Toronto, Ontario M5S 1A1, Canada

(Received 18 October 1988; accepted for publication 11 January 1989)

A generalization of the Perelomov procedure for the construction of coherent states is proposed. The new procedure is used to construct systems of coherent states in the carrier spaces of unitary irreducible representations of groups $G = S \circledS V$, where V is a vector space and $S \subset \mathrm{GL}(V)$. The coherent states are shown to be labeled by the points in cotangent bundles $T^*\mathcal{O}^*$ of orbits $\mathcal{O}^*$ of S in V^*, the dual of V; it is proven that $T^*\mathcal{O}^*$ is a symplectic homogeneous space of G. The generalized procedure for the construction of coherent states presented in this paper is shown to encompass as special cases the constructions known in the literature for the coherent states of the Weyl–Heisenberg, the "$ax + b$," and the Galilei and Poincaré groups. Moreover, completely new sets of coherent states are constructed for the Euclidean group $E(n)$, where the Perelomov construction fails.

I. INTRODUCTION

Continuously labeled, overcomplete sets of vectors in Hilbert space, referred to as coherent states, are used extensively in the physics literature for a variety of purposes,[1–3] as well as in signal analysis (see Refs. 3 and 4 and the references therein).

In this paper we construct coherent states in irreducible representation spaces of noncompact Lie groups $G = S \circledS V$, where V is a real vector space and $S \subset \mathrm{GL}(V)$. Our construction is a generalization of the Perelomov construction[5] and works in a number of cases where the latter fails. The coherent states for the Weyl–Heisenberg group,[1,2,5] the "$ax + b$" group,[1,3,4,6] and the Galilei and Poincaré groups[7,8] occur as special cases of our construction. The coherent states we construct are in all cases labeled by points in $X = T^*\mathcal{O}^*$, where $\mathcal{O}^*$ is an orbit of S in V^*, the dual of V. The cotangent bundle $T^*\mathcal{O}^*$ is proven to be a symplectic homogeneous space of G (Theorem 2.1) and the importance of the role of symplectic geometry will be stressed repeatedly. Interesting new sets of coherent states obtained from our construction include coherent states in $L^2(S^{(n)},\omega)$, where $S^{(n)}$ is the n sphere and ω is its Riemannian volume element. In this case the coherent states are labeled by the points in $X = T^*S^{(n)}$, the cotangent bundle to the sphere (Sec. IV). The construction also provides coherent states in $L^2(H^n,\omega)$, where H^n is the n-dimensional Poincaré half-space and ω is its Riemannian volume element; as before, the labeling is done with points in T^*H^n. In particular, this should allow one to study quantization problems for systems that have $S^{(n)}$, H^n or more generally, $\mathcal{O}^*$ as configuration spaces, much in the same way as previously done for $\mathbb{R}^n$ (Refs. 9 and 10); this point was briefly elaborated in Ref. 11 and we hope to return to it in a later publication.

Before outlining the generalized construction we propose, we first recall the general definition of coherent states.[1] Let $\mathcal{H}$ be a Hilbert space and X a smooth oriented manifold with volume form Ω. Let

$$C: x \in X \to C(x) \in \mathcal{H} \tag{1.1}$$

be a weakly C^∞ map such that

$$\int_X \Omega(x) \langle C(x), \psi \rangle C(x) = \psi, \quad \forall \psi \in \mathcal{H}, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathcal{H}$. Then we call the set $C(X)$ a collection of coherent states on $\mathcal{H}$. Note that as a result of (1.2), the map

$$W: \psi \in \mathcal{H} \mapsto W\psi \in L^2(X,\Omega), \tag{1.3}$$

defined by $(W\psi)(x) = \langle C(x), \psi \rangle$, is a partial isometry. For the link between this definition, the notion of POV measures on X, and reproducing kernel Hilbert spaces we refer to Ref. 12.

The above definition does not, of course, by itself give any indication on how to construct coherent states given a Hilbert space $\mathcal{H}$. In other words, the definition does not state where to find the manifold X or how to construct the map C in (1.1). In this paper, we are interested in constructing coherent states in the case where $\mathcal{H}$ carries a unitary irreducible representation U of a Lie group G. We propose the following construction, which generalizes the Perelomov construction, outlined below. Choose a fixed regular (i.e., C^∞) vector[13] $\eta \in \mathcal{H}$ and consider $\mathcal{O}_\eta = \{U(g)\eta \mid g \in G\} \subset \mathcal{H}$, the orbit of U through η. Here $\mathcal{O}_\eta$ carries in a natural way a degenerate symplectic (i.e., presymplectic[14]) structure as follows. Consider the symplectic form

$$\omega: (\phi,\psi) \in \mathcal{H} \times \mathcal{H} \to \mathrm{Im}\langle \phi, \psi \rangle \in \mathbb{R} \tag{1.4}$$

on $\mathcal{H}$ and define ϵ_η to be the restriction of ω to $T\mathcal{O}_\eta$; it then follows from general theory that ϵ_η is a presymplectic form on $\mathcal{O}_\eta$.[15] The corresponding moment map is computed to be

$$J_\eta: U_g\eta \in \mathcal{O}_\eta \mapsto \langle U_g\eta, \Gamma(\cdot)U_g\eta \rangle \in \mathbf{g}^*, \tag{1.5}$$

where $\mathbf{g}^*$ is the dual of the Lie algebra $\mathbf{g}$ of G and Γ is the representation of $\mathbf{g}$ on $\mathcal{H}$, obtained via the Stone theorem from U. The image under J_η of $\mathcal{O}_\eta$ is an orbit P_η of the coadjoint action of G on $\mathbf{g}^*$ and as such, is naturally a symplectic homogeneous space of G.[15] Defining $H_\eta \subset G$ by $h \in H_\eta$ iff $U(h)\eta = \eta$, we have $\mathcal{O}_\eta \cong G/H_\eta$. Moreover, defining $K_\eta \subset G$ by $k \in K_\eta$ iff $\mathrm{Ad}_k^* J_\eta(\eta) = J_\eta(\eta)$, one has $P_\eta \cong G/K_\eta$. Choosing a smooth section $\beta: G/K_\eta \to G/H_\eta$, provided it exists, we construct

 0022-2488/89/071401-07$02.50

$$C_\eta:\ x\in G/K_\eta \mapsto C_\eta(x)\in\mathscr{H}, \tag{1.6a}$$

where

$$C_\eta(x)\equiv U(g)\eta, \quad \text{for } g\in\beta(x). \tag{1.6b}$$

The map C_η is well-defined since η is H_η invariant. Setting $C\equiv C_\eta$ and $X\equiv P_\eta$ in (1.1), we see that $C_\eta(P_\eta)$ is a collection of coherent states, provided (1.2) holds, where Ω is now chosen to be (a constant multiple of) the symplectic volume form on the coadjoint orbit P_η. The requirement that (1.2) must hold imposes restrictions on the choices of $\eta\in\mathscr{H}$ and the section β. Given a Lie group G and an irreducible unitary representation U of G on $\mathscr{H}$, it is not clear that such choices can indeed be made: Our main result in this paper is the identification (in Sec. III) of admissible choices of η and β in the case where the group G is of the form $G=S\circledS V$ (Definition 3.1 and Theorem 3.2).

In the Perelomov construction,[5] one replaces (1.6) by

$$\hat{C}_\eta: x\in G/H'_\eta \mapsto U(\gamma(x))\eta\in\mathscr{H}, \tag{1.7}$$

where $H'_\eta\subset G$ is defined by $h\in H'_\eta$ iff $U(h)\eta=e^{i\alpha(h)}\eta$ for some $\alpha(h)\in\mathbb{R}$ and $\gamma: G/H'_\eta\to G$ is a Borel section. Note that $H_\eta\subset H'_\eta\subset K_\eta\subset G$. Assuming that G/H'_η carries an invariant measure ν, a necessary and sufficient condition for $\hat{C}_\eta(G/H'_\eta)$ to be a collection of coherent states is that[5]

$$\int_{G/H'_\eta}\nu(x)\,|\langle\eta,U(\gamma(x))\eta\rangle|^2=\|\eta\|^2, \tag{1.8}$$

which puts restrictions on the admissible η.

Our proposal (1.6) has the following advantages over the Perelomov construction (1.7). First, the symplectic structure on G/K_η makes G/K_η a natural object to consider and guarantees the existence of an invariant volume element, which might be absent on G/H'_η. Second, since $K_\eta\supset H'_\eta$, we see that typically, G/K_η is a manifold of lower dimension than G/H'_η; as a result, a coherent state representation of a vector in $\mathscr{H}$, labeled by points in G/K_η, is more parsimonious than one labeled by points in G/H'_η. This observation is of importance in applications. Finally, and most important, there are cases, as we shall see, where the integral in (1.8) diverges for all $\eta\in\mathscr{H}$, whereas (1.6) nevertheless yields coherent states for an appropriate choice of $\eta\in\mathscr{H}$. Examples of this phenomenon are given in Sec. IV.

The paper is organized as follows. In Sec. II, we analyze those symplectic orbits of $G=S\circledS V$ that are needed in the coherent state construction. In Sec. III, we identify necessary conditions on $\eta\in\mathscr{H}$, an irreducible representation space of G, in order for (1.6) to yield a collection of coherent states. Section IV is devoted to examples. We advise the reader to read Sec. IV A in conjunction with Secs. II and III: An effort has been made to clearly identify the mathematical objects introduced in these sections in the example worked out in Sec. IV A. The results of this paper were announced in Ref. 11, where an application to quantization was also briefly discussed.

II. SYMPLECTIC ORBITS OF $G=S\circledS V$

The central result of this section is Theorem 2.1: It is crucial for the construction of the coherent states in Sec. III. Let $\mathbf{g}=\mathbf{s}\times V$ be the Lie algebra of $G=S\circledS V$, with $\mathbf{s}$ the Lie algebra of S. Here G acts on V and hence, by lifting,[16] on T^*V. Moreover, this lifted action is symplectic with respect to the natural symplectic form ω_0 on $T^*V\cong V\times V^*$.[16] The corresponding Ad*-equivariant moment map J is[15]

$$J: p_v\in T^*V\mapsto J(p_v)\in\mathbf{g}^*, \tag{2.1}$$

with

$$J(p_v)\cdot\gamma=p_v((\gamma)_V(v)), \quad \forall\gamma\in\mathbf{g}, \tag{2.2}$$

where $\mathbf{g}^*$ is the dual of the Lie algebra of G and $(\gamma)_V$ is the generator of $\gamma\in\mathbf{g}$ on V.[15] The orbits of G in T^*V are of the form $V\times\mathscr{O}^*$, with $\mathscr{O}^*$ as in Sec. I, i.e., $\mathscr{O}^*\cong S/H$ is an orbit of the action of S on V^*. Hence $V\times\mathscr{O}^*\cong G/H$. Let $v_0^*\in\mathscr{O}^*\subset V^*$ be a fixed point in $\mathscr{O}^*$ such that $Hv_0^*=v_0^*$; in the identification of $\mathscr{O}^*$ with the coset space S/H, we identify v_0^* with $eH\in S/H$ ($e\in S$, the unit element of S). The following vector sub-bundle of $V\times\mathscr{O}^*\to\mathscr{O}^*$ presents itself naturally:

$$W\equiv\{(v,v^*)\in V\times\mathscr{O}^*\,|\,w^*(v)=0, \quad \forall w^*\in T_{v^*}\mathscr{O}^*\subset V^*\}. \tag{2.3}$$

We write W_{v^*} for the fiber of W at $v^*\in\mathscr{O}^*$ and call W the *normal bundle* over $\mathscr{O}^*$. We call a vector sub-bundle Σ of $V\times\mathscr{O}^*$ *parallel* iff

$$\Sigma\oplus W\cong V\times\mathscr{O}^*, \tag{2.4}$$

where $\oplus$ denotes the Whitney sum of vector bundles. Parallel bundles always exist; it suffices to introduce a metric g on V and to choose $\Sigma_{v^*}\equiv W_{v^*}^{\perp}$, the orthogonal complement of W_{v^*} with respect to g. We now have the following theorem.

Theorem 2.1: (i) $(V\times\mathscr{O}^*,\omega_0|_{V\times\mathscr{O}^*})$ is a presymplectic submanifold of T^*V. (ii) $I: V\times\mathscr{O}^*\to T^*\mathscr{O}^*$, defined by

$$I(v,v^*)\cdot w^*=-w^*(v), \quad \forall w^*\in T_{v^*}\mathscr{O}^*\subset V^*, \tag{2.5}$$

is the symplectic reduction of $V\times\mathscr{O}^*$, i.e., I is a surjective submersion and

$$\text{Ker }\omega_0|_{V\times\mathscr{O}^*}=\text{Ker }TI, \tag{2.6}$$

$$I^*\omega=\omega_0|_{V\times\mathscr{O}^*}, \tag{2.7}$$

where ω is the canonical symplectic form on $T^*\mathscr{O}^*$. (iii) $T^*\mathscr{O}^*\cong G/K$ with $K=H\circledS W_{v_0^*}$. (iv) If $\Sigma\subset V\times\mathscr{O}^*$ is a parallel bundle, then

$$I|_\Sigma:\Sigma\to T^*\mathscr{O}^* \tag{2.8}$$

is a symplectic diffeomorphism, where Σ is equipped with the symplectic form $\omega_0|_\Sigma$.

Remarks: Before turning to the proof of Theorem 2.1, we note that it follows from (i)–(iii) that $T^*\mathscr{O}^*$ is a symplectic homogeneous space of G. Let

$$K: T^*\mathscr{O}^*\to\mathbf{g}^*=\mathbf{s}^*\times V^* \tag{2.9}$$

be defined by

$$K(p_{v^*})\cdot\xi=p_{v^*}((\xi)_{V^*}(v^*)), \quad \forall\xi\in\mathbf{s}, \tag{2.10}$$

$$K(p_{v^*})\cdot w=v^*(w), \quad \forall w\in V. \tag{2.11}$$

Then one verifies that on $V\times\mathscr{O}^*$, with J as in (2.1),

$$J=K\circ I \tag{2.12}$$

and that K is a moment map[15] for the action of G on $T^*\mathscr{O}^*$. Moreover, K is a diffeomorphism onto its image, so that $T^*\mathscr{O}^*$ is symplectomorphic with a G orbit in $\mathbf{g}^*$.

The above observation is of interest in itself in the context of quantization. If $\mathcal{O}^*$ is thought of as the configuration space of a physical system, then $T^*\mathcal{O}^*$ is its classical phase space. Theorem 2.1 identifies a group G, i.e., $G = S\circledS V$, for which $T^*\mathcal{O}^*$ is a homogeneous symplectic space. Examples include $\mathcal{O}^* = \mathbb{R}^n$, $\mathcal{O}^* = S^{(n)}$, or $\mathcal{O}^* = H^{(n)}$, the Poincaré half-space with G given, respectively, by the Weyl–Heisenberg group, the Euclidean group in $(n+1)$ dimension, and the Poincaré group $SO(n,1)\circledS\mathbb{R}^{n+1}$.

Proof of Theorem 2.1: (i) Note that $\omega_0|_{V\times\mathcal{O}^*}$ is a closed two-form on the orbit $V\times\mathcal{O}^*$; its kernel has constant dimension on $V\times\mathcal{O}^*$ since ω_0 is invariant under the action of G. (ii) We need the following lemma.[17]

Lemma 2.2: Let $i: N\to M$ be an injective immersion of a smooth manifold N into a smooth manifold M. Define

$$\bar{N} = \pi_M^{-1}(i(N)), \tag{2.13}$$

where $\pi_M: T^*M\to M$ is the natural projection. Writing ω_M for the canonical symplectic form on T^*M, we have that $(\bar{N},\omega_{\bar{N}}\equiv\omega_M|_N)$ is a presymplectic submanifold of T^*M and

$$\bar{i}:\bar{N}\to T^*N, \tag{2.14}$$

defined by $\bar{i}(p_{i(n)})\cdot v_n = p_{i(n)}(Ti\cdot v_n)$; $(v_n\in T_nN)$ is a surjective submersion satisfying

$$\bar{i}^*\omega_N = \omega_{\bar{N}}. \tag{2.15}$$

Hence T^*N is the symplectic reduction of $\bar{N}$.

Proof of Lemma 2.2: It is readily seen that $\bar{N}$ is presymplectic and $\bar{i}$ is a surjective submersion. To prove (2.15), notice that locally, on some open set, U, $i(N)$ is determined by

$$f^i(m) = 0, \tag{2.16}$$

where the f^i $(i = 1,...,k = \dim M - \dim N)$ are smooth functions defined on $U\subset M$ such that the df_i are linearly independent one-forms on U. Hence, if x^j $(j = 1,2,...,\dim N)$ are local coordinates on N, we can use (x^j, f^i) as local coordinates on M. Then, locally,

$$\omega_M \equiv dx^j\wedge dp_j + df^i\wedge d\hat{p}_i. \tag{2.17}$$

Hence,

$$\omega_{\bar{N}} = dx^j\wedge dp_j = \bar{i}^*\omega_N. \tag{2.18}$$

□

Returning to the proof of Theorem 2.1, we apply Lemma 2.2 as follows. With $i:\mathcal{O}^*\to V^*$ as the natural imbedding, we have $\bar{\mathcal{O}}^* = \mathcal{O}^*\times V$ since $T^*V^*\equiv V^*\times V$. Hence, $T^*\mathcal{O}^*$ is the symplectic reduction of $\mathcal{O}^*\times V$, regarded as a presymplectic submanifold of T^*V^*. On the other hand, T^*V^* is symplectically diffeomorphic to T^*V; explicitly,

$$\Lambda:\ (v,v^*)\in T^*V\to(v^*,-v)\in T^*V^* \tag{2.19}$$

is a symplectomorphism.

Note that $\Lambda(V\times\mathcal{O}^*) = \mathcal{O}^*\times V$. Combining (2.5), (2.14), and (2.19), we conclude that $I = \bar{i}\circ\Lambda|_{V\times\mathcal{O}^*}$, which proves (ii).

(iii) As the symplectic reduction of a presymplectic homogeneous manifold, $T^*\mathcal{O}^*$ is itself a symplectic homogeneous manifold of G. We now determine its isotropy group K. From (2.5), it follows that

$$I(v,v^*) = I(v',v^{*\prime}) \tag{2.20}$$

iff

$$v^* = v^{*\prime} \tag{2.21}$$

and

$$w^*(v) = w^*(v'),\quad \forall w^*\in T_{v^*}\mathcal{O}^*\subset V^*, \tag{2.22}$$

i.e., by (2.3),

$$v - v'\in W_{v^*}. \tag{2.23}$$

Now $V\times\mathcal{O}^*\cong G/H$, where $(0, v_0^*)\in V\times\mathcal{O}^*$ is identified with $(e,0)H(e\in S$, the unit element in $S)$ and $v_0^*\in\mathcal{O}^*$ is as in the beginning of this section. Hence, $T^*\mathcal{O}^*\cong G/K$ with $K = H\circledS W_{v_0^*}$.

(iv) This statement follows from a dimensional argument using (2.4) and (ii). □

In Sec. III, we shall need to make extensive use of the symplectic volume form Ω on the parallel bundle Σ over $\mathcal{O}^*$. Let μ denote a volume form on $\mathcal{O}^*$: We need to express the relationship between μ and Ω. To do so, let $U_{v^*}\subset\mathcal{O}^*$ be a neighborhood of a point $v^*\in\mathcal{O}^*$ and let $\{\theta^1(v^{*\prime}),...,\theta^p(v^{*\prime})\}$, $p = \dim\mathcal{O}^*$, and $v^{*\prime}\in U_{v^*}$, be a moving frame in $T\mathcal{O}^*$ such that

$$\mu(\theta^1(v^{*\prime}),...,\theta^p(v^{*\prime})) = 1. \tag{2.24}$$

Denote by $\{e_1(v^{*\prime}),...,e_p(v^{*\prime})\}$ the unique moving frame in the vector bundle $\Sigma\to\mathcal{O}^*$ [i.e., $e_i(v^{*\prime})\in\Sigma_{v^{*\prime}}\subset V$, $i = 1,...,p$] determined by

$$\theta^i(v^{*\prime})(e_j(v^{*\prime})) = \delta^i_j,\quad i,j = 1,...,p. \tag{2.25}$$

Note that in (2.25) $\theta^i(v^{*\prime})$ is regarded as an element of V^* and $e_j(v^{*\prime})$ is regarded as an element of V. Then, (2.24) implies

$$\mu = e_1\wedge e_2\wedge\cdots\wedge e_p, \tag{2.26}$$

where now $e_i(v'_*)\in\Sigma_{v^{*\prime}}\subset V = (V^*)^*$ is regarded in the natural way as a linear function on $T_{v^{*\prime}}\mathcal{O}^*\subset V^*$. Choosing a moving frame $\{e_{p+1}(v^{*\prime}),...,e_n(v^{*\prime})\}$, $v^{*\prime}\in U_{v^*}$ in the normal bundle $W\to\mathcal{O}^*$ [see (2.3)], one obtains a moving frame $\{e_1(v^{*\prime}),...,e_n(v^{*\prime})\}$, $n = \dim V$ of the trivial bundle $V\times\mathcal{O}^*\to\mathcal{O}^*$. Denote by $\{\theta^1(v^{*\prime}),..,\theta^n(v^{*\prime})\}$ the basis of V^* dual to $\{e_1(v^{*\prime}),...,e_n(v^{*\prime})\}$. Recall now that every vector $w\in V$ is in an obvious way a linear function on $T_{(v',v^{*\prime})}(T^*V)\cong V\times V^*$; namely, if $(v,v^*)\in T_{(v',v^{*\prime})}\times(V\times V^*)$, then we define, with some abuse of notation,

$$w\cdot(v,v^*) = v^*(w). \tag{2.27a}$$

Analogously, every $w^*\in V^*$ is a one-form on T^*V:

$$w^*\cdot(v,v^*) = w^*(v). \tag{2.27b}$$

With this in mind, one can write, for each $(v,v^{*\prime})\in V\times\mathcal{O}^*$,

$$\omega_0(v,v^{*\prime}) = \sum_{i=1}^{n}\theta^i(v^{*\prime})\wedge e_i(v^{*\prime}), \tag{2.28}$$

where we recall that ω_0 is the symplectic form on T^*V. A simple calculation then shows that

$$\Omega\equiv(-1)^p(1/p!)\omega_0^p\big|_\Sigma = \theta^1\wedge\cdots\wedge\theta^p\wedge e_1\cdots\wedge e_p. \tag{2.29}$$

To further rewrite expression (2.29) we introduce the functions

$$\alpha_i: v^{*\prime}\in U_{v^*}\subset\mathcal{O}^*\mapsto(v^{*\prime}-v^*)(e_i(v^*))\in\mathbb{R}, \tag{2.30}$$

$(i=1,2,...,p)$, i.e., $\alpha_i(v^{*\prime})$ is the component of $v^{*\prime}-v^*$ along $\theta^i(v^*)$. Using the fact that $\theta^1(v^*),...,\theta^p(v^*)$ span $T_{v^*}\mathcal{O}^*$ it is readily shown that by choosing U_{v^*} sufficiently small around v^*, the α_i, $i=1,2,...,p$ can be used as coordinate functions on U_{v^*}. Hence, there exists a positive function $f_{v^{*\prime}}$ on U_{v^*} such that

$$\mu_{v^{*\prime}} = f_{v^*}(v^{*\prime})\,d\alpha_1\wedge\cdots\wedge d\alpha_p. \tag{2.31}$$

Moreover, defining for $i=1,2,...,p$ the functions

$$w^i:\ (v,v^{*\prime})\in\Sigma\cap(V\times U_{v^*})\to\theta^i(v^{*\prime})(v)\in\mathbb{R}, \tag{2.32a}$$

one verifies that

$$dw^i(e_j)=\delta^i_j, \tag{2.32b}$$

where $e_j(v^{*\prime})\in\Sigma_{v^{*\prime}}\subset V$ is regarded as a tangent vector to Σ. Consequently, combining (2.32b) with (2.26), (2.29), and (2.31), one finds

$$\Omega = f_{v^*}\,dw^1\wedge\cdots\wedge dw^p\wedge d\alpha_1\wedge\cdots\wedge d\alpha_p. \tag{2.33}$$

Recall that (2.33) is valid on the neighborhood U_{v^*} chosen above and note that the function f_{v^*} depends on the choice made for the parallel bundle Σ, but not on the choice of the moving frame $\{\theta^1(v^{*\prime}),...,\theta^p(v^{*\prime})\}$ in (2.24).

Definition 2.2: Let Σ be a parallel bundle over an orbit $\mathcal{O}^*$ of S in V^* and let μ be a volume element on $\mathcal{O}^*$. We shall say $(\Sigma,\mathcal{O}^*,\mu)$ is *admissible* if the following conditions hold.

(i) There exists a neighborhood $U_{v_0^*}$ of $v_0^*\in\mathcal{O}^*$ (where v_0^* is as in the beginning of this section) which is H invariant, i.e., $HU_{v_0^*}=U_{v_0^*}$ and on which there exists a function $f_{v_0^*}$ such that (2.31) and hence, (2.33) holds.

(ii) Defining $U_{v^*}\equiv s\cdot U_{v_0^*}$, where $s\in S$ is chosen such that $s\cdot v_0^*=v^*\in\mathcal{O}^*$, there exists a function f_{v^*} on U_{v^*} such that (2.33) holds. When dealing with an admissible triple $(\Sigma,\mathcal{O}^*,\mu)$, we introduce the notation

$$f:\ \{(v^*,v^{*\prime})\in\mathcal{O}^*\times\mathcal{O}^*\,|\,v^{*\prime}\in U_{v^*}\}\mapsto f_{v^*}(v^{*\prime})\in\mathbb{R}^+. \tag{2.34}$$

In Sec. IV we shall see that in many cases $U_{v_0^*}$ and hence U_{v^*}, $\forall v^*\in\mathcal{O}^*$ can be taken equal to $\mathcal{O}^*$ itself.

III. COHERENT STATES

Up to unitary equivalence, the unitary irreducible representations of $G=S\circledS V$ are obtained, via induction, as follows.[13] Let $\mathcal{O}^*\cong S/H$ be an orbit of S in V^* and L an irreducible unitary representation of H on a Hilbert space $\mathcal{K}$. Set $\mathcal{H}=L^2(\mathcal{O}^*,\mu;\mathcal{K})$, where μ is a quasi-invariant measure on $\mathcal{O}^*$. Then,

$$(U(s,v)\psi)(v^*)=e^{iv^*(v)}({}_LU^S(s)\psi)(v^*), \tag{3.1}$$

where ${}_LU^S$ is the representation of S induced from L and $\psi\in\mathcal{H}$ defines a unitary irreducible representation of G. Moreover, all such representations are unitarily equivalent to one of the above, so that they are labeled by a pair $(\mathcal{O}^*,L)$. In the following, we always assume that the triple $(\Sigma,\mathcal{O}^*,\mu)$ is admissible (Definition 2.2) and that the family of neighborhoods $U_{v^*}, v^*\in\mathcal{O}^*$ has been chosen fixed.

Definition 3.1: A vector $\eta\in\mathcal{H}=L^2(\mathcal{O},\mu;\mathcal{K})$ is said to be a *resolution generator* (or *an admissible vector*) with respect to a parallel bundle Σ iff:

(i) η is a regular vector[13];

(ii) $H\eta=\eta$, with H as above, i.e., $\mathcal{O}^*\cong S/H$;

(iii) $J_\eta(\eta)\cdot(\xi,v)=v_0^*(v)$, $\forall\xi\in\mathbf{s}$, $\forall v\in V$, i.e., $J_\eta(\eta)=(0,v_0^*)\in\mathbf{s}^*\times V^*\cong\mathbf{g}^*$; and

(iv) the support of η is contained in $U_{v_0^*}$ and there exists a nonzero constant N such that

$$\forall\phi,\quad\psi\in\mathcal{K},\quad\forall w^*\in\mathcal{O}^*,$$

$$N\langle\phi,\psi\rangle_{\mathcal{K}} = (2\pi)^p\int_{\mathcal{O}^*}\mu(v^*)f(v^*,w^*)\langle\phi,{}_LU^S(s(v^*))\eta(w^*)\rangle_{\mathcal{K}}\times\langle{}_LU^S(s(v^*))\eta(w^*),\psi\rangle_{\mathcal{K}}, \tag{3.2}$$

with $s(v^*)\in S$ such that $s(v^*)\cdot v_0^*=v^*$.

Remark: In the special case where G is the Weyl–Heisenberg or "$ax+b$" group, the resolution generator is what is usually called the "analyzing wavelet" in the literature, as seen in Sec. IV. We can now formulate and prove the main result of this paper.

Theorem 3.2: If $\eta\in\mathcal{H}=L^2(\mathcal{O}^*,\mu;\mathcal{K})$ is a resolution generator with respect to a parallel bundle $\Sigma\subset V\times\mathcal{O}^*$, then

(i) $\tilde{C}_\eta:\ (v,v^*)\in V\times\mathcal{O}^*\mapsto U(s(v^*),v)\eta\in\mathcal{O}_\eta$, with $s(v^*)\in S$ such that $s(v^*)\cdot v_0^*=v^*$, is a presymplectic diffeomorphism; in particular,

$$\tilde{C}_\eta^*\epsilon_\eta=\omega_0\big|_{V\times\mathcal{O}^*}. \tag{3.3}$$

(ii) Define $C_\eta\equiv\tilde{C}_\eta|_\Sigma$. Then, $C_\eta(\Sigma)\subset\mathcal{H}$ is a collection of coherent states on $\mathcal{H}$; in particular, the map

$$W_\eta:\ \psi\in\mathcal{H}\mapsto W_\eta\psi\in L^2(\Sigma,N^{-1}\Omega), \tag{3.4a}$$

defined by

$$(W_\eta\psi)(v,v^*)\equiv\langle C_\eta(v,v^*)\eta,\psi\rangle, \tag{3.4b}$$

is an isometry, where Ω is the invariant symplectic volume form on Σ. [Here N is defined in Definition 3.1 (iv).] Also,

$$\langle\phi,\psi\rangle = N^{-1}\int_\Sigma\Omega(v,v^*)\langle\phi,C_\eta(v,v^*)\eta\rangle\times\langle C_\eta(v,v^*)\eta,\psi\rangle. \tag{3.4c}$$

Proof: (i) Because of Definition 3.1 (ii), $\tilde{C}_\eta$ is both well-defined and bijective. That $\tilde{C}_\eta$ preserves the presymplectic structures of $\mathcal{O}_\eta$ and $V\times\mathcal{O}^*$ follows from Definition 3.1 (iii) and (2.10)–(2.12).

(ii) Let $\phi,\psi\in\mathcal{H}=L^2(\mathcal{O}^*,\mu;\mathcal{K})$. Introducing the notation

$$I=\int_\Sigma N^{-1}\Omega(v,v^*)\langle\psi,C_\eta(v,v^*)\rangle\langle C_\eta(v,v^*),\phi\rangle, \tag{3.5}$$

we have to show that

$$\langle\psi,\phi\rangle=I. \tag{3.6}$$

First, consider

$$\langle\psi,C_\eta(v,v^*)\rangle = \langle\psi,U(v,s(v^*)\eta\rangle = \int\mu(w^*)\langle\psi(w^*),{}_LU^S(s(v^*))\eta(w^*)\rangle_{\mathcal{K}}e^{iw^*(v)}, \tag{3.7}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ denotes the inner product on $\mathcal{K}$. Now, since η is a resolution generator, Definition 3.1 (iv) guarantees

that supp $\eta \subset U_{v_0^*}$ and thus supp ${}_L U^S(s(v^*))\eta \subset U_{v^*}$. Hence, the integral in (3.7) is only over U_{v^*}. Also,

$$\langle C_\eta(v,v^*),\phi\rangle = \int_{\mathcal{O}^*} \mu(w^{*\prime})\langle {}_L U^S(s(v^*))\eta(w^{*\prime}),\phi(w^{*\prime})\rangle_{\mathscr{H}} e^{-iw^{*\prime}(v)}, \tag{3.8}$$

where, again, the integral is over U_{v^*}. We now insert (3.7) and (3.8) into (3.5) and use (2.32) to first calculate the integral over the fiber Σ_{v^*} of Σ above v^*, i.e.,

$$\int_{\Sigma_{v^*}\cong\mathbb{R}^p} dw^1\cdots dw^p \exp i(w^* - w^{*\prime})_i w^i = (2\pi)^p \delta((w^* - w^{*\prime})_1)\cdots\delta((w^* - w^{*\prime})_p), \tag{3.9}$$

where we wrote $(w^* - w^{*\prime})(v) = (w^* - w^{*\prime})_i \theta^i(v^*)(v)$ and used the definition (2.32a) of w^i. Now, recall that for w^*, $w^{*\prime}\in U_{v^*}$, if $(w^* - w^{*\prime})_i = 0$, $\forall i = 1,\ldots,p$, then $w^* = w^{*\prime}$. Hence, using (3.9), we find

$$I = \frac{(2\pi)^p}{N}\int \mu(v^*)\mu(w^*) f(v^*,w^*) \times\langle\psi(w^*),{}_L U^S(s(v^*))\eta(w^*)\rangle_{\mathscr{H}} \times\langle {}_L U^S(s(v^*))\eta(w^*),\phi(w^*)\rangle_{\mathscr{H}}.$$

Using Definition 3.1 (iv), (3.6) follows. □

Remarks: (i) As a result of Definition 3.1 (ii), we have $H_\eta = H$, so $V\times\mathcal{O}^*\cong G/H_\eta$. Moreover, Theorem 3.2(i) guarantees that the following diagram is commutative:

$$\begin{array}{ccc} & \overset{C_\eta}{\longrightarrow} & \\ V\times\mathcal{O}^*\cong G/H_\eta & & \mathcal{O}_\eta\cong G/H_\eta \\ J\searrow & & \swarrow J_\eta \\ & P^\eta\cong G/K_\eta & \end{array}$$

Here, J is defined in (2.1) and J_η is defined in (1.5). As a result, Σ is a global section of $G/H_\eta \to G/K_\eta$ and, comparing to (1.6), we see that different choices of Σ correspond to different choices of the section β.

(ii) Condition (i) in Definition 3.1 is needed to guarantee that condition (iii) makes sense, i.e., that η is in the domain of the generators of the representation U of G on $\mathscr{H}$. If one is not interested in Theorem 3.2(i), then conditions (i) and (iii) of Definition 3.1 can be omitted and Theorem 3.2(ii) can be proven directly from (ii) and (iv) of Definition 3.1.

IV. EXAMPLES

A. Coherent states for the Euclidean group $E(n)$

We first illustrate the construction of Sec. III on the Euclidean group $E(n) = \mathrm{SO}(n)\circledS\mathbb{R}^n$: As we shall see, this is an example where the Perelomov construction fails and the full machinery developed here is needed. We shall do the calculations explicitly in the case of $E(2)$, indicating at the end the changes needed for the general case.

We shall write $\theta\in\mathrm{SO}(2)\equiv S$, $v = (v_1,v_2)\in\mathbb{R}^2\equiv V$, and $x = (x_1,x_2)\in\mathbb{R}^{2*}\equiv V^*$. The action of SO(2) on $\mathbb{R}^2$ is the usual one, i.e.,

$$\theta\cdot v\equiv(\cos\theta v_1 - \sin\theta v_2,\quad \sin\theta v_1 + \cos\theta v_2). \tag{4.1}$$

Hence, the action on V^* is

$$\theta\cdot x\equiv(\cos\theta x_1 - \sin\theta x_2,\quad \sin\theta v_1 + \cos\theta v_2), \tag{4.2}$$

so that the orbits $\mathcal{O}^*$ of S in V^* are circles. We shall choose $\mathcal{O}^*$ to be the unit circle S^1,

$$x_1^2 + x_2^2 = 1, \tag{4.3}$$

in V^*. There is now only one unitary irreducible representation of $E(2)$ associated with $\mathcal{O}^*$ since $H = \{e\}$: Its carrier space is $L^2(S^1,d\alpha)$ and, in accordance with (3.1), is given by

$$(U(\theta,v_1,v_2)\psi)(\alpha) = \exp i(v_1\cos\alpha + v_2\sin\alpha)\psi(\alpha - \theta), \tag{4.4}$$

where we introduce the obvious angular coordinate α on S^1, i.e., $x_1 = \cos\alpha$, $x_2 = \sin\alpha$. The last ingredient needed in order to be able to identify admissible vectors in $L^2(S^1,d\alpha)$ is a parallel bundle Σ in $V\times\mathcal{O}^*$.

Although many choices of Σ are in principle possible, the following imposes itself naturally. First, one verifies readily, using the definition in (2.3), that the normal bundle W is given in this case by

$$W\equiv\{(v,x)\in\mathbb{R}^2\times S^1 \mid \exists a\in\mathbb{R} \text{ such that } v = ax\}. \tag{4.5}$$

We then choose for Σ the bundle for which the fibers Σ_x, $x\in S^1$ are the orthogonal complement of W_x with respect to the usual inner product on $V = \mathbb{R}^2$:

$$\Sigma\equiv\{(v,x)\in\mathbb{R}^2\times S^1 \mid v_1x_1 + v_2x_2 = 0\}. \tag{4.6}$$

If we identify V and V^* using the inner product, we can represent the situation as shown in Fig. 1.

Note that Σ is invariant under the action of SO(2) on $\mathbb{R}^2\times\mathbb{R}^{2*}$. The results of Theorem 2.1 can now quite easily be verified by direct calculation using, for example, the Dirac theory of constraints.[18] In particular, since $H = \{e\}$, we find $K = W_{x_0}$, where $x_0\equiv(1,0)\in S^1$. Although it would suffice for the verification of condition (iv) in Definition 3.1 to find the function $f(x,x')$ in (2.34), we shall explicitly compute some of the other objects introduced in Sec. II in the present case. First, we introduce coordinates on Σ by

$$(a,\alpha)\in\mathbb{R}\times S^1 \to (v_1(a,\alpha),v_2(a,\alpha),\alpha)\in\Sigma\subset\mathbb{R}^2\times S^1, \tag{4.7}$$

with

$$v_1(a,\alpha) = -a\sin\alpha, \tag{4.8a}$$

$$v_2(a,\alpha) = a\cos\alpha, \tag{4.8b}$$

so that

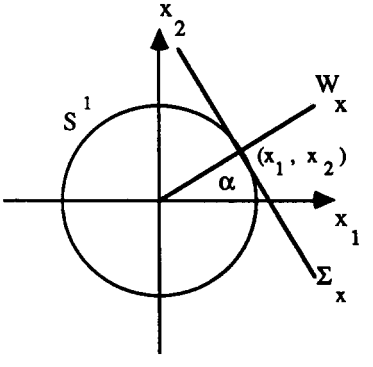


FIG. 1. The choice of $\mathcal{O}^*$, Σ, and W for $E(2)$.

$$\omega_0|_\Sigma = dv_1 \wedge dx_1 + dv_2 \wedge dx_2$$
$$= da \wedge d\alpha, \tag{4.9}$$

which is indeed seen to be a symplectic form on Σ. Second,

$$\theta^1(\alpha) = \frac{\partial}{\partial\alpha} \in T_\alpha S^1, \tag{4.10a}$$

$$e_1(\alpha) = (-\sin\alpha, \cos\alpha) \in \mathbb{R}^2 \tag{4.10b}$$

satisfy (2.34) (with $\mu = d\alpha$) and (2.25). Third, let $U_\alpha \subset S^1$ be the open half-circle $(\alpha - \pi, \alpha + \pi) \subset S^1$; then, following (2.30), we define

$$\alpha_1 : \alpha' \in U_\alpha \subset S^1 \mapsto \sin(\alpha' - \alpha) \in \mathbb{R}, \tag{4.11}$$

so that

$$d\alpha' = [1/\cos(\alpha' - \alpha)]\, d\alpha_1. \tag{4.12}$$

It then follows from (2.31) that

$$f(\alpha,\alpha') = [\cos(\alpha' - \alpha)]^{-1}, \tag{4.13}$$

where f is defined provided that

$$|\alpha - \alpha'| < \pi (\mathrm{mod}\, 2\pi). \tag{4.14}$$

Finally, we can verify that (2.33) indeed holds by first computing

$$w^1 : (a,\alpha') \in \Sigma \mapsto a = \theta^1(\alpha')(-a\sin\alpha, a\cos\alpha) \tag{4.15}$$

from (2.32a) and then inserting (4.13) into (2.33) and comparing to (4.9) using (4.12). We conclude that the triple $(\Sigma, S^1, d\alpha)$ is indeed admissible as defined in Definition 2.2.

We now have all the ingredients to check which vectors $\eta \in L^2(S^1, d\alpha)$, if any, are admissible with respect to Σ as defined in Definition 3.1: For (i), it suffices that η is in $C^\infty(S^1)$. Condition (ii) in Definition 3.1 is empty in the present case, whereas (iii) becomes

$$\int_{S^1} \left(\eta(\alpha) \frac{1}{i} \frac{\partial}{\partial\alpha} \eta(\alpha)\right) d\alpha = 0, \tag{4.16a}$$

$$\int_{S^1} \sin\alpha |\eta(\alpha)|^2\, d\alpha = 0, \tag{4.16b}$$

$$\int_{S^1} \cos\alpha |\eta(\alpha)|^2\, d\alpha = 1, \tag{4.16c}$$

where we used (1.5) and (4.4). For condition (iv) of Definition 3.1 we have the double requirement that η is supported in the half-circle $\alpha \in (-\pi/2, \pi/2)$ and that there must exist an N such that

$$N = (2\pi) \int_{\alpha' - \pi/2}^{\alpha' + \pi/2} d\alpha [\cos(\alpha' - \alpha)]^{-1} |\eta(\alpha' - \alpha)|^2,$$
$$\forall \alpha' \in S^1; \tag{4.17a}$$

this is readily rewritten as

$$\int_{-\pi/2}^{\pi/2} d\alpha (\cos\alpha)^{-1} |\eta(\alpha)|^2 < \infty. \tag{4.17b}$$

Hence, condition (iv) of Definition 3.1 in this case only requires the above support property of η and convergence of the integral in (4.17b): This will be assured provided η decays fast enough to zero at the edges of its support. We conclude that any regular $\eta \in L^2(S^1, d\alpha)$ with support in $(-\pi/2, \pi/2)$ and satisfying (4.16) and (4.17b) is admissible. In view of the second remark following Theorem 3.2, it is in fact only condition (4.17b) that is needed if one is only interested in Theorem 3.2(ii). At any rate, (4.16c) is easily satisfied by appropriately normalizing η and (4.16b) will follow if η is chosen symmetric about $\alpha = 0$; for (4.16a), we remark that if we are given $\tilde\eta \in L^2(S^1, d\alpha)$, satisfying (4.16b) and (4.16c), we can always find a $v_2 \in \mathbb{R}$ such that

$$\eta \equiv e^{iv_2 \sin\alpha} \tilde\eta(\alpha) \tag{4.18}$$

satisfies (4.16a)–(4.16c). Moreover, this will not affect the coherent states obtained since $C_{\tilde\eta}|_\Sigma = C_\eta|_\Sigma$. We conclude that $L^2(S^1, d\alpha)$ contains a large number of resolution generators; the map W_η in (3.4) is explicitly given by

$$(W_\eta \psi)(a,\alpha)$$
$$= \int_{S^1} \exp - ia \sin(\alpha' - \alpha) \tilde\eta(\alpha' - \alpha) \psi(\alpha')\, d\alpha'. \tag{4.19}$$

We note that for the representation (4.4) of $E(2)$, the Perelomov construction does not work since it would require integrating over all of $V \times \mathcal{O}^* \cong \mathbb{R}^2 \times S^1$, rather than over Σ alone, which would lead to a divergent integral. Finally, we remark that the irreducibility of (4.4) implies that the von Neumann algebra generated by the generators

$$P_1 = \cos\alpha, \tag{4.20a}$$

$$P_2 = \sin\alpha, \tag{4.20b}$$

$$L = -\frac{1}{i} \frac{\partial}{\partial\alpha}, \tag{4.20c}$$

of the representation (4.4) is $B(L^2(S^1, d\alpha))$; however, the same is already true for the von Neumann algebra generated by L and P_1 or L and P_2. This observation clarifies why it is sufficient to integrate over $(a,\alpha) \in \Sigma$, rather than over the whole group, to obtain a resolution of the identity as in (3.4c).

In the case of $E(n)$, the orbits $\mathcal{O}^*$ are spheres $S^{(n-1)}$ and H becomes $SO(n-1)$. The main differences with the case of $E(2)$ are then that representations of H other than the trivial one can be used to define the unitary irreducible representations of $E(n)$ and that $\eta \in \mathcal{H}$ has to be chosen $SO(n-1)$ invariant.

B. Coherent states for the Weyl–Heisenberg, "$ax+b$," and Galilei and Poincaré groups

We briefly identify the ingredients needed for our construction of coherent states for each of the Weyl–Heisenberg, "$ax + b$," and Galilei and Poincaré groups. For the Weyl–Heisenberg group, we have $S \cong \mathbb{R}^n$, $V = \mathbb{R}^{n+1}$ and $\mathbf{b} \in S$ acts on $v = (\mathbf{v}, v_{n+1}) \in V$ via

$$(\mathbf{b} \cdot v) \equiv \begin{pmatrix} I & 0 \\ -\mathbf{b} & 1 \end{pmatrix} v. \tag{4.21}$$

A typical orbit of S in V^* is given by $x^{n+1} = 1$, where we wrote $x = (\mathbf{x}, x^{n+1}) \in V^* \cong \mathbb{R}^n$. An obvious choice for Σ is

$$\Sigma \equiv \{(x,v) \in V^* \times V \mid x^{n+1} = 1, \quad v_{n+1} = 0\}. \tag{4.22}$$

The machinery of Sec. III now leads to the standard coherent states of the Weyl–Heisenberg group.[1,5] For the "$ax + b$" group, $S = \mathbb{R}_+$ and $V = \mathbb{R}$, $\Sigma = \mathbb{R} \times \mathbb{R}_+$. Again, the known coherent states are recovered; in particular, condition (iv) in Definition 3.1 now reads as

$$\int_{\mathbb{R}^+} dx\, x^{-1}|\eta(x)|^2\, dx < \infty, \tag{4.23}$$

which is indeed the standard condition.[1,6] Similarly, the coherent states for the Galilei and Poincaré groups considered in Refs. 7 and 8 can be obtained from our construction. By making a choice for Σ other than the one used implicitly in Refs. 7 and 8, in close analogy with the choice of Σ for $E(n)$ in Sec. IV A, we obtain new sets of coherent states for the Poincaré group as follows. Let $G = \mathrm{SO}(n,1) \circledcirc \mathbb{R}^{n+1}$, so that $S = \mathrm{SO}(n,1)$ and $V = \mathbb{R}^{n+1}$. Write $v = (v^i, v_0) \in V$ and $x = (x_i, x_0) \in V^*$. Then,

$$\Sigma = \{(v,x) \in V \times V^* | x_0 > 0,\ x \cdot x = -1,\ x \cdot v = 0\}, \tag{4.24}$$

where the center dot indicates the Lorentzian inner product on $\mathbb{R}^{n+1}$. The analogy with (4.6) is obvious. We recall that the mass hyperboloid $x \cdot x = -1$ is isometric to the Poincaré half-space H^n; we expect that the coherent states built on Σ in (4.24) will prove useful in the study of quantization problems of a system with H^n as configuration space.

As a result of the success of the wavelet transform in signal analysis,[3,4] the interest in coherent states for new groups is growing. We cite Ref. 19 as an example, where the group $(\mathrm{SO}(n) \times \mathbb{R}^+) \circledcirc \mathbb{R}^n$ is considered. Here $\mathbb{R}^+$ acts on $\mathbb{R}^n$ by dilation. Again, the construction proposed in this paper applies to this group.

ACKNOWLEDGMENTS

The author gratefully acknowledges stimulating and helpful conversations on the subject matter of this paper with S. T. Ali and N. Wildberger.

This work was supported in part by NSERC Grant Nos. A5206 and A7701.

[1] For a recent review, see J. R. Klauder and B. S. Skagerstam, *Coherent States—Applications in Physics and Mathematical Physics* (World Scientific, Singapore, 1985).
[2] A. Perelomov, *Generalized Coherent States and their Applications* (Springer, Berlin, 1986).
[3] I. Daubechies and T. Paul, "Wavelets and applications," preprint (1986).
[4] J. Morlet, G. Arens, I. Fourgeau, and D. Giard, Geophysics **47**, 203 (1982); A. Grossmann and J. Morlet, SIAM J. Math. Anal. **15**, 723 (1984); I. Daubechies, *Time-frequency Localization Operators, a Geometric Phase Space Approach*, preprint (1986).
[5] A. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[6] E. W. Aslaksen and J. R. Klauder, J. Math. Phys. **10**, 2267 (1969).
[7] S. T. Ali and E. P. Prugovecki, Acta Appl. Math. **6**, 19, 47 (1986).
[8] S. T. Ali and J. P. Antoine, *Coherent States of the 1 + 1-Dimensional Poincaré Group: Square Integrability and a Relativistic Weyl Transform*, preprint, Université Catholique de Louvain 1PT-87-39.
[9] A. Unterberger, Contemp. Math. **27**, 237 (1984).
[10] S. T. Ali and G. G. Emch, J. Math. Phys. **27**, 2936 (1986).
[11] S. T. Ali and S. DeBièvre, in *Proceedings of the XVIth International Colloquium on Group-Theoretical Methods in Physics*, edited by H. D. Doebner and T. D. Palev (Varna, Bulgaria, 1987).
[12] S. T. Ali, in Proceedings of a Workshop (University of Sherbrooke, July 1986) Suppl. Rend. Circ. Mat. Palermo Ser. II **17**, 12–46 (1987).
[13] A. O. Barut and R. Raczka, *Theory of Group Representations and Applications* (Polish Scientific, Warsaw, 1977).
[14] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin Cummings, London, 1978), Secs. 3.2 and 3.3; N. Woodhouse, *Geometric Quantization* (Clarendon, Oxford, 1980), Chaps. I and II; C. Günther, *Präsymplektische Geometrie und Quasikanonische Systeme*, master's thesis, Berlin, 1974.
[15] R. Abraham and J. E. Marsden, Ref. 14, Chap. 4.
[16] R. Abraham and J. E. Marsden, Ref. 14, Chap. 3.
[17] S. DeBièvre, "Scattering in Relativistic Particle Mechanics," Ph.D. thesis, Dept. of Physics, Univ. of Rochester, Rochester, New York, 1986.
[18] The literature on this subject is huge. For a concise and geometric introduction, we refer to J. Sniatycki, Ann. Inst. H. Poincaré A **20**, 365 (1974).
[19] R. Murenzi, *Proceedings of the Conference "Ondelettes"; Méthodes Temps-Fréquence* (Marseille, 1987), to appear.

Conformal branching rules from Kač–Moody automorphisms

Mark A. Walton
Stanford Linear Accelerator Center, Stanford University, Stanford, California 94309

(Received 2 December 1988; accepted for publication 22 March 1989)

The branching rules for the conformal subalgebra $\widehat{\mathrm{SU}}(2)^N \times \widehat{\mathrm{SO}}(N)^4 \subset \widehat{\mathrm{Sp}}(2N)^1$ are calculated. The method used relies on the outer automorphisms of affine Kač–Moody algebras, and was first applied by the author to the conformal subalgebra $\widehat{\mathrm{SU}}(p)^q \times \widehat{\mathrm{SU}}(q)^p \subset \widehat{\mathrm{SU}}(pq)^1$. The results presented here demonstrate the general applicability of the method.

I. INTRODUCTION

Affine Kač–Moody algebras[1] have made a mark in theoretical physics (for a review see Ref. 2). They are realized in many two-dimensional conformal field theories. These theories describe both the critical points of second-order phase transitions and the basic building blocks of classical string theories. In fact, a string theory can only have a local spacetime gauge symmetry if there is a corresponding Kač–Moody algebra realized in the conformal field theory on its world sheet.

The subalgebras of affine Kač–Moody algebras are therefore important. Even a small subclass of affine subalgebras, the so-called conformal subalgebras, are remarkably useful.[3–8] Conformal subalgebras are subalgebras having central charge equal to that of the algebra in which they are embedded. Lists of these subalgebras have been compiled,[9] but there is no universally applicable method for calculating their branching rules. Hence all of the conformal branching rules have not been worked out.

In this paper we show that there does exist a quite general procedure for calculating conformal branching rules. It makes use of the outer automorphisms of affine Kač–Moody algebras. The method was first applied to the subalgebra $\widehat{\mathrm{SU}}(p) \times \widehat{\mathrm{SU}}(q) \subset \widehat{\mathrm{SU}}(pq)$ in Ref. 8. Here we calculate the branching rules for $\widehat{\mathrm{SU}}(2) \times \widehat{\mathrm{SO}}(N) \subset \widehat{\mathrm{Sp}}(2N)$, demonstrating the general applicability of the outer automorphism method.

The layout of this paper is as follows. Section II contains a short review of affine Kač–Moody algebras and conformal subalgebras (serving mainly to establish notation) and a general description of the outer automorphism method. Section III contains the explicit calculation of the branching rules for the conformal embedding $\widehat{\mathrm{SU}}(2) \times \widehat{\mathrm{SO}}(N) \subset \mathrm{Sp}(\widehat{2N})$; Sec. III A treats N odd and Sec. III B treats N even. Finally, Sec. IV is a short conclusion.

II. REVIEW AND NOTATION

Let $\hat{g}$ denote the affine Kač–Moody algebra that is the central extension of the loop algebra of the finite-dimensional Lie algebra $\bar{g}$. (In general, we use the convention that carets and square brackets denote objects associated with affine algebras, and bars and parentheses their finite algebra counterparts.) The algebra $\hat{g}$ is

$$[J^a_m, J^b_n] = f^{abc} J^c_{m+n} + k\delta^{ab}\delta_{m+n,0} \cdot \tag{2.1}$$

We will often include the value of $k \in Z$ by writing $\hat{g} = \hat{g}^k$. Setting the integral indices m and n in (2.1) to zero reduces this algebra to the finite one $\bar{g} \subset \hat{g}$.

The Sugawara construction,

$$L_m = \frac{1}{2(k+h^\vee)} \sum_{n \in Z} K_{ab} : J^a_{m+n} J^b_{-n} : - \delta_{m0} \frac{c(g)}{24}, \tag{2.2}$$

associates with $\hat{g}$, the Virasoro algebra,

$$[L_m, L_n] = (m-n)L_{m+n} + [c(g)/12](m-1)m(m+1)\delta_{m+n,0}, \tag{2.3}$$

with central charge

$$c(g) = k \cdot \dim \bar{g}/(k+h^\vee) . \tag{2.4}$$

Here $K_{ab}, h^\vee$ are the Killing form and dual Coxeter number of $\bar{g}$, and the normal ordering is defined in the usual way.

The Cartan subalgebra $\hat{h}$ of $\hat{g}$ contains the Cartan subalgebra $\bar{h}$ of $\bar{g}$, with elements h_i $(i = 1,...,r)$ plus the extra element h_0. We denote the elements of $\hat{h}$ by h_μ $(\mu = 0,1,...,r)$. Dual to these elements, living in the weight space $\hat{h}^*$, are the fundamental weights ω^μ:

$$\omega^\mu(h_\nu) = \delta^\mu_\nu . \tag{2.5}$$

Associated with each h_μ is a coroot $\alpha^\vee_\mu$, also living in the weight space $\hat{h}^*$, so that we have

$$\omega^\mu \cdot \alpha^\vee_\nu = \omega^\mu(h_\nu) = \delta^\mu_\nu . \tag{2.6}$$

The dot product is determined from the Cartan matrix A by the definition of its elements:

$$A_{\mu\nu} = \alpha_\mu \cdot \alpha^\vee_\nu , \tag{2.7}$$

where a root α_μ and its coroot $\alpha^\vee_\mu$ are multiples of each other,

$$\alpha^\vee_\mu = 2\alpha_\mu/\alpha_\mu \cdot \alpha_\mu .$$

There is an extra operator that commutes with the h_μ; it is L_0 of (2.3). The Cartan subalgebra can be extended to $\hat{h}^e$, having elements h_μ and $d = -L_0$. We denote the weight dual to d by δ,

$$\delta(d) = \delta(-L_0) = 1 \tag{2.8}$$

and the coroot, corresponding to d, by Λ_0,

$$\delta \cdot \Lambda_0 = \delta(d) = 1 . \tag{2.9}$$

With the usual conventions,

$$\Lambda_0(h_i) = \delta_{i0}, \quad i = 1,...,r,$$

$$\Lambda_0(d) = \Lambda_0(-L_0) = \Lambda_0 \cdot \Lambda_0 = 0 ,$$

 0022-2488/89/071408-07$02.50

the scalar product on the extended weight space $\hat{h}^{*e}$ is Minkowskian. Let

$$\bar{\psi} = k^i \bar{\alpha}_i = k^{\vee i} \bar{\alpha}_i^{\vee} , \tag{2.10}$$

be the highest root of $\bar{g}$. The $k^{\mu}, k^{\vee\mu}$ are known as marks and comarks, respectively, with $k^0 \equiv k^{\vee 0} \equiv 1$. Then if $B, C \in \hat{h}^{*e}$, we write

$$\begin{aligned} B &= \beta_\mu \omega^\mu + \beta_\delta \delta = [\beta_i \omega^i = \bar{B}, k[B], \beta_\delta] , \\ C &= \gamma_\mu \omega^\mu + \gamma_\delta \delta = [\gamma_i \omega^i = \bar{C}, k[C], \gamma_\delta] , \end{aligned} \tag{2.11}$$

where $k[B], k[C]$ are the levels of B and C, respectively,

$$k[B] = \beta_\mu k^{\vee\mu}, \quad k[C] = \gamma_\mu k^{\vee\mu} . \tag{2.12}$$

Then the dot product on the expanded weight space $\hat{h}^{*e}$ takes the form

$$B \cdot C = \bar{B} \cdot \bar{C} + k[B]\gamma_\delta + \beta_\delta k[C] . \tag{2.13}$$

With the above notation, the simple roots and fundamental weights of $\hat{g}$ can be written as

$$\begin{aligned} \alpha_i &= [\bar{\alpha}_i, 0, 0], \quad \alpha_0 = [-\bar{\psi}, 0, 1] , \\ \omega^i &= [\bar{\omega}^i, k^{\vee i}, 0], \quad \omega^0 = [0,1,0] . \end{aligned}$$

So the Dynkin diagram of $\hat{g}$ is the extended Dynkin diagram of $\bar{g}$. Outer automorphisms act as symmetries of the Dynkin diagram of $\hat{g}$. The Dynkin diagrams of $\widehat{\mathrm{Sp}}(2N)$, $\widehat{\mathrm{SU}}(2)$, and $\widehat{\mathrm{SO}}(N)$ are shown in Fig. 1; their outer automorphisms will be discussed later.

The highest weight representations of $\hat{g}$ are generated from a "vacuum" state $|M\rangle$, labeled by a dominant weight,

$$M = M_\mu \omega^\mu = [\bar{M}, M_\mu k^{\vee\mu}, 0] . \tag{2.14}$$

The vacuum satisfies

$$\begin{aligned} J_n^a |M\rangle &= 0, \quad n > 0 , \\ J_0^a |M\rangle &= T_{\bar{M}}^a |M\rangle , \end{aligned}$$

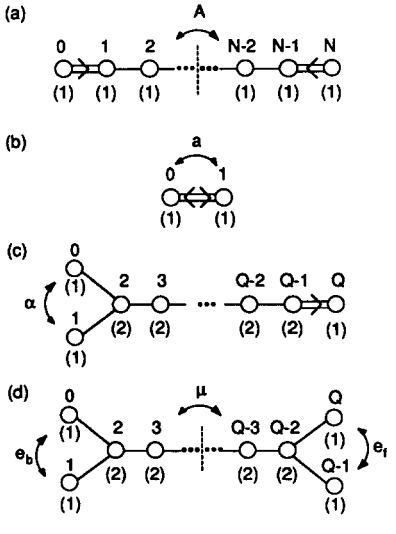


FIG. 1. The Dynkin diagrams of (a) $\widehat{\mathrm{Sp}}(2N) = \hat{C}_N$, (b) $\mathrm{SU}(2) = \hat{A}_1$, (c) $\widehat{\mathrm{SO}}(2Q+1) = \hat{B}_Q$, and (d) $\mathrm{SO}(2Q) = \hat{D}_Q$ are shown. The nodes are labeled above and the corresponding comark $k^{\vee\mu}$ is written below in brackets. Depicted are symmetries of Dynkin diagrams. As explained in the text, these are either outer automorphisms themselves or can be used to define them.

where $T_{\bar{M}}^a$ are the matrices representing the generators of $\bar{g}$ in the finite-dimensional representation labeled by $\bar{M}$. We will use the notation $M = [M_0 M_1 \cdots M_r] \equiv [M]$ to denote the weight M and corresponding highest weight representation. Similarly, representations of $\bar{g}$ will be denoted by $\bar{M} = (M_1 M_2 \cdots M_r) = (M)$. For unitary highest weight representations $[M]$ of (2.1), we must require

$$k[M] = M_\mu k^{\vee\mu} = k . \tag{2.15}$$

The states in $[M]$ having a fixed eigenvalue of L_0 fall into a finite sum of irreducible representations of $\bar{g}$. For example, the lowest value of L_0 is

$$L_0 = h[M] - \frac{c(g)}{24} = \frac{\bar{M} \cdot (\bar{M} + 2\bar{\rho})}{2(k + h^{\vee})} - \frac{c(g)}{24} , \tag{2.16}$$

where $\bar{\rho}$ is the half-sum of positive roots of $\bar{g}$. The states of $[M]$ with this value of L_0 fill out the representation $\bar{M} = (M)$.

The character of a highest weight representation $M = [M]$ is defined by

$$ch_M(\tau, z) = tr_{[M]} e^{2\pi i(\tau L_0 + z \cdot \bar{h})} . \tag{2.17}$$

The characters at $z = 0$,

$$\chi[M] = \chi[M_0 M_1, \ldots, M_r] = ch_M(\tau, 0) , \tag{2.18}$$

are called specialized characters. The modular transformation properties of the characters were found in Ref. 10. Transforming by $S: \tau \to -1/\tau$ reveals the asymptotic behavior of the characters:

$$\chi[M] \sim l[M] e^{2\pi i c(g)/24\tau} \quad \text{as } \tau \to 0. \tag{2.19}$$

The $l[M]$ can be calculated entirely from objects relevant to the finite algebra $\bar{g}$:

$$l[M] = \left| \frac{\bar{P}}{\bar{Q}} \right|^{1/2} [k[M] + h^{\vee}]^{-r/2} \times \prod_{\alpha \in \bar{\Delta}_+} 2 \sin\left[\frac{\pi(\bar{M} + \bar{\rho}) \cdot \alpha}{k[M] + h^{\vee}} \right] . \tag{2.20}$$

Here $\bar{\Delta}_+$ denotes the set of positive roots, $\bar{M} = (M_1 \cdots M_r)$, and $\bar{P}$ and $\bar{Q}$ are, respectively, the weight and root lattices of $\bar{g}$.

Knowledge of affine subalgebras of affine algebras is nowhere near as extensive as that of finite subalgebras. However, each subalgebra $\bar{j}$ of a finite Lie algebra $\bar{g}$ induces a subalgebra $\hat{j}$ of $\hat{g}$ (see, for example, Ref. 3). One identifying feature of a finite subalgebra $\bar{j} \subset \bar{g}$ is the index of embedding σ, equal to the ratio of the length squared of the highest root of $\bar{g}$ to that of $\bar{j}$ embedded in $\bar{g}$. The affine subalgebra induced by $\bar{j}^{\sigma} \subset \bar{g}$ (obvious notation) is $\hat{j}^{\sigma k} \subset \hat{g}^k$.

All other information concerning a finite subalgebra $\bar{j}^{\sigma} \subset \bar{g}$ is contained in the so-called projection matrix $\bar{F}$.[11] Here $\bar{F}$ is a (rank $\bar{j} = \bar{r}$) $\times$ (rank $\bar{g} = r$) matrix, relating weights of $\bar{g}$ to the weights of $\bar{j}$ onto which they are projected. If $(M) = (M_1 M_2 \cdots M_r)$ is a weight of $\bar{g}$, it is projected onto the weight $(M)\bar{F}^T$.

We can define an affine projection matrix $\hat{F}$, containing $\bar{F}$, so that an affine weight $[M] = [M_0 M_1 \cdots M_r]$ is projected onto the weight $[M]\hat{F}^T$. For our purposes, we can assume that $\bar{j}$ is semisimple, with f simple terms, $\bar{j} = \Sigma_{i=1}^f \bar{j}_i$, the terms having embedding indices σ_i. Then $\hat{F}$ will be a

$(\tilde{r}+f)\times(r+1)$ matrix. The extra column is determined by the projection of the weight $[100\cdots0]=\omega^0$ of $\hat{g}^1$. Clearly, ω^0 is projected onto $\sum_{i=1}^{f}\sigma_i\omega^0_{(i)}$, where $\omega^0_{(i)}$ is the zeroth fundamental weight of $\hat{j}_i$. The extra rows are easily determined by the values σ_i.

A special class of affine subalgebras is induced by $\bar{j}^\sigma\subset\bar{g}$ when $c(j)=c(g)$. These are the so-called conformal subalgebras and we denote them by $\hat{j}^{\sigma k}\triangleleft\hat{g}^k$. Now this situation is only possible for $k=1$,[9] so without loss of generality, we write $\hat{j}^\sigma\triangleleft\hat{g}^1$. The name conformal subalgebra is appropriate because the Sugawara stress tensors of $\hat{j}^\sigma$ and $\hat{g}^1$ are equivalent.[2] Complete lists of conformal subalgebras have been compiled.[9]

Consider a conformal subalgebra, $\hat{j}^k\triangleleft\hat{g}^1$. Suppose $[M]$ is a highest weight representation of $\hat{g}$ satisfying $k[M]=1$, i.e., it is a level-one representation. Then the branching rule for $[M]$ takes the form

$$[M]\rightarrow\sum_{k[m]=k}N_m[m]\,, \tag{2.21}$$

where $0\leqslant N_m\in Z$ and the sum is a finite one,[6] over all highest weight representations of $\hat{j}^k$ satisfying $k[m]=k$. Furthermore, there is a branching rule for each level-one representation of $\hat{g}$.

Since the Sugawara stress tensors for $\hat{g}^1$ and $\hat{j}^k$ are equivalent, so are their lowest moments L_0. Each state in $[M]$ must be represented by a state in one of the $[m]$ for which $N_m\neq0$, and having the same eigenvalue of L_0. Every state in $[M]$ has an L_0 eigenvalue equal to $h[M]-c/24$ of (2.16) mod an integer; and similarly for $[m]$. Therefore every $[m]$ for which $N_m\neq0$ in (2.21) must satisfy the level matching condition:

$$0\leqslant h[m]-h[M]\in Z\,. \tag{2.22}$$

Clearly, (2.21) implies

$$\chi[M]=\sum_{k[m]=k}N_m\chi[m]\,. \tag{2.23}$$

By (2.21), as $\tau\rightarrow0$, this yields the asymptotic constraint

$$l[M]=\sum_{k[m]=k}N_m l[m]\,. \tag{2.24}$$

For many conformal subalgebras, asymptotics and level matching are sufficient to determine the positive integers N_m and therefore the branching rules (2.21).[5,12] But there are many others for which this is not true.

Note that the outer automorphisms of $\hat{g}$ map level-one representations into each other. One can hope to obtain from the branching rule of one representation of $\hat{g}^1$ into $\hat{j}^k$ the branching rule for another level-one representation. This will be possible when there is an image of the outer automorphism in the outer automorphisms of the subalgebra $\hat{j}$. In that case there exists a projection matrix $\hat{F}$ manifesting the relation between the algebra and subalgebra automorphism.

Pieces of each branching rule can be computed simply from finite Lie algebra theory. The states with the lowest eigenvalue of L_0 in the level-one representation $[M]$ of $\hat{g}^1$ fill out the representation (M) of $\bar{g}$. Here (M) branches into several representations (m) of the subalgebra $\bar{j}$:

$$(M)\rightarrow\sum_m\bar{N}_m(m)\,. \tag{2.25}$$

These latter retain the lowest eigenvalue of L_0, so they must fill out the lowest L_0 level of highest weight representations of $\hat{j}$. The finite branching rules (2.25) provide part of the full affine branching rule (2.21).

The outer automorphisms may be sufficient to generate the full affine branching rule from the parts of them just mentioned. This is the case for the infinite series of conformal subalgebras $\widehat{SU}(p)\times\widehat{SU}(q)\triangleleft\widehat{SU}(pq)$.[8] In the next section we show that this conformal subalgebra is not special, by calculating in the same manner the branching rules for $\widehat{SU}(2)^N\times\widehat{SO}(N)^4\triangleleft\widehat{Sp}(2N)$.[1]

III. BRANCHING RULES FOR $\widehat{Sp}(2N)\triangleright\widehat{SU}(2)^N\times\widehat{SO}(N)^4$

The finite subalgebra

$$Sp(2N)\supset SU(2)^N\times SO(N)^4 \tag{3.1}$$

is defined by the branching rule

$$\bar{\omega}^1\rightarrow(\bar{\omega}^1,\bar{\omega}^1)\,. \tag{3.2}$$

The first entry in the parentheses is the first fundamental weight of $SU(2)$ and the second that of $SO(N)$. Equation (3.2) says that the fundamental $2N$-dimensional representation of $Sp(2N)$ branches into the direct product of an $SU(2)$ doublet with a vector of $SO(N)$.

The $Sp(2N)$ representations at the lowest eigenvalues of L_0 in the highest weight representations ω^μ $(\mu=0,1,\ldots,N)$ of $\widehat{Sp}(2N)$ are the scalar and basic representations, 0, $\bar{\omega}^i$ $(i=1,\ldots,N)$. The branching rules for the latter into $SU(2)\times SO(N)$ are

$$\bar{\omega}^i\rightarrow\sum_{s=0}^{[i/2]}((i-2s)\bar{\omega}^1,\bar{v}^s+\bar{v}^{i-s})\,. \tag{3.3}$$

Here the $\bar{\omega}^1$, on the right-hand side, is the fundamental weight of $SU(2)$, and the $\bar{v}^i$ are weights of $SO(N)$. The definitions of the $\bar{v}^i$ differ for N odd and N even. For $N=2Q+1$,

$$\begin{aligned}&\bar{v}^j=\bar{\omega}^j\quad(1\leqslant j\leqslant Q-1)\,,\\&\bar{v}^Q=2\bar{\omega}^Q,\quad\bar{v}^0=0\,.\end{aligned} \tag{3.4}$$

The weights with indices larger than the rank Q in (3.3) are handled by duality:

$$\bar{v}^{2Q+1-j}=\bar{v}^j\,. \tag{3.5}$$

For $N=2Q$, the definitions are

$$\begin{aligned}&\bar{v}^j=\bar{\omega}^j\quad(1\leqslant j\leqslant Q-2)\,,\\&\bar{v}^0=0,\quad\bar{v}^{Q-1}=\bar{\omega}^Q+\bar{\omega}^{Q-1}\,,\\&\bar{v}^Q=2\bar{\omega}^Q\oplus2\bar{\omega}^{Q-1}\,.\end{aligned} \tag{3.6}$$

The symbol $\oplus$ indicates that there are *two* separate representations in $\bar{v}^Q$. Weights with indices larger than Q are again handled by duality:

$$\bar{v}^{2Q-j}=\bar{v}^j,\quad j>Q\,. \tag{3.7}$$

Note that (3.3) also correctly gives the branching rule for the scalar representation when $i=0$. For the reader's convenience we sketch the derivation of the finite branching rules in an Appendix.

The outer automorphisms of $\hat{Sp}(2N)$, $\hat{SU}(2)$, and

$\widehat{SO}(N)$ can be explained with the help of Fig. 1. $Sp(2N) = \hat{C}_N$ has a Z_2 outer automorphism group generated by A:

$$\hat{C}_N: \quad A\omega^s = \omega^{N-s}, \quad 0 \leqslant s \leqslant N. \tag{3.8}$$

The action of A is illustrated in Fig. 1(a). $\widehat{SU}(2) = \hat{A}_1$ also has a Z_2 outer automorphism group. Its generator a is depicted in Fig. 1(b):

$$\hat{A}_1: \; a\omega^0 = \omega^1, \quad a\omega^1 = \omega^0. \tag{3.9}$$

Since the root structure of $SO(N)$ differs for odd $N = 2Q + 1$, and even $N = 2Q$, we will discuss them separately.

A. $\hat{C}_{2Q+1} \triangleright \hat{A}_1^{2Q+1} \times \hat{B}_Q^4$

With $N = 2Q + 1$, the subalgebra is that shown above. The outer automorphism group of $\hat{B}_Q$ is Z_2, generated by α:

$$\alpha\omega^0 = \omega^1, \quad \alpha\omega^1 = \omega^0, \quad \alpha\omega^s = \omega^s, \quad s \neq 0,1. \tag{3.10}$$

Figure 1(c) illustrates the action of α.

There is considerable freedom in the choice of affine projection matrix $\hat{F}$. One restriction is that it should contain an acceptable projection matrix $\overline{F}$ for the finite Lie algebra embedding. Here $\overline{F}$ is obtained from $\hat{F}$ by deleting the rows and columns associated with the zeroth fundamental weights of the "large" algebra and all embedded algebras. For $\overline{F}$ to be satisfactory for (3.1), it is necessary and sufficient that it reproduce (3.2).

An acceptable projection matrix $\hat{F}$ is

$$\begin{bmatrix}
N & N-1 & N-2 & \cdots & Q+2 & Q+1 & Q & Q-1 & \cdots & 2 & 1 & 0 \\
0 & 1 & 2 & \cdots & Q-1 & Q & Q+1 & Q+2 & \cdots & N-2 & N-1 & N \\
- & - & - & & - & - & - & - & & - & - & - \\
4 & 3 & 3 & & 3 & 2 & 2 & 3 & & 3 & 3 & 4 \\
0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\
0 & 0 & 1 & & 0 & 0 & 0 & 0 & & 1 & 0 & 0 \\
 & & & \ddots & & & & & \iddots & & & \\
0 & 0 & 0 & \cdots & 1 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\
0 & 0 & 0 & & 0 & 2 & 2 & 0 & & 0 & 0 & 0
\end{bmatrix} \tag{3.11}$$

This matrix makes manifest the relation

$$A = a \times 1. \tag{3.12}$$

The finite branching rules give us part of the full branching rules. Equation (3.3) implies

$$\omega^\mu \to \sum_{s=0}^{[\mu/2]} [(N - \mu + 2s)\omega^0 + (\mu - 2s)\omega^1, \nu^s + \nu^{\mu - s}] + \ldots, \tag{3.13}$$

where we define the weights ν^s such that $[\nu^\rho + \nu^\lambda]$ is always a level-four representaion of $\hat{B}_Q$:

$$\begin{aligned} &\nu^\mu = \omega^\mu, \quad 2 \leqslant \mu \leqslant Q - 1, \\ &\nu^Q = 2\omega^Q, \quad \nu^1 = \omega^0 + \omega^1, \end{aligned} \tag{3.14}$$

and also $\nu^0 = 2\omega^0$. Weights with negative indices in (3.13) are defined by

$$\nu^\mu = \nu^{-\mu}. \tag{3.15}$$

Equation (3.13) also implies

$$\omega^{N-\mu} \rightsquigarrow \sum_{s=0}^{Q-[\mu/2]} [(\mu + 2s)\omega^0 + (N - \mu - 2s)\omega^1, \nu^s - \nu^{N-\mu-s}] + \ldots. \tag{3.16}$$

Using duality,

$$\nu^{N-\mu-s} = \nu^{\mu+s} \tag{3.17}$$

and flipping the sign of s we get

$$\omega^{N-\mu} \to \sum_{s=[\mu/2]-Q}^{0} [(\mu - 2s)\omega^0 + (N - \mu + 2s)\omega^1, \nu^s + \nu^{\mu - s}] + \ldots.$$

Now applying (3.12) to this last equation and adding the result to (3.13) yields

$$\omega^\mu \to \sum_{s=[\mu/2]-Q}^{[\mu/2]} [(N - \mu + 2s)\omega^0 + (\mu - 2s)\omega^1, \nu^s + \nu^{\mu - s}] + \ldots. \tag{3.18}$$

To see whether or not (3.18) is the complete affine branching rule one can check numerically to see if the asymptotic sum rule (2.24) is satisfied. We find that it is not; there are representations missing from (3.18). But we find the unique way to satisfy the asymptotic sum rule by adding representations obeying the level matching condition (2.22) is to modify the definition of ν^0 to

$$\nu^0 = 2\omega^0 \oplus 2\omega^1. \tag{3.19}$$

Again, the symboll $\oplus$ indicates ν^0 consists of *two* representations. Then the following are the complete branching rules for the conformal embedding $\hat{C}_{2Q+1} \triangleright \hat{A}_1^{2Q+1} \times \hat{B}_Q^4$:

$$\omega^\mu \to \sum_{s=[\mu/2]-Q}^{[\mu/2]} [(N - \mu + 2s)\omega^0 + (\mu - 2s)\omega^1, \nu^s + \nu^{\mu - s}] \tag{3.20}$$

with the definitions (3.14) and (3.19).

To make the procedure perfectly clear, let us go through a specific example: the branching rule of

$\omega^2 = [0010000000]$ of $\hat{C}_9$ into $\hat{A}_1^9 \times \hat{B}_4^4$. Here (3.3) tells us the branching rules for the representations $\bar{\omega}^2, \bar{\omega}^7$ of the finite Lie algebra C_9 into representations of $A_1 \times B_4$:

$$\bar{\omega}^2 \to (2-0100) + (0-2000)\,,$$
$$\bar{\omega}^7 \to (7-0100) + (5-1010) + (3-0102) + (1-0012)\,.$$

Each of these tells us part of the corresponding affine branching rule:

$$\omega^2 \to [72-20100] + [90-22000] + \dots\,,$$
$$\omega^7 \to [27-20100] + [45-11010] + [63-00102] + [81-00012] + \dots\,.$$

Applying the automorphism (3.12) to the second equation and combining the result with the first gives

$$\omega^2 \to [90-22000] + [72-20100] + [54-11010] + [36-00102] + [18-00012] + \dots\,.$$

The unique way to satisfy (2.24) by adding representations obeying (2.22) is to include $[72-02100]$. Then the result is that dictated by (3.20).

B. $\hat{C}_{2Q} \triangleright \hat{A}_1^{2Q} \times \hat{D}_Q^4$

For even $N = 2Q$ in $\widehat{\mathrm{Sp}}(2N) \triangleright \widehat{\mathrm{SU}}(2)^N \times \widehat{\mathrm{SO}}(N)^4$, the embedding is $\hat{C}_{2Q} \triangleright \hat{A}_1^{2Q} \times \hat{D}_Q^4$. The outer automorphisms of $\hat{D}_Q$ differ for Q odd and even, and can be explained using the Dynkin diagram symmetries of Fig. 1(d). The symmetries e_b, e_f, and μ are defined by

$$\begin{aligned} &e_b\omega^0 = \omega^1, \quad e_b\omega^1 = \omega^0, \quad e_b\omega^\lambda = \omega^\lambda, \quad \lambda \neq 0,1\,,\\ &e_f\omega^{Q-1} = \omega^Q, \quad e_f\omega^Q = \omega^{Q-1},\\ &e_f\omega_\lambda = \omega^\lambda, \quad \lambda \neq Q-1, Q\,,\\ &\mu\omega^\lambda = \omega^{Q-\lambda}, \quad 0 \leqslant \lambda \leqslant Q\,. \end{aligned} \tag{3.21}$$

For Q odd there is an outer automorphism,

$$\mu' = \mu e_f = e_b \mu\,, \tag{3.22}$$

of period four. For Q even the outer automorphism group is $Z_2 \times Z_2$, with generators μ and

$$\sigma = e_b e_f\,. \tag{3.23}$$

Notice that σ is also an outer automorphism for odd Q, since $(\mu')^2 = \sigma$.

We have found three different affine projection matrices manifesting the outer automorphisms of the embedding $\hat{A}_1^{2Q} \times \hat{D}_Q^4 \triangleleft \hat{C}_{2Q}^1$. One is

$$\hat{F} = \begin{bmatrix} N & N-1 & N-2 & \cdots & Q+1 & Q & Q-1 & \cdots & 2 & 1 & 0 \\ 0 & 1 & 2 & \cdots & Q-1 & Q & Q+1 & \cdots & N-2 & N-1 & N \\ 4 & 3 & 2 & & 2 & 2 & 2 & & 2 & 3 & 4 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 1 & & 0 & 0 & 0 & & 1 & 0 & 0 \\ & \vdots & & \ddots & & \vdots & & \iddots & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & & 0 & 2 & 0 & & 0 & 0 & 0 \end{bmatrix}. \tag{3.24}$$

This matrix shows the branching rules obey (3.12), as in the case when N is odd, discussed above. A second matrix is

$$\hat{F} = \begin{bmatrix} N & N-1 & N & N-1 & N & & N & N-1 & N & N-1 & N \\ 0 & 1 & 0 & 1 & 0 & \cdots & 0 & 1 & 0 & 1 & 0 \\ 4 & 3 & 2 & 1 & 0 & & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & & 0 & 0 & 0 & 0 & 0 \\ & & \vdots & & & \ddots & & & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & & 0 & 1 & 2 & 3 & 4 \end{bmatrix}. \tag{3.25}$$

For Q odd, this manifests

$$A = 1 \times \mu'\,, \tag{3.26}$$

while for Q even it realizes

$$A = 1 \times \mu\,. \tag{3.27}$$

The third matrix is obtained by interchanging the last two rows of (3.25). So we also have

$$A = 1 \times \sigma\mu' \tag{3.28}$$

for Q odd and

$$A = 1 \times \sigma\mu \tag{3.29}$$

for Q even, when acting on the affine branching rules.

As in Sec. III A, we have $A = a \times 1$. Also, the finite branching rules (3.3) take the same form for N odd or even. So we arrive immediately at the result (3.18). However, the $\bar{\nu}^i$ of (3.3) are defined differently for $N = 2Q$ [compare (3.6) and (3.4)], so the corresponding ν^μ must also be defined differently.

In fact, comparing (3.26) with (3.28) for Q odd and

(3.27) with (3.29) for Q even tells us that the affine branching rules must be invariant under σ. This in turn means that the affine weights ν^μ in (3.18) must be redefined to be σ invariant. The appropriate ν^μ are therefore

$$\begin{aligned} &\nu^\mu = \omega^\mu, \quad 2 \leqslant \mu \leqslant Q-2, \\ &\nu^0 = 2\omega^0 \oplus 2\omega^1, \quad \nu^1 = \omega^0 + \omega^1, \\ &\nu^{Q-1} = \omega^{Q-1} + \omega^Q, \quad \nu^Q = 2\omega^{Q-1} \oplus 2\omega^Q. \end{aligned} \tag{3.30}$$

We have verified numerically (for a large number of cases) that the asymptotic sum rule (2.24) is also satisfied by the branching rule (3.20) for N even. Again, weights with negative indices are defined by $\nu^\mu = \nu^{-\mu}$, and the ν^μ satisfy the duality relation (3.17).

IV. CONCLUSION

For ease of reference, let us first state that the branching rules for the conformal embedding $\widehat{SU}(2)^N \times \widehat{SO}(N)^4 \triangleleft \hat{S}p(2N)$ are

$$\omega^\mu \to \sum_{s=[\mu/2]-Q}^{[\mu/2]} [(N-\mu+2s)\omega^0 + (\mu-2s)\omega^1, \nu^s + \nu^{\mu-s}] .$$

The $\widehat{SO}(N)$ weights ν^μ are defined differently for N odd or even; for $N = 2Q+1$ the definitions are given in (3.14) and (3.19), and for $N = 2Q$ they are those written in (3.30). Weights ν^μ with indices μ too small or too large are to be understood using $\nu^{-\mu} = \nu^\mu$ and $\nu^\mu = \nu^{N-\mu}$, respectively.

Our results demonstrate the general utility of outer automorphisms in the calculation of conformal branching rules.

They also complete the calculation of the conformal branching rules for infinite series of nonsimple higher level embeddings. There are four such infinite series of conformal subalgebras.[9] $\widehat{SU}(2)^N \times \widehat{SO}(N)^4 \triangleleft \widehat{Sp}(2N)$ was treated here and $\widehat{SU}(p)^q \times \widehat{SU}(q)^p \triangleleft \widehat{SU}(pq)$ in Ref. 8. The remaining two are $\widehat{SO}(p)^q \times \widehat{SO}(q)^p \triangleleft \widehat{SO}(pq)$ and $\widehat{Sp}(2p)^q \times \widehat{Sp}(2q)^p \triangleleft \widehat{SO}(4pq)$. But their branching rules may be calculated using a theorem[13] applicable to conformal embeddings, $\hat{j} \triangleleft \widehat{SO}(D)$, when there exists a symmetric space $\bar{g}/\bar{j}$ of dimension D.

We believe the use of Kač–Moody outer automorphisms will allow the calculation of all conformal branching rules.

ACKNOWLEDGMENTS

This research was supported in part by a postdoctoral fellowship from NSERC of Canada and by the U.S. Department of Energy, under Contract No. DE-AC03-76SF00515.

APPENDIX: BRANCHING RULES FOR Sp(2N)⊃SU(2)×SO(N)

In this appendix we derive formula (3.3) for the branching of the basic representations of $Sp(2N)$ into $SU(2) \times SO(N)$ representations.

Here $Sp(2N)$ is the group of transformations of $2N$-dimensional vectors that leaves invariant the antisymmetric tensor T with nonzero components $T_{M,M+1} = 1 = -T_{M+1,M}$. The basic representations $\bar{\omega}^i$ can be represented by traceless antisymmetric tensors of rank i, the traces taken using the invariant two tensor T. As such, they can be represented by Young tableaux consisting of one column of i boxes.

Representations of $SU(2)$ can also be symbolized by Young tableaux. A row of m boxes realizes the representation (m). Since $SU(2)$ leaves invariant the ϵ tensor of rank 2, a column of two boxes is equivalent by duality to a scalar (0). A column of more than two boxes is impossible and must be excluded.

Totally antisymmetric tensors again correspond to irreducible representations of $SO(N)$. The representations are not the basic ones, however. In fact, an antisymmetric tensor of rank i transforms as the representation with highest weight $\bar{\nu}^i$ of (3.4) or (3.6), according to whether N is odd or even.

Since $SO(N)$ transformations preserve determinants, they leave invariant the ϵ tensor of rank N. So the concept of dual tensors applies to $SO(N)$ as well. This is the origin of the relations (3.5) and (3.7). In the language of Young tableaux, it says that a column of j boxes is dual to a column of $N-j$ boxes.

The embedding $Sp(2N) \supset SU(2) \times SO(N)$ is defined by the branching rule

$$\bar{\omega}^1 \to (\bar{\omega}^1, \bar{\omega}^1).$$

This is depicted in Fig. 2(a), and allows us to use Young tableaux to find the branching rules for (3.3).

Consider a basic representation $\bar{\omega}^i$ of $Sp(2N)$. The i boxes of the column that is the corresponding Young tableau break up into $2i$ boxes, i for $SU(2)$ and the other i for $SO(N)$. Each of the two sets of i boxes can form a Young tableau of any type, but the antisymmetry of the original rank i tensor must be respected. This is done by pairing $SU(2)$ and $SO(N)$ representations such that their Young tableaux can be obtained from each other by interchanging rows and columns.

As an example, the branching of the fourth basic representation $\bar{\omega}^4$ of $Sp(2N)$ is shown in Fig. 2(b). If the first Young tableau in each pair on the right-hand side is that of

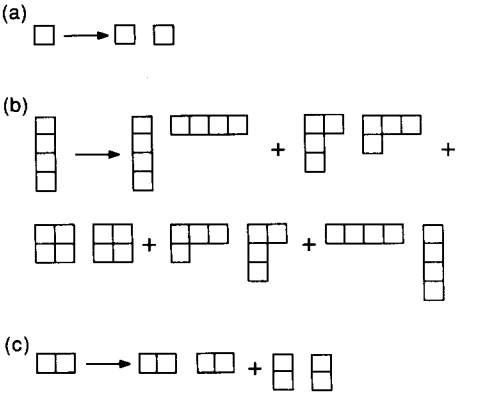


FIG. 2. $Sp(2N) \supset SU(2) \times SO(N)$ branching rules using Young tableaux. (a) Depicted is the defining branching rule of the fundamental representation $\bar{\omega}^1 \to (\bar{\omega}^1, \bar{\omega}^1)$. Branchings of the fourth basic representation and the adjoint representation of $Sp(2N)$ are shown in (b) and (c), respectively.

SU(2), the first two pairs are superfluous, since the tableaux do not correspond to SU(2) representations. It is easy to convince oneself that this procedure for any basic representation leads to (3.3).

For completeness we mention that Young tableaux can be used to find the branchings of other Sp($2N$) representations, but some care is needed. Consider the adjoint representation. It can be represented by a symmetric rank 2 tensor. Because of its symmetry, its branching is represented by identical Young tableaux for SU(2) and SO(N), as shown in Fig. 2(c). But the original Sp($2N$) tensor is not traceless, since there is no Sp($2N$) invariant symmetric tensor with which to contract. Therefore the row of two SO(N) boxes on the right-hand side of Fig. 2(c) represents two representations; $2\bar{\omega}^1$ and a scalar that is essentially the trace of the SO(N) tensor. The adjoint branching rule is therefore

$$2\bar{\omega}^1 \rightarrow (2\bar{\omega}^1, 2\bar{\omega}^1) + (2\bar{\omega}^1, 0) + (0, \bar{\omega}^2) .$$

[1]V. G. Kač, *Infinite-dimensional Lie Algebras* (Cambridge U. P., Cambridge, 1985); R. Slansky, Comments Nucl. Part. Phys. **18**, 175 (1988); S. Kass, R. V. Moody, J. Patera, and R. Slansky, *Affine Kač–Moody Algebras, Weight Multiplicities and Branching Rules* (University of California, Berkeley, to be published).

[2]P. Goddard and D. Olive, Int. J. Mod. Phys. A **1**, 303 (1986).

[3]F. A. Bais, F. Englert, A. Taormina, and P. Zizzi, Nucl. Phys. B **279**, 529 (1987).

[4]F. Bais and A. Taormina, Phys. Lett. B **181**, 87 (1986).

[5]P. Bouwknegt and W. Nahm, Phys. Lett. B **184**, 359 (1987).

[6]P. Bouwknegt, Nucl. Phys. B **290**, 507 (1987).

[7]D. Altschuler, Berkeley preprint LBL-TH-25125, 1988.

[8]M. A. Walton, preprint SLAC-PUB-4680, August, 1988; revised November, 1988; to be published in Nucl. Phys. B.

[9]A. N. Schellekens and N. P. Warner, Phys. Rev. D **34**, 3092 (1986); F. Bais and P. Bouwknegt, Nucl. Phys. B **279**, 561 (1987); R. Arcuri, J. Gomez, and D. Olive, *ibid.* **285**, 327 (1987).

[10]V. Kač and D. Peterson, Adv. Math. **53**, 125 (1984).

[11]A. Navon and J. Patera, J. Math. Phys. **8**, 489 (1967); W. McKay, J. Patera, and D. Sankoff, in *Computers in Nonassociative Rings and Algebras*, edited by J. Beck and B. Kolman (Academic, New York, 1977).

[12]V. Kač and M. Wakimoto, Adv. Math. **70**, 156 (1988); V. Kač and M. N. Sanielevici, Phys. Rev. D **37**, 2231 (1988).

[13]The theorem is due to V. Kač and D. Peterson, Proc. Natl. Acad. Sci. U.S.A. **78**, 3308 (1981), with a correction in W. Nahm, "Quantum field theories in 1 and 2 dimensions," Stony Brook preprint ITP-SB-87-7. This is a straightforward generalization of a theorem for finite Lie algebras, proved in R. Parthasarathy, Ann. Math. **96**, 1 (1972).

Highest weight representations for gl(m/n) and gl($m+n$)

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

(Received 24 June 1988; accepted for publication 22 February 1989)

It is shown how vector coherent state (VCS) theory enlightens the simultaneous discussion of representation theory for classical Lie algebras and superalgebras and provides an optimal framework for the explanation of the noted similarities and dissimilarities of their representations. Reducibility, atypicality, and the positive-definitiveness of inner products in VCS representation spaces are discussed. The discussion is exemplified through the parallel and explicit construction of highest weight ladder representations of the Lie algebra gl($m+n$) and superalgebra gl(m/n) in gl(m) $\oplus$ gl(n) bases.

I. INTRODUCTION

It has recently been shown in considerable detail[1-6] how vector coherent state (VCS) theory provides, under certain conditions, a means to construct ladder representations of a complex Lie algebra $\mathbf{g}$ in bases adapted to algebra–subalgebra chains of the type $\mathbf{g}\supset\mathbf{n}_0$, with $\text{rank}(\mathbf{n}_0)=\text{rank}(\mathbf{g})$. VCS theory then provides a systematic prescription by which one can induce irreps of $\mathbf{g}$ from highest or lowest weight irreps of $\mathbf{n}_0$. The VCS construction has much in common with the standard Chevalley–Harish-Chandra[7] method of inducing highest (or lowest) weight representations of a Lie algebra from a Cartan subalgebra. Similarly, its application to Lie superalgebras has much in common with the parallel inducing construction given by Kac.[8] However, there are notable differences particularly as regards the construction of the irreducible modules. In the Chevalley–Harish-Chandra–Kac construction, it is generally necessary to factorize the induced module with respect to its maximal invariant submodule. This latter step can be quite complicated and is entirely avoided in the VCS construction. The irreducibility of the VCS induced representations is illustrated in this paper but the proof will be given elsewhere.

Fundamental to VCS theory is the Z gradation of $\mathbf{g}$,

$$\mathbf{g}=\mathbf{n}_0+\sum_{i=1,2,\ldots}\mathbf{n}_{\pm i}, \tag{1.1}$$

where the gradation is performed by a grading operator $\hat{Z}$ belonging to the Cartan subalgebra and such that

$$[\hat{Z},x]=ix, \quad \forall x\in\mathbf{n}_i. \tag{1.2}$$

This gradation endows the Lie algebra with a Z-graded structure. By definition, a Lie algebra, endowed with a Z-graded structure, is a vector space $\mathbf{g}$ that (i) is a direct sum of vector subspaces $\mathbf{n}_i$, where the index i takes integer values; and (ii) has a bilinear product that satisfies

$$[x,y]\in\mathbf{n}_{i+j}, \tag{1.3a}$$

$$[x,y]=-[y,x], \tag{1.3b}$$

$$[x,[y,z]]=[[x,y],z]+[y,[x,z]], \tag{1.3c}$$

for $x\in\mathbf{n}_i$, $y\in\mathbf{n}_j$, and any z in $\mathbf{g}$.

The Z-graded stucture of a Lie algebra $\mathbf{g}$ has some important properties which are exploited in VCS theory.

(1) The zero grade component $\mathbf{n}_0$, usually referred to as the stability algebra, is a reductive subalgebra of $\mathbf{g}$ which is its own normalizer. It plays a central role in VCS theory. Obviously,the Cartan subalgebra $\mathbf{h}$ belongs to $\mathbf{n}_0$, $\mathbf{h}\subset\mathbf{n}_0$. However, unlike $\mathbf{h}$, $\mathbf{n}_0$ need not be Abelian. Note that the grading element $\hat{Z}$ belongs to $\mathbf{h}\subset\mathbf{n}_0$ and hence to the center of $\mathbf{n}_0$; i.e., $[\hat{Z},x]=0$ for any $x\in\mathbf{n}_0$.

(2) The subspaces

$$\mathbf{n}_+=\sum_{i>0}^{i_{\max}}\mathbf{n}_i, \quad \mathbf{n}_-=\sum_{i<0}^{-i_{\max}}\mathbf{n}_i \tag{1.4}$$

are nilpotent subalgebras of raising and lowering operators. The index $i_{\max}$ corresponds to the highest grade pertaining to the given Z gradation of the Lie algebra. Since each level $\mathbf{n}_i$ is invariant under the adjoint action $\text{ad}_{\mathbf{n}_0}$ of the stability algebra $\mathbf{n}_0$, the subalgebras $\mathbf{n}_\pm$ are generally reducible under $\text{ad}_{\mathbf{n}_0}$. For $i_{\max}=1$, the subalgebras $\mathbf{n}_\pm$ are necessarily Abelian. The first applications of VCS theory addressed the Abelian case,[1-3,5,6] but recent developments[3,4] have shown that it applies equally to non-Abelian cases with $i_{\max}\geqslant 2$.

Now, the Z-graded structure of the Lie algebra naturally imparts a Z-graded structure on a representation with highest or lowest weight of the Lie algebra. In particular, it defines a subspace of highest and/or lowest grade vectors which are, respectively, annihilated by the subset $\mathbf{n}_+$ or $\mathbf{n}_-$, and which transform irreducibly under the action of the stability algebra $\mathbf{n}_0$. The VCS construction then corresponds to the induction of highest/lowest weight irreps of $\mathbf{g}$ on a space of vector-valued holomorphic functions taking values in the highest/lowest grade irreducible $\mathbf{n}_0$ subspace.

Now, the gist of these considerations also applies to the so-called classical Lie superalgebras.[9] By definition, a Lie superalgebra, endowed with a Z-graded structure, is a vector space $\mathbf{g}$ that (i) is a direct sum of vector subspaces $\mathbf{n}_i$, where the index i takes integer values; and (ii) has a bilinear product that satisfies

$$[x,y]\in\mathbf{n}_{i+j}, \tag{1.5a}$$

$$[x,y]=-(-1)^{ij}[y,x], \tag{1.5b}$$

$$[x,[y,z]]=[[x,y],z]+(-1)^{ij}[y,[x,z]], \tag{1.5c}$$

for $x\in\mathbf{n}_i$, $y\in\mathbf{n}_j$, and any z in $\mathbf{g}$. One observes that the even grade sector $\mathbf{g}_{\bar{0}}$ of $\mathbf{g}$,

$$\mathbf{g}_{\bar{0}}=\sum_{i\,\text{even}}\mathbf{n}_i, \tag{1.6a}$$

 0022-2488/89/071415-18$02.50

defines a standard Lie algebra, and that the superalgebra $\mathbf{g}$ has a Z_2-graded structure

$$\mathbf{g} = \mathbf{g}_{\bar{0}} + \mathbf{g}_{\bar{1}}$$

with an odd grade sector defined by

$$\mathbf{g}_{\bar{1}} = \sum_{i\,\text{odd}} \mathbf{n}_i. \tag{1.6b}$$

The Z_2 gradation of such a superalgebra is thus given by its Z gradation modulo 2. The Z gradation is then said to be consistent with the Z_2 gradation defining the superalgebra.[9]

It can be shown that any classical Lie superalgebra can be assigned a convenient Z-graded structure, with either $i_{\max} = 1$ or 2, and where $\mathbf{n}_0$ is either the Lie algebra $\mathbf{g}_{\bar{0}}$ or a subalgebra thereof. We then have for classical Lie superalgebras that the even sector $\mathbf{g}_{\bar{0}}$ of the superalgebra is given by either $\mathbf{g}_{\bar{0}} = \mathbf{n}_0$ or $\mathbf{n}_0 + \mathbf{n}_{-2} + \mathbf{n}_{+2}$, and the odd sector $\mathbf{g}_{\bar{1}}$ is given by $\mathbf{g}_{\bar{1}} = \mathbf{n}_{-1} + \mathbf{n}_{+1}$.

The aim of this paper is to demonstrate that VCS theory applies without substantial modification to the representation theory of classical Lie superalgebras. In fact, the only significant departure amounts to a replacement of the vector-valued polynomials in Bargmann variables by vector-values polynomials in either Grassman variables ($i_{\max} = 1$) or both Bargmann and Grassman variables ($i_{\max} = 2$). All noted dissimilarities between the representation theory of Lie algebras and superalgebras are then readily explained in terms of the different algebraic properties of these variables. It is also shown how VCS theory enlightens the discussion of such concepts as reducibility, atypicality, and the positive-definitiveness of inner products in (graded) Hilbert spaces. For simplicity, we restrict our attention herein to the reductive Lie algebras $\mathrm{gl}(m+n) \supset \mathrm{gl}(m) \oplus \mathrm{gl}(n)$ and their superalgebraic counterparts, the superalgebras of the type $\mathrm{gl}(m/n) \supset \mathrm{gl}(m) \oplus \mathrm{gl}(n)$, $m,n \geqslant 1$, for which

$$\mathbf{g}_{\bar{0}} = \mathbf{n}_0 = \mathrm{gl}(m) \oplus \mathrm{gl}(n), \quad \mathbf{g}_{\bar{1}} = \mathbf{n}_{-1} + \mathbf{n}_{+1}.$$

Classical Lie superalgebras for which

$$\mathbf{g}_{\bar{0}} = \mathbf{n}_0 + \mathbf{n}_{-2} + \mathbf{n}_{+2}, \quad \mathbf{g}_{\bar{1}} = \mathbf{n}_{-1} + \mathbf{n}_{+1}$$

will be considered in a subsequent publication.[10]

II. Z GRADING OF THE ALGEBRAS

A. Z grading of the Lie algebra $\mathrm{gl}(m+n) \supset \mathrm{gl}(m) \oplus \mathrm{gl}(n)$

The canonical basis $\{E_{AB};\ 1 \leqslant A,B \leqslant m+n\}$ for the (complexification of the) $(m+n)^2$-dimensional Lie algebra $\mathrm{gl}(m+n)$, $m,n \geqslant 1$, satisfies the commutation relations

$$[E_{AB}, E_{CD}] = \delta_{BC} E_{AD} - \delta_{AD} E_{CB}, \tag{2.1}$$

where $A, B, \ldots = 1,2\ldots,m+n$.

The partial trace operator

$$\hat{Z} = \sum_{k=1}^{m} E_{kk} \tag{2.2}$$

is a convenient Z-grading operator for $\mathrm{gl}(m+n)$. It naturally performs a Z gradation of $\mathrm{gl}(m+n)$ into three subalgebras: (a) a *nilpotent Abelian* subalgebra $\mathbf{n}_{+1}$ of Z grade $g = +1$ spanned by the subset of raising operators

$$A_{i\alpha} = E_{i,\alpha+m}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant \alpha \leqslant n; \tag{2.3a}$$

(b) a *nilpotent Abelian* subalgebra $\mathbf{n}_{-1}$ of Z grade $g = -1$ spanned by the subset of lowering operators

$$B_{i\alpha} = E_{m+\alpha,i}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant \alpha \leqslant n; \tag{2.3b}$$

(c) a reducible stability subalgebra $\mathbf{n}_0$ of Z grade $g = 0$, the centralizer of $\hat{Z}$, isomorphic to $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ with

$$\begin{aligned} &\mathrm{gl}(m) = \mathrm{span}\{C_{ij} = E_{ij}, \quad 1 \leqslant i,j \leqslant m\}, \\ &\mathrm{gl}(n) = \mathrm{span}\{C_{\alpha\beta} = E_{\alpha+m,\beta+m}, \quad 1 \leqslant \alpha,\beta \leqslant n\}. \end{aligned} \tag{2.3c}$$

[We follow herein the convention that lowercase roman indices $(i,j,k\ldots)$ assume values from 1 to m; greek indices $(\alpha,\beta,\gamma\ldots)$ from 1 to n; and uppercase roman indices $(A,B,C\ldots)$, from 1 to $m+n$.]

The subset of diagonal elements

$$\mathbf{h} = \mathrm{span}\{E_{11}, E_{22}, \ldots, E_{m+n,m+n}\} \tag{2.4}$$

spans a splitting Cartan subalgebra $\mathbf{h}$ of $\mathrm{gl}(m+n)$.

From (2.1), we obtain the following nonvanishing commutators:

$$\begin{aligned} &[C_{ij}, C_{kl}] = \delta_{jk} C_{il} - \delta_{il} C_{kj}, \\ &[C_{\alpha\beta}, C_{\mu\nu}] = \delta_{\beta\mu} C_{\alpha\nu} - \delta_{\alpha\nu} C_{\mu\beta}, \\ &[C_{ij}, A_{k\mu}] = \delta_{jk} A_{i\mu}, \quad [C_{\alpha\beta}, A_{k\mu}] = -\delta_{\alpha\mu} A_{k\beta}, \\ &[C_{ij}, B_{k\mu}] = -\delta_{ik} B_{j\mu}, \quad [C_{\alpha\beta}, B_{k\mu}] = \delta_{\beta\mu} B_{k\alpha}, \\ &[A_{i\alpha}, B_{j\beta}] = \delta_{\alpha\beta} C_{ij} - \delta_{ij} C_{\beta\alpha}. \end{aligned} \tag{2.5}$$

From (2.5), we conclude that the set of raising operators $\{A_{i\alpha}\}$ spans an irreducible representation of the stability algebra $\mathbf{n}_0 \sim \mathrm{gl}(m) \oplus \mathrm{gl}(n)$ labeled by the partitions $\{1\}:\{-1\}$, while the set of lowering operators $\{B_{i\alpha}\}$ spans an irreducible representation of $\mathbf{n}_0$ labeled by the partitions $\{-1\}:\{1\}$. [For ease, we use the shorthand notation $\{1\}$ for the partition in l parts $\{1\dot{0}\} \sim \{100\cdots0\}$ with $(l-1)$ zeros and $\{-1\}$ for $\{0,\ldots,0,-1\}$.] Similarly, the $\mathrm{sl}(m)$ subalgebra

$$\mathrm{sl}(m) = \mathrm{span}\{C_{ij} - (1/m)\delta_{ij} C_{kk}\}$$

has rank $\{1\dot{0}-1\}:\{0\}$, the $\mathrm{sl}(n)$ subalgebra

$$\mathrm{sl}(n) = \mathrm{span}\{C_{\alpha\beta} - (1/n)\delta_{\alpha\beta} C_{\sigma\sigma}\}$$

has rank $\{0\}:\{1\dot{0}-1\}$, while the partial trace operators C_{kk} and $C_{\sigma\sigma}$ both have rank $\{0\}:\{0\}$. Following these identifications, we have that the Z grade of an element $X \in \mathrm{gl}(m+n)$, belonging to an $\mathbf{n}_0$ irreducible tensorial set $\{\mu\}:\{\nu\}$, is simply given by $\Sigma_{i=1}^{m} \mu_i$.

B. Z grading of the Lie superalgebra $\mathrm{gl}(m/n) \supset \mathrm{gl}(m) \oplus \mathrm{gl}(n)$

The above considerations apply almost verbatim to the Lie superalgebra $\mathrm{gl}(m/n)$, the essential modifications being as follows.

(i) The canonical basis $\{E_{AB};\ 1 \leqslant A,B \leqslant m+n\}$ for the (complexification of the) $(m+n)^2$-dimensional superalgebra $\mathrm{gl}(m/n)$, $m,n \geqslant 1$, now satisfies the *graded* commutation relations

$$\begin{aligned} [E_{AB}, E_{CD}] = {} & \delta_{BC} E_{AD} \\ & - \delta_{AD} (-1)^{(1/2)(\sigma_A - \sigma_B) \times (1/2)(\sigma_C - \sigma_D)} E_{CB}, \end{aligned} \tag{2.6}$$

where $A,B,\ldots = 1,2\ldots,m+n$, and where

$$\sigma_A = 1, \quad 1 \leqslant A \leqslant m,$$
$$\sigma_A = -1, \quad m+1 \leqslant A \leqslant m+n. \tag{2.7}$$

(ii) The nilpotent subalgebras $\mathbf{n}_{+1}$ and $\mathbf{n}_{-1}$ are now *super-Abelian*, i.e.,

$$[A_{i\alpha}, A_{j\beta}] = 0, \quad [B_{i\alpha}, B_{j\beta}] = 0$$

should be interpreted as vanishing anticommutators.

(iii) The anticommutator mapping $\mathbf{n}_{+1} \times \mathbf{n}_{-1} \to \mathbf{n}_0$ now reads

$$[A_{i\alpha}, B_{j\beta}] = \delta_{\alpha\beta} C_{ij} + \delta_{ij} C_{\beta\alpha}. \tag{2.8}$$

(iv) The algebra Z gradation is consistent with the Z_2 gradation

$$\mathrm{gl}(m/n)_{\bar{0}} = \mathbf{n}_0 = \mathrm{gl}(m) \oplus \mathrm{gl}(n),$$
$$\mathrm{gl}(m/n)_{\bar{1}} = \mathbf{n}_{+1} + \mathbf{n}_{-1}, \tag{2.9}$$

defining $\mathrm{gl}(m/n)$ as a superalgebra.

III. GRADED HIGHEST WEIGHT MODULES

A. Graded highest weight modules over gl(m+n)

We shall refer to the carrier space for a representation of either a Lie algebra or Lie superalgebra $\mathbf{g}$ as a $\mathbf{g}$ module. Let $M(\Lambda)$ be a module for an irreducible ladder representation of $\mathrm{gl}(m+n)$ with highest weight Λ (see also Sec. VI A). The Z gradation of $\mathrm{gl}(m+n)$ naturally imparts a Z gradation on this module.

Let $\overline{M}(\Lambda)$ be the highest Z-grade subspace with respect to the Z gradation, i.e., the subspace of weight vectors in $M(\Lambda)$ of homogeneous highest Z grade $g_{\max}$ annihilated by the subalgebra of raising operators $\mathbf{n}_{+1}$:

$$\overline{M}(\Lambda) = \left\{ |\eta\rangle \in M(\Lambda) \text{ such that } \begin{array}{c} \hat{Z}|\eta\rangle = g_{\max}|\eta\rangle, \\ A_{i\alpha}|\eta\rangle = 0, \quad \forall A_{i\alpha} \in \mathbf{n}_{+1} \end{array} \right\}. \tag{3.1}$$

This subspace carries an irreducible representation of the stability algebra and will be referred to as the *intrinsic* $\mathbf{n}_0$ module. It is assumed herein (although this is not necessary) that the intrinsic $\mathbf{n}_0$ module $\overline{M}(\Lambda)$ is finite dimensional and equivalent to a unitary representation of $\mathbf{n}_0$. It is conveniently labeled by its $\mathrm{gl}(m)$ and $\mathrm{gl}(n)$ highest weights

$$\{\mu^0\}:\{\nu^0\} \equiv \{\mu_1^0 \mu_2^0 \cdots \mu_m^0\}:\{\nu_1^0 \nu_2^0 \cdots \nu_n^0\},$$

where $\{\mu^0\}$ and $\{\nu^0\}$ refer to $\mathrm{gl}(m)$ and $\mathrm{gl}(n)$, respectively, and μ_i^0 and ν_α^0 are real numbers such that $(\mu_i^0 - \mu_{i+1}^0)$ and $(\nu_\alpha^0 - \nu_{\alpha+1}^0)$ are non-negative integers. By definition, $\{\mu^0\}:\{\nu^0\}$ are also the highest weights for the $\mathrm{gl}(m+n)$ irrep Λ; i.e., if $|\Lambda\rangle$ is the highest weight state in $M(\Lambda)$, we have

$$E_{ii}|\Lambda\rangle = C_{ii}|\Lambda\rangle = \mu_i^0|\Lambda\rangle,$$
$$E_{\alpha+m,\alpha+m}|\Lambda\rangle = C_{\alpha\alpha}|\Lambda\rangle = \nu_\alpha^0|\Lambda\rangle, \tag{3.2}$$
$$E_{i,\alpha+m}|\Lambda\rangle = A_{i\alpha}|\Lambda\rangle = 0$$

(no sum on i or α). Thus

$$(\Lambda) = (\mu_1^0 \mu_2^0 \cdots \mu_m^0 \nu_1^0 \nu_2^0 \cdots \nu_n^0). \tag{3.3}$$

Following the definition (2.2) of the grading operator, the highest Z grade of this representation is verified to be given by $g_{\max} = \sum_{i=1}^m \mu_i^0$.

B. Graded highest weight modules over gl(m/n)

Once more, all of the above considerations apply almost verbatim for $\mathrm{gl}(m/n)$: the module $M(\Lambda) = V_{\bar{0}} + V_{\bar{1}}$ is now defined as a Z_2-graded carrier space with highest weight Λ for a finite-dimensional irreducible representation of the Lie superalgebra $\mathrm{gl}(m/n)$ (see Secs. VI and IX for a discussion of the question of irreducibility). The Z gradation of $\mathrm{gl}(m/n)$ naturally imparts a Z gradation on this module compatible with the Z_2 gradation [see Eqs. (4.18) and (4.19)].

IV. VCS THEORY

A. VCS theory for gl(m+n)

A fundamental aspect of VCS theory for a Lie algebra $\mathbf{g}$ is the embedding of an irreducible graded highest weight $\mathbf{g}$ module $M(\Lambda)$ in a vector-Bargmann (VB) space. For $\mathrm{gl}(m+n)$, the VB space is the tensor product space $\mathscr{H}_{\mathrm{VB}} = V \oplus Bg$, where the following conditions hold.

(i) V is the intrinsic $\mathbf{n}_0$ module $V = \overline{M}(\Lambda)$ defined by Eq. (3.1). We recall that it carries a unitary irreducible finite-dimensional representation of the stability algebra $\mathbf{n}_0$. Let $\mathscr{B}_{\overline{M}} = \{|\eta\rangle\}$ be an orthonormal basis for V with respect to the inner product on V and let $\{\langle\eta|\}$ be a dual basis satisfying $\langle\eta|\eta'\rangle = \delta_{\eta\eta'}$.

(ii) Bg is the space of polynomials in the mn ($= \dim \mathbf{n}_{+1}$) Bargmann (complex) variables $\{z_{i\alpha}; 1 \leqslant i \leqslant m, 1 \leqslant \alpha \leqslant n\}$. The Bargmann space Bg is infinite dimensional. It is isomorphic to the symmetric tensor algebra over $\mathscr{C}^{mn}$, and has a natural inner product for which the nonzero polynomials

$$\left\{ \prod_{i=1}^m \prod_{\alpha=1}^n \frac{(z_{i\alpha})^{n_{i\alpha}}}{\sqrt{n_{i\alpha}!}} |\eta\rangle; \quad n_{i\alpha} = 0,1,2\ldots; \quad |\eta\rangle \in \mathscr{B}_{\overline{M}} \right\} \tag{4.1}$$

form an orthonormal basis for the VB space. The completion of this space is the well-known Bargmann–Segal Hilbert space of entire analytic functions introduced by Bargmann[11] as the carrier space for an irreducible representation of the mnth Heisenberg–Weyl algebra $\mathrm{hw}(mn)$,

$$\mathrm{hw}(mn) = \mathrm{span}\left\{ z_{i\alpha}, \nabla_{i\alpha} \equiv \frac{\partial}{\partial z_{i\alpha}}, \mathbf{1}; \quad 1 \leqslant i \leqslant m, \ 1 \leqslant \alpha \leqslant n \right\}, \tag{4.2}$$

defined by the commutation relations

$$[z_{i\alpha}, z_{j\beta}] = 0, \quad [\nabla_{i\alpha}, \nabla_{j\beta}] = 0, \quad [\nabla_{i\alpha}, z_{j\beta}] = \delta_{ij}\delta_{\alpha\beta}. \tag{4.3}$$

With respect to the inner product on Bg, we have

$$(z_{i\alpha})^\dagger = \nabla_{i\alpha}, \quad (\nabla_{i\alpha})^\dagger = z_{i\alpha}. \tag{4.4}$$

The Bargmann variables and their derivatives can be interpreted as boson annihilation and creation operators.

Levels can be defined on the VB basis (4.1) in terms of the eigenvalue n_z of the z-number operator $N_z = z_{i\mu}\nabla_{i\mu}$. There is an infinite number of such levels. We define the Z grade of a VB state by $g_{\max} - n_z$. This definition is consistent with the definition (2.2) for the grading operator and Eq. (4.11b) below.

The VCS embedding $M(\Lambda)\to\mathscr{H}_{\rm VB}$ is defined by

$$|\psi\rangle\to\psi(z)=\sum_{\eta}|\eta\rangle\langle\eta|\exp\mathscr{T}(z)|\psi\rangle, \tag{4.5a}$$

where

$$\mathscr{T}(z)=z\cdot A=\sum_{i\alpha}z_{i\alpha}A_{i\alpha}. \tag{4.5b}$$

This embedding invokes a projection $M(\Lambda)\to\overline{M}(\Lambda)$ in which $|\phi\rangle=\exp\mathscr{T}(z)|\psi\rangle$ projects to its highest grade component

$$|\phi\rangle\to\sum_{\eta}|\eta\rangle\langle\eta|\phi\rangle$$

and which, since $M(\Lambda)$ is a direct sum of graded subspaces, is well defined without the necessity of assuming that $M(\Lambda)$ is a Hilbert space. Because the variables $z_{i\alpha}$ belong to an algebra that is independent of the algebra $\mathrm{gl}(m+n)$, we have that all of the following commutators vanish:

$$[z_{i\alpha},X]=0,\quad \forall X\in\mathrm{gl}(m/n). \tag{4.6}$$

The VCS realization $\Gamma(X)$ of an arbitrary generator $X\in\mathrm{gl}(m+n)$ is defined by

$$\begin{aligned}\Gamma(X)\psi(z)&=\sum_{\eta}|\eta\rangle\langle\eta|\exp(\mathscr{T})X|\psi\rangle\\&=\sum_{\eta}|\eta\rangle\langle\eta|X+\frac{1}{1!}[\mathscr{T},X]\Big)\\&\quad+\frac{1}{2!}[\mathscr{T},[\mathscr{T},X]]+\cdots\Big)\exp(\mathscr{T})|\psi\rangle.\end{aligned} \tag{4.7}$$

The operator $\Gamma(X)$ can be expressed as a differential operator on $\psi(z)$. First note that

$$\langle\eta|B_{i\alpha}e^{z\cdot A}|\psi\rangle=0,\quad \forall B_{i\alpha}\in\mathbf{n}_{-}; \tag{4.8}$$

this is easily verified by considering the Z-graded structure of the highest weight module $M(\Lambda)$ with the understanding that states of different Z grade are orthogonal. We also have

$$\begin{aligned}\sum_{\eta}|\eta\rangle\langle\eta|C_{ij}e^{z\cdot A}|\psi\rangle&=\sum_{\eta}\mathbb{C}_{ij}|\eta\rangle\langle\eta|e^{z\cdot A}|\psi\rangle,\\\sum_{\eta}|\eta\rangle\langle\eta|C_{\alpha\beta}e^{z\cdot A}|\psi\rangle&=\sum_{\eta}\mathbb{C}_{\alpha\beta}|\eta\rangle\langle\eta|e^{z\cdot A}|\psi\rangle,\end{aligned} \tag{4.9}$$

where $\mathrm{span}\{\mathbb{C}_{ij}\}\oplus\mathrm{span}\{\mathbb{C}_{\alpha\beta}\}$ is the *intrinsic* $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ representation carried by the irreducible highest weight submodule $\overline{M}(\Lambda)$. Finally, we note that

$$\nabla_{i\alpha}e^{z\cdot A}=A_{i\alpha}e^{z\cdot A}=e^{z\cdot A}A_{i\alpha}. \tag{4.10}$$

Introducing these equalities in (4.7), we find, with the usual convention concerning the summation of repeated indices, the following VCS expansion for the generators of $\mathrm{gl}(m+n)$:

$$\Gamma(A_{i\alpha})=\nabla_{i\alpha}, \tag{4.11a}$$

$$\Gamma(C_{ij})=\mathbb{C}_{ij}-z_{j\mu}\nabla_{i\mu}, \tag{4.11b}$$

$$\Gamma(C_{\alpha\beta})=\mathbb{C}_{\alpha\beta}+z_{i\alpha}\nabla_{i\beta}, \tag{4.11c}$$

$$\Gamma(B_{i\alpha})=z_{k\alpha}\mathbb{C}_{ki}-z_{i\mu}\mathbb{C}_{\alpha\mu}-z_{k\alpha}z_{i\mu}\nabla_{k\mu}. \tag{4.11d}$$

Note that in the VCS realization, the $\mathrm{gl}(m)$ subalgebra consists of the piecewise sum by component of an intrinsic subalgebra ($\mathbb{C}_{ij}$) and a Bargmann realization ($-z_{j\mu}\nabla_{i\mu}$) of $\mathrm{gl}(m)$. Similarly, the $\mathrm{gl}(n)$ subalgebra consists of the sum of an intrinsic subalgebra ($\mathbb{C}_{\alpha\beta}$) and a Bargmann realization ($+z_{i\alpha}\nabla_{i\beta}$) of $\mathrm{gl}(n)$.

B. VCS theory for gl(*m*/*n*)

Conversely, the embedding of an irreducible graded $\mathrm{gl}(m/n)$ highest weight module $M(\Lambda)$ is in a vector-Grassman (VG) space. The VG space, which may be larger than $M(\Lambda)$, is the tensor product space $V\otimes G$, where the following conditions hold.

(i) V is the intrinsic $\mathbf{n}_0$ module described above.

(ii) G is the space of polynomials in the mn ($=\dim\mathbf{n}_{+1}$) anticommuting Grassman variables $\{\theta_{i\alpha};\ 1\leqslant i\leqslant m,\ 1\leqslant\alpha\leqslant n\}$;

$$[\theta_{i\alpha},\theta_{j\beta}]=0. \tag{4.12}$$

[We have assigned a Z_2 grade $\overline{1}$ to the variable $\theta_{i\alpha}$. The commutator in (4.12), interpreted as a graded commutator, thus stands for an anticommutator.] Eq. (4.12) implies

$$(\theta_{i\alpha})^2=0; \tag{4.13}$$

the Grassman space G is thus 2^{mn} dimensional. It is isomorphic to the antisymmetric (exterior) tensor algebra over $\mathscr{C}^{mn}$, and has a natural inner product for which the nonzero polynomials

$$\left\{\prod_{i=1}^{m}\prod_{\alpha=1}^{n}(\theta_{i\alpha})^{n_{i\alpha}}|\eta\rangle;\quad n_{i\alpha}=0,1;\quad |\eta\rangle\in\mathscr{D}_{\overline{M}}\right\} \tag{4.14}$$

form an orthonormal basis for the VG space. The space G carries an irreducible representation of the mnth Grassman algebra $\mathrm{Gr}(mn)$,

$$\mathrm{Gr}(mn)=\mathrm{span}\left\{\theta_{i\alpha},\partial_{i\alpha}\equiv\frac{\partial}{\partial\theta_{i\alpha}},\mathbf{1};\ 1\leqslant i\leqslant m,\quad 1\leqslant\alpha\leqslant n\right\}, \tag{4.15}$$

defined by the anticommutation relations

$$[\theta_{i\alpha},\theta_{j\beta}]=0,\quad [\partial_{i\alpha},\partial_{j\beta}]=0,\quad [\partial_{i\alpha},\theta_{j\beta}]=\delta_{ij}\delta_{\alpha\beta}. \tag{4.16}$$

With respect to the inner product on G, we have

$$(\theta_{i\alpha})^{\dagger}=\partial_{i\alpha},\quad (\partial_{i\alpha})^{\dagger}=\theta_{i\alpha}. \tag{4.17}$$

The Grassman variables and their derivatives can be interpreted as fermion annihilation and creation operators; G is thus isomorphic to a fermion Fock space. The dimension of the VG basis (4.14) is 2^{mn} times the dimension of the intrinsic $\mathbf{n}_0$ module V. Levels can be defined on this basis in terms of the eigenvalue n_θ of the θ-number operator $N_\theta=\theta_{i\mu}\partial_{i\mu}$. There are $mn+1$ such levels. The number of VG states on a given level is then given by $\binom{mn}{n_\theta}$ times the dimension of the intrinsic module.

We define the Z grade of a VG state by

$$g=g_{\max}-n_\theta. \tag{4.18}$$

This definition is consistent with definition (2.2) for the grading operator and Eq. (4.23b) below. Now, if the intrinsic space is assigned a Z_2 grade $\overline{0}$, consistency requires us to identify the even (odd) subspace $V_{\overline{0}}$ ($V_{\overline{1}}$) of the VG space with the set of all VG states with n_θ even (odd). The Z_2 grade σ of a state is then given by

$$\sigma = n_\theta \pmod 2. \tag{4.19a}$$

Conversely, if the intrinsic space is assigned a Z_2 grade $\overline{1}$, we identify the even (odd) subspace $V_{\overline{0}}$ ($V_{\overline{1}}$) of the VG space with the set of all VG states with n_θ odd (even). The Z_2 grade σ of a state is then given by

$$\sigma = n_\theta + 1 \pmod 2. \tag{4.19b}$$

The dimension of the even subspace $V_{\overline{0}}$ is equal to the dimension of the odd subspace $V_{\overline{1}}$ since

$$\sum_{n_\theta \text{ even}} \binom{mn}{n_\theta} = \sum_{n_\theta \text{ odd}} \binom{mn}{n_\theta}.$$

The VCS embedding $M(\Lambda) \to \mathscr{H}_{\text{VG}}$ is now defined by

$$|\psi\rangle \to \psi(\theta) = \sum_\eta |\eta\rangle\langle\eta|\exp \mathscr{T}(\theta)|\psi\rangle, \tag{4.20a}$$

where

$$\mathscr{T}(\theta) = \theta \cdot A = \sum_{i\alpha} \theta_{i\alpha} A_{i\alpha}. \tag{4.20b}$$

Once more, it is not necessary to assume that $M(\Lambda)$ is a Hilbert space but only that it is a Z-graded space; the projection $M(\Lambda) \to \overline{M}(\Lambda)$, defined by (4.20a), is thus meaningful. Because the variables $\theta_{i\alpha}$ belong to an algebra that is independent of the superalgebra $\mathrm{gl}(m/n)$, we have that all of the following (graded) commutators vanish:

$$[\theta_{i\alpha}, X] = 0, \quad \forall X \in \mathrm{gl}(m/n). \tag{4.21}$$

For example, the set $\{\theta_{i\alpha}\}$ of Grassman variables *anticommutes* with the set of raising and lowering operators $\{A_{i\alpha}\}$ and $\{B_{i\alpha}\}$, and (4.20a) should be developed accordingly.

The VCS realization $\Gamma(X)$ of an arbitrary generator $X \in \mathrm{gl}(m/n)$ is once more defined by

$$\begin{aligned}\Gamma(X)\psi(\theta) &= \sum_\eta |\eta\rangle\langle\eta|\exp(\mathscr{T})X|\psi\rangle \\ &= \sum_\eta |\eta\rangle\langle\eta|\Big(X + \frac{1}{1!}[\mathscr{T}, X] \\ &\quad + \frac{1}{2!}[\mathscr{T},[\mathscr{T}, X]] + \cdots\Big)\exp(\mathscr{T})|\psi\rangle,\end{aligned} \tag{4.22}$$

and, following a procedure similar to the $\mathrm{gl}(m/n)$ case, we find

$$\Gamma(A_{i\alpha}) = \partial_{i\alpha}, \tag{4.23a}$$

$$\Gamma(C_{ij}) = \mathbb{C}_{ij} - \theta_{j\mu}\,\partial_{i\mu}, \tag{4.23b}$$

$$\Gamma(C_{\alpha\beta}) = \mathbb{C}_{\alpha\beta} + \theta_{i\alpha}\,\partial_{i\beta}, \tag{4.23c}$$

$$\Gamma(B_{i\alpha}) = \theta_{k\alpha}\mathbb{C}_{ki} + \theta_{i\mu}\mathbb{C}_{\alpha\mu} - \theta_{k\alpha}\theta_{i\mu}\,\partial_{k\mu}. \tag{4.23d}$$

V. CONSTRUCTION OF $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$-COUPLED BASIS STATES FOR $\mathscr{H}$

A. Construction of $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$-coupled basis states for $\mathscr{H}_{\text{VB}}$

We now seek to construct a VB basis that [unlike the basis (4.1)] reduces the stability subalgebra $\mathrm{gl}(m) \oplus \mathrm{gl}(n) \subset \mathrm{gl}(m+n)$ and that, as a consequence, facilitates identification of its invariant VCS subspace.

We start by noting that, since

$$[\Gamma(C_{ij}), z_{k\sigma}] = -\delta_{ik} z_{j\sigma}, \tag{5.1a}$$

$$[\Gamma(C_{\alpha\beta}), z_{k\sigma}] = \delta_{\beta\sigma} z_{k\alpha}, \tag{5.1b}$$

the Bargmann variable $z_{i\alpha}$ transforms as the i component of a $\mathrm{gl}(m)$ tensor $\{-1\}$ and the α component of a $\mathrm{gl}(n)$ tensor $\{1\}$. The set $\{z_{i\alpha}\}$ thus transforms contragradiently to the set of raising operators $\{A_{i\alpha}\}$ which has a VCS realization given by the set of partial derivatives $\{\partial_{i\alpha}\}$ comprising a $\{1\}:\{-1\}$ $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ tensor.

The fully *symmetric* polynomials of degree n_z in the Bargmann variables $\{z_{i\alpha}\}$ span the set of irreps $\Sigma_\tau\{-\tau\}:\{\tau\}$ of $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$, where

$$\sum_{i=1}^{m} \tau_i = \sum_{\alpha=1}^{n} \tau_\alpha = n_z. \tag{5.2}$$

We shall denote these polynomials by

$$Z^{\{-\tau\}(m_{-\tau})}_{\{\tau\}(m_\tau)}(z), \tag{5.3}$$

where $(m_{-\tau})$ and (m_τ) stand for basis labels for the $\mathrm{gl}(m)$ and $\mathrm{gl}(n)$ irreps $\{-\tau\}$ and $\{\tau\}$, respectively. Assuming without loss of generality that $m \leqslant n$, the partition $\{\tau\}$ has m rows of length $\tau_\alpha \geqslant 0$ and, for a given $\{\tau\} = \{\tau_1\tau_2\cdots\tau_m\}$, $\{-\tau\}$ is given by $\{-\tau\} = \{-\tau_m, -\tau_{m-1}, \ldots, -\tau_1\}$.

A basis for the VB space, which reduces the stability subalgebra $\mathrm{gl}(m) \oplus \mathrm{gl}(n) \subset \mathrm{gl}(m+n)$, is then given by the (note) $U(m) \otimes U(n)$ coupling of the basis of Bargmann polynomials (5.3) with an orthonormal basis

$$\mathscr{B}_{\overline{M}} = \{|\eta\rangle \equiv |^{\{\mu^0\}(m_{\mu^0})}_{\{\nu^0\}(m_{\nu^0})}\rangle\} \tag{5.4}$$

for the intrinsic $\mathbf{n}_0$ module $\overline{M}(\Lambda)$ defined by Eq. (3.1). Such a basis is denoted

$$|^{\{\mu^0\}\{-\tau\}\{\mu\}\rho_\mu m_\mu}_{\{\nu^0\}\{\tau\}\{\nu\}\rho_\nu(m_\nu)}\rangle = [Z^{\{-\tau\}}_{\{\tau\}}(z) \times |^{\{\mu^0\}}_{\{\nu^0\}}\rangle]^{\{\mu\}\rho_\mu(m_\mu)}_{\{\nu\}\rho_\nu(m_\nu)}, \tag{5.5}$$

where the labels ρ_μ and ρ_ν denote multiplicity labels resolving any multiplicity that could possibly arise in the $U(m)$ and $U(n)$ couplings

$$\begin{aligned}\{\mu^0\} \times \{-\tau\} &\underset{\rho_\mu}{\to} \{\mu\}, \\ \{\nu^0\} \times \{\tau\} &\underset{\rho_\nu}{\to} \{\nu\}.\end{aligned} \tag{5.6}$$

B. Construction of $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$-coupled basis states for $\mathscr{H}_{\text{VG}}$

Our strategy is the same as for the $\mathrm{gl}(m+n)$ case. Again, we start by noting that, since

$$[\Gamma(C_{ij}), \theta_{k\sigma}] = -\delta_{ik}\theta_{j\sigma}, \tag{5.7a}$$

$$[\Gamma(C_{\alpha\beta}), \theta_{k\sigma}] = \delta_{\beta\sigma}\theta_{k\alpha}, \tag{5.7b}$$

the Grassman variable $\theta_{i\alpha}$ also transforms as the i component of a $\mathrm{gl}(m)$ tensor $\{-1\}$ and the α component of a $\mathrm{gl}(n)$ tensor $\{1\}$. The set $\{\theta_{i\alpha}\}$ thus transforms contragradiently to the set of raising operators $\{A_{i\alpha}\}$ which has a VCS realization given by the set of partial derivatives $\{\partial_{i\alpha}\}$ comprising a $\{1\}:\{-1\}$ $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ tensor.

The fully *antisymmetric* polynomials of degree n_θ in the Grassman variables $\{\theta_{i\alpha}\}$ span the set of irreps

$\Sigma_\tau\{-\tau\}:\{\tilde\tau\}$ of $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$, where, as a result of the antisymmetry of the polynomial, $\{\tilde\tau\}$ is the partition conjugate to $\{\tau\}$, and

$$\sum_{i=1}^{m}\tau_i=\sum_{\alpha=1}^{n}\tilde\tau_\alpha=n_\theta. \tag{5.8}$$

We shall denote these polynomials by

$$\Theta^{\{-\tau\}(m_{-\tau})}_{\{\tilde\tau\}(m_{\tilde\tau})}(\theta), \tag{5.9}$$

where $(m_{-\tau})$ and $(m_{\tilde\tau})$ stand for basis labels for the $\mathrm{gl}(m)$ and $\mathrm{gl}(n)$ irreps $\{-\tau\}$ and $\{\tilde\tau\}$, respectively. The n partition $\{\tilde\tau\}$ has rows of length $\tilde\tau_\alpha$ in the range $0\leqslant\tilde\tau_\alpha\leqslant m$, while the m partition $\{-\tau\}$ has rows of length τ_i in the range $0\leqslant\tau_i\leqslant n$.

A basis for the VG space, which reduces the stability subalgebra $\mathrm{gl}(m)\oplus\mathrm{gl}(n)\subset\mathrm{gl}(m/n)$, is then given by coupling the basis of Grassman polynomials (5.9) with the orthonormal basis $\{|\eta\rangle\}$ for the intrinsic $\mathbf{n}_0$ module $\overline{M}(\Lambda)$ defined by Eqs. (3.1) and (5.4). Such a basis is denoted

$$|^{\{\mu^0\}\{-\tau\}\{\mu\}\rho_\mu(m_\mu)}_{\{\nu^0\}\{\tilde\tau\}\{\nu\}\rho_\nu(m_\nu)}\rangle=[\Theta^{\{-\tau\}}_{\{\tilde\tau\}}(\theta)\times|^{\{\mu^0\}}_{\{\nu^0\}}\rangle]^{\{\mu\}\rho_\mu(m_\mu)}_{\{\nu\}\rho_\nu(m_\nu)}, \tag{5.10}$$

where, once more, the labels ρ_μ and ρ_ν denote multiplicity labels resolving any multiplicity that could possibly arise in the $U(m)$ and $U(n)$ couplings

$$\begin{aligned}&\{\mu^0\}\times\{-\tau\}\underset{\rho_\mu}{\to}\{\mu\},\\&\{\nu^0\}\times\{\tilde\tau\}\underset{\rho_\nu}{\to}\{\nu\}.\end{aligned} \tag{5.11}$$

For generic couplings [i.e., couplings for which the $\mathrm{U}(m)$ and $\mathrm{U}(n)$ multiplicities assume their maximal values], Eq. (5.10) defines a subset of 2^{mn} (highest weight) $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ modules.

VI. THE VCS Γ-MATRIX REPRESENTATIONS

A. VCS Γ-matrix representations for gl(m/n)

In this section, we argue that the VCS embedding $M(\Lambda)\to\mathscr{H}_{\mathrm{VG}}$ defines a VCS subspace of the VG space on which the VCS realization Γ of $\mathrm{gl}(m/n)$ acts *irreducibly*. This subspace is generated by the repeated action of the VCS operators $\{\Gamma(X);\ X\in\mathrm{gl}(m/n)\}$ on the highest grade VG states. However, one observes that the VCS operators have a well-defined action on the whole of the VG space and it is instructive to examine the structure of this (possibly reducible) *extended* $\overline{\Gamma}$ representation. It will be convenient to express it in matrix form with respect to the $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ coupled basis defined by Eq. (5.10).

Since we are interested in the superstructure of the algebra, it is useful to exploit the Wigner–Eckart theorem in order to compute $\mathbf{n}_0$-reduced matrix elements of the (VCS realization of the) $\mathbf{n}_0$ tensors $A^{\{1\}}_{\{-1\}}$ and $B^{\{-1\}}_{\{1\}}$ of the odd sector $\mathrm{gl}(m/n)_{\overline{1}}$ of the superalgebra. It is assumed that matrix elements of the generators of the Lie algebra $\mathbf{n}_0=\mathrm{gl}(m/n)_{\overline{0}}\simeq\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ are known[6,12]; the reduced matrix elements of the tensors A and B thus contain all the relevant information concerning the representation. Since we have not yet introduced adjoint relations for the superalgebra, we need to compute the $\mathbf{n}_0$-reduced matrix elements of A and B independently.

First, in order to facilitate the computation of $\mathbf{n}_0$-reduced matrix elements for B, note that Eq. (4.23d) can be rewritten

$$\Gamma(B_{i\alpha})=[\Omega,\theta_{i\alpha}], \tag{6.1a}$$

where the $\mathbf{n}_0$-invariant operator Ω is given by

$$\Omega=\tfrac{1}{4}(2I_{\mathrm{gl}(n)}-2I_{\mathrm{gl}(m)}-I^{(\theta)}_{\mathrm{gl}(n)}+I^{(\theta)}_{\mathrm{gl}(m)}+(m-n)N_\theta), \tag{6.1b}$$

and where

$$\begin{aligned}&I_{\mathrm{gl}(m)}=\Gamma(C_{ij})\Gamma(C_{ji}),\quad I^{(\theta)}_{\mathrm{gl}(m)}=(-\theta_{j\mu}\,\partial_{i\mu})(-\theta_{i\nu}\,\partial_{j\nu}),\\&I_{\mathrm{gl}(n)}=\Gamma(C_{\alpha\beta})\Gamma(C_{\beta\alpha}),\quad I^{(\theta)}_{\mathrm{gl}(n)}=(\theta_{i\alpha}\,\partial_{i\beta})(\theta_{j\beta}\,\partial_{j\alpha}),\end{aligned} \tag{6.1c}$$

are $\mathrm{gl}(m)$ and $\mathrm{gl}(n)$ quadratic Casimir invariants pertaining to the four independent (intrinsic and Grassmanian) algebras defined by the expansions (4.23c) and (4.23d). The $\mathbf{n}_0$-reduced matrix elements for the tensor B are then given by

$$\begin{aligned}&\langle^{\{\mu^0\}\{-\tau'\}\{\mu'\}\rho_{\mu'}}_{\{\nu^0\}\{\tilde\tau'\}\{\nu'\}\rho_{\nu'}}\|\Gamma(B^{\{-1\}}_{\{1\}})\|^{\{\mu^0\}\{-\tau\}\{\mu\}\rho_\mu}_{\{\nu^0\}\{\tilde\tau\}\{\nu\}\rho_\nu}\rangle\\&=\{\omega(^{\{\mu^0\}\{-\tau'\}\{\mu'\}}_{\{\nu^0\}\{\tilde\tau'\}\{\nu'\}})-\omega(^{\{\mu^0\}\{-\tau\}\{\mu\}}_{\{\nu^0\}\{\tilde\tau\}\{\nu\}})\}\times\begin{bmatrix}\{\mu^0\}&\{-\tau\}&\{\mu\}\rho_\mu\\\{0\}&\{-1\}&\{-1\}\\\{\mu^0\}&\{-\tau'\}&\{\mu'\}\rho_{\mu'}\end{bmatrix}\times\begin{bmatrix}\{\nu^0\}&\{\tilde\tau\}&\{\nu\}\rho_\mu\\\{0\}&\{1\}&\{1\}\\\{\nu^0\}&\{\tilde\tau'\}&\{\nu'\}\rho_\nu\end{bmatrix}\times(^{\{-\tau'\}}_{\{\tilde\tau'\}}\|\theta^{\{-1\}}_{\{1\}}\|^{\{-\tau\}}_{\{\tilde\tau\}}),\end{aligned} \tag{6.2}$$

where (1) a symbol like

$$\begin{bmatrix}\{\mu^0\}&\{-\tau\}&\{\mu\}\rho_\mu\\\{0\}&\{-1\}&\{-1\}\\\{\mu^0\}&\{-\tau'\}&\{\mu'\}\rho_{\mu'}\end{bmatrix} \tag{6.3}$$

stands for a unitary $9j$ recoupling coefficient [the appearance of the partition $\{0\}$ indicates that one can replace the unitary $9j$ symbol by a $6j$ symbol; we prefer the former notation since it makes readily apparent the various couplings, their order, and their assigned multiplicity labels], including the relevant multiplicity labels resolving the multiplicities appearing in the couplings (5.11);

(2) $(^{\{-\tau'\}}_{\{\tilde\tau'\}}\|\theta^{\{-1\}}_{\{1\}}\|^{\{-\tau\}}_{\{\tilde\tau\}})$

is the $\mathbf{n}_0$-reduced matrix element of the Grassman variables $\{\theta_{i\alpha}\}$ between two Grassman polynomials $\Theta(\theta)$ labeled, respectively, by $\{-\tau'\}:\{\tilde\tau'\}$ and $\{-\tau\}:\{\tilde\tau\}$ (explicitly computed in Appendix A); and

(3) $\omega\left(\begin{smallmatrix}\{\mu^0\}\{-\tau\}\{\mu\}\\ \{\nu^0\}\{\tilde{\tau}\}\{\nu\}\end{smallmatrix}\right)$

is the eigenvalue of the operator Ω on the VG state (5.10). [Since Ω is $\mathbf{n}_0$ invariant, its eigenvalues are independent of the basis labels (m_μ) and (m_ν). They are also, by construction, independent of the multiplicity labels ρ_μ, ρ_ν.] The $\mathbf{n}_0$-reduced matrix elements for the tensor A are more simply given by

$$\left\langle \begin{smallmatrix}\{\mu^0\}\{-\tau'\}\{\mu'\}\rho_{\mu'}\\ \{\nu^0\}\{\tilde{\tau}'\}\{\nu\}\rho_\nu\end{smallmatrix} \middle\| \Gamma\left(A\begin{Bmatrix}1\\-1\end{Bmatrix}\right) \middle\| \begin{smallmatrix}\{\mu^0\}\{-\tau\}\{\mu\}\rho_\mu\\ \{\nu^0\}\{\tilde{\tau}\}\{\nu\}\rho_\nu\end{smallmatrix} \right\rangle$$

$$= \begin{bmatrix}\{\mu^0\} & \{-\tau\} & \{\mu\}\rho_\mu\\ \{0\} & \{1\} & \{1\}\\ \{\mu^0\} & \{-\tau'\} & \{\mu'\}\rho_{\mu'}\end{bmatrix} \times \begin{bmatrix}\{\nu^0\} & \{\tilde{\tau}\} & \{\nu\}\rho_\nu\\ \{0\} & \{-1\} & \{-1\}\\ \{\nu^0\} & \{\tilde{\tau}'\} & \{\nu'\}\rho_{\nu'}\end{bmatrix} \times \left(\begin{Bmatrix}-\tau'\\ \tilde{\tau}'\end{Bmatrix} \middle\| \partial\begin{Bmatrix}1\\-1\end{Bmatrix} \middle\| \begin{Bmatrix}-\tau\\ \tilde{\tau}\end{Bmatrix}\right), \tag{6.4}$$

where

(4) $\left(\begin{Bmatrix}-\tau'\\ \tilde{\tau}'\end{Bmatrix} \middle\| \partial\begin{Bmatrix}1\\-1\end{Bmatrix} \middle\| \begin{Bmatrix}-\tau\\ \tilde{\tau}\end{Bmatrix}\right)$

is the $\mathbf{n}_0$-reduced matrix element of the partial derivatives $\{\partial_{i\alpha}\}$ (also computed in Appendix A).

Lowering down from the intrinsic $\mathbf{n}_0$-module, the VCS realization Γ of gl(m/n) generates an *irreducible* subspace of the VG space. If the representation is atypical,[8,13] it is a proper subspace of $\mathscr{H}_{\rm VG}$: its extended $\overline{\Gamma}$ representation (defined at the beginning of this section) is then indecomposable, i.e., it is reducible but not fully reducible. Its matrix representation is therefore of the form

$$\begin{pmatrix}\bullet & \bullet\\ \circ & \bullet\end{pmatrix}$$

as some of the states defined by (5.10) cannot be reached from the intrinsic module (as exemplified below). Meanwhile, the Γ representation itself consists only of the upper left matrix and is, as noted in the Introduction, always irreducible.

Before deriving explicitly the atypicality conditions, we find it useful to introduce some extra notation. Associated with the set of generators $\{E_{AB};\ 1\leqslant A,B\leqslant m+n\}$ of the superalgebra gl(m/n) is the set of roots $\{\pm(\epsilon_A-\epsilon_B);\ 1\leqslant A<B\leqslant m+n\}$, where

$$\epsilon_A(h_B)=\delta_{AB},\quad \epsilon_A\in\mathbf{h}^*,\quad h_B\in\mathbf{h}, \tag{6.5}$$

so that

$$[h,E_{AB}]=\epsilon_A(h)-\epsilon_B(h),\quad h\in\mathbf{h}. \tag{6.6}$$

The set of even roots is given by

$$\Delta_0=\left\{\begin{matrix}\pm(\epsilon_i-\epsilon_j), & 1\leqslant i<j\leqslant m\\ \pm(\delta_\alpha-\delta_\beta), & 1\leqslant\alpha<\beta\leqslant n\end{matrix}\right\},\quad \delta_\alpha=\epsilon_{m+\alpha}, \tag{6.7a}$$

and the odd roots by

$$\Delta_1=\{\pm(\epsilon_i-\delta_\alpha),\quad 1\leqslant i\leqslant m,\quad 1\leqslant\alpha\leqslant n\}. \tag{6.7b}$$

A supersymmetric invariant and nondegenerate bilinear form on $\mathbf{h}^*$ is given by [see Eq. (2.7)]

$$(\epsilon_A,\epsilon_B)=\sigma_A\delta_{AB}. \tag{6.8}$$

The linear form ρ, defined as half the sum of the positive even roots minus half the sum of the positive odd roots, is given by

$$\rho=\rho_0-\rho_1=\frac{1}{2}\sum_{i=1}^{m}(m-n+1-2i)\epsilon_i + \frac{1}{2}\sum_{\alpha=1}^{n}(m+n+1-2\alpha)\delta_\alpha. \tag{6.9}$$

The conditions for atypicality are easily derived in the present framework. Consider the matrix element of the $\mathbf{n}_0$-invariant operator T_-,

$$\begin{aligned}T_- &= \Gamma(B_{1n})\Gamma(B_{2n})\cdots\Gamma(B_{mn})\\ &\quad\times\Gamma(B_{1,n-1})\Gamma(B_{2,n-1})\cdots\Gamma(B_{m,n-1})\\ &\quad\times\cdots\\ &\quad\times\Gamma(B_{11})\Gamma(B_{21})\cdots\Gamma(B_{m1}) \qquad (6.10a)\\ &= [\Omega,\theta_{1n}]\times[\Omega,\theta_{2n}]\times\cdots\times[\Omega,\theta_{mn}]\\ &\quad\times[\Omega,\theta_{1,n-1}]\times[\Omega,\theta_{2,n-1}]\times\cdots\times[\Omega,\theta_{m,n-1}]\\ &\quad\times\cdots\\ &\quad\times[\Omega,\theta_{11}]\times[\Omega,\theta_{21}]\times\cdots\times[\Omega,\theta_{m1}], \qquad (6.10b)\end{aligned}$$

between the highest weight state $|\Lambda\rangle$ and the VG state $|\Lambda'\rangle$ defined by

$$|\Lambda'\rangle=\Theta\begin{Bmatrix}-n^m\\ m^n\end{Bmatrix}(\theta)|\Lambda\rangle.$$

The latter is the $\mathbf{n}_0$ highest weight state of the unique $\mathbf{n}_0$ module on the mnth θ level. The sequence of Grassman variables in (6.10b) determines a unique path in the weight space from Λ to Λ': it treks from highest-highest-weight state to highest-highest-weight state of the stability algebra as it successively goes up the various n_θ levels. The successive couplings are then fully stretched, the recoupling coefficients in (6.2) are all unity, and the only contribution of a given commutator in (6.10b) is given by the corresponding difference term on the right-hand side of (6.2) found to be given by the expression

$$\omega\left(\begin{smallmatrix}\{\mu^0\}\{-\tau'\}\{\mu'\}\\ \{\nu^0\}\{\tilde{\tau}'\}\{\nu'\}\end{smallmatrix}\right)-\omega\left(\begin{smallmatrix}\{\mu^0\}\{-\tau\}\{\mu\}\\ \{\nu^0\}\{\tilde{\tau}\}\{\nu\}\end{smallmatrix}\right) = \mu_l+m-l+\nu_\alpha-\alpha+\beta-\tilde{\tau}_\beta, \tag{6.11}$$

for $\mu'=\mu-\Delta(l)$, $\nu'=\nu+\Delta(\alpha)$, and $\tilde{\tau}'=\tilde{\tau}+\Delta(\beta)$ [where, e.g., $\Delta(l)$ is a null m vector except for its lth entry, which has value unity]. We then easily find, for

$$\Lambda=\sum_{i=1}^{m}\mu_i^0\epsilon_i+\sum_{\alpha=1}^{n}\nu_\alpha^0\delta_\alpha$$

[see also Eq. (3.3)],

$$\begin{aligned}\langle\Lambda'|T_-|\Lambda\rangle &= \prod_{k=1}^{m}\prod_{\beta=1}^{n}(\Lambda+\rho,\epsilon_k-\delta_\beta)\\ &= \prod_{k=1}^{m}\prod_{\beta=1}^{n}(\ell_{\mu_k^0}-\ell_{\nu_\beta^0}) \qquad (6.12)\end{aligned}$$

in terms of the quantities $\ell_{\mu_i^0}$ and $\ell_{\nu_\alpha^0}$ defined by [see Eq. (6.9)]

$$\ell_{\mu_i^0} = \mu_i^0 + \tfrac{1}{2}(m-n+1-2i),$$
$$-\ell_{\nu_\alpha^0} = \nu_\alpha^0 + \tfrac{1}{2}(m+n+1-2\alpha). \tag{6.13}$$

This matrix element vanishes if and only if the highest weight extended Γ representation under consideration is atypical.[8,13] We thus have atypicality whenever any of the factors in (6.12) vanishes, i.e., if

$$(\Lambda+\rho,\epsilon_k-\delta_\beta) = \ell_{\mu_k^0} - \ell_{\nu_\beta^0} = 0,$$
$$\text{for } 1\leqslant k\leqslant m, \quad 1\leqslant\beta\leqslant n. \tag{6.14}$$

The $\mathbf{n}_0$-invariant operator Ω, derived entirely within the VCS framework, thus conveniently summarizes the mn atypicality conditions (6.14).

It is interesting to identify which VG states can be reached from the intrinsic highest Z-grade module through the application of polynomials involving a single power of the lowering operators $\Gamma(B_{i\alpha})$. The possible $\mathbf{n}_0$ highest weight labels are then given by $\Lambda'' = \Lambda - (\epsilon_k - \delta_\beta)$, $1\leqslant k\leqslant m$, $1\leqslant\beta\leqslant n$, and each such $\mathbf{n}_0$ highest weight always appears in a multiplicity-free fashion. One then easily finds from (6.2) and (6.11) the reduced matrix elements

$$\langle\Lambda''\|\Gamma(B)\|\Lambda\rangle = (\Lambda+\rho,\epsilon_k-\delta_\beta). \tag{6.15}$$

Thus the $\mathbf{n}_0$ and consequently the $\mathrm{gl}(m/n)$ subrepresentations of atypical representations identified by the $\mathbf{n}_0$ highest weights $\Lambda = \Lambda - \epsilon_k + \delta_\beta$, such that $(\Lambda+\rho,\epsilon_k-\delta_\beta)=0$, decouple from the highest weight irreducible representation Λ of $\mathrm{gl}(m/n)$.

We note that the nonreducibility of atypical representations is not peculiar to superalgebras; extended $\overline{\Gamma}$ VCS representations of Lie algebras have a similar structure as we now illustrate.

B. Γ-matrix representations for gl(m+n)

As for $\mathrm{gl}(m/n)$, the VCS embedding $M(\Lambda)\to\mathscr{H}_{\mathrm{VB}}$ for $\mathrm{gl}(m+n)$ defines a VCS subspace of the VB space, generated by the repeated action of the VCS operators $\Gamma(X)$, $X\in\mathrm{gl}(m+n)$, on the highest grade VB states, on which the VCS realization Γ of $\mathrm{gl}(m+n)$ acts *irreducibly*.

We again exploit the Wigner–Eckart theorem in order to compute $\mathbf{n}_0$-reduced matrix elements of the (VCS realization of the) $\mathbf{n}_0$ tensors $A^{\{1\}}_{\{-1\}}$ and $B^{\{-1\}}_{\{1\}}$ composing the $\mathbf{n}_\pm$ nilpotent subalgebras of raising and lowering operators of the algebra $\mathrm{gl}(m+n)$. Similarly, in order to facilitate the computation of $\mathbf{n}_0$-reduced matrix elements for B, note that Eq. (4.11d) can be rewritten

$$\Gamma(B_{i\alpha}) = [\overline{\Omega}, z_{i\alpha}], \tag{6.16a}$$

where the $\mathbf{n}_0$-invariant operator $\overline{\Omega}$ is given by

$$\overline{\Omega} = \tfrac{1}{4}\{-2I_{\mathrm{gl}(m)} - 2I_{\mathrm{gl}(n)} + I^{(z)}_{\mathrm{gl}(m)} + I^{(z)}_{\mathrm{gl}(n)} + (m+n)N_z\}, \tag{6.16b}$$

where

$$I_{\mathrm{gl}(m)} = \Gamma(C_{ij})\Gamma(C_{ji}), \quad I^{(z)}_{\mathrm{gl}(m)} = (-z_{j\mu}\nabla_{i\mu})(-z_{i\nu}\nabla_{j\nu}),$$
$$I_{\mathrm{gl}(n)} = \Gamma(C_{\alpha\beta})\Gamma(C_{\beta\alpha}), \quad I^{(z)}_{\mathrm{gl}(n)} = (z_{i\alpha}\nabla_{i\beta})(z_{j\beta}\nabla_{j\alpha}). \tag{6.16c}$$

The $\mathbf{n}_0$-reduced matrix elements for the tensor B are given by

$$\langle {\{\mu^0\}\{-\tau'\}\{\mu'\}\rho_{\mu'} \atop \{\nu^0\}\{\tau'\}\{\nu'\}\rho_{\nu'}} \|\Gamma(B^{\{-1\}}_{\{1\}})\| {\{\mu^0\}\{-\tau'\}\{\mu\}\rho_\mu \atop \{\nu^0\}\{\tau\}\{\nu\}\rho_\mu}\rangle$$
$$= \{\overline{\omega}({\{\mu^0\}\{-\tau'\}\{\mu'\} \atop \{\nu^0\}\{\tau'\}\{\nu'\}}) - \overline{\omega}({\{\mu^0\}\{-\tau\}\{\mu\} \atop \{\nu^0\}\{\tau\}\{\nu\}})\}$$
$$\times \begin{bmatrix} \{\mu^0\} & \{-\tau\} & \{\mu\}\rho_\mu \\ \{0\} & \{-1\} & \{-1\} \\ \{\mu^0\} & \{-\tau'\} & \{\mu'\}\rho_{\mu'} \end{bmatrix} \times \begin{bmatrix} \{\nu^0\} & \{\tau\} & \{\nu\}\rho_\nu \\ \{0\} & \{1\} & \{1\} \\ \{\nu^0\} & \{\tau'\} & \{\nu'\}\rho_{\nu'} \end{bmatrix} \times ({\{-\tau'\} \atop \{\tau'\}} \| z^{\{-1\}}_{\{1\}} \| {\{-\tau\} \atop \{\tau\}}), \tag{6.17}$$

where

(1) $({\{-\tau'\} \atop \{\tau'\}} \| z^{\{-1\}}_{\{1\}} \| {\{-\tau\} \atop \{\tau\}})$

is the $\mathbf{n}_0$-reduced matrix element of the Bargmann variables $\{z_{i\alpha}\}$ between two Bargmann polynomials $Z(z)$ labeled, respectively, by $\{-\tau\}:\{\tau\}$ and $\{-\tau'\}:\{\tau'\}$ (given in Appendix B), and

(2) $\overline{\omega}({\{\mu^0\}\{-\tau\}\{\mu\} \atop \{\nu^0\}\{\tau\}\{\nu\}})$

is the eigenvalue of the operator $\overline{\Omega}$ on the VB state (5.5). The $\mathbf{n}_0$-reduced matrix elements for the tensor A are given by

$$\langle {\{\mu^0\}\{-\tau'\}\{\mu'\}\rho_{\mu'} \atop \{\nu^0\}\{\tau'\}\{\nu'\}\rho_{\nu'}} \|\Gamma(A^{\{1\}}_{\{-1\}})\| {\{\mu^0\}\{-\tau\}\{\mu\}\rho_\mu \atop \{\nu^0\}\{\tau\}\{\nu\}\rho_\mu}\rangle$$
$$= \begin{bmatrix} \{\mu^0\} & \{-\tau\} & \{\mu\}\rho_\mu \\ \{0\} & \{1\} & \{1\} \\ \{\mu^0\} & \{-\tau'\} & \{\mu'\}\rho_{\mu'} \end{bmatrix} \times \begin{bmatrix} \{\nu^0\} & \{\tau\} & \{\nu\}\rho_\nu \\ \{0\} & \{-1\} & \{-1\} \\ \{\nu^0\} & \{\tau'\} & \{\nu'\}\rho_{\nu'} \end{bmatrix} \times ({\{-\tau'\} \atop \{\tau'\}} \| \nabla^{\{1\}}_{\{-1\}} \| {\{-\tau\} \atop \{\tau\}}), \tag{6.18}$$

where

(3) $({\{-\tau'\} \atop \{\tau'\}} \| \nabla^{\{1\}}_{\{-1\}} \| {\{-\tau\} \atop \{\tau\}})$ (6.19)

is the $\mathbf{n}_0$-reduced matrix element of the partial derivatives $\{\nabla_{i\alpha}\}$ (also given in Appendix B).

C. A simple example: The gl(2) and gl(1/1) complex algebras

In this subsection, we show through a simple example, namely, the VCS construction of highest weight ladder representations of the (complex) Lie algebra $gl(2) \supset gl(1) \oplus gl(1)$, that the nonreducibility of highest weight atypical representations is not a phenomena peculiar to superalgebras: more specifically, we show that extended $\overline{\Gamma}$ representations of Lie algebraic structures usually present the same peculiarity. There is nevertheless a pertinent distinction to be made between the VCS theory for Lie algebras and superalgebras. For the former, basis states are vector-valued polynomials of Bargmann (complex) variables while, in the latter, they are polynomials of Grassman variables. As a consequence, the vector-Bargmann (VB) representation space is always infinite dimensional while the vector-Grassman (VG) representation space is finite dimensional whenever V is finite.

Setting $m = n = 1$ in the results of the preceding subsections on $gl(m+n)$, we find the following VCS representation for the complex Lie algebra gl(2):

$$\begin{aligned} &\Gamma(C_{11}) = \mu^0 - z\nabla, \quad \Gamma(C_{22}) = \nu^0 + z\nabla, \\ &\Gamma(B_{12}) = (\mu^0 - \nu^0)z - z^2\nabla, \quad \Gamma(A_{12}) = \nabla. \end{aligned} \tag{6.20}$$

Defining the VB basis states

$$|n\rangle = \frac{z^n}{\sqrt{n!}}|0\rangle \equiv \left\langle z \middle| \begin{matrix} \{\mu^0\} & \{\mu\} \\ \{\nu^0\} & \{\nu\} \end{matrix} \right\rangle,$$

where

$$|0\rangle = \left| \begin{matrix} \{\mu^0\} & \{\mu^0\} \\ \{\nu^0\} & \{\nu^0\} \end{matrix} \right\rangle$$

is the highest weight state, we obtain the matrix elements

$$\begin{aligned} &\langle n|\Gamma(C_{11})|n\rangle = \mu = \mu^0 - n, \\ &\langle n|\Gamma(C_{22})|n\rangle = \nu = \nu^0 + n, \\ &\langle n+1|\Gamma(B_{12})|n\rangle = (\mu^0 - \nu^0 - n)\sqrt{n+1}, \\ &\langle n-1|\Gamma(A_{12})|n\rangle = \sqrt{n}. \end{aligned} \tag{6.21}$$

We thus find the following *infinite-dimensional extended* $\overline{\Gamma}$-matrix representations for the raising and lowering operators A_{12} and B_{12} of $gl(2) \supset gl(1) \oplus gl(1)$:

(1) for $\mu^0 - \nu^0 = 2J \geqslant 0$,

$$\overline{\Gamma}(A_{12}) = \begin{pmatrix} 0 & \sqrt{1} & 0 & 0 & 0 & \\ 0 & 0 & \sqrt{2} & 0 & 0 & \\ 0 & 0 & 0 & \sqrt{3} & 0 & \\ 0 & 0 & 0 & 0 & \sqrt{4} & \\ 0 & 0 & 0 & 0 & 0 & \\ & & & & & \ddots \end{pmatrix},$$

$$\overline{\Gamma}(B_{12}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \\ 2J\cdot\sqrt{1} & 0 & 0 & 0 & 0 & \\ 0 & (2J-1)\cdot\sqrt{2} & 0 & 0 & 0 & \\ 0 & 0 & (2J-2)\cdot\sqrt{3} & 0 & 0 & \\ 0 & 0 & 0 & (2J-3)\cdot\sqrt{4} & 0 & \\ & & & & & \ddots \end{pmatrix}, \tag{6.22a}$$

(2) for $\mu^0 - \nu^0 = -2\mathscr{J} \leqslant 0$,

$$\overline{\Gamma}(A_{12}) = \begin{pmatrix} 0 & \sqrt{1} & 0 & 0 & 0 & \\ 0 & 0 & \sqrt{2} & 0 & 0 & \\ 0 & 0 & 0 & \sqrt{3} & 0 & \\ 0 & 0 & 0 & 0 & \sqrt{4} & \\ 0 & 0 & 0 & 0 & 0 & \\ & & & & & \ddots \end{pmatrix}, \tag{6.22b}$$

$$\overline{\Gamma}(B_{12}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \\ -2\mathscr{J}\cdot\sqrt{1} & 0 & 0 & 0 & 0 & \\ 0 & -(2\mathscr{J}+1)\cdot\sqrt{2} & 0 & 0 & 0 & \\ 0 & 0 & -(2\mathscr{J}+2)\cdot\sqrt{3} & 0 & 0 & \\ 0 & 0 & 0 & -(2\mathscr{J}+3)\cdot\sqrt{4} & 0 & \\ & & & & & \ddots \end{pmatrix},$$

in an increasing n-ordered basis.

It is easily seen that for $2J$ a positive integer the representation (6.22a) is reducible but not fully reducible, i.e., it has the matrix structure

$$\overline{\Gamma}\sim\begin{pmatrix} X & Y \\ O & Z \end{pmatrix}, \tag{6.23}$$

where the invariant subspace X is given by the upper-left $(2J+1)\times(2J+1)$ matrix. It carries a finite-dimensional ladder representation of gl(2) and is shown in the next section to be equivalent to a finite-dimensional unitary irreducible ladder representation (unirrep) of the compact real form u(2) of gl(2). We note that the submatrix Z carries an infinite-dimensional representation of gl(2) which shares with X the same values for the gl(2) Casimir operators. This representation is obviously not equivalent to a unitary representation of the real compact form u(2). These results though do not nullify Weyl's theorem concerning the complete reducibility of representations of compact Lie algebras since the theorem applies only to the finite-dimensional representations of these algebras. The representation (6.22b) is a highest weight infinite-dimensional irreducible representation of gl(2) and is shown in the next section to be equivalent to an infinite-dimensional unitary representation of the noncompact real form u(1,1) when $2\mathscr{J}$ is a positive integer.

Similar results hold for the generic gl$(m+n)$ case: the various reduced matrix elements for the raising operators $\Gamma(A)$ never vanish, but the same does not hold true for reduced matrix elements of $\Gamma(B)$ when starting from a *highest* grade $\mathbf{n}_0$ module since some states (5.5) (or linear combinations thereof) may not be accessible. For example, the difference term

$$\begin{aligned}&\overline{\omega}\left(\begin{smallmatrix}\{\mu^0\}\{-\tau'\}\{\mu'\}\\ \{\nu^0\}\{\tau'\}\{\nu'\}\end{smallmatrix}\right)-\overline{\omega}\left(\begin{smallmatrix}\{\mu^0\}\{-\tau\}\{\mu\}\\ \{\nu^0\}\{\tau\}\{\nu\}\end{smallmatrix}\right)\\ &\quad=\mu_l+m-l+\tau_\beta-\beta-\nu_\alpha+\alpha,\end{aligned} \tag{6.24}$$

for $\mu'=\mu-\Delta(l)$, $\nu'=\nu+\Delta(\alpha)$, and $\tau'=\tau+\Delta(\beta)$ in (6.17) may vanish for some simple $\mathbf{n}_0$ submodules $\{\mu'\}:\{\nu'\}$ in $\{\mu^0\}$. [A $\mathbf{n}_0$ submodule $\{\mu\}:\{\nu\}$ is called simple if no multiplicity occurs in the couplings (5.5) or (5.10) *and* if the specification of the partition $\{\mu\}:\{\nu\}$ uniquely determines the partition $\{-\tau\}:\{\tau\}$ or $\{-\tau\}:\{\tilde{\tau}\}$. For nonsimple $\mathbf{n}_0$ submodules, it is usually specific linear combinations of VB or VG states that may be inaccessible.] Whenever such an inaccessibility occurs, the (necessarily infinite-dimensional) extended $\overline{\Gamma}$ representation for gl$(m+n)$ is of the form

$$\begin{pmatrix} \bullet & \bullet \\ \circ & \bullet \end{pmatrix},$$

i.e., it is not fully reducible, and some of the states defined by (5.5) cannot be reached from the intrinsic module. Obviously, an irreducible finite-dimensional highest weight representation of a compact real Lie algebra $\mathbf{g}$ always extends to a non-fully-reducible infinite-dimensional $\overline{\Gamma}$ representation of the complexification of $\mathbf{g}$.

Finally, for comparison, we give the VCS representation for the Lie algebra gl(1/1):

$$\begin{aligned}&\Gamma(C_{11})=\nu^0-\theta\,\partial\\ &\Gamma(C_{22})=\mu^0+\theta\,\partial,\\ &\Gamma(B_{12})=(\mu^0+\nu^0)\theta,\\ &\Gamma(A_{12})=\partial\end{aligned} \tag{6.25}$$

[note that there is no quadratic term in θ in the expression for $\Gamma(B_{12})$]. Defining the VG basis states

$$|n\rangle=\theta^n|0\rangle\equiv\langle\theta|\begin{smallmatrix}\{\mu^0\}\{\mu\}\\ \{\nu^0\}\{\nu\}\end{smallmatrix}\rangle,\quad n=0,1,$$

we obtain the matrix elements

$$\begin{aligned}&\langle n|\Gamma(C_{11})|n\rangle=\mu=\mu^0-n,\\ &\langle n|\Gamma(C_{22})|n\rangle=\nu=\nu^0+n,\\ &\langle 1|\Gamma(B_{12})|0\rangle=(\mu^0+\nu^0),\\ &\langle 0|\Gamma(A_{12})|1\rangle=1.\end{aligned} \tag{6.26}$$

We thus find, for $\mathscr{I}=\mu^0+\nu^0$, the following two-dimensional extended $\overline{\Gamma}$-matrix representation

$$\begin{aligned}&\overline{\Gamma}(A_{12})=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\\ &\overline{\Gamma}(B_{12})=\begin{pmatrix}0 & 0\\ \mathscr{I} & 0\end{pmatrix},\end{aligned} \tag{6.27}$$

in an increasing n-ordered basis for the raising and lowering operators A_{12} and B_{12} of gl$(1/1)\supset$gl$(1)\oplus$gl(1). The representation is atypical for $\mathscr{I}=0$.

VII. VCS INNER PRODUCTS

To every weight λ, there exists, for a complex semisimple Lie algebra $\mathbf{g}$, an irreducible $\mathbf{g}$ module unique up to isomorphism and containing a highest weight vector of weight λ [14]. If $\mathbf{g}^{\mathfrak{R}}$ is a real compact form of $\mathbf{g}$ and the $\mathbf{g}$ module is finite dimensional, an inner product is known to exist on the $\mathbf{g}$ module for which the corresponding representation is Hermitian. Recall that if $g=e^{iX}$ is an element of a Lie group with infinitesimal generator $X\in\mathbf{g}^{\mathfrak{R}}$, then unitarity of the representation γ on the group implies Hermiticity of the representation on the algebra, i.e.,

$$\gamma(g^{-1})=\gamma^\dagger(g)\Rightarrow\gamma(X)=\gamma^\dagger(X).$$

It has been shown herein (see also Refs. 1–6) how VCS theory gives a realization of irreducible highest weight $\mathbf{g}$ modules of reductive complex Lie algebras as invariant subspaces of VB spaces. It has also been shown[1–3] how K-matrix theory enables one to evaluate inner products. Inner products for discrete series representations can be defined in integral form in terms of invariant measures on coset spaces.[2] Unfortunately, the integral form is in most cases unknown and/or difficult to evaluate. The K-matrix theory does not suffer such limitations as it defines the inner products in terms of the simple and well-defined VB inner product. We show here that the theory also applies (with minor modifications) to classical Lie superalgebras.

A. K-matrix theory for gl$(m+n)$ and gl(m/n)

In selecting a basis for a complex Lie algebra $\mathbf{g}$, it is customary and convenient to choose bases of raising $\{A_\nu\}$ and lowering $\{B_\nu\}$ operators such that a representation γ of $\mathbf{g}$ will be Hermitian on restriction to some particular real

form $\mathfrak{g}^{\mathfrak{R}}$ of $\mathfrak{g}$ when the Hermitian adjoint relations $\gamma(B_\nu) = \gamma(A_\nu)^\dagger$ are satisfied. For example, the basis given in Sec. II for $\mathrm{gl}(m+n)$ is such that a representation γ of $\mathrm{gl}(m+n)$ is Hermitian on restriction to the compact real form $\mathrm{u}(m+n)$ if

$$\gamma^\dagger(A_{i\alpha}) = \gamma(B_{i\alpha}). \tag{7.1a}$$

However, one is aware that representations that are not Hermitian for one real form may nevertheless be Hermitian for some other for which different Hermitian adjoint relations are satisfied. For example, in the given $\mathrm{gl}(m+n)$ basis, a representation γ is Hermitian on restriction to the noncompact real form $\mathrm{u}(m,n)$ if

$$\gamma^\dagger(A_{i\alpha}) = -\gamma(B_{i\alpha}). \tag{7.1b}$$

To determine if a given (irreducible) VCS representation Γ of $\mathrm{gl}(m+n)$ is equivalent to a Hermitian representation on restriction to $\mathrm{u}(m+n)$ or $\mathrm{u}(m,n)$, we inquire if there exists a similarity transformation K: $\Gamma \to \gamma = K^{-1}\Gamma K$, for which

$$\begin{aligned}\gamma(B_{i\alpha}) &= K^{-1}\Gamma(B_{i\alpha})K = \pm (K^{-1}\Gamma(A_{i\alpha})K)^\dagger \\ &= \pm K^\dagger\Gamma^\dagger(A_{i\alpha})K^{-1\dagger} = \pm\gamma^\dagger(A_{i\alpha}),\end{aligned} \tag{7.2a}$$

i.e., for which

$$KK^\dagger\Gamma^\dagger(A_{i\alpha}) = \epsilon\Gamma(B_{i\alpha})KK^\dagger, \tag{7.2b}$$

where $\epsilon = \pm 1$, where the Hermitian adjoints in (7.2) are understood to be defined with respect to the simple VB measure, and where the upper (lower) sign refers to $\mathrm{u}(m+n)$ [$\mathrm{u}(m,n)$]. Since $KK^\dagger$ is a positive definite operator, there can be only one solution (to within equivalence) to Eq. (7.2), i.e., the representation Γ is equivalent to a Hermitian representation of either real form of $\mathrm{gl}(m+n)$ but not both simultaneously.

It can be shown[3] that the VCS representation Γ is, by construction, Hermitian on restriction to the stability algebra $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$; it is then convenient to require that K commutes with the VCS representation of the stability algebra, i.e.,

$$\Gamma(X)K = K\Gamma(X), \quad \forall X \in \mathrm{gl}(m) \oplus \mathrm{gl}(n), \tag{7.3}$$

Having determined the existence of a similarity transformation K that satisfies the above equations (7.2) and (7.3), the equivalent representation γ, defined by

$$\gamma(X) = K^{-1}\Gamma(X)K, \quad \forall X \in \mathrm{gl}(m+n), \tag{7.4}$$

is observed to be explicitly Hermitian on restriction to $\mathrm{u}(m+n)$ if $\epsilon = +1$, and on restriction to $\mathrm{u}(m,n)$ if $\epsilon = -1$, in the solution to (7.2b).

Now, for a representation γ of a classical Lie superalgebra $\mathfrak{g}$ on a Hilbert space, one can define at least two types of adjoint operation.[15] A star adjoint $\gamma^\dagger(X)$ of the operator $\gamma(X)$ for $X \in \mathfrak{g}$ is defined by the usual Hermitian adjoint rule

$$\langle \gamma(X)x|y\rangle = \langle x|\gamma^\dagger(X)y\rangle.$$

Similarly, a grade star adjoint is defined by

$$\langle \gamma(X)x|y\rangle = (-1)^{\sigma(X)\cdot\sigma(x)}\langle x|\gamma^\ddagger(X)y\rangle,$$

where $\sigma(X)$ is the Z_2 grade of the element $X \in \mathfrak{g}$ and $\sigma(y)$ is the Z_2 grade of the state $|y\rangle$. A representation γ of a Lie classical superalgebra $\mathfrak{g}$ is then said to be a star representation if, for every $X \in \mathfrak{g}$, there is some $Y_X \in \mathfrak{g}$ for which

$$\gamma^\dagger(X) = \pm\gamma(Y_X).$$

For $\mathrm{gl}(m/n)$, we have two possibilities (up to equivalence) for the star adjoint operation:

$$\gamma^\dagger(A_{i\alpha}) = \pm\gamma(B_{i\alpha}), \quad \gamma^\dagger(B_{i\alpha}) = \pm\gamma(A_{i\alpha}). \tag{7.5}$$

Thus a star representation of a classical Lie superalgebra corresponds to a Hermitian representation of a standard Lie algebra. Similarly, a representation γ of a Lie classical superalgebra $\mathfrak{g}$ is said to be a grade star representation if, for every $X \in \mathfrak{g}$, there is some $Z_X \in \mathfrak{g}$ for which

$$\gamma^\ddagger(X) = \pm\gamma(Z_X), \quad \gamma^\ddagger(Z_X) = \mp\gamma(X).$$

Again, for $\mathrm{gl}(m/n)$, we have two possibilities (up to equivalence) for the grade star adjoint operation:

$$\gamma^\ddagger(A_{i\alpha}) = \pm\gamma(B_{i\alpha}), \quad \gamma^\ddagger(B_{i\alpha}) = \mp\gamma(A_{i\alpha}). \tag{7.6}$$

Note that it may be possible for a representation to be equivalent to both a star and grade star representation simultaneously. Representations may also exist that are neither star nor grade star.

We seek to determine which representations of $\mathrm{gl}(m/n)$ are equivalent to a star representation or a grade star representation. We thus seek solutions for a generalization of Eq. (7.2), namely,

$$\langle x|KK^\dagger\Gamma^\dagger(A_{i\alpha})|y\rangle = \epsilon\langle x|\Gamma(B_{i\alpha})KK^\dagger|y\rangle, \tag{7.7}$$

where ϵ takes values ± 1 for the $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ basis and is determined by the (critical) requirement that $KK^\dagger$ be positive semidefinite (otherwise K would not be equivalent to a similarity transformation). It is possible that such a solution does not exist for a given highest weight. Conversely, if a solution does exist, one can then verify that the representation $\gamma = K^{-1}\Gamma K$ is explicitly a star representation if and only if ϵ takes the constant value $+1$ or -1 for all matrix elements (7.7). Similarly, the representation is equivalent to a grade star representation if and only if ϵ is given by $-(-1)^{\sigma(x)}$ or $(-1)^{\sigma(x)}$ for all matrix elements (7.7). Otherwise, it is neither a star nor a grade star representation. The algorithm defined by (7.7) thus provides us with a simple mean to uncover (parametric) conditions for which a given representation $\{\mu^0\}:\{\nu^0\}$ can be declared equivalent to a star or grade star representation. We illustrate this point in Sec. VIII.

The explicit prescription for the computation of the (semipositive definite Hermitian) operator $\kappa^2 = KK^\dagger$ for $\mathrm{gl}(m/n)$ is the same as for Lie algebras with Abelian nilpotent algebra of raising operators.[1,3] From (7.7), we have

$$\langle x|\kappa^2\theta_{i\alpha}|y\rangle = \epsilon\langle x|\Gamma(B_{\alpha i}\kappa^2|y\rangle. \tag{7.8}$$

Substituting Eq. (6.1a) in (7.8), we thus have the equality

$$\langle x|\kappa^2\theta_{i\alpha}|y\rangle = \epsilon\langle x|[\Omega,\theta_{\alpha i}]\kappa^2|y\rangle \tag{7.9}$$

yielding

$$\begin{aligned}\langle x|\kappa^2 N_\theta|z\rangle &= \langle x|\kappa^2\theta_{\alpha i}\,\partial_{\alpha i}|z\rangle \\ &= \sum_y \epsilon\langle x|[\Omega,\theta_{\alpha i}]\kappa^2|y\rangle\langle y|\partial_{\alpha i}|z\rangle,\end{aligned} \tag{7.10}$$

which is the desired recursion formula for κ^2. For a generic problem, κ^2 is a matrix, diagonal in $\mathfrak{n}_0 = \mathrm{gl}(m) \oplus \mathrm{gl}(n)$, and of dimension given by the multiplicity of the $\mathrm{gl}(m/n) \downarrow \mathrm{gl}(m) \oplus \mathrm{gl}(n)$ branching. For simplicity, we write

$$\left\langle \begin{matrix} \{\mu^0\}\{-\tau'\}\{\mu'\}\rho'_{\mu'}(m_{\mu'}) \\ \{\nu^0\}\{\bar{\tau}'\}\{\nu'\}\rho'_{\nu'}(m_{\nu'}) \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{-\tau\}\{\mu\}\rho_{\mu}(m_{\mu}) \\ \{\nu^0\}\{\bar{\tau}\}\{\nu\}\rho_{\nu}(m_{\nu}) \end{matrix} \right\rangle = \delta_{\{\mu'\}\{\mu\}} \delta_{(m_{\mu'})(m_{\mu})} \delta_{\{\nu'\}\{\nu\}} \delta_{(m_{\nu'})(m_{\nu})} \left\langle \begin{matrix} \{\mu^0\}\{-\tau'\}\{\mu\}\rho'_{\mu} \\ \{\nu^0\}\{\bar{\tau}'\}\{\nu\}\rho'_{\nu} \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{-\tau\}\{\mu\}\rho_{\mu} \\ \{\nu^0\}\{\bar{\tau}\}\{\nu\}\rho_{\nu} \end{matrix} \right\rangle. \tag{7.11}$$

From (7.10), we easily derive the recursion formula

$$\begin{aligned} \left\langle \begin{matrix} \{\mu^0\}\{-\tau'\}\{\mu\}\rho'_{\mu} \\ \{\nu^0\}\{\bar{\tau}'\}\{\nu\}\rho'_{\nu} \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{-\tau\}\{\mu\}\rho_{\mu} \\ \{\nu^0\}\{\bar{\tau}\}\{\nu\}\rho_{\nu} \end{matrix} \right\rangle = {} & \frac{\epsilon}{n_\theta\left(\begin{smallmatrix}\{\mu\}\\\{\nu\}\end{smallmatrix}\right)} \times \sum_{\substack{\mu^{(a)}\tau^{(a)}\tau^{(b)}\rho^{(a)}\rho^{(b)} \\ \nu^{(a)}\rho^{(a)}\rho^{(\beta)}}} \left\{ \left\langle \begin{matrix} \{\mu^0\}\{-\tau'\}\{\mu\}\rho'_{\mu} \\ \{\nu^0\}\{\bar{\tau}'\}\{\nu\}\rho'_{\nu} \end{matrix} \middle\| \theta \middle\| \begin{matrix} \{\mu^0\}\{-\tau^{(a)}\}\{\mu^{(a)}\}\rho^{(a)} \\ \{\nu^0\}\{\bar{\tau}^{(a)}\}\{\nu^{(a)}\}\rho^{(a)} \end{matrix} \right\rangle \right. \\ & \times \left[\omega\left(\begin{smallmatrix}\{\mu^0\}\{-\tau'\}\{\mu\}\\\{\nu^0\}\{\bar{\tau}'\}\{\nu\}\end{smallmatrix}\right) - \omega\left(\begin{smallmatrix}\{\mu^0\}\{-\tau^{(a)}\}\{\mu^{(a)}\}\\\{\nu^0\}\{\bar{\tau}^{(a)}\}\{\nu^{(a)}\}\end{smallmatrix}\right) \right] \times \left\langle \begin{matrix} \{\mu^0\}\{-\tau^{(a)}\}\{\mu^{(a)}\}\rho^{(a)} \\ \{\nu^0\}\{\bar{\tau}^{(a)}\}\{\nu^{(a)}\}\rho^{(a)} \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{-\tau^{(b)}\}\{\mu^{(a)}\}\rho^{(b)} \\ \{\nu^0\}\{\bar{\tau}^{(b)}\}\{\nu^{(a)}\}\rho^{(\beta)} \end{matrix} \right\rangle \\ & \left. \times \left\langle \begin{matrix} \{\mu^0\}\{-\tau\}\{\mu\}\rho_{\mu} \\ \{\nu^0\}\{\bar{\tau}\}\{\nu\}\rho_{\nu} \end{matrix} \middle\| \theta \middle\| \begin{matrix} \{\mu^0\}\{-\tau^{(b)}\}\{\mu^{(a)}\}\rho^{(b)} \\ \{\nu^0\}\{\bar{\tau}^{(b)}\}\{\nu^{(a)}\}\rho^{(\beta)} \end{matrix} \right\rangle \right\}, \end{aligned} \tag{7.12a}$$

for κ^2, where

$$n_\theta\left(\begin{smallmatrix}\{\mu\}\\\{\nu\}\end{smallmatrix}\right) = \left\langle \begin{matrix} \{\mu^0\}\{-\tau\}\{\mu\}\rho_{\mu} \\ \{\nu^0\}\{\bar{\tau}\}\{\nu\}\rho_{\nu} \end{matrix} \middle\| N_\theta \middle\| \begin{matrix} \{\mu^0\}\{-\tau\}\{\mu\}\rho_{\mu} \\ \{\nu^0\}\{\bar{\tau}\}\{\nu\}\rho_{\nu} \end{matrix} \right\rangle = \sum_{i=1}^{m} (\mu_i^0 - \mu_i) = \sum_{i=1}^{m} \tau_i = \sum_{\alpha=1}^{n} (\nu_\alpha - \nu_\alpha^0) = \sum_{\alpha=1}^{n} \bar{\tau}_\alpha. \tag{7.12b}$$

Given a basis for the intrinsic $\mathfrak{n}_0$ module that is properly orthonormalized, the one-dimensional matrix

$$\left\langle \begin{matrix} \{\mu^0\}\{0\}\{\mu^0\} \\ \{\nu^0\}\{0\}\{\nu^0\} \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{0\}\{\mu^0\} \\ \{\nu^0\}\{0\}\{\nu^0\} \end{matrix} \right\rangle = 1 \tag{7.12c}$$

provides the starting point for the recursion process.

Although Eq. (7.12a) may appear quite formidable, it is usually quite straightforward in practice. Furthermore, for simple $\{\mu\}:\{\nu\}$ and $\{\mu'\}:\{\nu'\}$ $\mathfrak{n}_0$ modules (such modules have been defined in Sec. VI A), Eq. (7.9) yields directly the much simpler recursion formula

$$\begin{aligned} &\left\langle \begin{matrix} \{\mu^0\}\{\mu'\} \\ \{\nu^0\}\{\nu'\} \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{\mu'\} \\ \{\nu^0\}\{\nu'\} \end{matrix} \right\rangle \\ &\quad = \epsilon \{ \omega\left(\begin{smallmatrix}\{\mu^0\}\{m'\}\\\{\nu^0\}\{\nu'\}\end{smallmatrix}\right) - \omega\left(\begin{smallmatrix}\{\mu^0\}\{\mu\}\\\{\nu^0\}\{\nu\}\end{smallmatrix}\right) \} \left\langle \begin{matrix} \{\mu^0\}\{\mu\} \\ \{\nu^0\}\{\nu\} \end{matrix} \middle| \kappa^2 \middle| \begin{matrix} \{\mu^0\}\{\mu\} \\ \{\nu^0\}\{\nu\} \end{matrix} \right\rangle. \end{aligned} \tag{7.13}$$

The VG spaces for representations of the superalgebras $\mathrm{gl}(1/n)$ or $\mathrm{gl}(m/1)$ are composed entirely of simple states: VCS computations are therefore straightforward for these cases as we illustrate in Sec. VIII.

Finally, it is worth mentioning that the κ^2 matrices can be shown[16] to be related in a simple manner to the inner products (overlaps) of states belonging to a given $\{\mu\}:\{\nu\}$ multiplicity set in a representation of the superalgebra $\mathrm{gl}(m/n)$.

B. VCS inner products for real forms of gl(2) and gl(1/1)

In order to illustrate the K-matrix theory, we compute in this section inner products for real forms of the complex algebras gl(2) and gl(1/1). We first examine the two possible real forms u(2) and u(1,1) of gl(2) and seek to solve the recursion formula (7.2b).

(1) For $\mu^0 - \nu^0 = 2J \geqslant 0$, we find (see Sec. VI C for notation)

$$\kappa^2(n+1) = (2J - n)\kappa^2(n), \tag{7.14}$$

where

$$\kappa^2(n) = \langle n|\kappa^2|n\rangle.$$

Setting $\kappa^2(0) = 1$, we obtain, for $n < 2J$,

$$\kappa^2(n) = [(2J)!]/(2J - n)! \tag{7.15}$$

[we choose the positive square root $\kappa(n)$ of $\kappa^2(n)$]. It also follows from (7.14) that $\kappa^2(n) = 0$ for $n \geqslant 2J$ and the κ operator thus effects a truncation of the infinite-dimensional VB space to its irreducible invariant subspace. It is easily verified that matrix elements of the γ representation reproduce the well-known angular momentum matrix elements

$$\begin{aligned} &\langle n|\gamma(C_{11})|n\rangle = \mu^0 - n = \mu^0 - J + M, \\ &\langle n|\gamma(C_{22})|n\rangle = \nu^0 + n = \nu^0 + J - M, \\ &\langle n+1|\gamma(B)|n\rangle = \sqrt{(J+M)(J-M+1)}, \\ &\langle n-1|\gamma(A)|n\rangle = \sqrt{(J-M)(J+M+1)}, \end{aligned} \tag{7.16}$$

where $2J = \mu^0 - \nu^0$ and $M = J - n$. The γ representation is therefore finite-dimensional and satisfies the hermiticity condition (7.1a) for a unitary irreducible representation of u(2).

(2) For $\mu^0 - \nu^0 = -2\mathscr{J} \leqslant 0$, we find

$$\kappa^2(n) = [(2\mathscr{J} + n - 1)!]/(2\mathscr{J} - 1)!. \tag{7.17}$$

The matrix elements

$$\begin{aligned} &\langle n|\gamma(C_{11})|n\rangle = \mu^0 - n = \mu^0 + \mathscr{J} + \mathscr{M}, \\ &\langle n|\gamma(C_{22})|n\rangle = \nu^0 + n = \nu^0 - \mathscr{J} - \mathscr{M}, \\ &\langle n+1|\gamma(B)|n\rangle = -\sqrt{(\mathscr{M} + \mathscr{J} - 1)(\mathscr{M} - \mathscr{J})}, \\ &\langle n-1|\gamma(A)|n\rangle = \sqrt{(\mathscr{M} - \mathscr{J} + 1)(\mathscr{M} + \mathscr{J})}, \end{aligned} \tag{7.18}$$

of the γ representation, for $\mathscr{M} = -\mathscr{J} - n$, now pertain to the so-called D^- series[17] of Hermitian representations of the real noncompact form u(1,1).

For the gl(1/1) example of the end of Sec. VI C, we obtain from (7.13) and (6.5)

$$\kappa^2(0) = 1, \quad \kappa^2(1) = \pm(\mu^0 + \nu^0), \tag{7.19}$$

where the sign in the expression for $\kappa^2(1)$ is determined by the requirement that $\kappa^2(1)$ be positive. The two-dimensional γ representations for A and B are then given by

$$\gamma(A) = \begin{pmatrix} 0 & \kappa(1) \\ 0 & 0 \end{pmatrix}, \quad \gamma(B) = \begin{pmatrix} 0 & 0 \\ \pm\kappa(1) & 0 \end{pmatrix}.$$

The star adjoints

$$\begin{pmatrix} W & X \\ Y & Z \end{pmatrix}^{\dagger} = \begin{pmatrix} W^{\dagger} & Y^{\dagger} \\ X^{\dagger} & Z^{\dagger} \end{pmatrix}$$

and grade star adjoints

$$\begin{pmatrix} W & X \\ Y & Z \end{pmatrix}^{\ddagger} = \begin{pmatrix} W^{\dagger} & -Y^{\dagger} \\ X^{\dagger} & Z^{\dagger} \end{pmatrix}$$

of these matrices then verify that

$$\gamma^{\ddagger}(A) = \gamma^{\dagger}(A) = \pm\gamma(B),$$
$$\gamma^{\ddagger}(B) = -\gamma^{\dagger}(B) = \mp\gamma(A).$$

Any $\{\mu^0\}:\{\nu^0\}$ representation of gl(1/1) is thus simultaneously star and grade star, and trivial (one-dimensional) when atypical ($\mu^0 + \nu^0 = 0$).

VIII. THE LIE SUPERALGEBRA gl(1/n)

The Lie superalgebras gl(m/n) and gl(n/m) are isomorphic. We consider in this section the superalgebras of type gl($1/n$). These are the only gl(m/n) type of superalgebras whose representations are multiplicity-free on restriction to gl(m) $\oplus$ gl(n) and for which, as a consequence, major simplifications arise.

(1) The partitions $\{-\tau\}:\{\tilde{\tau}\}$ for the Grassman polynomial $\Theta^{\{-\tau\}}_{\{\tilde{\tau}\}}(\theta)$ restrict to the set $\{-k\}:\{1^k\}$, $0 \leqslant k \leqslant n$.

(2) The gl(1) algebra is isomorphic to $\mathfrak{R}$, with trivial one-dimensional representations labeled by the number μ^0, which is restricted here to real values.

(3) The $U(n)$ coupling

$$\{\nu^0\} \times \{1^k\} \to \{\nu^0 + \Delta(\alpha_1,\alpha_2,\ldots,\alpha_k)\}$$

[where $\Delta(\alpha_1,\alpha_2,\ldots,\alpha_k)$, $1 \leqslant \alpha_1 < \alpha_2 < \cdots < \alpha_k \leqslant n$, is a null n vector except for the numeral 1 in its $(\alpha_1,\alpha_2,\ldots,\alpha_k)$ entries] is multiplicity-free.

We ignore the τ, $\tilde{\tau}$ (redundant) and ρ_μ, ρ_ν (unnecessary) labels in the following. For ease of notation, we also suppress the labels $\{\mu^0\}$, $\{\nu^0\}$, and $\{\mu\} = \{\mu^0 - \kappa\}$.

A. The Lie superalgebra gl(1/2)

For simplicity of exposition, we first consider the finite-dimensional representations of the Lie superalgebra gl(1/2).

The orthonormal VG basis (5.10) consists of the $(2^{2\cdot 1} = 4)$gl(1) $\oplus$ gl(2) sub-bases:

$$|\{\nu_1^0,\nu_2^0\}(m_\nu)\rangle, \tag{8.1a}$$
$$|\{\nu_1^0 + 1,\nu_2^0\}(m_\nu)\rangle, \tag{8.1b}$$
$$|\{\nu_1^0,\nu_2^0 + 1\}(m_\nu)\rangle, \tag{8.1c}$$
$$|\{\nu_1^0 + 1,\nu_2^0 + 1\}(m_\nu)\rangle, \tag{8.1d}$$

When $\nu_1^0 = \nu_2^0$, the state (8.1c) is not allowed.

Using (6.2), (6.4), and (6.5), we find the following u(2)-reduced extended $\overline{\Gamma}$ matrix representation

$$\langle\{\nu'\}\|\overline{\Gamma}(B)\|\{\nu\}\rangle = \begin{pmatrix} 0 & 0 & 0 & 0 \\ (\mu^0 + \nu_1^0) & 0 & 0 & 0 \\ (\mu^0 + \nu_2^0 - 1) & 0 & 0 & 0 \\ 0 & (\mu^0 + \nu_2^0 - 1)\left[\dfrac{\nu_1^0 - \nu_2^0 + 2}{\nu_1^0 - \nu_2^0 + 1}\right]^{1/2} & -(\mu^0 + \nu_1^0)\left[\dfrac{\nu_1^0 - \nu_2^0}{\nu_1^0 - \nu_2^0 + 1}\right]^{1/2} & 0 \end{pmatrix}, \tag{8.2a}$$

$$\langle\{\nu'\}\|\overline{\Gamma}(A)\|\{\nu\}\rangle = \begin{pmatrix} 0 & \left[\dfrac{\nu_1^0 - \nu_2^0 + 2}{\nu_1^0 - \nu_2^0 + 1}\right]^{1/2} & -\left[\dfrac{\nu_1^0 - \nu_2^0}{\nu_1^0 - \nu_2^0 + 1}\right]^{1/2} & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{8.2b}$$

in the ordered basis (8.1). The atypicality conditions (6.14) are given by $\mu^0 + \nu_1^0 = 0$ or $\mu^0 + \nu_2^0 - 1 = 0$. It is verified that these matrices obey the U(2)-reduced commutator algebra

$$\sum_{\nu''} U(\{\nu\}\{1\}\{\nu'\}\{-1\};\{\nu''\}\{0\})\langle\{\nu'\}\|A\|\{\nu''\}\rangle\langle\{\nu''\}\|B\|\{\nu\}\rangle$$
$$-\sum_{\nu''} U(\{\nu\}\{-1\}\{\nu'\}\{1\};\{\nu''\}\{0\})\langle\{\nu'\}\|B\|\{\nu''\}\rangle\langle\{\nu''\}\|A\|\{\nu\}\rangle$$
$$= \delta_{\nu\nu'}(1/\sqrt{2})\langle\{\nu'\}\|(2C_{11} + C_{\alpha\alpha})\|\{\nu\}\rangle$$
$$= \delta_{\nu\nu'}(1/\sqrt{2})[2(\mu^0 + \nu_1^0 + \nu_2^0) - (\nu_1 + \nu_2)] \tag{8.3a}$$

and

$$\sum_{\nu''} U(\{\nu\}\{1\}\{\nu'\}\{-1\};\{\nu''\}\{1,-1\})\langle\{\nu'\}\|A\|\{\nu''\}\rangle\langle\{\nu''\}\|B\|\{\nu\}\rangle$$
$$+\sum_{\nu''} U(\{\nu\}\{-1\}\{\nu'\}\{1\};\{\nu''\}\{1,-1\})\langle\{\nu'\}\|B\|\{\nu''\}\rangle\langle\{\nu''\}\|A\|\{\nu\}\rangle$$
$$= \delta_{\nu\nu'}\langle\{\nu'\}\|\sqrt{2}\,\mathbf{J}\|\{\nu\}\rangle = \delta_{\nu\nu'}[(\nu_1 - \nu_2)(\nu_1 - \nu_2 + 2)/2]^{1/2}, \tag{8.3b}$$

where the symbols $U(\cdots)$ are unitary U(2) $6j$ recoupling coefficients, and **J** is the angular momentum operator of the su(2)⊂u(2) Lie subalgebra.

With the help of (6.5) and (7.13), we find the following solutions for the κ^2 matrices.

(1) For star equivalent representations,

$$\begin{aligned}
&\langle\{\nu_1^0,\nu_2^0\}|\kappa^2|\{\nu_1^0,\nu_2^0\}\rangle = 1, \quad \langle\{\nu_1^0+1,\nu_2^0\}|\kappa^2|\{\nu_1^0+1,\nu_2^0\}\rangle = \pm(\mu^0+\nu_1^0),\\
&\langle\{\nu_1^0,\nu_2^0+1\}|\kappa^2|\{\nu_1^0,\nu_2^0+1\}\rangle = \pm(\mu^0+\nu_2^0-1),\\
&\langle\{\nu_1^0+1,\nu_2^0+1\}|\kappa^2|\{\nu_1^0+1,\nu_2^0+1\}\rangle = (\mu^0+\nu_1^0)(\mu^0+\nu_2^0-1).
\end{aligned} \tag{8.4a}$$

(2) For grade star equivalent representations,

$$\begin{aligned}
&\langle\{\nu_1^0,\nu_2^0\}|\kappa^2|\{\nu_1^0,\nu_2^0\}\rangle = 1, \quad \langle\{\nu_1^0+1,\nu_2^0\}|\kappa^2|\{\nu_1^0+1,\nu_2^0\}\rangle = \pm(\mu^0+\nu_1^0),\\
&\langle\{\nu_1^0,\nu_2^0+1\}|\kappa^2|\{\nu_1^0,\nu_2^0+1\}\rangle = \pm(\mu^0+\nu_2^0-1),\\
&\langle\{\nu_1^0+1,\nu_2^0+1\}|\kappa^2|\{\nu_1^0+1,\nu_2^0+1\}\rangle = (\mu^0+\nu_1^0)(1-\mu^0-\nu_2^0).
\end{aligned} \tag{8.4b}$$

Recalling that $\nu_1^0 \geqslant \nu_2^0$, implying $\mu^0+\nu_1^0 > \mu^0+\nu_2^0-1$, we have that the VCS representation is equivalent to a star representation if and only if $\mu^0+\nu_2^0-1 \geqslant 0$ [upper sign in (8.4a)] or $-(\mu^0+\nu_1^0) \geqslant 0$ [lower sign in (8.4a)]. It is equivalent to a grade star representation if and only if $\nu_1^0 = \nu_2^0$ [the state (8.1c) is then not allowed] and $0 < (\mu^0+\nu_1^0) < 1$ [only the upper sign is allowed in (8.4b)]. To within a different parametrization ($\mu^0+\nu_1^0 = b+q$, $\mu^0+\nu_2^0-1 = b-q$,) these results duplicate the analysis carried out by Scheunert *et al.*[18] for su(1/2).

In the ordered basis (8.1), we find, for the star matrix representation,

$$\langle\{\nu'\}\|\gamma(B)\|\{\nu\}\rangle = \pm\begin{pmatrix} 0 & 0 & 0 & 0\\ [\pm(\mu^0+\nu_1^0)]^{1/2} & 0 & 0 & 0\\ [\pm(\mu^0+\nu_2^0-1)]^{1/2} & 0 & 0 & 0\\ 0 & \left[\pm(\mu^0+\nu_2^0-1)\left(\frac{\nu_1^0-\nu_2^0+2}{\nu_1^0-\nu_2^0+1}\right)\right]^{1/2} & -\left[\pm(\mu^0+\nu_1^0)\left(\frac{\nu_1^0-\nu_2^0}{\nu_1^0-\nu_2^0+1}\right)\right]^{1/2} & 0 \end{pmatrix}, \tag{8.5a}$$

$$\langle\{\nu'\}\|\gamma(A)\|\{\nu\}\rangle = \begin{pmatrix} 0 & \left[\pm(\mu^0+\nu_1^0)\left(\frac{\nu_1^0-\nu_2^0+2}{\nu_1^0-\nu_2^0+1}\right)\right]^{1/2} & -\left[\pm(\mu^0+\nu_2^0-1)\left(\frac{\nu_1^0-\nu_2^0}{\nu_1^0-\nu_2^0+1}\right)\right]^{1/2} & 0\\ 0 & 0 & 0 & -[\pm(\mu^0+\nu_2^0-1)]^{1/2}\\ 0 & 0 & 0 & -[\pm(\mu^0+\nu_1^0)]^{1/2}\\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{8.5b}$$

and, for the grade star matrix representation [with rows and columns referring to the unallowed state (8.1c) deleted],

$$\langle\{\nu'\}\|\gamma(B)\|\{\nu\}\rangle = \begin{pmatrix} 0 & 0 & 0\\ [\mu^0+\nu_1^0]^{1/2} & 0 & 0\\ 0 & -[2(1-\mu^0-\nu_1^0)]^{1/2} & 0 \end{pmatrix}, \tag{8.6a}$$

$$\langle\{\nu'\}\|\gamma(A)\|\{\nu\}\rangle = \begin{pmatrix} 0 & [2(\mu^0+\nu_1^0)]^{1/2} & 0\\ 0 & 0 & -[1-\mu^0-\nu_1^0]^{1/2}\\ 0 & 0 & 0 \end{pmatrix}. \tag{8.6b}$$

It is verified that these matrix representations both obey the reduced-commutator algebra (8.3).

Finally, for star representations, we have the star adjoint relationship

$$\langle\{\nu'\}\|\gamma(B)\|\{\nu\}\rangle = \pm(-1)^{\phi_2(\{\nu\})+\phi_2(\{1\})-\phi_2(\{\nu'\})}[\dim\{\nu\}/\dim\{\nu'\}]^{1/2}\langle\{\nu\}\|\gamma(A)\|\{\nu'\}\rangle \tag{8.7}$$

(see Appendix A for the definition of the phase factors) while, for the grade star representation, we have the grade star adjoint relationship

$$\langle\{\nu'\}\|\gamma(B)\|\{\nu\}\rangle = \mp(-1)^{\sigma(\{\nu'\})}(-1)^{\phi_2(\{\nu\})+\phi_2(\{1\})-\phi_2(\{\nu'\})}\times[\dim\{\nu\}/\dim\{\nu'\}]^{1/2}\langle\{\nu\}\|\gamma(A)\|\{\nu'\}\rangle, \tag{8.8}$$

where the sign in (8.8) (and, consequently, the grade of the intrinsic module) is chosen such that $\mp(-1)^{\sigma(\{\nu_1^0+1,\nu_2^0\})} = +1$.

B. The Lie superalgebra gl(1/n)

In this subsection on the superalgebra gl(1/n), we parallel the developments of the preceding subsection on gl(1/2), quoting only the most important results. For simplicity, we set $\{\nu\}\equiv\{\nu^0+\Delta(\alpha_1,\alpha_2,\ldots,\alpha_k)\}$ in the following.

From (6.2) and (6.5), we have

$$\langle\{\nu\}\|\Gamma(B)\|\{\nu-\Delta(\alpha_i)\}\rangle=(\mu^0+\nu_{\alpha_i}-\alpha_i+1)\times U(\{\nu^0\}\{1^{k-1}\}\{\nu\}\{1\};\{\nu-\Delta(\alpha_i)\}\{1^k\})\times(\{1^k\}\|\theta\|\{1^{k-1}\})$$

$$=(\mu^0+\nu_{\alpha_i}-\alpha_i+1)\times\left(\prod_{\substack{j=1\\ \neq i}}^{k}S(i-j)\right)\left[\prod_{\substack{j=1\\ \neq i}}^{k}\left(\frac{\nu^0_{\alpha_j}-\nu^0_{\alpha_i}+\alpha_i-\alpha_j+1}{\nu^0_{\alpha_j}-\nu^0_{\alpha_i}+\alpha_i-\alpha_j}\right)\right]^{1/2},\tag{8.9}$$

where we have used[19]

$$U(\{\nu^0\}\{1^{k-1}\}\{\nu\}\{1\};\{\nu-\Delta(\alpha_i)\}\{i^k\})=\left(\prod_{\substack{j=1\\ \neq i}}^{k}S(i-j)\right)\left[\frac{1}{k}\prod_{\substack{j=1\\ \neq i}}^{k}\left(\frac{\nu^0_{\alpha_j}-\nu^0_{\alpha_i}+\alpha_i-\alpha_j+1}{\nu^0_{\alpha_j}-\nu^0_{\alpha_i}+\alpha_i-\alpha_j}\right)\right]^{1/2},$$

with $S(i-j)$ the sign of the difference $i-j$, and, from Appendix A,

$$(\{1^k\}\|\theta\|\{1^{k-1}\})=\sqrt{k}\,.$$

Similarly, from (6.4), we have

$$\langle\{\nu-\Delta(\alpha_i)\}\|\Gamma(A)\|\{\nu\}\rangle=U(\{\nu^0\}\{1^k\}\{\nu-\Delta(\alpha_i)\}\{-1\};\{\nu\}\{1^{k-1}\})\times(\{1^{k-1}\}\|\partial\|\{1^k\})$$

$$=(-1)^{\alpha_i-1}\left(\prod_{\substack{j=1\\ \neq i}}^{k}S(i-j)\right)\left[\prod_{\substack{\beta=1\\ \neq\alpha_1,\alpha_2,\ldots,\alpha_k}}^{n}\left(\frac{\nu^0_{\alpha_i}-\nu^0_\beta+\beta-\alpha_i+1}{\nu^0_{\alpha_i}-\nu^0_\beta+\beta-\alpha_i}\right)\right]^{1/2},\tag{8.10}$$

where we have used[6]

$$U(\{\nu^0\}\{1^k\}\{\nu-\Delta(\alpha_i)\}\{-1\};\{\nu\}\{1^{k-1}\})$$

$$=(-1)^{\phi_n(\{\nu\})+\phi_n(\{1^{k-1}\})-\phi_n(\{\nu-\Delta(\alpha_i)\})-\phi_n(\{1^k\})}\times\left[\frac{\dim\{\nu\}\dim\{1^{k-1}\}}{\dim\{\nu-\Delta(\alpha_i)\}\dim\{1^k\}}\right]^{1/2}$$

$$\times U(\{\nu^0\}\{1^{k-1}\}\{\nu\}\{1\};\{\nu-\Delta(\alpha_i)\}\{1^k\})\,,$$

and where, from Appendix A,

$$(\{1^{k-1}\}\|\partial\|\{1^k\})=(-1)^{k-1}\sqrt{n-k+1}\,.$$

The atypicality conditions (6.9) are given by

$$\mu^0+\nu_\alpha-\alpha+1=0,\quad 1\leqslant\alpha\leqslant n\,.\tag{8.11}$$

The results are consistent with those of Palev[20] who used the method of induced representations,[8] which we discuss further in Sec. IX.

From (7.13), we easily derive the solutions

$$\langle\{\nu\}|\kappa^2|\{\nu\}\rangle=(-1)^{\Phi}\prod_{l=1}^{k}(\mu^0+\nu^0_{\alpha_l}-\alpha_l+1)\tag{8.12}$$

for κ^2, where $(-1)^{\Phi}$ is defined such that the right-hand side of Eq. (8.12) is positive. A representation is then determined to be equivalent to a star representation if $(-1)^{\Phi}=(\pm1)^k$ and equivalent to a grade star representation if $(-1)^{\Phi}=(-1)^{(1/2)k(k\mp1)}$. Since $\nu^0_\alpha\geqslant\nu^0_\beta$, for $\alpha>\beta$, we have that a highest weight representation $\{\mu^0\}:\{\nu^0\}$ is star if and only if $\mu^0+\nu^0_n-n+1\geqslant0$ $[(-1)^{\Phi}=(+1)^k=+1]$ or $-(\mu^0+\nu^0_1)\geqslant0$ $[(-1)^{\Phi}=(-1)^k]$. The conditions for positive-definitiveness of κ^2 in the grade star case are so stringent that they are the exception rather than the rule. Consequently, we do not investigate the grade star case any further.

For star representations, we then have

$$\langle\{\nu\}\|\gamma(B)\|\{\nu-\Delta(\alpha_i)\}\rangle=\pm\left(\prod_{\substack{j=1\\ \neq i}}^{k}S(i-j)\right)\left[\pm(\mu^0+\nu_{\alpha_i}-\alpha_i+1)\times\prod_{\substack{j=1\\ \neq i}}^{k}\left(\frac{\nu^0_{\alpha_j}-\nu^0_{\alpha_i}+\alpha_i-\alpha_j+1}{\nu^0_{\alpha_j}-\nu^0_{\alpha_i}+\alpha_i-\alpha_j}\right)\right]^{1/2},\tag{8.13a}$$

and

$$\langle\{\nu-\Delta(\alpha_i)\}\|\gamma(A)\|\{\nu\}\rangle$$

$$=(-1)^{\alpha_i-1}\left(\prod_{\substack{j=1\\ \neq i}}^{k}S(i-j)\right)\left[\pm(\mu^0+\nu_{\alpha_i}-\alpha_i+1)\times\prod_{\substack{\beta=1\\ \neq\alpha_1,\alpha_2,\ldots,\alpha_k}}^{k}\left(\frac{\nu^0_{\alpha_i}-\nu^0_\beta+\beta-\alpha_i+1}{\nu^0_{\alpha_i}-\nu^0_\beta+\beta-\alpha_i}\right)\right]^{1/2},\tag{8.13b}$$

with star adjoint relationship

$$\langle\{\nu\}\|\gamma(B)\|\{\nu-\Delta(\alpha_i)\}\rangle = \pm(-1)^{\phi_n(\{\nu-\Delta(\alpha_i)\})+\phi_n(\{1\})-\phi_n(\{\nu\})} \times[\dim\{\nu-\Delta(\alpha_i)\}/\dim\{\nu\}]^{1/2}\langle\{\nu-\Delta(\alpha_i)\}\|\gamma(A)\|\{\nu\}\rangle\ . \tag{8.14}$$

The Wigner coefficients needed to evaluate from these reduced matrix elements the full matrix elements in a Gel'fand–Zetlin basis are given in Refs. 19 and 21.

IX. DISCUSSION

It has been shown, herein and elsewhere, that, using VCS theory, one can explicitly construct irreducible representations of semisimple (and, more generally, reductive) Lie algebras starting from representations of a stability subalgebra with highest or lowest weight. We have further shown, in this paper, that, with minor adjustments, the same construction applies to classical Lie superalgebras, thus providing us with the means to construct all irreducible representations with highest/lowest weight of the latter. However, a few qualifications concerning this claim are in order.

Recall that to every weight for a complex semisimple Lie algebra, there exists an irreducible representation that is unique up to equivalence and for which the given weight is its highest.[14] It can be proved that VCS theory gives an explicit construction of this represenation on an irreducible subspace of a VB space. We have shown that application of the VCS construction to the classical Lie superalgebra $\mathrm{gl}(m/n)$, albeit using Grassman as opposed to Bargmann variables, leads to parallel results.

To expose the essence of the VCS construction, it is useful to compare it with the standard inducing construction given for superalgebras by Kac[8] and applied by Palev[20] to $\mathrm{gl}(1/n)$. In the standard theory, the module for the induced representation of $\mathrm{gl}(m/n)$ is easily shown to be isomorphic to the vector-Grassman space $V\otimes G$, where V is the irreducible highest grade module for the $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ subalgebra. But whereas in the extended VCS representation, the odd generators are realized by

$$\overline{\Gamma}(A_{i\alpha})=\partial_{i\alpha},\quad \overline{\Gamma}(B_{i\alpha})=[\Omega,\theta_{i\alpha}]\ ,$$

it can be shown that they have the expression

$$\overline{T}(A_{i\alpha})=[\partial_{i\alpha},\Omega],\quad \overline{T}(B_{i\alpha})=\theta_{i\alpha}\ ,$$

in the standard induced representation. One sees that the $\overline{T}(B_{i\alpha})$ lowering operators can lower all the way from the highest to the lowest grade Grassman states. It follows that the space $V\otimes G$ could have no proper submodule for $\overline{T}$ which contains the highest weight state. To obtain an irreducible module for a representation with the desired highest weight, one must factorize the $V\otimes G$ space with respect to its maximal invariant submodule. In contrast, the VCS $\Gamma(B_{i\alpha})$ lowering operators generate an irreducible submodule of the $V\otimes G$ space and there is no need for factorization. For typical representations of superalgebras, the standard induced representation is irreducible and the two constructions become equivalent. However, for the atypical representations, the VCS construction is simpler and more direct.

Another distinctive difference between classical Lie superalgebras and standard Lie algebras, which is clearly exposed in VCS theory, is the existence of three kinds (star, grade star, and neither) of representations as opposed to two kinds (Hermitian and non-Hermitian) for Lie algebras. Recall that for a real Lie group, and irreducible representation is either equivalent to a unitary representation or it is not. The corresponding representation of the real Lie algebra is then Hermitian or not. Now if γ is a representation of a (real or complex) Lie algebra $\mathbf{g}$, one has the rather remarkable (but familiar) result that, with respect to any nondegenerate (Hermitian, positive) inner product on the carrier space for γ, there always exists some $Y_X\in\mathbf{g}$ for which $\gamma(Y_X)=\gamma^\dagger(X)$. The concept of a Hermitian adjoint operation is therefore well defined for a Hilbert space representation of a Lie algebra. A corollary is that every representation of a complex Lie algebra is (equivalent to) a Hermitian representation of some real form of the Lie algebra.

For a classical Lie superalgebra $\mathbf{g}$, we can similarly define the Hermitian adjoint $\gamma^\dagger(X)$ as a linear operator on the representation space for any $X\in\mathbf{g}$. However, we cannot in general guarantee that Y_X defined by $\gamma(Y_X)=\gamma^\dagger(X)$ is an element of $\mathbf{g}$. If it is, the representation is called a "star" representation. For a representation of a superalgebra, Scheunert *et al.* have shown[15] that one can also define a grade adjoint operation $\gamma(x)\to\gamma^\ddagger(X)$ and is therefore meaningful to enquire if there exists some $Z_X\in\mathbf{g}$, corresponding to an $X\in\mathbf{g}$, for which $\gamma(Z_X)=\gamma^\ddagger(X)$. If there does, the representation is called a "grade star" representation. As we have shown in this paper, K-matrix theory provides a simple straightforward technique for determining if a given irreducible representation is of the star or grade star type. Note that representations can be of one or the other type, of both types simultaneously, or of neither type.

In conclusion, we submit that VCS and K-matrix theory have been proved to be of substantial pedagogical and practical value for classical Lie superalgebras as well as for standard Lie algebras. They not only enable one to construct and examine the structures and properties of any irreducible ladder representations of these algebras, they also provide the only simple systematic procedure that we are aware of for explicitly determining the inner products and computing matrix elements for these representations. This is, of course, an essential step in the practical application of algebraic structures to physical problems. Furthermore, it is achieved in a very physical way by embedding representation in a simple boson Fock space for Lie algebras, and in a simple fermion or combined boson/fermion Fock space for classical Lie superalgebras.

APPENDIX A: THE $\Theta(\theta)$ POLYNOMIALS

As argued in Sec. V, the set of Grassman variables $\{\theta_{i\alpha}\}$ transforms under $\mathbf{n}_0=\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ as a $\{-1\}:\{1\}$ tensor. The set of all polynomials of rank n_θ in the Grassman variables $\{\theta_{i\alpha}\}$ reduces, under $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$, to the direct sum of all tensor irreps labeled by partitions $\{-\tau\}:\{\tilde{\tau}\}$ in m and n parts, respectively, where in order to insure full

antisymmetry of the polynomial, $\{\tilde{\tau}\}$ is the partition conjugate to $\{\tau\}$.

We set the convention (essentially a phase convention) that the θ polynomial corresponding to the highest weight component of the irreducible tensor labeled by $\{-\tau\}:\{\tilde{\tau}\}$, and properly normalized with respect to the Grassman inner product, is given by

$$\begin{aligned}\Theta^{\{-\tau\}(\mathrm{hw})}_{\{\tilde{\tau}\}(\mathrm{hw})}(\theta) &= |^{\{-\tau\}(\mathrm{hw})}_{\{\tilde{\tau}\}(\mathrm{hw})}) \\ &= \theta_{m-\tilde{\tau}_n+1,n}\theta_{m-\tilde{\tau}_n,n}\cdots\theta_{m,n} \\ &\quad\times\theta_{m-\tilde{\tau}_{n-1}+1,n-1}\theta_{m-\tilde{\tau}_{n-1},n-1}\cdots\theta_{m,n-1} \\ &\quad\times\cdots \\ &\quad\times\theta_{m-\tilde{\tau}_1+1,1}\theta_{m-\tilde{\tau}_1,1}\cdots\theta_{m,1}\,.\end{aligned} \tag{A1}$$

That the expression (A1) is of highest weight with respect to both $\mathrm{gl}(m)$ and $\mathrm{gl}(n)$ is easily verified by the vanishing action of the raising operators $\Gamma(C_{\alpha\beta}),\alpha<\beta$, and $\Gamma(C_{ij}),i<j$, on it.

Appealing to the Wigner–Eckart theorem, we now compute the $\mathfrak{n}_0$-reduced matrix elements of the Grassman tensor $\theta^{\{-1\}}_{\{1\}}$ between states belonging to the tensor representations $\{-\tau\}:\{\tilde{\tau}\}$ and $\{-\tau-\Delta(\tilde{\tau}_\beta+1)\}:\{\tilde{\tau}+\Delta(\beta)\}$, where, e.g., $\Delta(\beta)$ is the n vector $(00\cdots010\cdots0)$ with null entries everywhere except for the numeral 1 in its βth entry. Using the antisymmetry properties of the Grassman variables, one easily obtains the following matrix element:

$$(^{\{-\tau-\Delta(\tilde{\tau}_\beta+1)\}(\mathrm{hw})}_{\{\tilde{\tau}+\Delta(\beta)\}(\mathrm{hw})}|(-1)^{m-\tilde{\tau}_\beta-1}\theta_{m-\tilde{\tau}_\beta,\beta}|^{\{-\tau\}(\mathrm{hw})}_{\{\tilde{\tau}\}(\mathrm{hw})}) = (-1)^{m-\tilde{\tau}_\beta-1}(-1)^{\sum_{\gamma=\beta+1}^{n}\tilde{\tau}_\gamma}. \tag{A2}$$

The corresponding $\mathrm{U}(m)$ and $\mathrm{U}(n)$ Wigner coefficients needed to isolate the reduced matrix element from the matrix element (A2) are obtained following the pattern calculus of Biedenharn and Louck[19,21] and are given, for $\mathrm{U}(m)$, by

$$\left[\frac{1}{(\tilde{p}_{\beta n}+1)}\prod_{\sigma=\beta+1}^{n}\left(\frac{\tilde{p}_{\beta n}-\tilde{p}_{\sigma n}+1}{\tilde{p}_{\beta n}-\tilde{p}_{\sigma n}}\right)\right]^{1/2} \tag{A3a}$$

and, for $\mathrm{U}(n)$, by

$$\left[\prod_{\sigma=1}^{\beta-1}\left(\frac{\tilde{p}_{\beta n}-\tilde{p}_{\sigma n}+1}{\tilde{p}_{\beta n}-\tilde{p}_{\sigma n}}\right)\right]^{1/2}, \tag{A3b}$$

where the partial hooks $\tilde{p}_{\alpha n}$ are defined by

$$\tilde{p}_{\alpha n}=\tilde{\tau}_\alpha+n-\alpha\,.$$

Dividing the right-hand side of (A2) by (A3a) and (A3b), we find

$$(^{\{-\tau-\Delta(\tilde{\tau}_\beta+1)\}}_{\{\tau+\Delta(\beta)\}}\|\theta^{\{-1\}}_{\{1\}}\|^{\{-\tau\}}_{\{\tilde{\tau}\}}) = (-1)^{m-\tilde{\tau}_\beta-1}(-1)^{\sum_{\gamma=\beta+1}^{n}\tilde{\tau}_\gamma}\left[(\tilde{p}_{\beta n}+1)\prod_{\substack{j=1\\ \neq\beta}}^{n}\left(\frac{\tilde{p}_{\beta n}-\tilde{p}_{\sigma n}}{\tilde{p}_{\beta n}-\tilde{p}_{\sigma n}+1}\right)\right]^{1/2}. \tag{A4}$$

The reduced matrix elements for the tensor operator $\partial^{\{1\}}_{\{1\}}$ are similarly given by

$$\begin{aligned}(^{\{-\tau'\}}_{\{\tilde{\tau}'\}}\|\partial^{\{1\}}_{\{-1\}}\|^{\{-\tau\}}_{\{\tilde{\tau}\}}) &= (-1)^{\phi_m(\{-\tau\})-\phi_m(\{1\})-\phi_m(\{-\tau'\})}(-1)^{\phi_n(\{\tilde{\tau}\})-\phi_n(\{-1\})-\phi_n(\{\tilde{\tau}'\})} \\ &\quad\times\left[\frac{\dim\{\tilde{\tau}\}}{\dim\{\tilde{\tau}'\}}\cdot\frac{\dim\{-\tau\}}{\dim\{-\tau'\}}\right]^{1/2}\times(^{\{-\tau\}}_{\{\tilde{\tau}\}}\|\theta^{\{-1\}}_{\{1\}}\|^{\{-\tau'\}}_{\{\tilde{\tau}'\}}),\end{aligned} \tag{A5}$$

where the dimension of the $\mathrm{U}(l)$ partition $\{\xi\}$ is given by

$$\dim(\{\xi\}) = \frac{\Pi_{1\le i<j\le l}(p_{il}-p_{jl})}{1!2!\cdots(l-1)!}, \tag{A6}$$

and

$$\phi_l\{\xi\}) = \frac{1}{2}\sum_{k=1}^{l}(l+1-2k)\xi_k\,. \tag{A7}$$

APPENDIX B: THE $Z(z)$ POLYNOMIALS

Also as argued in Sec. V, the set of Bargmann variables $\{z_{i\alpha}\}$ transforms under $\mathfrak{n}_0=\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ as a $\{-1\}:\{1\}$ tensor. The set of all polynomials of rank n_z in the Bargmann variables $\{z_{i\alpha}\}$ reduces, under $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$, to the direct sum of all tensor irreps labeled by partitions $\{-\tau\}:\{\tau\}$ in m and n parts, respectively.

We set the convention that the z polynomial corresponding to the highest weight component of the irreducible tensor labeled by $\{-\tau\}:\{\tau\}$, and properly normalized with respect to the Bargmann inner product, is given by[19,22]

$$Z^{\{-\tau\}(\mathrm{hw})}_{\{\tau\}(\mathrm{hw})}(z) = N(\tau)(z^m_1)^{\tau_1-\tau_2}(z^{m\ m-1}_{1\ 2})^{\tau_2-\tau_3}\cdots(z^{m\ m-1\cdots1}_{1\ 2\cdots m})^{\tau_m} \tag{B1a}$$

(recall that we assume $m\leqslant n$), where

$$z_{1\ 2\ \cdots l}^{m\ m-1\ \cdots m-l+1} = \begin{vmatrix} z_{m1} & z_{m2} & \cdots & z_{ml} \\ z_{m-1,1} & z_{m-1,2} & \cdots & z_{m-1,m} \\ \vdots & & \ddots & \\ z_{m-l+1,1} & z_{m-l+1,2} & \cdots & z_{m-l+1,l} \end{vmatrix} \tag{B1b}$$

$$N(\tau) = \left[\frac{\Pi_{1\le i<j\le m}(p_{im}-p_{jm})}{\Pi_{k=1}^{m} p_{km}!} \right]^{1/2}, \tag{B1c}$$

and

$$p_{im} = \tau_i + m - i\,. \tag{B1d}$$

That the expression (B1a) is highest weight with respect to both gl(m) and gl(n) is easily verified by the vanishing action of the raising operators $\Gamma(C_{\alpha\beta})$, $\alpha<\beta$, and $\Gamma(C_{ij})$, $i<j$, on it.

Following a procedure similar to the one used in Appendix A, we find that the $\mathbf{n}_0$-reduced matrix elements for the Bargmann tensor $z^{\{-1\}}_{\{1\}}$ between states belonging to the tensor representations $\{-\tau\}$:$\{\tau\}$ and $\{-\tau-\Delta(i)\}$:$\{\tau+\Delta(i)\}$, where $\Delta(i)$ is the m vector $(00\cdots010\cdots0)$ with null entries everywhere except for the numeral 1 in its ith entry, are given by

$$(^{\{-\tau-\Delta(i)\}}_{\{\tau+\Delta(i)\}}\|z^{\{-1\}}_{\{1\}}\|^{\{-\tau\}}_{\{\tau\}}) = \left[(p_{im}+1) \prod_{\substack{k=1 \\ \neq i}}^{m} \left(\frac{p_{im}-p_{km}}{p_{im}-p_{km}+1} \right) \right]^{1/2}. \tag{B2}$$

The reduced matrix elements for the tensor operator $\nabla^{\{1\}}_{\{-1\}}$ are then given by

$$(^{\{-\tau'\}}_{\{\tau'\}}\|\nabla^{\{1\}}_{\{-1\}}\|^{\{-\tau\}}_{\{\tau\}}) = (-1)^{\phi_m(\{-\tau\})-\phi_m(\{1\})-\phi_m(\{-\tau'\})}(-1)^{\phi_n(\{\tau\})-\phi_n(\{-1\})-\phi_n(\{\tau'\})} \times \left[\frac{\dim\{\tau\}}{\dim\{\tau'\}} \cdot \frac{\dim\{-\tau\}}{\dim\{-\tau'\}} \right]^{1/2} \times (^{\{-\tau\}}_{\{\tau\}}\|z^{\{-1\}}_{\{1\}}\|^{\{-\tau'\}}_{\{\tau'\}})\,. \tag{B3}$$

[1]D. J. Rowe, J. Math. Phys. **25**, 2662 (1984); "Coherent states, contractions and classical limits of the non-compact symplectic groups," in *Proceedings of the XIII International Colloquium on Group Theoretical Methods in Physics*, edited by W. W. Zachary (World Scientific, Singapore, 1984).

[2]D. J. Rowe, G. Rosensteel, and R. Gilmore, J. Math. Phys. **26**, 2787 (1985).

[3]D. J. Rowe, R. Le Blanc, and K. T. Hecht, J. Math. Phys. **29**, 287 (1988).

[4]R. Le Blanc and D. J. Rowe, J. Math. Phys. **29**, 758, 767 (1988).

[5]J. Deenen and C. Quesne, J. Math. Phys. **25**, 1638, 2354 (1984); **26**, 2705 (1985); K. T. Hecht, and J. P. Elliott, Nucl. Phys. A **438**, 29 (1985); C. Quesne, J. Math. Phys. **27**, 428, 869 (1986); R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **19**, 1111 (1986); D. J. Rowe, and J. Carvalho, Phys. Lett. B **175**, 243 (1986); K. T. Hecht, R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **20**, 257 (1987); K. T. Hecht, *The Vector Coherent State Method and its Application to Problems of Higher Symmetry* (Max-Planck-Institut für Kernphysic, Heildelberg, 1987), MPIH-1987, Vol. 19.

[6]K. T. Hecht, R. Le Blanc, and D. J. Rowe, J. Phys. A: Math. Gen. **20**, 2241 (1987).

[7]C. Chevalley, C. R. Acad. Sci. Paris **227**, 1136 (1948); Harish-Chandra Trans. Am. Math. Soc. **70**, 28 (1951).

[8]V. G. Kac, Lect. Notes Math. **626**, 597 (1978).

[9]M. Scheunert, *The Theory of Lie Superalgebras, Lecture Notes in Mathematics*, Vol. 716 (Springer, Berlin, 1979).

[10]R. Le Blanc and D. J. Rowe, "Superfield and matrix realizations of highest weight representations for osp (m/2m)" preprint, 1989.

[11]V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961); I. Segal, in *Mathematical Problems of Relativistic Physics* (Am. Math. Soc., Providence, RI, 1963).

[12]G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1443 (1963); I. M. Gel'fand and M. L. Tseitlin, Dokl. Akad. Nauk. **71**, 825, 1017 (1950).

[13]M. Scheunert, *Proceedings of a NATO Advanced Study Institute on Supersymmetry*, 20–31 August, 1984, Bonn, West Germany, edited by K. Dietz, R. Flume, G. v. Gehlen, and V. Rittenberg (Plenum, New York, 1985), NATO Asi Series, p. 421.

[14]J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer, New York, 1972).

[15]M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 146 (1977).

[16]R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **18**, 1891 (1985).

[17]B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).

[18]M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977).

[19]R. Le Blanc and K. T. Hecht, J. Phys. A: Math. Gen. **20**, 4613 (1987).

[20]T. D. Palev, Funct. Anal. Appl. **21**, 245 (1987); J. Math. Phys. **28**, 2280 (1987).

[21]L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **8**, 89 (1968).

[22]L. C. Biedenharn, A. Giovanninni, and J. D. Louck, J. Math. Phys. **8**, 691 (1967).

Irreducible finite-dimensional representations of the Lie superalgebra gl($n/1$) in a Gel'fand–Zetlin basis

Tchavdar D. Palev[a),b)]
International Centre for Theoretical Physics, Trieste, Italy

(Received 30 September 1988; accepted for publication 4 January 1989)

All finite-dimensional irreducible representations of the general linear Lie superalgebra gl($n/1$) are studied. For each representation, a concept of a Gel'fand–Zetlin basis is defined. Expressions for the transformation of the basis under the action of the generators are written down.

I. INTRODUCTION

In the present paper we study all finite-dimensional irreducible representations of the general linear Lie superalgebra gl($n/1$) for any $n>1$. To this end we first extend the concept of a Gel'fand–Zetlin basis (GZ basis) for this Lie superalgebra (LS) and then write down explicit expressions for the transformation of the basis under the action of the algebra generators. The algebra under consideration is a central extension of the special linear LS sl($n/1$), which is also denoted as $A(n-1,0)$.[1] The latter belongs to the class of the basic Lie superalgebras.[2] Each finite-dimensional irreducible module (fidirmod) of $A(n-1,0)$ is either typical or nontypical.[2] The modules over gl($n/1$), which we consider, are such that they remain irreducible when restricted to sl($n/1$). These modules describe all typical and nontypical representations of this LS in a unified form. In this respect the relations between the irreducible representations of gl($n/1$) and sl($n/1$) are in complete analogy with the corresponding relations for the Lie algebras gl(n) and sl(n).

The properties and the transformation of the GZ basis for gl($n/1$) have been partially announced in Ref. 3. In the present paper we derive all results. Our considerations are based on the material contained in Refs. 4 and 5, where we have studied the finite-dimensional irreducible modules of the special linear LS sl($1/n$).

The Lie superalgebra gl($n/1$) can be defined as the set of all squared $(n+1)$-dimensional matrices, whose rows and columns we label with indices $A,B,C,D,\ldots = 1,2,\ldots,n+1$. As a basis in gl($n/1$) we choose all Weyl matrices e_{AB}, $A,B=1,\ldots,n+1$. Assign to each index A a degree (A), which is zero for $A=1,\ldots,n$ and 1 for $A=n+1$. Then the generator e_{AB} is even (resp. odd), if $(A)+(B)$ is an even (resp. odd) number. The multiplication (= the supercommutator) $[,]$ of gl($n/1$) is given with the linear extension of the relations

$$[e_{AB},e_{CD}] = \delta_{BC}e_{AD} - (-1)^{[(A)+(B)][(C)+(D)]}\delta_{AD}e_{CB}. \tag{1}$$

The even subalgebra gl$(n/1)_0$ of gl($n/1$) is

$$\mathrm{gl}(n/1)_0 = \text{lin. env.}\{e_{ij},e_{n+1,n+1}\,|\,i,j=1,\ldots,n\}, \tag{2}$$

and it is isomorphic to

$$\mathrm{gl}(n/1)_0 = \mathrm{gl}(n)\oplus\mathbb{C}, \tag{3}$$

where

$$\mathrm{gl}(n) = \text{lin. env.}\{e_{ij}\,|\,i,j=1,\ldots,n\}, \tag{4}$$

$$\mathbb{C} = \text{lin. env.}\{e_{n+1,n+1}\}. \tag{5}$$

As an ordered basis in the Cartan subalgebra H of gl($n/1$), we choose $e_{11}, e_{22},\ldots,e_{n+1,n+1}$, and denote by

$$e^1,e^2,\ldots,e^{n+1} \tag{6}$$

the dual to its basis in the space of all linear functionals H^* of H, i.e.,

$$e^A(e_{BB}) = \delta^A_B. \tag{7}$$

The Lie algebra gl(n), defined by (4), contains several copies of the subalgebra gl(k) for $k=1,2,\ldots,n-1$. Unless otherwise stated, by gl(k) we shall understand the subalgebra

$$\mathrm{gl}(k) = \text{lin. env.}\{e_{ij}\,|\,i,j=1,\ldots,k\}. \tag{8}$$

The possibility to introduce a Gel'fand–Zetlin basis in any finite-dimensional irreducible gl(m) module V stems from the following proposition.

Proposition 1: Consider the finite-dimensional irreducible gl(m) module V as a gl($m-1$) module. Then one can always represent V as a direct sum of gl($m-1$) modules,

$$V = \sum_i \oplus V_i, \tag{9}$$

with the following properties. (1) All V_i have different highest weights, i.e., the decomposition (9) is multiplicity-free; (2) All V_i are irreducible gl($m-1$) modules.

Consider the chain of subalgebras

$$\mathrm{gl}(n)\supset\mathrm{gl}(n-1)\supset\cdots\supset\mathrm{gl}(k)\supset\cdots\mathrm{gl}(2)\supset\mathrm{gl}(1), \tag{10}$$

and let

$$V\equiv V(n)\supset V(n-1)\supset\cdots\supset V(k)\supset\cdots\supset V(2)\supset V(1) \tag{11}$$

be a flag of subspaces of the gl(n) fidirmod V, where for each $k=1,\ldots,n$, $V(k)$ is an irreducible gl(k) module. Since any irreducible gl(1) module $V(1)$ is a one-dimensional space, the flag (11) determines a one-dimensional subspace in V. Choose an arbitrary vector in this subspace and denote it by (m). Let

a) Permanent address: Institute for Nuclear Research and Nuclear Energy, boul. Lenin 72, 1184 Sofia, Bulgaria.
b) On leave of absence from the Arnold-Sommerfeld Institute for Mathematical Physics, 3392 Clausthal Zellerfeld, West Germany.

 0022-2488/89/071433-10$02.50

$$\Lambda(k) = m_{1k}e^1 + m_{2k}e^2 + \cdots + m_{kk}e^k \tag{12}$$

be the highest weight of $V(k)$. From (9) it follows that the highest weights,

$$\Lambda(n),\Lambda(n-1),...,\Lambda(2),\Lambda(1), \tag{13}$$

determine up to a multiplicative constant the vector (m). Therefore, one can set

$$(m) = \begin{bmatrix} m_{1n}, & m_{2n}, & \cdot & \cdot & \cdot & m_{n-1,n}, & m_{nn} \\ m_{1,n-1}, & m_{2,n-1}, & \cdot & \cdot & \cdot & m_{n-1,n-1}, & \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ m_{1k}, & m_{2k}, & \cdot & m_{kk} & & & \\ \vdots & \vdots & \vdots & \vdots & & & \\ m_{12}, & m_{22} & & & & & \\ m_{11} & & & & & & \end{bmatrix}, \tag{14}$$

where the k th row $m_{1k}, m_{2k},...,m_{kk}$ gives the signature of $V(k)$, i.e., these numbers are the coordinates of $\Lambda(k)$ [see (12)]. The vectors (14), corresponding to all possible flags (11), constitute a basis in V, which was introduced by Gel'fand and Zetlin[6] and is now called a Gel'fand–Zetlin basis in the gl(n) module V.

In the present paper we show that the above approach can be used in a similar way in order to introduce a basis in every fidirmod W of the LS gl$(n/1)$. The crucial point in this respect stems from the observation (see Proposition 4) that Proposition 1 holds also for any gl$(n/1)$ fidirmod W, considered as a representation space of the Lie algebra gl(n), namely

$$W = \sum_i \oplus W_i, \tag{15}$$

where all W_i are irreducible gl(n) submodules with different highest weights. It may be worthwhile to point out that similar property does not hold for an arbitrary gl(n/m) fidirmod W, $m > 1$, considered as a gl$(n/m-1)$ module. In this more general case some of the gl$(n/m-1)$ submodules W_i in the decomposition (15) may be indecomposible.[7] The gl$(n/m-1)$ highest weight Λ_i of W_i does not carry information of whether W_i is irreducible or indecomposible. Therefore, the very idea to introduce a GZ basis in an arbitrary gl(n/m) fidirmod fails. In this respect the algebras gl$(n/1)$ and gl$(1/m)$ are the only exceptions among all Lie superalgebras gl(n/m).

II. FINITE-DIMENSIONAL IRREDUCIBLE REPRESENTATIONS OF sl$(n/1)$

A. Some abbreviation and notation

LS, LS's—Lie superalgebra, Lie superalgebras.

LA, LA's—Lie algebra, Lie algebras.

Fidirmod(s)—finite-dimensional irreducible module(s).

GZ basis—Gel'fand-Zetlin basis.

lin. env. $\{X\}$—the linear envelope of X.

$\mathbb{C}$—the complex numbers.

$\mathbb{Z}_+$—all non-negative integers.

$[\,,\,]$—product (= supercommutator) in the LS.

Let $m_{ij} \in \mathbb{C}$. Then we set

$$[m_{1,n+1}, m_{2,n+1},..., m_{n,n+1}] = [m]_{n+1}, \tag{16}$$

$$[m_{1k}, m_{2k},..., m_{kk}] = [m]_k, \quad k = 1,...,n, \tag{17}$$

$$[m_{1k} + c, m_{2k} + c,..., m_{kk} + c] = [m + c]_k, \quad c \in \mathbb{C}, \tag{18}$$

$$[m_{1k} \pm \delta_{1i}, m_{2k} \pm \delta_{2i},..., m_{kk} \pm \delta_{ki}] = [m]_k^{\pm i}. \tag{19}$$

Moreover, if $M_{AB} \in \mathbb{C}$, then

$$[M_{1A}, M_{2A},..., M_{AA}] = [M]_A, A = 1,...,n+1, \tag{20}$$

$$[M_{1A} \pm \delta_{1i}, M_{2A} \pm \delta_{2i},..., M_{AA} \pm \delta_{Ai}] = [M]_A^{\pm i}, \tag{21}$$

$$l_{ij} = m_{ij} - i, \tag{22}$$

$$L_{ij} = M_{ij} - i. \tag{23}$$

B. Transformation of the sl$(n/1)$ fidirmods

The special linear LS sl$(n/1)$ is a subalgebra of gl$(n/1)$. It consists of all those $(n+1)$-dimensional squared matrices $a \in$ gl$(n/1)$, whose supertrace (= str) vanishes, i.e.,

$$\text{sl}(n/1) = \left\{ a \,|\, a \in \text{gl}(n/1), \text{str}(a) = \sum_{A=1}^{n+1} (-1)^{(A)} a_{AA} = 0 \right\}. \tag{24}$$

The even subalgebra

$$\text{sl}(n/1)_0 = \text{lin. env.}\{E_{ij} \,|\, E_{ij} = e_{ij} + \delta_{ij} e_{n+1,n+1}, \quad i,j = 1,...,n\} \tag{25}$$

is isomorphic to the general linear Lie algebra gl(n). In this case E_{ij} are the Weyl generators of the algebra:

$$[E_{ij}, E_{kl}] = \delta_{jk} E_{il} - \delta_{li} E_{kj}, \quad i,j,k,l = 1,...,n. \tag{26}$$

We shall denote this particular gl(n) by $\overline{\text{gl}(n)}$, i.e., we set

$$\overline{\text{gl}(n)} = \text{lin. env.}\{E_{ij} \,|\, E_{ij} = e_{ij} + \delta_{ij} e_{n+1,n+1}, \quad i,j, = 1,...,n\}. \tag{27}$$

The notation gl(n) is reserved for [see (4)]

$$\text{gl}(n) = \text{lin. env.}\{e_{ij} \,|\, i,j = 1,...,n\}. \tag{28}$$

In Refs. 4 and 5 we have introduced a basis within each sl$(1/n)$ fidirmod and have written down explicit relations for its transformation under the action of the algebra. Using those results and the circumstance that the mapping $(i,j = 1,...,n)$

$$\varphi(e_{0i}) = e_{n+1,i}, \quad \varphi(e_{i0}) = e_{i,n+1}, \quad \varphi(e_{ij}) = e_{ij} \tag{29}$$

defines an isomorphism of $\mathrm{gl}(1/n)$ onto $\mathrm{gl}(n/1)$, we can write down immediately the corresponding relations for $\mathrm{sl}(n/1)$. We formulate the result as a proposition [see Proposition 14, Eqs. (2.5)–(2.7), (3.64), and (3.66) of Ref. 5].

Proposition 2: The finite-dimensional irreducible modules $W([m]_{n+1})$ of the Lie superalgebra $\mathrm{sl}(n/1)$ are in one-to-one correspondence with the set of all complex n-tuples,

$$[m]_{n+1} \equiv [m_{1,n+1}, m_{2,n+1}, \ldots, m_{n,n+1}],$$

$$m_{i,n+1} - m_{j,n+1} \in \mathbb{Z}_+, \tag{30}$$

for all $i<j=1,\ldots,n$. The $\mathrm{sl}(n/1)$ basis $\Gamma([m]_{n+1})$ in $W([m]_{n+1})$ can be chosen to consist of all patterns,

$$(m) \equiv \begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \\ [m]_i \\ \vdots \\ [m]_2 \\ [m]_{11} \end{bmatrix} \equiv \begin{bmatrix} m_{1,n+1}, & m_{2,n+1}, & \cdot & \cdot & \cdot & m_{n,n+1} \\ m_{1n}, & m_{2n}, & \cdot & \cdot & \cdot & m_{nn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ m_{1i}, & m_{2i}, & \cdot & m_{ii} & & \\ \vdots & \vdots & \vdots & \vdots & & \\ m_{12}, & m_{22} & & & & \\ m_{11} & & & & & \end{bmatrix}, \tag{31}$$

which are consistent with the conditions:

(1) The numbers $m_{1,n+1}, m_{2,n+1}, \ldots, m_{n,n+1}$ are fixed and label the $\mathrm{sl}(n/1)$ module $W([m]_{n+1})$;

(2) $$m_{in} = m_{i,n+1} + \theta_i - \sum_{k=1}^{n} \theta_k, \quad \theta_1, \theta_2, \ldots, \theta_n = 0,1; \tag{32}$$

(3) If $m_{j,n+1} = j-1$, then θ_j is fixed to be either 0 or 1; (33)

(4) $$m_{i,j+1} - m_{ij} \in \mathbb{Z}_+, \quad m_{ij} - m_{i+1,j+1} \in \mathbb{Z}_+ \quad \forall i \leqslant j = 1,\ldots,n-1. \tag{34}$$

The transformation of the $\mathrm{sl}(n/1)$ basis (31) is completely determined from the action of the even generators ($l_{ij} = m_{ij} - i$),

$$E_{kk} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ \vdots \\ m_{11} \end{bmatrix} = (m_{1k} + \cdots + m_{kk} - m_{1,k-1} - \cdots - m_{k-1,k-1}) \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ \vdots \\ m_{11} \end{bmatrix}, \tag{35}$$

$$E_{k,k-1} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ [m]_{k-2} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{j=1}^{k-1} \left| \frac{\Pi_{i=1}^{k} (l_{ik} - l_{j,k-1} + 1) \Pi_{i=1}^{k-2} (l_{i,k-2} - l_{j,k-1})}{\Pi_{i \neq j=1}^{k-1} (l_{i,k-1} - l_{j,k-1} + 1)(l_{i,k-1} - l_{j,k-1})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1}^{-j} \\ [m]_{k-2} \\ \vdots \\ m_{11} \end{bmatrix}, \tag{36}$$

$$E_{k-1,k} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ [m]_{k-2} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{j=1}^{k-1} \left| \frac{\Pi_{i=1}^{k} (l_{ik} - l_{j,k-1}) \Pi_{i=1}^{k-2} (l_{i,k-2} - l_{j,k-1} - 1)}{\Pi_{i \neq j=1}^{k-1} (l_{i,k-1} - l_{j,k-1})(l_{i,k-1} - l_{j,k-1} - 1)} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1}^{j} \\ [m]_{k-2} \\ \vdots \\ m_{11} \end{bmatrix}, \tag{37}$$

and the action of the odd generators,

$$e_{n,n+1} \begin{bmatrix} [m]_{n+1} \\ [m]_n \\ [m]_{n-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{i=1}^{n} (1-\theta_i)(-1)^{\theta_1 + \cdots + \theta_{i-1}} (l_{i,n+1} + 1)^{1/2} \left| \frac{\Pi_{k=1}^{n-1} (l_{k,n-1} - l_{in} - 1)}{\Pi_{k \neq i=1}^{n} (l_{k,n+1} - l_{i,n+1})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ [m-1]_n^i \\ [m-1]_{n-1} \\ \vdots \\ m_{11} - 1 \end{bmatrix}, \tag{38}$$

$$e_{n+1,n}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{i=1}^{n}\theta_i(-1)^{\theta_1+\cdots+\theta_{i-1}}(l_{i,n+1}+1)^{1/2}\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}. \quad (39)$$

The expressions for the other generators [see Ref. 5, Eqs. (3.56) and (3.57)] can be derived from the above relations (35)–(39) and the supercommutation relations (1).

The condition (3) [see (33)] indicates that whenever $m_{j,n+1}=j-1$ and $\theta_j=0,1$, i.e., if one skips (33), then the linear envelope of all sl($n/1$) basis vectors (31) leads to a reducible module. It is a direct sum of two fidirmods. Fixing with (33) θ_j to be 0 or 1, one selects one of the fidirmods.

The possibility to define the basis $\Gamma([m]_{n+1})$ stems from Proposition 4 in Ref. 4, which asserts that the decomposition of $W([m]_{n+1})$ into $\overline{\mathrm{gl}(n)}$ fidirmods $W([m]_n)$ is simple and multiplicity-free:

$$W([m]_{n+1})=\sum_{[m]_n}\oplus W([m]_n). \quad (40)$$

The sum in (40) is over all $\overline{\mathrm{gl}(n)}$ signatures $[m]_n$, which satisfy the conditions (32) and (33). Each signature $[m]_n$ in (40) denotes the coordinates of the $\overline{\mathrm{gl}(n)}$ highest weight $\Lambda([m]_n)$ of $W([m]_n)$ in the basis

$$E^1,E^2,\ldots,E^n, \quad (41)$$

dual to the Cartan basis

$$E_{11},E_{22},\ldots,E_{nn} \quad (42)$$

of $\overline{\mathrm{gl}(n)}$, i.e.,

$$\Lambda([m]_n)=\sum_{i=1}^{n}m_{in}E^i. \quad (43)$$

The basis $\Gamma([m]_{n+1})$ is introduced in a similar way as the GZ basis for gl(n). In this case instead of the chain (10) we take

$$\mathrm{sl}(n/1)\supset\overline{\mathrm{gl}(n)}\supset\overline{\mathrm{gl}(n-1)}\supset\cdots\supset\overline{\mathrm{gl}(k)}\supset\cdots\supset\overline{\mathrm{gl}(1)}, \quad (44)$$

where

$$\overline{\mathrm{gl}(k)}=\text{lin. env.}\{E_{ij}|i,j=1,\ldots,k\}. \quad (45)$$

Each vector (31) is determined by a flag of subspaces

$$W([m]_{n+1})\supset W([m]_n)\supset\cdots\supset W([m]_k)\supset\cdots\supset W(m_{11}), \quad (46)$$

where $W([m]_k)$ is a $\overline{\mathrm{gl}(k)}$ fidirmod with a highest weight

$$\Lambda([m]_k)=\sum_{i=1}^{k}m_{ik}E^i. \quad (47)$$

Thus the rows $[m]_n,\ldots,[m]_k,\ldots,m_{11}$ labeling the sl($n/1$) basis vectors (31) are the coordinates of the highest weights $\Lambda([m]_n),\ldots,\Lambda([m]_k),\ldots,\Lambda(m_{11})$ of the chain (46) corresponding to this vector.

It is natural to expect that the top row $[m]_{n+1}$ in any pattern (31) gives the coordinates of the sl($n/1$) highest weight in $W([m]_{n+1})$. Here, however, this is not the case. The reason for this peculiarity is due to the fact that the basis $\Gamma([m]_{n+1})$ was first introduced for the LS sl($1/n$).[4,5] Here we have described its properties directly in terms of sl($n/1$), using the isomorphism φ of sl($1/n$) onto sl($n/1$) [see (29)]. In the case of sl($1/n$), the weight vector x_Λ with a weight Λ, for which $[m]_n=[m]_{n+1}$ and $m_{kk}=m_{k,k+1}=\cdots=m_{kn}$ for all $k=1,\ldots,n$, is annihilated by all sl($1/n$) simple root vectors,

$$e_{01},e_{12},\ldots,e_{n-1,n}. \quad (48)$$

Therefore x_Λ is the sl($1/n$) highest weight vector in $W([m]_{n+1})$, i.e., in the case of sl($1/n$), the top row of each pattern (31) gives the coordinates of the sl($1/n$) highest weight in $W([m]_{n+1})$. The isomorphism φ, however, does not transform the simple root vectors (48) of sl($1/n$) onto the simple root vectors

$$e_{12},\ldots,e_{n-1,n},e_{n,n+1} \quad (49)$$

of sl($n/1$). Indeed,

$$\varphi(e_{01})=e_{n+1,n} \quad (50)$$

is a negative root vector in sl($n/1$). Consequently x_Λ is not annihilated by the generators (49) and is not, therefore, the sl($n/1$) highest weight vector in the sl($n/1$) module $W([m]_{n+1})$. In the next section we improve this anomaly of the notation, introducing appropriate notations directly for gl($n/1$).

III. FINITE-DIMENSIONAL IRREDUCIBLE REPRESENTATIONS OF gl($n/1$)

A. Gel'fand–Zetlin basis

The first task of the present subsection is to enlarge each sl($n/1$) module $W([m]_{n+1})$ to a gl($n/1$) fidirmod. To this end, it is sufficient to define the transformation of $\Gamma([m]_{n+1})$ under the action, for instance, of the generator

$$I=e_{11}+e_{22}+\cdots+e_{n+1,n+1}\in\mathrm{gl}(n/1), \quad (51)$$

which in the defining realization of gl($n/1$) is the unit matrix.

Since I is a central element in gl($n/1$),

$$[I,\mathrm{sl}(n/1)]=0, \quad (52)$$

and $W([m]_{n+1})$ is an irreducible space under the action of sl($n/1$), the generator I [considered as an operator in $W([m]_{n+1})$] has to be proportional to the unit operator, i.e.,

$$I\begin{bmatrix}[m]_{n+1}\\ \vdots\\ m_{11}\end{bmatrix}=c\begin{bmatrix}[m]_{n+1}\\ \vdots\\ m_{11}\end{bmatrix},\quad c\in\mathbb{C}, \quad (53)$$

for every pattern (31). The relations (35)–(39), together with (53), turn $W([m]_{n+1})$ into an irreducible $\mathrm{gl}(n/1)$ module, which is characterized by the sequence of the numbers

$$(m_{1,n+1},m_{2,n+1},\ldots,m_{n+1,n+1},c)\equiv([m]_{n+1},c).$$

If (53) holds, we set

$$W([m]_{n+1}) = W([m]_{n+1},c),$$
$$\Gamma([m]_{n+1}) = \Gamma([m]_{n+1},c). \tag{54}$$

This solves the problem of how to enlarge any $\mathrm{sl}(n/1)$ fidirmod to a $\mathrm{gl}(n/1)$ fidirmod. We wish, however, more. We wish to introduce within each $W([m]_{n+1},c)$ a basis which will be a natural generalization of the GZ basis for $\mathrm{gl}(n)$ with respect to the chain,

$$\mathrm{gl}(n/1)\supset\mathrm{gl}(n)\supset\mathrm{gl}(n-1)\supset\cdots\supset\mathrm{gl}(k)\supset\cdots\mathrm{gl}(2)\supset\mathrm{gl}(1). \tag{55}$$

Consider for a certain p, $1\leqslant p\leqslant n$, the subalgebras $\overline{\mathrm{gl}(p)}$ and $\mathrm{gl}(p)$ of $\mathrm{gl}(n/1)$:

$$\overline{\mathrm{gl}(p)} = \text{lin. env.}\{e_{ij},E_{kk} = e_{kk}+e_{n+1,n+1}\,|\,i\neq j=1,\ldots,p,\quad k=1,\ldots,p\}, \tag{56}$$
$$\mathrm{gl}(p) = \text{lin. env.}\{e_{ij},e_{kk}\,|\,i\neq j=1,\ldots,p,\quad k=1,\ldots,p\}. \tag{57}$$

Proposition 3: Let W be a (finite-dimensional) module over both $\overline{\mathrm{gl}(p)}$ and $\mathrm{gl}(p)$. Then W is an irreducible $\overline{\mathrm{gl}(p)}$ module iff it is an irreducible $\mathrm{gl}(p)$ module. In the natural ordering $E_{11},\ldots,E_{pp}$ of the Cartan generators of $\overline{\mathrm{gl}(p)}$ and $e_{11},\ldots,e_{pp}$ of $\mathrm{gl}(p)$, both algebras have one and the same highest weight vector, if W is irreducible.

Proof: Since the Cartan generators of $\overline{\mathrm{gl}(p)}$ and $\mathrm{gl}(p)$ commute, the basis in W can be chosen to consist of weight vectors with respect to both algebras. Then the irreducibility of W depends only on the actions of the other generators, which are the same for $\overline{\mathrm{gl}(p)}$ and $\mathrm{gl}(p)$. Hence W is simultaneously irreducible or reducible with respect to both algebras. Moreover, these algebras have the same simple root vectors e_{12}, e_{23},..., $e_{p-1,p}$ and therefore one and the same highest weight vector, if the module W is irreducible.

If W is a fidirmod over $\mathrm{gl}(p)$ and $\overline{\mathrm{gl}(p)}$, then we denote by

$$[M]_p = [M_{1p},\ldots,M_{pp}] \text{ and } [m]_p = [m_{1p},\ldots,m_{pp}] \tag{58}$$

the $\mathrm{gl}(p)$ and the $\overline{\mathrm{gl}(p)}$ signatures (= the coordinates of the highest weight Λ) of W, respectively. In order to compute them, one has to determine the eigenvalues of the Cartan generators on the highest weight vector $x_\Lambda\in W$:

$$e_{ii}x_\Lambda = M_{ip}x_\Lambda,\quad i=1,\ldots,p, \tag{59}$$
$$E_{ii}x_\Lambda = m_{ip}x_\Lambda,\quad i=1,\ldots,p. \tag{60}$$

We write

$$W\equiv W([M]_p)\equiv W([m]_p)\equiv W([M]_p;[m]_p) \tag{61}$$

if we wish to indicate the signature of W with respect to $\mathrm{gl}(p)$, $\overline{\mathrm{gl}(p)}$, or both of them.

Proposition 4: A decomposition,

$$W([m]_{n+1},c) = \sum_s \oplus W([M]_n)_s, \tag{62}$$

of $W([m]_{n+1},c)$ into a direct sum of $\mathrm{gl}(n)$ fidirmods $W([M]_n)_s$ is multiplicity-free, i.e., the $\mathrm{gl}(n)$ signatures $[M]_n$ of the different terms in (62) are different.

Proof: Clearly, each term $W([M]_n)_s$ is a $\overline{\mathrm{gl}(n)}$ module, and hence (Proposition 3) each $W([M]_n)_s$ is a $\overline{\mathrm{gl}(n)}$ fidirmod. Denote its $\overline{\mathrm{gl}(n)}$ signature by $[m]_n$. Then [see (61)]

$$W([M]_n)_s = W([M]_n;[m]_n)_s, \tag{63}$$

and we can write (62) as

$$W([m]_{n+1},c) = \sum_s \oplus W([M]_n;[m]_n)_s. \tag{64}$$

Hence (64) can be viewed also as a decomposition of $W([m]_{n+1},c)$ into fidirmods of $\overline{\mathrm{gl}(n)}$. This decomposition is multiplicity-free. Therefore (64) is nothing but the decomposition (40). The latter is unique, since the $\overline{\mathrm{gl}(n)}$ signatures $[m]_n$ of all its terms are different. Then the decomposition (62) is also unique, which is possible only if the $\mathrm{gl}(n)$ signatures $[M]_n$, corresponding to different terms in the sum (62), are also different. This completes the proof.

From the above proposition, it follows that there exists one-to-one correspondence between s, $[M]_n$, and $[m]_n$, i.e., for every two terms in the sum (64)

$$W([M]_n;[m]_n)_s,\quad W([M']_n;[m']_n)_{s'}, \tag{65}$$

any two of the inequalities

$$s\neq s',\quad [M]_n\neq[M']_n,\quad [m]_n\neq[m']_n \tag{66}$$

are a consequence of the third inequality. Therefore, in particular, (62) can be written as

$$W([m]_{n+1},c) = \sum_{[M]_n} \oplus W([M]_n). \tag{67}$$

Consider the chain of subalgebras (55) and let

$$W([m]_{n+1},c)\equiv W(n+1)\supset W(n)\supset\cdots\supset W(k)\supset\cdots\supset W(2)\supset W(1) \tag{68}$$

be a flag of subspaces, where for each $k=1,\ldots,n$, $W(k)$ is a $\mathrm{gl}(k)$ fidirmod. Let

$$[M]_{n+1} = [M_{1,n+1},M_{2,n+1},\ldots,M_{n+1,n+1}] \tag{69}$$

be the $\mathrm{gl}(n/1)$ signature of $W([m]_{n+1},c)$,

$$W([m]_{n+1},c)\equiv W([M]_{n+1}). \tag{70}$$

From Proposition 1 and Proposition 4, it follows that each $W(k)$ in (68) is uniquely defined by its $\mathrm{gl}(k)$ signature $[M]_k$. Therefore an arbitrary flag (68) can be written as

$$W([m]_{n+1},c)\equiv W([M]_{n+1})\supset W([M]_n)\supset\cdots\supset W([M]_k)\supset\cdots\supset W([M]_2)\supset W(M_{11}). \tag{71}$$

This flag defines (up to a multiplicative constant, which we fix) a vector (M). This vector is characterized uniquely by the signatures $[M]_n,\ldots,[M]_k,\ldots,[M]_2,M_{11}$. Therefore, in a complete analogy with the LA $\mathrm{gl}(n)$, we set

$$(M)\equiv\begin{bmatrix}[M]_{n+1}\\ [M]_n\\ \vdots\\ [M]_k\\ \vdots\\ [M]_2\\ M_{11}\end{bmatrix}\equiv\begin{bmatrix}M_{1,n+1}, & M_{2,n+1}, & \cdot & \cdot & \cdot & M_{n,n+1}, & M_{n+1,n+1}\\ M_{1n}, & M_{2n}, & \cdot & \cdot & \cdot & M_{nn} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ M_{1k}, & M_{2k}, & \cdot & M_{kk} & & & \\ \vdots & \vdots & \vdots & \vdots & & & \\ M_{12}, & M_{22} & & & & & \\ M_{11} & & & & & & \end{bmatrix}. \tag{72}$$

Definition: The vectors (72), corresponding to all possible flags (71), constitute a basis in the $\mathrm{gl}(n/1)$ module $W([M]_{n+1})$, which we call a $\mathrm{gl}(n/1)$ Gel'fand–Zetlin basis in $W([M]_{n+1})$, and denote it as $\Gamma([M]_{n+1})$.

According to Proposition 3, every $\mathrm{gl}(k)$ fidirmod $W([M]_k)$ in (71) is also a $\overline{\mathrm{gl}(k)}$ fidirmod. Let its $\overline{\mathrm{gl}(k)}$ signature be $[m]_k$. Then [see (61)]

$$W([M]_k)\equiv W([M]_k;[m]_k), \tag{73}$$

and the flag (71) can be written as

$$\begin{aligned}W([m]_{n+1},c) &= W([M]_{n+1})\supset W([M]_n;[m]_n)\\ &\supset\cdots\supset W([M]_k;[m]_k)\\ &\supset\cdots\supset W([M]_2;[m]_2)\supset W(M_{11};m_{11}).\end{aligned} \tag{74}$$

According to (71) and (72), the flag (74) defines a GZ basis vector (M); according to (46) it defines a basis vector $(m)\in\Gamma([m]_{n+1},c)$. Therefore $(M)=(m)$, i.e.,

$$\begin{bmatrix}[M]_{n+1}\\ [M]_n\\ \vdots\\ [M]_k\\ \vdots\\ [M]_2\\ M_{11}\end{bmatrix}=\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\\ [m]_k\\ \vdots\\ [m]_2\\ m_{11}\end{bmatrix}. \tag{75}$$

An essential part of what remains to be done is (1) to find the relations between the labels of the GZ basis vector (M) and the labels of (m); (2) to write down the transformations (35)–(39) in terms of the notations of the GZ basis.

B. Transformations of the GZ basis

We first proceed to establish a connection between the different labels of the vector (75). From (25) and (51) one derives

$$e_{n+1,n+1}=(n-1)^{-1}(E_{11}+\cdots+E_{nn}+I). \tag{76}$$

Let $(m)_N\in\Gamma([m]_{n+1},c)$ be a vector (31) of degree N (Ref. 4, Def. 3), i.e., the θ-tuple of this vector is such that

$$\theta_1+\theta_2+\cdots+\theta_n=N. \tag{77}$$

The relation (35) yields

$$(E_{11}+\cdots+E_{nn})(m)_N=\left[\sum_{i=1}^{n}m_{in}\right](m)_N,$$

which, using (32), can be written as

$$(E_{11}+\cdots+E_{nn})(m)_N=\left[\sum_{i=1}^{n}m_{i,n+1}-(n-1)N\right](m)_N. \tag{78}$$

From (53), (76), and (77) we derive that

$$e_{n+1,n+1}(m)_N=(d-N)(m)_N, \tag{79}$$

where we have replaced the constant c by another constant,

$$d=(n-1)^{-1}(m_{1,n+1}+\cdots+m_{n,n+1}-c). \tag{80}$$

Taking into account that $E_{ii}=e_{ii}+e_{n+1,n+1}$, $i=1,\ldots,n$, from (35) and (79) we receive

$$e_{ii}(m)_N=\left[\sum_{k=1}^{i}m_{ki}-\sum_{k=1}^{i-1}m_{k,i-1}+N-d\right](m)_N. \tag{81}$$

Consider the vector (of degree n)

$$(m)_n=\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\\ [m]_i\\ \vdots\\ [m]_2\\ m_{11}\end{bmatrix}\in\Gamma([m]_{n+1},c), \tag{82}$$

for which

$$\begin{aligned}&\theta_1=\theta_2=\cdots=\theta_n=1,\\ &m_{ii}=m_{i,i+1}=\cdots=m_{in}=m_{i,n+1}+1-n,\\ &\qquad\forall i=1,\ldots,n,\end{aligned} \tag{83}$$

i.e., for all $i=1,\ldots,n$,

$$\begin{aligned}[m]_i=[&m_{1,n+1}+1-n,m_{2,n+1}\\ &+1-n,\ldots,m_{i,n+1}+1-n].\end{aligned} \tag{84}$$

Using Eqs. (37) and (38), and taking into account that $E_{k-1,k}=e_{k-1,k}$, one derives that $(m)_n$ is annihilated by all simple root vectors (49),

$$e_{i,i+1}(m)_n=0,\quad\forall i=1,\ldots,n. \tag{85}$$

Therefore, $(m)_n$ is the $\mathrm{gl}(n/1)$ [and also $\mathrm{sl}(n/1)$] highest weight vector. In the case of $(m)_N=(m)_n$, the relations (81) and (79) reduce to

$$\begin{aligned}&e_{ii}(m)_n=(m_{i,n+1}+1-d)(m)_n,\quad i=1,\ldots,n,\\ &e_{n+1,n+1}(m)_n=(d-n)(m)_n.\end{aligned} \tag{86}$$

Since by definition [see (69)]

$$e_{ii}(m)_n=M_{i,n+1}(m)_n$$

and

$$e_{n+1,n+1}(m)_n = M_{n+1,n+1}(m)_n, \tag{87}$$

from (86) and (87) we obtain a relation between the gl($n/1$) signatures $[M]_{n+1}$ and $[m]_{n+1}$,

$$M_{i,n+1} = m_{i,n+1} + 1 - d, \quad \forall i = 1,...,n,$$
$$M_{n+1,n+1} = d - n. \tag{88}$$

In order to relate $[M]_p$ with $[m]_p$, $p = 1,...,n$ for the vector (72), recall that $M_{1p},...,M_{pp}$ is the signature of the gl(p) module $W([M]_p)$ in the flag (71). In other words, $M_{1p},...,M_{pp}$ are the eigenvalues of the generators $e_{11},...,e_{pp}$ on the highest weight vector $x([M]_p)$ of $W([M]_p)$. Since $x([M]_p)$ is also the highest weight vector of gl(p) (Proposition 3), it has to be a vector (31) for which

$$m_{kk} = m_{k,k+1} = \cdots = m_{ki} = \cdots = m_{kp}, \quad \forall k = 1,...,p. \tag{89}$$

Equations (89) are known from the properties of the GZ basis for gl(p). They can be also derived from (37). Inserting (89) into (81), we obtain

$$e_{ii}x([M]_p) = (m_{ip} + N - d)x([M]_p), \quad \forall i = 1,...,p, \tag{90}$$

where [see (77)]

$$\deg(x([M]_p)) = N = \theta_1 + \theta_2 + \cdots + \theta_n. \tag{91}$$

Thus

$$M_{ip} = m_{ip} + N - d, \quad \forall i = 1,...,p. \tag{92}$$

In particular, if $p = n$, (92) and (32) yield

$$M_{in} = m_{i,n+1} + \theta_i - d. \tag{93}$$

Now we are able to draw conclusions about the parameters M_{AB}, $A \leqslant B = 1,...,n+1$, which label the GZ basis (75). From (30) and (88) it follows that the gl($n/1$) signature $M_{1,n+1},M_{2,n+1},...,M_{n+1,n+1}$, which labels the representations' space and is therefore fixed within $W([M]_{n+1})$, can consist of any complex numbers, for which

$$M_{i,n+1} - M_{j,n+1} \in \mathbb{Z}_+ \quad \forall i \leqslant j = 1,...,n. \tag{94}$$

Equations (33) and (88) yield that

$$m_{j,n+1} = j - 1 \text{ iff } M_{j,n+1} + M_{n+1,n+1} = j - n. \tag{95}$$

The irreducible representations for which (95) holds are [when restricted to sl($n/1$)] called nontypical representations.[2] All other irreducible representations are said to be typical.

As a consequence of (93) and (88), one concludes that (32) holds if and only if

$$M_{in} = M_{i,n+1} + \theta_i - 1, \quad \forall i = 1,...,n, \tag{96}$$

where in the typical case

$$\theta_1,\theta_2,...,\theta_n = 0,1. \tag{97}$$

If (95) is fulfilled, then [see (33)] we have to fix θ_j to be either 0 or 1. In view of the choice (83), it is more convenient to set $\theta_j = 1$.

From (34) and (93), it follows that within a given GZ pattern (75), the entries $M_{ij}, i \leqslant j = 1,...,n$ are such that

$$M_{i,j+1} - M_{ij} \in \mathbb{Z}_+, \quad \forall i \leqslant j = 1,...,n-1,$$
$$M_{ij} - M_{i+1,j+1} \in \mathbb{Z}_+. \tag{98}$$

In order to have more natural notation for the highest weight vector [see (83) and (96)], it is convenient to replace θ_i by

$$\varphi_i = 1 - \theta_i. \tag{99}$$

We summarize the results obtained so far for the GZ basis.

Proposition 5: The finite-dimensional irreducible modules $W([M]_{n+1})$ of the Lie superalgebra gl($n/1$) are in one-to-one correspondence with the set of all complex $n+1$ tuples

$$[M]_{n+1} = [M_{1,n+1},M_{2,n+1},...,M_{n+1,n+1}], \tag{100}$$

for which

$$M_{i,n+1} - M_{j,n+1} \in \mathbb{Z}_+, \quad \forall i \leqslant j = 1,...,n. \tag{101}$$

Within a given gl($n/1$) fidirmod $W([M]_{n+1})$ the numbers (100) are fixed. The Gel'fand–Zetlin basis $\Gamma([M]_{n+1})$ in $W([M]_{n+1})$ consists of all patterns,

$$(M) \equiv \begin{bmatrix} [M]_{n+1} \\ [M]_n \\ \vdots \\ [M]_i \\ \vdots \\ [M]_2 \\ M_{11} \end{bmatrix} \equiv \begin{bmatrix} M_{1,n+1}, & M_{2,n+1}, & \cdot & \cdot & \cdot & M_{n,n+1}, & M_{n+1,n+1}, \\ M_{1n}, & M_{2n}, & \cdot & \cdot & \cdot & M_{nn} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ M_{1i}, & M_{2i}, & \cdot & M_{ii} & & & \\ \vdots & \vdots & \vdots & \vdots & & & \\ M_{12}, & M_{22} & & & & & \\ M_{11} & & & & & & \end{bmatrix}, \tag{102}$$

which are consistent with the conditions:

(1) $M_{in} = M_{i,n+1} - \varphi_i, \quad \varphi_1,\varphi_2,...,\varphi_n = 0,1,$ (103)

(2) if $M_{j,n+1} + M_{n+1,n+1} = j - n$, then $\varphi_j = 0$, (104)

(3) $M_{i,j+1} - M_{ij} \in \mathbb{Z}_+, \quad \forall i \leqslant j = 1,...,n-1,$
$M_{ij} - M_{i+1,j+1} \in \mathbb{Z}_+.$ (105)

The gl($n/1$) highest weight vector is the one from (102), for which

$$[M_{1i},M_{2i},...,M_{ii}] = [M_{1,n+1},M_{2,n+1},...,M_{i,n+1}] \tag{106}$$

for all $i = 1,2,...,n$.

Proposition 6: If Eq. (75) holds, then

$$\begin{bmatrix} [m]_{n+1} \\ [m-1]_n^i \\ [m-1]_{n-1} \\ \vdots \\ m_{11} - 1 \end{bmatrix} = \begin{bmatrix} [M]_{n+1} \\ [M]_n^i \\ [M]_{n-1} \\ \vdots \\ M_{11} \end{bmatrix}, \tag{107}$$

and

$$\begin{bmatrix} [m]_{n+1} \\ [m+1]_n^{-i} \\ [m+1]_{n-1} \\ \vdots \\ m_{11}+1 \end{bmatrix} = \begin{bmatrix} [M]_{n+1} \\ [M]_n^{-i} \\ [M]_{n-1} \\ \vdots \\ M_{11} \end{bmatrix}. \tag{108}$$

Proof: According to Proposition 7 in Ref. 4,

$$\text{if } \deg\left(\begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \end{bmatrix}\right) = N, \quad \text{then } \deg\left(\begin{bmatrix} [m]_{n+1} \\ [m \mp 1]_n^{\pm i} \\ \vdots \end{bmatrix}\right) = N+1 \equiv \tilde{N}. \tag{109}$$

Consider (107) and denote

$$\begin{bmatrix} [m]_{n+1} \\ [\tilde{m}]_n \\ [\tilde{m}]_{n-1} \\ \vdots \\ \tilde{m}_{11} \end{bmatrix} \equiv \begin{bmatrix} [m]_{n+1} \\ [m-1]_n^i \\ [m-1]_{n-1} \\ \vdots \\ m_{11}-1 \end{bmatrix} = \begin{bmatrix} [M]_{n+1} \\ [\tilde{M}]_n \\ [\tilde{M}]_{n-1} \\ \vdots \\ \tilde{M}_{11} \end{bmatrix}. \tag{110}$$

According to (92) and (110) [see (18) and (19) for $k = 1,2,...,n$],

$$\tilde{M}_{kn} = \tilde{m}_{kn} + \tilde{N} - d = m_{kn} - 1 + \delta_{ki} + N + 1 - d = M_{kn} + \delta_{ki}, \tag{111}$$

and if $p < n$,

$$\tilde{M}_{kp} = \tilde{m}_{kp} + \tilde{N} - d = m_{kp} - 1 + N + 1 - d = M_{kp}. \tag{112}$$

Therefore

$$[\tilde{M}]_n = [M]_n^i, \quad [\tilde{M}]_p = [M]_p$$

for $p = 1,...,n-1$, which proves (107). In a similar way one proves (108).

Let (M) be an arbitrary GZ pattern (102). Denote by

$$(M)_{\pm ij}, \quad i \leqslant j = 1,...,n \tag{113}$$

the scheme obtained from (M) by the replacement

$$M_{ij} \to M_{ij} \pm 1. \tag{114}$$

Set

$$L_{AB} = M_{AB} - A, \quad A \leqslant B = 1,...,n+1. \tag{115}$$

Then using the relations (88), (92), (99), (107), and (108), we easily rewrite the transformation (35)–(39) in terms of the GZ basis (102). We write the result in somewhat more general form.

The relations written below give transformations of the GZ basis (102), if the new variable q, which we have introduced for convenience, takes value

$$q = \tfrac{1}{2}, \tag{116}$$

$$e_{AA}(M) = \left(\sum_{B=1}^{A} M_{BA} - \sum_{B=1}^{A-1} M_{B,A-1}\right)(M), \quad A = 1,2,...,n+1, \tag{117}$$

$$e_{k,k-1}(M) = \sum_{j=1}^{k-1} \left| \frac{\prod_{i=1}^{k}(L_{ik} - L_{j,k-1} + 1)\prod_{i=1}^{k-2}(L_{i,k-2} - L_{j,k-1})}{\prod_{i\neq j=1}^{k-1}(L_{i,k-1} - L_{j,k-1} + 1)(L_{i,k-1} - L_{j,k-1})} \right|^{1/2} (M)_{-j,k-1}, \tag{118}$$

$$e_{k-1,k}(M) = \sum_{j=1}^{k-1} \left| \frac{\prod_{i=1}^{k}(L_{ik} - L_{j,k-1})\prod_{i=1}^{k-2}(L_{i,k-2} - L_{j,-1} - 1)}{\prod_{i\neq j=1}^{k-1}(L_{i,k-1} - L_{j,k-1})(L_{i,k-1} - L_{j,k-1} - 1)} \right|^{1/2} (M)_{j,k-1}, \tag{119}$$

$$e_{n,n+1}(M) = \sum_{i=1}^{n} \varphi_i(-1)^{i-1}(-1)^{\varphi_1+\cdots+\varphi_{i-1}}(L_{i,n+1} + L_{n+1,n+1} + 2n + 1)^q \times \left| \frac{\prod_{k=1}^{n-1}(L_{k,n-1} - L_{i,n+1})}{\prod_{k\neq i=1}^{n}(L_{k,n+1} - L_{i,n+1})} \right|^{1/2} (M)_{i,n}, \tag{120}$$

$$e_{n+1,n}(M) = \sum_{i=1}^{n} (1-\varphi_i)(-1)^{i-1}(-1)^{\varphi_1+\cdots+\varphi_{i-1}}(L_{i,n+1} + L_{n+1,n+1} + 2n + 1)^{1-q} \times \left| \frac{\prod_{k=1}^{n-1}(L_{k,n-1} - L_{i,n+1})}{\prod_{k\neq i=1}^{n}(L_{k,n+1} - L_{i,n+1})} \right|^{1/2} (M)_{-i,n}. \tag{121}$$

Using the relations (117)–(121) and the supercommutation relations (1), one can write down expressions for the transformations of the GZ basis under the action of the other generators of $\mathrm{gl}(n/1)$. In particular, taking into account that

$$\begin{bmatrix} [m]_{n+1} \\ [m-1]_n^{i_n} \\ \vdots \\ [m-1]_p^{i_p} \\ [m-1]_{p-1} \\ \vdots \\ m_{11}-1 \end{bmatrix} = \begin{bmatrix} [M]_{n+1} \\ [M]_n^{i_n} \\ \vdots \\ [M]_p^{i_p} \\ [M]_{p-1} \\ \vdots \\ M_{11} \end{bmatrix}, \quad \begin{bmatrix} [m]_{n+1} \\ [m+1]_n^{-i_n} \\ \vdots \\ [m+1]_p^{-i_p} \\ [m+1]_{p-1} \\ \vdots \\ m_{11}+1 \end{bmatrix} = \begin{bmatrix} [M]_{n+1} \\ [M]_n^{-i_n} \\ \vdots \\ [M]_p^{-i_p} \\ [M]_{p-1} \\ \vdots \\ M_{11} \end{bmatrix}, \tag{122}$$

using the relations (3.116) and (3.117) from Ref. 4 and going to the second basis [see Sec. III C in Ref. 5], one obtains the expressions for all odd generators ($p = 1,...,n$),

$$e_{p,n+1}\begin{bmatrix}[M]_{n+1}\\ [M]_n\\ \vdots\\ [M]_p\\ [M]_{p-1}\\ \vdots\\ M_{11}\end{bmatrix} = \sum_{i_n=1}^{n}\sum_{i_{n-1}=1}^{n-1}\cdots\sum_{i_p=1}^{p}\varphi_{i_n}(-1)^{\varphi_1+\cdots+\varphi_{i_n-1}}(L_{i_n,n+1}+L_{n+1,n+1}+2n+1)^q$$

$$\times\prod_{r=p+1}^{n}S(i_r,i_{r-1})\left|\frac{\Pi_{k\neq i_{r-1}=1}^{r-1}(L_{k,r-1}-L_{i_r,r}-1)\Pi_{k\neq i_r=1}^{r}(L_{kr}-L_{i_{r-1},r-1})}{\Pi_{k\neq i_r=1}^{r}(L_{kr}-L_{i_r,r})\Pi_{k\neq i_{r-1}=1}^{r-1}(L_{k,r-1}-L_{i_{r-1},r-1}-1)}\right|^{1/2}$$

$$\times(-1)^{i_n-1}\left|\prod_{k\neq i_n=1}^{n}\frac{(L_{kn}-L_{i_n,n})}{(L_{k,n+1}-L_{i_n,n+1})}\right|^{1/2}\left|\frac{\Pi_{k=1}^{p-1}(L_{k,p-1}-L_{i_p,p}-1)}{\Pi_{k\neq i_p=1}^{p}(L_{kp}-L_{i_p,p})}\right|^{1/2}\begin{bmatrix}[M]_{n+1}\\ [M]_n^{i_n}\\ \vdots\\ [M]_p^{i_p}\\ [M]_{p-1}\\ \vdots\\ M_{11}\end{bmatrix}, \tag{123}$$

$$e_{n+1,p}\begin{bmatrix}[M]_{n+1}\\ [M]_n\\ \vdots\\ [M]_p\\ [M]_{p-1}\\ \vdots\\ M_{11}\end{bmatrix} = \sum_{i_n=1}^{n}\sum_{i_{n-1}=1}^{n-1}\cdots\sum_{i_p=1}^{p}(1-\varphi_{i_n})(-1)^{\varphi_1+\cdots+\varphi_{i_n-1}}(L_{i_n,n+1}+L_{n+1,n+1}+2n+1)^{1-q}$$

$$\times\prod_{r=p+1}^{n}S(i_r,i_{r-1})\left|\frac{\Pi_{k\neq i_{r-1}=1}^{r-1}(L_{k,r-1}-L_{i_r,r})\Pi_{k\neq i_r=1}^{r}(L_{kr}-L_{i_{r-1},r-1}+1)}{\Pi_{k\neq i_r=1}^{r}(L_{kr}-L_{i_r,r})\Pi_{k\neq i_{r-1}=1}^{r-1}(L_{k,r-1}-L_{i_{r-1},r-1}+1)}\right|^{1/2}$$

$$\times(-1)^{i_n-1}\left|\prod_{k\neq i_n=1}^{n}\frac{(L_{kn}-L_{i_n,n})}{(L_{k,n+1}-L_{i_n,n+1})}\right|^{1/2}\left|\frac{\Pi_{k=1}^{p-1}(L_{k,p-1}-L_{i_p,p})}{\Pi_{k\neq i_p=1}^{p}(L_{kp}-L_{i_p,p})}\right|^{1/2}\begin{bmatrix}[M]_{n+1}\\ [M]_n^{-i_n}\\ \vdots\\ [M]_p^{-i_p}\\ [M]_{p-1}\\ \vdots\\ M_{11}\end{bmatrix}, \tag{124}$$

where

$$S(i,j)=\begin{cases}1, & \text{for } i<j\\ -1, & \text{for } i>j\end{cases}. \tag{125}$$

The expressions for the even generators e_{ij}, $i\neq j = 1,...,n$, are the same as for the transformation of a gl(n) module with signature $[M]_n$. These expressions are available from the literature.[8]

If $q = 0$, then the Eqs. (117)–(121), (123), and (124) define a representation of gl($n/1$) in the same space $W([M]_{n+1})$. The relations between the GZ basis (corresponding to $q = \frac{1}{2}$) and the $q = 0$ basis can be easily written down and are similar to the relations between the first and the second basis in Ref. 5, Eqs. (3.61) and (3.68).

If $q = 1$, then the basis is defined with all patterns (102) and the conditions (103) and (105). In this case, the representation is indecomposible if

$$M_{j,n+1}+M_{n+1,n+1}=j-n. \tag{126}$$

The maximal invariant subspace consists of all those vectors (102), for which $\varphi_j = 1$.

ACKNOWLEDGMENTS

The author would like to thank Professor Abdus Salam and the mathematical section of ICTP, the International

Atomic Agency, and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste. He is thankful to Professor H. D. Doebner for the valuable discussions on different points of the present investigation.

[1]V. G. Kac, Adv. Math. **26**, 8 (1977).
[2]V. G. Kac, *Lecture Notes in Mathematics*, Vol. 626 (Springer, Berlin, 1978), p. 597.
[3]T. D. Palev, Funkt. Anal. Prilozh. **21**, 85 (1987) [Funct. Anal. Appl. **21**, 245 (1987)].
[4]T. D. Palev, J. Math. Phys. **28**, 2280 (1987).
[5]T. D. Palev, J. Math. Phys. **29**, 2589 (1988).
[6]I. M. Gel'fand and M. L. Zetlin, Dokl. Akad. Nauk SSSR **71**, 825 (1950) (in Russian); see also G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963).
[7]T. D. Palev, "Essentially typical representations of the Lie superalgebras gl(n/m) in a Gel'fand–Zetlin basis," preprint JINR, Dubna, P5-88-169, 1988.
[8]J. D. Louck, Am. J. Phys. **28**, 1 (1970).

The role of discontinuities in conservation laws

F. I. Cooperstock and P. H. Lim
Department of Physics, University of Victoria, Victoria, British Columbia V8W 2Y2, Canada

P. C. Peters
Department of Physics, FM-15, University of Washington, Seattle, Washington 98195

(Received 12 March 1987; accepted for publication 15 March 1989)

Two types of discontinuities are considered in conjunction with the conservation laws: those that arise from a limit of finite and differentiable functions in a boundary layer and those that do not. It is shown that the former, which prevail in electromagnetism, do not play a role in the conservation laws, but the situation in other areas of physics, such as general relativity, may be more complex. A necessary and sufficient condition for the existence of the first type of discontinuity is derived. For the second type, the conservation laws are no longer of the conventional form.

I. INTRODUCTION

In a recent paper,[1] the authors presented various integral theorems for functions with discontinuities. The theorems are of two types: (A) those for "instantaneous" volume integrals, i.e., for integrals defined on constant-time slices, and (B) those for retarded integrals. In this paper the theorems are used to study the conservation laws associated with functions τ^m satisfying[2]

$$\tau^m{}_{,m} = 0. \tag{1}$$

In three-vector notation, (1) has the form

$$\frac{\partial \tau^0}{\partial t} + \nabla\cdot\boldsymbol{\tau} = 0.$$

Should the functions τ^m be continuous everywhere, then the conservation laws associated with (1) are, of course, simply

$$\frac{d}{dt}\int_V \tau^0\, dV = -\oint_S dS_\mu\, \tau^\mu, \tag{2}$$

where the surface S bounds V. The point of interest in this work is to find the corresponding conservation laws for functions τ^m with discontinuities.

For convenience, the theorems of Ref. 1 that are relevant to this paper are noted below in Sec. II. In Sec. III, the conservation laws arising from (1) are derived, and the role played by discontinuities is discussed for the cases of electrodynamics and general relativity. In Sec. IV, it is shown that there exist two classes of discontinuities, and the necessary and sufficient conditions are found for the class that always yields the conventional conservation laws of the form of (2).

II. INTEGRAL THEOREMS WITH DISCONTINUITIES

Consider a time-independent volume V bounded by a surface S in which functions $f(\mathbf{r},t)$ and $\mathbf{F}(\mathbf{r},t)$ are defined to be continuous everywhere, except on a closed time-dependent surface (or a finite set of such surfaces) D. The jumps (or discontinuities) in f and $\mathbf{F}$ on D are denoted by a slash:

$$f| = f_{\text{in}} - f_{\text{ex}},$$
$$\mathbf{F}| = \mathbf{F}_{\text{in}} - \mathbf{F}_{\text{ex}}, \tag{3}$$

where in and ex refer to functions inside and outside D, respectively. The theorems (of type a and b) to be used below are

$$\frac{d}{dt}\int_V f\, dV = \int_V \left(\frac{\partial f}{\partial t}\right) dV + \oint_D d\mathbf{S}\cdot\mathbf{v} f|, \tag{4a}$$

$$\frac{d}{dt}\int_V [f]\, dV = \int_V \left[\frac{\partial f}{\partial t}\right] dV + \oint_D d\mathbf{S}\cdot\left[\frac{\mathbf{v} f}{W}\right|], \tag{4b}$$

where the surface element $d\mathbf{S}$ is outwardly directed from D, $\mathbf{v}$ is the velocity of an element of D, square brackets denote retardation, and W is defined as

$$W = 1 - \mathbf{v}\cdot\hat{R}, \tag{5}$$

where $\hat{R}$ is a unit vector from a source point on D to the field point in question.[3] Equation (4a) has a simple counterpart to a closed time-dependent surface S bounding a volume V, where f is continuous and need not have a discontinuity on S:

$$\frac{d}{dt}\int_V f\, dV = \int_V \left(\frac{\partial f}{\partial t}\right) dV + \oint_S d\mathbf{S}\cdot\mathbf{v} f. \tag{4'}$$

In fact, (4a) can be derived from (4') by taking S to be a surface of discontinuity D, applying (4') first to the volume interior to D, then to the volume exterior to D, and adding the results.

Two forms of Gauss' theorem will also be employed:

$$\int_V (\nabla\cdot\mathbf{F})\, dV = \oint_S d\mathbf{S}\cdot\mathbf{F} + \oint_D d\mathbf{S}\cdot\mathbf{F}|, \tag{6a}$$

$$\int_V [\nabla\cdot\mathbf{F}]\, dV = \oint_S d\mathbf{S}\cdot[\mathbf{F}] + \oint_D d\mathbf{S}\cdot[\mathbf{F}|] - \int_V dV\left[\frac{\partial}{\partial t}(\hat{R}\cdot\mathbf{F})\right]. \tag{6b}$$

It is stressed that the partial derivatives of f and $\mathbf{F}$ in Eq. (4) and (6) are not defined[4] on D; for this reason, the volume of integration excludes the surface D.[1]

 0022-2488/89/071443-04$02.50

III. CONSERVATION LAWS

Suppose the functions τ^m are defined in V to have discontinuities on the surface (or surfaces) D and to satisfy (1) everywhere except on D; on this surface, partial derivatives of τ^m are not defined. Equations (1) and (4a) show that

$$\frac{d}{dt}\int_V \tau^0\,dV = -\int_V \tau^\mu{}_{,\mu}\,dV + \oint_D d\mathbf{S}\cdot\mathbf{v}\tau^0 \Big| ,$$

and this expression is reduced by (6a) to the form

$$\frac{d}{dt}\int_V \tau^0\,dV = -\oint_S dS_\mu\,\tau^\mu - \oint_D dS_\mu(\tau^\mu - v^\mu\tau^0) \Big| . \quad (7)$$

Thus the D integral in (7) negates the usual interpretation of τ^μ as the μ component of flux of a portion of the quantity $\int_V^P \tau^0\,dV$ across the surface S as in (2) or, alternatively, that the S integral does not account for the entire rate of change within S of the total content of τ^0. It is instructive to consider this D integral for the particular cases of electromagnetic and gravitational fields.

In electromagnetism, the four-current j^m satisfies an equation of the form of (1) and the corresponding D integral in (7) is then

$$\oint_D dS_\mu(j^\mu - v^\mu j^0) \Big| .$$

For each charge species, the four-current components are related to the charge density as[5]

$$j^0 = \rho, \quad j^\mu = \rho v_c^\mu, \quad (8)$$

where v_c^μ is the three-velocity of the respective charge species. On any given D surface, the charge velocity would differ from the velocity of the surface of discontinuity v^μ by, at most, a component tangential to the surface. Hence (8) and the orthogonality of dS_μ with any residual tangential component reduces (7) to the usual form of (2) for charge conservation:

$$\frac{d}{dt}\int_V \rho\,dV = -\oint_S d\mathbf{S}\cdot\mathbf{j}. \quad (9)$$

In general relativity one may define symmetric (nontensorial) functions τ^{lm}, which satisfy equations of the form of (1), i.e.,

$$\tau^{lm}{}_{,m} = 0.$$

The total four-momentum of the system is then

$$P^l = \int_V \tau^{l0}\,dV, \quad (10)$$

where V is all space. The conservation laws (7) then take the form

$$\frac{dP^l}{dt} = -\oint_S dS_\mu\tau^{l\mu} - \oint_D dS_\mu(\tau^{l\mu} - v^\mu\tau^{l0}) \Big| . \quad (11)$$

Although it may be tempting to conjecture that the D integral vanishes in (11), as in the electromagnetic case, the functions τ^{lm} do not satisfy relations of the form of (8). This is because τ^{lm} includes not only a matter part, which does satisfy a relation analogous to (8), but also a pseudotensorial field part which does not.[6] Thus the precise role of discontinuities in the energy-momentum conservation laws of general relativity remains to be found, and this is discussed in some detail in a separate paper.[7]

IV. CONCEPT OF STRUCTURED JUMPS

In Sec. III, the fields τ^m satisfy (1) only inside and outside the surface D; partial derivatives of the τ^m are undefined on the surface of discontinuity. Without further knowledge of τ^m, it would be meaningless to assume that (1) holds on D. One may, however, construct a special class of discontinuities for which (1) could hold on D as follows. Define an infinitesimal boundary layer, of thickness ϵ and volume V_ϵ, containing the surface D. Inside this layer, define auxiliary fields $\bar{\tau}^m$ that are continuous with the fields τ^m on each surface of the boundary layer (i.e., inside and outside D), and are finite and differentiable throughout V_ϵ. Such fields $\bar{\tau}^m$ are henceforth designated as "smoothing fields." (Note that, for any fields τ^m with finite discontinuities on D, the associated smoothing fields $\bar{\tau}^m$ can *always* be defined.) Since the τ^m satisfy (1), here it is taken that the smoothing fields also satisfy this equation in V_ϵ, i.e., $\bar{\tau}^m{}_{,m} = 0$. Finite discontinuities in τ^m are then obtained by taking the limit $\epsilon\to 0$ on the smoothing fields $\bar{\tau}^m$ satisfying Eq. (1). It is convenient to designate such discontinuities as "structured jumps."[8] The necessary and sufficient conditions for the fields τ^m, satisfying (1), to have structured jumps are now found using the theorems of Sec. II.

First, assume that the jumps are structured. Consider an arbitrary open section of D, denoted by D_1, and the boundary layer of thickness ϵ and volume V_ϵ around D_1. Now, the limit $\epsilon\to 0$ of the integral

$$\int_{V_\epsilon} dV\,\bar{\tau}^\mu{}_{,\mu}$$

may be found in two ways. First, using (6a),[9] with the conditions of continuity and finiteness of $\bar{\tau}^m$ it follows that

$$\lim_{\epsilon\to 0}\int_{V_\epsilon}\bar{\tau}^\mu{}_{,u}\,dV = \int_{D_1} dS_\mu(\tau^\mu_{\text{ex}} - \tau^\mu_{\text{in}}) \equiv -\int_{D_1} dS_\mu\,\tau^\mu \Big| . \quad (12)$$

(Note that the continuity of $\bar{\tau}^m$ with τ^m on each surface of the boundary layer, with Gauss' theorem, forces the boundary surface integral of $\bar{\tau}^m$ in the limit $V_\epsilon\to 0$ to become the τ^m surface discontinuity integral because τ^m itself is discontinuous there. Also, since the $\bar{\tau}^m$ are finite, the Gauss-theorem surface contributions with unit normal tangential to the surface D are of order ϵ and vanish on taking the limit.) The left side of (12) could also be found from (4') with Eq. (1) (and the finiteness condition) for the $\bar{\tau}^m$:

$$\lim_{\epsilon\to 0}\int_{V_\epsilon}\bar{\tau}^{\mu}{}_{,\mu}\,dV = -\lim_{\epsilon\to 0}\int_{v_\epsilon}\bar{\tau}^{0}{}_{,0}\,dV$$

$$= -\lim_{\epsilon\to 0}\left\{\frac{d}{dt}\int_{v_\epsilon}\bar{\tau}^{0}\,dV\right\}$$

$$+\lim_{\epsilon\to 0}\int_{S_\epsilon} d\mathbf{S}\cdot\mathbf{v}\bar{\tau}^{0}$$

$$= -\int_{D_1} d\mathbf{S}\cdot\mathbf{v}\tau^{0}\Big|\,, \tag{13}$$

where S_ϵ is the surface of the infinitesimal boundary layer that approaches D_1 in the limit. The limit applied to the S_ϵ integral produces the result in (13), provided the expression

$$\frac{d}{dt}\int_{V_\epsilon}\bar{\tau}^{0}\,dV$$

vanishes in the limit $\epsilon\to 0$. If that is the case, (12) and (13) yield

$$\int_{D_1} dS_\mu(\tau^{\mu}-v^{\mu}\tau^{0})\Big| = 0, \tag{14}$$

and since D_1 is arbitrary, (14) implies that the integrand is orthogonal to the unit normal $\hat{n}_\mu$ on D:

$$\hat{n}_\mu(\tau^{\mu}-v^{\mu}\tau^{0})| = 0. \tag{15}$$

Although the finiteness of the integrand for all times assures that the integral of $\bar{\tau}^0$ is of order ϵ for all times, it is not obvious that, in the absence of any further constraining conditions, the time derivative of this integral must also be of order ϵ.

To analyze this further, assume that τ^m satisfies (1), but *not* the equivalent of (8); i.e., that $\tau^{\mu}=\tau^{0}v^{\mu}$ does not hold. Consider an arbitrary point P on the surface of discontinuity at time t, at which point the surface has a velocity $\mathbf{v}$ with respect to the original reference frame K. Let K' denote the reference frame that is instantaneously comoving with the surface at point P at time t, in which we consider an element of four-volume dU' of thickness ϵ in the direction perpendicular to the surface, area dA' tangential to the surface, and of a time interval dt'. In dU', the structure functions $\bar{\tau}'^{m}$ join smoothly with τ'^{m} at the boundaries and satisfy (1). Thus by Gauss' theorem in four dimensions,

$$0=\int\frac{\partial\bar{\tau}'^{m}}{\partial x'^{m}}\,dU' = \oint\bar{\tau}'^{m}\,dS'_{m} = \int\boldsymbol{\tau}'\cdot\hat{n}'\Big|\,dA'\,dt',$$

where the last expression follows in the limit as $\epsilon\to 0$. Since dA' and dt' are arbitrary, it follows that structured jumps imply that $\boldsymbol{\tau}'\cdot\hat{n}'| = 0$.

This can be expressed in Lorentz-covariant form as follows: Let $\Omega(\mathbf{r},t)=0$ be the equation of the surface of discontinuity, Ω being a scalar. Then in any Lorentz frame, the four-gradient of Ω, evaluated at $\Omega=0$, has components

$$\left(\frac{\partial\Omega}{\partial x^{i}}\right)_{\Omega=0} = |\nabla\Omega|\cdot(-\hat{n}\cdot\mathbf{v},\hat{n}).$$

Thus in K', $\boldsymbol{\tau}'\cdot\hat{n}'| = 0$ is equivalent to

$$\tau'^{i}\frac{\partial\Omega}{\partial x'^{i}}\Big| = 0,$$

since $\mathbf{v}'=\mathbf{0}$ at the given point and time. Moreover, since the latter is a scalar equation, it can be applied in the K frame as well, where it reduces to

$$0=\tau^{i}\frac{\partial\Omega}{\partial x^{i}}| = (\boldsymbol{\tau}-\tau^{0}\mathbf{v})\cdot\hat{n}|\,,$$

which is the result of (15).

This latter approach avoided a confrontation with the evaluation of

$$\frac{d}{dt}\int_{V_\epsilon}\bar{\tau}^{0}\,dV.$$

However, an implicit assumption employed in the latter derivation was that of continuity of the velocity of the surface of discontinuity. A comparison of the two derivations thus shows that the same condition constrains the time derivative of the $\bar{\tau}^0$ integral, in addition to the integral itself, to be of order ϵ.

Equation (15) is the necessary condition for the fields τ^m to have structured jumps. It is interesting to note that (15) shows that structured jumps play no role whatsoever in the conservation laws because the D integral in (7) vanishes identically in that case.

It is now shown that (15) is also a sufficient condition for the τ^m to have structured jumps. Suppose (15) holds. As noted above, one can always define a set of smoothing fields $\bar{\tau}^m$ associated with the τ^m. The above argument leading to (15) can be run backward, and it is easy to show that (15) then implies that the smoothing fields satisfy

$$\bar{\tau}^{m}{}_{,m}=0$$

in V_ϵ. Thus (15) is the necessary and sufficient condition for the τ^m to have structured jumps.

The entire discussion in this section has considered volume integrals on constant time slices. Thus (14) and its consequence (15) are conditions on D for a constant time t. In radiation theory one often has to deal with retarded functions,[6] and it is then important to find the conditions corresponding to (14) and (15) for retarded fields $[\tau^m]$. An argument similar to the one above, but using the retarded theorems (4b) and (6b), shows that the necessary and sufficient conditions for fields $[\tau^m]$, satisfying (1), to have structured jumps is

$$\hat{\hat{n}}_\mu\left[\tau^{\mu}-(v^{\mu}/W)\tau^{0}+(v^{\mu}/W)\,\hat{R}_\alpha\tau^{\alpha}\right]| = 0, \tag{16}$$

where $\hat{\hat{n}}$ is the unit normal to the retarded surface of discontinuity,

$$\tilde{\Omega}(\mathbf{r},t)\equiv\Omega(\mathbf{r},t-R)=0, \quad |\mathbf{R}|\equiv|\mathbf{R}_0-\mathbf{r}|, \tag{17}$$

where, as before, $\Omega(\mathbf{r},t)=0$ is the equation of the discontinuity surface on a constant time slice and $\mathbf{R}_0$ is the vector defining a fixed field point. Thus

$$\hat{\hat{n}}\equiv\nabla\tilde{\Omega}/|\nabla\boldsymbol{\Omega}|. \tag{18}$$

It is straightforward to show that

$$\hat{\hat{n}}_\mu\left[\tau^{\mu}+(v^{\mu}/W)\hat{R}_\alpha\tau^{\alpha}-(v^{\mu}/W)\tau^{0}\right]|$$

$$=\left[|\nabla\Omega|/|\nabla\tilde{\Omega}|\right]\left[\hat{n}_\mu\right]\left[\tau^{\mu}-v^{\mu}\tau^{0}\right]|, \tag{19}$$

and hence the test for the existence of structured jumps with retarded fields, Eq. (16), can be more easily applied by de-

termining whether or not the right-hand side of Eq. (19) vanishes.

These considerations have an immediate application in the formula for gravitational radiation energy loss in general relativity.[6] In that formula, with the vector fields τ^i of this paper replaced by pseudotensor field constructs τ^{ki}, surface integrals appear whose integrands contain factors of the form of the left-hand side of Eq. (19). Thus if the jumps are structured, it immediately follows that all of these surface integrals vanish and the formula is simplified considerably. If τ^i were a four-current, as in electromagnetism, then the vanishing of such effects would be immediate since $\tau^\mu = v_c^\mu \tau^0$. However, in more complicated theories such as general relativity, where the fields serve in a nonlinear way as sources, the situation is not as straightforward.

A simple example is readily constructed in which the orthogonality in Eq. (16) does not hold, but in which both global and local [Eq. (1)] conservation are satisfied. Consider a flow in the x direction of a source confined to a cylinder of cross section A and length $L + M$:

$$\begin{aligned}
&\tau^0 = h_1 x/L, \quad 0 \leqslant x \leqslant L, \\
&\tau^0 = h_2\{x/M + (1 + L/M)\}, \quad L < x \leqslant L + M, \\
&\tau = (\hat{i}/2)(x^2/L)\dot{h}_1, \quad 0 \leqslant x \leqslant L, \\
&\tau = \hat{i}\dot{h}_2\{x^2/2M - x(1 + L/M)\}, \quad L < x \leqslant L + M, \\
&\tau^0 = 0 = \tau, \quad x < 0, \quad x > L + M.
\end{aligned} \tag{20}$$

The local conservation laws (1) are seen to hold, and if

$$L\dot{h}_1 + M\dot{h}_2 = 0, \tag{21}$$

there is also a global conservation of $\int \tau^0 \, dV$. The normal jumps of $\tau|\cdot\hat{n}$ are nonzero at $x = L$ and at $x = L + M$. However, consistently with the conservation law (7), we find that $\oint_D d\mathbf{S}\cdot\boldsymbol{\tau}| = 0$ from the *sum* of the contributions over the faces at $x = L$ and at $x = L + M$. The S integral at infinity vanishes and the D surfaces have no velocity, hence the global conservation demands that the D integrals must sum to zero, which they do consistently under the condition (21). It should be noted, however, that because of the fact that there is effectively a separated source and sink, it is not surprising that the conservation fails to hold under a Lorentz transformation, i.e., the simultaneity of the events at L and at $L + M$ is not absolute.

Such a model would be impossible for ordinary currents: there is no buildup of surface layers of τ^0 at the D surfaces, and hence in a theory such as electromagnetism we would require that the three-current flux have continuity across D. The interesting question is whether or not the analog of such a model could exist with τ^i constructed in a complicated manner from fields, such as the case in general relativity. Sample calculations by the authors on model two-body systems in general relativity have, to this point, revealed only structured jumps with the condition of (16) holding. However, these examples have not yet probed complex field sources to the order where the field has served as part of the source itself. It will be interesting to determine whether nature provides exceptions to (16). It is conceivable that sources in nature that are constructed in a causal manner have a natural currentlike behavior, and so allow jumps that are, at most, structured.

ACKNOWLEDGMENT

The authors are grateful to an anonymous referee for very valuable criticism.

[1] F. I. Cooperstock and P. H. Lim, J. Math. Phys. **27**, 458 (1986).

[2] Latin indices run from 0 to 3 and Greek indices run from 1 to 3. Partial derivatives with respect to the space-time coordinate x^m (with the x^α here being Cartesian coordinates) are denoted by a comma. The notation used throughout this paper is essentially that of Ref. 1.

[3] Retardation in (4b) means setting the time in square-bracketed quantities to $t - R$, where $R = |\mathbf{R}_0 - \mathbf{r}|$ is the distance from the field point $\mathbf{R}_0$ to a source point $\mathbf{r}$ (and $c = 1$).

[4] C. Truesdell, *The Classical Field Theories*, in *Encyclopedia of Physics, Vol. III/1*, edited by S. Flugge (Springer, Berlin, 1960), pp. 491–529.

[5] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 4th ed. (Pergamon, Oxford, 1975), p. 69.

[6] F. I. Cooperstock and P. H. Lim, Phys. Rev. Lett. **55**, 265 (1986); Astrophys. J. **304**, 671 (1986); Can. J. Phys. **64**, 134 (1986).

[7] F. I. Cooperstock and P. H. Lim, Phys. Rev. D **36**, 330 (1987).

[8] Note that jump conditions in electromagnetism are usually deduced by assuming, at least implicity, structured jumps in the fields where Eq. (1) is replaced by Maxwell's equations; see M. Born and E. Wolf, *Principles of Optics*, 2nd ed. (Pergamon, New York, 1959), pp. 4–7. Similar implicit assumptions are made in relativistic fluid mechanics; see A. H. Taub, Annu. Rev. Fluid Mech. **10**, 301 (1978).

[9] For the integral on the left-hand side of (12), Eq. (6a) takes the conventional form of Gauss' theorem without discontinuities prior to the limit $\epsilon \to 0$.

A new perturbative approach to nonlinear problems

C. M. Bender
Department of Physics, Washington University, St. Louis, Missouri 63130

Kimball A. Milton
Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019

Stephen S. Pinsky
Department of Physics, The Ohio State University, Columbus, Ohio 43210

L. M. Simmons, Jr.
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 7 December 1988; accepted for publication 8 March 1989)

A recently proposed perturbative technique for quantum field theory consists of replacing nonlinear terms in the Lagrangian such as ϕ^4 by $(\phi^2)^{1+\delta}$ and then treating δ as a small parameter. It is shown here that the same approach gives excellent results when applied to difficult nonlinear differential equations such as the Lane–Emden, Thomas–Fermi, Blasius, and Duffing equations.

I. INTRODUCTION

A recent series of papers has introduced a new perturbative method specifically designed to confront the nonlinear aspects of quantum field theory.[1–9] The idea is quite simple. Consider, for example, a scalar quantum field theory having a $g\phi^4$ self-interaction term. One replaces this interaction term by another which contains the parameter δ:

$$g\phi^4 \to g(\phi^2)^{1+\delta}.$$

The parameter δ is a measure of the nonlinearity of the self-interaction term. When $\delta = 0$ this term becomes quadratic, the field equations are linear, and the theory (which is now free) can be solved analytically. As δ increases smoothly from zero the nonlinear processes gradually turn on.

The new approach consists of expanding the $g(\phi^2)^{1+\delta}$ theory as a formal perturbation series in powers of δ. Then having computed several terms in this series one can set $\delta = 1$ to obtain numerical results for the $g\phi^4$ theory. The most immediate advantage of this procedure is that the δ-perturbation series usually has a finite (nonzero) radius of convergence.[10] This is surprising because it is normally the case that quantum field theoretic perturbation series have zero radii of convergence. Indeed, the conventional perturbative approach is to expand a $g\phi^4$ theory as a series in powers of the coupling constant parameter g. Such a series is called a weak-coupling expansion and the coefficients of such a series can be represented as sums of Feynman diagrams. It was Dyson[11] who first pointed out that this series has a vanishing radius of convergence. Dyson's argument is simply that for any $g > 0$ the energy is bounded below, but when $g < 0$ the energy is not bounded below. Thus in every neighborhood of $g = 0$ there is an *abrupt* transition from a theory having a ground state to one that does not have a ground state. Therefore, there must be a singularity in the complex-g plane situated at $g = 0$ and the radius of convergence of the perturbation series in powers of g is zero.[12] In contrast, the series in powers of δ is convergent because in the vicinity of $\delta = 0$ there is usually no abrupt transition; the dependence on δ when $|\delta|$ is small is smooth. In slightly different words, a change from $g = 0$ to $g \neq 0$ causes the linear theory to jump to a fully nonlinear theory, no matter how small $|g|$ is. However, in most of the models we have studied, the effect of changing from $\delta = 0$ to $\delta \neq 0$ is not associated with any sudden nonanalytic effects.

While the δ-perturbation expansion method was specifically developed to solve quantum field theory problems, we have come to realize that it can be a powerful tool in the analysis of *any* nonlinear problem. The purpose of this paper is to show how to use the δ expansion to solve nonlinear differential equation problems. We apply it to a panoply of well-known and difficult nonlinear ordinary differential equations and from just a few terms in the δ series, we obtain uniformly excellent numerical results.

This paper is organized as follows. In Sec. II we use the δ expansion to examine some elementary problems; we illustrate the properties of the δ series and demonstrate the kind of numerical results one can expect to obtain. Then, in Secs. III–VI we apply the δ-expansion method in turn to four nonlinear ordinary differential equation problems: The Lane–Emden, Thomas–Fermi, Blasius, and Duffing equations (the classical anharmonic oscillator). In future work we will apply the δ-expansion method to two nonlinear partial differential equations: The Burgers and Korteweg–deVries equations.

II. ELEMENTARY ILLUSTRATIVE EXAMPLES

The idea of introducing a small parameter δ in the exponent of a nonlinear term as a calculational tool is sufficiently new that it is of value to solve some elementary problems. In doing so we hope to achieve at least an intuitive understanding of the perturbative procedure. We consider in this section two elementary illustrative problems: The first involves finding the roots of a fifth-degree polynomial and the second concerns a very elementary nonlinear differential equation.

 0022-2488/89/071447-09$02.50

A. Roots of a fifth-degree polynomial

We are concerned here with finding the real root x_0 of the polynomial equation

$$x^5 + x = 1 . \tag{2.1}$$

We have chosen the degree of this polynomial to be 5 because this is just high enough to be sure that there is no quadrature formula for the roots. However, one can be sure that there is a *unique* real root x_0 and that this root is positive because the function $x^5 + x$ is monotone increasing. Using Newton's method we compute that

$$x_0 = 0.75487767\ldots . \tag{2.2}$$

There are several conventional perturbative approaches that we could use to find x_0. One such approach, which we will call the *weak-coupling* perturbation theory, requires that we introduce a perturbative parameter ϵ in front of the x^5 term:

$$\epsilon x^5 + x = 1 . \tag{2.3}$$

Now, x depends on ϵ and we assume that $x(\epsilon)$ has a formal power series expansion in ϵ:

$$x(\epsilon) = a_0 + a_1\epsilon + a_2\epsilon^2 + a_3\epsilon^3 + \cdots . \tag{2.4}$$

To find the coefficients a_i we substitute (2.4) into (2.3) and expand the result as a series in powers of ϵ. We find that the coefficients a_n are integers which oscillate in sign and grow rapidly as n increases:

$$\begin{aligned} &a_0 = 1, \quad a_1 = -1, \quad a_2 = 5, \quad a_3 = -35 , \\ &a_4 = 285, \quad a_5 = -2530, \quad a_6 = 23751 , \end{aligned} \tag{2.5}$$

etc. In fact, we can find a closed-form expression for a_n valid for all n,

$$a_n = [(-1)^n(5n)!]/[n!(4n+1)!] , \tag{2.6}$$

from which we can determine the radius of convergence R of the series in (2.4):

$$R = 4^4/5^5 = 0.08192 . \tag{2.7}$$

Evidently, if we try to use the weak-coupling series in (2.4) directly to calculate $x(1)$ we will fail miserably. Indeed, using the seven coefficients in (2.5) at $\epsilon = 1$ gives

$$x(1) = \sum_{n=0}^{6} a_n = 21476 ,$$

which is a poor approximation to the true value of $x(1)$ in (2.2)!

Of course, we can improve the prediction enormously by first computing the (3,3) Padé and then evaluating the result at $\epsilon = 1$. Now we obtain the result

$$x(1) = 0.76369 , \tag{2.8}$$

which differs from the correct answer in (2.2) by 1.2%.

A second conventional perturbative approach is to use a strong-coupling expansion. Here, we introduce a perturbative parameter ϵ in front of the x term in (2.1):

$$x^5 + \epsilon x = 1 . \tag{2.9}$$

As before, x depends on ϵ and we assume that $x(\epsilon)$ has a formal series expansion in powers of ϵ:

$$x(\epsilon) = b_0 + b_1\epsilon + b_2\epsilon^2 + b_3\epsilon^3 + \cdots . \tag{2.10}$$

Determining the coefficients of this series is routine and we find that

$$\begin{aligned} &b_0 = 1, \quad b_1 = -\tfrac{1}{5}, \quad b_2 = -\tfrac{1}{25}, \quad b_3 = -1/125 , \\ &b_4 = 0, \quad b_5 = 21/15625, \quad b_6 = 78/78125 , \end{aligned} \tag{2.11}$$

etc. Again, we can find a closed-form expression for b_n valid for all n,

$$b_n = -\{\Gamma[(4n-1)/5]\}/\{5\Gamma[(4-n)/5]n!\} , \tag{2.12}$$

from which we can determine the radius of convergence R of the series in (2.10):

$$R = 5/4^{4/5} = 1.64938\ldots . \tag{2.13}$$

Now, $\epsilon = 1$ lies inside the circle of convergence so it is easy to compute $x(1)$ by summing the series (2.10) directly. Using the coefficients listed in (2.11), we have

$$x(1) = \sum_{n=0}^{6} b_n = 0.75434 , \tag{2.14}$$

which differs from the true result in (23.2) by 0.07%, a vast improvement over the weak-coupling approach.

Now we use the δ-expansion method to find the root x_0. We introduce a small parameter δ in the exponent of the nonlinear term in (2.1),

$$x^{1+\delta} + x = 1 , \tag{2.15}$$

and seek an expansion for $x(\delta)$ as a series in powers of δ:

$$x(\delta) = c_0 + c_1\delta + c_2\delta^2 + c_3\delta^3\ldots . \tag{2.16}$$

The coefficients of this series may be computed easily. The first few are

$$\begin{aligned} &c_0 = \tfrac{1}{2}, \quad c_1 = \tfrac{1}{4}\ln 2, \quad c_2 = -\tfrac{1}{8}\ln 2 , \\ &c_3 = -\tfrac{1}{48}\ln^3 2 + \tfrac{1}{32}\ln^2 2 + \tfrac{1}{16}\ln 2 , \\ &c_4 = \tfrac{1}{32}\ln^3 2 - \tfrac{3}{64}\ln^2 2 - \tfrac{1}{32}\ln 2 , \\ &c_5 = \frac{1}{480}\ln^5 2 - \frac{7}{768}\ln^4 2 - \frac{3}{128}\ln^3 2 + \frac{3}{64}\ln^2 2 \\ &\qquad + \frac{1}{64}\ln 2 , \\ &c_6 = -\frac{1}{192}\ln^5 2 + \frac{35}{1536}\ln^4 2 + \frac{5}{768}\ln^3 2 - \frac{5}{128}\ln^2 2 \\ &\qquad - \frac{1}{128}\ln 2 , \end{aligned}$$

etc.

The radius of convergence of the δ series in (2.16) is 1. A heuristic argument for this conclusion is as follows. The radius of convergence is determined by the location of the nearest singularity of $x(\delta)$ in the complex-δ plane. To find this singularity we differentiate (2.15) with respect to δ and solve the resulting equation for $x'(\delta)$:

$$x'(\delta) = -(x^{1+\delta}\ln x)/[1 + x^{\delta}(1+\delta)] .$$

Since $x(\delta)$ is singular where its derivative ceases to exist we look for zeros of the denominator:

$$1 + x^{\delta}(1+\delta) = 0 .$$

We solve this equation simultaneously with (2.15) to eliminate δ and obtain a single equation satisfied by x:

$$0 = x\ln x + (1-x)\ln(1-x) .$$

The solution to this equation corresponding to the smallest value of $|\delta|$ is $x=0$. From (2.15) we therefore see that $\delta=-1$ is the location of the nearest singularity in the complex-δ plane. In fact, as δ decreases below -1, (2.15) abruptly ceases to have a real root. This abrupt transition accounts for the singularity in the function $x(\delta)$.

Clearly, to compute x_0 it is necessary to evaluate the series (2.16) at $\delta=4$. For this large value of δ we use the coefficients in (2.1) and convert the Taylor series to a (3,3) Padé. Evaluating the Padé at $\delta=4$ gives

$$x(\delta=4)=0.75448\,, \tag{2.17}$$

which differs from the exact answer in (2.2) by 0.05%, the *best* result of the three perturbation series methods we have considered.

The δ series continues to provide excellent numerical results as we increase the order of perturbation theory. If we compute all the coefficients up through c_{12} and then convert (2.16) to a (6,6) Padé we obtain

$$x(\delta=4)=0.75487654\,, \tag{2.18}$$

which differs from x_0 in (2.2) by 0.00015%.

B. Solution of a simple nonlinear differential equation

Consider the nonlinear ordinary differential equation problem

$$y'(x)=[y(x)]^n,\quad y(0)=1\,. \tag{2.19}$$

The exact solution to this problem is

$$y(x)=[1-(n-1)x]^{-1/(n-1)}\,. \tag{2.20}$$

To solve (2.19) approximately using the δ expansion we let $n=1+\delta$ and solve

$$y'(x)=[y(x)]^{1+\delta},\quad y(0)=1\,. \tag{2.21}$$

To solve (2.21) perturbatively we can seek a solution $y(x)$ in the form of a series in powers of δ:

$$y(x)=y_0(x)+\delta y_1(x)+\delta^2 y_2(x)+\cdots\,. \tag{2.22}$$

For example, $y_0(x)$ satisfies the *linear* differential equation problem

$$y_0'=y_0(x),\quad y_0(0)=1\,,$$

whose solution is

$$y_0=e^x\,.$$

Indeed, all functions $y_n(x)$ satisfy *linear* differential equations which are easy to solve. We find that

$$y_1(x)=\tfrac{1}{2}e^x x^2,\quad y_2(x)=e^x\left[\tfrac{1}{3}x^3+\tfrac{1}{8}x^4\right]\,,$$

etc. The reason for using a perturbative approach is that, in general, even when one cannot solve the nonlinear differential equation, the differential equations for the perturbation coefficients $y_0(x)$, $y_1(x)$, $y_2(x)$,... are always linear and therefore can be solved in quadrature form.

For the particular problem (2.21) a closed-form solution exists. Therefore, we can determine the radius of convergence R of the series (2.22):

$$R=1/|x|\,.$$

We have computed the series in (2.22) out to the δ^{10} term. Let us examine the numerical accuracy of the δ series. The exact value of $y(x)$ at $x=\frac{1}{4}$ for the case $n=4$ ($\delta=3$) is

$$y(\tfrac{1}{4})=1.587401\,. \tag{2.23}$$

Directly summing the δ series $\Sigma_0^n\delta^n y_n(\frac{1}{4})$ gives 1.284 when $n=0$ (19% error), 1.404 when $n=1$ (11.5% error), 1.470 when $n=2$ (7.4% error), 1.5099 when $n=3$ (4.9% error), 1.5626 when $n=6$ (1.6% error), and 1.58128 when $n=10$ (0.39% error). We can also compute a Padé approximant from the δ series and then set $\delta=3$. The (3,3) Padé gives 1.58692 (0.03% error) and the (5,5) Padé gives 1.587395 (3.7×10^{-4}% error). It is numerical results such as these that encourage us to use the δ expansion to solve difficult nonlinear differential equations.

III. LANE–EMDEN EQUATION

The Lane–Emden equation is a nonlinear ordinary differential equation which describes the equilibrium density distribution in a self-gravitating sphere of polytropic isothermal gas. It is thus of fundamental importance in the field of stellar structure.[13] The Lane–Emden equation reads as

$$y''(x)+(2/x)y'(x)+[y(x)]^n=0\,. \tag{3.1a}$$

The differential equation (3.1a) must be solved subject to the initial conditions

$$y(0)=1,\quad y'(0)=0\,. \tag{3.1b}$$

The parameter n corresponds to the particular choice of equation of state.

The objective is to find the first zero ξ of $y(x)$ as a function of n. (The radius of the star is proportional to ξ.) The initial-value problem (3.1) can be solved analytically in closed form for the special cases $n=0$, 1, and 5. However, closed-form analytical solutions are not known for any other values of n. Thus it is conventional to obtain ξ using numerical integration. In Ref. 13 one can find a table of ξ for various values of n.

The δ expansion predicts ξ accurately for an entire range of values of n. We let $n=1+\delta$ in (3.1a) and try to solve

$$y''(x)+(2/x)y'(x)+[y(x)]^{1+\delta}=0\,. \tag{3.2}$$

We seek a solution $y(x)$ as a series in powers of δ:

$$y(x)=y_0(x)+\delta y_1(x)+\delta^2 y_2(x)+\delta^3 y_3(x)+\cdots\,. \tag{3.3}$$

Substituting (3.3) into (3.2) and comparing powers of δ, we obtain a sequence of *linear* equations and associated initial conditions. The first few read as

$$y_0''(x)+(2/x)y_0'(x)+y_0(x)=0\,,$$
$$y_0(0)=1,\quad y_0'(0)=0\,, \tag{3.4a}$$
$$y_1''(x)+(2/x)y_1'(x)+y_1(x)=-y_0(x)\ln[y_0(x)]\,,$$
$$y_1(0)=y_1'(0)=0\,, \tag{3.4b}$$
$$y_2''(x)+(2/x)y_2'(x)+y_2(x)$$
$$\quad=-y_1(x)\ln[y_0(x)]-y_1(x)-\tfrac{1}{2}y_0(x)\ln^2[y_0(x)]\,,$$
$$y_2=y_2'(0)=0\,, \tag{3.4c}$$

$$y_3''(x) + (2/x)y_3'(x) + y_3(x)$$
$$= -y_2(x)\ln[y_0(x)] - y_2(x) - y_1^2(x)/2y_0(x)$$
$$-\tfrac{1}{2}y_1(x)\ln^2[y_0(x)] - y_1(x)\ln[y_0(x)]$$
$$-\tfrac{1}{6}y_0(x)\ln^3[y_0(x)], \quad y_3(0) = y_3'(0) = 0. \quad (3.4d)$$

Note that (3.4a) is a *homogeneous* linear equation having inhomogeneous initial conditions, while (3.4b)–(3.4d) are inhomogeneous linear equations satisfying *homogeneous* initial conditions. This pattern is typical of the δ expansion and indeed of perturbation theory in general.

The solution to (3.4a) is

$$y_0(x) = \sin x/x, \quad (3.5)$$

whose first zero is at

$$\xi = \pi. \quad (3.6)$$

A. First-order perturbation theory

We solve (3.4b) using the method of reduction of order. We let

$$y_1(x) = (\sin x/x)u_1(x)$$

and obtain

$$\frac{\sin x}{x}u_1''(x) + \frac{2\cos x}{x}u_1'(x) = -\frac{\sin x}{x}\ln\frac{\sin x}{x},$$

which can be solved by multiplying by the integrating factor $x\sin x$. The final result for $y_1(x)$ is

$$y_1(x) = \frac{\cos x}{2x}\int_0^x ds\,\ln(\sin s) - \frac{\sin x}{2x}\ln\left(\frac{\sin x}{x}\right)$$
$$+\frac{3}{4}\cos x + \frac{\sin x}{4x} - \frac{1}{2}\cos x\ln x$$
$$-(\cos x/4x)Si(2x) - (\sin x/4x)Cin(2x), \quad (3.7)$$

where

$$Si(x) \equiv \int_0^x dt\,\frac{\sin t}{t}, \quad Cin(x) \equiv \int_0^x dt\,\frac{1-\cos t}{t}.$$

Now we can compute the first zero of $y(x)$ as a series in powers of δ. We let

$$\xi = \pi + \delta a.$$

Then

$$a = \pi y_1(\pi)$$
$$= (\pi/2)\ln 2 - 3\pi/4 + (\pi/2)\ln\pi + \tfrac{1}{4}Si(2\pi)$$
$$= 0.885273956.$$

B. Second-order calculation

The second-order calculation requires more algebra than the first-order calculation, but the procedure is routine and straightforward. The result for the location of the zero ξ of $y(x)$ to second order in δ is

$$\pi + 0.885273956\delta + 0.24222\delta^2. \quad (3.8)$$

There are two applications we can make of (3.8). First, we can predict the value of ξ for various values of δ. In Table I we give a comparison between the value of ξ predicted by converting the three-term series in (3.8) to a (1,1) Padé and the true value of ξ obtained numerically and given in Ref. 13. Second, we can use the (1,1) Padé to predict the value of δ for which $\xi = \infty$. The exact value of δ is 4, while the predicted value [the value of δ for which the denominator of the (1,1) Padé vanishes] is 3.65. Note that even for this very large value of δ the series in (3.8) gives a relative error of less than 9%.

TABLE I. Comparison of the predicted values of ξ obtained by converting the δ series in (3.8) to a (1,1) Padé, with the exact value of ξ taken from Ref. 12. The number ξ is the first of the Lane–Emden equation.

δ	(1,1) Padé prediction for ξ	Exact value of ξ
0	π	π
−0.5	2.4465	2.4494
0.5	4.3603	4.3529
1.0	7.0521	6.8969
1.5	17.967	14.972

IV. THOMAS–FERMI EQUATION

In Sec. III we saw how the δ expansion can be used to solve an initial-value problem for a nonlinear differential equation: In this section we use it to solve a delicate nonlinear boundary-value problem. We consider here the Thomas–Fermi equation

$$y'' = y^{3/2}x^{-1/2}, \quad y(0) = 1, \quad y(\infty) = 0, \quad (4.1)$$

which describes the charge density in atoms of high atomic number. The solution to (4.1) can be found numerically with great difficulty: To integrate from $x = 0$ (using Runge–Kutta, for example) we must assume a value for $y'(0)$. If $y'(0)$ is chosen too large the solution will eventually become singular at some finite value of x; the leading behavior of $y(x)$ near this singularity is that of a fourth-order pole

$$y(x) \sim 400a/(x-a)^4, \quad (x \to a).$$

On the other hand, if $y'(0)$ is chosen too small the solution will cross below the x axis at some finite value of x and become complex. The number $y'(0)$ can be regarded as a kind of eigenvalue; at the correct value

$$y'(0) = -1.5880710..., \quad (4.2)$$

the solution (see Fig. 1) decays smoothly and monotonically from $y(0) = 1$ to $y(\infty) = 0$, decaying like $144x^{-3}$ for large x.[14] Finding the value of $y'(0)$ accurately is a tedious process which requires a large amount of computer time. However, this is not surprising because (4.1) is a *global* problem whose solution is determined by boundary data from the widely separated points $x = 0$ and $x = \infty$. By contrast, the Lane–Emden equation is relatively easy to solve numerically because the solution is *locally* determined by initial data given at $x = 0$.

Our objective here is to use the δ expansion to predict the numerical value of $y'(0)$. To do so we introduce δ in such a way that the unperturbed problem has a simple solution:

$$y''(x) = [y(x)]^{1+\delta}x^{-\delta}, \quad y(0) = 1, \quad y(\infty) = 0. \quad (4.3)$$

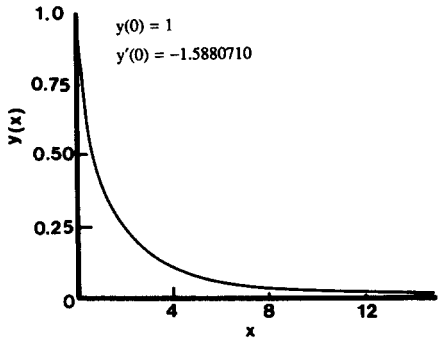


FIG. 1. The solution to the Thomas–Fermi equation $y'' = y^{3/2}x^{-1/2}$, $y(0) = 1$, and $y(\infty) = 0$.

The Thomas–Fermi equation is recovered by setting $\delta = \frac{1}{2}$. Now, if we assume that $y(x)$ can be represented as a series in powers of δ:

$$y(x) = y_0(x) + \delta y_1(x) + \delta^2 y_2(x) + \delta^3 y_3(x) + \cdots, \tag{4.4}$$

then $y_0(x)$ satisfies the boundary-value problem

$$y_0''(x) = y_0(x), \quad y_0(0) = 1, \quad y_0(\infty) = 0, \tag{4.5}$$

whose solution is simply

$$y_0(x) = e^{-x}. \tag{4.6}$$

Thus to leading order in powers of δ we have

$$y'(0) = -1,$$

which is not a bad approximation to the true value of $y'(0)$ in (4.2).

It is important to emphasize here the philosophy of our perturbative approach. Of course, when $\delta > 0$, (4.3) is not analytically solvable because it is a nonlinear equation. However, we expect that as δ varies, the solution $y(x)$ changes slowly and smoothly as a function of δ. Thus the solution at $\delta = 0$ should be a reasonable approximation to the solution at $\delta = \frac{1}{2}$. Furthermore, the solution at $\delta = 0$ is easy to obtain and expand around because, at $\delta = 0$, (4.3) becomes linear. Indeed, the graph of the elementary function $y_0(x) = e^{-x}$ bears a strong resemblance to the exact solution in Fig. 1.

A. First-order calculation

The equation for $y_1(x)$ is

$$y_1''(x) - y_1(x) = e^{-x}\ln(e^{-x}/x), \quad y_1(0) = y_1(\infty) = 0. \tag{4.7}$$

Note that $y_1(x)$ satisfies *homogeneous* boundary conditions. We solve (4.7) by substituting

$$y_1(x) = e^{-x}u_1(x)$$

and solving the equation for $u_1(x)$,

$$u_1''(x) - 2u_1'(x) = \ln(e^{-x}/x),$$

by multiplying by the integrating factor e^{-2x}. The final result is

$$y_1(x) = e^{-x}\int_0^x ds\, e^{2s}\left[y_1'(0) + \int_0^s dt\, e^{-2t}\ln\left(\frac{e^{-t}}{t}\right)\right]. \tag{4.8}$$

The condition that $y_1(\infty) = 0$ implies that

$$\begin{aligned} y_1'(0) &= -\int_0^\infty dt\, e^{-2t}\ln\left(\frac{e^{-t}}{t}\right) \\ &= \frac{1}{4} - \frac{1}{2}\ln 2 - \frac{1}{2}\gamma \\ &= -0.385181423\ldots. \end{aligned} \tag{4.9}$$

Hence, to first order in powers of δ our prediction for $y'(0)$ is

$$-1 + \tfrac{1}{2}y_1'(0) = -1.192590711,$$

which differs from the exact result in (4.2) by 25%.

B. Second-order calculation

Solving for $y_2(x)$ is routine and only slightly more complicated than the first-order calculation. The procedure involves nothing more than repeatedly interchanging orders of integration. The exact formula for $y_2'(0)$ is a simple expression in terms of $y_1'(0)$:

$$y_2'(0) = \tfrac{1}{2}\{-1 + y_1'(0) - [y_1'(0)]^2\} = -0.766773076. \tag{4.10}$$

Hence, our second-order prediction for $y'(0)$ obtained by setting $\delta = \frac{1}{2}$ in $-1 + \delta y_1'(0) + \delta^2 y_2'(0)$ is -1.384283980, which differs from the exact result in (4.2) by 13%.

C. Third-order calculation

The third-order calculation is straightforward, but requires some moderately lengthy algebra. We find that

$$\begin{aligned} y_3'(0) = &-\pi^2/32 - \tfrac{7}{96}\zeta(3) - [y_1'(0)]^2 + y_1'(0) \\ &+ [y_1'(0)]^3/8 + 23/128 = -0.757077189. \end{aligned} \tag{4.11}$$

Thus the third-order prediction for $y'(0)$ is obtained by setting $\delta = \frac{1}{2}$ in

$$-1 + \delta y_1'(0) + \delta^2 y_2'(0) + \delta^3 y_3'(0). \tag{4.12}$$

We obtain -1.478918629, which differs from the exact result in (4.2) by 6.9%.

We can improve this prediction by first converting (4.12) to a (1,2) Padé and then evaluating the Padé at $\delta = \frac{1}{2}$. The result is -1.616287138; now the relative error has decreased to 1.8%. Alternatively, we can compute a (2,1) Padé at $\delta = \frac{1}{2}$. The result is -1.571189843 and the relative error is now 1.1%.

This is an extremely good result for the Thomas–Fermi equation. We know of no other analytical approach to the Thomas–Fermi equation that is productive. Indeed, the Thomas–Fermi equation is quite unique in that the asymptotic methods that one would conventionally use to find an approximate solution to a nonlinear differential equation (a large-x expansion, for example) are dismal failures. The accuracy is so poor that such series are virtually worthless.[14]

V. BLASIUS EQUATION

The Blasius equation is a famous third-order nonlinear differential equation that describes the velocity profile of the fluid in the boundary layer which forms when fluid flows along a flat plate.[15] The Blasius equation reads as

$$y'''(x)+y''(x)y(x)=0,\quad y(0)=y'(0)=0,\quad y'(\infty)=1. \tag{5.1}$$

Equation (5.1) is similar to the Thomas–Fermi equation in that it is posed as a boundary-value problem, but since it is higher order it is even more difficult to solve.

The graph of the numerical solution to this problem begins at $y(0)=0$, rises monotonically with increasing x, and asymptotes at a slope of 1 as $x\to\infty$. From the numerical solution we can determine the value of $y''(0)$:

$$y''(0)=0.46960.... \tag{5.2}$$

It is not known how to calculate this number analytically.

Our objective here is to use the δ expansion to obtain a good approximation to $y''(0)$. We introduce the parameter δ by considering the boundary-value problem

$$y'''(x)+y''(x)[y(x)]^{\delta}=0,\quad y(0)=y'(0)=0,\quad y'(\infty)=1. \tag{5.3}$$

We assume that $y(x)$ has a series expansion in powers of δ:

$$y(x)=y_0(x)+\delta y_1(x)+\delta^2 y_2(x)+\cdots. \tag{5.4}$$

The form of (5.3) is chosen so that the leading term in the series (5.4) satisfies a *linear* boundary-value problem

$$y_0'''(x)+y_0''(x)=0,\quad y_0(0)=y_0'(0)=0,\quad y_0'(\infty)=1. \tag{5.5}$$

The solution to (5.5) is simple:

$$y_0(x)=x-1+e^{-x}. \tag{5.6}$$

Note that $y_0''(0)=1$. This is already a fairly good approximation to $y''(0)$ in (5.2). Moreover, the graph $y_0(x)$ has a strong qualitative resemblance to the exact solution to the Blasius equation.

A. First-order calculation

The function $y_1(x)$ satisfies an *inhomogeneous* linear boundary-value problem

$$y_1'''(x)+y_1''(x)=-y_0''(x)\ln[y_0(x)],\quad y(0)=y_1'(0)=y_1'(\infty)=0. \tag{5.7}$$

Note that $y_0(x)>0$ for $x>0$; thus the argument of the logarithm is never negative.

From (5.7) we obtain directly a formula for the second derivative of $y_1(x)$ at $x=0$:

$$y_1''(0)=\int_0^{\infty} dt\, e^{-t}\ln(t-1+e^{-t}). \tag{5.8}$$

We do not know how to evaluate this integral analytically. (In the Appendix we describe a serious but abortive attempt to evaluate this integral in the form of a series.) However, a numerical integration gives

$$y_1''(0)=-2.1332745. \tag{5.9}$$

We do not obtain a good approximation to $y''(0)$ in (5.2) if we evaluate the δ series

$$1-2.1332745\delta \tag{5.10}$$

at $\delta=1$. Apparently, the radius of convergence of the δ series in this problem is smaller than 1. However, if we convert (5.10) to a (0,1) Padé and evaluate the Padé at $\delta=1$ we obtain

$$\left.\frac{1}{1+2.1332745\delta}\right|_{\delta=1}=0.31915.$$

This is a good approximation to the exact value of $y''(0)$ in (5.2). It differs from the true value of $y''(0)$ by 32%.

B. Second-order calculation

A straightforward but lengthy calculation gives

$$\begin{aligned} y_2''(0)=&\frac{1}{2}[y_1''(0)]^2-y_1''(0)\\ &+\frac{1}{2}+\int_0^{\infty} dt\, te^{-t}\ln[y_0(t)]\\ &+\frac{1}{2}\int_0^{\infty} dt(1-t)e^{-t}\ln^2[y_0(t)]+\int_0^{\infty} dt\,\frac{e^{-t}}{y_0(t)}\\ &\times\int_0^{t} ds\, se^{-s}\ln[y_0(s)]\\ =&\,5.831. \end{aligned} \tag{5.11}$$

We convert the δ series $y_0''(0)+\delta y_1''(0)+\delta^2 y_2''(0)$ to a (1,1) Padé,

$$\frac{1+\delta[y_1''(0)-y_2''(0)/y_1''(0)]}{1-\delta[y_2''(0)/y_1''(0)]}, \tag{5.12}$$

and evaluate the result at $\delta=1$. We obtain 0.429, which differs from the exact answer in (5.2) by a relative error of 8.7%. This is a dramatic improvement over the result obtained to first order in δ.

VI. CLASSICAL ANHARMONIC OSCILLATOR

The classical anharmonic oscillator is defined by the nonlinear ordinary differential equation[16]

$$\frac{d^2y}{dt^2}+y+\epsilon y^3=0, \tag{6.1a}$$

also known as the Duffing equation. We impose the conventional initial conditions

$$y(0)=1,\quad y'(0)=0. \tag{6.1b}$$

Our objective here will be to find the period of the anharmonic oscillator. It is well known that the initial-value problem (6.1) can be solved exactly in terms of elliptic functions and that the period T can be expressed exactly as an elliptic integral[16]

$$T=4\int_0^{\pi/2} d\theta\left[1+\frac{\epsilon}{2}(1+\sin^2\theta)\right]^{-1/2}. \tag{6.2}$$

The integral in (6.2) can be expanded as a series in powers of ϵ:

$$T=2\pi[1+\tfrac{3}{8}\epsilon-(21/256)\epsilon^2+\cdots]^{-1}. \tag{6.3}$$

One cannot use conventional perturbation theory to find the period T for small $|\epsilon|$. It is true that when $|\epsilon|$ is small the exact solution $y(t)$ approximates the motion of a harmonic oscillator of period 2π. However, solving the Duffing equation perturbatively requires some subtlety. If we seek a conventional perturbative solution for $y(t)$ as a series in powers of ϵ we find that there is a resonant coupling between successive orders in perturbation theory. As a result the coefficient of ϵ in the perturbation series for $y(t)$ grows linearly with t, the coefficient of ϵ^2 grows quadratically with t, the coefficient of ϵ^3 grows like t^3, etc. Thus the perturbative solution is only valid for times t which are small compared with $1/\epsilon$. At such short times we cannot use the perturbation expansion for $y(t)$ to obtain the series expansion in (6.3).

More sophisticated perturbative methods have been devised which enable us to calculate $y(t)$ perturbatively for times $t \sim 1/\epsilon$ and thus to obtain the series in (6.3). One such method is called multiple-scale perturbation theory (see Ref. 16).

We will attack (6.1) using the δ expansion and will find that here, too, the methods of multiple-scale perturbation theory must be used. To use the δ expansion we replace y^3 by $y^{1+2\delta}$ and consider the differential equation

$$\frac{d^2y}{dt^2} + y + (\omega^2 - 1)y^{1+2\delta} = 0, \quad y(0) = 1, \quad y'(0) = 0. \tag{6.4}$$

In (6.4) we have found it convenient to set

$$\epsilon = \omega^2 - 1, \tag{6.5}$$

so that when $\delta = 0$, (6.4) describes a classical harmonic oscillator whose frequency is ω. Note, also, that $y^{2\delta}$ is to be interpreted as the positive quantity $(y^2)^\delta$. Thus when we expand $y^{2\delta}$ as a series in powers of δ we obtain

$$y^{2\delta} = 1 + \delta \ln(y^2) + \frac{\delta^2}{2}[\ln(y^2)]^2 + \frac{\delta^3}{6}[\ln(y^2)]^3 + \cdots, \tag{6.6}$$

in which the argument of the logarithm is always positive and no complex numbers appear.

Let us begin by trying to find a conventional perturbative solution to (6.4) as a series in powers of δ:

$$y(t) = \sum_{n=0}^{\infty} \delta^n y_n(t). \tag{6.7}$$

Substituting (6.7) into (6.4) and using (6.6) we obtain a sequence of *linear* equations and associated initial conditions which must be solved. The first few read as

$$\frac{d^2y_0}{dt^2} + \omega^2 y_0 = 0, \quad y_0(0) = 1, \quad y_0'(0) = 0, \tag{6.8a}$$

$$\frac{d^2y_1}{dt^2} + \omega^2 y_1 = -(\omega^2 - 1)y_0 \ln(y_0^2),$$

$$y_1(0) = y_1'(0) = 0, \tag{6.8b}$$

$$\frac{d^2y_2}{dt^2} + \omega^2 y_2 = -(\omega^2 - 1) \times \left\{ y_1 \ln(y_0^2) + 2y_1 + \frac{1}{2} y_0 [\ln(y_0^2)]^2 \right\}, \tag{6.8c}$$

$$y_2(0) = y_2'(0) = 0,$$

$$\frac{d^2y_3}{dt^2} + \omega^2 y_3 = -(\omega^2 - 1)\left\{ y_2 \ln(y_0^2) + 2y_2 + \frac{y_1^2}{y_0} + \frac{1}{2} y_1 [\ln(y_0^2)]^2 + 2y_1 \ln(y_0^2) + \frac{1}{6} y_0 [\ln(y_0^2)]^3 \right\}, \tag{6.8d}$$

$$y_3(0) = y_3'(0) = 0.$$

The solution to (6.8a) is

$$y_0(t) = \cos(\omega t). \tag{6.9}$$

A. Conventional first-order perturbation theory

We can solve (6.8b) using the method of order. We let

$$y_1(t) = \cos(\omega t) u_1(t). \tag{6.10}$$

The equation satisfied by $u_1(t)$ is

$$\cos(\omega t)\frac{d^2u_1}{dt^2} - 2\omega \sin(\omega t)\frac{du_1}{dt} = -(\omega^2 - 1)\cos(\omega t)\ln[\cos^2(\omega t)], \tag{6.11}$$

which has $\cos(\omega t)$ as its integrating factor:

$$\frac{d}{dt}\left[\cos^2(\omega t)\frac{du_1}{dt}\right] = -(\omega^2 - 1)\cos^2(\omega t)\ln[\cos^2(\omega t)]. \tag{6.12}$$

Two integrations of (6.12) give, from (6.10),

$$y_1(t) = -\cos(\omega t)(\omega^2 - 1)\int_0^t \frac{ds}{\cos^2(\omega s)} \times \int_0^s dr \cos^2(\omega r)\ln[\cos^2(\omega r)]. \tag{6.13}$$

The integral with respect to s in (6.13) can be performed by interchanging the orders of integration:

$$\begin{aligned} y_1(t) &= \frac{-\cos(\omega t)(\omega^2 - 1)}{\omega}\int_0^t dr \cos^2(\omega r)\ln[\cos^2(\omega r)] \\ &\quad \times [\tan(\omega t) - \tan(\omega r)] \\ &= \cos(\omega t)[(\omega^2 - 1)/2\omega^2]\{\cos^2(\omega t) - 1 \\ &\quad - \ln[\cos^2(\omega t)]\cos^2(\omega t)\} - \sin(\omega t)\frac{\omega^2 - 1}{\omega^2} \\ &\quad \times \int_0^{\omega t} dx \cos^2 x \ln(\cos^2 x). \end{aligned} \tag{6.14}$$

Note that the integral in (6.14) grows linearly with t for large t because the integrand is a positive periodic function. However, we know that the exact solution to (6.4) is a *bounded* function. Hence, (6.14) can only be valid for times that are short compared with $1/\delta$. This problem appears because each successive order in perturbation theory is resonantly coupled to the previous orders. To see this, note that (6.8b) is the differential equation for a driven harmonic oscillator of natural frequency ω. The driving term (the inhomogeneous part of the differential equation) *also* has frequency ω because it is a functional of y_0. Thus the oscillator described by (6.8b) is driven on resonance and the solution

exhibits secular behavior (it grows with t).

We can still try to use the expression in (6.14) to infer a value for the period of the oscillator accurate to order δ. We assume that the period T of (6.4) is itself a series in powers of δ:

$$T = 2\pi/\omega + (a/\omega)\delta + \cdots . \tag{6.15}$$

To determine the coefficient a in (6.15) we require that after a quarter-period the amplitude of the oscillator

$$y(t) = y_0(t) + \delta y_1(t)$$

will fall from $y = 1$ at $t = 0$ to $y = 0$ at $t = T/4$. Evaluating the expressions for y_0 in (6.9) and y_1 in (6.14) we obtain, to order δ,

$$0 = \cos\left(\frac{\pi}{2} + \frac{a\delta}{4}\right) - \delta\,\frac{\omega^2 - 1}{\omega^2}\int_0^{\pi/2} dx \cos^2 x \ln(\cos^2 x) ,$$

or, neglecting terms which are higher order in δ,

$$\begin{aligned} a &= -4\,\frac{\omega^2 - 1}{\omega^2}\int_0^{\pi/2} dx \cos^2 x \ln(\cos^2 x) \\ &= \pi\,\frac{1-\omega^2}{\omega^2}(1 - 2\ln 2) . \end{aligned} \tag{6.16}$$

Thus to leading order in δ the period of the oscillator is

$$T = (2\pi/\omega)\{1 + \delta[(\omega^2 - 1)/2\omega^2](2\ln 2 - 1)\} . \tag{6.17}$$

B. First-order multiple-scale analysis (MSA)

Let us reexamine the problem in (6.4) using the methods of multiple-scale perturbation theory. We assume that there are two time scales in the problem: a short-time scale described by the variable t and a long-time scale described by the variable

$$\tau \equiv \delta t .$$

We then seek a solution to (6.1) of the form

$$y(t) = Y_0(t,\tau) + \delta Y_1(t,\tau) + \cdots , \tag{6.18}$$

where the initial conditions in (6.1b) become

$$\begin{aligned} &Y_0(0,0) = 1, \quad \frac{\partial Y_0}{\partial t}(0,0) = 0 , \\ &Y_1(0,0) = 0, \quad \frac{\partial Y_0}{\partial \tau}(0,0) + \frac{\partial Y_1}{\partial t}(0,0) = 0 , \end{aligned} \tag{6.19}$$

etc.

If we substitute (6.18) into (6.4) we obtain

$$\frac{\partial^2}{\partial t^2} Y_0(t,\tau) + \omega^2 Y_0(t,\tau) = 0 \tag{6.20a}$$

to zeroth order in δ and

$$\begin{aligned} &\frac{\partial^2}{\partial t^2} Y_1(t,\tau) + \omega^2 Y_1(t,\tau) \\ &\quad = -2\,\frac{\partial^2 Y_0}{\partial t\,\partial \tau} + (1-\omega^2) Y_0 \ln(Y_0^2) \end{aligned} \tag{6.20b}$$

to first order in δ. The most general real solution to (6.20a) is

$$Y_0(t,\tau) = A(\tau)e^{i\omega t} + A^*(\tau)e^{-i\omega t} , \tag{6.21}$$

where as yet $A(\tau)$ is undetermined.

Using (6.21) we can evaluate the rhs of (6.20b):

$$\begin{aligned} &(1-\omega^2)[A(\tau)e^{i\omega t} + A^*(\tau)e^{-i\omega t}] \\ &\quad \times \ln\left[(2|A|^2)\left(1 + \frac{A^2 e^{2i\omega t}}{2|A|^2} + \frac{A^{*2}e^{-2i\omega t}}{2|A|^2}\right)\right] \\ &\quad - 2[i\omega e^{i\omega t}A'(\tau) - i\omega e^{-i\omega t}A^{*\prime}(\tau)] . \end{aligned} \tag{6.22}$$

Now, we expand the logarithm in (6.22) to identify all terms proportional to $e^{i\omega t}$ and $e^{-i\omega t}$. Such terms oscillate at the frequency ω and thus give rise to secular behavior in Y_1. The coefficient of $e^{i\omega t}$ is

$$\begin{aligned} &-2i\omega A'(\tau) + (1-\omega^2)A(\tau)\ln(2|A|^2) \\ &\quad - \frac{(1-\omega^2)}{2} A(\tau) \sum_{k=1}^{\infty} \frac{1}{k} 4^{-k} \begin{bmatrix} 2k \\ k \end{bmatrix} \\ &\quad + \frac{(1-\omega^2)}{2} A(\tau) \sum_{k=0}^{\infty} \frac{1}{2k+1} 4^{-k} \begin{bmatrix} 2k+1 \\ k+1 \end{bmatrix} . \end{aligned} \tag{6.23}$$

Evaluating the sums in (6.23) gives

$$\begin{aligned} &-2i\omega A'(\tau) + (1-\omega^2)A(\tau)\ln(2|A|^2) \\ &\quad - (1-\omega^2)A(\tau)\ln 2 + (1-\omega^2)A(\tau) . \end{aligned} \tag{6.24}$$

Thus the condition that there be no secular behavior in $Y_1(t,\tau)$ is that the expression in (6.24) (as well as its complex conjugate) vanishes:

$$-2i\omega A'(\tau) + (1-\omega^2)A(\tau)[1 + \ln(|A^2|)] = 0 . \tag{6.25}$$

To solve (6.25) we let

$$A(\tau) = R(\tau)e^{i\theta(\tau)} , \tag{6.26}$$

substitute (6.26) into (6.25), and decompose the result into its real and imaginary parts:

$$\begin{aligned} &R'(\tau) = 0 , \\ &\theta'(\tau) = [(\omega^2 - 1)/2\omega](1 + 2\ln R) . \end{aligned} \tag{6.27}$$

Hence, $R(\tau)$ is a constant,

$$R(\tau) = R_0 , \tag{6.28a}$$

and $\theta(\tau)$ is a linear function of τ,

$$\theta(\tau) = [(\omega^2 - 1)/2\omega](1 + 2\ln R_0)\tau + \theta_0 . \tag{6.28b}$$

The initial conditions in (6.19) imply that $R_0 = \frac{1}{2}$ and $\theta_0 = 0$; thus our final result for $Y_0(t,\tau)$ is

$$Y_0(t,\tau) = \cos\{\omega t + \tau[(\omega^2 - 1)/2\omega](1 - 2\ln 2)\} . \tag{6.29}$$

Finally, we eliminate τ in favor of δt to obtain the MSA result

$$T_{\rm MSA} = 2\pi/\{\omega - \delta[(\omega^2 - 1)/2\omega](2\ln 2 - 1)\} , \tag{6.30}$$

which we expand to order δ:

$$\begin{aligned} T_{\rm MSA} &= (2\pi/\omega)\{1 + \delta[(\omega^2 - 1)/\omega^2](2\ln 2 - 1)\} \\ &\quad + O(\delta^2) . \end{aligned} \tag{6.31}$$

To our surprise, (6.31) agrees exactly with the order-δ result we obtained in (6.17) using the δ-perturbation method at a quarter-period.

It is a long but routine calculation to carry the δ-perturbation series out to order δ^2. Using the quarter-period method we find that at $\delta = 1$,

$$T = \frac{1}{\omega}\left[2\pi + 0.5238\,\frac{\omega^2 - 1}{\omega^2} + 0.6041\left(\frac{\omega^2 - 1}{\omega^2}\right)^2\right] . \tag{6.32}$$

TABLE II. Comparison of the exact value of the period of the anharmonic oscillator with the period calculated from the order-δ quarter-period method (same as MSA) and the order-δ^2 quarter-period method.

ϵ	ω	T(exact)	T(order δ)	T(order δ^2)
1	$\sqrt{2}$	4.76802	4.87195	4.73488
3	2	3.52114	3.59669	3.50794
8	3	2.41289	2.45397	2.40871

C. Comparison between exact and approximate results

In Table II we compare three results: the exact numerical calculation of the period T; the order-δ quarter-period calculation, which is the same as the order-δ MSA result in (6.31); and the order-δ^2 quarter-period calculation in (6.32). We set $\delta = 1$ and look at three values of $\epsilon = \omega^2 - 1$. As expected, the MSA's and order-δ results are excellent, having an accuracy of about 2%. The order-δ^2 results are even better, having a relative error of less than 0.5%.

ACKNOWLEDGMENTS

We thank the Aspen Center for Physics, where part of this work was done, for its hospitality.

We also thank the US DOE for partial financial support.

APPENDIX: EVALUATING (5.8) IN SERIES FORM

In this Appendix we describe an interesting attempt to evaluate the integral in (5.8) in the form of a series. Unfortunately, the series we obtain is not rapidly convergent and thus the result is not numerically useful. To date, an analytical evaluation of this integral has eluded us.

The integral in (5.8) is

$$I = \int_0^\infty dt\, e^{-t} \ln(t - 1 + e^{-t}) . \tag{A1}$$

We rewrite this integral as

$$I = -\gamma + \int_0^\infty dt\, e^{-t} \ln\left(1 + \frac{e^{-t} - 1}{t}\right), \tag{A2}$$

where γ is Euler's constant. Expanding the logarithm in (A2) in a series and using the identity

$$\int_0^t ds\, e^{-s} = 1 - e^{-t},$$

we obtain

$$I = -\gamma - \sum_{n=1}^\infty \frac{1}{n} \prod_{i=1}^n \int_0^1 \frac{dx_i}{1 + x_1 + \cdots + x_n} . \tag{A3}$$

The multiple integrals in (A3) can be evaluated as finite sums for all values of n:

$$I = -\gamma - \sum_{n=1}^\infty \sum_{j=1}^n \frac{(j+1)^{n-1} \ln(j+1)(-1)^{n-j}}{j!(n-j)!} . \tag{A4}$$

Next, we replace the upper limit on the j summation in (A4) by ∞ and interchange orders of summation:

$$I = -\gamma - \sum_{j=0}^\infty \frac{(j+1)^{j-1} \ln(j+1)}{j!} \times \sum_{n=0}^\infty \frac{(j+1)^n (-1)^n}{n!} . \tag{A5}$$

The sum on n can be performed explicitly, giving the slowly converging series

$$I = -\gamma - \sum_{j=0}^\infty \frac{(j+1)^{j-1}}{j!} e^{-(j+1)} \ln(j+1) . \tag{A6}$$

It is remarkable that the series

$$f(x) = \sum_{j=0}^\infty \frac{(j+1)^{j-1}}{j!} x^j \tag{A7}$$

is known.[17] Its sum $f(x)$ satisfies

$$f(x) = e^{x f(x)}.$$

The radius of convergence of (A7) is $1/e$. Hence, the sum in (A6) is evaluated at *exactly* the radius of convergence of the series. However, the series in (A6) does converge because for large j the jth term in the series decays like $j^{-3/2} \ln j$. Unfortunately, this convergence rate is too slow for the series to be of much use numerically.

[1] C. M. Bender, K. A. Milton, M. Moshe, S. S. Pinsky, and L. M. Simmons, Jr., Phys. Rev. Lett. **58**, 2615 (1987).

[2] C. M. Bender, K. A. Milton, M. Moshe, S. S. Pinsky, and L. M. Simmons, Jr., Phys. Rev. D **37**, 1472 (1988).

[3] C. M. Bender, K. A. Milton, S. S. Pinsky, and L. M. Simmons, Jr., Phys. Lett. B **205**, 493 (1988); C. M. Bender and K. A. Milton, Phys. Rev. D **38**, 1310 (1988).

[4] S. S. Pinsky and L. M. Simmons, Jr., Phys. Rev. D **38**, 2518 (1988).

[5] C. M. Bender and H. F. Jones, Phys. Rev. D **38**, 2526 (1988).

[6] C. M. Bender and H. F. Jones, J. Math. Phys. **29**, 2659 (1988).

[7] N. Brown, Phys. Rev. D **38**, 723 (1988).

[8] H. F. Jones and M. Monoyios, Imperial College preprint TH/87-88/21.

[9] I. Yotsuyanagi, Kanazawa University preprint DPKU-8809.

[10] When the bare mass is negative the δ-perturbation series appears to be only asymptotic. See H. T. Cho, K. A. Milton, S. S. Pinsky, and L. M. Simmons, Jr., University of Oklahoma preprint, for a discussion of this in zero-dimensional field theory and see J. Cline, S. Pinsky, and L. M. Simmons, Jr., OSU preprint DOE/ER01545/414 for a discussion of this in the context of renormalized self-interacting scalar field theory in two dimensions.

[11] F. J. Dyson, Phys. Rev. **85**, 631 (1952).

[12] A. M. Jaffe, Commun. Math. Phys. **1**, 127 (1965).

[13] S. Chadrasekhar, *An Introduction to the Study of Stellar Structure* (Dover, New York, 1967), Chap. IV.

[14] See C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1978), Chap. 4.

[15] See H. Schlichting, *Boundary Layer Theory* (McGraw-Hill, New York, 1968), p. 129.

[16] See Ref. 13, Chap. 11.

[17] See Ref. 13, Chap. 4.

Dichotomies for linear periodic differential equations with impulses

N. V. Milev and D. D. Bainov
P. O. Box 45, 1504 Sofia, Bulgaria

(Received 4 October 1988; accepted for publication 2 February 1989)

In the present paper the exponential and ordinary dichotomies for linear periodic differential equations with impulses are investigated.

I. INTRODUCTION

In Ref. 1 we investigated for the first time the dichotomies for linear differential equations with impulses. The present paper represents a more detailed investigation of the dichotomies for linear periodic differential equations with impulses at fixed times.

II. PRELIMINARY NOTES

Let the basic vector space be $\mathbb{R}^n$ or $\mathbb{C}^n$ and denote by I the unit matrix. Let $\mathbb{Z}$ be the set of all integers.

Definition 1: The linear differential equation with impulses at fixed times,

$$dx/dt = A(t)x, \quad t \neq t_i, \quad i \in \mathbb{Z}, \quad (1)$$
$$\Delta x|_{t=t_i} = (B_i - I)x,$$

is called periodic with period T if the coefficient matrix $A(t)$ is T-periodic and there exists a positive k such that $t_{i+k} = t_i + T$ and $B_{i+k} = B_i$ for any $i \in \mathbb{Z}$. Without loss of generality we assume that $0 < t_1 < \cdots < t_k \leqslant T$. Moreover we assume that conditions (G) hold:

(G1) The coefficient matrix $A(t)$ is piecewise continuous with points of discontinuity of the first kind for $t = t_i$, $i \in \mathbb{Z}$.

(G2) The constant matrices B_i, $i \in \mathbb{Z}$, are nonsingular.

From (G1) it follows that the fundamental matrix $U(t)$ of the equation $dx/dt = A(t)x$ is continuously differentiable for $t \neq t_i$ with points of discontinuity of the first kind at $t = t_i$.

Let $X(t)$ be the fundamental matrix of Eq. (1), $X(0^+) = I$, where $0^+ = 0$ for $t_k < T$ and $0^+ = 0 + 0$ for $t_k = T$. For $t \in [t_i + 0, t_{i+1} - 0]$, the matrix $X(t)$ admits the representation

$$X(t) = U(t)U^{-1}(t_i + 0)B_i U(t_i - 0)U^{-1}(t_{i-1} + 0) \times B_{i-1} \cdots B_1 U(t_1 - 0)U^{-1}(0^+).$$

Hence the fundamental matrix $X(t)$ is continuously differentiable for $t \neq t_i$ with points of discontinuity of the first kind at $t = t_i$, i.e., $X(t_i + 0) = B_i X(t_i - 0)$. The matrix $X(t + T)$ is also fundamental because $X'(t + T) = A(t + T)X(t + T) = A(t)X(t + T)$, i.e., for any t we have $X(t + T) = X(t)C$ where C is a nonsingular constant matrix defined by the equality $X(T) = X(0^+ + T) = X(0^+)C = C$. Hence $X(t + T) = X(t)X(T)$. The nonsingular matrix $X(T)$ is called the matrix of monodromy. It admits the representation

$$X(T) = U(T^+)U^{-1}(t_k + 0)B_k U(t_k - 0) \times U^{-1}(t_{k-1} + 0)B_{k-1} \cdots B_1 U(t_1 - 0)U^{-1}(0^+),$$

where

$$T^+ = 0^+ + T = \begin{cases} T, & \text{for } t_k < T, \\ T + 0, & \text{for } t_k = T. \end{cases}$$

We shall note that the eigenvalues ζ of the matrix of monodromy $X(T)$ are called multiplicators of the linear periodic differential equation with impulses (1).

III. MAIN RESULTS

Let τ_0 be a fixed real number.

Definition 2 (Ref. 1): Equation (1) has an exponential dichotomy on the interval $[\tau_0, +\infty)$ if for some projector P ($P^2 = P$) there exist positive constants α, β, and L such that:

(D1) $|X(t)X^{-1}(\tau_0)PX(\tau_0)X^{-1}(s)| \leqslant Le^{-\alpha(t-s)}$, for $t \geqslant s \geqslant \tau_0$,

(D2) $|X(t)X^{-1}(\tau_0)(I-P)X(\tau_0)X^{-1}(s)| \leqslant Le^{-\beta(s-t)}$, for $s \geqslant t \geqslant \tau_0$.

Definition 3: If the exponents α and β of Def. 2 are equal to zero, we say that Eq. (1) has an ordinary dichotomy.

Definition 4: The solution of Eq. (1) will be called uniformly bounded away from the zero if for any $\epsilon > 0$ there exists $\delta = \delta(\epsilon)$ such that for any $t \geqslant \tau_0$ and any solution $x(t)$ for which $|x(t)| > \delta$ the inequality $|x(s)| > \epsilon$ holds for $s \geqslant t$.

Remark: Equation (1) has an ordinary (exponential) dichotomy with projector $P = T$ if and only if it is uniformly (asymptotically) stable. If Eq. (1) has an exponential dichotomy with projector $P = 0$, then all solutions tend to infinity uniformly and exponentially. If Eq. (1) has an ordinary dichotomy with projector $P = 0$, then all solutions are uniformly bounded away from the zero.

Lemma 1 (Ref. 1): Let the linear differential equation with impulses (1) satisfy condition (G). If Eq. (1) has an exponential (ordinary) dichotomy on the interval $[\tau_0, +\infty)$ with projector $P(\tau_0)$, then it will have an exponential (ordinary) dichotomy on any interval $[\tau, +\infty)$, $\tau > -\infty$, with projector $P(\tau) = X(\tau)X^{-1}(\tau_0)P(\tau_0)X(\tau_0)X^{-1}(\tau)$.

Proof: For $\tau \geqslant \tau_0$ the assertion is trivial since the fundamental matrix $X(t)$ is invertible.

Let $\tau < \tau_0$. By the inequality of Gronwall–Bellman for $\tau_1, \tau_2 \in [t_m + 0, t_{m+1} - 0]$,

$$|U(\tau_1)U^{-1}(\tau_2)| \leqslant \exp\left|\int_{\tau_1}^{\tau_2} |A(\theta)|d\theta\right|.$$

Let $t > s$ with $t \in [t_i + 0, t_{i+1} - 0]$ and $s \in [t_j + 0, t_{j+1} - 0]$. Then the fundamental matrix $X(t)$ has the form

$$X(t) = U(t)U^{-1}(t_i + 0)B_i U(t_i - 0)U^{-1}(t_{i-1} + 0) \times B_{i-1} \cdots B_{j+1} U(t_j + 0)U^{-1}(s)X(s),$$

hence

 0022-2488/89/071456-03$02.50

$$|X(t)X^{-1}(s)|\leqslant U(t)U^{-1}(t_i+0)\|B_i\|U(t_i-0)U^{-1}(t_{i-1}+0)\|B_{i-1}|\cdots|B_{j+1}\|U(t_j+0)U^{-1}(s)|$$
$$\leqslant \exp\int_{t_i}^{t}|A(\theta)|d\theta\, K_i \exp\int_{t_{i-1}}^{t_i}|A(\theta)|d\theta\, K_{i-1}\cdots K_{j+1}\exp\int_{s}^{t_j}|A(\theta)|d\theta$$
$$\leqslant \prod_{\tau<t_v<\tau_0} K_v \exp\int_{\tau}^{\tau_0}|A(\theta)|d\theta = K,$$

where $K_v = \max(|B_v|,|B_v^{-1}|)\geqslant 1$.

In the same way we prove the validity of the inequality

$$|X(s)X^{-1}(t)|$$
$$\leqslant|U(s)U^{-1}(t_j+0)\|B_{j+1}^{-1}|\cdots|B_{i-1}^{-1}|$$
$$\times|U(t_{i+1}+0)U^{-1}(t_i-0)\|B_i^{-}\|U(t_i+0)U^{-1}(t)|$$
$$\leqslant K_{j+1}\cdots K_{i-1}K_i\exp\int_s^t|A(\theta)|d\theta\leqslant K.$$

Hence for any $t,s\in[\tau,\tau_0]$ the following inequality holds:

$$|X(t)X^{-1}(s)|\leqslant K. \tag{2}$$

If $\tau\leqslant s<\tau_0\leqslant t$, then

$$|X(t)X^{-1}(\tau)P(\tau)X(\tau)X^{-1}(s)|$$
$$=|X(t)X^{-1}(\tau_0)P(\tau_0)X(\tau_0)X^{-1}(s)|$$
$$\leqslant|X(t)X^{-1}(\tau_0)P(\tau_0)X(\tau_0)X^{-1}(\tau_0)|$$
$$\times|X(\tau_0)X^{-1}(s)|\leqslant Le^{-\alpha(t-\tau_0)}K$$
$$\leqslant LKe^{\alpha(\tau_0-\tau)}e^{-\alpha(t-s)}.$$

If $\tau\leqslant s\leqslant t<\tau_0$, then

$$|X(t)X^{-1}(\tau)P(\tau)X(\tau)X^{-1}(s)|$$
$$\leqslant|X(t)X^{-1}(\tau_0)|L\,|X(\tau_0)X^{-1}(s)|$$
$$\leqslant K^2L\leqslant K^2Le^{\alpha(\tau_0-\tau)}e^{-\alpha(t-s)}.$$

Hence for any $t\geqslant s\geqslant\tau$,

$$|X(t)X^{-1}(\tau)P(\tau)X(\tau)X^{-1}(s)|\leqslant L_1e^{-\alpha(t-s)}$$

where $L_1 = LKe^{\alpha(\tau_0-\tau)}\max(1,K)$.

It is analogously verified that for $s\geqslant t\geqslant\tau$,

$$|X(t)X^{-1}(\tau)(I-P(\tau))X(\tau)X^{-1}(s)|\leqslant L_2e^{-\beta(s-t)}$$

where

$$L_2 = LKe^{\beta(\tau_0-\tau)}\max(1,K).$$

This completes the proof of Lemma 1.

Definition 5: The linear differential equations with impulses,

$$dx/dt = A(t)x,\quad t\neq t_i,\quad \Delta x|_{t=t_i} = (B_i-I)x, \tag{3}$$
$$d\tilde{x}/dt = \tilde{A}(t)\tilde{x},\quad t\neq t_i,\quad \Delta\tilde{x}|_{t=t_i} = (\tilde{B}_i-I)\tilde{x}, \tag{4}$$

are called kinematically similar on the interval $[\tau_0,+\infty)$ if between the sets of their solutions a bijective correspondence

$$x(t) = Q(t)\tilde{x}(t),\quad t\in[\tau_0,+\infty), \tag{5}$$

can be established where $Q(t)$ is a bounded matrix having a bounded inverse matrix, i.e., $|Q(t)|\leqslant q_1$, $|Q^{-1}(t)|\leqslant q_2$ for any $t\in[\tau_0,+\infty)$. We shall assume that for Eqs. (3) and (4), condition (G) holds.

Let $X(t)$ and $\tilde{X}(t)$ be the fundamental matrices of Eq. (3) and (4), respectively, for which $X(\tau_0)=\tilde{X}(\tau_0)=I$. By Eq. (5),

$$x(t) = Q(t)\tilde{x}(t) = Q(t)\tilde{X}(t)\tilde{x}(\tau_0),$$
$$x(t) = X(t)x(\tau_0) = X(t)Q(\tau_0)\tilde{x}(\tau_0),$$

whence we obtain that

$$Q(t) = X(t)Q(\tau_0)\tilde{X}^{-1}(t). \tag{6}$$

Equality (6) shows that the matrix $Q(t)$ is continuously differentiable for $t\neq t_i$ with points of discontinuity of the first kind at $t=t_i$. By Eq. (5),

$$x(t_i+0) = Q(t_i+0)\tilde{x}(t_i+0) = Q(t_i+0)\tilde{B}_i\tilde{x}(t_i-0),$$
$$x(t_i+0) = B_ix(t_i-0) = B_iQ(t_i-0)\tilde{x}(t_{i-0}),$$

whence we obtain that $Q(t_i+0) = B_iQ(t_i-0)\tilde{B}_i^{-1}$.

From the differentiability of $Q(t)$ for $t\neq t_i$, it follows that

$$\frac{dx}{dt} = (Q(t)\tilde{x}(t))'$$
$$= Q'(t)\tilde{x}(t) + Q(t)\frac{d\tilde{x}}{dt}$$
$$= (Q'(t) + Q(t)\tilde{A}(t))\tilde{x},$$
$$\frac{dx}{dt} = A(t)x = A(t)Q(t)\tilde{x},$$

i.e., for $t\neq t_i$ the following equality holds:

$$\tilde{A}(t) = Q^{-1}(t)(A(t)Q(t) - Q'(t)). \tag{7}$$

Lemma 2: If Eq. (3) has an exponential (ordinary) dichotomy on the interval $[\tau_0,+\infty)$ with projector P, then the kinematically similar Eq. (4) also has an exponential (ordinary) dichotomy on the interval $[\tau_0,+\infty)$ with a projector $\tilde{P} = Q^{-1}(\tau_0)PQ(\tau_0)$ and the same exponents α and β.

Proof: We express from Eq. (6) $\tilde{X}(t)$, and for $t\geqslant s\geqslant\tau_0$ we verify the validity of condition (D1),

$$|\tilde{X}(t)\tilde{P}\tilde{X}^{-1}(s)|$$
$$=|Q^{-1}(t)X(t)Q(\tau_0)Q^{-1}(\tau_0)PQ(\tau_0)$$
$$\times Q^{-1}(\tau_0)X^{-1}(s)Q(s)|$$
$$\leqslant|Q^{-1}(t)||X(t)PX^{-1}(s)||Q(s)|$$
$$\leqslant q_1q_2Le^{-\alpha(t-s)}.$$

Analogously for $s\geqslant t\geqslant\tau_0$ we verify the validity of condition (D2).

Theorem 1: If the linear periodic differential equation with impulses (1) satisfies condition (G), then it is kinematically similar on the interval $[\tau_0,+\infty)$ to a linear autonomous differential equation without impulses.

Proof: Set

$$H = (1/T)\operatorname{Ln} X(T). \tag{8}$$

This is possible since the matrix of monodromy $X(T)$ is nonsingular, hence it has a logarithm. Obviously, $X(T) = e^{HT}$. We set

$$\Phi(t) = X(t)e^{-Ht}. \tag{9}$$

The matrix $\Phi(t)$ is T-periodic;

$$\Phi(t+T) = X(t+T)e^{-H(t+T)} = X(t)X(T)e^{-HT}e^{-Ht}$$
$$= X(t)e^{-Ht} = \Phi(t),$$

i.e., a representation of Floquet $X(t) = \Phi(t)e^{Ht}$ is valid. From equality (9) it follows that the matrix $\Phi(t)$ is continuously differentiable for $t \neq t_i$ with points of discontinuity of the first kind at $t = t_i$,

$$\Phi(t_i + 0) = X(t_i + 0)e^{-Ht_i} = B_i X(t_i - 0)e^{-Ht_i}$$
$$= B_i \Phi(t_i - 0).$$

Moreover, from equality (9) it follows that

$$\min_{t\in[0,T]} |\det \Phi(t)| \geqslant \delta > 0.$$

In view of the periodicity of $\Phi(t)$ we obtain that $\Phi(t)$ and $\Phi^{-1}(t)$ are bounded, i.e., for any t, $|\Phi(t)| \leqslant q_1$, $|\Phi^{-1}(t)| \leqslant q_2$.

In Eq. (1) we perform the change $x = \Phi(t)\tilde{x}$. For $t \neq t_i$ we obtain that

$$\frac{d\tilde{x}}{dt} = \Phi^{-1}(t)[A(t)\Phi(t) - \Phi'(t)]\tilde{x}$$
$$= e^{Ht}X^{-1}(t)[A(t)X(t)e^{-Ht}$$
$$- X'(t)e^{-Ht} + X(t)e^{-Ht}H]\tilde{x}$$
$$= e^{Ht}X^{-1}(t)[A(t)X(t)e^{-Ht} - A(t)X(t)e^{-Ht}$$
$$+ X(t)e^{-Ht}H]\tilde{x} = H\tilde{x},$$

i.e., the transformed equation is autonomous,

$$d\tilde{x}/dt = H\tilde{x}, \tag{10}$$

and is without impulses because

$$\tilde{x}(t_i + 0) = \Phi^{-1}(t_i + 0)x(t_i + 0)$$
$$= (B_i\Phi(t_i - 0))^{-1}B_i x(t_i - 0)$$
$$= \Phi^{-1}(t_i - 0)x(t_i - 0) = \tilde{x}(t_i - 0).$$

Equality (8) implies that the eigenvalues λ of the constant matrix H are related to the multiplicators ζ of Eq. (1) by means of the equality

$$\operatorname{Re}\lambda(H) = (1/T)\ln|\zeta|. \tag{11}$$

Theorem 2 (Ref. 2, p. 10): Equation (10) has an exponential dichotomy on the interval $[0, +\infty)$ if and only if all eigenvalues of the constant matrix H have nonzero real parts. Equation (10) has an ordinary dichotomy if and only if the eigenvalues of H with zero real parts are semisimple.

Theorem 3: Let the linear differential equation with impulses at fixed times (1) be periodic and satisfy condition (G). Then the following assertions are valid.

(a) Equation (1) is uniformly asymptotically stable if and only if all multiplicators are inside the unit circle.

(b) Equation (1) is uniformly stable if and only if all multiplicators are inside or on the unit circle, and those on the unit circle are semisimple.

(c) Equation (1) has an exponential dichotomy if and only if some of the multiplicators lie inside the unit circle and the rest of them lie outside the unit circle but none on the unit circle.

(d) Equation (1) has an ordinary dichotomy if and only if the multiplicators lying on the unit circle are semisimple.

(e) If all multiplicators are outside the unit circle, then the solution of Eq. (1) tend to infinity uniformly and exponentially.

(f) If all multiplicators are outside or on the unit circle and those on the unit circle are semisimple, then the solutions of Eq. (1) are uniformly bounded away from the zero.

Proof: By Theorem 1, Eq. (1) is kinematically similar to the linear autonomous equation (10) which is without impulses and from equality (11) it follows that the eigenvalues of the constant matrix H have negative, zero, or positive real parts, depending on whether the respective multiplicator lies inside, on, or outside the unit circle, respectively. Hence in view of Theorem 2 and Remark 1, we obtain the assertions of Theorem 3.

ACKNOWLEDGMENT

The present investigation is partially supported by the Ministry of Culture, Science, and Education of the People's Republic of Bulgaria under Grant 61.

[1]N. V. Milev and D. D. Bainov, "Dichotomies of linear differential equations with variable structure and impulse effect" (to be published).
[2]W. A. Coppel, "Dichotomies in Stability Theory," in *Lecture Notes in Mathematics* (Springer, Berlin, 1978).

Solving differential equations by a maximum entropy–minimum norm method with applications to Fokker–Planck equations

James Baker-Jarvis[a] and Michael Racine
Physics Department, North Dakota State University, Fargo, North Dakota 58105
Jihad Alameddine
Electrical Engineering Department, North Dakota State University, Fargo, North Dakota 58105

(Received 6 December 1988; accepted for publication 8 March 1989)

The method of maximum entropy–minimum norm is utilized to produce a general method of solving differential equations. The technique is a generalization and extension of previous work performed by Baker-Jarvis [J. Math. Phys. **30**, 302 (1989)]. It is found that introducing an additional constraint on the norm of the solution vector produces a probability distribution that is integrable over the entire real axis. A number of simplifications occur. In this extended method the Lagrange multipliers and solution vector can be solved for explicitly, thus eliminating the necessity of solving systems of nonlinear equations for the Lagrange multipliers, as was required in the previous approach. It is shown that the solution obtained is equivalent to a minimum norm approximation. The maximum entropy solution of differential equations with Fourier moments is shown to be identical to a Fourier series solution. Additionally, the new method is applied to solving the random walk and Fokker–Planck equations.

I. INTRODUCTION

Maximum entropy methods (MAXENT) are very useful for approximating solutions to systems where there is a general paucity of data or nonunique solutions. The concept of information entropy originated with Shannon[1] as an algorithm for estimating the uncertainty in a signal. Jaynes[2,3] extended this work to statistical mechanics and data reduction by maximizing the information entropy. Since then many researchers have used the technique in a wide array of applications.[4,8] The maximum entropy algorithm determines a probability distribution for the data set which is maximally noncommittal with respect to missing information. The technique is particularly useful when the data is incomplete and/or noisy, in which case the method yields the most objective estimate consistent with *a priori* information.

It has been demonstrated in past research by Baker-Jarvis[9,10] that it is possible to approximate solutions to differential equations by a maximum entropy method. The method has been developed for both linear and nonlinear differential equations and provides a viable alternative to classical approaches. The method proceeds by finding the probability density distribution for the solution vector subject to moment constraints derived from the differential equation. In the previous method the probability density distribution is not integrable over the region $[-\infty, \infty]$ since the exponential in the probability contains only terms linear in the solution vector; thus a finite range of integration is utilized. Recently Poon[11] has found that a probability distribution that contains moments integrable to all orders over $[-\infty, \infty]$ can be obtained if a constraint on the norm of the solution vector is given. In this paper the results of Poon[11] are derived in a very different manner with a slightly different interpretation and the method is then used to approximate solutions to differential equations. In Sec. II the results of previous research[9,10] are generalized to the infinite interval by specifying an additional constraint on the vector norm and it is found that the new maximum entropy algorithm reduces to a minimum norm solution in an appropriate limit. In Sec. III the relationship between a Fourier series and MAXENT for differential equations is examined in light of the minimum norm solution. The method is also examined in the case of hybrid moments, that is, when the moment functions are any general acceptable function. Also, in Sec. III the newly developed technique is utilized to approximate solutions to stochastic differential equations occurring in statistical mechanics. In particular, equations for a random walk and the Fokker–Planck equations are studied. It appears that MAXENT is particularly well suited for studying such equations from a formal aspect.

[a] Present address: National Institute of Standards and Technology, Broadband Microwave Metrology Group, 723.02, Boulder, CO 80303.

II. THE MAXIMUM ENTROPY FORMALISM

We consider a general linear differential equation on the interval $[a, b]$ written as

$$L_1\langle V(r)\rangle = C(r), \tag{2.1}$$

where L_1 is a differential operator, $\langle V\rangle$ is the expectation value of a function, and C is a source term. We multiply Eq. (2.1) by an appropriate moment function f_n for $n = 1,2,3\ldots$ and then integrate over the solution region to obtain

$$\int_a^b f_n(L_1\langle V(r)\rangle - C(r))d^3r = 0. \tag{2.2}$$

The moment function can be an eigenfunction of the operator L_1 or any other set of functions which has the desired properties. Integration by parts and use of boundary conditions implies that

$$\int_a^b \alpha_n(r)\langle V(r)\rangle d^3r = A_n, \tag{2.3}$$

 0022-2488/89/071459-05$02.50

where A_n contains boundary information and information on $C(r)$. We see that in Eq. (2.3) all the derivatives have been transferred to the function α_n. Let us consider the function $V(r)$ as represented by a discrete set of N points $V(r_1)=V_1$, $V(r_2)=V_2,\ldots,V(r_N)=V_N$, so that in vector notation we can define the solution vector

$$\mathbf{V}=(V_1,V_2,V_3,\ldots,V_N)^t, \tag{2.4}$$

where t denotes transpose. We define the information entropy as

$$S=-\int_V P(\mathbf{V})\ln(P(\mathbf{V}))d\mathbf{V}, \tag{2.5}$$

where $\int d\mathbf{V}$ denotes $\int dV_1\int dV_2 dV_3\cdots dV_N$ and P is the probability density. We define the expectation value of a function V_j as

$$\langle V_j\rangle=\int P(\mathbf{V})V_j d\mathbf{V}. \tag{2.6}$$

Our constraint conditions on the allowable solutions are

$$B\langle\mathbf{V}\rangle=\mathbf{A}, \tag{2.7}$$

where we define the matrix $B=[\alpha_{ki}]$ and $\alpha_{ni}=\alpha_n(r_i)\Delta r$, where Δr is the grid size for a direct Riemann sum and A_k form a vector

$$\mathbf{A}=[A_1,\ldots,A_M]^t.$$

In component form we can write Eq. (2.7) as

$$\sum_{i=1}^{N}\alpha_{ni}\langle V_i\rangle=A_n. \tag{2.8}$$

We also have a constraint on the norm:

$$\langle\mathbf{V}\mathbf{V}^t\rangle=\|\mathbf{V}\|^2 \tag{2.9}$$

where

$$\|V\|^2=\sum_i V_i^2. \tag{2.10}$$

Finally we have the normalization condition

$$\int P(\mathbf{V})d\mathbf{V}=1. \tag{2.11}$$

The entropy with M constraints on the differential equation and the norm condition can be written as

$$\begin{aligned}S&=\int_{-\infty}^{\infty}\{-P(\mathbf{V})\ln(P(\mathbf{V}))-\lambda_0 P(\mathbf{V})\\&\quad-\lambda'B\mathbf{V}P(\mathbf{V})-\beta\mathbf{V}^t\mathbf{V}\}d\mathbf{V}\\&=\int_{-\infty}^{\infty}\Bigg\{-P(\mathbf{V})\ln(P(\mathbf{V}))-\lambda_0 P(\mathbf{V})\\&\quad-\sum_{n=1}^{M}\lambda_n\Bigg[\sum_{j=1}^{N}[\alpha_{nj}V_j P(\mathbf{V})]\Bigg]-\beta\sum_{i=1}^{N}V_i^2P(\mathbf{V})\Bigg\}d\mathbf{V},\end{aligned} \tag{2.12}$$

where λ_i and β are Lagrange multipliers, N is the number of points, and M is the number of moments. We note that the integrals run from $-\infty$ to ∞, which is possible for positive β (this will be shown later). In Eq. (2.12) the Lagrange multipliers are represented by the vector

$$\lambda=[\lambda_1,\ldots,\lambda_m]^t. \tag{2.13}$$

Performing a variation of P yields

$$\delta S=\int_{-\infty}^{\infty}\Bigg[-1-\ln(P(\mathbf{V}))-\lambda_0-\sum_j\Gamma_j V_j-\beta\sum_j V_j^2\Bigg]\delta P\,d\mathbf{V}=0,$$

where we have defined

$$\Gamma_j=\sum_{k=1}^{M}\alpha_{kj}\lambda_k=[\lambda'\mathbf{B}]. \tag{2.14}$$

We then obtain

$$\begin{aligned}P(\mathbf{V})&=\exp\{-(1+\lambda_0)\}\exp\{-[\lambda'B\mathbf{V}+\beta\mathbf{V}^t\mathbf{V}]\}\\&=\exp[-(1+\lambda_0)]\exp\Bigg\{-\sum_j[\Gamma_j V_j+\beta V_j^2]\Bigg\}.\end{aligned} \tag{2.15}$$

We define the partition function as

$$\begin{aligned}Z&=\exp\{1+\lambda_0\}\\&=\int_{-\infty}^{\infty}dV_1\cdots dV_N\exp\{-(\lambda'B\mathbf{V}+\beta\mathbf{V}^t\mathbf{V})\}\\&=\int_{-\infty}^{\infty}dV_1\cdots dV_N\exp\Bigg\{-\sum[\Gamma_j V_j+\beta V_j^2]\Bigg\}\\&=\prod_i\left(\frac{\pi}{\beta}\right)^{1/2}\exp\left\{-\frac{\Gamma_i^2}{4\beta}\right\}.\end{aligned} \tag{2.16}$$

We then obtain

$$\begin{aligned}P(\mathbf{V})&=\frac{\exp[-\{\lambda'B\mathbf{V}+\beta\mathbf{V}^t\mathbf{V}\}]}{Z}\\&=\frac{\exp[-\Sigma\{\Gamma_j V_j+\beta V_i^2\}]}{Z}.\end{aligned} \tag{2.17}$$

Now by Eq. (2.6) we have

$$\begin{aligned}\langle V_i\rangle&=\int_{-\infty}^{\infty}dV_1\cdots dV_N\,V_i\frac{\exp[-\Sigma\{\Gamma_j V_j+\beta V_i^2\}]}{Z}\\&=-\frac{\Gamma_i}{2\beta},\end{aligned} \tag{2.18}$$

$$\langle V_i^2\rangle=\frac{1}{2\beta}-\Gamma_i^2/4\beta^2, \tag{2.19}$$

$$\|V\|^2=\sum_i\langle V_i^2\rangle=\frac{N}{2\beta}-\sum_i\frac{\Gamma_i^2}{4\beta^2}. \tag{2.20}$$

The variance of the distribution can be found from

$$\sigma^2=\sum_{i=1}^{N}\frac{\langle V_i^2\rangle-\langle V_i\rangle^2}{N}=\frac{1}{2\beta} \tag{2.21}$$

and thus we see that β is positive. We note that the variance is independent of Γ_i, which prompts the following interpretation. If the constraint given by Eq. (2.21) is used, then the Lagrange multiplier β has the interpretation as one-half of $(\sigma^2)^{-1}$, whereas if Eq. (2.20) is used as the constraint, then the Lagrange multiplier can be determined:

$$\beta=\frac{N}{4\|V\|^2}+\sqrt{\frac{N^2}{16\|V\|^4}-\sum_i\frac{\Gamma_i^2}{4\|V\|^2}}\geqslant 0. \tag{2.22}$$

It is possible to solve for the Lagrange multipliers explicitly by use of the constraint condition in Eq. (2.7) using Eq. (2.18):

$$\lambda = -2\beta\,[BB^t]^{-1}\mathbf{A} \tag{2.23}$$

and therefore by Eq. (2.18):

$$\langle \mathbf{V} \rangle = B^t[BB^t]^{-1}\mathbf{A}. \tag{2.24}$$

As Poon[11] has noted, this solution for V is exactly the minimum norm solution[12] of the generalized inverse problem; therefore, the maximum entropy method on the infinite interval minimizes the norm in the least-squares sense subject to the available information. We also note that Eq. (2.24) is independent of the Lagrange multiplier β. The explanation of this comes from the fact that the expectation value of a function should not depend on the variance of the distribution. This solution is very easy to implement in practice and does not require the solution of simultaneous nonlinear equations for Lagrange multipliers. In Sec. III we will apply the technique to solving linear differential equations and the Fokker–Planck equation.

III. APPLICATIONS

A. Linear differential equations

As a simple example, we use $f_n = \sin n\pi z$ {for $z \in [0,1]$} as moment functions for Eq. (2.2) for the following differential equation:

$$\frac{d^2\langle V \rangle}{dz^2} = -z^3\langle V \rangle. \tag{3.1}$$

The boundary condition here is the specification of V on boundaries. The procedure is to multiply Eq. (3.1) by the moment functions and then integrate by parts to yield the form of Eq. (2.3) and thus identify the matrix B and vector A. In Fig. 1 the finite difference solution is ploted against the MAXENT solution for the cases of varying numbers of moments.

In the next example the following differential equation is approximated by MAXENT, a Fourier series, and finite differences in an attempt to probe the relationship of the MAXENT solution and Fourier series:

$$\frac{d^2\langle V \rangle}{dz^2} = -C\langle V \rangle, \tag{3.2}$$

where C is a constant. In this case, the function's value is specified on the boundary. The Fourier series solution for this case is simple to derive. For the case where the moment functions are $\sin n\pi z$, the solution using MAXENT is found to be exactly the same as the Fourier series approximation. This is depicted in Fig. 2. This result has a simple explanation. Since the coefficients in a Fourier series approximation, C_m, are picked to minimize the least-squares norm

$$I = \int \left| V(z) - \sum_{m=1}^{M} C_m f_m(z) \right|^2 dz, \tag{3.3}$$

a Fourier series representation is therefore a minimum norm approximation to the function. This is commonly termed a least-squares fit of the function. Therefore, the maximum entropy method on the infinite interval picks the Fourier coefficients as the "best fit."

However, the present MAXENT solution technique is much more robust than a Fourier series in that it allows the implementation of additional information into the process of solution of equations with noisy coefficients. For example, information such as knowledge of the solution over certain regions or boundary values could be fed into the B matrix. Thus we see that although the MAXENT solution reduces to a Fourier series solution in the limit of only Fourier moments, the method is versatile and much more general than a Fourier series. In previous work[9,10] the probability distribution was integrated over finite limits and it was determined that the MAXENT approximation was better than the Fourier series approximation in many cases. We now understand this since in the case of infinite limits the MAXENT picks out Fourier coefficients, whereas in the case of finite

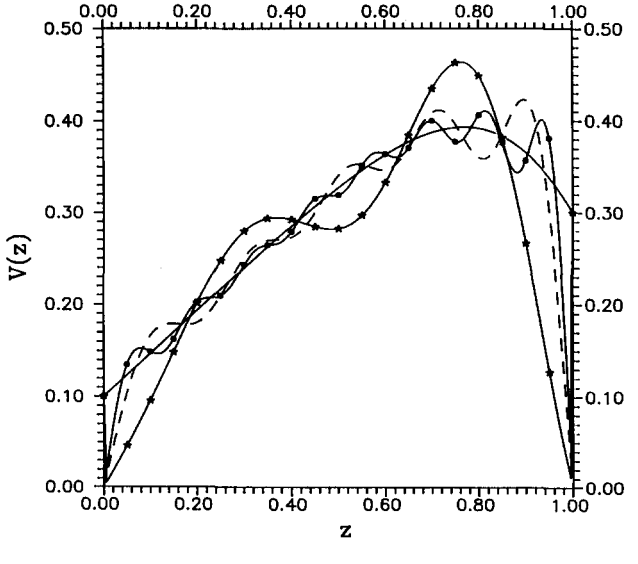


FIG. 1. The maximum entropy solution to Eq. (3.1) for six (- * - * -), 12 (- - -), and 18 moments (- ● - ● -) compared to the finite difference solution (———).

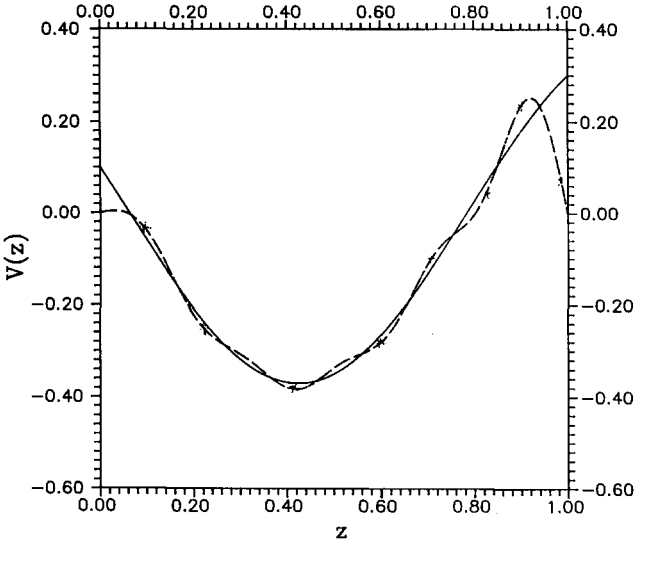


FIG. 2. The maximum entropy solution to Eq. (3.2) with eight Fourier moments (- - -) compared to the Fourier series solution with eight expansion functions ($\sin n\pi z$) (- * - * -) and the finite difference solution (———).

limits more information has been implemented into the problem and thus a better solution results.

As another example, it is of interest to examine the solution to differential equations when combinations of funtion types are used as moment functions (we term these as hybird moments). In Fig. 3 the solution to Eq. (3.2) is plotted for the case of the following moments: $f_n = \sin n\pi z$, $(n = 1,2,3)$, $f_4 = z(1-z)$, $f_5 = z^2(1-z)$, and $f_6 = z^3(1-z)$. The solution is compared to the Fourier series solution for six expansion functions and the exact solution of Eq. (3.2). In this case, this hybrid series expansion also minimizes the norm and actually approximates the solution better than the Fourier series.

B. Fokker–Planck equations

1. Example 1: Random walk

The Fokker–Planck equation is a differential equation for the probability density of a particle under the influence of external stochastic forces. The equation is usually derived by solution of master equations for probability density distributions. The random walk problem is a simple example of a Fokker–Planck equation:

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2}, \tag{3.4}$$

where D is a diffusion coefficient. If we multiply Eq. (3.4) alternately by x and x^2, integrate over $[-\infty, \infty]$, and use the fact that the probability vanishes at infinity, we obtain the following moments:

$$\frac{d\langle x\rangle}{dt} = 0, \tag{3.5}$$

$$\frac{d\langle x^2\rangle}{dt} = 2D. \tag{3.6}$$

If we assume a zero-mean process, then we obtain

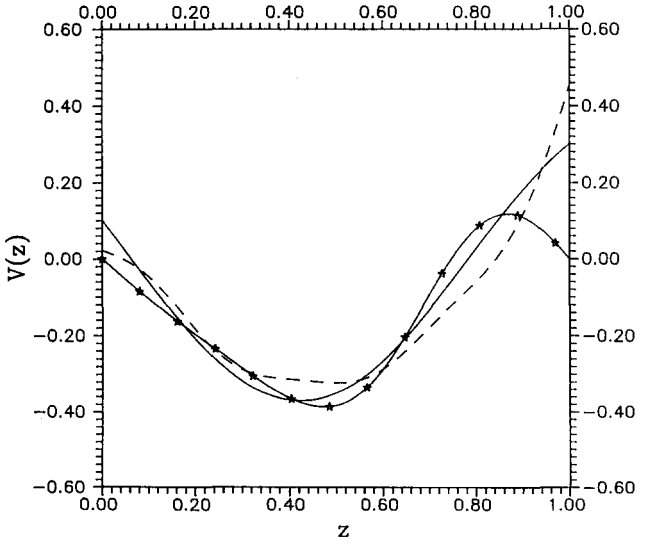


FIG. 3. The maximum entropy solution to Eq. (3.2) using hybrid moments compared to the finite difference solution (———) and the Fourier series solution (- * - * -) for six sin $n\pi z$ expansion functions.

$$\langle x\rangle = 0, \tag{3.7}$$

$$\langle x^2\rangle = 2Dt. \tag{3.8}$$

Now the probability density can be written as

$$P(x,t) = \exp(-\lambda x - \beta x^2)/Z. \tag{3.9}$$

However, we know $Z = (\pi/\beta)^{1/2}\exp(-\lambda^2/4\beta)$ and by Eq. (2.21) we obtain

$$\langle x^2\rangle - \langle x\rangle^2 = 2Dt = \frac{1}{2\beta} \Rightarrow \beta = \frac{1}{4Dt},$$

$$\langle x\rangle = \lambda/2\beta = 0 \Rightarrow \lambda = 0. \tag{3.10}$$

Therefore, we have the following solution for P:

$$P(x,t) = (4\pi Dt)^{-1/2}\exp(-x^2/4Dt), \tag{3.11}$$

which is exactly the solution obtained by classical Laplace transform techniques.

2. Example 2: General Fokker–Planck equation

In this case we consider the generalized Fokker–Planck equation

$$\frac{\partial P(x,t)}{\partial t} = \frac{-\partial\{\alpha_1(t)P(x,t)\}}{\partial x} + \left(\frac{1}{2}\right)\frac{\partial^2\{\alpha_2(t)P(x,t)\}}{\partial x^2}, \tag{3.12}$$

where $\alpha_n(t)$ is the nth-order jump moment. If we multiply Eq. (3.12) alternately by x and x^2, integrate over $[-\infty, \infty]$, and use the fact that the probability vanishes at infinity, we obtain the following moments:

$$\frac{d\langle x(t)\rangle}{dt} = \langle\alpha_1(t)\rangle, \tag{3.13}$$

$$\frac{d\langle x^2(t)\rangle}{dt} = 2\langle x(t)\alpha_1(t)\rangle + \langle\alpha_2(t)\rangle, \tag{3.14}$$

which we may integrate to yield

$$\langle x\rangle = \int_0^t \langle\alpha_1(\tau)\rangle d\tau = \frac{\lambda}{2\beta}, \tag{3.15}$$

$$\langle x^2\rangle = 2\int_0^t \langle x\alpha_1(\tau)\rangle d\tau + \int_0^t \langle\alpha_2(\tau)\rangle d\tau. \tag{3.16}$$

Here we assume that $\langle x\rangle$ and $\langle x^2\rangle$ are given functions of time. The variance is then

$$\langle x^2\rangle - \langle x\rangle^2 = \int_0^t \{2\langle x\alpha_1(\tau)\rangle + \langle\alpha_2(\tau)\rangle\}d\tau - \left[\int_0^t \langle\alpha_1(\tau)\rangle d\tau\right]^2 = \frac{1}{2\beta}. \tag{3.17}$$

Therefore, we can solve for the Lagrange multiplier:

$$\beta = \frac{1}{2[\langle x^2\rangle - \langle x\rangle^2]} \tag{3.18}$$

and from Eq. (3.15):

$$\lambda = 2\beta\int_0^t \langle\alpha_1(\tau)\rangle d\tau. \tag{3.19}$$

Thus the probability density is given by Eq. (3.9), where Z is given by Eq. (2.16).

IV. DISCUSSION

A general method for solving linear differential equations which provides a viable alternative to classical approaches has been developed. The method assumes that the solution is determined by the probability distribution for the solution vector, which is subject to moment constraints on the differential equation and norm. The associated Lagrange multiplier for the norm condition turns out to be related to the covariance, which is always positive; thus the probability distribution is integrable over $[-\infty, \infty]$. It is also found that with this method both the expectation value of the solution vector and the Lagrange multipliers can be obtained explicitly, thereby eliminating the solution of systems of nonlinear equations for the Lagrange multipliers. It is shown that for Fourier moments of the differential equation, the maximum entropy solution reduces precisely to the Fourier series solution. The method is also examined for the case of hybrid moments of differential equations and it is found that in these cases the solution obtained minimizes the norm and can be a very good approximation. Additional information of the differential equation can easily be inserted into the solution process, thus enhancing the accuracy and generality of the approximation. Additionally, solutions to the Fokker–Planck equation are obtained with the method. The method could be very useful when higher order moments of the Fokker–Planck equation are known. The method should be easily extended to solving, for example, Langevin equations and general master equations.

ACKNOWLEDGMENTS

We would like to thank Professor R. Inguva for many useful discussions.

Acknowledgment is made to the donors of the Petroleum Research Fund administered by ACS for partial support of this research, Grant No. ACS:PRF #20304-G2.

[1] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[2] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[3] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[4] J. P. Burg, Ph.D. thesis, Stanford University (1975).
[5] L.R. Mead, J. Math. Phys. **27**, 2903 (1986).
[6] L.R. Mead and N. Papanicolaou, J. Math. Phys. **25**, 2404 (1984).
[7] R. Inguva and J. Baker-Jarvis, in *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by James Justice (Cambridge, London, 1986), pp. 300–315.
[8] A. van den Bos, IEEE Trans. Inform. Theory **IT-17**, 493 (1971).
[9] J. Baker-Jarvis, J. Math. Phys. **30**, 302 (1989).
[10] J. Baker-Jarvis, J. Alameddine, and D. Schultz, *Numerical Methods for Partial Differential Equations*, Vol. 5, No. 2, pp. 133–142 (1989).
[11] C. Poon, Annual Report to National Science Foundation, Grant No. EET 8611970; also, to be published in IEEE Trans. Inform. Theory.
[12] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications* (Wiley, New York, 1974).

Grassmann-valued fluid dynamics

B. W. Fatyga and V. Alan Kostelecký
Physics Department, Indiana University, Bloomington, Indiana 47405

D. Rodney Truax
Department of Chemistry, University of Calgary, Calgary, Alberta, T2N 1N4 Canada

(Received 28 December 1988; accepted for publication 22 March 1989)

Certain systems of nonlinear partial differential equations can be written in a simple form as a single Grassmann-valued partial differential equation. Equations describing compressible fluid flow are of this type. A method for finding soft solutions of the Grassmann-valued partial differential equation arising in this context is presented. The method is a generalization of the Lagrangian-coordinates approach to the case of Grassmann variables. Generally, solutions obtained by this method have the form of infinite series, whose expansion yields new relations among the unknown variables. In some simple cases, the series can be summed. The equivalence of the Grassmann solutions to the usual solutions is shown for these cases.

I. INTRODUCTION

Grassmann-valued quantities are playing an increasingly significant role in modern theoretical physics. One reason is that they may be viewed as the classical limits of fermionic quantities. Another is that applications of supersymmetry, which relates anticommuting degrees of freedom to commuting ones, are now commonplace.[1] Grassmann-valued quantities are relevant to these applications, because the elements of a Grassmann algebra split naturally into an odd sector with mutually anticommuting variables and an even sector with mutually commuting ones.

Problems in physics are frequently expressed mathematically as differential equations, the solution of which yields desired physical quantities. When physical quantities involve anticommuting variables, the differential equations may have Grassmann-valued dependent variables. Grassmann-valued differential equations can also arise in the context of mathematical investigations of supersymmetry algebras and supergroups,[2,3] for example, in the derivation of Baker–Campbell–Hausdorff relations.[2]

A few attempts have been made[4,5] to investigate Grassmann-valued variables in contexts other than the supersymmetric or fermionic ones already mentioned. These have concentrated on exploratory studies of Grassmann-valued differential equations. Even this subfield is vast, with large areas remaining untouched. One goal of this paper is to present an application of Grassmann-valued differential equations that is independent of the notions of supersymmetry or fermions.

The basic idea is to consider a system of real- or complex-valued equations and to combine them into a single Grassmann-valued equation. One can then analyze directly this single equation using Grassmann methods, without referring to its expansion. In favorable cases, the complete solution to the Grassmann-valued equation may be found. Information about the original system of equations can subsequently be extracted. When the solution to the Grassmann-valued equation is known, the solution of the original system of equations may then be determined by expanding in the basis of Grassmann generators.

One evident advantage of this approach is that many equations may be handled simultaneously. The method does *not* gain significantly in complexity as the size of the original system is increased. It should therefore be of special interest for many-variable systems, for example. Another potentially significant consequence of the method is the possibility for the discovery of new solutions or approximation methods based on the Grassmann methods of solution.

A particularly interesting feature of this idea is that a single Grassmann equation may incorporate several physically inequivalent systems of equations, since we are free to define Grassmann-valued variables to suit the problem at hand. Solutions of these inequivalent systems may then all be expressed in terms of a single Grassmann-valued solution.

A suggestive analogy[6] is the use of complex numbers in the description of physical systems, which in effect allows the simultaneous handling of two real variables. Much of the importance of complex numbers stems from their properties as commutative division rings. Grassmann algebras do not enjoy similar properties because they contain nilpotent elements that act as divisors of zero. However, this difficulty may be partially overcome, as is shown in the examples below.

In this paper, we begin exploration of these ideas by testing them on several systems of nonlinear differential equations arising in fluid dynamics.[7] These systems of equations describe several compressible fluid flows under different conditions and in different dimensions. All can be expressed in terms of a single Grassmann-valued partial differential equation. We show that soft solutions to this single equation can be found by a method that is an extension to the case of Grassmann-valued variables of an established Lagrangian-coordinate approach. The resulting Grassmann-valued solution has the form of an infinite series. In some special cases, the series can be summed. For these cases, expansion of the solution yields solutions to the original systems that are equivalent to those found by standard methods.

In Sec. II, a simple example is presented to illustrate the basic methods and ideas. The general Grassmann-valued equation is presented in Sec. III, along with its solution. In

 0022-2488/89/071464-09$02.50

Sec. IV, the homogeneous version of the equation is considered and the solution found is compared to the usual case. Section V deals with several other special cases that can be solved exactly by standard methods. In Sec. V A, the case of three-dimensional flow under constant pressure in any coordinate system is discussed. In Sec. V B, the case of one-dimensional flows resulting from three-dimensional flows with planar, spherical, or cylindrical symmetry is studied. Section V C deals with the case of a polytropic gas law. We conclude in Sec. VI. For the convenience of the reader, the basics of Grassmann algebras and Grassmann-valued variables are summarized in Apendix A. Appendices B and C provide technical details of the derivations of some of the equations in the main text.

II. A SIMPLE EXAMPLE

We begin with a simple system illustrating the key ideas of our approach.

Consider a variable Z that is a function of two independent variables, position x and time t. Let Z satisfy the nonlinear partial differential equation

$$Z_t + ZZ_x = 0, \tag{2.1}$$

subject to the initial condition

$$Z(t=0,x) = F(x). \tag{2.2}$$

The subscripts t and x in Eq. (2.1) denote partial derivatives. The characteristic system for this equation is

$$\frac{dZ}{dt} = 0, \quad \frac{dx}{dt} = Z. \tag{2.3}$$

Solving this system gives

$$x = a + Zt, \quad Z = F(a) = F(x - Zt). \tag{2.4}$$

The variable a is called a Lagrange coordinate; it corresponds to a comoving frame. Solutions of the form (2.4) are called soft.[8]

Suppose now that Z is not the usual real or complex variable but a Grassmann-valued variable (see Appendix A) with expansion

$$Z = u + \rho\beta_1, \tag{2.5}$$

in terms of Grassmann generators, where u and ρ are functions of x and t. Then, Eq. (2.1) is equivalent to the system

$$u_t + uu_x = 0, \quad \rho_t + u\rho_x + u_x\rho = 0. \tag{2.6}$$

If we identify u with velocity and ρ with density, the system (2.6) describes the motion of a one-dimensional inviscid compressible fluid under constant pressure.

Take the initial conditions for u and ρ as

$$u(t=0,x) = f(x), \quad \rho(t=0,x) = g(x). \tag{2.7}$$

Then, the Grassmann-valued initial condition is

$$F(x) = f(x) + g(x)\beta_1. \tag{2.8}$$

The solution to Eq. (2.1) is still Eq. (2.4), but now the variables are Grassmann valued. Expanding the solution (2.4) gives

$$\begin{aligned} u + \rho\beta_1 = Z &= F(x - Zt) \\ &= f(x - ut - \rho t\beta_1) + g(x - ut - \rho t\beta_1)\beta_1. \end{aligned} \tag{2.9}$$

We can further expand f around the body $(x - ut)$ of its argument. Using $\beta_1^2 = 0$ yields

$$u + \rho\beta_1 = f(x - ut) - f'(x - ut)\rho t\beta_1 + g(x - ut)\beta_1. \tag{2.10}$$

Here, a prime denotes a derivative with respect to the argument. Separation of the body and soul parts provides the result

$$u = f(x - ut), \quad \rho = g(x - ut)/[1 + tf'(x - ut)]. \tag{2.11}$$

This is the soft solution[8] of the system (2.6).

Notice that in this simple example the partial differential equation for u and its solution are identical in form to the partial differential equation for Z and its solution. This occurs because u is the body of Z.

In contrast, ρ corresponds to the soul of Z; its partial differential equation and solution contain extra terms. The extra term $tf'(x - ut)$ in the denominator of the solution for ρ arises from the soul piece of the Grassmann-valued Lagrange coordinate a. Presumably, the latter must now be interpreted as a Grassmann-valued comoving frame.

Note also that no extra term arises from the expansion of $g(z - Zt)$ because g is already multiplied by β_1. The cancellations of terms proportional to β_1^2 are an important reason why Grassmann-valued variables are useful in nonlinear problems.

III. THE GENERAL EQUATION

Consider a first-order partial differential equation in D spatial variables $\mathbf{x}$ and one time variable t,

$$Z_t + (\mathbf{W}\cdot\nabla)Z = P. \tag{3.1}$$

Here, ∇ is the gradient operator in D dimensions with Euclidean coordinates, Z and P are Grassmann-valued functions, and $\mathbf{W}$ is a D-dimensional Grassmann-valued vector function. We shall consider this equation subject to the initial condition

$$Z(t=0,x) = F(x). \tag{3.2}$$

This Grassmann-valued equation is relevant to many compressible fluid flows under different conditions. In this section, its solution is presented. Subsequent sections discuss particular physical applications.

To avoid ordering problems, we take all Grassmann-valued variables to be even. This can be done for all cases of interest here.

The solution begins with the introduction of generalized Lagrange-coordinate variables $\mathbf{a}$ and s. The variables $\mathbf{x}$ and t then become functions of $\mathbf{a}$ and s; in particular,

$$\frac{\partial}{\partial s} = \frac{\partial t}{\partial s}\frac{\partial}{\partial t} + \sum_{i=1}^{D}\frac{\partial x^i}{\partial s}\frac{\partial}{\partial x^i}. \tag{3.3}$$

The key idea is to note that Eq. (3.1) has the form

$$\frac{\partial Z}{\partial s} = P, \tag{3.4}$$

if we set

$$\frac{\partial t}{\partial s} = 1, \quad \frac{\partial \mathbf{x}}{\partial s} = \mathbf{W}. \tag{3.5}$$

However, the last equation makes sense only if $\mathbf{x}$ is Grassmann valued as a function of $\mathbf{a}$ and s.

This problem may be circumvented by introducing a Grassmann-valued extension $\mathbf{X}$ of $\mathbf{x}$, to be determined below, that satisfies this equation. We further *define*

$$\frac{\partial}{\partial s} := \frac{\partial}{\partial t} + \sum_{i=1}^{D} \frac{\partial X^i}{\partial s} \frac{\partial}{\partial X^i}, \tag{3.6}$$

where the derivative $\partial/\partial R$ with respect to a Grassmann-valued variable R is an extension of differentiation *defined* by

$$\frac{\partial f(R)}{\partial R} = \left.\frac{\partial f(r)}{\partial r}\right|_{r \to R}, \tag{3.7}$$

for real r.

With these definitions, we obtain a characteristic system for Eq. (3.1):

$$\frac{\partial Z}{\partial s} = P, \quad \frac{\partial \mathbf{X}}{\partial s} = \mathbf{W}. \tag{3.8}$$

Note that Eqs. (3.5) do not completely specify the Lagrange coordinates $\mathbf{a}$ and s. In accordance with the usual Lagrangian description of fluid motion, we require in addition

$$s = t, \quad \mathbf{X}(s=0) = \mathbf{a}. \tag{3.9}$$

These equations complete the specification of the characteristic system (3.8).

It remains to determine $\mathbf{X}$ as a function of $\mathbf{a}$ and s. So far, we have

$$\mathbf{X}(s=0) = \mathbf{a}, \quad \frac{\partial \mathbf{X}}{\partial s} = \mathbf{W}. \tag{3.10}$$

Higher derivatives may be found from the definition (3.6), for example,

$$\begin{aligned} V^{(1)i} &:= \frac{\partial W^i}{\partial s} = W^i_t + (\mathbf{W}\cdot\nabla) W^i, \\ V^{(2)i} &:= \frac{\partial^2 W^i}{\partial s^2} = V^{(1)i}_t + (\mathbf{W}\cdot\nabla) V^{(1)i}. \end{aligned} \tag{3.11}$$

Then, X^i is given by[9]

$$X^i = a^i + W^i s - \tfrac{1}{2} V^{(1)i} s^2 + \tfrac{1}{6} V^{(2)i} s^3 - \cdots. \tag{3.12}$$

The solution for Z can be found similarly. We have

$$Z(s=0) = F(\mathbf{a}), \quad \frac{\partial Z}{\partial s} = P. \tag{3.13}$$

Higher derivatives can again be calculated, for example,

$$\begin{aligned} Q^{(1)} &:= \frac{\partial P}{\partial s} = P_t + (\mathbf{W}\cdot\nabla) P, \\ Q^{(2)} &:= \frac{\partial^2 P}{\partial s^2} = Q^{(1)}_t + (\mathbf{W}\cdot\nabla) Q^{(1)}. \end{aligned} \tag{3.14}$$

The final expression for Z has the form

$$Z = F(\mathbf{a}) + Ps - \tfrac{1}{2} Q^{(1)} s^2 + \tfrac{1}{6} Q^{(2)} s^3 - \cdots. \tag{3.15}$$

The solution of Eq. (3.1) can now be found by substitution into Eq. (3.15) of the expression for $\mathbf{a}$ obtained from Eq. (3.12). This solution is explored for particular choices of Z $\mathbf{W}$, and P in subsequent sections.

The general procedure is as follows. Choose the expansions in Grassmann generators of Z, $\mathbf{W}$, and P, such that Eq. (3.1) is equivalent to the system of real- or complex-valued partial differential equations that is of interest. The expansion of Eq. (3.15) then corresponds to the solution of this system.

Note that, in general, not only Z but also $\mathbf{W}$ and P contain the dependent variables. Therefore, the expansion of Eq. (3.15) in terms of Grassmann generators does not always lead to explicit solutions of the original system of equations. In the examples discussed below, explicit solutions can be found. However, even where this is not possible, the expansion of Eq. (3.15) gives new relations between the unknown variables, typically in the form of differential equations. These new relations often cannot be obtained from the initial system of equations by simple manipulations and may be useful in solving the initial system either by analytical or by approximate methods.

Note also that Eq. (3.15) contains two infinite series. These series can be summed in the examples we consider, although it is unlikely that they can be summed for arbitrary systems.

IV. THE HOMOGENEOUS CASE

In this section we give an example, including an explicit solution, of a system of partial differential equations that can be written as the homogeneous Grassmann-valued equation

$$Z_t + (\mathbf{W}\cdot\nabla) Z = 0. \tag{4.1}$$

In terms of analysis of Sec. III, this is the case $P = 0$. Therefore, only the first term on the right-hand side of the expression for Z, Eq. (3.15), remains different from zero:

$$Z = F(\mathbf{a}). \tag{4.2}$$

Our example concerns the problem of the motion of a D-dimensional compressible fluid under constant pressure. The fluid flow is described by the following partial differential equations arising from the conservation laws for momentum, mass, and energy:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = 0, \tag{4.3}$$

$$\frac{\partial \rho}{\partial t} + (\mathbf{u}\cdot\nabla)\rho + \rho\nabla\cdot\mathbf{u} = 0, \tag{4.4}$$

$$\frac{\partial e}{dt} + (\mathbf{u}\cdot\nabla)e + (e+p)\nabla\cdot\mathbf{u} = 0. \tag{4.5}$$

Here, $\mathbf{u}$ is velocity, ρ is the density, e is the internal energy per unit volume, and p is the pressure.

For convenience, we introduce Euclidean coordinates and define $\epsilon \equiv e + p$. Then, the system (4.3)–(4.5) is equivalent to $D+2$ scalar partial differential equations,

$$u^j_t + \sum_{k=1}^{D} u^k u^j_k = 0 \quad (j = 1,2,\ldots,D), \tag{4.6}$$

$$\rho_t + \sum_{k=1}^{D} (\rho u^k)_k = 0, \tag{4.7}$$

$$\epsilon_t + \sum_{k=1}^{D} (\epsilon u^k)_k = 0, \tag{4.8}$$

where the superscripts denote Euclidean components and the subscripts denote partial derivatives.

We wish to write Eqs. (4.6)–(4.8) in the form (4.1). To

avoid questions of the ordering of Grassmann-valued variables, we work with basis elements γ_j defined by

$$\gamma_j := \beta_{2j-1}\beta_{2j}, \tag{4.9}$$

where β_j are the generators of the Grassmann algebra. Then, Z and $\mathbf{W}$ may be taken to have the following expansions:

$$Z = u^1\gamma_1 + u^2\gamma_2 + \cdots + u^D\gamma_D + \rho\gamma_{12\cdots M} + e\gamma_{12\cdots M(M+1)}, \tag{4.10}$$

$$W_i = u^i + \rho\gamma_{1\cdots\hat{i}\cdots M} + \epsilon\gamma_{1\cdots\hat{i}\cdots M(M+1)}. \tag{4.11}$$

Here, $M \geqslant D$ except for the one-dimensional case where $M \geqslant 2$. The quantity $\gamma_{1\cdots\hat{i}\cdots M}$ denotes $\gamma_{123\cdots(i-1)(i+1)\cdots M}$, i.e., the subscript i is removed. Now, Eq. (4.1) with Z and $\mathbf{W}$ given by (4.10) and (4.11) is equivalent to the system (4.6)–(4.8).

We remark that there is some freedom in defining $\mathbf{W}$ because Z has no body and $\mathbf{W}$ appears in Eq. (4.1) only in the product $(\mathbf{W}\cdot\nabla)Z$. The choice of $\mathbf{W}$ made above is the simplest possible in the general case. Note also that for $D = 1$ the system (4.6) and (4.7) is the one discussed in Sec. II. However, the expansions (4.10) and (4.11) are somewhat different from the ones given in Eq. (2.5) of Sec. II. This illustrates the general fact that the representation of a system of real- or complex-valued partial differential equations in the form of a single Grassmann-valued equation is not unique, as a result of the anticommuting property of Grassmann-valued variables.

Next, consider the determination of an explicit solution for the case $D = 2$. Although the choice $M = D = 2$ is possible, computations are simpler if we take $M = 3$. Also, since Eq. (4.8) for ϵ has the same form as Eq. (4.7) for ρ, the solution for ϵ can be found from the solution for ρ by direct substitution. Therefore, the ϵ terms are omitted from the discussion below. They can be obtained from the ρ terms by substituting $\rho \to \epsilon\gamma_4$.

The Grassmann-valued equation is

$$Z_t + W^1 Z_1 + W^2 Z_2 = 0. \tag{4.12}$$

The expansions for Z and $\mathbf{W}$ are

$$\begin{aligned} Z &= u^1\gamma_1 + u^2\gamma_2 + \rho\gamma_{123}(+\epsilon\gamma_{123}), \\ W^1 &= u^1 + \rho\gamma_{23}(+\epsilon\gamma_{234}), \\ W^2 &= u^2 + \rho\gamma_{13}(+\epsilon\gamma_{134}). \end{aligned} \tag{4.13}$$

The initial conditions are

$$\begin{aligned} u^1(t=0,\mathbf{x}) &= f^1(\mathbf{x}), \\ u^2(t=0,\mathbf{x}) &= f^2(\mathbf{x}), \\ \rho(t=0,\mathbf{x}) &= h(\mathbf{x}). \end{aligned} \tag{4.14}$$

Therefore, the function F in Eq. (4.2) has the expansion

$$F(\mathbf{a}) = f^1(\mathbf{a})\gamma_1 + f^2(\mathbf{a})\gamma_2 + h(\mathbf{a})\gamma_{123}. \tag{4.15}$$

The expressions for the first and second s derivatives of $\mathbf{W}$, defined by Eqs. (3.11), are

$$\begin{aligned} V^{(1)1} &= W_t^1 + W^1 W_1^1 + W^2 W_2^1 = \rho[u_2^1\gamma_{13} - u_2^2\gamma_{23}], \\ V^{(1)2} &= \rho[u_1^2\gamma_{23} - u_1^1\gamma_{13}], \\ V^{(2)1} &= \rho[-2u_2^1(u_1^1 + u_2^2)\gamma_{13} \\ &\quad + (2u_2^2u_2^2 + u_1^1u_2^2 + u_2^1u_1^2)\gamma_{23}], \\ V^{(2)2} &= \rho[-2u_1^2(u_1^1 + u_2^2)\gamma_{23} \\ &\quad + (2u_1^1u_1^1 + u_2^2u_1^1 + u_1^2u_2^1)\gamma_{13}]. \end{aligned} \tag{4.16}$$

Higher s derivatives can also be calculated. Introducing the notation

$$\begin{aligned} S(\mathbf{V}^{(1)}) &\equiv -\frac{1}{2}\mathbf{V}^{(1)} + \frac{1}{3!}\frac{\partial}{\partial s}(\mathbf{V}^{(1)})t - \frac{1}{4!}\frac{\partial^2}{\partial s^2}(\mathbf{V}^{(1)})t^2 + \cdots \\ &= -\frac{1}{2}\mathbf{V}^{(1)} + \frac{1}{3!}\mathbf{V}^{(2)}t - \frac{1}{4!}\mathbf{V}^{(3)}t^2 + \cdots, \end{aligned} \tag{4.17}$$

we find

$$\mathbf{X} = \mathbf{a} + \mathbf{W}t + S(\mathbf{V}^{(1)})t^2. \tag{4.18}$$

According to Eq. (4.2), the solution of Eq. (4.12) is

$$Z(x^1,x^2,t) = f^1(\mathbf{A})\gamma_1 + f^2(\mathbf{A})\gamma_2 + h(\mathbf{A})\gamma_{123}. \tag{4.19}$$

The next step is to expand $\mathbf{A}$ around its body:

$$\begin{aligned} &A^1 = \mathbb{B}(A^1) + \Sigma(A^1), \quad A^2 = \mathbb{B}(A^2) + \Sigma(A^2), \\ &\mathbb{B}(A^1) = x^1 - u^1t, \quad \Sigma(A^1) = -\rho t\gamma_{23} - S(V^{(1)1})t^2, \\ &\mathbb{B}(A^2) = x^2 - u^2t, \quad \Sigma(A^2) = -\rho t\gamma_{13} - S(V^{(1)2})t^2. \end{aligned} \tag{4.20}$$

Noting that $[\Sigma(A^1)]^2 = [\Sigma(A^2)]^2 = 0$, we find

$$\begin{aligned} Z(\mathbf{x},t) &= f^1(\mathbf{A}^b)\gamma_1 + f^2(\mathbf{A}^b)\gamma_2 + f_{\mathrm{I}}^1(\mathbf{A}^b)A^{1s}\gamma_1 \\ &\quad + f_{\mathrm{II}}^1(\mathbf{A}^b)A^{2s}\gamma_1 + f_{\mathrm{I}}^2(\mathbf{A}^b)A^{1s}\gamma_2 \\ &\quad + f_{\mathrm{II}}^2(\mathbf{A}^b)A^{2s}\gamma_2 + h(\mathbf{A}^b)\gamma_{123}. \end{aligned} \tag{4.21}$$

Here, the subscripts I and II denote derivatives with respect to the first and second arguments, respectively.

Substituting the expressions for $\mathbb{B}(\mathbf{A})$ and $\Sigma(\mathbf{A})$ into these equations gives

$$\begin{aligned} Z(\mathbf{x},t) &= f^1(\mathbf{x}-\mathbf{u}t)\gamma_1 + f^2(\mathbf{x}-\mathbf{u}t)\gamma_2 + h(\mathbf{x}-\mathbf{u}t)\gamma_{123} \\ &\quad + \{-\rho(f_{\mathrm{I}}^1 + f_{\mathrm{II}}^2)t - [f_{\mathrm{I}}^1 S(V_{\gamma_{23}}^{(1)1}) \\ &\quad + f_{\mathrm{II}}^1 S(V_{\gamma_{23}}^{(1)2}) + f_{\mathrm{I}}^2 S(V_{\gamma_{13}}^{(1)1}) \\ &\quad + f_{\mathrm{II}}^2 S(V_{\gamma_{13}}^{(1)2})]t^2\}\gamma_{123}, \end{aligned} \tag{4.22}$$

where we have defined

$$\begin{aligned} S(V_{\gamma_{23}}^{(1)1})\gamma_{23} &\equiv -\frac{1}{2}(-\rho u_2^2\gamma_{23}) + \frac{1}{3!}\frac{\partial}{\partial s}(-\rho u_2^2\gamma_{23})t \\ &\quad - \frac{1}{4!}\frac{\partial^2}{\partial s^2}(-\rho u_2^2\gamma_{23})t^2 + \cdots. \end{aligned} \tag{4.23}$$

The first term in parentheses is the part of $V^{(1)1}$ proportional to γ_{23}. The quantity $S(V_{\gamma_{13}}^{(1)1})$ and others are defined similarly.

Comparing coefficients multiplying the same Grassmann generators on both sides of Eq. (4.22) gives

$$u^1 = f^1(x^1 - u^1t, x^2 - u^2t), \tag{4.24}$$

$$u^2 = f^2(x^1 - u^1t, x^2 - u^2t), \tag{4.25}$$

$$\begin{aligned} \rho &= h(x^1 - u^1t, x^2 - u^2t) + \rho(-f_{\mathrm{I}}^1t - f_{\mathrm{II}}^2t) \\ &\quad - [f_{\mathrm{I}}^1 S(V_{\gamma_{23}}^{(1)1}) + f_{\mathrm{II}}^1 S(V_{\gamma_{23}}^{(1)2}) + f_{\mathrm{I}}^2 S(V_{\gamma_{13}}^{(1)1}) \\ &\quad + f_{\mathrm{II}}^2 S(V_{\gamma_{13}}^{(1)2})]t^2. \end{aligned} \tag{4.26}$$

The expression in the brackets on the right-hand side of Eq. (4.26) contains a sum of four infinite series. This particular linear combination can be summed (see Appendix B). The result is

$$f_{\rm I}^1 S(V_{\gamma_{23}}^{(1)1}) + f_{\rm II}^1 S(V_{\gamma_{23}}^{(1)2}) + f_{\rm I}^2 S(V_{\gamma_{13}}^{(1)1}) + f_{\rm II}^2 S(V_{\gamma_{13}}^{(1)2}) = \rho(f_{\rm I}^1 f_{\rm II}^2 - f_{\rm II}^1 f_{\rm I}^2). \tag{4.27}$$

Using Eq. (4.27) and rearranging Eq. (4.26) we obtain the final expression for ρ:

$$\rho(x^1,x^2,t) = \frac{h(x^1 - u^1 t, x^2 - u^2 t)}{(1+f_{\rm I}^1 t)(1+f_{\rm II}^2 t) - f_{\rm II}^1 f_{\rm I}^2 t^2}. \tag{4.28}$$

The solution to the original system of equations is thus given by Eqs. (4.24), (4.25), and (4.28). It agrees with the solution found by standard methods in Ref. 8. Just as for the simple example described in Sec. II, the denominator in Eq. (4.28) arises from the expansion of Grassmann-valued functions around the body of their arguments.

V. THE INHOMOGENEOUS CASE

In this section, we consider several distinct physical situations for which the relevant Grassmann-valued differential equation is inhomogeneous. Inhomogeneous terms can arise in different ways. Here, we consider inhomogeneities arising from the choice of non-Euclidean coordinates and ones arising directly from the addition of terms to the original system of equations.

A. Three-dimensional compressible fluid flow under constant pressure in curvilinear coordinates

In this subsection, we formulate the Grassmann-valued differential equation suitable for the description of three-dimensional compressible fluid flow under constant pressure in curvilinear coordinates. The formulation has a straightforward generalization to $D>3$.

Consider an orthogonal curvilinear system of coordinates y_i with unit vectors $\mathbf{e}_i$ tangent to the coordinate curves and with curvilinear metric

$$ds^2 = (h_1)^2\,dy_1^2 + (h_2)^2\,dy_2^2 + (h_3)^2\,dy_3^2. \tag{5.1}$$

In this system of coordinates the expressions for the gradient and divergence are

$$\nabla f = \frac{\mathbf{e}_1}{h_1}\frac{\partial f}{\partial y_1} + \frac{\mathbf{e}_2}{h_2}\frac{\partial f}{\partial y_2} + \frac{\mathbf{e}_3}{h_3}\frac{\partial f}{\partial y_3},$$
$$\nabla\cdot\mathbf{F} = \frac{1}{\mathscr{J}}\left[\frac{\partial}{\partial y_1}(h_2h_3F_1) + \frac{\partial}{\partial y_2}(h_1h_3F_2) + \frac{\partial}{\partial y_3}(h_1h_2F_3)\right], \tag{5.2}$$

where $\mathscr{J} = h_1h_2h_3$.

The vector equations (4.3) and (4.4) that describe the fluid flow are valid in any system of coordinates. They can be written in a single inhomogeneous Grassmann-valued equation as

$$Z_t + (\mathbf{W}\cdot\nabla)Z = P, \tag{5.3}$$

where

$$Z = \sum_{i=1}^{3} u_i\gamma_i + \rho\gamma_{1234}, \quad W_i = u_i + \rho\gamma_{1\ldots\hat{i}\ldots4}, \tag{5.4}$$

with u_i representing velocity components in the curvilinear system. In the metric (5.1), the scalar product $\mathbf{W}\cdot\nabla$ is

$$\mathbf{W}\cdot\nabla = \sum_{i=1}^{3} W_i\frac{1}{h_i}\frac{\partial}{\partial y_i}, \tag{5.5}$$

and the inhomogeneity P is

$$P = -\frac{\rho}{\mathscr{J}}\left[u_1\frac{\partial}{\partial y_1}(h_2h_3) + u_2\frac{\partial}{\partial y_2}(h_1h_3) + u_3\frac{\partial}{\partial y_3}(h_1h_2)\right]\gamma_{1234}. \tag{5.6}$$

The characteristic system, Eqs. (3.8), must be modified to account for the change in the scalar product (5.5). The new system is

$$\frac{\partial Z}{\partial s} = P, \quad \frac{\partial X_i}{\partial s} = \frac{W_i}{h_i}, \tag{5.7}$$

where $i=1,2,3$ is not summed and where the s derivative is now defined by

$$\frac{\partial}{\partial s} := \frac{\partial}{\partial t} + \sum_{i=1}^{3} W_i\frac{1}{h_i}\frac{\partial}{\partial y_i}. \tag{5.8}$$

For Cartesian coordinates in $\mathbb{R}^3$, we have $h_1 = h_2 = h_3 = 1$, and so

$$P = 0. \tag{5.9}$$

For cylindrical coordinates $y_1 = r, y_2 = \phi, y_3 = z$ and $h_1 = 1, h_2 = r, h_3 = 1$, which gives

$$P = -(\rho/r)u_1\gamma_{1234}. \tag{5.10}$$

For spherical coordinates $y_1 = r, y_2 = \theta, y_3 = \phi$ and $h_1 = 1, h_2 = r, h_3 = r\sin\theta$, which gives

$$P = -\rho[u_1(2/r) + u_2(\cot\theta/r)]\gamma_{1234}. \tag{5.11}$$

B. One-dimensional flows arising from three-dimensional flows with planar, cylindrical, or spherical symmetry

In this subsection, we consider restrictions to one-dimensional flows arising from symmetry of the three-dimensional fluid flows described in Sec. V A. The idea is to allow only the radial component of the velocity to be nonzero. The expansion in terms of Grassmann generators of the solution of the inhomogeneous Grassmann-valued equation can in these cases be written in closed form.

For these cases, Eqs. (5.9)–(5.11) may be written as

$$P = -\nu(\rho u/x)\gamma_{1234}, \tag{5.12}$$

where we write u for u_1 and x for r. The coefficient ν is $\nu = 0$ for plane symmetry, $\nu = 1$ for cylindrical symmetry, and $\nu = 2$ for spherical symmetry.

Since the situation is effectively one dimensional, instead of working with the full equation $Z_t + (\mathbf{W}\cdot\nabla)Z = P$ we can use the simpler equation

$$Z_t + ZZ_x = P. \tag{5.13}$$

The expansions of Z and P are taken as

$$Z = u + \rho\gamma_1 \tag{5.14}$$

and

$$P = -\nu(\rho u/x)\gamma_1. \tag{5.15}$$

Instead of the characteristic system (3.8), we find

$$\frac{\partial Z}{\partial s} = P, \quad \frac{\partial X}{\partial s} = Z. \tag{5.16}$$

This simplifies the general solution described in Sec. III because the quantities $V^{(i)}$ and $Q^{(j)}$ defined in Eqs. (3.11) and (3.14) are now related by

$$Q^{(i)} = V^{(i+1)}, \quad V^{(1)} = P. \tag{5.17}$$

The expressions for X and Z, in terms of $V^{(i)}$ are

$$X = a + Zs - \tfrac{1}{2}V^{(1)}s^2 + \tfrac{1}{6}V^{(2)}s^3 - \cdots,$$
$$Z = f(a) + h(a)\gamma_1 + V^{(1)}s - \tfrac{1}{2}V^{(2)}s^2 + \tfrac{1}{6}V^{(3)}s^3 - \cdots. \tag{5.18}$$

Expanding both sides of the latter equation in terms of the Grassmann generators and expanding functions of Grassmann-valued variables around the body yields

$$u = f(x - ut),$$
$$\rho = h(x - ut) + f'\left\{ -\rho t + \frac{1}{2}V^{(1)}t^2 - \frac{1}{6}V^{(2)}t^3 + \frac{1}{4!}V^{(3)}t^4 + \cdots \right\} + V^{(1)}t - \frac{1}{2}V^{(2)}t^2 + \frac{1}{6}V^{(3)}t^3 - \frac{1}{4!}V^{(4)}t^4. \tag{5.19}$$

Here, as before, f' denotes the derivative with respect to the argument. In Appendix C, we show that the equation for ρ can be written as

$$\rho = h(x - ut)(1 + f't)^{-1}[1 - (f/x)t]^{\nu}. \tag{5.20}$$

This agrees with the soft solution we obtained by standard methods.

C. One-dimensional compressible fluid flow with polytropic gas law

In this subsection, we consider a generalization of the simple example given in Sec. II to the case where the pressure depends on the fluid density via the polytropic gas law[10]

$$p = c\rho^{\gamma}, \tag{5.21}$$

where c and γ are constants. We remark that for $\gamma = 2$, the polytropic gas equations have the same form as the equations describing wave motion in shallow fluids,[7] if u is interpreted as the horizontal wave velocity and ρ is replaced by the variable η measuring the fluid depth. This case is thus also described by the analysis below.

The fluid flow for a polytropic gas can be described by the inhomogeneous Grassmann-valued equation

$$Z_t + ZZ_x = P, \tag{5.22}$$

where Z has the same expansion as in Eq. (2.5) of Sec. II and

$$P = -c\gamma\rho^{\gamma-2}\rho_x. \tag{5.23}$$

Note that for this situation the inhomogeneity P is pure body. In our approach it is more convenient to deal with an inhomogeneity that is pure soul. Hence instead of solving Eq. (5.22) directly we tackle an equivalent problem for which the variable Z has been redefined so that the inhomogeneity is pure soul. The method is applicable for $\gamma \neq 1$.

To implement this approach, define a new variable M by imposing

$$P = -MM_x. \tag{5.24}$$

Then, add terms dependent on M to both sides of Eq. (5.22) to yield the expression

$$(Z + pM)_t + (Z + rM)(Z + pM)_x, \tag{5.25}$$

on the left-hand side and

$$Q := pM_t + pZM_x + rMZ_x + (pr - 1)MM_x \tag{5.26}$$

on the right-hand side, where p and r are arbitrary constants.

Explicitly, for the polytropic gas, we find

$$M = \alpha\rho^m, \tag{5.27}$$

where

$$m = (\gamma - 1)/2, \quad \alpha = \sqrt{2c[\gamma/(\gamma - 1)]}, \quad \gamma \neq 1. \tag{5.28}$$

Defining a new variable $\overline{Z}$ by

$$\overline{Z} = Z + pM, \tag{5.29}$$

we obtain the modified equation

$$\overline{Z}_t + (Z + rM)\overline{Z}_x = Q. \tag{5.30}$$

The idea is to choose the coefficients p and r so that Q has no body. Substituting the expansions of Z and M in terms of Grassmann generators into Eq. (5.26) gives

$$\frac{\partial \overline{Z}}{\partial s} \equiv Q = \alpha\rho^m u_x(-pm + r) + (pr - 1)\alpha^2 m\rho^{2m-1}\rho_x + \alpha\rho^m\rho_x(pm + r)\gamma_1. \tag{5.31}$$

The inhomogeneity Q has no body if $r = mp$ and $pr = 1$.

For the special case of an adiabatic gas, $\gamma = 3$, the soft solution has been explicitly found by other methods.[10] We demonstrate a method of deriving it using Grassmann analysis. For this case, $m = 1$; hence Q is pure soul, $\mathbb{B}(Q) = 0$, if either $p = r = 1$ or $p = r = -1$. Choosing the plus sign yields

$$\overline{Z} = u + \sqrt{3c}\rho + \rho\gamma_1. \tag{5.32}$$

Since $\mathbb{B}(\partial\overline{Z}/\partial s) = 0$, $\mathbb{B}(\overline{Z})$ for any s must be equal to $\mathbb{B}(\overline{Z})$ at $s = 0$. The latter is given by the initial conditions.

Next, consider the variable X. From Eq. (5.30) it follows that

$$\frac{\partial X}{\partial s} = Z + rM. \tag{5.33}$$

Since $r = p$, $Z + rM = \overline{Z}$. Therefore,

$$\mathbb{B}\left(\frac{\partial^2 X}{\partial s^2} = \frac{\partial \overline{Z}}{\partial s}\right) = 0. \tag{5.34}$$

This shows that

$$\mathbb{B}(X) = \mathbb{B}(a + (u + \sqrt{3c}\rho)t). \tag{5.35}$$

The body part of the equation for $\overline{Z}$ then gives

$$u + \sqrt{3c}\rho = f[x - (u + \sqrt{3c}\rho)t] + \sqrt{3c}\, g[x - (u + \sqrt{3c}\rho)t]. \tag{5.36}$$

The equation for ρ, which arises from the γ_1 part of the equation for $\overline{Z}$, is still quite cumbersome as it involves an infinite series. Rather than finding ρ from this equation, it is simpler to consider the second possible choice for p and r, namely, $p = r = -1$. This leads to another equation involv-

ing u and ρ. The calculation follows the previous one, except that now

$$\overline{Z} = u - \sqrt{3c}\,\rho + \rho\gamma_1. \tag{5.37}$$

The final equation is

$$u - \sqrt{3c}\,\rho = f[x - (u - \sqrt{3c}\,\rho)t] - \sqrt{3c}\,g[x - (u - \sqrt{3c}\,\rho)t]. \tag{5.38}$$

Equations (5.36) and (5.38) form a solution of the adiabatic gas problem. They have the same form as the solution obtained from the standard approach[10] if we make the change of variables

$$(u, \rho) \to (u + \sqrt{3c}\,\rho,\ u - \sqrt{3c}\,\rho). \tag{5.39}$$

In the context of the Grassmann analysis, this change of variables arises from the requirement that the inhomogeneity have the simplest possible form.

In the cases for which $\gamma \neq 3$, values of p and r may also be found such that $\mathbb{B}(\partial\overline{Z}/\partial s) = 0$. However,

$$\mathbb{B}\left(\frac{\partial^2 X}{\partial s^2}\right) \neq 0, \tag{5.40}$$

and so $\mathbb{B}(X)$ is, in general, represented by an infinite series. The expansion in terms of Grassmann generators then yields new relations between the unknown variables.

As an example, consider the case of the Chaplygin gas,[10] $\gamma = -1$. We take $p = \pm i$ and $r = \mp i$. Then, the body and soul parts of the equation for $\overline{Z}$ yield the equations

$$\begin{aligned} &u \pm i\sqrt{c}\rho^{-1} = F_{\pm}, \\ &\rho = h + C_{\pm}[f' \mp i\sqrt{c}h^{-2}h'] + D_{\pm}, \end{aligned} \tag{5.41}$$

where $F_{\pm} = +f \pm i\sqrt{c}h^{-1}$ and where the functions $F_{\pm}, f', h$, and h' have arguments $(x - B_{\pm})$. The capital letters B,C,D represent the infinite series

$$\begin{aligned} B_{\pm} &= (u \pm i\sqrt{c}\rho^{-1})t + (c\rho^{-3}\rho_x \mp i\sqrt{c}\rho^{-1}u_x)t^2 \\ &\quad + \tfrac{2}{3}c(\rho^{-4}u_x\rho_x \pm i\sqrt{c}\rho^{-1}(\rho^{-3}\rho_x)_x)t^3 + \cdots, \\ C_{\pm} &= -\rho t(c\rho^{-3}(\rho_x)^2 - \tfrac{1}{3}c\rho^{-2}\rho_{xx} \pm cu_{xx} \\ &\quad \mp i\sqrt{c}\rho^{-1}u_x\rho_x)t^3 + \cdots, \\ D_{\pm} &= \mp 2i\sqrt{c}\rho^{-1}\rho_x t \pm i\sqrt{c}[\rho^{-1}u_x\rho_x + u_{xx} \\ &\quad + i\sqrt{c}\rho^{-2}(\rho_{xx} - \rho^{-1}(\rho_x)^2]t^2 + \cdots. \end{aligned} \tag{5.42}$$

If the variables u and ρ are expanded in t then, as expected, these equations yield coefficients that agree at each order in t with the result of expanding the original system in t. However, information about this system is not encoded in a straightforward manner in Eqs. (5.40) and (5.41). For example, derivatives of the initial conditions appear.

VI. DISCUSSION AND CONCLUSIONS

In this paper, we presented a method for investigating systems of equations based on Grassmann-valued variables. It is an application of Grassmann-valued analysis that is independent of supersymmetry or fermions. Features of the method include its ability to handle simultaneously in a single Grassmann-valued equation many real- or complex-valued equations together with many physically different situations.

The idea was applied to systems of partial differential equations arising in the context of fluid dynamics. Various compressible fluid flows are described by a single Grassmann-valued equation, whose solution can be obtained using Grassmann-valued methods. Expansion of the solution in terms of a Grassmann basis yields solutions of the original systems of equations. We showed that these solutions are equivalent to ones obtained by standard approaches for simple cases. In general, the solution has the form of infinite series whose expansion yields new relations between the unknown variables. This investigation serves as a useful test of the potential role of Grassmann-valued variables in dealing with systems of nonlinear equations.

Several interesting areas remain for future research. Here, we have not addressed the question of the physical meaning of the use of Grassmann-valued variables. Also, many avenues for the direct application of the ideas to other systems of equations remain to be explored. Indeed, the Grassmann-valued differential equation we have studied may be relevant for physical problems other than fluid dynamics, for example, those involving the mechanics of continuous media. Another interesting possibility is the determination of new analytical or approximate solutions for involved systems of equations, including nonperturbative methods.

ACKNOWLEDGMENTS

This research was supported in part by the United States Department of Energy under Contract No. DE-AC02-84ER40125 and by the National Sciences and Engineering Research Council of Canada.

APPENDIX A: GRASSMANN ALGEBRAS AND GRASSMANN-VALUED QUANTITIES

In this Appendix, we provide a summary of key features of Grassmann algebras and Grassmann-valued quantities.

A real Grassmann algebra B_L over $\mathbb{R}^L$ is defined as an associative algebra that contains all vectors in $\mathbb{R}^L$ and that may be generated from them and from scalar multipliers with a product operation such that each pair of vectors $\beta_1, \beta_2 \in \mathbb{R}^L$ satisfies $\beta_1\beta_2 = -\beta_2\beta_1$.

A basis for B_L therefore consists of the identity $\beta_\omega \equiv I$, a set of L vectors β_j, $j = 1,...,L$, and all nonvanishing products of these vectors, denoted $\beta_{jk},...,\beta_{jk\cdots p},...,\beta_{12\ldots L}$, $j < k < \cdots < p$. There are 2^L basis elements for B_L, which we denote[2] collectively by $\{\beta_\mu\}$. The subset of vectors in B_L generated by I and by even products of β_j spans what we call the even part ${}^0B_L$ of B_L, while the subset generated by β_j and by odd products of β_j spans the odd part ${}^1B_L$ of B_L.

If a quantity A is B_L valued, then it may be expanded in terms of the basis $\{\beta_\mu\}$ as $A = \Sigma_\mu A_\mu\beta_\mu$, where $A_\mu \in \mathbb{R}$ are the components of A. It is convenient to define the body $\mathbb{B}(A)$ of A as the component A_I, and the soul $\Sigma(A)$ of A as $A - \mathbb{B}(A)I$. These projections are the analogs for Grassmann-valued variables of the real and imaginary parts of complex-valued variables.

As a simple example, consider the four-dimensional Grassmann algebra B_2 over $\mathbb{R}^2$. The basis for this algebra is the set $\{I,\beta_1,\beta_2,\beta_{12}=\beta_1\beta_2\}$. A Grassmann-valued variable $A\in B_2$ has expansion $A=aI+b\beta_1+c\beta_2+d\beta_{12}$, with $a,b,c,d\in\mathbb{R}$. Then, $\mathbb{B}(A)=a$ and $\Sigma(A)=b\beta_1+c\beta_2+d\beta_{12}$. Note that A is the sum of an even variable ${}^0A=aI+d\beta_{12}$ and an odd variable ${}^1A=b\beta_1+c\beta_2$. In the text, we omit explicitly writing the basis element I for convenience.

APPENDIX B: DERIVATION OF EQ. (4.27)

In this appendix, we present some details of the calculations involved in Eq. (4.27).

Introduce the notation

$$p=u_1^1+u_2^2,\quad q=u_1^1u_2^2-u_2^1u_1^2,\quad T=1-pt-qt^2,$$
$$D_k=u_1^1(\partial^k u_2^2)+u_2^2(\partial^k u_1^1)-u_2^1(\partial^k u_1^2)-u_1^2(\partial^k u_2^1). \tag{B1}$$

The symbol ∂^k is defined by $\partial^k(f)\beta_{23}=(\partial^k/\partial s^k)(f\beta_{23})$ or equivalently by $\partial^k(f)\beta_{13}=(\partial^k/\partial s^t)(f\beta_{13})$, with $\partial/\partial s$ determined as in Eq. (3.6) and with f representing $\rho,p,u_1^1,u_2^1,u_1^2,u_2^2$. We shall prove that

$$LT=RT, \tag{B2}$$

where L and R denote the left- and right-hand sides of Eq. (4.27), respectively.

Expand LT and RT in power series in t as

$$LT=\sum_{n=0}^{\infty}L(n)t^n,\quad RT=\sum_{n=0}^{\infty}R(n)t^n. \tag{B3}$$

Direct inspection shows that $L(n)=R(n)$ for $n=0,1,2$. For $n\geqslant 3$ we have

$$L(n)=\frac{1}{(n+2)!}(-1)^n\sum_{k=0}^{n}\binom{n}{k}(\partial^k\rho)D_{n-k}$$
$$+\frac{q}{(n+1)!}(-1)^n$$
$$\times\sum_{k=0}^{n-1}\binom{n-1}{k}(\partial^k\rho)(\partial^{n-k-1}p) \tag{B4}$$

and

$$R(n)=\rho q\Bigg\{\sum_{l=[\frac{1}{2}(n+1)]}^{n}\binom{l}{2l-n}p^{2l-n}(-q)^{n-1}$$
$$-\sum_{l=[\frac{1}{2}(n+1)]-1}^{n-1}\binom{l}{2l-n+1}p^{2l-n+2}(-q)^{n-l-1}$$
$$-\sum_{l=[\frac{1}{2}(n+1)]-2}^{n-2}\binom{l}{2l-n+2}p^{2l-n+2}(-q)^{n-l-1}\Bigg\}. \tag{B5}$$

In these expressions, the brackets in the summation limits indicate the integer part of the quantity enclosed.

The simplest expression to analyze is $R(n)$. Grouping terms containing the same powers of p and q, we find that $p^{2l-n}(-q)^{n-1}$ is multiplied by a coefficient

$$\binom{l}{2l-n}-\binom{l-1}{2l-n-1}-\binom{l-1}{2l-n}=0. \tag{B6}$$

Therefore, $R(n)=0$ for $n\geqslant 3$.

It thus remains to show that $L(n)=0$ for $n\geqslant 3$. For this we need expressions for the derivatives of p and q and for D_n, which are found to be

$$\partial^n\rho=(-1)^n n!\rho\sum_{k=0}^{[\frac{1}{2}n]}(-1)^k\binom{n-k}{k}p^{n-2k}q^k$$
$$\partial^n p=(-1)^n n!\sum_{k=0}^{[\frac{1}{2}(n+1)]}(-1)^k\left[\binom{n-k+1}{k}\right.$$
$$\left.+\binom{n-k}{k-1}\right]p^{n-2k+1}q^k,$$
$$D_n=nq\partial^{n-1}p. \tag{B7}$$

These expressions can be proved by induction. They may be used to cast Eq. (B4) in the form

$$L(n)=\frac{2n!}{(n+2)!}\rho qp^n\sum_{m=0}^{[\frac{1}{2}n]}(-1)^m p^{-2m}q^m$$
$$\times\Bigg\{\binom{n-m}{m}-\frac{1}{n}\sum_{k=0}^{n-1}\sum_{l=0}^{[\frac{1}{2}k]}\binom{k-l}{l}$$
$$\times\left[\binom{n-k-m+l}{m-l}\right.$$
$$\left.+\binom{n-k-m+l-1}{m-l-1}\right]\Bigg\}. \tag{B8}$$

The double sum in the braces can be rewritten as the sum of the following four terms:

$$\sum_{p=0}^{[\frac{1}{2}n]-1}\sum_{l=0}^{p}\binom{p}{l}\binom{n-m-p}{m-l}=\frac{1}{2}n\binom{n-m}{m},$$
$$\sum_{p=0}^{[\frac{1}{2}n]-1}\sum_{l=0}^{p}\binom{p}{l}\binom{n-m-p-1}{m-l-1}=\frac{1}{2}n\binom{n-m-1}{m-1},$$
$$\sum_{p=[\frac{1}{2}n]}^{n-1}\sum_{l=0}^{n-p-1}\binom{p}{l}\binom{n-m-p}{m-l}=\left(\frac{1}{2}n-m\right)\binom{n-m}{m},$$
$$\sum_{p=[\frac{1}{2}n]}^{n-1}\sum_{l=0}^{n-p-1}\binom{p}{l}\binom{n-m-p-1}{m-l-1}=\left(\frac{1}{2}n-m\right)\binom{n-m-1}{m-1}. \tag{B9}$$

Carrying out the sum of the four terms on the left-hand side of the above equations, we find it is equal to $n\binom{n-1}{m}$. Therefore, $L(n)=0$ as required.

APPENDIX C: DERIVATION OF EQ. (5.20)

In this appendix, we reduce Eq. (5.19) to the form of the standard soft solution.

Equation (5.19) can be written as the product

$$\rho = h(x-ut)D, \tag{C1}$$

where D is given by the infinite series

$$D = 1 + f't - v^{(1)}t - v^{(1)}t\left(1+\frac{1}{2}f't\right) + \frac{1}{2!}v^{(2)}t^2\left(1+\frac{1}{3}f't\right) + \cdots + (-1)^n\frac{1}{n!}v^{(n)}t^n\left(1+\frac{1}{n+1}f't\right) + \cdots, \tag{C2}$$

with

$$v^{(i)} = V^{(i)}/\rho. \tag{C3}$$

The goal is to sum the series for D.

By induction, we can show that

$$v^{(n)} = (-1)^n\nu\gamma\sum_{k=0}^{n-1}(n-1)!\binom{\nu+k}{k}\alpha^{n-k-1}\beta^k, \tag{C4}$$

where

$$\alpha = f'/(1+f't), \quad \beta = f/x. \tag{C5}$$

With the additional definition $v^{(0)}\equiv 1$, D can be written as

$$D = \sum_{n=0}^{\infty}(-1)^n\frac{1}{n!}v^{(n)}t^n\left[(1+f't)+f't\left(\frac{1}{n+1}-1\right)\right]. \tag{C6}$$

Writing the $n=0$ term explicitly and rearranging the remaining sum yields

$$D = 1 + f't + \sum_{n=1}^{\infty}(-1)^n\frac{1}{n!}(1+f't)t^n[v^{(n)} + (n-1)v^{(n-1)}\alpha]. \tag{C7}$$

However,

$$\begin{aligned} v^{(n)} &+ (n-1)\alpha v^{(n-1)} \\ &= (-1)^n\nu\beta\left[\sum_{k=0}^{n-1}(n-1)!\binom{\nu+k}{k}\alpha^{n-k-1}\beta^k - \sum_{k=0}^{n-2}(n-1)!\binom{\nu+k}{k}\alpha^{n-k-1}\beta^k\right] \\ &= (-1)^n\nu\beta(n-1)!\binom{\nu+n-1}{n-1}\alpha^0\beta^{n-1}. \end{aligned} \tag{C8}$$

Therefore, the expression for D becomes

$$D = 1 + f't + \sum_{n=1}^{\infty}\frac{\nu}{n}\binom{\nu+n-1}{n-1}(1+f't)\beta^n t^n. \tag{C9}$$

Since

$$\frac{\nu}{n}\binom{\nu+n-1}{n-1} = \binom{\nu+n-1}{n}, \tag{C10}$$

we finally obtain

$$\begin{aligned} D &= (1+f't)\sum_{n=0}^{\infty}\binom{\nu+n-1}{n}\beta^n t^n \\ &= (1+f't)(1-\beta t)^{-\nu}, \end{aligned} \tag{C11}$$

which is the desired result.

[1]For a range of applications of supersymmetry see, for example, *Supersymmetry in Physics*, edited by V. A. Kostelecký and D. K. Campbell (North-Holland, Amsterdam, 1985). These applications must now be supplemented with additional ones in atomic physics and in string theory.

[2]D. R. Truax, V. A. Kostelecký, and M. M. Nieto, J. Math. Phys. **27**, 354 (1986); V. A. Kostelecký, M. M. Nieto, and D. R. Truax, *ibid.* **27**, 1419 (1986); V. A. Kostelecký and D. R. Truax, *ibid.* **28**, 2480 (1987); B. W. Fatyga, V. A. Kostelecký, and D. R. Truax, *ibid.* **30**, 291 (1989).

[3]J. Beckers, L. Gagnon, V. Hussin, and P. Winternitz, Lett. Math. Phys. **13**, 113 (1987).

[4]C. Elphick, J. Math. Phys. **28**, 1243 (1987).

[5]R. Cianci, J. Math. Phys. **29**, 2156 (1988).

[6]V. A. Kostelecký and J. M. Rabin, J. Math. Phys. **25**, 2748 (1984).

[7]A good introductory text is H. Lamb, *Hydrodynamics* (Dover, New York, 1945).

[8]W. F. Noh and M. H. Protter, J. Math. Mech. **12**, 149 (1963).

[9]Note that this is an identity, *not* an expansion about $s=0$. The signs are not those of an expansion, and the quantities $\mathbf{W}$ and $\mathbf{V}^{(k)}$ are not evaluated at $s=0$.

[10]B. L. Rozdestvenskii and N. N. Janenko, *Systems of Quasilinear Equations and Their Applications to Gas Dynamics* (AMS, Providence, 1983).

On constrained mechanical systems: D'Alembert's and Gauss' principles

Franco Cardin and Giovanni Zanzotto
Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy

(Received 22 November 1988; accepted for publication 15 February 1989)

A geometric formulation of the classical principles of D'Alembert and Gauss in analytical mechanics is given, and their equivalence for possibly non-Riemannian mechanical systems is shown, in the case of ideal holonomic constraints. This is done by means of a Gauss' function, which is defined in a natural way on the bundle of two-jets on the configuration space, and which gives the "intensity" of the "reaction forces" of the constraints. It is originated by a metric structure on the bundle of semibasic forms on the phase space determined by the Finslerian kinetic energy functions of the mechanical system.

I. INTRODUCTION

There are several reasons for a revisitation, from a geometrical point of view, of the well known Gauss' principle of "least constraint." [1]

Indeed, it is remarkable that this topic seems to not have received a great deal of attention in the formulation of analytical mechanics, within the framework of differential geometry, that took place in the last decades. For instance, we can quote, among others, the books by Godbillon,[2] Libermann and Marle,[3] or Abraham and Marsden,[4] where this principle is not treated.

However, Gauss' principle seems to be of undoubtable foundational relevance and worthy of interest, especially in a broad generality of choice for the form of the kinetic energy function of the mechanical system and for the active forces, possibly nonconservative and dependent upon the distribution of the generalized velocities in the phase space.

We wish to stress that an accurate study of Gauss' principle is also interesting from the point of view of the possible applications. Indeed, for instance, in a rather recent work by Lilov and Lorer,[5] an algorithm for a dynamical investigation of a multirigid body system is proposed on the basis of Gauss' principle. The two authors remark that "the main advantage of this approach, ..., over the derivation and investigation of the nonlinear equation of motion, ..., is that, using Gauss' principle, the accelerations can be found out from the condition for minimum of a functional, ..., and there is the possibility to use effectively the mathematical programming methods, and especially the recent iterative algorithms for constraint and unconstraint minimization of quadratic functionals."

Usually, in the analytical mechanics textbooks, Gauss' principle is stated for mechanical systems composed of a finite number of material particles under the presence of ideal constraints. Such a procedure excludes the finite-dimensional systems with an infinite number of particles, like rigid bodies, unless some limiting processes are carried out, which are sometimes lacking the necessary rigor. About this, we agree with Wang (Ref. 6, p. vii), when he states that rigid bodies should be regarded as primitive concepts like mass points and treated as such.

The present version of Gauss' principle complies with these ideas, and can be applied as soon as the finite-dimensional mechanical systems are assigned a "free" configuration manifold and a constraint manifold, together with a kinetic energy function and an "active force" field, both independent of the constraints. Mechanical systems composed of a finite number of mass points and/or rigid bodies are thus equally treated in a natural way. Indeed, the general form for the kinetic energy that we adopt gives a generalization of Gauss' principle to the case of Finslerian (possibly non-Riemannian) systems.

In order to accomplish our goal, a suitable statement of D'Alembert's principle is needed. Our approach regarding the latter is close to that of Vershik and Faddeev.[7] However, since we focus on the holonomic ideal case, constraints in this work are treated in a somewhat different way. Here the point of view and techniques of Ref. 2 are adopted, so that the present versions of the principles of D'Alembert and Gauss follow the spirit of the construction in Ref. 2.

A Gauss function is introduced in a natural way on the bundle of two-jets on the configuration space, by means of a "kinetic" metric on the bundle of semibasic forms on the velocity phase space. This norm measures the "deviation forces" that are needed for the mechanical system to undergo the motions associated with *a priori* chosen semisprays. They are compared with the only dynamically possible motion compatible with the constraints, i.e., with the motion $\mathbb{M}$ associated with semisprays determined by D'Alembert's equation. The final result, that is, the equivalence of the principles of Gauss and D'Alembert, is basically a characterization of $\mathbb{M}$ in terms of either of the following properties: (a) $\mathbb{M}$ is the unique motion along which the deviation forces are of the kind that the ideal constraints are capable of exerting; and (b) along $\mathbb{M}$ the above forces minimize, in a certain sense, the Gaussian function. Local expressions of all the definitions and results are given.

II. CONSTRAINED MECHANICAL SYSTEMS

We begin with a brief summary of some fundamental notions and results, and an introduction of the notations. We refer mainly to Refs. 2 and 3 for details.

Let M be a differentiable (C^∞) manifold of dimension n, and without boundary: $\partial M = \emptyset$ (Ref. 2, pp. 57 and 58).

To any coordinate system (x^i) on M are canonically associated natural coordinate systems $(x^i, \dot{x}^i)$ and $(x^i, \dot{x}^i, \delta x^i, \delta \dot{x}^i)$ on TM and TTM, respectively. Here and in the sequel, latin indices run from 1 to n.

 0022-2488/89/071473-07$02.50

The canonical tangent projections, $\tau_M: TM \to M$ and $T_M: TTM \to TM$, thus have the local expressions

$$\tau_M: (x^i, \dot{x}^i) \mapsto x^i, \quad \tau_{TM}(x^i, \dot{x}^i, \delta x^i, \delta \dot{x}^i) \mapsto (x^i, \dot{x}^i), \tag{2.1}$$

respectively. The space TTM is fibered in two ways on TM: either by means of the projection $\tau_{TM}: TTM \to TM$ introduced above, or by means of $T\tau_M: TTM \to TM$, whose local expression is as follows:

$$T\tau_M: (x^k, \dot{x}^k, \delta x^k, \delta \dot{x}^k) \mapsto (x^k, \delta x^k). \tag{2.2}$$

As usual, Th denotes the tangent of a mapping h.

The kernel of $T\tau_M$ is a canonical subbundle of TTM, called the *vertical tangent bundle* to TM, and denoted by VTM (see Ref. 3, p. 54). The elements of VTM are termed *vertical* and have the local expression $(x^k, \dot{x}^k, 0, \delta \dot{x}^k)$. We denote by $V: TM \to VTM$ the Liouville vertical vector field on TM, generating the one-parameter group of positive dilations of TM. In natural coordinates, it has the local expression

$$V = \dot{x}^i \frac{\partial}{\partial \dot{x}^i}, \tag{2.3}$$

so that, for instance, $V \cdot f = \dot{x}^i (\partial f / \partial \dot{x}^i)$ for all $f \in D(TM)$, the algebra of the differentiable functions on TM.

In a similar way, we will denote by $\pi_M: T^*M \to M$ and $\pi_{TM}: T^*TM \to TM$ the cotangent projections. To (x^i), the systems of coordinates (x^i, p_i) and $(x^i, \dot{x}^i, p_i, r_i)$ on T^*M and T^*TM, respectively, are canonically associated. In this way, the above projections π_M and π_{TM} have the local expressions

$$\pi_M: (x^i, p_i) \mapsto x^i, \quad \pi_{TM}: (x^i, \dot{x}^i, p_i, r_i) \mapsto (x^i, \dot{x}^i). \tag{2.4}$$

Following Ref. 2, we introduce the vector bundle $\beta: \tau_M^* T^*M \to TM$ of semibasic forms on TM. The total space $\tau_M^* T^*M$ can be identified with the subspace $\cup_{y \in M} \tau_M^{-1}(y) \times \pi_M^{-1}(y)$ of $TM \times T^*M$, and the projection β is the restriction of the projection of $TM \times T^*M$ onto TM (see Ref. 2, p. 166).

By Proposition 2.2 and Remark 2 in Ref. 3, pp. 55 and 56, we will identify the vector bundle $\beta: \tau_M^* T^*M \to TM$ of the semibasic forms on TM with the subbundle of T^*TM, $\pi_{TM}|_{(VTM)^0}: (VTM)^0 \to TM$, the annihilator of the vertical bundle VTM.

Hence, for simplicity, we will also use π_{TM} for semibasic forms, instead of β. Of course, in natural coordinates the elements of $(VTM)^0$ have the expression $(x^i, \dot{x}^i, p_i, 0)$. A differential one-form σ on TM is semibasic, if and only if it has, in natural coordinates, the local expression

$$\sigma = \sigma_i(x^h, \dot{x}^h) dx^i, \tag{2.5}$$

where $\sigma_i(x^h, \dot{x}^h)$ are given functions on TM (see Ref. 2, p. 165 or Ref. 3, pp. 56–58).

The identification of the vector bundles $(VTM)^0$ and $(VTM)^*$ on TM (see Propositions 2.4, 2.5, and 3.11 in Ref. 3, pp. 55–58), allows for the definition of a vector bundle morphism $v^*: T^*TM \to (VTM)^0$ (also see Proposition 6.9 in Ref. 3, p. 70), whose local expression is

$$v^*: (x^i, \dot{x}^i, p_i, r_i) \mapsto (x^i, \dot{x}^i, r_i, 0). \tag{2.6}$$

The morphism v^* induces an endomorphism, still denoted by v^* and called *vertical*, of the $D(TM)$-algebra $\Lambda(TM)$ of the differential forms on TM. It is locally determined by the conditions (Ref. 2, p. 161)

$$v^* f = f, \quad \text{for any } f \text{ in } D(TM), \quad v^*(dx^i) = 0,$$
$$v^*(d\dot{x}^i) = dx^i. \tag{2.7}$$

The subalgebra $B(TM)$ of semibasic differential forms is the range and kernel of v^*. Beside the usual exterior differential d, by means of v^* the *vertical differential* d_v is also defined on $\Lambda(TM)$. It is uniquely characterized by the relations (see Ref. 2, p. 163)

$$d_v f = v^* df, \quad d_v(df) = -d(v^* df), \quad \text{for any } f \text{ in } D(TM). \tag{2.8}$$

Locally, d_v is determined by

$$d_v f = \frac{\partial f}{\partial \dot{x}^i} dx^i, \quad d_v(dx^i) = 0, \quad d_v(d\dot{x}^i) = 0, \tag{2.9}$$

and the relation $dd_v = -d_v d$ holds.

We now recall the following:

Definition (Ref. 2, p. 169): A mechanical system $\mathscr{M}$ is a triple (M, K, Φ) where (a) M is a differentiable manifold of dimension n, the *configuration space*; (b) K is a differentiable function on TM, the *kinetic energy*; and (c) Φ is a semibasic differential one-form on TM, the *force field*.

The differential two-form $dd_v K$ on TM is called the *fundamental form* of the mechanical system $\mathscr{M}$, which is called *regular* if $dd_v K$ is symplectic on TM. This happens if and only if locally we have (Ref. 2, p. 169)

$$\det\left(\frac{\partial^2 K}{\partial \dot{x}^i \partial \dot{x}^j}\right) \neq 0. \tag{2.10}$$

Now, the space T^2M of two-jets of M can be defined by (see Ref. 3, p. 372)

$$T^2M = \{w \in TTM: \tau_{TM}(w) = T\tau_M(w)\}, \tag{2.11}$$

and the canonical submersion $\tau_{TM}^{21}: T^2M \to TM$ can be identified with $\tau_{TM}|_{T^2M}$ or $T\tau_M|_{T^2M}$.

Since τ_{TM} and $T\tau_M$ have the local expressions (2.1) and (2.2), the elements of T^2M are given locally by $(x^k, \dot{x}^k, \dot{x}^k, \delta \dot{x}^k)$. As usual, they will be denoted by $(x^k, \dot{x}^k, \ddot{x}^k)$, with $\ddot{x}^k$ written for $\delta \dot{x}^k$. In this way, the local expression for the canonical submersion $\tau_{TM}^{21}: T^2M \to TM$ is

$$\tau_{TM}^{21}: (x^k, \dot{x}^k, \ddot{x}^k) \mapsto (x^k, \dot{x}^k). \tag{2.12}$$

A *semispray* Y is a section of τ_{TM}^{21}. Of course, Y can be seen as a vector field $Y: TM \to TTM$, satisfying the condition

$$\tau_{TM} \circ Y = T\tau_M \circ Y. \tag{2.13}$$

In other words, a semispray Y is a vector field on TM, which is at the same time a section of τ_{TM} and $T\tau_M$. Locally, such a Y is given by

$$Y = \dot{x}^i \frac{\partial}{\partial x^i} + b^i(x^k, \dot{x}^k) \frac{\partial}{\partial \dot{x}^i}, \tag{2.14}$$

with $b^i(x^k, \dot{x}^k)$ given functions on TM. From this, it is im-

mediately seen that the integral curves of $Y: TM \to T^2M$ are velocity curves of the curves on M of which Y [or $b^i(x^k, \dot{x}^k)$] is the acceleration at each point. These base curves on M are also called the *solutions* of Y, because they locally satisfy the system of equations

$$\frac{d^2x^i}{dt^2} = b^i\left(x^k, \frac{dx^k}{dt}\right). \tag{2.15}$$

For this reason, semisprays are also called second-order differential equations.

The following proposition holds (Ref. 2, p. 170). Let $\mathscr{M}$ be a regular mechanical system. Then there exists a unique vector field X on TM, such that

$$\mathbf{i}_X dd_v K = d(K - V \cdot K) + \Phi. \tag{2.16}$$

Here, the symbol $\mathbf{i}_X$ denotes as usual the interior product of a differential form by a vector field. The vector field X is called the *dynamical system* associated with $\mathscr{M}$.

Furthermore, it can be proved that the dynamical system X associated with a regular mechanical system $\mathscr{M}$ is a semispray (see Ref. 2, p. 170).

Let the local expression of the semibasic one-form Φ be $\Phi = \Phi_i(x^k, \dot{x}^k)dx^i$; then it can be proved that the solutions of the dynamical system X associated with $\mathscr{M} = (M, K, \Phi)$ locally satisfy the Lagrange equations (Ref. 2, p. 171),

$$\frac{d}{dt}\left(\frac{\partial K}{\partial \dot{x}^k}\right) - \frac{\partial K}{\partial x^k} = \Phi_k. \tag{2.17}$$

We now give the following:

Definition: A mechanical system with bilateral holonomic constraints is a quintuplet, $\mathscr{M}_c = (M, K, \Phi, Q, \mathscr{R})$, where (a) M, K, and Φ are as above; (b) Q, the *constraint*, is an m-dimensional ($m \leqslant n$) imbedded submanifold of M, with imbedding denoted by χ: $Q \to M$ and such that $\partial(cl_M Q) = \emptyset$; (c) $\mathscr{R} \subset (VTM)^0|_{TQ}$ is the total space of a subbundle $\pi_{TM}|_{\mathscr{R}}$: $\mathscr{R} \to TQ$ of the vector bundle $\pi_{TM}|_{(VTM)^0|_{TQ}}$: $(VTM)^0|_{TQ} \to TQ$ of the semibasic forms restricted to TQ.

Here and in the sequel, we of course identify Q with its image in M under χ, as well as TQ with its image in TM under $T\chi$, and so on. Also, for simplicity, in the sequel we will drop the restriction symbol from π_{TM}, since no confusion arises.

We explicitly notice that in (b), the condition $\partial(cl_M Q) = \emptyset$ expresses the notion that the constraints are "bilateral." The subbundle introduced in (c) describes the forces that the constraints are capable of exerting, which are called the *admissible constraint reaction forces.*

For brevity, $\mathscr{M}_c$ will be referred to as the *constrained mechanical system*; it will be called *regular* when both the systems $\mathscr{M} = (M, K, \Phi)$ and $\mathscr{Q} = (Q, \tilde{K}, \tilde{\Phi})$ are such, where

$$\tilde{K} = (T\chi)^* K = K \circ T\chi, \quad \tilde{\Phi} = (T\chi)^* \Phi. \tag{2.18}$$

Since χ is an imbedding and the fibers in TM are linear, it is easily verified that if $\mathscr{M}$ is regular, $\mathscr{Q}$ is also regular.

We will consider the case of *ideal* constraints, in which $\mathscr{R}$ is specified as follows:

$$\mathscr{R} = v^*((TTQ)^0). \tag{2.19}$$

An explicit equivalent description of $\mathscr{R}$ is the following:

$$\mathscr{R} = \{r \in (VTM)^0 : \pi_{TM}(r) = u \in TQ \subset TM, \quad \text{Ker } r = T_u TQ \subset T_{T\chi(u)} TM\}. \tag{2.20}$$

The above definitions of holonomic constraints and admissible constraint reaction forces strongly rely on the introduction of *one* assigned constraint submanifold Q of M. In Ref. 7, where anholonomic constraints are treated in a very general setting, holonomic constraints possibly emerge as foliations of M, introduced by suitable repeated integrations of distributions on TM. The simpler procedure we follow, which is closer to the classical treatments, seems more natural from a physical point of view.

A characterization of the sections of $\pi_{TM}|_{\mathscr{R}}: \mathscr{R} \to TQ$, which is important in the sequel, is given by the following:

Lemma: Let ρ be a differential semibasic one-form, i.e., a section of $\pi_{TM}|_{(VTM)^0}$: $(VTM)^0 \to TM$. Then $\rho \circ T\chi$ is a section of $\pi_{TM}|_{\mathscr{R}}: \mathscr{R} \to TQ$, if and only if

$$(T\chi)^* \rho = 0. \tag{2.21}$$

Proof: Indeed, denoting by $\tilde{\rho}$: $TQ \to T^*TQ$ the differential one-form $(T\chi)^*\rho$, it is $\tilde{\rho} = 0$, if and only if, for any arbitrarily fixed $u \in TQ$, it results that

$$\mathbf{i}_z \tilde{\rho}(u) = 0, \quad \text{for any } z \in T_u TQ. \tag{2.22}$$

But this is true, if and only if

$$\mathbf{i}_{TT\chi(z)} \rho(T\chi(u)) = 0, \quad \text{for any } z \in T_u TQ, \tag{2.23}$$

that is, if and only if $\rho(T\chi(u)) \in \mathscr{R}$ for all $u \in TQ$, or, if and only if $\rho \circ T\chi$ is a section of $\pi_{TM}|_{\mathscr{R}}: \mathscr{R} \to TQ$.

The local expressions will be useful, and we give them in detail. Let (x^i) and (q^α) be local coordinate systems on M and Q, respectively (here greek indices run from 1 to m). Furthermore, let $\rho = \rho_i(x^h, \dot{x}^h)dx^i$ be a semibasic differential one-form, and let $u = (q^\alpha, \dot{q}^\alpha)$ be a fixed arbitrary element of TQ, so that $z \in T_u TQ$ has coordinates $(q^\alpha, \dot{q}^\alpha, \delta q^\alpha, \delta \dot{q}^\alpha)$ and

$$TT\chi(z) = (\chi^k(q^\alpha), D_\sigma \chi^j(q^\alpha)\dot{q}^\sigma, D_\sigma \chi^j(q^\alpha)\delta q^\sigma, D_\gamma D_\sigma \chi^h(q^\alpha)\dot{q}^\sigma \delta q^\gamma + D_\sigma \chi^h(q^\alpha)\delta \dot{q}^\sigma). \tag{2.24}$$

As usual, D_σ denotes the partial derivative in $\mathbb{R}^m$. Then, since

$$\rho \circ T\chi = \rho_i(\chi^k(q^\alpha), D_\sigma \chi^j(q^\alpha)\dot{q}^\sigma)dx^i \tag{2.25}$$

and

$$(T\chi)^* \rho = \rho_i(\chi^k(q^\alpha), D_\sigma \chi^j(q^\alpha)\dot{q}^\sigma) D_\alpha \chi^i(q^\alpha) dq^\alpha, \tag{2.26}$$

(2.23)–(2.25) yield

$$\rho_i(\chi^k(q^\gamma), D_\sigma \chi^j(q^\gamma)\dot{q}^\sigma) D_\alpha \chi^i(q^\gamma)\delta q^\alpha = 0, \quad \text{for all } \delta q^\alpha \in \mathbb{R}, \tag{2.27}$$

which of course is true if and only if

$$\rho_i(\chi^k(q^\gamma), D_\sigma \chi^j(q^\gamma)\dot{q}^\sigma) D_\alpha \chi^i(q^\gamma) = 0, \quad \text{for all } (q^\alpha, \dot{q}^\alpha). \tag{2.28}$$

By (2.26), we see that (2.28) is equivalent to (2.21).

Remark 1: The characterization of the constraint reaction forces given by (2.21) or (2.28), easily leads to the following, which will also be used in the sequel.

Let ρ be a semibasic differential one-form. Then $\rho \circ T\chi$ is a section of $\pi_{TM}\big|_{\mathscr{R}}: \mathscr{R} \to TQ$, if and only if

$$(T\chi)^*(\mathbf{i}_Y \rho) = 0, \tag{2.29}$$

for all fields Y: $TM \to TTM$, such that there exists a field Z: $TQ \to TTQ$ for which the relation

$$Y \circ T\chi = TT\chi \circ Z \tag{2.30}$$

holds.

We will call the vector fields Y on TM, and Z on TQ, $T\chi$-related, when they satisfy (2.30). In this way, Y is an extension to TM of a vector field Z on TQ.

To prove the assertion of Remark 1, let us recall that we can write, for any $v \in TQ$,

$$\begin{aligned}(T\chi)^*(\mathbf{i}_Y \rho)(v) &= \mathbf{i}_{Y(T\chi(v))}[\rho(T\chi(v))] \\ &= \mathbf{i}_{TT\chi(U(v))}[\rho(T\chi(v))] = 0.\end{aligned} \tag{2.31}$$

Hence the first term in (2.31) is zero for any Y, if and only if $\rho(T\chi(v)) \in \mathscr{R}$ for any $v \in TQ$.

Since $\mathbf{i}_Y \rho$ obviously has the meaning of "power" of a force along a "path," Remark 1 shows that constraint forces are characterized by the fact that they do no work on vector fields on TM that extend vector fields tangent to TQ.

Remark 2: Before concluding this section, we mention without details that Remark 1 implies a further characterization of the admissible constraint forces.

Let ρ be a differential semibasic form. Then $\rho \circ T\chi$ is a section of $\pi_{TM}\big|_{\mathscr{R}}: \mathscr{R} \to TQ$, if and only if

$$(T\chi)^*(\mathbf{i}_{Y^c} \rho) = 0, \tag{2.32}$$

for all fields Y: $M \to TM$ that are χ-related to some vector field Z: $Q \to TQ$. In (2.32), Y^c indicates the *complete lift* to TM of a vector field Y on M [see Yano and Ishihara (Ref. 8, p. 14)]. If $Y = b^i(x^k)(\partial/\partial x^i)$ locally, it is $Y^c = b^i(x^k)(\partial/\partial x^i) + D_r b^i(x^k)\dot{x}^r(\partial/\partial \dot{x}^i)$. By Remark 1, to prove the assertion of Remark 2, we just need to notice that

$$Y \circ \chi = T\chi \circ Z \Leftrightarrow Y^c \circ T\chi = TT\chi \circ Z^c,$$

which we give without proof.

Remark 2 is interesting because it clarifies the "physical meaning" of the constraint forces. Indeed, it shows that to characterize them, it is enough that they do no work just on complete lifts to TM of vector fields on M that extend fields on Q. The above local expression of the complete lift clearly shows that the latter condition basically amounts to the classical one requiring that, in the ideal case, admissible reaction forces do no work on "displacements" tangent to the constraint Q.

III. D'ALEMBERT'S PRINCIPLE

We introduce, in connection with a given mechanical system $\mathscr{M} = (M, K, \Phi)$ and with a given, arbitrary, semispray Y: $TM \to T^2M$, the following *deviation* differential one-form ρ_Y:

$$\rho_Y = \mathbf{i}_Y dd_v K - d(K - V \cdot K) - \Phi. \tag{3.1}$$

These forms ρ_Y have the meaning of "forces to be added" to the given force field Φ, in order that a semispray Y, chosen *a priori*, be the dynamical system associated with the mechanical system $(M, K, \Phi + \rho_Y)$. In fact, the following holds:

Lemma: The deviation differential one-forms are semibasic.

Proof: If $Y = \dot{x}^i(\partial/\partial x^i) + b^i(x^k, \dot{x}^k)(\partial/\partial \dot{x}^i)$ locally, with $b^i(x^k, \dot{x}^k)$ given functions, it is not difficult to show that the local expression for ρ_Y is as follows:

$$\rho_Y = \left(\frac{\partial^2 K}{\partial \dot{x}^k \partial \dot{x}^i} b^k + \frac{\partial^2 K}{\partial x^k \partial \dot{x}^i} \dot{x}^k - \frac{\partial K}{\partial x^i} - \Phi_i\right) dx^i. \tag{3.2}$$

Definition: Two semisprays Y and Y' are said to be equivalent, $Y \approx Y'$, when their restrictions to TQ are equal, i.e., when $Y \circ T\chi = Y' \circ T\chi$.

Now, our goal is the construction of a dynamics for the constrained mechanical system $\mathscr{M}_c = (M, K, \Phi, Q, \mathscr{R})$. In order to do this, a twofold result must be obtained. Basically, we first need to select the semisprays X that (besides defining a dynamical system on the overall manifold M, also) define a dynamical system on the constraint manifold Q, meaning that the *solutions* of X must be all on Q when the initial data are in TQ.

Furthermore, the deviation differential one-forms ρ_X, that is, the forces necessary to "maintain" the system on the constraint, must be of the kind that the constraints are capable of exerting, i.e., $\rho_X \circ T\chi$ must be a section of the bundle $\pi_{TM}\big|_{\mathscr{R}}: \mathscr{R} \to TQ$ of admissible constraint forces, introduced in Sec. II.

We now show that, in the case of regular systems, the two properties above characterize X in a unique way on the constraint. In fact, the following theorem holds.

D'Alembert's principle: Let $\mathscr{M}_c$ be a regular constrained mechanical system. Then, up to equivalence, there is a unique semispray X: $TM \to T^2M$, such that the following D'Alembert's equation holds:

$$\mathbf{i}_X dd_v K - d(K - V \cdot K) = \Phi + \rho_X, \tag{3.3}$$

with $\rho_X \circ T\chi$ a section of $\pi_{TM}\big|_{\mathscr{R}}: \mathscr{R} \to TQ$. The solutions of X have the property that their image is all on Q as soon as the initial values are in TQ.

Proof: Let us set for brevity $\omega = dd_v K$, and $\sigma = d(K - V \cdot K)$, and let us consider the following pull-backs of ω, σ, and Φ:

$$\tilde{\omega} = (T\chi)^*\omega, \quad \tilde{\sigma} = (T\chi)^*\sigma, \quad \tilde{\Phi} = (T\chi)^*\Phi. \tag{3.4}$$

Since K is a function (a zero-form), both d and d_v commute with the pull-back, so that it is $\tilde{\omega} = dd_v \tilde{K}$ [see (2.18)]; hence the hypothesis that $\mathscr{M}_c$ is regular implies that $\tilde{\omega}$ is symplectic on TQ. Then, Theorems 1.4 and 1.6 in Ref. 1 (p. 170), applied to the mechanical system $\mathscr{Q} = (Q, \tilde{K}, \tilde{\Phi})$, guarantee the uniqueness of the semispray $\tilde{X}$: $TQ \to T^2Q$, such that the following equation holds:

$$\mathbf{i}_{\tilde{X}} \tilde{\omega} - \tilde{\sigma} = \tilde{\Phi}. \tag{3.5}$$

Now, let us consider an arbitrary semispray X: $TM \to T^2M$, $T\chi$-related to $\tilde{X}$:

$$X \circ T\chi = TT\chi \circ \tilde{X}. \tag{3.6}$$

Then, considering the deviation one-form ρ_X, we have [see (3.1) and (3.3)]

$$\mathbf{i}_X dd_v K - d(K - V\cdot K) = \Phi, \tag{3.7}$$

where ρ_X is such that

$$(T\chi)^*\rho_X = (T\chi)^*(\mathbf{i}_X\omega - \sigma - \Phi) \tag{3.8a}$$

$$= (T\chi)^*(\mathbf{i}_X\omega) - \tilde{\sigma} - \tilde{\Phi} \tag{3.8b}$$

$$= \mathbf{i}_{\tilde{X}}\tilde{\omega} - \tilde{\sigma} - \tilde{\Phi} \tag{3.8c}$$

$$= 0. \tag{3.8d}$$

Equality (3.8c) holds because of (3.6) and example 1.8 (iii) in Ref. 2, p. 89, whereas (3.8d) is just (3.5). By the lemma in Sec. II, (3.8) shows that $\rho_X \circ T\chi$ is a section of admissible reaction forces.

The uniqueness of X, up to equivalence, is a consequence of (3.6) and of the uniqueness of $\tilde{X}$. Finally, the last assertion in the statement is true because, again by (3.6), TQ is an integral manifold of the semispray X.

We will say that any of the equivalent semisprays X above is a *dynamical system* associated with the constrained mechanical system $\mathscr{M}_c$. Here it is worth noticing that the class of semisprays X, $T\chi$-related to $\tilde{X}$, is not empty.

We also remark that the theorem above shows that, for fixed $\mathscr{M}$, the deviation forms ρ_Y connected with a semispray Y, are not, in general, forces that the constraints are capable of exerting. Indeed, they are such, only when Y is a dynamical system associated with $\mathscr{M}_c$. This is the reason why we refrained altogether from calling the ρ_Y "constraint reaction forms."

The local expressions will be useful in Sec. IV. If, locally, the field $\tilde{X}$, uniquely determined by Eq. (3.5), is given by

$$\tilde{X} = \dot{q}^\alpha \frac{\partial}{\partial q^\alpha} + \tilde{a}^\sigma(q^\beta, \dot{q}^\beta)\frac{\partial}{\partial \dot{q}^\sigma},$$

let us consider a semispray

$$X = \dot{x}^i \frac{\partial}{\partial x^i} + a^i(x^k, \dot{x}^k)\frac{\partial}{\partial \dot{x}^i},$$

such that Eq. (3.6) holds, that is, such that

$$a^i(\chi^k(q^\alpha), D_\sigma\chi^j(q^\alpha)\dot{q}^\sigma) = \tilde{a}^\beta(q^\alpha, \dot{q}^\sigma)D_\beta\chi^i(q^\alpha) + D_\gamma D_\sigma\chi^i(q^\alpha)\dot{q}^\sigma\dot{q}^\gamma. \tag{3.9}$$

Then, writing $\rho_X = \rho_{X_i}(x^h, \dot{x}^h)dx^i$, with

$$\rho_{X_i} = \frac{\partial^2 K}{\partial \dot{x}^k \partial \dot{x}^i}a^k + \frac{\partial^2 K}{\partial x^k \partial \dot{x}^i}\dot{x}^k - \frac{\partial K}{\partial x^i} - \Phi_i \tag{3.10}$$

[see (3.2)], by (3.3), we conclude that the equation

$$\rho_{X_i}(\chi^k(q^\alpha), D_\sigma\chi^j(q^\alpha)\dot{q}^\sigma)D_\beta\chi^i(q^\alpha) = 0 \tag{3.11}$$

holds, if and only if the functions $a^i(x^k, \dot{x}^k)$ satisfy (3.9).

It is worth noticing explicitly that the semispray $\tilde{X}$: $TQ \to T^2Q$, introduced in the proof of D'Alembert's principle, is such that

$$\mathbf{i}_{\tilde{X}} dd_v \tilde{K} - d(\tilde{K} - \tilde{V}\cdot\tilde{K}) = \tilde{\Phi}, \tag{3.12}$$

where $\tilde{V}$ is the canonical Liouville field on TQ. Hence, by (3.12), the solutions of $\tilde{X}$ locally satisfy the equations of Lagrange relative to the mechanical system $\mathscr{Q} = (Q, \tilde{K}, \tilde{\Phi})$,

$$\frac{d}{dt}\left(\frac{\partial \tilde{K}}{\partial \dot{q}^\alpha}\right) - \frac{\partial \tilde{K}}{\partial q^\alpha} = \tilde{\Phi}_\alpha. \tag{3.13}$$

Equation (3.12) is easily seen to hold because $\tilde{\omega} = dd_v\tilde{K}$ (see above) and $(T\chi)^*(V\cdot K) = \tilde{V}\cdot\tilde{K}$. The latter equality is a consequence of the fact that the fields V and $\tilde{V}$ are $T\chi$-related. Thus (3.12) is immediately derived from (3.5).

We conclude this section with a statement of the classical "energy theorem."

Let X: $TM \to T^2M$ be a semispray solving the equation of D'Alembert (3.3) and let $\tilde{X}$ be the $T\chi$-related semispray $TQ \to T^2Q$, solving Eq. (3.12). Let η: $[a,b] \to TQ$ be an integral curve of $\tilde{X}$, so that $T\chi \circ \eta$ is an integral curve of X. Then

$$\int_a^b \eta^* \circ (T\chi)^*\Phi = [V\cdot K - K]_{T\chi[\eta(a)]}^{T\chi[\eta(b)]}$$

$$\equiv [(\tilde{V}\cdot\tilde{K} - \tilde{K})]_{\eta(a)}^{\eta(b)}. \tag{3.14}$$

To prove that Eq. (3.14) holds, we recall that, being the semisprays $\tilde{X}$ and X, $T\chi$-related, Remark 1 of Sec. II gives

$$(T\chi)^*(\mathbf{i}_Y\rho_Y) = 0. \tag{3.15}$$

Then, since X solves D'Alembert's equation (3.3), and taking into account that dd_vK is symplectic, evaluation on X of the forms appearing in (3.3), gives

$$(T\chi)^*[X\cdot(V\cdot K - K)] = (T\chi)^*(\mathbf{i}_X\Phi). \tag{3.16}$$

Equation (3.16) immediately yields (3.14).

IV. GAUSS' PRINCIPLE

As in the preceding sections, we do not necessarily consider K to be quadratic on the fibers $\tau_M^{-1}(x)$ of TM. This is the case, for instance, of Newtonian classical mechanics. Rather, we have in mind certain generalizations, such as Finslerian mechanics (see, for example, Ref. 2, p. 130, and also Rund,[9] Ruiz,[10] and Eringen[11]); we do not require here that K be a Riemannian metric.

Let $\mathscr{M}_c$ be a regular constrained mechanical system. Following Ref. 7, we introduce the (2,0)-tensor field Π on TM, defined by means of the relation

$$dd_vK[\Pi(\sigma), H] = \mathbf{i}_H\theta, \tag{4.1}$$

which is to hold for every differential one-form σ, and vector field H, on TM. Also, a new (2,0)-tensor field Γ on TM is defined, such that

$$\Gamma(\alpha, \gamma) = \Pi(\alpha, v^*\gamma), \tag{4.2}$$

for any differential one-forms α and γ on TM. Of course, we denote as usual with the same symbol both the morphism and the bilinear form induced by the (2,0)-tensor field Π.

A straightforward calculation shows that, setting for brevity,

$$K_{ij}(x^r, \dot{x}^s) = \frac{\partial^2 K}{\partial \dot{x}^i \partial \dot{x}^j}, \tag{4.3}$$

the local expression of Γ is

$$\Gamma = K^{ij}(x^r, \dot{x}^s)\,\frac{\partial}{\partial \dot{x}^i} \otimes \frac{\partial}{\partial \dot{x}^j}, \tag{4.4}$$

where

$$K_{ij}K^{jh} = \delta_i{}^h. \tag{4.5}$$

Hence Γ is connected to the Hessian of the function K along the fibers of TM (see Spivak, Ref. 12, Vol, 2, pp. 206 and 207).

From now on, we assume that K is the "energy" of a Finslerian structure on M; explicitly, we suppose that there exists a function F on TM such that $K = F^2$, with the following properties: (a) $F(v) \neq 0$, $F(\lambda v) = |\lambda|F(v)$, if $v \neq 0$, $v \in TM$, and $\lambda \in \mathbb{R}$; (b) the functions $K_{ij}(x^r, \dot{x}^s)$ define a positive-definite quadratic form on $V_v TM$, at every point $v = (x^r, \dot{x}^s)$ of TM with $\dot{x}^s \neq 0$.

It is clear that the (2,0)-tensor field Γ introduced in (4.2)–(4.5) is but the dual to the tensor $K_{ij}(x^r, \dot{x}^s) \times d\dot{x}^i \otimes d\dot{x}^j$ on TM, canonically associated with the Finslerian metric F on M (see Ref. 12, Vol. 2, p. 208).

Owing to its structure, Γ generates a metric for the quotient bundle $T^*TM/(VTM)^0$, which we will continue to indicate by Γ. Indeed, in natural local coordinates, letting $[\alpha] = [x^r, \dot{x}^s, r_i, p_j]$ be the equivalence class in $T^*TM/(VTM)^0$ of an element $\alpha = (x^r, \dot{x}^s, r_i . p_j) \in T^*TM$, we have

$$\Gamma([\alpha],[\alpha]) = \Gamma(\alpha, \alpha) = K^{ij}(x^r, \dot{x}^s)p_i p_j, \tag{4.6}$$

The number $\Gamma([\alpha],[\alpha]) = K^{ij}(x^r, \dot{x}^s)p_i p_j$ does not depend on the coordinates r_i of the element α, that is, it does not depend on the particular representative chosen for the equivalence class $[\alpha]$, so that the function Γ is well defined on equivalence classes.

By means of v^* (see Sec. II), we now construct the vector bundle isomorphism ω: $T^*TM/(VTM)^0 \to (VTM)^0$, such that $\omega^* = \omega \circ pr$, where pr: $T^*TM \to T^*TM/(VTM)^0$ is the usual quotient projection. As mentioned in Sec. II, $(VTM)^0$ is identified with the vector bundle of semibasic forms on TM by means of the results of Proposition 2.2 in Ref. 3 (p. 55).

Using natural coordinates, ω^{-1} has the local expression

$$\omega^{-1}\colon (x^r, \dot{x}^s, p_i, 0) \mapsto [x^r, \dot{x}^s, r_j, p_i], \tag{4.7}$$

where the r_j are arbitrarily fixed real numbers that label the elements in an equivalence class.

It is now clear that, through ω^{-1}, Γ determines a well-defined metric on the bundle $(VTM)^0$ of semibasic forms on TM. Let $w = (x^k, \dot{x}^k, \ddot{x}^k)$ be a given element of T^2M, with $\tau^{21}_{TM}(w) = v = (x^k, \dot{x}^k) \in TM$. In correspondence with w, let us consider the following element of $(V_v TM)^0$:

$$r_w = \mathbf{i}_w [dd_v K(v)] - d(K - V\cdot K)(v) - \Phi(v). \tag{4.8}$$

The coordinates of r_w are easily seen to be $(x^k, \dot{x}^k, p_i, 0)$, see (3.2), with

$$p_i = \frac{\partial^2 K}{\partial \dot{x}^k \partial \dot{x}^i}\ddot{x}^k + \frac{\partial^2 K}{\partial x^k \partial \dot{x}^i}\dot{x}^k - \frac{\partial K}{\partial x^i} - \Phi_i. \tag{4.9}$$

Then the following Gauss function:

$$\mathrm{G}\colon T^2M \to \mathbb{R}$$

$$w \mapsto \tfrac{1}{2}\Gamma[\omega^{-1}(r_w), \omega^{-1}(r_w)], \tag{4.10}$$

is well defined on T^2M.

To the element w in T^2M (describing the configuration and the distribution of velocities and accelerations of the system), G associates the Γ norm of the deviation force r_w "excited" by w. We recall that, by D'Alembert's principle, $r_w \in \mathscr{R}$, if and only if $w = X(v)$, with X a semispray associated with $\mathscr{M}_c$.

In the case of a system of N mass points, G has the classical expression

$$\mathrm{G} = \sum_1^N{}_i \frac{|m_i \mathbf{a}_i - \mathbf{F}_i(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{v}_1,\ldots,\mathbf{v}_N)|^2}{m_i}, \tag{4.11}$$

with a clear meaning of the symbols.

Let us now fix an arbitrary element u in TQ, and let us introduce the following restricted pull-back of Gauss' function (4.10):

$$G_u = [(T^2\chi)^*\mathrm{G}]\,|_{T^2_u Q}, \tag{4.12}$$

where $T^2\chi$ is the map $T^2Q \to T^2M$, canonically induced by χ, and $T^2_u Q = (\tau^{21}_{TQ})^{-1}(u)$ is the fiber in T^2Q over u. Let us notice that, when natural local coordinates are used in T^2Q, associated with a local coordinate system (q^α) in Q, then the elements of the m-dimensional vector space $T^2_u Q$ have the expression $(\bar{q}^\alpha, \dot{\bar{q}}^\alpha, \ddot{q}^\alpha)$, where $(\bar{q}^\alpha, \dot{\bar{q}}^\alpha)$ are the (fixed) coordinates of $u \in TQ$.

The following proposition holds.

Gauss' principle: The semispray $\tilde{X}$, associated with the mechanical system $\mathscr{Q} = (Q, \tilde{K}, \tilde{\Phi})$, is characterized by the following property for any fixed $u \in TQ$:

$$G_u[\tilde{X}(u)] < G_u(z), \quad \text{for all } z \in T^2_u Q \setminus \{\tilde{X}(u)\}. \tag{4.13}$$

Proof: Let an arbitrary element $z \in T^2_u Q$ be chosen; in natural local coordinates, we have $z = (\bar{q}^\alpha, \dot{\bar{q}}^\alpha, \ddot{q}^\alpha)$, so that

$$T^2\chi(z) = (\chi^k(\bar{q}^\alpha), D_\sigma\chi^j(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}, D_\gamma D_\sigma\chi^h(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}\,\overline{\dot{q}^\gamma} + D_\sigma\chi^h(\bar{q}^\alpha)\ddot{q}^\sigma). \tag{4.14}$$

Hence, by (4.9), the local expression of $r_{T^2\chi(z)} \in (V_{T\chi(u)}\,TM)^0$ is

$$r_{T^2\chi(z)} = (\chi^k(\bar{q}^\alpha), D_\sigma\chi^j(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}, p_i(\ddot{q}^\alpha), 0), \tag{4.15}$$

where, for brevity we have set

$$p_i(\ddot{q}^\alpha) = \bar{K}_{ih}\{D_\gamma D_\sigma\chi^h(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}\,\overline{\dot{q}^\gamma} + D_\sigma\chi^h(\bar{q}^\alpha)\ddot{q}^\sigma\} + \overline{K}'_{ki}D_\sigma\chi^k(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma} - \overline{K}_i - \overline{\Phi}_i, \tag{4.16}$$

with [see (4.4)]

$$\overline{K}_{is} = K_{is}(\chi^k(\bar{q}^\alpha), D_\sigma\chi^j(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}), \tag{4.17}$$

$$\overline{K}'_{ki} = \frac{\partial^2 K}{\partial x^k \partial \dot{x}^i}(\chi^h(\bar{q}^\alpha), D_\gamma\chi^j(\bar{q}^\alpha)\,\overline{\dot{q}^\gamma}), \tag{4.18}$$

$$\overline{K}_i = \frac{\partial K}{\partial x^i}(\chi^k(\bar{q}^\alpha), D_\sigma\chi^j(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}), \tag{4.19}$$

$$\overline{\Phi}_i = \Phi_i(\chi^k, D_\sigma\chi^j(\bar{q}^\alpha)\,\overline{\dot{q}^\sigma}). \tag{4.20}$$

Then, from (4.12), (4.10), (4.4)–(4.7), and (4.15), we explicitly get, for G_u,

$$G_u(\bar{q}^\alpha, \bar{\dot{q}}^\alpha, \bar{\ddot{q}}^\alpha) = \tfrac{1}{2}\bar{K}^{ih}p_i(\bar{\ddot{q}}^\alpha)p_h(\bar{\ddot{q}}^\alpha). \tag{4.21}$$

From this we see that G_u is trivially differentiable, and, by (4.16) and (4.21), that it is indeed a quadratic polynomial in $\ddot{q}^\alpha$. Hence we only need to prove that its differential dG_u vanishes at $\tilde{X}(u)$, and only there. It is sufficient to show this locally. A direct calculation yields the local expression for the differential dG_u at the point $z = (\bar{q}^\alpha, \bar{\dot{q}}^\alpha, \bar{\ddot{q}}^\alpha) \in T_u^2 Q$, as follows:

$$dG_u(z) = \bar{K}^{ij}\bar{K}_{jh}p_i(\bar{\ddot{q}}^\alpha)D_\alpha\chi^h(\bar{q}^\alpha)d\bar{\ddot{q}}^\alpha \tag{4.22a}$$

$$= p_i(\bar{\ddot{q}}^\alpha)D_\sigma\chi^i(\bar{q}^\alpha)d\bar{\ddot{q}}^\sigma, \tag{4.22b}$$

where (4.22b) holds because of (4.5).

Now, let $\tilde{X} = \dot{q}^\alpha(\partial/\partial q^\alpha) + \tilde{a}^\beta(q^\sigma, \dot{q}^\sigma)(\partial/\partial\dot{q}^\beta)$ be the local expression for the dynamical system $\tilde{X}$: $TQ \to T^2Q$, associated with $\mathcal{Q} = (Q, \tilde{K}, \tilde{\Phi})$, so that $\tilde{X}(u) = (\bar{q}^\alpha, \bar{\dot{q}}^\alpha, \tilde{a}^\beta(\bar{q}^\alpha, \overline{\dot{q}^\sigma}))$.

The local expressions (3.10)–(3.12) of D'Alembert's principle show that, for the deviation semibasic form $r_{T^2\chi(z)}$ [see (4.16)],

$$p_i(\bar{\ddot{q}}^\alpha)D_\sigma\chi^i(\bar{q}^\alpha) = 0 \tag{4.23}$$

holds, if and only if $\bar{\ddot{q}}^\beta = \tilde{a}^\beta(\bar{q}^\alpha, \overline{\dot{q}^\sigma})$, i.e., if and only if $z = (\bar{q}^\alpha, \bar{\dot{q}}^\alpha, \tilde{a}^\beta(\bar{q}^\alpha, \overline{\dot{q}^\sigma})) = \tilde{X}(u)$. Hence, by (4.22b) and (4.23), $dG_u(z) = 0$, if and only if $z = \tilde{X}(u)$.

To conclude that $\tilde{X}(u)$ is indeed a minimizing point for G_u, we recall that, as already noticed above, G_u is a quadratic polynomial in $\ddot{q}^\alpha$, whose leading term is

$$\bar{K}_{ih}D_\alpha\chi^i(\bar{q}^\sigma)D_\beta\chi^h(\bar{q}^\sigma)\ddot{q}^\alpha\ddot{q}^\beta, \tag{4.24}$$

which is positive-definite because of condition (b) following (4.5) and because χ is an imbedding.

Remark: Gauss' principle can be stated, in an equivalent way, directly in terms of Gauss' function G above, rather than in terms of its pull-back G. In this case, the wording turns out to be closer to the classical statements of the principle that can be found in the literature. Nevertheless, the statement itself becomes more involved and we omit the details here.

[1] K. F. Gauss, "Über ein neues allgemeines Grundgesetz der Mechanik," Crelle Journal für die reine und angewandte Mathematik **4**, 232 (1829).

[2] C. Godbillon, *Géométrie Différentielle et Mécanique Analitique* (Hermann, Paris, 1969).

[3] P. Libermann and C. M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Dordrecht, 1987).

[4] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin, London, 1978), 2nd ed.

[5] L. Lilov and M. Lorer, "Dynamic analysis of multirigid-body systems based on the Gauss Principle" Z. Angew. Math. Mech. **62**, 539 (1982).

[6] C.-C. Wang, *Mathematical Principles of Mechanics and Electromagnetism. Part A: Analytical and Continuum Mechanics* (Plenum, New York, 1979).

[7] A. M. Vershik and L. D. Faddeev, "Differential Geometry and Lagrangian Mechanics with constraints," Sov. Phys. Dokl. **17**, 34 (1972).

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

[9] H. Rund, *The Differential Geometry of Finsler Spaces* (Springer, Berlin, 1959).

[10] O. R. Ruitz, "Existence of brake orbits in Finsler mechanical systems," in *Geometry and Topology, Lecture Notes in Mathematics*, Vol. 597 (Springer, Berlin, 1977).

[11] A. C. Eringen, "Tensor Analysis," in *Continuum Physics* (Academic, New York, 1971), Vol. 1.

[12] M. Spivak, *A Comprehensive Introduction to Differential Geometry* (Publish or Perish, Berkeley, CA, 1979), Vol. 2.

Geometry of free fall and simultaneity

Holger Ewen and Heinz-Jürgen Schmidt
Fachbereich Physik, Universität Osnabrück, Postfach 4469, D-4500 Osnabrück, West Germany

(Received 14 July 1988; accepted for publication 2 February 1989)

It is shown that the differentiable, affine, and metric structure of Newton–Cartan space-time is uniquely determined by its projective–conformal–material structure. This means physically that—similarly as in general relativity—it is also possible in classical gravitation to define operations of parallel transport and measurements of length, time, and mass using only three kinds of world lines: world lines of freely falling test particles, of photons, and of gravitational matter.

I. INTRODUCTION

Physical theories that are completely mathematized can nevertheless be the object of further analysis and reformulation. One motif of such an analysis could be casting old theories in newly invented mathematical forms, as, for example, in reformulating classical mechanics in terms of symplectic manifolds or presenting Newtonian gravitation *à la* general relativity, which is known as Newton–Cartan (NC) theory.[1,2] Very often, this endeavor results in more than a mere reformulation but rather gives new insights into the theory. Another aim of revisiting a theory could be the physical analysis of its concepts and laws, concerning problems like "which concepts could be viewed as basic and which as derived ones" or "which part of a physical law is just a definition and which part is empirically restrictive" (for a detailed account of this kind of problem, see Ludwig's book[3]). Examples of this sort of approach are Giles' axiomatization of thermodynamics[4] or the paper of Ehlers, Pirani, and Schild on general relativity.[5]

This latter work shows that the affine and metric structure of space-time can be uniquely characterized in terms of its projective (geodesic) and conformal (light cone) structure and thus general relativistic space-time appears as a "geometry of free fall and light propagation." In this paper we apply the analogous approach to the nonrelativistic space-time of the NC theory. The nonrelativistic analog of the conformal structure is the simultaneity relation, which obviously has less "characterizing power" than the light cone structure (compare, for instance, the flat case). Therefore it is not surprising that the affine and metric structure *cannot* be uniquely characterized by the projective–conformal structure. More specific is the result that the remaining freedom in choosing the NC structure is given by just one real function of time. If, for example, the time metric is additionally fixed, then the remaining physical concepts of parallel transport, spatial metric, and mass density are unambiguously determined.

In order to get a unique characterization of NC geometry we extend the projective–conformal framework by what we will call a "material structure," given by the set of world lines of the material particles that act as the source of gravitation and are subject to a continuity equation. This entails a number of constraints on the mass function and yields the desired uniqueness.

So our final result intuitively reads as follows: Suppose as given a set of points (space-time) and three classes of lines: world lines of all possible freely falling test particles, a subclass of world lines with "infinite velocity," and another class of world lines of gravitational matter. Then, if there exists a NC structure compatible with these data, it is unique up to the choice of units. Thus one could decide whether a given clock or measuring rod is a correct one or not solely on the basis of world lines of particles. In order to derive this result one has, of course, to exclude some highly symmetrical models of NC theory.

The benefits of this approach lie, on the one hand, in the deeper understanding of the foundations of classical physics; on the other hand, it possibly simplifies the task of relating nonrelativistic to relativistic theory of gravitation, a task not yet completely solved. It has been pointed out by Ehlers[6] that this problem requires a common conceptual base of both theories, which is now boiled down to a system of three concepts that determine the other ones.

Our paper is organized as follows. Section II contains a short account of NC theory in a coordinate-free fashion. From the variants of this theory we choose the "strongest" by adopting the axioms of Newtonicity and simply connectedness thus avoiding global topological finesses. The essential well-known properties are concentrated in Theorem 2.3, the proof of which we include for convenience of the reader and because it provides some technical tools needed later. Section III is devoted to the projective and conformal aspects of NC theory. Especially in Theorem 3.3 we derive necessary and sufficient conditions for two NC structures to have the same set of unparametrized geodesics and the same relation of simultaneity.

To exclude the symmetric cases we define the "tidal algebra" of a NC theory, which is isomorphic to the 3×3 matrix algebra in the standard (generic) case and smaller in the exceptional (symmetric) cases. Section IV deals with the additional restrictions imposed by the material structure and the continuity equation. Theorem 4.2 contains our main uniqueness result indicated above.

An alternative way to presenting our approach would be to give operational definitions of length, time, mass, and parallel transport using only world lines of the different kinds. This will be achieved in a forthcoming paper.

We shall use the notations of differential geometry established, for instance, in the book of Kobayashi and Nomizu[7] with some minor modifications. A linear connection on a manifold is identified with the operator of covariant differentiation ∇. Only occasionally do we resort to the alter-

 0022-2488/89/071480-07$02.50

native view of a linear connection as a horizontal distribution in the frame bundle. Torsion, curvature, and Ricci tensors are defined as usual:

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y], \tag{1.1}$$

$$R(X,Y)Z = ([\nabla_X,\nabla_Y] - \nabla_{[X,Y]})Z, \tag{1.2}$$

$$\mathrm{Ric}(X,Y) = \mathrm{trace}\,(Z \mapsto R(Z,X)Y). \tag{1.3}$$

We shall make extensive use of the contraction operator $\lrcorner$ between vector fields X,Y or one-forms u,v and suitable tensor fields, and write, for example, $X \otimes Y \lrcorner g$ (or $X \wedge Y \lrcorner g$ for g antisymmetric) instead of $g(X,Y)$. The ith partial derivative of a function φ will be denoted by $\varphi_{,i}$. Greek indices run from 1 to 3, latin ones from 0 to 3. Gal will denote the Galilei group acting on $\mathbb{R}\times\mathbb{R}^3$.

II. NEWTON–CARTAN (NC) THEORY

A. General remarks

NC theory is based on the observation that a classical theory of gravitation can be formulated without fixing a class of global inertial systems. Such a class may be introduced as an auxiliary tool to solve concrete problems, but—up to special cosmological situations such as asymptotic flat space-time—this is not possible in a unique, physically founded way. This is similar to the situation in general relativity, and, as in that case, the proportionality between inertial and gravitational mass suggests a geometrization of gravity: the world lines of freely falling test particles are geodesics of a linear connection ∇ on the space-time manifold M. But unlike the relativistic case, space and time measures are not soldered into a single metric but give rise to two different geometric entities: a one-form τ measuring the proper time of world lines and a two-tensor h that induces a Euclidean metric on the null space of τ.

B. Mathematical structure of NC theory

Definition 2.1: A *NC structure* is a six-tuple $\alpha = (M;\mathscr{D},\tau,h,\nabla,\rho)$ satisfying the following conditions: (i) M is a set; (ii) $\mathscr{D}$ is a maximal C^∞ atlas on M such that $(M,\mathscr{D})$ is a four-dimensional connected and simply connected manifold; (iii) τ is a nowhere vanishing one-form on $(M,\mathscr{D})$; (iv) h is a symmetric contravariant two-tensor on $(M,\mathscr{D})$, positive semidefinite, of rank 3, such that $\tau \lrcorner h = 0$; (v) ∇ is a torsion-free, complete linear connection on $(M,\mathscr{D})$, compatible with τ and h, i.e., $\nabla\tau = 0$, $\nabla h = 0$, and its curvature tensor R satisfies

$$R(U\wedge X)U = 0 \quad (\text{"spatial flatness"}) \tag{2.1}$$

and

$$R(X\wedge U)Y \lrcorner v = R(Y\wedge V)X \lrcorner u \quad (\text{"Newtonicity"}) \tag{2.2}$$

for all vector fields U,V,X,Y and one-forms u,v such that $U = u \lrcorner h$ and $V = v \lrcorner h$; and (vi) $\rho \in C^\infty(M)$ is a non-negative function satisfying

$$\mathrm{Ric} = \rho\tau\otimes\tau \quad (\text{field equation with } 4\pi G = 1).$$

Remarks and further definitions:

(1) A vector (field) X such that $X \lrcorner \tau = 0$ will be called *spacelike*. In the case $X \lrcorner \tau = 1$ it will be called a *unit vector (field)*. Here $\mathscr{S}$ will denote the space of all spacelike vector fields and $\mathscr{S}_x$ the subspace of spacelike vectors of T_xM.

(2) If $f \in C^\infty(M)$ satisfies $\tau\wedge df = 0$ or, equivalently, $U(f) = 0$ for all $U\in\mathscr{S}$, the function $T(f)$ is the same for all unit vector fields T and will be denoted by f_t ("time derivation").

(3) The *gradient* of a function $\varphi\in C^\infty(M)$ is defined by $\mathrm{grad}\,\varphi = d\varphi \lrcorner h$ and hence a spacelike vector field. The *divergence* of a vector field X may be defined as $\mathrm{div}\,X = \mathrm{trace}(Y\mapsto\nabla_Y X)$ or, equivalently, by the equation

$$\mathfrak{L}_X\mu = (\mathrm{div}\,X)\mu,$$

where μ is a Galilean invariant volume four-form, unique up to a factor, and $\mathfrak{L}$ the Lie derivative. Thus the definition of "div" does not really presuppose a connection ∇. An analogous remark applies to the *Laplacian* $\Delta\varphi = \mathrm{div}\,\mathrm{grad}\,\varphi$.

(4) Let γ be an affinely parametrized geodesic of ∇, $\dot\gamma = U$, and the vector field X be an "infinitesimal variation" of γ (Jacobi field). Then there holds the equation of geodetic variation or Jacobi equation[8]

$$\nabla_U\nabla_U X = R(U\wedge X)U \tag{2.3}$$

and the vanishing of $R(U\wedge X)U$ for spacelike U means that there is no relative acceleration between neighboring spacelike geodesics (for this the term "spatial flatness").

It will turn out in the proof of Theorem 2.3 that the conditions in Definition 2.1 (i)–(v) already constrain the Ricci tensor being of the form $\mathrm{Ric} = \lambda\tau\otimes\tau$. So the only restriction implied by Definition 2.1 (vi) is $\lambda \geqslant 0$. We, however, adhere to ρ as a NC structural component because of its physical importance.

Definition 2.2: An *extended NC structure* is a seven-tuple $\alpha = (M;\mathscr{D},\tau,h,\nabla,\rho,S)$ such that $(M;\mathscr{D},\tau,h,\nabla,\rho)$ is a NC structure and S is a vector field on M satisfying $S \lrcorner \tau = 1$ and $\mathrm{div}(\rho S) = 0$ (continuity equation).

In the following theorem we recall the essential properties of NC structures and how to retain the usual formulation of classical gravity.

Theorem 2.3: Let α be an (extended) NC structure, then the following hold.

(1) τ is exact, i.e., $\tau = dt$ with $t\in C^\infty(M)$. The three-dimensional submanifolds $M_c = \{x\in M \mid t(x) = c\}$, $c\in\mathbb{R}$, will be called *time slices*. $M = \cup\{M_c \mid c\in\mathbb{R}\}$ defines a regular foliation of M.

(2) h induces a positive-definite metric $\tilde h$ on time slices. The Levi-Civita connection of $\tilde h$ coincides with the restriction of ∇ and is flat.

(3) There exist isomorphisms between $(M;\mathscr{D},\tau,h)$ and the standard Galilean space-time $\mathbb{R}\times\mathbb{R}^3$ (without the possibility of singling out a canonical one). These isomorphisms are also isomorphisms of globally trivial Galilean frame bundles.

(4) The parallel transport of spacelike vectors yields spacelike vectors and is path independent, so ∇ is completely parallelizable when acting on spacelike vectors.

(5) ∇ acting on unit vectors can be regarded as a potential force, i.e., there exists a flat, torsion-free connection $\mathring\nabla$, compatible with τ and h, and a C^∞ function $\Phi: M\to\mathbb{R}$ such that

$$\nabla = \mathring{\nabla} + (d\Phi \lrcorner h) \otimes \tau \otimes \tau. \tag{2.4}$$

Moreover,

$$\Delta\Phi = \rho. \tag{2.5}$$

(6) The geodesic equation in Galilean coordinates (i.e., in coordinates such that $\mathring{\Gamma}^a_{bc} = 0, \tau_a = \delta_{a,0}, h^{\alpha\beta} = \delta_{\alpha,\beta}$) reads

$$\ddot{x}^\alpha = -(\text{grad}\,\Phi)^\alpha. \tag{2.6}$$

If α is an extended NC structure, the continuity equation in these coordinates reads

$$\rho_{,0} + (\rho S^\alpha)_{,\alpha} = 0. \tag{2.7}$$

Proof:

(1) $(X \wedge Y) \lrcorner d\tau$

$= X(Y \lrcorner \tau) - Y(X \lrcorner \tau) - [X,Y] \lrcorner \tau$

$= \nabla_X(Y \lrcorner \tau) - \nabla_Y(X \lrcorner \tau) - [X,Y] \lrcorner \tau$

$= (\nabla_X Y - \nabla_Y X - [X,Y]) \lrcorner \tau$

$\quad + X \lrcorner \nabla_Y \tau + Y \lrcorner \nabla_X \tau = 0,$

since $T(X,Y) = 0$ and $\nabla\tau = 0$. So τ is closed and, because M is assumed to be simply connected, also exact: $\tau = dt$, $t \in C^\infty(M)$. The foliation property follows, for example, by Frobenius' theorem.[9] Consider a geodesic γ with tangent vector T, $T \lrcorner \tau \neq 0$. Because of $\nabla_T(T \lrcorner \tau) = 0$ and γ being complete, $t: M \to \mathbb{R}$ is surjective, and the set of time slices is a manifold diffeomorphic to $\mathbb{R}$. Thus the foliation is regular.[8]

(2) If u is a one-form, $\tau \lrcorner (u \lrcorner h) = -u \lrcorner (\tau \lrcorner h) = 0$, thus $U = u \lrcorner h$ will be a spacelike vector field. Conversely, for each $U \in \mathscr{S}$ there exists a one-form u, unique modulo τ, such that $U = u \lrcorner \tau$. So the equation

$$U \otimes V \lrcorner \tilde{h} = u \otimes v \lrcorner h \tag{2.8}$$

defines a covariant positive-definite two-tensor $\tilde{h}$ on time slices. Here and henceforward, U,V,W will denote spacelike vector fields corresponding to one-forms u,v,w in the way indicated above.

Because of $0 = \nabla_U(V \lrcorner \tau) = \nabla_U V \lrcorner \tau$, ∇ can be restricted to time slices and is easily shown to be the Levi-Civita connection of $\tilde{h}$. Polarization of (2.1) together with the general symmetry property of the curvature tensor,

$$0 = R(X \wedge Y)Z + R(Z \wedge X)Y + R(Y \wedge Z)X, \tag{2.9}$$

gives $R(U \wedge V)W = 0$ and proves the flatness of ∇ when restricted to time slices.

(4) Let X be parallel along an integral curve of Y, $\nabla_Y X = 0$. Then we have

$$\nabla_Y(X \lrcorner \tau) = (\nabla_Y X) \lrcorner \tau + X \lrcorner \nabla_Y \tau = 0$$

and the first claim follows. The path independence of parallel transport of spacelike vectors can be reduced to the equation $R(X \wedge Y)U = 0$, using the Ambrose–Singer theorem[10] and that M is simply connected. To prove this identity we note that the differential operator

$$R(X \wedge Y) = \nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]}$$

gives zero when applied to functions or to τ and h. Hence

$$R(X \wedge Y)Z \lrcorner \tau = R(X \wedge Y)(Z \lrcorner \tau) - Z \lrcorner R(X \wedge Y)\tau = 0$$

and

$$R(X \wedge Y)U \lrcorner u = R(X \wedge Y)(u \otimes u \lrcorner h) - u \otimes u \lrcorner R(X \wedge Y)\tau = 0.$$

Polarizing the last identity and using (2.1), (2.2), and (2.9) lead to $R(U \wedge X) = 0$ and further $(R(U \wedge V)X = 0$.

To evaluate $R(X \wedge Y)V$, we decompose Y in the form $Y = \lambda X + U$ and obtain

$$R(X \wedge Y)V = \lambda R(X \wedge X)V + R(X \wedge U)V = 0.$$

This also proves that $\text{Ric}(X,Y) = 0$ if X or Y are spacelike.

(3) We will construct a flat torsion-free connection $\mathring{\nabla}$, compatible with τ and h, that coincides with ∇ for the parallel transport of spacelike vectors and in spacelike directions: $\nabla_X U = \mathring{\nabla}_X U$, $\nabla_U X = \mathring{\nabla}_U X$. Recall that a flat connection is given by an involutive horizontal distribution in the Galilean frame bundle P, or, via Frobenius' theorem, by a foliation of P by horizontal submanifolds, both satisfying a Galilean covariance condition. By virtue of this covariance condition it suffices to select a single submanifold of P, i.e., a Galilean frame field E_a, in order to determine $\mathring{\nabla}$. Moreover, $\mathring{\nabla}$ is torsion-free iff $[E_a, E_b] = 0$.

The frame field E_a is uniquely determined by some nonspacelike geodesic γ, a Galilean frame at some point x on γ, and the equations above: The geodesic γ yields a tangent unit vector field E_0 along γ, which is uniquely transported into spacelike directions according to $\mathring{\nabla}_U E_0 = \nabla_U E_0$ and $R(U \wedge V) = 0$. The spacelike three-frame E_q at x is analogously transferred everywhere according to $\mathring{\nabla}_X E_\alpha = \nabla_X E_\alpha$ and $R(X \wedge Y)E_\alpha = 0$. Thus a global Galilean frame field and a corresponding flat connection $\mathring{\nabla}$ are defined. The equation $\nabla_X U = \mathring{\nabla}_X U$ is true by construction and $\nabla_U X = \mathring{\nabla}_U X$ follows from the decomposition $X = \lambda E_0 + V$. In general, $\mathring{\nabla}$ is not unique, because the construction depends on the choice of the geodesic γ. In order to show $[E_a, E_b] = 0$, we note that $[E_\alpha, E_\beta] = 0$ because $\nabla = \mathring{\nabla}$ on time slices is torsion-free and flat. Moreover, E_α is ∇ parallel by construction, hence $\nabla_{E_0} E_\alpha = 0$. Similarly, $\nabla_{E_\alpha} E_0 = 0$ and $[E_0, E_\alpha] = \nabla_{E_0} E_\alpha - \nabla_{E_\alpha} E_0 = 0$, because ∇ is torsion-free.

(5) Assertion (5) now follows from the fact[11] that a connected, simply connected manifold with a complete, flat, torsion-free connection is affinely isomorphic to $\mathbb{R}^n$. Since $\mathring{\nabla}$ is given by a Galilean frame field E_a, the isomorphism I: $M \to \mathbb{R}^4$ maps τ and h onto the standard Galilean structure of $\mathbb{R}$ and can be extended to a Galilean frame-bundle isomorphism $\tilde{I}$: $P \to \mathbb{R}^4 \times \text{Gal}$. Next we consider the difference tensor

$$D_X Y = \nabla_X Y - \mathring{\nabla}_X Y \tag{2.10}$$

It satisfies $D_U X = 0$, $D_X U = 0$, $D_X Y \lrcorner \tau = 0$, and $D_X Y = D_Y X$. Thus the only nontrivial component of D is $D_{E_0} E_0$, which is spacelike and will be shown to be of the form $D_{E_0} E_0 = d\Phi \lrcorner h$. Since $\mathring{R} = 0$, the curvature tensor R may be written as

$$R(X \wedge Y) = D_X \mathring{\nabla}_Y - \mathring{\nabla}_Y D_X + \mathring{\nabla}_X D_Y - D_Y \mathring{\nabla}_X + D_X D_Y - D_Y D_X - D_{[X,Y]}.$$

In particular, if X and Y are unit vector fields, we have

$$R(X \wedge U)Y = D_X \mathring{\nabla}_U Y - \mathring{\nabla}_U D_X Y - D_{[X,U]} Y.$$

It can be shown that the first and the third term of the rhs vanish. The first term vanishes, since $\mathring{\nabla}_U Y \in \mathscr{S}$:

$$0 = \mathring{\nabla}_U(1) = \mathring{\nabla}_U(Y \lrcorner \tau) = \mathring{\nabla}_U Y \lrcorner \tau.$$

The third term vanishes, since $[X,U] \in \mathscr{S}$:

$$0 = \nabla_U(U \lrcorner \tau) = \nabla_X U \lrcorner \tau,$$

$$0 = \nabla_U(X \lrcorner \tau) = \nabla_U X \lrcorner \tau,$$

$$[X,U] \lrcorner \tau = (\nabla_X U - \nabla_U X) \lrcorner \tau = 0.$$

Hence

$$R(X,U)Y = -\mathring{\nabla}_U D_X Y. \tag{2.11}$$

We set $W = D_{E_0} E_0$. Because W is spacelike, it may be written as $W = w \lrcorner h$, where w is a one-form unique modulo τ. By using Newtonicity (2.2) we will show that the "curl" of W vanishes, i.e., $(U \wedge V) \lrcorner dw = 0$, or, equivalently, $dw \wedge \tau = 0$. Let $U = u \lrcorner h$, $V = v \lrcorner h$, then it follows that

$$(U \wedge V) \lrcorner dw = \mathring{\nabla}_U(V \lrcorner w) - \mathring{\nabla}_V(U \lrcorner w) - [U,V] \lrcorner w$$
$$= R(E_0 \wedge V)E_0 \lrcorner u - R(E_0 \wedge U)E_0 \lrcorner v = 0. \tag{2.12}$$

Hence w is closed modulo τ, and, by a generalization[12] of Poincaré's lemma, also exact modulo τ, i.e.,

$$w \lrcorner \tau = d\Phi \wedge \tau. \tag{2.13}$$

This holds even globally, since M is contractible.[10] We conclude $D_{E_0} E_0 = d\Phi \lrcorner h$ and $\nabla = \mathring{\nabla} + (d\Phi \lrcorner h) \otimes \tau \otimes \tau$. Further,

$$\mathrm{Ric}(E_0,E_0) = \mathrm{trace}\, R(\cdot,E_0)E_0 = \mathrm{trace}\, \mathring{\nabla}_{(D_{E_0}E_0)}$$
$$= \mathrm{trace}\, \nabla_{(d\Phi \lrcorner h)} = \Delta\Phi.$$

Hence $\Delta\Phi = \rho$. We note that, although Φ depends on the choice of $\mathring{\nabla}$, $\Delta\Phi$ does not.

(6 + 7) These identities follow immediately by noting that in Galilean coordinates we have $\Gamma^{\alpha}_{00} = h^{\alpha\beta}\Phi_{,\beta} = (\mathrm{grad}\,\Phi)^{\alpha}$ and all other components of Γ vanish. ■

III. PROJECTIVE–CONFORMAL EQUIVALENCE

As indicated in the Introduction we will perform an analysis of classical gravitation theory analogous to that of Ehlers, Pirani, and Schild[5] for general relativity. This means that we regard the projective–conformal structure of spacetime as basic and seek to derive the metrical concepts and the concept of mass from this structure. The projective structure Π will be given by the world lines of all freely falling test particles without imposing an affine parameter on these world lines. The conformal structure $\mathscr{C}$ is represented by the world lines of photons with infinite velocity ("$c \to \infty$") or, equivalently, by the nonrelativistic concept of simultaneity.

In the general relativistic case, one has two classes of linear connections compatible with Π (resp. $\mathscr{C}$) and a single connection in the intersection of the two classes. In our case, to the contrary, all relevant components of the NC theory, τ, h, ∇, ρ, are not completely fixed by the derived projective–conformal structures. The remaining freedom consists of the choice of a gravitational clock, and hence of an arbitrary C^{∞} function a: $\mathbb{R} \to \mathbb{R}$ (without zeros). Independent of a is the Euclidean structure of time slices and parallel transport of spacelike directions, which was shown to be path independent. Interestingly enough, the Ricci tensor, and hence the mass density, also is unique up to a time-dependent affine transformation.

Definition 3.1: Let $\alpha = (M;\mathscr{D},\tau,h,\nabla,\rho)$ be a NC structure. By $\Pi(\alpha)$ we will denote the set of (unparametrized) ∇ geodesics of $(M,\mathscr{D})$. Thus $\Pi(\alpha)$ is a family of subsets of M and will be called the *projective structure* of α. The subset of $\Pi(\alpha)$ of spacelike geodesics will be denoted by $\mathscr{C}(\alpha)$ and called the *conformal structure* of α. Two NC structures α, α' over the same set $M = M'$ are called projectively conformal equivalent, or, for short, $\Pi\mathscr{C}$ *equivalent*, if $\Pi(\alpha) = \Pi(\alpha')$ and $\mathscr{C}(\alpha) = \mathscr{C}(\alpha')$.

Before exploring the concept of $\Pi\mathscr{C}$ equivalence we will consider a classification of NC structures which will turn out to be relevant for this purpose. Let T be a fixed unit vector field and consider the map

$$G\colon \mathscr{S} \to \mathscr{S}, \quad U \mapsto G(U) = R(T \wedge U)T. \tag{3.1}$$

It is independent of T by the properties of R considered in the last chapter.

Let $\mathscr{S}\mathscr{P}$ denote the space of parallel spacelike vector fields, and $ev_x\colon \mathscr{S}\mathscr{P} \to \mathscr{S}_x$ the evaluation map $ev_x(U) = U(x)$, which is a linear isomorphism. Thus if $U \in \mathscr{S}\mathscr{P}$, then

$$G_x(U)(x) = (G(U))(x) \tag{3.2}$$

defines a field of linear maps $G_x\colon \mathscr{S}\mathscr{P} \mapsto \mathscr{S}\mathscr{P}$, $x \in M$. Another representation is the field $g_y\colon \mathscr{S}_x \to \mathscr{S}_x$, where

$$g_y = ev_x \circ G_y \circ ev_x^{-1}, \quad x,y \in M.$$

If $U,V \in \mathscr{S}$, then $\langle U,V \rangle := U \otimes V \lrcorner \tilde{h} \in C^{\infty}(M)$. Especially, for $U,V \in \mathscr{S}\mathscr{P}$, $\langle U,V \rangle$ will be a constant function and $(\mathscr{S}\mathscr{P}, \langle\ ,\ \rangle)$ may be considered as a Euclidean vector space. If T is tangent to a family of geodesics, $G(U) = R(T \wedge U)T$ occurs on the right-hand side of the Jacobi equation (2.3) and thus can be interpreted as the linearized acceleration field into the direction U. Therefore we will call the topologically closed $\mathbb{C}$-linear algebra generated by the operators $G_x \in \mathrm{Lin}(\mathscr{S}\mathscr{P} \to \mathscr{S}\mathscr{P})$ the *tidal algebra* $\mathscr{G}$ of the NC structure α. The dimension of $\mathscr{G}$ as a $\mathbb{C}$-linear space can serve as a crude classification of NC structures and will be called the *type* of α. Flatness of ∇ implies $\mathrm{type}(\alpha) = 0$, but any physically interesting NC structure will have a lot of different tidal forces and thus $\mathrm{type}(\alpha) = 9$. Therefore we will refer to the latter as the *standard case*, and to $\mathrm{type}(\alpha) < 9$ as the *exceptional cases*.

Lemma 3.2:
(1) $\langle G(U),V \rangle = \langle U,G(V) \rangle$;
(2) the operators $G_y\colon \mathscr{S}\mathscr{P} \to \mathscr{S}\mathscr{P}$ and $g_y\colon \mathscr{S}_x \to \mathscr{S}_x$ are symmetric for all $x,y \in M$;
(3) $\mathrm{type}(\alpha) \in \{0,1,2,3,4,5,9\}$.

Proof:
(1) $\langle G(U),V \rangle$
$= R(T \wedge U)T \otimes V \lrcorner \tilde{h} = R(T \wedge U)T \lrcorner v$
$= R(T \wedge V)T \lrcorner u$
$= \langle G(V),U \rangle$ by Newtonicity (2.2).

(2) By evaluation of (1) for $U,V\in\mathscr{S}\mathscr{P}$ at some point $y\in M$.

(3) $\mathscr{G}$ is a von Neumann algebra and hence generated by its projections, if not $\mathscr{G}=\{0\}$. The lattice of projections is either Boolean with 2^n elements ($n=1,2,3$) and n linearly independent projections [type$(\alpha)=n$], or it is irreducible and isomorphic to the lattice of a two- or three-dimensional Hilbert space [type$(\alpha)=4$ or 9], or it is the direct sum of those lattices, which gives the additional possibility type$(\alpha)=4+1$. ■

Theorem 3.3: In the standard case two NC structures $\alpha=(M;\mathscr{D},\tau,h,\nabla,\rho)$ and $\alpha'=(M;\mathscr{D}',\tau',h',\nabla',\rho')$ are $\Pi\mathscr{C}$ equivalent iff there exists a function $a\in C^\infty(M)$, constant on time slices and nowhere zero, such that

(1) $\mathscr{D}'=\mathscr{D}$;

(2) $\tau'=a\tau$;

(3) $h'=\lambda a^{-1}h$, $\lambda>0$ constant on M;

(4) $\nabla'_X Y=\nabla_X Y+b\{(Y\lrcorner\tau)X+(X\lrcorner\tau)Y\}$,

where $b=\frac{1}{2}a^{-1}a_t$;

(5) $\rho'=a^{-2}\rho+3a^{-2}(b^2-b_t)$.

Remarks: (1) In the exceptional case, only condition (3) is modified such that $U\otimes V\lrcorner\tilde{h}'=(\lambda U)\otimes V\lrcorner\tilde{h}$ holds for $U,V\in\mathscr{S}\mathscr{P}$, where λ is a positive definite operator commuting with $\mathscr{G}$.

(2) The proof will also show that $\mathscr{G}^C$, the commutant of $\mathscr{G}$, is a projective–conformant invariant. So the above distinction standard case/exceptional case is meaningful without reference to α or α'.

Proof: (1) We will sketch the construction of a chart $q\in\mathscr{D}\cap\mathscr{D}'$ around any point $x\in M$ using only projective–conformal means. Here α and α' have the same time slices (x and y are simultaneous iff they can be joined by a geodesic $\gamma\in\mathscr{C}$). Each time slice together with its geodesics satisfies the axioms of synthetic affine geometry, and can be endowed with affine coordinates.[13] These three coordinates are C^∞ functions (w.r.t. $\mathscr{D}$ and $\mathscr{D}'$) on M, if the corresponding affine frames are, for example, generated by four geodesics, which do not lie in a plane of a time slice (at least locally). The fourth coordinate q^0 could be the arclength of the curve in $\mathbb{R}^3$ which is given by the q^α coordinates of another suitable geodesic (a "clock").

(2) $\tau'=a\tau$, $a\in C^\infty(M)$, $a(x)\neq 0$, because τ and τ' have the same kernel. Moreover, $0=d\tau'=da\wedge\tau+ad\tau=da\wedge\tau$, so a is constant on time slices, $a=a(t)$.

(3) It is well known[14] that projective equivalence of two torsion-free connections ∇',∇ is equivalent to

$$\nabla'_X Y=\nabla_X Y+(X\lrcorner\omega)Y+(Y\lrcorner\omega)X, \tag{3.3}$$

where ω is a one-form on M [the projective geodesic equations $T\wedge\nabla'_T T=0$ and $T\wedge\nabla_T T=0$ are equivalent iff the difference tensor is of the form

$$\nabla'_T T-\nabla_T T=T(T\lrcorner\Psi), \tag{3.4}$$

polarization and symmetry of ∇,∇' give the result above]. Now consider the identity

$$\begin{aligned}\nabla'_X(Z\lrcorner\tau')&=\nabla'_X Z\lrcorner\tau'\\&=\nabla_X Z\lrcorner\tau'+(Z\lrcorner\omega)(X\lrcorner\tau')\\&\quad+(X\lrcorner\omega)(Z\lrcorner\tau').\end{aligned}$$

If $Z\lrcorner\tau=1$ and $X\lrcorner\tau=X\lrcorner\tau'=0$, all terms vanish; hence $X\lrcorner\omega=0$ and $\omega=b\tau$, $b\in C^\infty(M)$. If in the above equation $Z\lrcorner\tau=X\lrcorner\tau=1$, we have $\nabla_X Z\lrcorner\tau'=0$ and the remaining terms read $a_t=2ab$, whence the claim follows.

(4) If $\nabla_X U=0$, U spacelike, then $\nabla'_X(a^{-1/2}U)=0$:

$$\begin{aligned}\nabla'_X(a^{-1/2}U)&=\nabla_X(a^{-1/2}U)+(X\lrcorner\omega)a^{-1/2}U\\&=-\tfrac{1}{2}a^{-3/2}(X\lrcorner da)U+a^{-1/2}\nabla_X U\\&\quad+a^{-1/2}b(X\lrcorner\tau)U\\&=a^{-1/2}(-\tfrac{1}{2}a^{-1}a_t(X\lrcorner\tau)U\\&\quad+b(X\lrcorner\tau)U)=0.\end{aligned}$$

Hence $U\mapsto U'=a^{-1/2}U$ maps parallel spacelike vector fields of α onto those of α'. The analogous transformation for unit vector fields is $T\mapsto T'=a^{-1}T$. Therefore it is enough to check the identity $h'=\lambda a^{-1}h$ at one point $x\in M$. Since $\mathscr{S}_x$ and $\mathscr{S}'_x$ are isomorphic as Euclidean spaces with isomorphism $q_x:\mathscr{S}_x\to\mathscr{S}'_x$, we have

$$\langle U(x),V(x)\rangle'=\langle q_x U(x),q_x V(x)\rangle.$$

Next we will consider the relation between G and G'. The two curvature tensors are related by

$$\begin{aligned}R'(X\wedge Y)Z&=R(X\wedge Y)Z+(Z\lrcorner\nabla_X\omega)Y\\&\quad+(Y\lrcorner\nabla_X\omega)Z\\&\quad-(Z\lrcorner\nabla_Y\omega)-(X\lrcorner\nabla_Y\omega)Z\\&\quad+(Z\lrcorner\omega)(Y\lrcorner\omega)X-(Z\lrcorner\omega)(X\lrcorner\omega)Y.\end{aligned}$$

Setting $Z=X=T'$, $Y=U'$ and regarding $\nabla_{U'}\omega=0$, $U'\lrcorner\omega=0$, $U'\lrcorner\nabla_T\omega=0$, and the above results we obtain

$$\begin{aligned}G'(U')&=R'(T'\wedge U')T'\\&=R(T'\wedge U')T'+(T'\wedge\nabla_{T'}\omega)U'\\&\quad-(T'\lrcorner\omega)(T'\lrcorner\omega)U'\\&=a^{-2}(G(U')+(b_t-b^2)U');\end{aligned} \tag{3.5}$$

hence $\mathscr{G}^C\cong\mathscr{G}'^C$, or even $\mathscr{G}^C=\mathscr{G}'^C$ if $\mathscr{G}$ and $\mathscr{G}'$ are both represented in $\mathrm{Lin}(\mathscr{S}_x\to\mathscr{S}_x)$. Moreover, (3.5) shows that g_y enjoys the symmetry property of Lemma 3.2 also with respect to the inner product $\langle\ ,\ \rangle'$ in $\mathscr{S}_x$. For $U,V\in\mathscr{S}_x$ and $x,y\in M$ we conclude from this and Lemma 3.2,

$$\begin{aligned}\langle g_y U,V\rangle&=\langle U,g_y V\rangle=\langle q_x\,g_y U,q_x V\rangle'\\&=\langle q_x U,q_x\,g_y V\rangle'=\langle g_y U,q_x^T q_x V\rangle'\\&=\langle U,q_x^T q_x g_y V\rangle'=\langle U,g_y q_x^T q_x V\rangle'.\end{aligned}$$

This gives $[g_y,q_x^T q_x]=0$, i.e., $q_x^T q_x$ is an operator in $\mathscr{G}^C$. In the case type$(\alpha)=9$ this could only be a constant factor times the identity

$$q_x^T q_x=c\mathbb{1},$$

which completes the proof of 3.

(5) Similarly as above we calculate

$$R'(U\wedge Y)Z = R(U\wedge Y)Z - (Z\lrcorner\nabla_Y\omega)U + (Z\lrcorner\omega)(Y\lrcorner\omega)U$$
$$= R(U\wedge Y)Z + (b^2 - b_t) \times (Z\lrcorner\tau)(Y\lrcorner\tau)U,$$
$$\mathrm{Ric}'(Y,Z) = \mathrm{Ric}(Y,Z) + 3(b^2 - b_t)(Z\lrcorner\tau)(Y\lrcorner\tau),$$

and, by the two field equations,

$$\rho' = a^{-2}\rho + 3a^{-2}(b^{-2} - b_t).$$

The reverse statement of the iff claim can be verified by direct calculation. ■

Proposition 3.4: In the exponential case there exists $U\in\mathcal{S}\mathcal{P}$ such that for all $V\in\mathcal{S}$ with $\langle U,V\rangle = 0$ we have $U(V(\rho)) = 0$. Or, equivalently, there exist Galilean coordinates x^α such that

$$\rho(x^\alpha) = \rho_1(x^0,x^1) + \rho_2(x^0,x^2,x^3).$$

Proof: Let us assume for the moment that the eigenvectors U_α and eigenvalues ϵ_α of the symmetric operator G_x can be chosen to depend smoothly on x, such that we obtain the representation

$$G = \sum_\alpha \epsilon_\alpha U_\alpha \otimes u_\alpha, \quad \text{with} \quad \langle U_\alpha,U_\beta\rangle = \delta_{\alpha\beta}. \tag{3.6}$$

Recall from Sec. II the identities

$$W := D_{E_0}E_0,$$
$$\nabla_U W = \mathring{\nabla}_U D_{E_0}E_0 = -R(E_0\wedge U)E_0 = G(U) \quad \text{by (2.13) and (3.1),}$$

and further

$$0 = R(U\wedge V)W = (\nabla_U\nabla_V - \nabla_V\nabla_U - \nabla_{[U,V]})W$$
$$= \nabla_U G(V) - \nabla_V G(U) - G([U,V]).$$

Setting $U = U_\alpha$, $V = U_\beta$ and using the above representation we obtain

$$0 = U_\alpha(\epsilon_\beta)U_\beta + \epsilon_\beta\nabla_{U_\alpha}U_\beta - U_\beta(\epsilon_\alpha)U_\alpha - \epsilon_\alpha\nabla_{U_\beta}U_\alpha + \sum_\gamma \epsilon_\gamma U_\gamma\langle\nabla_{U_\beta}U_\alpha - \nabla_{U_\alpha}U_\beta, U_\gamma\rangle.$$

In the exceptional case the commutant of $\mathcal{G}$ strictly encompasses $\{\lambda\mathbb{1}\,|\,\lambda\in\mathbb{R}\}$; hence there is a common eigenvector of all G_x, say $U_1\in\mathcal{S}\mathcal{P}$. Thus all $\nabla_X U_1$ terms vanish and, taking into account

$$0 = \nabla_X\langle U_\alpha,U_\beta\rangle = \langle\nabla_X U_\alpha,U_\beta\rangle + \langle U_\alpha,\nabla_X U_\beta\rangle,$$

we conclude

$$U_2(\epsilon_1) = U_3(\epsilon_1) = U_1(\epsilon_2) = U_1(\epsilon_3) = 0.$$

Finally

$$\rho = \mathrm{Ric}(E_0,E_0) = \mathrm{trace}\,G = \epsilon_1 + (\epsilon_2 + \epsilon_3) =: \rho_1 + \rho_2.$$

We may choose $U = U_1$ and express V as a linear combination of U_2 and U_3 in order to prove the assertion.

Now we will consider the case in which not all eigenvalues and eigenvectors of G_x can be chosen to depend smoothly on x, which occurs if G_x is degenerated. Again, let $U_1\in\mathcal{S}\mathcal{P}$ be a common eigenvector of all G_x and (U_1,V_2,V_3) an orthonormal basis in $\mathcal{S}\mathcal{P}$. Degeneration of G_x occurs iff both C^∞ functions $v_2\otimes V_3\lrcorner G$ and $(v_2\otimes V_2 - v_3\otimes V_3)\lrcorner G$ vanish; hence for x in a closed subset $\mathcal{K}\subset M$. Let $\mathcal{O}_1 = M\setminus\mathcal{K}$ and $\mathcal{O}_2 = \mathring{\mathcal{K}}$, the open set of interior points of $\mathcal{K}$; then $M = \overline{\mathcal{O}_1\cup\mathcal{O}_2}$. Within $\mathcal{O}_1$ and $\mathcal{O}_2$ the eigenvectors and eigenvalues of G_x can be chosen smoothly and the above result $U(V(\rho)) = 0$ holds in $\mathcal{O}_1\cup\mathcal{O}_2$; hence by continuity also in $\overline{\mathcal{O}_1\cup\mathcal{O}_2} = M$. ■

The last proposition again shows that the exceptional case corresponds to a highly symmetrical mass density and thence to a class of unphysical situations.

IV. $\Pi\mathcal{C}\mathcal{M}$ EQUIVALENCE

At first sight, the result of the last section seems to contradict the existence of celestial "clocks" and "measuring rods." Consider, for instance, a celestial body revolved by a satellite at such a small distance that the influence of other bodies could be neglected. Then the satellite will move periodically on Kepler ellipses with constant size and the whole system might serve as a clock and a measuring rod simultaneously. What would this system look like for another astronomer who uses a different time scale, and accordingly, different measures of length and mass? Let us consider, for the sake of simplicity, the new time scale $t' = \exp(t)$, where t is the old one. A short computation then gives the transformation for length and mass density (cf. Theorem 3.3):

$$L' = L\exp(t/2), \tag{4.1}$$

$$\rho' = (\rho + \tfrac{3}{4})\exp(-2t). \tag{4.2}$$

This proves that the second astronomer will describe the system differently: An expanding body, which permanently loses mass, is embedded into a uniform background density, equally fading away, and the satellite consequently spirals out with exponentially growing periods.

On the grounds of projective–conformal geometry, no distinction can be drawn between the two interpretations of time, length, and mass, and the apparent contradiction disappears.

Nevertheless, this example shows that we need additional physical information about the mass distribution in order to exclude exotic interpretations if we want to define gravitational clocks and "rigid" rods.

From several possibilities we choose as an additional basic structure the "material structure" $\mathcal{M}$, given as the set of world lines of gravitational matter. We do not require $\mathcal{M}\subset\Pi$, thus allowing for other forces such as pressure, electromagnetic fields, etc. acting on matter. This has also the advantage that we need not bother about singularities of ρ and caustics of S, which would be produced by pure gravitation.

Definition 4.1: Let $\alpha = (M;\mathcal{D},\tau,h,\nabla,\rho,S)$ be an extended NC structure. Then we will denote by $\mathcal{M}(\alpha)$ the set of maximal integral curves (viewed as subsets of M) of S. Two extended NC structures α and α' over the same set

$M = M'$ are called $\Pi\mathscr{C}\mathscr{M}$ *equivalent* if $\Pi(\alpha) = \Pi(\alpha')$, $\mathscr{C}(\alpha) = \mathscr{C}(\alpha')$, and $\mathscr{M}(\alpha) = \mathscr{M}(\alpha')$.

Theorem 4.2: In the standard case two extended NC structures α,α' are $\Pi\mathscr{C}\mathscr{M}$ equivalent iff there are constants $a \neq 0$, $\lambda > 0$ such that

$$\mathscr{D}' = \mathscr{D},\ \tau' = a\tau,\ h' = \lambda h,$$
$$\nabla' = \nabla,\quad \rho' = a^{-2}\rho,\quad S' = a^{-1}S.$$

Proof: The corresponding NC structures are $\Pi\mathscr{C}$ equivalent and, by Theorem 3.3, related by $a \in C^\infty(M)$. We only have to show that a is constant. We will calculate the continuity equation of α'. From $\tau' = a\tau$ and $\tilde{h}' = \lambda^{-1}a\tilde{h}$ we conclude $\mu' = \lambda^{-3/2}a^{5/2}\mu$ for the volume forms, where the constant $\lambda^{-3/2}$ may be skipped. Hence

$$(\mathrm{div}'\, X)\mu' = \mathfrak{L}_X\mu' = \mathfrak{L}_X(a^{3/2})\mu + a^{3/2}\mathfrak{L}_X$$
$$= (\tfrac{5}{2}a^{3/2}a_t(X \lrcorner \tau) + a^{5/2}\,\mathrm{div}\, X)\mu$$

or

$$\mathrm{div}'\, X = \mathrm{div}\, X + 5b(X \lrcorner \tau). \tag{4.3}$$

Together with $S' = a^{-1}S$ (because $S \lrcorner \tau = S' \lrcorner \tau' = 1$) and

$$\rho' = a^{-2}(\rho + 3(b^2 - b_t)) = a^{-2}\rho + a^{-3/2}f, \tag{4.4}$$

where $f := 3a^{-1/2}(b^2 - b_t)$, we obtain

$$0 = \mathrm{div}'(\rho' S')\mu = \mathfrak{L}_{\rho' S'}\mu + 5b\rho'(S' \lrcorner \tau)\mu.$$

The first term gives

$$\mathfrak{L}_{\rho' S'}\mu = a^{-3}\mathfrak{L}_{\rho S}\mu + d(a^{-3}) \wedge (\rho S \lrcorner \mu)$$
$$+ a^{-3/2}f\mathfrak{L}_S\mu + d(a^{-5/2}f) \wedge (S \lrcorner \mu).$$

Using $\mathfrak{L}_{\rho S}\mu = 0$ and

$$dg \wedge (S \lrcorner \mu) = g_t(\tau \wedge (S \lrcorner \mu))$$
$$= g_t(S \lrcorner (\tau \wedge \mu) + (S \lrcorner \tau)\mu)$$
$$= g_t\mu,$$

we get

$$\mathfrak{L}_{\rho' S'}\mu = -6a^{-3}b\rho\mu + a^{-5/2}f\mathfrak{L}_S\mu$$
$$+ (-5a^{-5/2}bf + a^{-5/2}f_t)\mu.$$

The second term gives

$$5b\rho'(S' \lrcorner \tau)\mu = 5a^{-3}b\rho\mu + 5a^{-5/2}bf\mu,$$

and after some simplification we obtain

$$0 = -a^{-1/2}b\rho + f\,\mathrm{div}\, S + f_t. \tag{4.5}$$

We will show that this equation can only be satisfied in the trivial sense $f = b \equiv 0$, which implies $a = \text{const}$. To this end consider an integral curve of S passing through $m \in M_t$, and write ρ as a function of m and t: $\rho = \rho(m,t)$. Let $\dot{\rho} := \partial\rho/\partial t$, then the continuity equation reads: $\dot{\rho} = -\rho\,\mathrm{div}\, S$ and Eq. (4.5) becomes

$$0 = a^{-1/2}b\rho^2 + f\dot{\rho} - \rho\dot{f}. \tag{4.6}$$

If $\rho(m,t_0) = 0$, $\rho(m,t) \equiv 0$ is the trivial solution of this differential equation. If $\rho(m,t) \neq 0$ and $f(t) \neq 0$, the substitution $y = \rho f^{-1}$ leads to $y^{-2}\,dy = d(a^{-1/2})$ and hence to the local solutions

$$\rho(m,t) = f(t)(C(m) - a^{-1/2}(t))^{-1},\quad f(t) \neq 0. \tag{4.7}$$

What happens, if $f(t_0) = 0$, but $f(t_0 + \epsilon) \neq 0$ for small $\epsilon > 0$? We define the decomposition $M_{t_0} = \mathscr{E} \cup \mathscr{U}$ by $m \in \mathscr{E}$ iff $C(m) = a^{-1/2}(t_0)$, otherwise $m \in \mathscr{U}$. Here $C(m)$ is the integration constant of the solution (4.7) in the interval $(t_0, t_0 + \epsilon)$. For $m \in \mathscr{U}$ we have $\lim_{t \to t_0} \rho(m,t) = 0$ for all t. On the other hand, C is constant on $\mathscr{E}$ and therefore ρ is constant on local time slices $\mathscr{E} \times \{t\}$, $t_0 < t < t_0 + \epsilon$. Thus ρ vanishes on M_t or is constant on M_t, if $\mathscr{U} = \emptyset$, and we are in the exceptional case. So $f(t)$ is either constantly 0 (and the proof is finished) or $f > 0$ or $f < 0$ and the solution (4.7) is globally valid for some $m \in M_{t_0}$. Let $f > 0$, then $a^{-1/2}(t) < C(m)$ since $\rho(m,t) > 0$. But $a^{-1/2}$ is strictly convex by virtue of $(a^{-1/2})_{tt} = a^{-1/2}(b^2 - b_t) = f/3 > 0$ and defined for all $t \in \mathbb{R}$, which yields a contradiction. The case $f < 0$ is analogous, since then $a^{-1/2}$ is strictly concave and bounded from below. ■

[1] E. Cartan, Ann. Sci. Ec. Norm. Sup. **40**, 325 (1923); **41**, 1 (1924).
[2] H. P. Künzle, Ann. Inst. H. Poincaré **17**, 337 (1972).
[3] G. Ludwig, *Die Grundstrukturen einer physikalischen Theorie* (Springer, Berlin, 1978).
[4] R. Giles, *Mathematical Foundations of Thermodynamics* (Macmillan, New York, 1964).
[5] J. Ehlers, F. A. E. Pirani, and A. Schild, "The geometry of free fall and light propagation," in *General Relativity, Papers in Honour of L. Synge* (Oxford U.P., Oxford, 1972).
[6] J. Ehlers, "On limit relations between, and approximative explanations of, physical theories," in *Proceedings of the 7th International Congress of Logic, Methodology and Philosophy of Science, Studies in Logic and the Foundations of Mathematics,* Vol. 114, edited by P. Weingartner (Elsevier, New York, 1986), 2 volumes.
[7] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963).
[8] See Ref. 7, VIII, Theorem 1.2.
[9] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications* (Addison-Wesley, London, 1983), cf. Theorem 4.4.7.
[10] Y. Choquet-Bruhat and C. Dewitt-Morette, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1982), p. 389.
[11] See Ref. 7, V, Theorem 4.2
[12] P. Libermann and C.-M. Marle, *Symplectic Geometry and Analytical Mechanics* (Reidel, Dordrecht, 1987), cf. Proposition 7.4 (2).
[13] E. Artin, "Coordinates in Affine Geometry," Notre Dame Math. Coll. **1940**, 15 [reprinted in *Emil Artin, Collected Papers* (Springer, Berlin, 1965)].
[14] H. Weyl, "Zur Infinitesimalgeometrie: Einordnung der projektiven und der konformen Auffassung," Nachr. Ges. Wiss. (Goettingen **1921**, 99 [reprinted in *Selecta Hermann Weyl* (Birkhäuser, Basel, 1956)].

A systematic approach to the soliton equations of a discrete eigenvalue problem

Randolph James Schilling
Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana 70803-4913

(Received 28 December 1988; accepted for publication 8 March 1989)

The Ablowitz–Ladik (AL) problem is a linear vector difference equation whose isospectral flow equations include several important soliton equations; e.g., the discrete nonlinear Schrödinger equation: $i\dot{q}_n = q_{n-1} - 2q_n + q_{n+1} + |q_n|^2(q_{n-1} + q_{n+1})$. There is an established procedure for describing the soliton hierarchy of the more familiar AKNS (Ablowitz, Kaup, Newell, and Segur) problem. It is based on the notion of a *generator* for the hierarchy. In this paper the soliton equations of the AL hierarchy are described and characterized by a generator pair. A new continuous spectral problem is introduced and the AKNS hierarchy is embedded in its hierarchy as a specialization.

I. INTRODUCTION

We shall begin this paper using a familiar example—the AKNS eigenvalue problem:

$$\psi_x = (\epsilon\zeta + P)\psi, \text{ where } P = \begin{pmatrix} 0 & q(x) \\ r(x) & 0 \end{pmatrix}, \tag{1.1}$$

$\psi \in \mathbb{C}^2$, $\epsilon = \operatorname{diag}(\epsilon_1, \epsilon_2)$ is constant (historically $\epsilon_1 = i$ and $\epsilon_2 = -i$), ζ is the eigenvalue parameter, and $q(x)$ and $r(x)$ are any C^∞ functions. Flaschka, Newell, and Ratiu[1] (FNR) and Wilson[2] used the idea of a generator to describe the hierarchy of soliton equations for (1.1). Their approach may be described in the following way. We adjoin to (1.1) another linear equation,

$$\psi_t = B\psi \text{ subject to } \psi_{xt} = \psi_{tx}$$
$$\text{or } B_x - P_t = [\epsilon\zeta + P, B]. \tag{1.2}$$

Then there exist a family of matrices $Q^{(j)}, j \in \mathbb{N}$, with $Q^{(0)} = \epsilon$ and $Q^{(1)} = P$, and satisfying the following property. Let

$$Q = Q^{(0)} + Q^{(1)}\zeta^{-1} + \cdots \text{ and } Q_n = \pi_+ \cdot \zeta^n Q, \tag{1.3}$$

where π_+ denoted the polynomial part of its argument. Then

$$Q_x = [\epsilon\zeta + P, Q], \tag{1.4}$$

and if B is any solution to (1.2) that depends polynomially on ζ, then there exist constants C_n such that

$$B = \sum_n C_n Q_n. \tag{1.5}$$

With this result in mind, we shall refer to Q as the *generator* of the AKNS hierarchy. Intrigued by Q and inspired by Adler and van Moerbeke,[3] FNR found two Kirillov–Poisson brackets and rederived the conservation laws and τ functions of the AKNS hierarchy from Lie algebraic considerations. The generator provides us with the most efficient way of computing soliton equations and conservation laws. Thus the generator is an important computational and theoretical tool.

The paper[4] which inspired this work dealt with a vector difference eigenvalue problem of the form

$$p(n)f(n+1) = (E_z + q(n))f(n) + r(n)f(n-1), \tag{1.6}$$

where $p(n)$, B $q(n)$, and $r(n)$ are 2×2 matrices, $E_z = \operatorname{diag}(z_1, z^{-1})$, and z is the eigenvalue parameter. We shall refer to (1.6) as the Ablowitz–Ladik (AL) problem. The authors discovered soliton equations for (1.6) that are important in certain areas of applied research. However, their method for finding soliton equations is algebraically complicated. They do not make a clear statement regarding the existence of higher soliton equations. The problem of characterizing all the equations of the hierarcy is not addressed. This paper presents a complete description of the AL hierarchy through the construction of a generator pair for the hierarchy. The algebra leading to the generators is quite complicated but, once it is understood, one may easily rederive the important soliton equations of the AL problem. This is accomplished in Sec. III.

In Sec. II we use an interesting modification of a method due to Krichever[5] to obtain a linear differential equation like (1.1), except that the eigenvalue and its inverse appear; the matrix E_z replaces $\epsilon\zeta$. The differential problem, being algebraically simpler than the difference problem yet more complicated than the AKNS problem, shall serve to illustrate our method. The soliton hierarchy of the general problem consists of a doubly-infinite sequence of compatible equations of the form

$$\psi_{t_m} = B_m\psi, \quad m \in \mathbb{Z}.$$

Under a certain specialization of the general problem, the soliton hierarchy reduces to a semi-infinite series of equations that may be identified with the AKNS hierarchy itself.

Many authors have struggled with the problem of showing that the entries of the AKNS generator Q in (1.4) belong to the differential algebra β generated by the entries of P with respect to the derivation $\partial = \partial/\partial x$. This is a nontrivial fact because an entry-wise analysis of (1.4) would suggest that the multiple x integrals of the entries of P must also be considered. For instance, Wilson[2] appeals to a clever but indirect argument.[6] The problem introduced in Sec. II contains the eigenvalue parameter z and its inverse. The term eigenvalue is used somewhat loosely here in that z is not in general the eigenvalue of a differential operator. Thus we cannot appeal to Wilson's argument. However, the result has a direct proof. We observed that it is an immediate consequence of these two facts: (i) $\det(Q)$ is constant in x and (ii) $Q^{(0)}$ is a constant diagonal matrix. The details are in the proof

 0022-2488/89/071487-15$02.50

of Theorem 2.3. This observation may be traced back to Ref. 7. The notion of generator for a soliton hierarchy has appeared in many disguises.[1,2,7–9]

This paper is written in the spirit of papers like Refs. 1, 2, 6–12 in that it is concerned primarily with algebraic aspects of soliton equations. It is an essential preliminary step in a much broader study. We would like to explain our research plans in terms of a few related papers.

We are primarily concerned with the periodic inverse spectral transform which is based on a difference analog of Floquet theory.[13] The basis ideas as they apply to the Toda lattice can be found in Refs. 14–18. The theory leads naturally to algebraic curves and Jacobian varieties as in Refs. 3, 19–21. Strictly speaking, we are not lead to tridiagonal or banded matrices as in Refs. 19–22 because we are dealing with a vector (not a scalar) difference equation. However, one may extract a rather intriguing generalization in the following way. We let f and $\mathbb{E}_z$ be the $2N$ vector and the $2N\times 2N$ diagonal matrix given by

$$f=(f_1(1),f_2(1),\ldots,f_1(N),f_2(N))^T$$

and

$$\mathbb{E}_z=\operatorname{diag}(z,z^{-1},\ldots,z,z^{-1}).$$

If $f(N+1)=\rho f(1)$, ρ being the Floquet multiplier, then by (1.6) one has

$$\mathbb{E}_z f=L_\rho f, \tag{1.7}$$

where L_ρ may be described as a $2N\times 2N$ periodic banded matrix in which each band consists of 2×2 matrices. The so-called multiplier curve[13] is parametrized by (ρ,z) and it is given by the equation

$$|L_\rho-\mathbb{E}_z|=0. \tag{1.8}$$

In another paper we shall use a divisor map (a composition of the Abel map followed by the auxiliary divisor[3,13,19,20]) to show that the equation derived in this paper correspond to linear flows on the Jacobian of the multiplier curve and to express the solution in terms of theta functions.

We are also interested in proving the following Lie theoretical conjectures. (a) The equations derived in this paper may be described in terms of the Kirillov–Poisson bracket coming from a decomposition of a loop extension in z of gl(2, $\mathbb{C}$). (b) The time dependence of L_ρ, like that of the periodic tridiagonal matrices of the Toda problem, may be described in terms of the Kirillov–Poisson bracket[1,3,12,23,24] coming from a decomposition of a loop extension in ρ of gl($2N,\mathbb{C}$).

In the spirit of Ref. 25, we plan to derive an algebraically completely integrable oscillator system from the spectral theory of (1.6). Lastly, and in view of Refs. 17, 26, and 27, we plan to perform numerical experiments, implementing the theory developed above, to analyze the solution.

II. ANOTHER TWIST TO THE KRICHEVER SETUP

In this section we shall be concerned with the following 2×2 linear system of equations:

$$(I-p(x))\psi_x=(E_z+q(x))\psi \text{ or } \pi\psi_x=A\psi, \tag{2.1a}$$

where

$$p=\begin{pmatrix}0 & p_0\\ p_\infty & 0\end{pmatrix},\quad q=\begin{pmatrix}0 & q_\infty\\ q_0 & 0\end{pmatrix},$$

$$E_z=\begin{pmatrix}z & 0\\ 0 & z^{-1}\end{pmatrix},\quad A=(I+p)(E_z+q),$$

$\pi=|I-p|$ and the entries of p and q are given smooth functions of x. We let $\pi'=|E_z+q|=|1-q_0\cdot q_\infty|$. We have $A=\alpha z+\beta+\gamma z^{-1}$, where

$$\alpha=\begin{pmatrix}1 & 0\\ p_\infty & 0\end{pmatrix},\quad \beta=\begin{pmatrix}p_0\cdot q_0 & q_\infty\\ q_0 & p_\infty q_\infty\end{pmatrix},\text{ and }\gamma=\begin{pmatrix}0 & p_0\\ 0 & 1\end{pmatrix}. \tag{2.1b}$$

It may at times be convenient to consider the problem obtained from (2.1a) by dropping π; namely,

$$\tilde\psi_x=A(x,z)\tilde\psi. \tag{2.1c}$$

The relationship between solutions to (2.1a) and solutions to (2.1c) is not a simple one:

$$\psi(x)=Q(x)\tilde\psi(x), \tag{2.1d}$$

where Q is any solution to this Lax-like equation,

$$Q'(x)=[1/\pi(x)]A(x)Q(x)-Q(x)A(x).$$

We have several reasons for looking at (2.1a). Its formal similarity to the AL problem (Sec. III) has already proven to be suggestive. We shall see that (2.1a) has some interesting soliton equations. The Lie algebraic interpretation of these equations should be quite interesting. We shall now show how (2.1a) fits into the Krichever setup.

Let R be a hyperelliptic Riemann surface of genus g. Let 0_α and ∞_α ($\alpha=1,2$) be distinct points of R and suppose that z is a rational function on R with

$$(z)=0_1+0_2-(\infty_1+\infty_2). \tag{2.2}$$

Let δ be a positive nonspecial divisor of degree $g+1$ with support in $R-(z)$. Then, according to Refs. 5 and 20, the linear space K over $\mathbb{C}$ defined by the following two conditions has dimension 2.

Condition 2.1: In the open set $R-(\infty_1+0_1)$, χ is meromorphic and any pole of χ lies in δ.

Condition 2.2: In a neighborhood of ∞_1 (respectively, 0_1), χe^{-zx} (respectively, $\chi e^{-z^{-1}x}$) is holomorphic.

Furthermore, if $L(\delta-0_2-\infty_2)=0$, then there exist a basis χ_1 and χ_2 for K such that

$$\lim_{\rho\to\infty_1}\chi_1 e^{-zx}=1,\quad \lim_{\rho\to 0_1}\chi_2 e^{-z^{-1}x}=1, \tag{2.3a}$$

$$\chi_1(\infty_2)=0 \quad\text{and}\quad \chi_2(0_2)=0. \tag{2.3b}$$

We see then that there exist x-dependent scalars ξ_j,ξ_j^v,η_j and η_j^v, $j\in\mathbb{N}$, such that

$$\chi_1=\begin{cases}e^{zx}(1+\xi_1 z^{-1}+\cdots) & \text{at } \infty_1,\\ e^{z^{-1}}(\xi_0^v+\xi_1^v z+\cdots) & \text{at } 0_1,\end{cases} \tag{2.3c}$$

and

$$\chi_2=\begin{cases}e^{zx}(\eta_0+\eta_1 z^{-1}+\cdots) & \text{at } \infty_1,\\ e^{z^{-1}x}(1+\eta_1^v z+\cdots) & \text{at } 0_1.\end{cases} \tag{2.3d}$$

We set

$$\psi_\alpha(x)=e^{s(x)}\chi_\alpha$$

where

$$\frac{ds}{dx} = \eta_0 \frac{\xi_1^v + \xi_{0x}^v - \xi_0^v \eta_1^v}{1 - \eta_0 \xi_0^v} + \xi_0^v \frac{\eta_1 + \eta_{0x} - \eta_0 \xi_1}{1 - \eta_0 \xi_0^v}. \tag{2.3e}$$

We shall refer to $\psi = (\psi_1, \psi_2)^T$ as a Baker function. The essential singularities of ψ lie at 0_1 and ∞_1, a zero and a pole of z. Our construction departs from that of Krichever in the following way. The spectral parameter in any one of Krichever's eigenvalue problems is a rational function whose poles coincide with the essential singularities.

We shall now derive a linear system of differential equations for ψ. The derivation contains an explanation for our rather mysterious use of $s(x)$. Using the formulas in (2.3e) and (2.3d), we find that

$$\psi_{1x} - z\psi_1 - z^{-1}\xi_0^v \psi_2 = \begin{cases} (s_x + O(z^{-1}))e^{zx+s} & \text{at } \infty_1, \\ (s_x \xi_0^v + \xi_1^v + \xi_{0x}^v - \xi_0^v \eta_1^v + O(z))e^{zx+s} & \text{at } 0_1, \end{cases} \tag{2.4a}$$

and

$$\psi_{2x} - z^{-1}\psi_2 - \eta_0 z\psi_1 = \begin{cases} (s_x \eta_0 + \eta_1 + \eta_0 x - \eta_0 \xi_1 + O(z^{-1}))e^{zx+s} & \text{at } \infty_1, \\ (s_x + O(z))e^{zx+s} & \text{at } 0_1. \end{cases} \tag{2.4b}$$

By (2.3b), the functions on the left-hand side in these formulas belong to the linear space $e^s K$. Thus there exist scalars α, α', β and β' such that

$$\psi_{1x} - z\psi_1 - z^{-1}\xi_0^v \psi_2 = \alpha\psi_1 + \beta\psi_2, \tag{2.4c}$$

$$\psi_{2x} - z^{-1}\psi_2 - \eta_0 z\psi_1 = \alpha'\psi_1 + \beta'\psi_2. \tag{2.4d}$$

Letting $p \to \infty_1$ then 0_1 and using (2.3e) we find that

$$\alpha = \xi_0^v \frac{\eta_1 + \nu_{0x} - \eta_0 \xi_1}{1 - \eta_0 \xi_0^v}, \quad \beta = \frac{\xi_1^v + \xi_{0x}^v - \xi_0^v \eta_1^v}{1 - \eta_0 \xi_0^v}, \tag{2.4e}$$

$$\alpha' = \frac{\eta_1 + \eta_{0x} - \eta_0 \xi_1}{1 - \eta_0 \xi_0^v}, \text{ and } \beta' = \eta_0 \frac{\xi_1^v + \xi_{0x}^v - \xi_0^v \eta_1^v}{1 - \eta_0 \xi_0^v}. \tag{2.4f}$$

We have $\alpha = \xi_0^v \alpha'$ and $\beta' = \eta_0 \beta$. Now let

$$p_\infty = \eta_0, \quad p_0 = \xi_0^v, \quad q_\infty = \beta, \quad \text{and} \quad q_0 = \alpha'. \tag{2.4g}$$

Then we find that

$$s_x = p_0 q_0 + p_\infty q_\infty$$

and ψ satisfies this linear system:

$$\psi_{1x} = (z + p_0 q_0)\psi_1 + (q_\infty + p_0 z^{-1})\psi_2,$$

$$\psi_{2x} = (p_\infty z + q_0)\psi_1 + (p_\infty q_\infty + z^{-1})\psi_2, \tag{2.4h}$$

a system of the form (2.1c).

A. Soliton equations

We suppose that q and p depend on another parameter; say t. Then one should expect that the eigenfunctions of (2.1c) depend on time. We suppose that the t dependence of ψ is prescribed by a linear equation of the form

$$\psi_t = B\psi, \tag{2.5a}$$

and that ψ satisfies $\psi_{xt} = \psi_{tx}$ that is,

$$B_x - A_t = [A,B]. \tag{2.5b}$$

We shall consider (2.5b) as a linear inhomogeneous equation for B. Our purpose in this subsection is to characterize all possible solutions to (2.5b) which are polynomials in z and z^{-1}. We let

$$m = \max\{\deg_z(B), \deg_{z^{-1}}(B)\}.$$

Then there exist z independent 2×2 matrices $B(j = 0,\ldots,m-1)$ and $B(m)$ such that

$$B = \sum_{j=0}^{m-1} (B^+(j)z^{m-j} + B^-(j)z^{-m+j}) + B(m). \tag{2.6}$$

We shall consider first the linear homogeneous part of (2.5b); namely $A_t = 0$ or

$$F_x = [A,F]. \tag{2.7}$$

Theorem 2.3: The Lax equation (2.7) admits a pair of formal series solutions

$$F(z) = \sum_{j=0}^{\infty} F^{(j)} z^{-j} \text{ and } G(z) = \sum_{j=0}^{\infty} G^{(j)} z^j \tag{2.8a}$$

that are determined in a unique way by the following two conditions.

Condition 2.4: The matrices $F^{(0)}$ and $G^{(0)}$ are x-independent diagonal matrices whose entries are given by

$$F_{1,1}^{(0)} = \eta \quad F_{2,2}^{(0)} = 1 + \eta,$$

$$G_{1,1}^{(0)} = 1 + \xi, \text{ and } G_{2,2}^{(0)} = \xi,$$

where η and ξ are any scalars. [When it becomes necessary to indicate the dependence of F on η we shall write $F(z) = F(z;\eta)$.]

Condition 2.5: The diagonal entries of the $F^{(j)}$ and $G^{(j)}$ $(j = 1,2,\ldots)$ are polynomials in (p,q) of positive degree. [Equation (2.7) determines $F^{(k)}$ up to an x-independent diagonal matrix. The ubiquitous condition (2.5) amounts to choosing this matrix to be zero.]

The determinant of F (respectively, G) is $\eta(1+\eta)(\xi(1+\xi))$. [When it becomes necessary to indicate the x dependence of F we shall write $F(z) = F(x,z)$.] Any solution F (respectively G) that is a formal series in z^{-1}/z can be written in the form

$$F(x,z) = \sum_{j=0}^{\infty} \rho_j z^{-j} F(x,z,\eta_j)[G(x,z)]^{-1}$$

$$= \sum_{j=0}^{\infty} \sigma_j z^j G(x,z,\xi_j).$$

Proof: If we substitute the series for F in (2.8a) into (2.7), we find that the equation in the coefficient of z^{-1} is [α, β, and γ are given in (2.1b)]

$$F_x^{(j)} = [\alpha, F^{(j+1)}] + [\beta, F^{(j)}] + [\gamma, F^{(j-1)}]. \tag{2.8b}$$

We must analyze (2.8b) entrywise. One finds that

$$F_{2,1,x}^{(j-1)} = -(p_\infty \Delta(j) + F_{2,1}^{(j)}) - q_0 \Delta(j-1) + (p_\infty q_\infty - p_0 q_0) F_{2,1}^{(j-1)} + F_{2,1}^{(j-2)}, \tag{2.8c}$$

$$F^{(j)}_{1,2,x} = F^{(j+1)}_{1,2} + q_\infty \Delta(j) + (p_0 q_0 - p_\infty q_\infty) F^{(j)}_{1,2} + p_0 \Delta(j-1) - F^{(j-1)}_{1,2}, \tag{2.8d}$$

$$F^{(j)}_{1,1,x} = -F^{(j)}_{2,2x} = -p_\infty F^{(j+1)}_{1,2} + q_\infty F^{(j)}_{2,1} - q_0 F^{(j)}_{1,2} + p_0 F^{(j-1)}_{2,1}, \tag{2.8e}$$

where

$$\Delta(j) = F^{(j)}_{2,2} - F^{(j)}_{1,1}.$$

We proceed by (i) replacing j by $j-1$ in (2.8d) and solving for $F^{(j)}_{1,2}$, (ii) eliminating $F^{(j+1)}_{1,2}$ from (2.8e) using (2.8d) and eliminating $p_\infty \Delta(j) + F^{(j)}_{2,1}$ from the result using (2.8a), and (iii) solving for $F^{(j)}_{2,1)}$ in (2.8c). This leads to these formulas:

$$F^{(j)}_{1,2} = F^{(j-1)}_{1,2,x} - q_\infty \Delta(j-1) + (p_\infty q_\infty - p_0 q_0) F^{(j-1)}_{1,2} - p_0 \Delta(j-2) + F^{(j-2)}_{1,2}, \tag{2.8f}$$

$$\begin{aligned} F^{(j)}_{1,1,x} &= -F^{(j)}_{2,2,x} \\ &= -(p_\infty F^{(j)}_{1,2} + q_\infty F^{(j-1)}_{2,1})_x \\ &\quad + (p_\infty (p_0 q_0 - p_\infty q_\infty) - q_0 + p_{\infty,x}) F^{(j)}_{1,2} \\ &\quad + ((p_\infty q_\infty - p_0 q_0) q_\infty + p_0 + q_{\infty,x}) F^{(j-1)}_{2,1} \\ &\quad + (p_0 p_\infty - q_0 q_\infty) \Delta(j-1) - p_\infty F^{(j-1)}_{1,2} \\ &\quad + q_\infty F^{(j-2)}_{2,1}, \end{aligned} \tag{2.8g}$$

$$F^{(j)}_{2,1} = -F^{(j-1)}_{2,1,x} - p_\infty \Delta(j) - q_0 \Delta(j-1) + (p_\infty q_\infty - p_0 q_0) F^{(j-1)}_{2,1} + F^{(j-2)}_{2,1}. \tag{2.8h}$$

If the matrices $F^{(0)},\ldots,F^{(j-1)}$ are known, then we can solve for $F^{(j)}$ in the order indicated by the previous formulas.

Remark: The matrix F_j does not depend on the diagonal entries of $F_0,\ldots,F_{j-1}$; rather, it depends upon the differences $\Delta(0),\ldots,\Delta(j-1)$. By condition 2.4 we have $\Delta(0) = 1$ for the solutions in (2.8a).

It is not yet clear, as least from (2.8g), that $F_{1,1}$ is a polynomial in $(p_0,p_\infty,q_0,q_\infty)$. Indeed (2.8g) suggests that $F_{1,1}$ contains integrals involving $(p_0,p_\infty,q_0,q_\infty)$. We shall show now that this is not the case. The fundamental consequence of the Lax equation (2.7) is that the spectrum of F is independent of x; in particular,

$$\frac{d}{dx}|F| = F_{1,1x} - 2F_{1,1}F_{1,1x} - (F_{1,2}F_{2,1,x} + F_{1,2,x}F_{2,1}) = 0, \tag{2.9a}$$

where we have used the formula

$$F_{1,1} + F_{2,2} = 1 + 2\eta. \tag{2.9b}$$

The formula in the coefficient of z^{-j} implies that

$$F^{(j)}_{1,1} = -p_\infty F^{(j)}_{1,2} + \sum_{s=1}^{j-1} (F^{(s)}_{1,2}F^{(j-s)}_{2,1} - F^{(s)}_{1,1}F^{(j-s)}_{2,2}) \tag{2.9c}$$

plus a constant. In accordance with condition 2.5, we have taken constant to be zero. It follows then that

$$|F| = |F^{(0)}| = \eta(1+\eta). \tag{2.9d}$$

This completes our construction of F.

We shall now describe the fundamental solution G in terms of F using a symmetry in our eigenvalue problem. If we let v denote the transformation given by

$$v:(p_0,p_\infty,q_0,q_\infty) \to (p_\infty,p_0,q_\infty,q_0), \tag{2.10a}$$

then we have

$$A(x,z) = JA(x,z^{-1})^v J, \text{ where } J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{2.10b}$$

It follows then that if $F(x,z)$ is any solution to (2.7) then $JF(x,z^{-1})^v J^{-1}$ is also a solution to (2.7). We set

$$G(x,z,\xi) = JF(x,z^{-1},\xi)^v J^{-1}. \tag{2.11}$$

Then G satisfies (2.7) condition 2.4, 2.5, and it is a formal series in z. This completes our construction of G.

Let F be a formal series in z^{-1} satisfying (2.7). [We may assume that the diagonal entries of $F^{(0)}$ are distinct; otherwise, we could replace F by $F - F^{(0)}$ and thereby lower the degree of F without upsetting (2.7).] Then there exists constants ρ_0 and η_0 such that $F - \rho_0 F(\eta_0)$ is equal to z^{-1} times a solution to (2.7) of the same form. The last statement of our theorem is proven by continuing in this way. ■

We wish to list the first three matrices of the F series for use in the examples below. These matrices are given by

$$F^{(0)} = \begin{pmatrix} \eta & 0 \\ -p_\infty & 1+\eta \end{pmatrix},$$

$$F^{(1)} = \begin{pmatrix} p_\infty q_\infty & -q_\infty \\ p_{\infty,x} + p_\infty^2 q_\infty - q_0\pi & -p_\infty q_\infty \end{pmatrix}, \tag{2.12a}$$

and

$$F^{(2)} = \begin{pmatrix} p_0 p_\infty \pi' + q_0 q_\infty \pi - p_\infty^2 q_\infty^2 + p_\infty q_{\infty,x} - p_{\infty,x} q_\infty & p_\infty q_\infty^2 - q_{\infty,x} - p_0 \pi' \\ F^{(2)}_{2,1} & -F^{(2)}_{1,1} \end{pmatrix}$$

where

$$F^{(2)}_{2,1} = -p_{\infty,xx} - 3p_\infty q_\infty p_{\infty,x} + p_\infty^2 q_{\infty,x} - 2p_0 q_0 p_{\infty,x} + q_{0,x}\pi - q_0 p_\infty p_{0,x} + p_\infty (2p_0 p_\infty - 4p_0 p_\infty q_0 q_\infty - q_0 q_\infty - 1) + p_0 q_0^2 \pi.$$

The matrices of the G series are given in accordance with (2.11) by the formula

$$G^{(j)}(x,\xi) = JF^{(j)}(x,\xi)^{v}J^{-1}. \tag{2.12b}$$

The formula in the z^j coefficient of (2.5b) with B given by (2.6) is this ($\cdot = \partial/\partial t$)

$$B^{+}(j)_x = [\alpha,B^{+}(j+1)] + [\beta,B^{+}(j)] + [\gamma,B(j-1)]$$
$$\text{if } (j=0,...,m-2), \tag{2.13a}$$

$$B^{+}(m-1)_x - \dot{\alpha} = [\alpha,B(m)] + [\beta,B^{+}(m-1)] + [\gamma,B^{+}(m-2)], \tag{2.13b}$$

$$B(m)_x - \dot{\beta} = [\alpha,B^{-}(m-1)] + [\beta,B(m)] + [\gamma,B^{+}(m-1)], \tag{2.13c}$$

$$B^{-}(m-1)_x - \dot{\gamma} = [\alpha,B^{-}(m-2)] + [\beta,B^{-}(m-1)] + [\gamma,B(m)], \tag{2.13d}$$

$$B^{-}(j)_x = [\alpha,B^{-}(j-1)] + [\beta,B^{-}(j)] + [\gamma,B^{-}(j+1)]$$
$$\text{if } (j=0,...,m-2). \tag{2.13e}$$

By comparing (2.13a) with (2.8b) and using (2.8f)–(2.8h), one can see that there exists a formal series solution to (2.7) of the form

$$F = \sum_{j=0}^{\infty} F^{(j)}z^{-j} \text{ with } F^{(j)} = B^{+}(j)$$
$$\text{if } (j=0,...,m-1). \tag{2.13f}$$

By (2.13e) and by analogy with the previous statement, there exists a formal series solution to (2.7) of the form

$$G = \sum_{j=0}^{\infty} G^{(j)}z^{j} \text{ with } G^{(j)} = B^{-}(j)$$
$$\text{if } (j=0,...,m-1). \tag{2.13g}$$

The matrix $F^{(m)}$ is determined up to an x-independent diagonal matrix by (2.8f), (2.8h) and a computation like the derivation of (2.9c). This would give $F^{(m)}$ in terms of the $F^{(j)} = B^{+}(j)$ $(j=0,...,m-1)$. We shall now derive a simpler formula for $F^{(m)}$; one that involves $B(m)$. Now F satisfies (2.8b) for all j. In particular, we have

$$B^{+}(m-1)_x = [\alpha,F^{(m)}] + [\beta,B^{+}(m-1)] + [\gamma,B^{+}(m-2)] \tag{2.13h}$$

and

$$F^{(m)}_x = [\alpha,F^{(m+1)}] + [\beta,F^{(m)}] + [\gamma,B^{+}(m-1)]. \tag{2.13i}$$

The analogous equations involving $G^{(m)}$ are these:

$$B^{-}(m-1)_x = [\alpha,B^{-}(m-2)] + [\beta,B^{-}(m-1)] + [\gamma,G^{(m)}] \tag{2.13j}$$

and

$$G^{(m)}_x = [\alpha,B^{-}(m-1)] + [\beta,G^{(m)}] + [\gamma,G^{(m+1)}]. \tag{2.13k}$$

We combine (2.13b) with (2.13k) to obtain the formula

$$\dot{\alpha} = [\alpha,F^{(m)} - B(m)]. \tag{2.13l}$$

This formula contains these two bits of information:

$$F^{(m)}_{1,2} = B_{1,2}(m), \tag{2.13m}$$

$$\dot{p}_\infty = (F^{(m)}_{1,1} - F^{(m)}_{2,2} + B_{2,2}(m) - B_{1,1}(m))p_\infty + B_{2,1}(m) - F^{(m)}_{2,1}. \tag{2.13n}$$

Using (2.13j) and (2.13d) in the same way, we find that

$$\dot{\gamma} = [\gamma,G^{(m)} - B(m)]; \tag{2.13o}$$

i.e.,

$$G^{(m)}_{2,1} = B_{2,1}(m) \tag{2.13p}$$

$$\dot{p}_0 = (G^{(m)}_{2,2} - G^{(m)}_{1,1} + B_{1,1}(m) - B_{2,2}(m))p_0 + B_{1,2}(m) - G^{(m)}_{1,2}. \tag{2.13q}$$

Using the (1,2) entry in (2.13c), (2.13m), and the (1,2) entry in (2.13i), we find that

$$\dot{q}_\infty = F^{(m+1)}_{1,2} + q_\infty(F^{(m)}_{2,2} - F^{(m)}_{1,1} + B_{1,1}(m) - B_{2,2}(m)) - B^{-}_{1,2}(m-1). \tag{2.13r}$$

Using the (2,1) entry in (2.13c), (2.13p), and the (2,1) entry in (2.13k), we find that

$$\dot{q}_0 = G^{(m+1)}_{2,1} + q_0(G^{(m)}_{1,1} - G^{(m)}_{2,2} + B_{2,2}(m) - B_{1,1}(m)) - B^{+}_{2,1}(m-1). \tag{2.13s}$$

When we compare the (1,1) entry of (2.13c) to the (1,1) and (2,2) entries in (2.13k), we find that

$$B_{1,1}(m)_x = G^{(m)}_{1,1,x} = -G^{(m)}_{2,2x}.$$

Similarly, one is lead to the formula

$$B_{2,2}(m)_x = -F^{(m)}_{1,1,x} = F^{(m)}_{2,2,x}$$

using the (2,2) entry of (2.13c) and the (1,1) and (2,2) entries of (2.13i). The matrices $F^{(m)}$ and $G^{(m)}$ are given for some constants $(c_1, c_2, \tilde{c}_1, \tilde{c}_2)$ by

$$F^{(m)} = \begin{pmatrix} -B_{2,2}(m) + c_1 & B_{1,2}(m) \\ F^{(m)}_{2,1} & B_{2,2}(m) + c_2 \end{pmatrix},$$

$$G^{(m)} = \begin{pmatrix} B_{1,1}(m) + \tilde{c}_1 & G^{(m)}_{1,2} \\ B_{2,1}(m) & -B_{1,1}(m) + \tilde{c}_2 \end{pmatrix}, \tag{2.13t}$$

where, by (2.8h) or the (2,1) entry in (2.13h), we have

$$F^{(m)}_{2,1} = -B^{+}_{2,1}(m-1)_x - p_\infty(F^{(m)}_{2,2} - F^{(m)}_{1,1}) - q_0(B^{+}_{2,2}(m-1) - B^{+}_{1,1}(m-1)) + (p_\infty q_\infty - p_0 q_0)B^{+}_{2,1}(m-1) + B^{+}_{2,1}(m-2), \tag{2.13u}$$

and, using the symmetry (2.10b), we have

$$G^{(m)}_{1,2} = -B^{-}_{1,2}(m-1)_x - p_0(G^{(m)}_{1,1} - G^{(m)}_{2,2}) - q_\infty(B^{-}_{1,1}(m-1) - B^{-}_{2,2}(m-1)) + (p_0 q_0 - p_\infty q_\infty)B^{-}_{1,2}(m-1) + B^{-}_{1,2}(m-2). \tag{2.13v}$$

Since the diagonal of any $F^{(j)}$ is determined by Eq. (2.8g) only up to a constant diagonal matrix, one could set all the c's to zero. By substituting the previous formulas for $F_{2,1}^{(m)}$ and $G_{1,2}^{(m)}$ into (2.13n) and (2.13q), respectively, one can see that $\dot{p}_\infty$ and $\dot{p}_0$ do not depend on the c's. Now to complete our description of Eq. (2.5b), we need the following formulas for $F_{1,2}^{(m+1)}$ and $G_{2,1}^{(m+1)}$ in terms of the entries of B:

$$\begin{aligned} F_{1,2}^{(m+1)} &= B_{1,2}(m)_x - q_\infty (F_{2,2}^{(m)} - F_{1,1}^{(m)}) \\ &\quad + (p_\infty q_\infty - p_0 q_0) B_{1,2}(m) \\ &\quad - p_0 (B_{2,2}^{+}(m-1) - B_{1,1}^{+}(m-1)) \\ &\quad + B_{1,2}^{+}(m-1), \end{aligned} \tag{2.13w}$$

$$\begin{aligned} G_{2,1}^{(m+1)} &= B_{2,1}(m)_x - q_0 (G_{1,1}^{(m)} - G_{2,2}^{(m)}) \\ &\quad + (p_0 q_0 - p_\infty q_\infty) B_{2,1}(m) \\ &\quad - p_\infty (B_{1,1}^{-}(m-1) - B_{2,2}^{-}(m-1)) \\ &\quad + B_{2,1}^{-}(m-1). \end{aligned} \tag{2.13x}$$

It will be convenient for use in the examples later on to write our soliton equations completely in terms of the entries of B. If we substitute (2.13u), (2.13v) into (2.13n), (2.13q), respectively, we obtain

$$\begin{aligned} \dot{p}_\infty &= (B_{2,2}(m) - B_{1,1}(m)) p_\infty + B_{2,1}(m) \\ &\quad + B_{2,1}^{+}(m-1)_x + (B_{2,2}^{+}(m-1) \\ &\quad - B_{1,1}^{+}(m-1)) q_0 + (p_0 q_0 - p_\infty q_\infty) \\ &\quad \times B_{2,1}^{+}(m-1) - B_{2,1}^{+}(m-2) \end{aligned} \tag{2.13y}$$

and

$$\begin{aligned} \dot{p}_0 &= (B_{1,1}(m) - B_{2,2}(m)) p_0 + B_{1,2}(m) \\ &\quad + B_{1,2}^{-}(m-1)_x + (B_{1,1}^{-}(m-1) \\ &\quad - B_{2,2}^{-}(m-1)) q_0 + (p_\infty q_\infty - p_0 q_0) \\ &\quad \times B_{2,1}^{-}(m-1) - B_{1,2}^{-}(m-2). \end{aligned}$$

If we substitute (2.13w), (2.13x) into (2.13r), (2.13s) respectively, we find that

$$\begin{aligned} \dot{q}_\infty &= (B_{1,1}(m) - B_{2,2}(m)) q_\infty - B_{1,2}^{-}(m-1) \\ &\quad + B_{1,2}(m)_x + B_{1,2}^{+}(m-1) \\ &\quad + (B_{1,1}^{+}(m-1) - B_{2,2}^{+}(m-1)) p_0 \\ &\quad + (p_\infty q_\infty - p_0 q_0) B_{1,2}(m) \end{aligned} \tag{2.13z}$$

and

$$\begin{aligned} \dot{q}_0 &= (B_{2,2}(m) - B_{1,1}(m)) q_0 - B_{2,1}^{+}(m-1) \\ &\quad + B_{2,1}(m)_x + B_{2,1}^{-}(m-1) \\ &\quad + (B_{2,2}^{-}(m-1) - B_{1,1}^{-}(m-1)) p_\infty \\ &\quad + (p_0 q_0 - p_\infty q_\infty) B_{2,1}(m) \end{aligned}$$

Let $e_{\alpha,\beta}$ denote the 2×2 unit matrix with a 1 in entry (α,β). For any series $F = \Sigma_{j=0}^{\infty} F^{(j)} z^j$ we let

$$\pi_+ F = F_> = \sum_{j=0}^{\infty} F^{(j)} z^j$$

and

$$\pi_- F = F_< = \sum_{j=0}^{\infty} F^{(-j)} z^{-j}.$$

Using (2.13f), (2.13g), and (2.13f) and the formula

$$B(m) = G^{(m)} e_{1,1} + F^{(m)} e_{2,2}, \tag{2.14a}$$

we find that

$$B = B^+ + B^-, \tag{2.14b}$$

where

$$B^+ = (z^m F)_> - F^{(m)} e_{1,1}$$

and

$$B^- = (z^{-m} G)_< - G^{(m)} e_{2,2}. \tag{2.14c}$$

We are now in a position to define the soliton hierarchy of our eigenvalue problem (2.1a). We shall derive these equations in the following way. Let F and G as in (2.13f), (2.13g), and (2.13t). There exist sequences ρ_j and η_j of constants such that

$$F = \sum_{j=0}^{\infty} \rho_j z^{-j} F(\eta_j).$$

Since B does not involve the terms of F of order $-(m+1)$ or less, we may assume that $\rho_j = \sigma_j = 0$ if $j = m+1, m+2, \ldots$. We can apply a similar construction to G. We have

$$F = \sum_{j=0}^{m} \rho_j z^{-j} F(\eta_j)$$

and

$$G = \sum_{j=0}^{m} \sigma_j z^j G(\xi_j). \tag{2.15a}$$

If we define for $j \in \mathbb{N}$,

$$F_j(\eta) = (z^j F(\eta))_> - F(\eta)^{(j)} e_{1,1}$$

and

$$G_j(\eta) = (z^{-j} G(\eta))_< - G(\eta)^{(j)} e_{2,2} \tag{2.15b}$$

then by (2.4c) we have

$$B^+ = \sum_{j=0}^{m} \rho_j F_{m-j}(\eta_j)$$

and

$$B^- = \sum_{j=0}^{m} \sigma_j G_{m-j}(\xi_j). \tag{2.15c}$$

We define the basic soliton equations by these formulas:

$$B_{m,x} - A_{t_m} = [A,B],$$

$$\text{where } B_m = \rho F_m(\eta) + \sigma G_m(\xi) \tag{2.16}$$

and $\rho, \sigma \in \{0,1\}$. The soliton equations do not depend on η. One could take ρ and σ to be any constants. We shall now list Eq. (2.16) and (2.13y), (2.13z) for small m to the extent that our preliminary calculation (2.12) will allow.

(m=0)

$$B_0 = \begin{pmatrix} \sigma(1+\xi) & 0 \\ 0 & \rho(1+\eta) \end{pmatrix} = B^{(m)},$$

$$\dot{p}_\infty = -c p_\infty, \quad \dot{p}_0 = c p_0, \quad \dot{q}_\infty = c q_\infty, \quad \dot{q}_0 = -c q_0,$$

$$c := \sigma(1+\xi) - \rho(1+\eta). \tag{2.16a}$$

(m=1)

$$B_1=\rho\begin{pmatrix}\eta & 0\\ -\rho_\infty & 1+\eta\end{pmatrix}z-\begin{pmatrix}p_0q_0\sigma & q_\infty\rho\\ q_0\sigma & p_\infty q_\infty\rho\end{pmatrix}+\sigma\begin{pmatrix}1+\xi & -p_0\\ 0 & \xi\end{pmatrix}z^{-1},$$

$$\begin{aligned}\dot p_\infty &= -\rho p_{\infty,x}+(\rho-\sigma)q_0\pi,\\ \dot p_0 &= -\sigma p_{0,x}+(\sigma-\rho)q_\infty\pi,\\ \dot q_\infty &= -\rho q_{\infty,x}+(\sigma-\rho)p_0\pi',\\ \dot q_0 &= -\sigma q_{0,x}+(\rho-\sigma)p_\infty\pi'.\end{aligned}\tag{2.16b}$$

(m=2)

$$\begin{aligned}B_2=&\rho\begin{pmatrix}\eta & 0\\ -p_\infty & 1+\eta\end{pmatrix}z^2+\rho\begin{pmatrix}p_\infty q_\infty & -q_\infty\\ p_{\infty,x}+p_\infty^2q_\infty-q_0\pi & -p_\infty q_\infty\end{pmatrix}z\\ &+\begin{pmatrix}(p_0^2q_0^2-p_0p_\infty\pi'-q_0q_\infty\pi-p_0q_{0,x}+p_{0,x}q_0)\sigma & (-q_{\infty,x}-p_\infty q_\infty^2-p_0\pi')\rho\\ (-q_{0,x}+p_0q_0^2-p_\infty\pi')\sigma & (p_\infty^2q_\infty^2-p_0p_\infty\pi'-q_0q_\infty\pi-p_0q_{\infty,x}+p_{\infty,x}q_\infty)\rho\end{pmatrix}\\ &+\sigma\begin{pmatrix}-p_0q_0 & q_{0,x}+p_0^2q_0-q_\infty\pi\\ -q_0 & p_0q_0\end{pmatrix}z^{-1}+\sigma\begin{pmatrix}1+\xi & -p_0\\ 0 & \xi\end{pmatrix}z^{-2}.\end{aligned}\tag{2.16c}$$

The soliton equations are so complicated that we shall not write them down.

B. ($p=0$). The AKNS hierarchy

In this subsection we shall be concerned with the eigenvalue problem (2.1a) with $p=0$. In particular, we want to compare its soliton hierarchy to the usual AKNS hierarchy.

The equations (2.8f), (2.8h), and (2.9c) simplify in the following way under the $p=0$ assumption:

$$F_{1,2}^{(j)}=F_{1,2,x}^{(j-1)}-q_\infty\Delta(j-1)+F_{1,2}^{(j-2)},\tag{2.17a}$$

$$F_{1,1}^{(j)}+-F_{2,2}^{(j)}=\sum_{s=1}^{j-1}(F_{1,2}^{(s)}F_{2,1}^{(j-s)}-F_{1,1}^{(s)}F_{2,2}^{(j-s)}),\tag{2.17b}$$

$$F_{2,1}^{(j)}=-F_{2,1,x}^{(j-1)}-q_0\Delta(j-1)+F_{2,1}^{(j-2)}.\tag{2.17c}$$

The first few matrices $F(\eta)^{(j)}$ are given in these formulas:

$$F^{(0)}=\begin{pmatrix}\eta & 0\\ 0 & 1+\eta\end{pmatrix},\quad F^{(1)}=\begin{pmatrix}0 & -q_\infty\\ -q_0 & 0\end{pmatrix},\tag{2.18}$$

$$F^{(2)}=\begin{pmatrix}q_0q_\infty & -q_{\infty,x}\\ q_{0,x} & -q_0q_\infty\end{pmatrix},$$

$$F^{(3)}=\begin{pmatrix}q_0q_{\infty,x}-q_{0,x}q_\infty & 2q_0q_\infty^2-q_\infty-q_{\infty,xx}\\ 2q_0^2q_\infty-q_0-q_{0,xx} & q_{0,x}q_\infty-q_0q_{\infty,x}\end{pmatrix},$$

and

$$F^{(4)}=\begin{pmatrix}-3q_0^2q_\infty^2+2q_0q_\infty+q_{0,xx}q_\infty+q_0q_{\infty,xx}-q_{0,x}q_{\infty,x} & 6q_0q_\infty q_{\infty,x}-2q_{\infty,x}-q_{\infty,xxx}\\ -6q_0q_{0,x}q_\infty+2q_{0,x}+q_{0,xxx} & 3q_0^2q_\infty^2-2q_0q_\infty-q_{0,xx}q_\infty-q_0q_{\infty,xx}+q_{0,x}q_{\infty,x}\end{pmatrix}.$$

Using the equations (2.17a)–(2.17c) and the symmetry (2.10b), we find that the $F(\eta)$ and $G(\xi)$ series in Theorem 2.3 satisfy these equations:

$$\begin{aligned}&F_{2,1}^{(j)}=(-1)^{j+1}F_{1,2}^{(j)v}=(-1)^{j+1}G_{2,1}^{(j)},\\ &F_{1,2}^{(j)}=(-1)^{j+1}G_{1,2}^{(j)},\quad F_{1,1}^{(j)v}=(-1)^jF_{1,1}^{(j)}.\end{aligned}\tag{2.19a}$$

Let $B=B_m$ as in (2.16). Our constraint $p=0$ is compatible with the soliton equations (2.13n) and (2.13q) if and only if

$$B_{2,1}(m)=F_{2,1}^{(m)}\ \text{and}\ B_{1,2}(m)=G_{1,2}^{(m)}.\tag{2.19b}$$

On the other hand, by (2.13m) and (2.13o) we have

$$B_{2,1}(m)=\sigma G_{2,1}(m)\ \text{and}\ B_{1,2}(m)=\rho F_{1,2}(m).\tag{2.19c}$$

The consistency condition (2.19b) comes down to the matrices (2.16) of our soliton hierarchy as

$$\sigma=(-1)^{m+1}\rho.\tag{2.19d}$$

Remark: The previous formula has an interesting consequence. In the absence of constraints, we can think of (2.16) as defining a doubly infinite hierarchy; namely,

$$B_m=F_m\ \text{if}\ m\geqslant 0\ \text{and}\ B_m=G_m\ \text{if}\ m<0.$$

However, in the presence of the constraint $p=0$ our hierachy becomes semi-infinite:

$$B_m=F_m(\eta)+(-1)^{m+1}G_m(\xi)\ \text{if}\ m>0.\tag{2.20}$$

We will see below that the $p=0$ hierarchy contains the nonlinear Schrödinger system and the modified Korteweg–de Vries system. Our previous remark is consistent with the completeness (in the sense of Hamiltonian mechanics) of the AKNS hierarchy.

($m=0$) There is no consistency condition. B_0, $\dot q_\infty$, and $\dot q_0$ are as in (2.16a),

$$\dot{q}_\infty = -cq_\infty \text{ and } \dot{q}_0 = cq_0. \tag{2.20a}$$

(m=1) Consistency: $\sigma = \rho$.

$$B_1 = \rho\begin{pmatrix}\eta & 0\\ 0 & 1+\eta\end{pmatrix}z - \begin{pmatrix}0 & \rho q_\infty\\ \rho q_0 & 0\end{pmatrix} + \rho\begin{pmatrix}1+\xi & 0\\ 0 & \xi\end{pmatrix}z^{-1},$$

$$\dot{q}_\infty = -\rho q_{\beta,x}, \quad \dot{q}_0 = -\rho q_{0,x}. \tag{2.20b}$$

(m=2) Consistency: $\sigma = -\rho$.

$$B_2 = \rho\begin{pmatrix}\eta & 0\\ 0 & 1+\eta\end{pmatrix}z^2 - \rho\begin{pmatrix}0 & q_\infty\\ q_0 & 0\end{pmatrix}z + \rho\begin{pmatrix}q_0q_\infty & -q_{\infty,x}\\ q_{0,x} & q_0q_\infty\end{pmatrix} - \rho\begin{pmatrix}1+\xi & 0\\ 0 & \xi\end{pmatrix}z^{-2} + \rho\begin{pmatrix}0 & q_\infty\\ q_0 & 0\end{pmatrix}z^{-1}, \tag{2.20c}$$

$$\dot{q}_\infty = \rho(2q_0q_\infty^2 - q_{\infty,xx} - 2q_\infty),$$
$$\dot{q}_0 = -\rho(2q_0^2q_\infty - q_{0,xx} - 2q_0).$$

(m=3) Consistency: $\sigma = \rho$

$$B_3 = \rho\begin{pmatrix}\eta & 0\\ 0 & 1+\eta\end{pmatrix}z^3 - \rho\begin{pmatrix}0 & q_\infty\\ q_0 & 0\end{pmatrix}a^2 + \rho\begin{pmatrix}q_0q_\infty & -q_{\infty,x}\\ q_{0,x} & -q_0q_\infty\end{pmatrix}z + \rho\begin{pmatrix}q_0q_{\infty,x} - q_{0,x}q_\infty & 2q_0q_\infty^2 - q_\infty - q_{\infty,xx}\\ 2q_0^2q_\infty - q_0 - q_{0,xx} & q_{0,x}q_\infty - q_0q_{\infty,x}\end{pmatrix} + \rho\begin{pmatrix}1+\xi & 0\\ 0 & \xi\end{pmatrix}z^{-3} - \rho\begin{pmatrix}0 & q_\infty\\ q_0 & 0\end{pmatrix}z^{-2} + \rho\begin{pmatrix}-q_0q_\infty & q_{\infty,x}\\ -q_{0,x} & q_0q_\infty\end{pmatrix}z^{-1}, \tag{2.20d}$$

$$\dot{q}_\infty = \rho(6q_0q_\infty q_{\infty,x} - q_{\infty,xxx} - 2q_{\infty,x}),$$
$$\dot{q}_0 = \rho(6q_0q_\infty q_{\infty,x} - q_{0,xxx} - 2q_{0,x}).$$

We can argue that the soliton hierarchy of the eigenvalue problem (2.1a) with $p = 0$ agrees with the AKNS hierarchy as follows. The previous calculations shows that the $(p = 0)$ hierarchy includes the first few members of the AKNS hierarchy: scaling, translation, NLS, and MKDV. We may conclude that the $p = 0$ hierarchy is the AKNS hierarchy.

III. THE SOLITON EQUATIONS OF THE ABLOWITZ–LADIK PROBLEM

In this section we are concerned with the discrete 2×2 linear eigenvalue problem given by the equation

$$(I_2 - p(n))f(n+1) = (E_z + q(n))f(n)$$
$$\text{or } \pi_n f(n+1) = A(n)f(n) \tag{3.1a}$$

where I_2 is the 2×2 identity matrix,

$$\pi_n = |I_2 + p(n)|, \quad \pi_n' = |E_z + q(n)|,$$
$$p(n) = \begin{pmatrix}0 & p_0(n)\\ p_\infty(n) & 0\end{pmatrix}, \quad q(n) = \begin{pmatrix}0 & q_\infty(n)\\ q_0(n) & 0\end{pmatrix},$$
$$E_z = \begin{pmatrix}z & 0\\ 0 & z^{-1}\end{pmatrix}, \quad f(n) = \begin{pmatrix}f_1(n)\\ f_2(n)\end{pmatrix},$$

and

$$A(n) = (I_2 + p(n))(E_z + q(n)) = \begin{pmatrix}z + p_0(n)q_0(n) & q_\infty(n) + p_0(n)z^{-1}\\ p_\infty(n)z + q_0(n) & p_\infty(n)q_\infty(n) + z^{-1}\end{pmatrix}.$$

We let

$$\alpha(n) = \begin{pmatrix}1 & 0\\ p_\infty(n) & 0\end{pmatrix}, \quad \gamma(n) = \begin{pmatrix}0 & p_0(n)\\ 0 & 1\end{pmatrix},$$

and

$$\beta(n) = (\alpha(n) + \gamma(n))q(n) = (I + p(n))q(n).$$

In this notation, we have

$$A(n) = \alpha(n)z + \beta(n) + \gamma(n)z^{-1}. \tag{3.1b}$$

We shall consider the difference eigenvalue problem obtained from (3.1a) by dropping π_{n+1},

$$h(n+1) = A(n)h(n). \tag{3.1c}$$

We introduce another dependent variable t (time) by adjoining a linear problem to (3.1c); namely,

$$\dot{h}(n) = B(n)h(n) \tag{3.2a}$$

where $\cdot = d/dt$. We insist that the t problem be compatible with (3.1c); that is,

$$\dot{A}(n) = B(n+1)A(n) - A(n)B(n). \tag{3.2b}$$

This is the difference analog of the Lax equation (2.5b). Our problem is to describe all sequences $\langle B(n)\rangle$ satisfying (3.2b) in which each $B(n)$ depends polynomially on z and z^{-1}. It follows immediately from (3.2b) that $\deg_z(B(n))$ and $\deg_{z^{-1}}(B(n))$ are independent of n. We let

$$m = \max\{\deg_z(B(n)), \deg_{z^{-1}}(B(n))\}. \tag{3.2c}$$

Then there exists z independent 2×2 matrices $B^{\pm}(n,j)$ $(j = 0,\ldots,m-1)$ and $B(n,m)$ such that

$$B(n) = B(n,m) + \sum_{j=0}^{m-1} B^+(n,j)z^{m-j} + B^-(n,j)z^{-m+j}. \tag{3.2d}$$

When it becomes necessary to indicate the z dependence of $B(n)$, we shall write

$$B(n) = B(n;z).$$

Let us consider the linear homogeneous part of (3.2b),

$$0 = F(n+1)A(n) - A(n)F(n)$$

or

$$F(n+1) = A(n)F(n)A(n)^{-1}. \tag{3.3}$$

It is interesting to note that the set of all sequences $\langle F(n)\rangle$ satisfying (3.3) is an algebra with respect to component wise matrix multiplication.

Remark 3.1: Our own research is primarily concerned with the periodic inverse spectral transform where it is assumed that $p(n)$ and $q(n)$ are periodic in n; say,

$$p(n+N) = p(n) \quad \text{and} \quad q(n+N) = q(n), \tag{3.4a}$$

for some $N \in \mathbb{N}$. The problem (3.1a) has, for all but finitely many values of z, two linearly independent Floquet solutions; that is, solutions $\langle f(n)\rangle$ satisfying

$$f(n+N) = \rho f(n) \tag{3.4b}$$

for some $\rho \in \mathbb{C}$. In this case it is natural to restrict our attention to $\langle f(n)|\nu = 1,\ldots,N\rangle$. The relationship between the eigenspaces (3.1a) and (3.1c) can now be written down:

$$h(n) = \left(\sum_{j=1}^{n-1} \pi_j\right) f(n). \tag{3.4c}$$

If $h(n)$ satisfies (3.2a) then

$$f(n) = \tilde{B}(n)f(n),$$

$$\text{where } \tilde{B}(n) = B(n) - \sum_{j=1}^{n-1} \dot{\pi}_j \pi_j^{-1} I_2. \tag{3.4d}$$

We will see below that the sequence $\langle B(n)\rangle$ is periodic in n and that $\tilde{B}(n)$ is not in general periodic in n. Indeed, if $\alpha = 1$ or 2 then we have

$$\tilde{B}_{\alpha,\alpha}(n+1) - \tilde{B}_{\alpha,\alpha}(n) = B_{\alpha,\alpha}(n+1) - B_{\alpha,\alpha}(n) - \dot{\pi}_n/\pi_n; \tag{3.4e}$$

and then, by summing over a period, one finds that

$$\tilde{B}_{\alpha,\alpha}(N+1) - \tilde{B}_{\alpha,\alpha}(1) = -\left(\ln \prod_{n=1}^{N} \pi_n\right)^{\cdot}. \tag{3.4f}$$

Our results should have applications to inverse scattering. We shall not provide further details even though finding the correct analytical assumptions may be troublesome.

Theorem 3.2: Equation (3.3) admits a pair of formal series solutions

$$F(n;z) = \sum_{j=0}^{\infty} F(n,j)z^{-j}$$

and

$$G(n;z) = \sum_{j=0}^{\infty} G(n,j)z^{j} \tag{3.5a}$$

[We shall often drop the argument z and write simply $F(n) = F(n;z)$], which are determined in a unique way by the following two conditions.

Condition 3.3: The matrices $F(n,0)$ and $G(n,0)$ are n-independent diagonal matrices whose entries are given by

$$F_{1,1}(n,0) = \eta, \quad F_{2,2}(n,0) = 1+\eta,$$

$$G_{1,1}(n,0) = 1+\xi, \text{ and } G_{2,2}(n,0) = \xi.$$

[When it is necessary to indicate the dependence of F on η we shall write $F(n) = F(n;z;\eta)$.]

Condition 3.4: The diagonal entries of the $F(n,j)$ and $G(n,j)$ $(j = 1,2,\ldots)$ are polynomials in (q,p) of positive degree. [Equation (3.3) determines $F(n,j)$ up to an n-independent diagonal matrix. For the series in (3.5a) we have taken this diagonal matrix to be zero.] Any solution $\langle F(n)\rangle$ to (3.3) that is a formal series in z^{-1} (respectively, z) can be written in the form

$$F(n) = \sum_{j=0}^{\infty} \rho_j z^{-j} F(n;z;\eta_j);$$

respectively,

$$G(n) = \sum_{j=0}^{\infty} \sigma_j z^j G(n,z;\xi_j),$$

for some n-independent constants σ_j, ρ_j, η_j, and ξ_j.

Proof: We begin by writing the second equation in (3.3) in terms of the entries of F. One can show, using a direct but hard computation, that it is equivalent to this formula:

$$\begin{aligned}
&\pi_n \pi_n' \begin{pmatrix} F_{1,1}(n+1) - F_{1,1}(n) & F_{1,2}(n+1) \\ F_{2,1}(n+1) & F_{2,2}(n+1) - F_{2,2}(n) \end{pmatrix} \\
&= \begin{pmatrix} 1 & -p_0 \\ p_\infty & -1 \end{pmatrix}_n \pi_n' \Delta(n) + (F_{1,2}(n)z^2 + q_\infty(n)\Delta(n)z - q_\infty(n)^2 F_{2,1}(n)) \begin{pmatrix} -p_\infty & 1 \\ -p_\infty^2 & p_\infty \end{pmatrix}_n \\
&\quad + (q_0(n)F_{1,2}(n)z + \Delta(n) - q_\infty(n)F_{2,1}(n)z^{-1}) \begin{pmatrix} -(1+p_0p_\infty) & 2p_0 \\ -2p_\infty & 1+p_0p_\infty \end{pmatrix}_n \\
&\quad + (q_0(n)^2 F_{1,2}(n) + q_0(n)\Delta(n)z^{-1} - F_{2,1}(n)z^{-2}) \begin{pmatrix} -p_0 & p_0^2 \\ -1 & p_0 \end{pmatrix}_n,
\end{aligned} \tag{3.5b}$$

where

$$\Delta(n) := F_{2,2}(n) - F_{1,1}(n).$$

We substitute the series for $F(n)$ in (3.5a) into (3.5b) and we examine the coefficient of z^{-j} $(j \geqslant -2)$; each coefficient must vanish. From the z^2 and z terms in condition 3.4, we find that

$$F_{1,2}(n,0) = 0 \text{ and } F_{1,2}(n,1) + q_\infty(n)\Delta(n,0) = 0. \tag{3.5c}$$

From the z^0 equation in (3.5b), we eliminate the coefficient of the matrix

$$\begin{pmatrix} -p_\infty & 1 \\ -p_\infty^2 & p_\infty \end{pmatrix}_n$$

using the (1,2) entry:

$$\pi_n\pi_n' F_{1,2}(n+1,0) = -p_0(n)\pi_n'\Delta(n,0) + (F_{1,2}(n,2) + q_\infty(n)\Delta(n,1) - q_\infty(n)^2 F_{2,1}(n,0)) + 2p_0(n)(q_0(n)F_{1,2}(n,1) + \Delta(n,0)) + p_0(n)^2 q_0(n)^2 F_{1,2}(n,0)$$

or by (3.5c),

$$F_{1,2}(n,2) + q_\infty(n)\Delta(n,1) - q_\infty(n)^2 F_{2,1}(n,0) = -p_0(n)\pi_n'\Delta(n,0).$$

This leaves us with this formula:

$$\pi_n \begin{pmatrix} F_{1,1}(n+1,0) - F_{1,1}(n,0) & F_{1,2}(n+1,0) \\ F_{2,1}(n+1,0) & F_{2,2}(n+1,0) - F_{2,2}(n,0) \end{pmatrix}$$
$$= \begin{pmatrix} p_\infty & -1 \\ p_\infty^2 & -p_\infty \end{pmatrix}_n p_0(n)\Delta(n,0) + \begin{pmatrix} -(1+p_0p_\infty) & 2p_0 \\ -2p_\infty & 1+p_0p_\infty \end{pmatrix}_n \Delta(n,0) + \begin{pmatrix} 1 & -p_0 \\ p_\infty & -1 \end{pmatrix}_n \Delta(n,0)$$
$$= \begin{pmatrix} 0 & 0 \\ p_\infty & 0 \end{pmatrix}_n \pi_n\Delta(n,0).$$

We can make the following conclusions: (a) For each ($\alpha = 1,2$), $F_{\alpha,\alpha}(n,0)$ is independent of n; we choose

$$\eta = F_{1,1}(n,0), \quad (\eta+1) = F_{2,2}(n,0) \quad \text{and then} \quad \Delta(0) = 1$$

in accordance with condition 3.3,

$$\text{(b)} \quad F_{1,2}(n,1) = -q_\infty(n), \tag{3.5d}$$

$$\text{(c)} \quad F_{2,1}(n,0) = -p_\infty(n-1), \tag{3.5e}$$

$$\text{(d)} \quad F_{1,2}(n,2) + q_\infty(n)\Delta(n,1) = q_\infty(n)^2 p_\infty(n-1) - p_0(n)\pi_n'. \tag{3.5f}$$

The equation in the coefficient z^{-j} in the first formula in (3.3) is this:

$$F(n+1,j+1)\alpha(n) - \alpha(n)F(n,j+1) + F(n+1,j)\beta(n) - \beta(n)F(n,j) + F(n+1,j-1)\gamma(n) - \gamma(n)F(n,j-1) = 0 \quad (j=0,1,\ldots). \tag{3.5g}$$

The following formulas will be used repeatedly throughout the rest of this paper; if $\langle C(n)\rangle$ is any sequence of 2×2 matrices then

$$C(n+1)\alpha(n) - \alpha(n)C(n) = \begin{pmatrix} C_{1,1}(n+1) - C_{1,1}(n) + p_\infty(n)C_{1,2}(n+1) & -C_{1,2}(n) \\ C_{2,1}(n+1) + p_\infty(n)(C_{2,2}(n+1) - C_{1,1}(n)) & -p_\infty(n)C_{1,2}(n) \end{pmatrix},$$

$$C(n+1)\beta(n) - \beta(n)C(n) = (p_0(n)q_0(n)(C_{1,1}(n+1) - C_{1,1}(n)) + q_0(n)C_{1,2}(n+1) - q_\infty(n)C_{2,1}(n))e_{1,1}$$
$$\times(p_0(n)q_0(n)C_{2,1}(n+1) - p_\infty(n)q_\infty(n)C_{2,1}(n) + q_0(n)(C_{2,2}(n+1) - C_{1,1}(n)))e_{2,1}$$
$$\times(p_\infty(n)q_\infty(n)C_{1,2}(n+1) - p_0(n)q_0(n)C_{1,2}(n) + q_\infty(n)(C_{1,1}(n+1) - C_{2,2}(n)))e_{1,2}$$
$$\times(p_\infty(n)q_\infty(n)(C_{2,2}(n+1) - C_{2,2}(n)) + q_\infty(n)C_{2,1}(n+1) - q_0(n)C_{1,2}(n))e_{2,2},$$

$$C(n+1)\gamma(n) - \gamma(n)C(n) = \begin{pmatrix} -p_0(n)C_{2,1}(n) & C_{1,2}(n+1) + p_0(n)(C_{1,1}(n+1) - C_{2,2}(n)) \\ -C_{2,1}(n) & p_0(n)C_{2,1}(n+1) + C_{2,2}(n+1) - C_{2,2}(n) \end{pmatrix}.$$

The three formulas (3.5j)–(3.5l) below contain a procedure for constructing the $F(n,j)$ ($j\geqslant 1$). The first formula is equivalent to the (1,2) of (3.5g). The second formula (3.5k) is the discrete analog of (2.9c), our solution to the problem of integrating Eq. (2.8g). By (3.3), trace $(F(n))$ and $|F(n)|$ are n-independent functions:

$$F_{1,1}(n+1) + F_{2,2}(n+1) = F_{1,1}(n) + F_{2,2}(n), \tag{3.5h}$$

$$F_{1,1}(n+1)F_{2,2}(n+1) - F_{1,2}(n+1)F_{2,1}(n+1) = F_{1,1}(n)F_{2,2}(n) - F_{1,2}(n)F_{2,1}(n). \tag{3.5i}$$

The equation in the coefficient of z^{-j} in (3.5i) is equivalent to [using the z^{-j} coefficient in (3.5h) and condition 3.3]

$$F_{1,1}(n+1,j) - F_{1,1}(n,j) = -\sum_{s=1}^{j-1}(F_{1,1}(n+1,s)F_{2,2}(n+1,j-s) - F_{1,1}(n,s)F_{2,2}(n,j-s)) + \sum_{s=0}^{j}(F_{1,2}(n+1,s)F_{2,1}(n+1,j-s) - F_{1,2}(n,s)F_{2,1}(n,j-s)).$$

The formula (3.5k) contains a particular solution to this equation and the equation in the z^{-j} coefficient in (3.5h). One could add an n-independent constant to the right-hand side in (3.5k). This constant is taken to be zero in accordance with condition 3.4. Lastly, (3.5l) is equivalent to the equation in the (2,1) entry of (3.5g) with (n,j) replaced by $(n-1,j-1)$,

$$F_{1,2}(n,j) = p_\infty(n)q_\infty(n)F_{1,2}(n+1,j-1) - p_0(n)q_0(n)F_{1,2}(n,j-1)$$
$$+ q_\infty(n)(F_{1,1}(n+1,j-1) - F_{2,2}(n,j-1)) + F_{1,2}(n+1,j-2) \tag{3.5j}$$
$$+ p_0(n)(F_{1,1}(n+1,j-2) - F_{2,2}(n,j-2)),$$

$$F_{1,1}(n,j) = -F_{2,2}(n,j)$$
$$= -p_\infty(n-1)F_{1,2}(n,j) + \sum_{s=1}^{j-1}(F_{1,2}(n,s)F_{2,1}(n,j-s) - F_{1,1}(n,s)F_{2,2}(n,j-s)), \tag{3.5k}$$

$$F_{2,1}(n,j) = -p_\infty(n-1)(F_{2,2}(n,j) - F_{1,1}(n-1,j))$$
$$- p_0(n-1)q_0(n-1)F_{2,1}(n,j-1)$$
$$+ p_\infty(n-1)q_\infty(n-1)F_{2,1}(n-1,j-1) - q_0(n-1)$$
$$\times(F_{2,2}(n,j-1) - F_{1,1}(n-1,j-1)) + F_{2,1}(n-1,j-2). \tag{3.5l}$$

Suppose that the matrices $F(n,0),\ldots,F(n,j-1)$ are known for all n. Then using these formulas in the indicated order one may compute $F(n,j)$ for all n. This completes our construction of F.

Just as in (2.11), the fundamental solution G may be written in terms of F using a symmetry in our eigenvalue problem. We let v denote the transformation given for all n by

$$v\colon (p_0(n),p_\infty(n),q_0(n),q_\infty(n)) \to (p_\infty(n),p_0(n),q_\infty(n),q_0(n)).$$

Then we have

$$A(n,z) = JA(n;z^{-1})^v J^{-1}, \tag{3.6a}$$

where J is given in (2.10b). It follows then that if $\langle F(n)\rangle$ a solution to (3.3) then so is $\langle JF(n;z^{-1})^v J^{-1}\rangle$. We set

$$G(n;z;\xi) = JF(n;z^{-1};\xi)^v j^{-1}. \tag{3.6b}$$

Then G satisfies (3.3), conditions, 3.3, 3.4, and it is a formal series in z. This completes our construction of G.

The last statement in Theorem (3.2) follows from the argument leading to (2.15a). ■

Let us list the first few matrices of the F series for use in the examples below:

$$F(n,0) = \begin{pmatrix} \eta & 0 \\ -p_\infty(n-1) & 1+\eta \end{pmatrix},$$
$$F(n,1) = \begin{pmatrix} p_\infty(n-1)q_\infty(n) & -q_\infty(n) \\ p_\infty(n-1)^2 q_\infty(n) - q_0(n-1)\pi_{n-1} & -p_\infty(n-1)q_\infty(n) \end{pmatrix} \tag{3.7a}$$

and

$$F(n,2) = \begin{pmatrix} -F_{2,2}(n,2) & q_\infty(n)^2 p_\infty(n-1) - p_0(n)\pi_n' \\ F_{2,1}(n,2) & p_\infty(n-1)^2 q_\infty(n)^2 - p_0(n)p_\infty(n-1)\pi_n' - q_0(n-1)q_\infty(n)\pi_{n-1} \end{pmatrix},$$

where

$$F_{2,1} = p_\infty(n-1)^2(p_0(n)\pi_n{}^q - p_\infty(n-1)q_\infty(n))^2$$
$$+ 2p_\infty(n-1)q_0(n-1)q_\infty(n)\pi_{n-1}$$
$$+ p_0(n-1)q_0(n-1)^2\pi_{n-1}$$
$$- p_\infty(n-2)\pi_{n-1}\pi_{n-1}'.$$

The matrices of the G series are given in accordance with (3.6b) by the formula

$$G(n,j;\xi) = JF(n,j,\xi)^v J^{-1}. \tag{3.7b}$$

Let $\langle B(n)\rangle$ be a sequence of 2×2 matrices of rational form (3.2d) and compatible with (3.2b). The formula in the z^j coefficient (3.2b) is this ($\cdot = \partial/\partial t$):

$$0 = B^+(n+1,j+1)\alpha(n) - \alpha(n)B^+(n,j+1)$$
$$+ B(n+1,j)\beta(n) - \beta(n)B^+(n,j)$$
$$+ B^+(n+1,j-1)\gamma(n) - \gamma(n)B^+(n,j-1)$$
$$\text{if } (j=0,\ldots,m-2), \tag{3.8a}$$

$$\dot\alpha(n) = B(n+1,m)\alpha(n) - \alpha(n)B(n,m)$$
$$+ B^+(n+1,m-1)\beta(n) - \beta(n)B^+(n,m-1)$$
$$+ B^+(n+1,m-2)\gamma(n) - \gamma(n)B^+(n,m-2), \tag{3.8b}$$

$$\dot\beta(n) = B^-(n+1,m-1)\alpha(n) - \alpha(n)B^-(n,m-1)$$
$$+ B(n+1,m)\beta(n) - \beta(n)B(n,m)$$
$$+ B^+(n+1,m-1)\gamma(n) - \gamma(n)B^+(n,m-1), \tag{3.8c}$$

$$\dot\gamma(n) = B^-(n+1,m-2)\alpha(n) - \alpha(n)B^-(n,m-2)$$
$$+ B^-(n+1,m-1)\beta(n) - \beta(n)B^-(n,m)$$
$$+ B(n+1,m)\gamma(n) - \gamma(n)B(n,m), \tag{3.8d}$$
$$0 = B^-(n+1,j-1)\alpha(n) - \alpha(n)B^-(n,j-1)$$
$$+ B^-(n+1,j)\beta(n) - \beta(n)B^-(n,j)$$

$$+B^-(n+1,j+1)\gamma(n)-\gamma(n)B^-(n,j+1)$$

$$\text{if } (j=0,\ldots,m-2). \tag{3.8e}$$

By comparing (3.5g) to (3.8a) one can see that there exists a formal series solution to (3.3) of the form

$$F(n)=\sum_{j=0}^{\infty}F(n,j)z^{-j} \text{ with } F(n,j)=B^+(n,j)$$

$$\text{if } (j=0,\ldots,m-1), \tag{3.8f}$$

for all n. It follows then that there exists a formal series solution to (3.3) of the form

$$G(n)=\sum_{j=0}^{\infty}G(n,j)z^{j} \text{ with } G(n,j)=B^-(n,j)$$

$$\text{if } (j=0,\ldots,m-1). \tag{3.8g}$$

The matrix $F(n,m)$ is determined up to an n-independent diagonal matrix by (3.5j), (3.5l) and a calculation like the derivation of (3.5k). This would give $F(n,m)$ in terms of the $F(n,j)=B^+(n,j)$ $(j=0,\ldots,m-1)$. We shall now derive a simpler formula for $F(n,m)$; one that involves $B(n,m)$. According to (3.8c) with $(j=m-1)$ and then $(j=m)$, we have

$$\begin{aligned}&F(n+1,m)\alpha(n)+B^+(n+1,m-1)\beta(n)\\&\quad+B^+(n+1,m-2)\gamma(n)\\&\quad-\alpha(n)F(n,m)-\beta(n)B^+(n,m-1)\\&\quad-\gamma(n)B^-(n,m-2)=0\end{aligned} \tag{3.8h}$$

and

$$\begin{aligned}&F(n+1,m+1)\alpha(n)+F(n+1,m)\beta(n)\\&\quad+B^+(n+1,m-1)\gamma(n)\\&\quad-\alpha(n)F(n,m+1)-\beta(n)F(n,m)\\&\quad-\gamma(n)B^+(n,m-1)=0.\end{aligned} \tag{3.8i}$$

We substitute G into the first formula in (3.3); the equations in the coefficient of z^j ($j=m-1$ and $j=m$) are

$$\begin{aligned}&B^-(n+1,m-2)\alpha(n)+B^-(n+1,m-1)\beta(n)\\&\quad+G(n+1,m)\gamma(n)-\alpha(n)B^-(n,m-2)\\&\quad-\beta(n)B^-(n,m-1)-\gamma(n)G(n,m)=0\end{aligned} \tag{3.8j}$$

and

$$\begin{aligned}&B^-(n+1,m-1)\alpha(n)+G(n+1,m)\beta(n)\\&\quad+G(n+1,m+1)\gamma(n)-\alpha(n)B^-(n,m-1)\\&\quad-\beta(n)G(n,m)-\gamma(n)G(n,m+1)=0.\end{aligned} \tag{3.8k}$$

Using the (2,1) entry in (3.8h) [or (3.5l) with $j=m$] and the (1,2) entry in (3.8j), we obtain these formulas:

$$\begin{aligned}&F_{2,1}(n+1,m)\\&\quad=-p_\infty(n)(F_{2,2}(n+1,m)-F_{1,1}(n,m))\\&\qquad-p_0(n)q_0(n)B^+_{2,1}(n+1,m-1)\\&\qquad+p_\infty(n)q_\infty(n)B^+_{2,1}(n,m-1)\\&\qquad-q_0(n)(B^+_{2,2}(n+1,m-1)\\&\qquad-B^+_{1,1}(n,m-1))+B^+_{2,1}(n+1,m-2),\end{aligned} \tag{3.8l}$$

$$\begin{aligned}&G_{1,2}(n+1,m)\\&\quad=-p_0(n)(G_{1,1}(n+1,m)-G_{2,2}(n,m))\\&\qquad-p_\infty(n)q_\infty(n)B^-_{1,2}(n+1,m-1)\\&\qquad+p_0(n)q_0(n)B^-_{1,2}(n,m-1)\\&\qquad-q_\infty(n)(B^-_{1,1}(n+1,m-1)\\&\qquad-B^-_{2,2}(n,m-1))+B^-_{1,2}(n+1,m-2).\end{aligned} \tag{3.8m}$$

We combine (3.8b) with (3.8h) and (3.8d) with (3.8j) to obtain these formulas:

$$\begin{aligned}\dot\alpha(n)&=(B(n+1,m)-F(n+1,m))\alpha(n)\\&\quad-\alpha(n)(B(n,m)-F(n,m)),\end{aligned} \tag{3.8n}$$

$$\begin{aligned}\dot\gamma(n)&=(B(n+1,m)-G(n+1,m))\gamma(n)\\&\quad-\gamma(n)(B(n,m)-G(n,m)).\end{aligned} \tag{3.8o}$$

These two formulas (3.8l) and (3.8m) contain the following bits of information:

$$F_{1,2}(n,m)=B_{1,2}(n,m),\quad G_{2,1}(n,m)=B_{2,1}(n,m), \tag{3.8p}$$

$$\begin{aligned}\dot p_\infty(n)&=p_\infty(n)(B_{2,2}(n+1,m)-B_{1,1}(n,m)\\&\quad-F_{2,2}(n+1,m)+F_{1,1}(n,m))\\&\quad+B_{2,1}(n+1,m)-F_{2,1}(n+1,m)\\&=p_\infty(n)(B_{2,2}(n+1,m)-B_{1,1}(n,m))\\&\quad+B_{2,1}(n+1,m)+p_0(n)q_0(n)\\&\quad\times B^+_{2,1}(n+1,m-1)\\&\quad-p_\infty(n)q_\infty(n)B^+_{2,1}(n,m-1)\\&\quad+q_0(n)(B^+_{2,2}(n+1,m-1)-B^+_{1,1}(n,m-1))\\&\quad-B^+_{2,1}(n+1,m-2)\end{aligned} \tag{3.8q}$$

and

$$\begin{aligned}\dot p_0(n)&=p_0(n)(B_{1,1}(n+1,m)-B_{2,2}(n,m)\\&\quad-G_{1,1}(n+1,m)+G_{2,2}(n,m))\\&\quad+B_{1,2}(n+1,m)-G_{1,2}(n+1,m)\\&=p_0(n)(B_{1,1}(n+1,m)-B_{2,2}(n,m))\\&\quad+B_{1,2}(n+1,m)\\&\quad+p_\infty(n)q_\infty(n)B^-_{1,2}(n+1,m-1)\\&\quad-p_0(n)q_0(n)B^-_{2,1}(n,m-1)\\&\quad+q_\infty(B^-_{1,1}(n+1,m-1)-B^-_{2,2}(n,m-1))\\&\quad-B^-_{1,2}(n+1,m-2).\end{aligned} \tag{3.8r}$$

The (1,2) and (2,1) entries of (3.8c) are equivalent to these formulas:

$$\begin{aligned}\dot q_\infty(n)=&-B^-_{1,2}(n,m-1)+q_\infty(n)(B_{1,1}(n+1,m)-B_{2,2}(n,m))+p_\infty(n)q_\infty(n)B_{1,2}(n+1,m)\\&-p_0(n)q_0(n)B_{1,2}(n,m)+B^+_{1,2}(n+1,m-1)+p_0(n)(B^+_{1,1}(n+1,m-1)-B^+_{2,2}(n,m-1))\end{aligned} \tag{3.8s}$$

and

$$\dot{q}_0(n) = B^{-}_{2,1}(n+1,m-1) + p_\infty(n)(B^{-}_{2,2}(n+1,m-1) - B^{-}_{1,1}(n,m-1)) + p_0(n)q_0(n)B_{2,1}(n+1,m)$$
$$- p_\infty(n)q_\infty(n)B_{2,1}(n,m) + q_0(n)(B_{2,2}(n+1,m) - B_{1,1}(n,m)) - B^{+}_{2,1}(n,m-1). \tag{3.8t}$$

The (1,1) entry of (3.8c) contains a formula for $(q_0(n)p_0(n))^{\cdot}$ that we wish to compare to the one arising from (3.8r) and (3.8t). This comparison plus an entry-wise analysis of (3.8k) shows that one may choose

$$G_{2,2}(n,m) = -G_{1,1}(n,m) = B_{2,2}(n,m). \tag{3.8u}$$

Using the (2,2) entry of (3.8c) and (3.8i), we find that we may choose

$$F_{1,1}(n,m) = -F_{2,2}(n,m) = B_{1,1}(n,m). \tag{3.8v}$$

We have shown that the matrices $F(n,m)$ and $G(n,m)$ are given by

$$F(n,m) = \begin{pmatrix} B_{1,1}(n,m) & B_{1,2}(n,m) \\ F_{2,1}(n,m) & -B_{1,1}(n,m) \end{pmatrix} \tag{3.8w}$$

and

$$G(n,m) = \begin{pmatrix} -B_{2,2}(n,m) & G_{1,2}(n,m) \\ B_{2,1}(n,m) & B_{2,2}(n,m) \end{pmatrix},$$

where $F_{1,2}(n,m)$ and $G_{1,2}(n,m)$ may be gotten from (3.8l) and (3.8m). It follows then that

$$B(n,m) = e_{1,1}F_{(n,m)} + e_{2,2}G(n,m). \tag{3.8x}$$

We are now in a position to define the soliton hierarchy of our discrete eigenvalue problem (3.1a). We shall motivate our definition in the following way. Let $F(n)$ and $G(n)$ as in (3.8f), (3.8g) and (3.8w). Then, as in (2.15a), there exists constants ρ_j,σ_j,η_j, and ξ_j such that

$$F(n,z) = \sum_{j=0}^{m} \rho_j z^{-j} F(n;z;\eta_j)$$

and

$$G(n,z) = \sum_{j=0}^{m} \sigma_j z^{j} G(n;z;\xi_j). \tag{3.9a}$$

Let us define two sequences of matrices by these formulas:

$$F_j(n;z;\eta) = (z^j F(n;z;\eta))_{\geqslant} - e_{2,2}F(n,j;\eta) \tag{3.9b}$$

and

$$G_j(n;z;\eta) = (z^{-j}G(n;z;\eta))_{\leqslant} - e_{1,1}G(n,j;\eta). \tag{3.9c}$$

Then by (3.8f) (3.8g), and (3.8x) we have

$$B(n) = B^{+}(n) + B^{-}(n), \tag{3.9d}$$

where

$$B^{+}(n) := \sum_{j=0}^{m} \rho_j F_{m-j}(n;z;\eta_j)$$

and

$$B^{-}(n) := \sum_{j=0}^{m} \sigma_j G_{m-j}(n;z;\xi_j). \tag{3.9e}$$

We define the basic soliton equations for (3.1a) by these formulas:

$$\frac{dA(n)}{dt_m} = B_m(n+1)A(n) - A(n)B_m(n) \tag{3.10}$$

and $\sigma,\rho\in\{0,1\}$. (The soliton equation themselves do not depend on η). We shall now list Eqs. (3.10) and (3.8q)–(3.8t) to the extent that our preliminary calculations (3.7a) will allow. One could take σ and ρ to be any constants in the formulas that follow.

(m=0)

$$B_0(n) = B(n,0) = \begin{pmatrix} \rho\eta & 0 \\ 0 & \sigma\xi \end{pmatrix},$$
$$\dot{p}_\infty = -cp_\infty(n), \quad \dot{p}_0(n) = cp_0(n),$$
$$\dot{q}_\infty(n) = cq_\infty(n), \quad \dot{q}_0(n) = -cq_0(n),$$
$$c = \pi\eta - \sigma\xi. \tag{3.10a}$$

(m=1)

$$B_1(n) = \begin{pmatrix} \eta & 0 \\ -p_\infty(n-1) & 1+\eta \end{pmatrix}\rho z + \begin{pmatrix} p_\infty(n-1)q_\infty(n)\rho & -q_\infty(n)\rho \\ -q_0(n)\sigma & p_0(n-1)q_0(n)\sigma \end{pmatrix} + \begin{pmatrix} 1+\xi & -p_0(n-1) \\ 0 & \xi \end{pmatrix}\sigma z^{-1}, \tag{3.10b}$$
$$\dot{p}_\infty(n) = (q_0(n)\rho - q_0(n+1)\sigma)\pi_n,$$
$$\dot{p}_0(n) = (q_\infty(n)\sigma - q_\infty(n+1)\rho)\pi_n,$$
$$\dot{q}_\infty = (p_0(n-1)\sigma - p_0(n)\rho)\pi'_n,$$
$$\dot{q}_0(n) = (p_\infty(n-1)\rho - p_\infty(n)\sigma)\pi'_n.$$

(m=2)

$$B_2(n) = \begin{pmatrix} \eta & 0 \\ -p_\infty(n-1) & 1+\eta \end{pmatrix}\rho z^2 + \begin{pmatrix} p_\infty(n-1)q_\infty(n) & -q_\infty(n) \\ p_\infty(n-1)^2 q_\infty(n) - q_0(n-1)\pi_{n-1} & -q_\infty(n)p_\infty(n-1) \end{pmatrix}\rho z + \begin{pmatrix} \rho & 0 \\ 0 & \sigma \end{pmatrix}\begin{pmatrix} p_0(n)p_\infty(n-1)\pi'_n + q_0(n-1)q_\infty(n)\pi_{n-1} & q_\infty(n)^2 p_\infty(n-1) - p_0(n)\pi'_n \\ -p_\infty(n-1)^2 q_\infty(n)^2 & \\ q_0(n)^2 p_0(n-1) - p_\infty(n)\pi'_n & B_{1,1}(n,m)^{v} \end{pmatrix} + \begin{pmatrix} -p_0(n-1)q_0(n) & p_0(n-1)^2 q_0(n) - q_\infty(n-1)\pi_{n-1} \\ -q_0(n) & p_0(n-1)q_0(n) \end{pmatrix}\sigma z^{-1} + \begin{pmatrix} 1+\xi & -p_0(n-1) \\ 0 & \xi \end{pmatrix}\sigma z^{-2}. \tag{3.10c}$$

The soliton equations are so complicated that we shall not write them down.

A. ($p=0$). The discrete nonlinear Schrödinger problem

We wish to consider the specialization

$$p_0(n)=0 \text{ and } p_\infty(n)=0,$$

and the soliton flows associated with the eigenvalue problem

$$f(n+1)=(E_z+q(n))f(n). \tag{3.11}$$

Equations (3.5j)–(3.5l) simplify in this way:

$$\begin{aligned}F_{1,2}(n,j)&=q_\infty(n)(F_{1,1}(n+1,j-1)-F_{2,2}(n,j-1))\\&\quad+F_{1,2}(n+1,j-2),\end{aligned} \tag{3.12a}$$

$$\begin{aligned}F_{2,1}(n,j)&=q_0(n-1)(F_{1,1}(n-1,j-1)\\&\quad-F_{2,2}(n,j-1))\\&\quad+F_{2,1}(n-1,j-2),\end{aligned} \tag{3.12b}$$

$$F_{1,1}(n,j)=-F_{2,2}(n,j)=\sum_{s=1}^{j-1}F_{1,2}(n,s)F_{2,1}(n,j-s). \tag{3.12c}$$

The generator F of Theorem 3.2 satisfies these equations:

$$\begin{aligned}0&=F_{1,1}(n,2j+1)=F_{2,2}(n,2j+1)\\&=F_{1,2}(n,2j)=F_{2,1}(n,2j),\end{aligned} \tag{3.12d}$$

if $j=0,1,2,\ldots$. The first few terms of the $F(n,\eta)$ series are given by these formulas:

$$F(n,0;\eta)=\begin{pmatrix}\eta & 0\\ 0 & 1+\eta\end{pmatrix}, \tag{3.13}$$

$$F(n,1;\eta)=\begin{pmatrix}0 & -q_\infty(n)\\ -q_0(n-1) & 0\end{pmatrix}$$

and

$$F(n,2;\eta)=\begin{pmatrix}q_\infty(n)q_0(n-1) & 0\\ 0 & -q_\infty(n)q_0(n-1)\end{pmatrix}.$$

Let $B(n)=B_m(n)$ as in (3.10). Our constraint $p=0$ is consistent with the soliton equations (3.8r) if and only if

$$B_{2,1}(n,m)=\rho F_{2,1}(n,m;\eta) \text{ and } B_{1,2}(n,m)=\sigma G_{1,2}(n,m;\xi) \tag{3.14a}$$

for all n. On the other hand, by (3.8p) we have

$$B_{2,1}(n,m)=\sigma G_{2,1}(n,m;\xi)$$

and

$$B_{1,2}(n,m)=\rho F_{1,2}(n,m;\xi). \tag{3.14b}$$

These equations are not consistent [for nonzero (ρ,σ)] unless m is even. If $m=2s$ then the soliton equations (3.8s) and (3.8t) come down as

$$\begin{aligned}\dot q_\infty(n)&=-\sigma F_{2,1}(n,m-1)^\nu+\rho F_{1,2}(n+1,m-1)\\&\quad+q_\infty(n)(\rho F_{1,1}(n+1,m)-\sigma F_{1,1}(n,m)\nu)\end{aligned} \tag{3.15a}$$

and

$$\begin{aligned}\dot q_0(n)&=\sigma F_{1,2}(n+1,m-1)^\nu-\rho F_{2,1}(n,m-1)\\&\quad+q_0(n)(\sigma F_{1,1}(n+1,m)^\nu-\rho F_{1,1}(n,m)).\end{aligned} \tag{3.15b}$$

(m=2). (DNLS):

$$\begin{aligned}B_2(n)&=\begin{pmatrix}\eta & 0\\ 0 & 1+\eta\end{pmatrix}\rho z^2-\begin{pmatrix}0 & q_\infty(n)\\ q_0(n-1) & 0\end{pmatrix}\rho z\\&\quad+\begin{pmatrix}q_0(n-1)q_\infty(n)\rho & 0\\ 0 & q_\infty(n-1)q_0(n)\sigma\end{pmatrix}\\&\quad-\begin{pmatrix}0 & q_\infty(n-1)\\ q_0(n) & 0\end{pmatrix}\sigma z^{-1}+\begin{pmatrix}1+\xi & 0\\ 0 & \xi\end{pmatrix}\sigma z^{-2},\end{aligned} \tag{3.16}$$

$$\dot q_\infty(n)=(q_\infty(n-1)\sigma-q_\infty(n+1)\rho)\pi_n',$$

$$\dot q_0(n)=(q_0(n-1)\rho-q_0(n+1)\sigma)\pi_n'.$$

These equations reduce to the DNLS equation under the following substitutions:

$$\sigma=-i,\quad \rho=i,\quad B(n)=\begin{pmatrix}i & 0\\ 0 & -i\end{pmatrix}+B_2(n),$$

and

$$q_0(n)=-q_\infty(n)^*.$$

B. Specialization to the Toda lattice

It is a well known fact that the Toda lattice, the nonlinear system of differential equations given by

$$a_n=a_n(b_{n+1}-b_n)/2 \text{ and } \dot b_n=(a_n^2-a_{n-1}^2), \tag{3.17a}$$

can be described as a spectrum preserving deformation of the following second-order scalar difference equation:

$$\lambda u(n)=a_{n-1}u(n-1)+b_nu(n)+a_nu(n+1). \tag{3.17b}$$

The following transformation was discovered by Ablowitz and Ladik:

$$\left.\begin{aligned}u(n)&=a_{n-1}\cdots a_1f_1(n)\\ f_1(n)&=f_2(n)-z^{-1}f_2(n-1)\end{aligned}\right\}$$

$$\Rightarrow\begin{pmatrix}1 & -1+a_n^2\\ -1 & 1\end{pmatrix}f(n+1)=\begin{pmatrix}z & -b_n\\ 0 & z^{-1}\end{pmatrix}f(n), \tag{3.17c}$$

where z satisfies

$$\lambda=z+z^{-1}.$$

Equations (3.10b), with p and q defined in accordance with (3.17c):

$$p_0(n)=1-a_n^2,\quad p_\infty(n)=1,\quad \pi_n=a_n^2,$$

$$q_0(n)=0,\quad q_\infty(n)=-b_n, \text{ and } \pi_n'=1,$$

and with $\rho=\sigma=1$ for consistency, are equivalent to (3.17a).

ACKNOWLEDGMENTS

I am especially grateful to H. Flaschka who introduced me to the AL problem and who spent many hours in communication with me while in Kyoto and Potsdam. I would also like to thank Loc Stewart for a very professional preparation of the manuscript.

This research was supported by a grant from the Louisiana State University Council of Research and by a grant

from the Louisiana Education Quality Support Fund [No. 86-LBR(16)-016-04].

[1]H. Flaschka, A. C. Newell, and T. Ratiu, Physics D **9**, 300 (1983).
[2]G. Wilson, Ergodic Theory Dynamical Systems **1**, 361 (1981).
[3]M. Adler and P. van Moerbeke, Adv. Math, **38**, 267 (1980).
[4]M. Abolowitz and J. Ladik, J. Math. Phys. **16**, 598 (1975).
[5]I. M. Krichever, Funkt. Anal. Appl. **11**, 12 (1977).
[6]G. Wilson, Proc. Comb. Phil. Soc. **86**, 131 (1979).
[7]I. M. Gel'fand and L. A. Dikii, Russ. Math. Surveys **30**, 77 (1975).
[8]M. Adler, Invent. Math. **50**, 219 (1979).
[9]B. A. Kuperschmidt, Asterisque **123** (1985).
[10]B. A. Kuperschmidt and G. Wilson Invent. Math. **62**, 403 (1981).
[11]Y. I. Manin, J. Sov. Math. **1**, 1 (1979).
[12]V. G. Drinfel'd and V. V. Sokolov, J. Sov. Math. **30**, 1975 (1985).
[13]H. P. McKean Commun. Pure Appl. Math. **34**, 599 (1981).
[14]K. M. Case, J. Math. Phys. **14**, 916 (1973).
[15]K. M. Case and M. Kac, J. Math. Phys. **14**, 594 (1973).
[16]H. Flaschka, Prog. Theor. Phys. **51**, 703 (1974).
[17]H. Flaschka and D. W. McLauglin, Prog. Theor. Phys. **55**, 438 (1976).
[18]M. Kac and P. van Moerbeke, Proc. Natl. Acad. Sci. USA **72**, 2879 (1975).
[19]P. van Moerbeke, *Lecture Notes in Mathematics*, Vol. 755 (Springer, Berlin, 1979).
[20]P. van Moerbeke and D. Mumford, Acta Math. **134**, 93 (1979).
[21]D. Mumford, Proceedings of a Conference in Algebraic Geometry, Kyoto (Japan Math. Soc., 1977), p. 115.
[22]P. Deift, L. C. Li, and C. Tomei, J. Funct. Anal. **64**, 358 (1985).
[23]M. A. Semenov-Tian-Shansky, Rims Kyoto **21**, 1237 (1985).
[24]Symes, Invent. Math. **59**, 13 (1980).
[25]P. Deift, F. Lund, and E. Trubowitz, Commun. Math. Phys. **74**, 141 (1980).
[26]W. E. Ferguson, Math. Comp. **35**, 1203 (1980).
[27]W. E. Ferguson, H. Flaschka, and D. W. Mclaughlin, J. Comput. Phys. **45**, 157 (1982).
[28]R. J. Schilling, Proc. Am. Math. Soc. **98**, 671 (1986).

Lagrangian formalism and retarded classical electrodynamics

X. Jaen, J. Llosa, and A. Molina
Grup de Relativitat, Societat Catalana de Física (I.E.C.) and Departament Física Fonamental, Universitat de Barcelona, Diagonal 647, E-08028 Barcelona, Spain

(Received 28 June 1988; accepted for publication 2 February 1989)

Unlike the $1/c^2$ approximation, where classical electrodynamics is described by the Darwin Lagrangian, here there is no Lagrangian to describe retarded (resp., advanced) classical electrodynamics up to $1/c^3$ for two-point charges with different masses.

I. INTRODUCTION

The noninteraction theorem of Currie *et al.*[1] and its further generalizations[2,3] established that if position coordinates are to be taken as canonical, then there is no Poincaré invariant Hamiltonian system of directly interacting particles other than the trivial case of free particles.

The original result was actually proven in the instant form of Hamiltonian relativistic mechanics[4] and has only recently been extended to the other two Dirac approaches to relativistic Hamiltonian dynamics, namely, the front- and point-form approaches[5] and, also, to more general approaches.[6,7]

By means of a Legendre transformation, that negative result can be translated into its Lagrangian counterpart, namely, the nonexistence of Poincaré invariant Lagrangian systems of directly interacting particles, apart from the above-mentioned case of free particles. (Note that here, the "Poincaré invariant Lagrangian system" only means that the Euler–Lagrange equations are Poincaré invariant; it does not imply the Poincaré invariance of the Lagrangian function.)

However, this result does not exclude the existence of Lagrangian (resp., Hamiltonian) systems that are Poincaré invariant up to terms $1/c^2$, that is, modulo $1/c^3$. This case encompasses several well-known Lagrangians, e.g., Darwin,[8] Einstein *et al.*,[9] Bopp,[10] and Breit[11] for classical spin charges. Going further into this approach, Martin and Sanz[12] proved that *there exist nontrivial Lagrangian systems of directly interacting particles that are Poincaré invariant up to* $1/c^n$, but only if $n<6$.

In a later paper,[13] Martin and Sanz derive the most general form of a Lagrangian function such that (i) it is invariant under the Aristotle group (i.e., space rotations and space and time translations), (ii) it admits a Newtonian limit, (iii) it is separable, and (iv) it yields a system of equations of motion that is Poincaré invariant up to $1/c^3$. Martin and Sanz[13] also obtain some conditions to be fulfilled by the $1/c^4$ part of the Lagrangian in order to guarantee the Poincaré invariance of equations of motion up to this order: They finally prove that the approximated Lagrangians derived for systems of particles interacting through a classical field are not Poincaré invariant up to $1/c^4$.

Reference 13 agrees with the well-known fact that although classical electrodynamics of point charges is described up to $1/c^4$ terms by the Darwin Lagrangian and the equations of motion are relativistic invariant up to this order, the same does not hold for the Golubenkov–Smorodinskii Lagrangian,[14] i.e., the Lagrangian one would obtain from Fokker symmetric electrodynamics of two charges[15] by a convenient $1/c$ expansion.[16,17]

One point, which in our opinion is interesting, has not been considered in Ref. 13: *Is there a Lagrangian system fulfilling conditions (i)–(iv) and describing retarded (resp., advanced) electrodynamics up to* $1/c^3$*?* [Note that this question would not make sense for symmetric electrodynamics—half-retarded plus half-advanced—because time reversal invariance implies that only even powers of $1/c$ occur in the Lagrangian.)

In the present paper we give a negative answer to the above question, taking as equations of motion those given by predictive relativistic retarded electrodynamics[18] of two point charges up to $1/c^3$.

II. THE LAGRANGIAN

The search is for a Lagrangian approximated up to $1/c^3$. Meeting conditions (i)–(iv) of Sec. I is done as follows.[13] First, an analytical dependence on the "small" parameter $1/c$ is assumed:

$$L=\sum_{n=0}^{\infty} c^{-n} L^{(n)}(\mathbf{x}_a,\mathbf{v}_b,t)\ . \tag{1}$$

Then, the Aristotle invariance condition (i) implies

$$L^{(n)}(\mathbf{x}_a,\mathbf{v}_b,t)=L^{(n)}(r,s,q,v_a^2,v^2)\ , \tag{2}$$

where the Aristotle invariant variables

$$\begin{aligned} &r\equiv|\mathbf{x}_1-\mathbf{x}_2|,\quad s=\tfrac{1}{2}(\mathbf{x}_1-\mathbf{x}_2)\cdot(\mathbf{v}_1-\mathbf{v}_2)\ ,\\ &q\equiv\tfrac{1}{2}(\mathbf{x}_1-\mathbf{x}_2)\cdot(\mathbf{v}_1+\mathbf{v}_2),\quad v^2\equiv(\mathbf{v}_1-\mathbf{v}_2)^2 \end{aligned} \tag{3}$$

have been introduced.

The Newtonian limit condition (ii) and the separability condition (iii) read, respectively, as

$$L^{(0)}=\tfrac{1}{2}m_1v_1^2+\tfrac{1}{2}m_2v_2^2-V(r) \tag{4}$$

and

$$\lim_{r\to\infty} L=\sum_a m_a c^2\left\{1-\sqrt{1-v_a^2/c^2}\right\}. \tag{5}$$

Equation (4) ensures that the Lagrangian (1) is nonsingular, at least for "small" values of $1/c$; indeed,

$$\frac{\partial^2 L}{\partial v_a^i\,\partial v_b^j}=m_a\delta_{ab}\delta_{ij}+O\left(\frac{1}{c}\right), \tag{6}$$

where the subscripts $a,b=1,2$.

Hence the Euler–Lagrange equations can be solved in the particle accelerations, thus yielding

 0022-2488/89/071502-03$02.50

$$\dot{\mathbf{v}}_a = \boldsymbol{\mu}_a(\mathbf{x}_b,\mathbf{v}_c) = \sum_{n=0}^{\infty} \frac{1}{c^n} \boldsymbol{\mu}_a^{(n)}(\mathbf{x}_b,\mathbf{v}_c), \tag{7}$$

where the subscripts $a, b, c = 1, 2$.

The condition of Poincaré invariance up to $1/c^3$ [i.e., condition (iv)] of the equations of motion (7) is then ensured by requiring $\boldsymbol{\mu}_a(\mathbf{x}_b,\mathbf{v}_c)$ to meet the Currie–Hill equations[19] up to $1/c^3$:

$$\frac{v_a^k}{c^2}(x_{aj} - x_{bj}) \frac{\partial \mu_b^i}{\partial x_a^k} + \left[\frac{v_a^k v_{aj}}{c^2} + \frac{\mu_a^k}{c^2}(x_{aj} - x_{bj}) - \epsilon_a \delta_j^k\right] \times \frac{\partial \mu_b^i}{\partial x_a^k} = \frac{1}{c^2}(2\mu_b^i v_{bj} + v_b^i \mu_{bj}), \tag{8}$$

where $\epsilon_a = 1$. (Summation over repeated indices is understood.)

It has been proven elsewhere[20] that these equations (8) are a consequence of requiring that the whole family of solutions of the equations of motion (7) are invariant under the action of the Poincaré group on the space of the initial data $D = (\mathbf{x}_{10},\mathbf{x}_{20};\mathbf{v}_{10},\mathbf{v}_{20})$. This is a sort of "world line condition" that basically states that provided that $\phi_a(t,D)$, where $a = 1,2$, are the particle trajectories from some given initial data D and D' are the transformed initial data (i.e., the initial data that would be "seen" from another inertial frame), then the world lines $[t,\phi_a(t,D)]$ transform into the world lines $[t',\phi_a(t',D')]$.

The conditions (2), (4), (5), and (8) constrain the Lagrangian to have the special form

$$L = \tfrac{1}{2}(m_1 v_1^2 + m_2 v_2^2) - V(r) + \frac{1}{c^2}\left\{\frac{1}{8}(m_1 v_1^4 + m_2 v_2^4) - V'(r)\frac{q^2}{2r} + \frac{V}{4}(v_1^2 + v_2^2) + \frac{\alpha(r)}{4}(v_1^2 - v_2^2) + \alpha'(r)\frac{sq}{r} + f(r,s,v^2)\right\} + \frac{1}{c^3}\left\{\frac{\gamma(r)}{4}(v_1^2 - v_2^2) + \gamma'(r)\frac{sq}{r} + g(r,s,v^2)\right\} + O\left(\frac{1}{c^4}\right), \tag{9}$$

where a prime means derivative and $\alpha(r)$, $\Gamma(r)$, $f(r,s,y)$, and $g(r,s,y)$ are arbitrary functions subject to the limit conditions (for $r \to \infty$)

$$\lim \alpha(r) = \lim(\alpha'(r)/r) = \lim f(r) = \lim \gamma(r) = \lim(\gamma'(r)/r) = \lim g(r) = 0. \tag{10}$$

[See Ref. 13, Sec. III, for the intermediate steps leading to Eq. (9); indeed, there Eq. (9) is labeled (3.13).]

III. RETARDED (ADVANCED) ELECTRODYNAMICS OF TWO-POINT CHARGES

The equations of motion for retarded (resp., advanced) electrodynamics of two directly interacting charges (without an intermediate field) are given by

$$m_a \frac{d}{dt}(\gamma_a \mathbf{v}_a) = e_a\left[\mathbf{E}_{a'}(\mathbf{x}_a,\epsilon) + \frac{\mathbf{v}_a}{c} \times \mathbf{B}_{a'}(\mathbf{x}_a,\epsilon)\right], \tag{11}$$

where the subscripts $a \neq a'$ and $a,a' = 1, 2$ and where $\gamma_a = \sqrt{1 - v_a^2/c^2}$. Here, $\mathbf{E}_{a'}(\mathbf{x}_a,\epsilon)$ and $\mathbf{B}_{a'}(\mathbf{x}_a,\epsilon)$ are the retarded ($\epsilon = -1$) [resp., advanced ($\epsilon = +1$)] Lienard–Wiechert electric and magnetic fields associated (*adjunct*) to the charge $a' \neq a$. The trouble with Eq. (11) is that it is not an ordinary differential system because $\mathbf{E}_{a'}(\mathbf{x}_a,\epsilon)$ and $\mathbf{B}_{a'}(\mathbf{x}_a,\epsilon)$ are only defined for null configurations of particle a' retarded (resp., advanced) relative to a, i.e., $(x_1^\mu - x_2^\mu) \times (x_{1_\mu} - x_{2_\mu}) = 0$ and $(x_a^0 - x_{a'}^0)\cdot\epsilon < 0$.

Hence (11) is a difference-differential system and initial positions and velocities of particles do not determine a unique future evolution.

However, Eq. (11) can be taken as a boundary condition for solving the Currie–Hill equation (8).[20] (The Currie–Hill equation (8) acts as a partial differential condition on the particle accelerations in order to ensure the Poincaré invariance of the world lines.) Introducing the additional requirement that world lines depend analytically on the small parameter $1/c$, the resulting equations of motion up to the order $1/c^3$ are derived in Ref. 18 and read as

$$m_a \boldsymbol{\mu}_a = \eta_a\left[-\frac{V'}{r}\mathbf{r} + \frac{1}{c^2}\left\{-\mathbf{r}\left(\frac{V'}{2r}(v^2 - 2v_a^2) + \frac{1}{2r}\left(\frac{V'}{r}\right)'(\mathbf{r}\cdot\mathbf{v}_{a'}) + \frac{VV'}{m_{a'}r}\right) + \mathbf{v}\frac{V'}{r}(\mathbf{r}\cdot\mathbf{v}_a)\right\} + \frac{1}{c^3}\epsilon\frac{2e_1e_2}{3m_{a'}}\left[\mathbf{r}\frac{2s}{r}\left(\frac{V'}{r}\right)' + \frac{V'}{r}\mathbf{v}\right]\right] + O\left(\frac{1}{c^4}\right), \tag{12}$$

where $\mathbf{r} = \mathbf{x}_1 - \mathbf{x}_2$, $\mathbf{v} = \mathbf{v}_1 - \mathbf{v}_2$, and $V(r) = e_1e_2/r$ is the Coulomb potential energy.

The accelerations $\boldsymbol{\mu}_a(\mathbf{x},\mathbf{v})$ given by (12) must now be compared with those that one derives from the Lagrangian (9). Expanding the Euler–Lagrange equations in a $1/c$ series and taking Eq. (6) into account, we obtain from the $1/c^n$ term in the expansion that

$$m_a \mu_a^{i(n)} = \frac{\partial L^{(n)}}{\partial x_a^i} - \sum_{b=1}^{2} \times \left(v_b^j \frac{\partial^2 L^{(n)}}{\partial x_b^j \partial v_a^i} + \sum_{m=0}^{n-1} \mu_b^{j(m)} \frac{\partial^2 L^{(n-m)}}{\partial v_b^j \partial v_a^i}\right). \tag{13}$$

As is well known, for $n \leqslant 2$, the equations of motion (12) can be derived from the Darwin Lagrangian[8]

$$L_{\mathrm{DW}} = L^{(0)} + (1/c^2)L^{(2)} + O(1/c^3).$$

(Notice that $L^{(1)} = 0$.) A short calculation proves that the above equation is fitted by the Lagrangian (9) for

$$f = -(e_1e_2/4)(v^2/r) - (e_1e_2/2r^3)s^2, \quad \alpha = 0.$$

For $n = 3$, we have

$$m_a \mu_a^{i(3)} = \frac{\partial L^{(3)}}{\partial x_a^i} - \sum_{b=1}^{2}\left(v_b^j \frac{\partial^2 L^{(3)}}{\partial x_b^j \partial v_a^i} + \mu_b^{j(0)} \frac{\partial^2 L^{(3)}}{\partial v_b^j \partial v_a^i}\right), \tag{14}$$

which using (9), yields

$$m_a \boldsymbol{\mu}_a^{(3)}$$

$$= \mathbf{r}\left[-\left(\frac{\gamma'}{r}\right)' \frac{s^2\eta_a}{r} - \frac{\gamma'}{r}\left(\frac{y}{4} + \frac{1}{2}\frac{e_1e_2}{r}\frac{\eta_a}{m_a}\right) \right.$$

$$- \gamma \frac{1}{2}\frac{e_1e_2}{r^3}\frac{\eta_a}{m_a} - g_{rs}\cdot\frac{s}{r} - g_{ss}\cdot\frac{1}{4}\left(y + \frac{e_1e_2}{r}\frac{1}{\mu}\right)$$

$$\left. - g_{sy}\frac{2e_1e_2}{r^3}\frac{s}{\mu} - g_y\frac{2e_1e_2}{\mu r^3} + \frac{g_r}{r}\right] - \mathbf{v}\left[\frac{\gamma'}{r}s\eta_a + g_{ry}\right.$$

$$\left. \cdot\frac{4s}{r} + g_{yy}\cdot\frac{8e_1e_2}{r^3}\frac{s}{\mu} + g_{sy}\left(y + \frac{e_1e_2}{\mu r}\right)\right], \qquad (15)$$

where $g_r = \partial g/\partial r, g_{rs} = \partial^2 g/\partial r\,\partial s$, etc., and the new variable $y = v^2$ has been introduced. Moreover, $\mu = m_1m_2/(m_1 + m_2)$ stands for the reduced mass.

Then comparing (15) with the $1/c^3$ term in the rhs of Eq. (12) and after some manipulation we arrive at

$$\gamma(r) = 0, \quad m_1 = m_2 . \qquad (16)$$

Consequently, the retarded (resp., advanced) electrodynamics of two-point charges does not admit a Lagrangian description approximated up to $1/c^3$ unless both particles have the same mass.

In that case, the comparison of Eqs. (12) and (15) also yields the further condition

$$\frac{\partial g}{\partial \mathbf{r}} - D\left(\frac{\partial g}{\partial \mathbf{v}}\right) = \frac{\epsilon e_1^2 e_2^2}{3\mu}\left[\frac{6s}{r^5}\mathbf{r} - \frac{1}{r^3}\mathbf{v}\right], \qquad (17)$$

with

$$D \equiv \mathbf{v}\frac{\partial}{\partial \mathbf{r}} + \frac{e_1e_2}{\mu}\frac{\mathbf{r}}{r^3}\frac{\partial}{\partial \mathbf{v}} . \qquad (18)$$

Equation (17) splits into two scalar equations: (*along* $\mathbf{r}$),

$$\frac{2g_r}{r} - \frac{1}{2}h_s = \frac{3e_1^2e_2^2}{\mu}\frac{2s}{r^5}; \qquad (19a)$$

(*along* $\mathbf{v}$),

$$g_s - 2h_y = -\epsilon(e_1e_2/3\mu r^3); \qquad (19b)$$

where

$$h = Dg = g_r\frac{2s}{r} + g_s\frac{1}{2}\left(v^2 + \frac{e_1e_2}{\mu r}\right)$$

$$= g_y\frac{4se_1e_2}{\mu r^3} . \qquad (20)$$

We stop at this point because whether or not Eqs. (19a) and (19b) are integrable has very little significance; indeed, they will be relevant only in the very special case of equal masses.

IV. CONCLUSION AND OUTLOOK

We have proven that under rather unrestrictive conditions (invariance under space-time translations and rotations) there is no Lagrangian up to $1/c^3$ for the retarded (resp., advanced) predictive electrodynamics of two-point charges with different masses; therefore, the no-interaction theorem for $1/c$ expansions[12] applies already at this order $(1/c^3)$ for these theories. Nevertheless, it seems that this negative result only occurs for different masses.

As is usually done in relativistic theories of directly interacting particles, a possible way out would consist in dropping the condition that the configuration space is spanned by the particle positions $\mathbf{x}_a$, where $a = 1,2$, and introducing a new set of configuration space coordinates $\mathbf{q}_a$ related to the former ones by

$$\mathbf{x}_a = \mathbf{q}_a + [(m_1 - m_2)/c^3]\, f_a(\mathbf{q},\dot{\mathbf{q}}) ,$$

where f_a is a set of suitably chosen functions.

ACKNOWLEDGMENT

This work has been partially supported by CAICyT under Contract No. 0649-84.

[1] D. G. Currie, T. F. Jordan, and E. C. Sudarshan, Rev. Mod. Phys. **35**, 350 (1963).
[2] H. Van Dam and E. P. Wigner, Phys. Rev. **142**, 838 (1966); H. Leutwyler, Nuovo Cimento **37**, 556 (1965).
[3] From the Hamiltonian point of view, see, R. N. Hill, J. Math. Phys. **8**, 1756 (1967); L. Bel, Ann. Inst. H. Poincaré **XIV**, 189 (1971). From the Lagrangian point of view, see, E. H. Kerner, J. Math. Phys. **9**, 222 (1968); J. Martin and J. L. Sanz, *ibid.* **19**, 780 (1978); R. P. Gaida, Y. B. Klyuchkouskii and V. I. Tretyak, Teor. Mat. Fiz. **44**, 194 (1980).
[4] P. A. M. Dirac, Rev. Mod. Phys. **21**, 392 (1949).
[5] X. Jaen, A. Molina, and V. Iranzo, J. Math. Phys. **27**, 512 (1986).
[6] X. Jaen, J. Llosa, and F. Marques, J. Math. Phys. **27**, 519 (1986).
[7] S. de Bievre, J. Math. Phys. **27**, 7 (1986).
[8] C. G. Darwin, Philos. Mag. **39**, 537 (1920).
[9] A. Einstein, L. Infeld, and B. Hoffman, Ann. Math. **39**, 65 (1938).
[10] F. Bopp, Ann. Phys. **38**, 345 (1940); **42**, 573 (1943).
[11] G. Breit, Phys. Rev. **34**, 553 (1929); **36**, 383 (1930); **39**, 616 (1932).
[12] J. Martin and J. L. Sanz, J. Math. Phys. **19**, 780 (1978).
[13] J. Martin and J. L. Sanz, J. Math. Phys. **20**, 25 (1979).
[14] V. N. Golubenkov and A. I. Smorodinski, Zh. Eksp. Teor. Fiz. **4**, 442 (1957) [Sov. Phys. JETP **31**, 330 (1958)].
[15] A. D. Fokker, Physica **58**, 386 (1929).
[16] E. H. Kerner, J. Math. Phys. **3**, 35 (1962).
[17] R. P. Gaida and V. I. Tetriak, Acta Phys. Pol. B **11**, 509 (1980).
[18] L. Bel, A. Salas, and J. M. Sanchez Ron, Phys. Rev. D **7**, 1099 (1973); R. Lapiedra and A. Molina, J. Math. Phys. **20**, 1308 (1979).
[19] D. G. Currie, Phys. Rev. **142**, 817 (1966); R. N. Hill, J. Math. Phys. **8**, 201 (1967).
[20] L. Bel, Ann. Inst. H. Poincaré **XII**, 307 (1970).

The concept of a time-of-sojourn operator and spreading of wave packets

W. Jaworski[a)]
Department of Chemistry, Queen's University, Kingston, Ontario, K7L 3N6 Canada

(Received 28 July 1987; accepted for publication 8 February 1989)

The concept of sojourn time and a sojourn time operator, aimed at describing the length of time spent by a quantum mechanical system in a given subspace of states, is investigated. A general rigorous definition to the sojourn time operator is given and some of its properties are studied. In particular, it is shown that the usual Born's probability interpretation of the associated spectral measure yields strange, if not paradoxical, results, resembling the well-known quantum mechanical Zeno paradox. Also a specific example of a free nonrelativistic particle is considered. Here it is proven that the probability $P_t(\Omega)$ of the particle being present in a volume Ω at time t cannot vanish on a set of t having nonzero measure. This implies that the sojourn time in Ω never vanishes, and that zero is never an eigenvalue of the sojourn time operator. It is also shown that for a very general class of sets Ω, including all bounded sets, the sojourn time turns out to be bounded with a bound independent of the initial state of the particle. Correspondingly, the sojourn time operator turns out to be a bounded one.

I. INTRODUCTION

Let $\Psi \in L^2(\mathbb{R}^3)$ be a wave function describing an initial state of a quantum mechanical particle with Hamiltonian H. When $\Omega \subseteq \mathbb{R}^3$ is an arbitrary Borel subset, then the quantity

$$\tau_q(\Omega,t_1,t_2;\Psi) = \int_{t_1}^{t_2} dt \int_\Omega dx |\Psi_t(x)|^2 = \int_{t_1}^{t_2} dt \, \|E_q(\Omega)\Psi_t\|^2 \tag{1.1}$$

is usually interpreted as the mean sojourn time of the particle in volume Ω during the time interval (t_1,t_2). Here $-\infty \leqslant t_1 < t_2 \leqslant \infty$, $\Psi_t = \exp(-itH)\Psi$ ($\hbar = 1$), and E_q denotes the joint spectral measure of the position operators $q = (q_1,q_2,q_3)$.

Generally, one can consider an arbitrary quantum mechanical system with Hilbert space $\mathscr{H}$ and Hamiltonian H. Let $A = (A_1,...,A_n)$ be a set of commuting observables in $\mathscr{H}$ and let $E_A = E_{A_1} \times \cdots \times E_{A_n}$ be their joint spectral measure. When $\Omega \subseteq \mathbb{R}^n$ is an arbitrary Borel subset, then the quantity

$$\tau_A(\Omega,t_1,t_2;\Psi) = \int_{t_1}^{t_2} \|E_A(\Omega)\Psi_t\|^2 dt, \tag{1.2}$$

can again be interpreted as a sojourn time—the time spent by the system with the values of the observables $A = (A_1,...,A_n)$ remaining in Ω.

The origin of formulas (1.1) and (1.2) and their interpretation in quantum theory can be looked upon twofold. The first approach makes use of an analogy with classical statistical mechanics or, more generally, with the theory of stochastic processes. When $f(t) = (f_1(t),...f_n(t))$ is a stochastic process [t–time, $f(t)$–real-valued random variables] and when $P'_t(\Omega)$ is the probability that at time t the value of $f(t)$ belongs to a set Ω, then

$$\tau'(\Omega,t_1,t_2) = \int_{t_1}^{t_2} P'_t(\Omega) dt \tag{1.3}$$

can be easily shown to be the proper formula for the mean sojourn time of $f(t)$ in Ω. In fact, $\tau'(\Omega,t_1,t_2)$ is an average of

$$\int_{t_1}^{t_2} \chi_\Omega(f(t)) dt$$

over an ensemble of trajectories $f(t)$. Here χ_Ω is the characteristic function of the set Ω. Thus if one assumes the existence of trajectories of a quantum particle in (1.1) or the existence of trajectories (maybe discontinuous) for the observables $A_1,...,A_n$ in (1.2), then the interpretation of (1.1) or (1.2) as mean sojourn time seems to be inescapable. The trajectory assumption is certainly in agreement with Feynman's approach to quantum mechanics.[1] It is also in the spirit of the theory of stochastic mechanics.[2,3] In general, however, it is well-known to be questionable. Note that in (1.1) and (1.2) we deal with the system undisturbed by observation.

The second approach, adopted by Ekstein and Siegert[4] in connection with the theory of decay of unstable states, consists in constructing a sojourn-time operator as a quantum image of the corresponding classical quantity. In classical mechanics the sojourn time for observables $A = (A_1,...,A_n)$ in volume Ω during the time interval (t_1,t_2) is

$$\int_{t_1}^{t_2} \chi_\Omega(A(t)) dt. \tag{1.4}$$

The corresponding quantum mechanical operator reads

$$T_A(\Omega,t_1,t_2) = \int_{t_1}^{t_2} \exp(itH) E_A(\Omega) \exp(-itH) dt \tag{1.5}$$

and the translation of (1.4) into (1.5) seems to be unambiguous. The mean values $\langle\Psi|T_A(\Omega,t_1,t_2)\Psi\rangle$ are equal to $\tau_A(\Omega,t_1,t_2;\Psi)$. When $t_2 - t_1 < \infty$, expression (1.5) is well-defined via the Bochner integral theory. However, one must be careful in the case $t_2 - t_1 = \infty$ when the integral is not absolutely convergent.

In Sec. II we give a rigorous definition to the sojourn time operator $T_A(\Omega,t_1,t_2)$. In general, it cannot be claimed to be densely defined in the whole $\mathscr{H}$. It is, however, a self-

[a)] On leave from Institute of Physics, Nicholas Copernicus University, Torun, Poland.

 0022-2488/89/071505-10$02.50

adjoint operator in a closed subspace of $\mathscr{H}$ which reduces the Hamiltonian. This subspace can be equal to $\mathscr{H}$—see the example of a free particle in Sec. IV.

$T_A(\Omega,t_1,t_2)$ as a self-adjoint operator has a spectral resolution

$$T_A(\Omega,t_1,t_2)=\int_{\mathbb{R}}\lambda F_A(\Omega,t_1,t_2;d\lambda). \tag{1.6}$$

The usual interpretation of the spectral measure $F_A(\Omega,t_1,t_2;\cdot)$, according to the Born rule, would be the following: When Ψ is a state of the system, then $\langle\Psi|F_A(\Omega,t_1,t_2;\Delta)\Psi\rangle$ is the probability that the sojourn time will take a value from the set $\Delta\subseteq\mathbb{R}$ (in the course of the time evolution Ψ_t of the state $\Psi=\Psi_0$). It must be stressed that what we are using here is, in fact, an extension of the Born rule since, conventionally, the latter applies to operators representing instantaneous observables (measurements). The sojourn time is certainly not of this kind. It corresponds rather to a continuous observation in the limit of weak disturbance of the system by the measurement procedure.[5]

Classically, a measuring device for the sojourn time (1.1) would be a sensor that allows a clock to run when the particle is inside Ω. The sensor should be sufficiently gentle that it does not alter the motion of the particle. Whether a quantum mechanical analog of such a device can be constructed is a challenging question which is, however, outside the scope of this paper. Here let us only mention the idea of a spin clock of Baz.[6,7] Let us also remark that the concept of sojourn time plays an important role in a rigorous definition of time delay[8,9] in scattering theory. There is no doubt that the latter quantity is, at least in principle, measurable. The time delay operator is, essentially, a difference of two sojourn-time operators—one for a particle interacting with a scattering center, and one for a free particle.

Finally, one can also look at the quantities $\tau_A(\Omega,t_1,t_2;\Psi)$ and $T_A(\Omega,t_1,t_2)$ as at a sort of "ideal" quantities describing the inner continually existent quantum world—when one is inclined to believe in such an existence.

With all the qualifications made above, the interpretation of $\langle\Psi|F_A(\Omega,t_1,t_2;\Delta)\Psi\rangle$ in terms of probabilities is, nevertheless, attractive, and merits some attention. Thus, for example, $\langle\Psi|F_A(\Omega,t_1,t_2;(0,\infty))\Psi\rangle$ is the probability that the values of the observables $A=(A_1,\ldots,A_n)$ can be found in Ω for a nonzero fraction of time, or that the "trajectory" of A enters Ω for a nonzero fraction of time. Then, $\langle\Psi|F_A(\Omega,t_1,t_2;\{t_2-t_1\})\Psi\rangle$ $(t_2-t_1<\infty)$ would be the probability that the trajectory stays in Ω for almost all $t\in(t_1,t_2)$. Finally, $\langle\Psi|F_A(\Omega,t_1,t_2;\{0\})\Psi\rangle$ is the probability that the trajectory will not enter Ω during the time interval (t_1,t_2), except maybe for a set of times $t\in(t_1,t_2)$ having measure zero.

In the following we will rather adopt the equivalent language of the theory of unstable states. $\mathscr{M}=E_A(\Omega)\mathscr{H}$ will be designated as the subspace of undecayed states, and $\mathscr{M}^\perp=E_A(\mathbb{R}^n\setminus\Omega)\mathscr{H}$ as the subspace of decay products. Then, e.g., $\langle\Psi|F_A(\mathbb{R}^n\setminus\Omega,\,t_1,t_2;\{0\})\Psi\rangle$ or $\langle\Psi|F_A(\Omega,t_1,t_2;\{t_2-t_1\})\Psi\rangle$ is the probability that the system is undecayed during the time interval (t_1,t_2) (except maybe for a set of t having measure zero). $\langle\Psi|F_A(\mathbb{R}^n\setminus\Omega,t_1,t_2;(0,\infty))\Psi\rangle$ is the probability that the system will stay decayed for a nonzero fraction of time $t\in(t_1,t_2)$.

We will show that the properties of all the above probabilities are striking, if not paradoxical.

Our paper is organized as follows. In Sec. II we define rigorously the sojourn time operators and point out some of their properties, in particular, the commutation relations with the Hamiltonian. In Sec. III we study the spectral projectors $F_A(\Omega,t_1,t_2;\{0\})$, $F_A(\Omega,t_1,t_2;\{t_2-t_1\})$, $F_A(\Omega,t_1,t_2;(0,\infty))$, and the corresponding probabilities. Section IV is devoted to the specific example of a free particle $[\mathscr{H}=L^2(\mathbb{R}^n),A=q=(q_1,\ldots,q_n),H=H_0=-\Delta/2]$.

II. DEFINITION OF THE SOJOURN TIME OPERATORS

To simplify the notation a little bit we will write

$$\tau(\mathscr{M},t_1,t_2;\Psi)=\int_{t_1}^{t_2}\|P_{\mathscr{M}}\Psi_t\|^2\,dt, \tag{2.1}$$

where $\mathscr{M}\subseteq\mathscr{H}$ is a closed subspace of the Hilbert space $\mathscr{H}$, and $P_{\mathscr{M}}$ is the corresponding projector. We will also write

$$\tau(\mathscr{M},t_1,t_2;\Phi,\Psi)=\int_{t_1}^{t_2}\langle P_{\mathscr{M}}\Phi_t|\,P_{\mathscr{M}}\Psi_t\rangle dt. \tag{2.2}$$

When $t_2-t_1<\infty$, then the integrals (2.1) and (2.2) are always finite, and $\tau(\Omega,t_1,t_2;\Phi,\Psi)$ as a function of $(\Phi,\Psi)\in\mathscr{H}\times\mathscr{H}$ is a bounded sesquilinear form. Since there is a one-to-one correspondence between such forms and bounded operators in $\mathscr{H}$, we have immediately the following theorem.

Theorem 2.1: Let $(t_1,t_2)\subseteq\mathbb{R}$ be a bounded interval. There is then a unique self-adjoint operator $T(\mathscr{M},t_1,t_2)$: $\mathscr{H}\to\mathscr{H}$ such that

$$\langle\Psi|T(\mathscr{M},t_1,t_2)\Psi\rangle=\tau(\mathscr{M},t_1,t_2;\Psi) \tag{2.3}$$

for all $\Psi\in\mathscr{H}$.

The case $t_2-t_1=\infty$ is a little more subtle, although it can be easily dealt with using the theory of bounded below, symmetric, sesquilinear forms.[10]

Lemma 2.2: Let $(t_1,t_2)\subseteq\mathbb{R}$ be an unbounded interval and let

$$\mathscr{H}_0=\left\{\Psi\in\mathscr{H}\,\middle|\,\int_{t_1}^{t_2}\|P_{\mathscr{M}}\Psi_t\|^2\,dt<\infty\right\}. \tag{2.4}$$

Then $\mathscr{H}_0$ is a linear subspace of $\mathscr{H}$. The sesquilinear form

$$\mathscr{H}_0\times\mathscr{H}_0\ni(\Phi,\Psi)\to\tau(\mathscr{M},t_1,t_2;\Phi,\Psi)$$
$$=\int_{t_1}^{t_2}\langle P_{\mathscr{M}}\Phi_t|\,P_{\mathscr{M}}\Psi_t\rangle dt \tag{2.5}$$

is well-defined on $\mathscr{H}_0\times\mathscr{H}_0$. Moreover, it is positive and closed.

Proof: The linearity of $\mathscr{H}_0$ and the well-definiteness of $\tau(\mathscr{M},t_1,t_2;\Phi,\Psi)$ follow immediately from Schwarz inequality. Positivity is trivial. It remains to check that $\tau(\mathscr{M},t_1,t_2;\Phi,\Psi)$ is a closed form. Let $\Psi^{(n)}\in\mathscr{H}_0$ be a sequence such that $\Psi^{(n)}\underset{n\to\infty}{\to}\Psi\in\mathscr{H}$, and

$$\tau(\mathscr{M},t_1,t_2;\Psi^{(n)}-\Psi^{(m)})\underset{m,n\to\infty}{\to}0.$$

We are to show that $\Psi\in\mathscr{H}_0$ and

$$\tau(\mathscr{M},t_1,t_2;\Psi^{(n)}-\Psi)\underset{n\to\infty}{\to}0.$$

For any $\epsilon>0$ and sufficiently large m,n: $m,n>N_\epsilon$, we have

$$\int_{t_1}^{t_2}\|P_{\mathscr{M}}(\Psi_t^{(n)}-\Psi_t^{(m)})\|^2\,dt\leqslant\epsilon. \tag{2.6}$$

By the Fatou lemma,

$$\begin{aligned}\int_{t_1}^{t_2}&\|P_{\mathscr{M}}(\Psi_t^{(n)}-\Psi_t)\|^2\,dt\\&=\int_{t_1}^{t_2}\lim_{m\to\infty}\|P_{\mathscr{M}}(\Psi_t^{(n)}-\Psi_t^{(m)})\|^2\,dt\\&\leqslant\liminf_{m\to\infty}\int_{t_1}^{t_2}\|P_{\mathscr{M}}(\Psi_t^{(n)}-\Psi_t^{(m)})\|^2\,dt\leqslant\epsilon.\end{aligned} \tag{2.7}$$

This proves that $\Psi\in\mathscr{H}_0$, and $\tau(\mathscr{M},t_1,t_2;\Psi^{(n)}-\Psi)\underset{n\to\infty}{\to}0$. □

As for the subspace $\mathscr{H}_0$, it need not be dense in $\mathscr{H}$. One can easily construct an example of this when the point spectrum of H is nonempty.

Using the first representation theorem for bounded below, closed sesquilinear forms[10] we immediately obtain the following theorem.

Theorem 2.3: With the notation and assumptions of Lemma 2.2, there is a unique self-adjoint operator $T(\mathscr{M},t_1,t_2)$ acting in the Hilbert space $\overline{\mathscr{H}}_0\subseteq\mathscr{H}$ with domain of definition $D(T(\mathscr{M},t_1,t_2))\subseteq\mathscr{H}_0$ such that $D(T(\mathscr{M},t_1,t_2))$ is a core of the sesquilinear form $\tau(\mathscr{M},t_1,t_2;\Phi;\Psi)$ and

$$\tau(\mathscr{M},t_1,t_2;\Psi)=\langle\Psi|T(\mathscr{M},t_1,t_2)\Psi\rangle \tag{2.8}$$

for all $\Psi\in D(T(\mathscr{M},t_1,t_2))$.

$T(\mathscr{M},t_1,t_2)$ is the unique self-adjoint operator acting in $\overline{\mathscr{H}}_0$ with domain $D(T(\mathscr{M},t_1,t_2))$ satisfying $D(T(\mathscr{M},t_1,t_2))\subseteq\mathscr{H}_0$ and such that

$$\langle\Phi|T(\mathscr{M},t_1,t_2)\Psi\rangle=\tau(\mathscr{M},t_1,t_2;\Phi,\Psi), \tag{2.9}$$

for all $\Phi\in\mathscr{H}_0$ and $\Psi\in D(T(\mathscr{M},t_1,t_2))$.

Theorems 2.1 and 2.3 provide a general rigorous definition to the sojourn time operators. They assign a precise meaning to the formal expression (1.5). For $t_2-t_1<\infty$ the Bochner integral

$$\int_{t_1}^{t_2}\exp(itH)E_A(\Omega)\exp(-itH)\Psi\,dt,\quad\Psi\in\mathscr{H},$$

evidently yields the same result.

Although the subspace $\overline{\mathscr{H}}_0$ need not be equal to $\mathscr{H}$, it has the remarkable property of reducing the Hamiltonian H. More precisely, we have the following result.

Remark 2.4: $\exp(itH)\mathscr{H}_0=\mathscr{H}_0$ and $\exp(itH)\overline{\mathscr{H}}_0=\overline{\mathscr{H}}_0$ for all $t\in\mathbb{R}$. Hence $\overline{\mathscr{H}}_0$ reduces H.

Proof: Inclusion $\exp(itH)\mathscr{H}_0\subseteq\mathscr{H}_0$ can be easily deduced from the definition of $\mathscr{H}_0$ in Lemma 2.2. Since t is arbitrary, the above equalities follow. Reduction of H is then a well-known fact. □

The following statements are immediate consequences of Theorems 2.1, 2.3, and the definition of $\tau(\mathscr{M},t_1,t_2;\Psi)$.

Theorem 2.5: (a) $T(\mathscr{M},t_1,t_2)\leqslant T(\mathscr{M},t_1',t_2')\leqslant t_2'-t_1'$, for $-\infty<t_1'\leqslant t_1<t_2\leqslant t_2'<\infty$.

(b) $T(\mathscr{M},t_1,t_2)+T(\mathscr{M}^\perp,t_1,t_2)=t_2-t_1$, for $-\infty<t_1<t_2<\infty$.

(c) $T(\mathscr{M},t_1,t_2)+T(\mathscr{M},t_2,t_3)=T(\mathscr{M},t_1,t_3)$, for $-\infty<t_1<t_2<t_3\leqslant\infty$, or $-\infty\leqslant t_1<t_2<t_3<\infty$.

(d) $\exp(itH)T(\mathscr{M},t_1,t_2)\exp(-itH)=T(\mathscr{M},t_1+t,t_2+t)$, for $-\infty\leqslant t_1<t_2\leqslant\infty$, and $t\in\mathbb{R}$.

Combining (c) and (d) we get the following supplement to Remark 2.4.

Remark 2.6: $\exp(itH)D(T(\mathscr{M},t_1,t_2))=D(T(\mathscr{M},t_1,t_2))$ for any choice of $-\infty\leqslant t_1<t_2\leqslant\infty$, and $t\in\mathbb{R}$.

Point (d) of Theorem 2.5 constitutes in fact a commutation relation. *Thus for $t_1=-\infty$, $t_2=\infty$, the operators $T(\mathscr{M},-\infty,\infty)$ and H do commute.* This is not the case when $t_1\neq-\infty$ or $t_2\neq\infty$. To put the commutation relations into a more conventional form we note the following result.

Remark 2.7: The operator valued function

$$\{(x,y)\in\mathbb{R}^2|x<y\}\ni(t_1,t_2)\to T(\mathscr{M},t_1,t_2)$$

is weakly continuous. Moreover, the derivatives $\partial T(\mathscr{M},t_1,t_2)/\partial t_1$ and $\partial T(\mathscr{M},t_1,t_2)/\partial t_2$ exist in the sense of weak topology, and

$$w-\partial T(\mathscr{M},t_1,t_2)/\partial t_1=-\exp(it_1H)P_{\mathscr{M}}\exp(-it_1H), \tag{2.10}$$

$$w-\partial T(\mathscr{M},t_1,t_2)/\partial t_2=\exp(it_2H)P_{\mathscr{M}}\exp(-it_2H). \tag{2.11}$$

Now, appropriately differentiating formula (d) of Theorem 2.5, we obtain the following theorem.

Theorem 2.8: (a) $i[H,T(\mathscr{M},t_1,t_2)]=\exp(it_2H)P_{\mathscr{M}}\times\exp(-it_2H)-\exp(it_1H)P_{\mathscr{M}}\exp(-it_1H)$, for $-\infty<t_1<t_2<\infty$;

(b) $i[H,T(\mathscr{M},t_1,\infty)]=-P_{\overline{\mathscr{H}}_0}\exp(it_1H)P_{\mathscr{M}}\times\exp(-it_1H)$, for $|t_1|<\infty$,

(c) $i[H,T(\mathscr{M},-\infty,t_2)]=P_{\overline{\mathscr{H}}_0}\exp(it_2H)P_{\mathscr{M}}\times\exp(-it_2H)$, for $|t_2|<\infty$, where equalities (b) and (c) hold under the condition that $D(T(\mathscr{M},t_1,t_2)H)\cap D(T(\mathscr{M},t_1,t_2))$ is dense in $\overline{D(T(\mathscr{M},t_1,t_2))}=\overline{\mathscr{H}}_0$. $P_{\overline{\mathscr{H}}_0}$ is the projector onto $\overline{\mathscr{H}}_0$.

Proof: The case $t_2-t_1<\infty$ does not present any problems. However, one must be careful when $t_2-t_1=\infty$, because of the possible unboundedness of the operator $T(\mathscr{M},t_1,\infty)$ or $T(\mathscr{M},-\infty,t_2)$. Let us prove (b). For $\Phi\in D(T(\mathscr{M},t_1,\infty)H)\cap D(T(\mathscr{M},t_1,\infty))$ and $\Psi\in D(T(\mathscr{M},t_1,\infty)H)\cap D(HT(\mathscr{M},t_1,\infty))$ we have

$$\begin{aligned}t^{-1}[&\langle\Phi|\exp(itH)T(\mathscr{M},t_1,\infty)\exp(-itH)\Psi\rangle\\&-\langle\Phi|T(\mathscr{M},t_1,\infty)\Psi\rangle]\\&=\langle[iH+(\exp(-itH)-1)/t]\Phi|T(\mathscr{M},t_1,\infty)\\&\quad\times\exp(-itH)\Psi\rangle+\langle-iT(\mathscr{M},t_1,\infty)H\Phi|\\&\quad\times\exp(-itH)\Psi\rangle+\langle T(\mathscr{M},t_1,\infty)\Phi|\\&\quad\times[(\exp(-itH)-1)/t]\Psi\rangle.\end{aligned} \tag{2.12}$$

To deal with the first term in the limit $t\to0$, we apply the Schwarz inequality and relations (c) and (d) of Theorem 2.5:

$$\begin{aligned}&\langle[iH+(\exp(-itH)-1)/t]\Phi|\\&\quad\times\exp(-itH)T(\mathscr{M},t_1+t,\infty)\Psi\rangle\\&\leqslant\|[iH+(\exp(-itH)-1)/t]\Phi\|\\&\quad\times\|T(\mathscr{M},t_1+t,\infty)\Psi\|\end{aligned}$$

$$\leqslant \|[iH + (\exp(-itH) - 1)/t]\Phi\| \times (\|T(\mathscr{M},t_1 + t,t_0)\Psi\| + \|T(\mathscr{M},t_0,\infty)\Psi\|)$$
$$\leqslant \|[iH + (\exp(-itH) - 1)/t]\Phi\| \times (\|T(\mathscr{M},t_1 + t,t_0)\| \, \|\Psi\| + \|T(\mathscr{M},t_0,\infty)\Psi\|). \quad (2.13)$$

Here $t_1 < t_0 < \infty$. Since inequality (a) of Theorem 2.5 implies $\|T(\mathscr{M},t_1 + t,t_0)\| \leqslant t_0 - t_1 - t$ for sufficiently small t, we infer that the first term in (2.12) vanishes in the limit $t \to 0$. The remaining terms produce

$$\langle -iT(\mathscr{M},t_1,\infty)H\Phi|\Psi\rangle + \langle iT(\mathscr{M},t_1,\infty)\Phi| - iH\Psi\rangle = i\langle\Phi|[H,T(\mathscr{M},t_1,\infty)]\Psi\rangle. \quad (2.14)$$

Calculating the same limit $t \to 0$ for the right-hand side of (d) of Theorem 2.5, using point (c) of this theorem and (2.10), we get

$$\langle\Phi| - \exp(it_1H)P_{\mathscr{M}} \exp(-it_1H)\Psi\rangle. \quad (2.15)$$

Since Φ belongs to $D(T(\mathscr{M},t_1,\infty)H) \cap D(T(\mathscr{M},t_1,\infty))$ which is assumed to be dense in $\mathscr{H}_0$, and since $\mathscr{H}_0$ reduces H, then by comparing (2.15) and (2.14) we arrive at (b). The proof of (c) is evidently analogous. □

The commutation relations of Theorem 2.8 yield uncertainty relations between the sojourn time and energy expressible in terms of mean square deviations (for sufficiently regular state vectors). Only for a special choice of state vectors are these uncertainty relations of the familiar form of time-energy uncertainty relations.

III. SPECTRAL MEASURES OF THE SOJOURN-TIME OPERATORS

The spectral measure associated with the sojourn time operator $T(\mathscr{M},t_1,t_2)$ of Theorem 2.1 or 2.3 will be denoted by $F(\mathscr{M},t_1,t_2;\cdot)$, i.e.,

$$T(\mathscr{M},t_1,t_2) = \int_{\mathbb{R}} \lambda F(\mathscr{M},t_1,t_2;d\lambda). \quad (3.1)$$

We proceed to investigate the properties of the spectral projectors $F(\mathscr{M},t_1,t_2;\{0\})$, $F(\mathscr{M},t_1,t_2;(0,\infty))$, $F(\mathscr{M},t_1,t_2;\{t_2 - t_1\})$, and the corresponding probabilities $\langle\Psi|F(\mathscr{M},t_1,t_2;\cdot)\Psi\rangle$. The main tool to this end is the following lemma.

Lemma 3.1: Let H be a bounded below self-adjoint operator in a Hilbert space $\mathscr{H}$ and let $P: \mathscr{H} \to \mathscr{H}$ be a projector. Let $\Psi \in \mathscr{H}$ and suppose that the set

$$A_\Psi = \{t \in \mathbb{R} | P\exp(-itH)\Psi = 0\} \quad (3.2)$$

has nonzero measure. Then $A_\Psi = \mathbb{R}$.

Proof: Let $\mathscr{N} = P\mathscr{H}$ and let $\{\varphi_\beta\}_{\beta \in B}$ be an orthonormal basis in $\mathscr{N}$. Fix $\beta \in B$ and consider the function

$$G_\beta(z) = [1/(z - i)]\langle\varphi_\beta|\exp[-iz(H - \zeta)]\Psi\rangle = [\exp(iz\zeta)/(z - i)]\langle\varphi_\beta|\exp(-izH)\Psi\rangle, \quad (3.3)$$

where $z \in \mathbb{C}$, $\operatorname{Im} z \leqslant 0$, and $\zeta \in \mathbb{R}$ is chosen so that $H - \zeta \geqslant 0$. The function $G_\beta(z)$ can be seen to be analytic in the open lower half-plane and continuous in the closed lower half-plane. Moreover,

$$\int_{\mathbb{R}} |G_\beta(x + iy)|^2 \, dx \leqslant \|\varphi_\beta\| \, \|\Psi\| \int_{\mathbb{R}} \frac{dx}{x^2 + (1 - y)^2} \leqslant \|\varphi_\beta\| \, \|\Psi\| \pi, \quad (3.4)$$

for $y < 0$. This means that G_β belongs to the class of Hardy H^{2-} functions.[11,12] An important property of a H^{2-} function f is that the limit

$$f(x) = \lim_{y \to -0} f(x + iy)$$

exists in the sense of $L^2(\mathbb{R})$, i.e.,

$$\lim_{y \to -0} \int_{\mathbb{R}} |f(x) - f(x + iy)|^2 \, dx = 0,$$

and $f(x) \neq 0$ almost everywhere, unless $f(z) = 0$ for all z, $\operatorname{Im} z < 0$. In our case, the above limit, due to continuity of G_β, is equal to

$$G_\beta(x) = [\exp(i\zeta x)/(x - i)]\langle\varphi_\beta|\exp(-ixH)\Psi\rangle. \quad (3.5)$$

At the same time,

$$\langle\varphi_\beta|\exp(-ixH)\Psi\rangle = \langle P\varphi_\beta|\exp(-ixH)\Psi\rangle = \langle\varphi_\beta|P\exp(-ixH)\Psi\rangle = 0 \quad (3.6)$$

for $x \in A_\Psi$. Since A_Ψ has a nonzero measure, it follows that $G_\beta \equiv 0$ and, consequently, $\langle\varphi_\beta|\exp(-itH)\Psi\rangle = 0$ for all $t \in \mathbb{R}$.

Since β is arbitrary, to finish the proof it is enough to notice that

$$P\exp(-itH)\Psi = \sum_{\beta \in B} \langle\varphi_\beta|\exp(-itH)\Psi\rangle\varphi_\beta. \quad (3.7)$$

□

$F(\mathscr{M},t_1,t_2;\{0\})\mathscr{H}$ and $F(\mathscr{M},t_1,t_2;\{t_2 - t_1\})\mathscr{H}$ ($t_2 - t_1 < \infty$ in the latter case) are eigenspaces of $T(\mathscr{M},t_1,t_2)$ to the eigenvalues 0 and $t_2 - t_1$, respectively. By Theorem 2.5(b), we have

$$F(\mathscr{M},t_1,t_2;\{0\}) = F(\mathscr{M}^\perp,t_1,t_2;\{t_2 - t_1\}).$$

Also,

$$F(\mathscr{M},t_1,t_2;(0,\infty)) = 1 - F(\mathscr{M},t_1,t_2;\{0\}).$$

Therefore it is enough to investigate the projector $F(\mathscr{M},t_1,t_2;\{0\})$ ($t_2 - t_1 \leqslant \infty$).

Lemma 3.2: $T(\mathscr{M},t_1,t_2)\Psi = 0$ if and only if $\tau(\mathscr{M},t_1,t_2;\Psi) = 0$.

Proof: The implication $\Rightarrow$ is trivial. To prove $\Leftarrow$ in the case $t_2 - t_1 < \infty$ we note that

$$\tau(\mathscr{M},t_1,t_2;\Psi) = \langle\Psi|T(\mathscr{M},t_1,t_2)\Psi\rangle = \|\sqrt{T(\mathscr{M},t_1,t_2)}\,\Psi\|^2 = 0$$

implies $T(\mathscr{M},t_1,t_2)\Psi = 0$. In the case $t_2 - t_1 = \infty$ one must be certain that $\Psi \in D(T(\mathscr{M},t_1,t_2))$ when $\tau(\mathscr{M},t_1,t_2;\Psi) = 0$. But this follows from the second representation theorem for sesquilinear forms[10]: The domain $\mathscr{H}_0$ of the form $\tau(\mathscr{M},t_1,t_2;\Phi,\Psi)$ turns out to be equal to the domain of the operator $\sqrt{T(\mathscr{M},t_1,t_2)}$, and

$$\tau(\mathscr{M},t_1,t_2;\Phi,\Psi) = \langle\sqrt{T(\mathscr{M},t_1,t_2)}\,\Phi|\sqrt{T(\mathscr{M},t_1,t_2)}\,\Psi\rangle.$$

Hence we can use the same argument as in the case $t_2 - t_1 < \infty$. □

As an immediate consequence of Lemmas 3.1 and 3.2 we have the following theorem.

Theorem 3.3: Let $T(\mathscr{M},\bar{t}_1,\bar{t}_2)\Psi = 0$ for some $-\infty \leqslant \bar{t}_1 < \bar{t}_2 \leqslant \infty$. Then $T(\mathscr{M},t_1,t_2)\exp(-itH)\Psi = 0$ for all $-\infty \leqslant t_1 < t_2 \leqslant \infty$, and $t \in \mathbb{R}$.

Theorem 3.4: (a) $F(\mathscr{M},t_1,t_2;\{0\}) = F(\mathscr{M},t'_1,t'_2;\{0\})$ for all $-\infty \leqslant t_1 < t_2 \leqslant \infty$, $-\infty \leqslant t'_1 < t'_2 \leqslant \infty$.

(b) $\exp(itH)\, F(\mathscr{M},t_1,t_2;\{0\})\exp(-itH) = F(\mathscr{M},t_1,t_2;\{0\})$, for all $-\infty \leqslant t_1 < t_2 \leqslant \infty$, and $t \in \mathbb{R}$. Hence the operators H and $F(\mathscr{M},t_1,t_2;\{0\})$ commute.

Let us now shortly discuss the physical implications of the above results. This will be done in the language of the theory of unstable states. Let us designate $\mathscr{M} \subseteq \mathscr{H}$ as the subspace of undecayed states, and $\mathscr{M}^\perp$ as the subspace of decay products. The physical message of Lemma 3.1 is: If $\Psi \in \mathscr{H}$ is an initial state ($t = 0$) and if the system is undecayed $[P_{\mathscr{M}^\perp}\exp(-itH)\Psi = 0]$ for a nonzero fraction of times $t > 0$, then the system will not decay at all $[P_{\mathscr{M}^\perp}\exp(-itH)\Psi = 0$ for all $t]$.

In Sec. I we already discussed the limitations of the interpretation of $\langle\Psi|F(\mathscr{M},t_1,t_2;\cdot)\Psi\rangle$ as probabilities. When we, nevertheless, assume such an interpretation, then $\langle\Psi|F(\mathscr{M}^\perp,t_1,t_2;\{0\})\Psi\rangle$ is the probability that the system will be undecayed during the time interval (t_1,t_2), except maybe for a set of $t \in (t_1,t_2)$ having measure zero. Remarkably, by Theorem 3.4(a), this probability is independent of the time interval (t_1,t_2). In particular, when it is equal to 1 for some interval $(\bar{t}_1,\bar{t}_2)$ then it is equal to one for all intervals (t_1,t_2), i.e., the system will never decay at all.

The results described in the preceding two paragraphs can be recognized as a kind of a quantum mechanical Zeno paradox. The well-known form of this paradox[13] is obtained when the system is theoretically considered as being under continuous observation treated as a limit of infinitely densely spaced instantaneous measurements, each causing a collapse of the wave function. The result then is that the system undecayed initially ($\Psi = \Psi_0 \in \mathscr{M}$) will never be found to decay. It is now interesting to note, that a kind of Zeno's paradox persists in the opposite extreme situation when the system is essentially undisturbed by observation.

IV. AN EXAMPLE—A FREE NONRELATIVISTIC PARTICLE

The free particle is certainly one of the simplest quantum mechanical systems. Nevertheless, its mathematical analysis is not completely trivial, and there does not exist a single, simple, explicit formula from which all the physically interesting properties can be deduced with ease (as is the case with the classical counterpart). Here using the concept of the sojourn time and the sojourn time operator we want to shed some more light onto the well-known phenomenon of spreading of wave packets.

We put $\mathscr{H} = L^2(\mathbb{R}^n)$, $H = H_0 = -\Delta/2$, $\mathscr{M} = E_q(\Omega)L^2(\mathbb{R}^n)$, where $\Omega \subseteq \mathbb{R}^n$ is a Borel subset and E_q is the joint spectral measure of the position operators $q = (q_1,\ldots,q_n)$.

The spreading can be formulated as follows.[14] Let $\Omega \subseteq \mathbb{R}^n$ be an arbitrary Borel subset of finite measure, and let $v \in \mathbb{R}^n$. Then for any $\Psi \in \mathbb{R}^n$, $\|\Psi\| = 1$, the probability $P_t(\Omega + vt,\Psi) = \|E(\Omega + vt)\Psi_t\|^2$ of finding the particle in the set $\Omega + vt$ at time t, vanishes as $t \to \pm\infty$,

$$P_t(\Omega + vt,\Psi) = \|E(\Omega + vt)\Psi_t\|^2 = \int_{\Omega + vt} |\Psi_t(x)|^2\, dx \underset{t\to\pm\infty}{\to} 0. \tag{4.1}$$

Formula (4.1) says that in contrast with classical theory, spreading is inevitable in quantum mechanics. It can be intuitively perceived as a manifestation of the uncertainty principle. A more detailed description of this phenomenon and its consequences is certainly of physical interest.

In Appendix A we give a proof to the following theorem.

Theorem 4.1: Let $\Psi \in L^2(\mathbb{R}^n)$, $\Psi \neq 0$ and let $\Omega \subseteq \mathbb{R}^n$ be a Borel subset with nonempty interior. Then the set

$$\{t \in \mathbb{R} | E_q(\Omega)\Psi_t = 0\} = \{t \in \mathbb{R} | P_t(\Omega,\Psi) = 0\} \tag{4.2}$$

has zero Lebesgue measure.

The physical content of the theorem is that for a given Ω one cannot prepare the system (i.e., find an initial state $\Psi = \Psi_0$) in such a way that the particle will with certainty avoid entering Ω for a nonzero fraction of times $t > 0$. In particular, when the particle is initially localized in a set $\Omega_1 \subseteq \mathbb{R}^n$ [$\Psi_0(x) = 0$ outside Ω_1], then the probability distribution $|\Psi_t(x)|^2$ spreads immediately in such a way that there is a nonzero chance of finding the particle in any $\Omega \subseteq \mathbb{R}^n$ (specified in Theorem 4.1) for almost all $t > 0$.[15,16] This can again be intuitively perceived as a manifestation of the uncertainty principle. [Note that for $\Psi(x)$ vanishing outside a bounded Ω_1, the momentum representation wave function $\hat{\Psi}(p)$ is an entire analytic function of $p \in \mathbb{C}^n$. The probability distribution for momentum $|\hat{\Psi}(p)|^2$ cannot therefore vanish on any open subset of $\mathbb{R}^n$, unless $\Psi \equiv 0$.]

Theorem 4.1 implies that the sojourn time $\tau_q(\Omega,t_1,t_2;\Psi)$ is never zero for $\Psi \neq 0$ and Ω with nonempty interior. Thus, for $t_2 - t_1 < \infty$ and $\|\Psi\| = 1$, $\tau_q(\mathbb{R}^n \setminus \Omega,t_1,t_2;\Psi)$ is always strictly smaller than $t_2 - t_1$. This is just another formulation of the impossibility to well-localize the particle in the course of its time evolution.

Another, even more striking result proven in Appendix B is the following theorem.

Theorem 4.2: Let $n \geqslant 2$ and let $\Omega \subseteq \mathbb{R}^n$ be a Borel subset with the property that $\Omega \subseteq A(\mathbb{R}^{n-2} \times W)$, where A: $\mathbb{R}^n \to \mathbb{R}^n$ is an orthogonal transformation and $W \subseteq \mathbb{R}^2$ is Borel of finite two-dimensional Lebesgue measure $\lambda_2(W) < \infty$. Then for any $\Psi \in L^2(\mathbb{R}^n)$

$$\tau_q(\Omega, -\infty, \infty;\Psi) \leqslant \lambda_2(W)\|\Psi\|^2. \tag{4.3}$$

Since the effect of the Planck constant $\hbar$ and the mass m of the particle on the bound in (4.3) is of some interest, we note that

$$\tau(\Omega, -\infty, \infty;\Psi) \leqslant (m/\hbar)\lambda_2(W)\|\Psi\|^2 \tag{4.4}$$

when $\hbar$ and m are explicitly introduced into the theory [i.e., $H = -(\hbar^2/2m)\Delta$, $\Psi_t = \exp(-itH/\hbar)\Psi$]. Theorem 4.2 evidently applies to bounded sets as a particular case.

For quantum states $\|\Psi\| = 1$, and we can see that the

sojourn time is bounded for sets Ω satisfying assumptions of Theorem 4.2. This property is hardly understandable in terms of any classical analogy. If one assumes the existence of trajectories of a quantum particle, then the result (4.3), (4.4) and the nonvanishing of the sojourn time stated before, would be a manifestation of the complexity of such hypothetical trajectories and of their statistical ensemble corresponding to a given wave function Ψ. Let us also note that finiteness of the sojourn time $\tau(\Omega,-\infty,\infty;\Psi)$ demands that the decay of the probability $P_t(\Omega,\Psi)=\|E(\Omega)\Psi_t\|^2$ cannot be too slow.

It must be stressed that the condition $n\geqslant 2$ in the assumptions of Theorem 4.2 is essential. In Appendix B we prove the following theorem.

Theorem 4.3: Let $n=1$ and let $\Omega\subseteq\mathbb{R}$ be Borel of nonzero measure. Suppose that the interval (t_1,t_2) is unbounded. Then

$$\sup_{\substack{\Psi\in L^2(\mathbb{R})\\ \|\Psi\|=1}} \tau_q(\Omega,t_1,t_2;\Omega)=\infty. \tag{4.5}$$

However, for Ω of finite measure $\tau_q(\Omega,t_1,t_2;\Psi)$ is finite for a dense subspace of vectors $\Psi\in L^2(\mathbb{R})$.

We now proceed to discuss some of the properties of the sojourn time operator $T_q(\Omega,t_1,t_2)$.

By Lemma 3.2, nonvanishing of $\tau_q(\Omega,t_1,t_2;\Psi)$ implies the following theorem.

Theorem 4.4: Let $\Omega\subseteq\mathbb{R}^n$ be a Borel subset with nonempty interior. Then for any choice of $-\infty\leqslant t_1<t_2\leqslant\infty$, $F_q(\Omega,t_1,t_2;\{0\})=0$, i.e., there are no eigenstates of $T_q(\Omega,t_1,t_2)$ to the eigenvalue zero.

Theorems 4.2 and 4.3 combined with the representation theorem for sesquilinear forms yield the following theorem.

Theorem 4.5: (a) Let $n\geqslant 2$ and let $\Omega\subseteq\mathbb{R}^n$ be of the form $\Omega=A(\mathbb{R}^{n-2}\times W)$ where $A:\mathbb{R}^n\to\mathbb{R}^n$ is an orthogonal transformation and $W\subseteq\mathbb{R}^2$ has finite Lebesgue measure $\lambda_2(W)<\infty$. Then the sojourn-time operator $T_q(\Omega,t_1,t_2)$ is bounded for any choice of $-\infty\leqslant t_1<t_2\leqslant\infty$, and $\|T_q(\Omega,t_1,t_2)\|\leqslant\lambda_2(W)$ $[\|T_q(\Omega,t_1,t_2)\|\leqslant(m/\hbar)\lambda_2(W)]$.

(b) Let $n=1$, let Ω be Borel of finite measure, and let $t_2-t_1=\infty$. Then the sojourn time operator $T_q(\Omega,t_1,t_2)$ is densely defined in $L^2(\mathbb{R})$, but unbounded.

If we adopt the interpretation of $\langle\Psi|F_q(\Omega,t_1,t_2;\{0\})\Psi\rangle$ as the probability that during the time interval (t_1,t_2) the particle will not enter Ω, except for a set of $t\in(t_1,t_2)$ of measure zero, then Theorem 4.4 says that this probability is always equal to zero. Correspondingly, the probability $\langle\Psi|F_q(\Omega,t_1,t_2;(0,\infty))\Psi\rangle$ that the particle will spend some time in Ω is always equal to 1. Since Ω is an arbitrary Borel subset with nonempty interior, and since the time interval (t_1,t_2) is also arbitrary, that would indicate that the concept of a trajectory of the particle has to be abandoned.

Boundedness of the sojourn time $\tau_q(\Omega,-\infty,\infty;\Psi)$ and the sojourn time operator $T_q(\Omega,-\infty,\infty)$ under the conditions of Theorem 4.5(a) is a striking quantum mechanical result. It is of course equivalent to the boundedness of the spectrum of $T_q(\Omega,-\infty,\infty)$. Hence the probability

$$\langle\Psi|F_q(\Omega,-\infty,\infty;[0,(m/\hbar)\lambda_2(W)])\Psi\rangle$$

that the total time spent by the particle in Ω does not exceed $(m/\hbar)\lambda_2(W)$, is equal to 1; the particle cannot spend more time in Ω than $(m/\hbar)\lambda_2(W)$. This adds even more curiosity to what is described in the preceding paragraph.

APPENDIX A: PROOF OF THEOREM 4.1

For $\Psi\in L^2(\mathbb{R}^n)$ we will write $\Psi_t=\exp(-itH_0)\Psi$, and $\hat{\Psi}=F\Psi$, $\check{\Psi}=F^{-1}\Psi$, with $F: L^2(\mathbb{R}^n)\to L^2(\mathbb{R}^n)$ being the Fourier–Plancherel transform.

Lemma A.1: Let $\Psi\in L^2(\mathbb{R}^n)$. Then there exists a Borel measurable function $\tilde{\Psi}(t,x)$ defined on $\mathbb{R}^{1+n}$ and having the following properties: (a) for almost all $t\in\mathbb{R}$ the function $\mathbb{R}^n\ni x\to\tilde{\Psi}(t,x)$ belongs to $L^2(\mathbb{R}^n)$ and equals $\Psi_t(x)$ almost everywhere, i.e.,

$$\int_{\mathbb{R}^n}|\tilde{\Psi}(t,x)-\Psi_t(x)|^2\,dx=0, \tag{A1}$$

(b) for each $\varphi\in L^1(\mathbb{R})$,

$$\lim_{R\to\infty}\int_{\mathbb{R}^{1+n}}|\tilde{\Psi}(t,x)-\tilde{\Psi}_R(t,x)|^2|\,\varphi(t)|\,dt\,dx=0,$$

where

$$\tilde{\Psi}_R(t,x)=(2\pi)^{-n/2}\int_{\|k\|<R}\exp\left[ikx-it\left(\frac{k^2}{2}\right)\right]\hat{\Psi}(k)\,dk. \tag{A2}$$

Proof: Let $\hat{\Psi}_{Rt}(k)=\hat{\Psi}(k)\exp(-itk^2/2)\times\chi_{\{k\,|\,\|k\|<R\}}(k)$. Clearly, for each $t\in\mathbb{R}$, $\hat{\Psi}_{Rt}\in L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$ and $\lim_{R\to\infty}\hat{\Psi}_{Rt}=\hat{\Psi}_t$ (in the L^2 norm). Hence $\Psi_t=\lim_{R\to\infty}\Psi_{Rt}$, where $\Psi_{Rt}=F^{-1}\hat{\Psi}_{Rt}$, or

$$\Psi_{Rt}(x)=(2\pi)^{-n/2}\int_{\mathbb{R}^n}\exp(ikx)\hat{\Psi}_{Rt}(k)\,dk=\tilde{\Psi}_R(t,x). \tag{A3}$$

The functions $\tilde{\Psi}_R(t,x)$ are continuous functions of $(t,x)\in\mathbb{R}^{1+n}$ (hence Borel measurable). By the Fubini theorem and from the unitarity of the Fourier transform it follows that

$$\begin{aligned}\int_{\mathbb{R}^{1+n}}&|\tilde{\Psi}_R(t,x)-\tilde{\Psi}_{R'}(t,x)|^2(1+t^2)^{-1}\,dt\,dx\\&=\int_{\mathbb{R}}dt(1+t^2)^{-1}\int_{\mathbb{R}^n}dx|\tilde{\Psi}_R(t,x)-\tilde{\Psi}_{R'}(t,x)|^2\\&=\pi\|\hat{\Psi}_{R0}-\hat{\Psi}_{R'0}\|^2.\end{aligned} \tag{A4}$$

Since $\hat{\Psi}_{R0}\to\hat{\Psi}$ as $R\to\infty$, we can see that the functions $\tilde{\Psi}_R$ converge in the norm of $L^2(\mathbb{R}^{1+n},(1+t^2)^{-1}\,dt\,dx)$. Let us take $\tilde{\Psi}$ as the above limit of the sequence $\tilde{\Psi}_R$. We can assume that $\tilde{\Psi}$ is defined everywhere in $\mathbb{R}^{1+n}$. Moreover, from the properties of the L^2 spaces, there is a subsequence $r_N\to\infty$ such that $\tilde{\Psi}_{r_N}(t,x)\to\tilde{\Psi}(t,x)$ almost everywhere.

To check (a) notice that

$$\begin{aligned}\int_{\mathbb{R}}dt(1+t^2)^{-1}&\int_{\mathbb{R}^n}dx|\tilde{\Psi}(t,x)|^2\\&=\int_{\mathbb{R}^{1+n}}|\tilde{\Psi}(t,x)|^2(1+t^2)^{-1}\,dt\,dx<\infty\end{aligned} \tag{A5}$$

implies that for almost every $t \in \mathbb{R}$ the function $x \to \tilde{\Psi}(t,x)$ belongs to $L^2(\mathbb{R}^n)$. Applying then the Fatou lemma and the fact that $\Psi_{Rt} \to \Psi_t$, we obtain

$$0 = \lim_{R\to\infty} \int_{\mathbb{R}^{1+n}} |\tilde{\Psi}(t,x) - \tilde{\Psi}_R(t,x)|^2 (1+t^2)^{-1}\, dt\, dx$$

$$= \lim_{R\to\infty} \int_{\mathbb{R}} dt (1+t^2)^{-1} \int_{\mathbb{R}^n} dx |\tilde{\Psi}(t,x) - \tilde{\Psi}_R(t,x)|^2$$

$$\geqslant \int_{\mathbb{R}} dt (1+t^2)^{-1} \liminf_{R\to\infty} \int_{\mathbb{R}} dx |\tilde{\Psi}(t,x) - \tilde{\Psi}_R(t,x)|^2$$

$$= \int_{\mathbb{R}} dt (1+t^2)^{-1} \int_{\mathbb{R}^n} dx |\tilde{\Psi}(t,x) - \tilde{\Psi}_t(x)|^2, \quad \text{(A6)}$$

so that (a) holds.

To prove (b) rewrite the left-hand side of (A4) inserting $|\varphi(t)|$ in the place of $(1+t^2)^{-1}$. The right-hand side will thereby take the form

$$\|\varphi\|_1 \|\hat{\Psi}_{R0} - \hat{\Psi}_{R'0}\|^2. \quad \text{(A7)}$$

Hence the sequence $\tilde{\Psi}_R$ converges in the norm of $L^2(\mathbb{R}^{1+n}, |\varphi(t)|\,dt\,dx)$. The same is true for the subsequence $\tilde{\Psi}_{r_N}$ defined above. It easily follows that $\tilde{\Psi}$ is the limit of $\tilde{\Psi}_R$ in the norm of $L^2(\mathbb{R}^{1+n}, |\varphi(t)|\,dt\,dx)$. So (b) is proved. □

Lemma A.2: Let $\Psi \in L^2(\mathbb{R}^n)$ and let $\tilde{\Psi}$ and $\tilde{\Psi}_R$ be defined as in Lemma A.1. Let $\varphi \in L^1(\mathbb{R})$. Then for almost every $x \in \mathbb{R}^n$ (more precisely, for $x \in S_\varphi$ with $\mathbb{R}^n \setminus S_\varphi$ having measure zero),

$$\int_{\mathbb{R}} |\tilde{\Psi}(t,x)|^2 |\varphi(t)|\,dt < \infty, \quad \text{(A8)}$$

and there is a sequence $R_N = R_N(\varphi,x) \to \infty$ such that

$$\lim_{N\to\infty} \int_{\mathbb{R}} |\tilde{\Psi}_{R_N}(t,x) - \tilde{\Psi}(t,x)|^2 |\varphi(t)|\,dt = 0. \quad \text{(A9)}$$

Proof:

$$\int_{\mathbb{R}^n} dx \int_{\mathbb{R}} dt\, |\tilde{\Psi}(t,x)|^2 |\varphi(t)|$$

$$= \int_{\mathbb{R}} dt\, |\varphi(t)| \int_{\mathbb{R}^n} dx |\Psi_t(x)|^2$$

$$= \|\varphi\|_1 \|\Psi\| < \infty. \quad \text{(A10)}$$

The above implies (A8) for almost every $x \in \mathbb{R}^n$. From (b) of Lemma A.1 and the Fatou lemma we have,

$$\int_{\mathbb{R}^n} dx \liminf_{R\to\infty} \int_{\mathbb{R}} dt\, |\tilde{\Psi}(t,x) - \tilde{\Psi}_R(t,x)|^2 |\varphi(t)|$$

$$\leqslant \lim_{R\to\infty} \int_{\mathbb{R}^n} dx \int_{\mathbb{R}} dt\, |\tilde{\Psi}(t,x) - \tilde{\Psi}_R(t,x)|^2 |\varphi(t)| = 0. \quad \text{(A11)}$$

The existence of the sequence $R_N(\varphi,x)$ follows immediately. □

We will use spherical coordinates in $\mathbb{R}^n$ writing,

$$k = \kappa\eta(\omega), \quad dk = d\kappa\, d\omega\, \kappa^{n-1}, \quad \text{(A12)}$$

where $\kappa = \|k\|$, ω is the set of angular variables and $\eta(\omega)$ is the unit vector defined by these variables. We will not need the explicit expressions for $\eta(\omega)$ and $d\omega$.

Lemma A.3: Let $\Psi \in L^2(\mathbb{R}^n)$. There is then a set $\mathscr{E} \subseteq (0,\infty)$ such that $(0,\infty)\setminus\mathscr{E}$ has measure zero and for every $E \in \mathscr{E}$ and every $z \in \mathbb{C}^n$

$$\int d\omega |\hat{\Psi}(\sqrt{2E}\,\eta(\omega)) \exp(iz\eta(\omega)\sqrt{2E})| < \infty. \quad \text{(A13)}$$

The function

$$f(E,z) = (2\pi)^{-(n-1)/2} (2E)^{(n/2)-1} \times \int d\omega\, \hat{\Psi}(\sqrt{2E}\,\eta(\omega)) \exp(iz\eta(\omega)) \quad \text{(A14)}$$

is measurable for $(E,z) \in \mathscr{E} \times \mathbb{R}^n$, and for each fixed $E \in \mathscr{E}$ the function $\mathbb{C}^n \ni z \to f(E,z)$ is analytic.

Proof: By changing variables we obtain

$$\infty > \int_{\mathbb{R}^n} |\hat{\Psi}(k)|^2\, dk = \int_0^\infty d\kappa\, \kappa^{n-1} \int d\omega |\hat{\Psi}(\kappa\eta(\omega))|^2$$

$$= \int_0^\infty dE\, E^{(n/2)-1} \times \int d\omega |\hat{\Psi}(\sqrt{2E}\,\eta(\omega))|^2. \quad \text{(A15)}$$

Hence for almost all $E > 0$, i.e., for $E \in \mathscr{E}$

$$\int d\omega |\hat{\Psi}(\sqrt{2E}\,\eta(\omega))|^2 < \infty. \quad \text{(A16)}$$

Since $\int d\omega < \infty$ and $|\exp(iz\eta(\omega)\sqrt{2E})| \leqslant \exp(\|z\|\sqrt{2E})$, we infer that (A13) holds for $E \in \mathscr{E}$. Measurability follows from the Fubini theorem, and analyticity can be established without trouble calculating derivatives with the aid of majorized convergence. □

Lemma A.4: Let $\Psi \in L^2(\mathbb{R}^n)$ and let $\Omega \subseteq \mathbb{R}^n$ be a Borel subset with nonzero measure. Suppose that

$$\int_{\mathbb{R}} dt \int_\Omega dx |\Psi_t(x)|^2 < \infty. \quad \text{(A17)}$$

Then for almost every $x \in \Omega$

$$\int_{\mathbb{R}} dt\, |\tilde{\Psi}(t,x)|^2 < \infty, \quad \int_0^\infty dE\, |f(E,x)|^2 < \infty, \quad \text{(A18)}$$

and

$$\tilde{\Psi}(\cdot,x) = F(f(\cdot,x)), \quad \text{(A19)}$$

i.e., the function $\mathbb{R} \ni t \to \tilde{\Psi}(t,x)$ is the Fourier–Plancherel transform of the function $\mathbb{R} \ni E \to f(E,x)$. Here $\tilde{\Psi}$ is defined in Lemma A.1, and f in Lemma A.3 with the natural extension: $f(E,x) = 0$ for $E \leqslant 0$.

Proof: Let $\mathscr{F}$ be the family of all functions $\sigma_\alpha : \mathbb{R} \to \mathbb{R}$ having the form

$$\sigma_\alpha(t) = \exp[-(t-\alpha)^2/2] \quad \text{(A20)}$$

with α rational. Since $\mathscr{F}$ is countable, Lemma A.2 implies the existence of a set $S \subseteq \mathbb{R}^n$, such that $\mathbb{R}^n \setminus S$ has measure zero, and for each $x \in S$ and each $\sigma_\alpha \in \mathscr{F}$ there is a sequence $R_N(\sigma_\alpha,x) \to \infty$ with

$$\lim_{N\to\infty} \int_{\mathbb{R}} |\tilde{\Psi}_{R_N(\sigma_\alpha,x)}(t,x) - \tilde{\Psi}(t,x)|^2 \sigma_\alpha(t)\,dt = 0. \quad \text{(A21)}$$

Now,

$$\infty > \int_{\mathbf{R}} dt \int_{\Omega} dx |\Psi_t(x)|^2 = \int_{\mathbf{R}\times\Omega} |\tilde{\Psi}(t,x)|^2 \, dt\, dx$$
$$= \int_{\Omega} dx \int_{\mathbf{R}} dt \, |\tilde{\Psi}(t,x)|^2. \quad \text{(A22)}$$

So for $x \in \Omega_1 \subseteq \Omega$, $\Omega \setminus \Omega_1$ being of measure zero,

$$\int_{\mathbf{R}} dt \, |\tilde{\Psi}(t,x)|^2 < \infty. \quad \text{(A23)}$$

Define $\Omega_0 = \Omega_1 \cap S$. Clearly, $\Omega \setminus \Omega_0$ has measure zero.

Now fix an $x \in \Omega_0$, and define a linear functional μ_x: $L^2(\mathbf{R}) \to \mathbf{C}$,

$$\mu_x \phi = \int_{\mathbf{R}} dt \, \check{\Phi}(t) \tilde{\Psi}(t,x), \quad \Phi \in L^2(\mathbf{R}). \quad \text{(A24)}$$

By (A23) this is a bounded linear functional. Thus from the Riesz representation theorem there is a unique $g_x \in L^2(\mathbf{R})$ such that

$$\mu_x \Phi = \int_{\mathbf{R}} dE \, \Phi(E) g_x(E), \quad \Phi \in L^2(\mathbf{R}). \quad \text{(A25)}$$

We are going to show that $g_x(E) = f(E,x)$ for almost all E. Consider the action of μ_x on $\sigma_\alpha(E) = \exp[-i\alpha x - (E^2/2)]$. By (A21) and the Schwarz inequality in $L^2(\mathbf{R}, \sigma_\alpha(t)dt)$,

$$\lim_{N\to\infty} \int_{\mathbf{R}} dt \, \sigma_\alpha(t) |\tilde{\Psi}_{R_N(\sigma_\alpha,x)}(t,x) - \tilde{\Psi}(t,x)|$$
$$\leqslant \lim_{N\to\infty} \sqrt{2\pi} \left\{ \int_{\mathbf{R}} dt \, \sigma_\alpha(t) |\tilde{\Psi}_{R_N(\sigma_\alpha,x)}(t,x) - \tilde{\Psi}(t,x)|^2 \right\}^{1/2} = 0. \quad \text{(A26)}$$

Hence

$$\mu_x \hat{\sigma}_\alpha = \lim_{N\to\infty} \int_{\mathbf{R}} dt \, \sigma_\alpha(t) \tilde{\Psi}_{R_N(\sigma_\alpha,x)}(t,x). \quad \text{(A27)}$$

Now,

$$\tilde{\Psi}_R(t,x) = (2\pi)^{-n/2} \int_{\|k\|<R} \exp\left[ikx - it\left(\frac{k^2}{2}\right)\right] \hat{\Psi}(k) dk$$
$$= (2\pi)^{-1/2} \int_0^{R^2/2} \exp(-itE) f(E,x) dE, \quad \text{(A28)}$$

with $f(E,x)$ defined by (A14).

Since by Schwarz inequality,

$$|f(E,x)| \leqslant \left(\int d\omega \right)^{1/2} \left\{ \int d\omega |\hat{\Psi}(\sqrt{2E}\, \eta(\omega))|^2 \right\}^{1/2} \times \frac{(2E)^{(n/2)-1}}{(2\pi)^{(n-1)/2}}, \quad \text{(A29)}$$

the function $E \to f(E,x)\chi_{\{E' | 0<E'<R^2/2\}}(E)$ belongs to $L^2(\mathbf{R}) \cap L^1(\mathbf{R})$ as a consequence of $\hat{\Psi} \in L^2(\mathbf{R}^n)$ and $R^2/2 < \infty$. Therefore, the function $t \to \infty \Psi_R(t,x)$ is its Fourier transform. By Parseval's identity we thus have

$$\mu_x \hat{\sigma}_\alpha = \lim_{N\to\infty} \int_0^{R_N^2(\sigma_\alpha,x)/2} \exp\left(-i\alpha x - \left(\frac{E^2}{2}\right)\right) f(E,x) dE. \quad \text{(A30)}$$

Since

$$\int_0^\infty \exp\left(\frac{-E^2}{2}\right) |f(E,x)| dE < \infty \quad \text{(A31)}$$

as a consequence of (A29) and $\hat{\Psi} \in L^2(\mathbf{R}^n)$, so we obtain

$$u_x \hat{\sigma}_\alpha = \int_0^\infty \exp\left(-i\alpha x - \left(\frac{E^2}{2}\right)\right) f(E,x) dE. \quad \text{(A32)}$$

Comparison with (A25) yields

$$\int_{\mathbf{R}} \exp(-iE\alpha) \exp\left(\frac{-E^2}{2}\right) \times (f(E,x)\chi_{\{E'|E'>0\}}(E) - g_x(E)) dE = 0 \quad \text{(A33)}$$

for all rational α. But due to the factor $\exp(-E^2/2)$ and the properties of $f(E,x)$ and $g_x(E)$, the integrant is from $L^1(\mathbf{R})$. Therefore (A33) holds for all $\alpha \in \mathbf{R}$. Hence by the uniqueness property of the Fourier transform in $L^1(\mathbf{R})$, we obtain

$$g_x(E) = f(E,x)\chi_{\{E'|E'>0\}}(E) \quad \text{(A34)}$$

for almost all $E \in \mathbf{R}$. So (A18) is true [cf. also (A23)].

To finish the proof we apply Parseval's identity to (A24). This, and comparison with (A25), yields $F^{-1}\tilde{\Psi}(\cdot,x) = g_x$, or $\tilde{\Psi}(\cdot,x) = \hat{g}_x = Ff(\cdot,x)$ by (A34). □

Proof of Theorem 4.1: Suppose that the measure of the set $\{t \in \mathbf{R} | E(\Omega)\Psi_t = 0\}$ is nonzero. Then, by Lemma 3.1 $E(\Omega)\Psi_t = 0$ for all $t \in \mathbf{R}$. Hence

$$\int_{\mathbf{R}} dt \int_{\Omega} dx |\Psi_t(x)|^2 = 0. \quad \text{(A35)}$$

Applying Lemmas A.1, A.4, A.3, and Parseval's identity, we have

$$0 = \int_{\mathbf{R}} dt \int_{\Omega} dx |\tilde{\Psi}(t,x)|^2 = \int_{\Omega} dx \int_{\mathbf{R}} dt \, |\tilde{\Psi}(t,x)|^2$$
$$= \int_{\Omega} dx \int_0^\infty dE \, |f(E,x)|^2 = \int_0^\infty dE \int_{\Omega} dx |f(E,x)|^2. \quad \text{(A36)}$$

This means that for almost every $E > 0$, the set

$$\{x \in \Omega | f(E,x) \neq 0\} = B_E \quad \text{(A37)}$$

has zero measure. From the analytic properties of $f(E,z)$ (Lemma A.3) and the fact that Int $\Omega \neq 0$, it follows that $\{x \in \mathbf{R}^n | f(E,x) \neq 0\} = \phi$. (Note that for $n = 1$ the condition Int $\Omega \neq \phi$ is superfluous.)

Now, from Lemma A.1 or the proof of it, there is a sequence $r_N \to \infty$ such that $\tilde{\Psi}_{r_N}(t,x) \to \tilde{\Psi}(t,x)$ almost everywhere in $\mathbf{R}^{1+n}$. But

$$\tilde{\Psi}_{r_N}(t,x) = (2\pi)^{-n/2} \int_{\|k\|<r_N} \times \exp\left[ikx - it\left(\frac{k^2}{2}\right)\right] \hat{\Psi}(k) dk$$
$$= (2\pi)^{-1/2} \int_0^{r_N^2/2} \exp(-itE) f(E,x) = 0. \quad \text{(A38)}$$

Hence $\tilde{\Psi}(t,x) = 0$ almost everywhere. This implies

$$0=\int_{\mathbf{R}^{1+n}}|\tilde{\Psi}(t,x)|^2\,dt\,dx=\int_{\mathbf{R}}dt\int_{\mathbf{R}^n}dx|\tilde{\Psi}(t,x)|^2$$
$$=\int_{\mathbf{R}}dt\int_{\mathbf{R}^n}dx|\Psi_t(x)|^2=\int_{\mathbf{R}}dt\,\|\Psi\|^2, \tag{A39}$$

so that $\|\Psi\|^2=0$, in contradiction with the assumptions of Theorem 4.1. The set $\{t\in\mathbf{R}|E(\Omega)\Psi_t=0\}$ must have zero measure. □

Remark A.5: For $n=1$, Theorem 4.1 holds under the weaker assumption that $\Omega\subseteq\mathbf{R}$ has nonzero measure.

APPENDIX B: PROOF OF THEOREMS 4.2 AND 4.3

Proof of Theorem 4.2: Without loss of generality we assume that $\Omega=\mathbf{R}^{n-2}\times W$. Let us take $\Psi\in L^2(\mathbf{R}^n)$ with $\hat{\Psi}\in C_0^\infty(\mathbf{R}^n)$ (infinitely differentiable with compact support). Then all the operations below are certainly legitimate:

$$\int_{\mathbf{R}}dt\int_{\mathbf{R}^{n-2}\times W}dx|\Psi_t(x)|^2=\int_{\mathbf{R}}dt\int_{\mathbf{R}^{n-2}\times W}dx\left|(2\pi)^{-n/2}\int_{\mathbf{R}^n}dk\exp\left(ikx-it\left(\frac{k^2}{2}\right)\right)\hat{\Psi}(k)\right|^2$$
$$=\int_{\mathbf{R}}dt\int_{\mathbf{R}^{n-2}}dx_1\cdots dx_{n-2}\int_W dx_{n-1}\,dx_n\left|(2\pi)^{-n/2}\int_{\mathbf{R}^{n-2}}dk_1\cdots dk_{n-2}\int_0^\infty d\kappa\,\kappa\int_0^{2\pi}d\varphi\right.$$
$$\times\hat{\Psi}(k_1,\cdots,k_{n-2},\kappa\cos\varphi,\kappa\sin\varphi)\exp(ik_1x_1+\cdots+ik_{n-2}x_{n-2}-it(k_1^2+\cdots+k_{n-2}^2)/2)$$
$$\left.\times\exp(ix_{n-1}\kappa\cos\varphi+ix_n\kappa\sin\varphi-it(\kappa^2/2))\right|^2=(2\pi)^{-1}\int_W dx_{n-1}\,dx_n\int_{\mathbf{R}}dt$$
$$\times\int_{\mathbf{R}^{n-2}}dx_1\cdots dx_{n-2}\left|(2\pi)^{-(n-1)/2}\int_{\mathbf{R}^{n-2}}dk_1\cdots dk_{n-2}\exp(ik_1x_1+\cdots+ik_{n-2}x_{n-2})\right.$$
$$\times\exp(-it(k_1^2+\cdots+k_{n-2}^2)/2)\int_0^\infty dE\exp(-itE)\int_0^{2\pi}d\varphi\exp(ix_{n-1}\sqrt{2E}\cos\varphi$$
$$\left.+ix_n\sqrt{2E}\sin\varphi)\hat{\Psi}(k_1,\cdots k_{n-2},\sqrt{2E}\cos\varphi,\sqrt{2E}\sin\varphi\right|^2=(2\pi)^{-1}\int_W dx_{n-1}\,dx_n\int_{\mathbf{R}}dt$$
$$\times\int_{\mathbf{R}^{n-2}}dk_1\cdots dk_{n-2}\left|\int_0^\infty dE\exp(-itE)\int_0^{2\pi}d\varphi\exp(ix_{n-1}\sqrt{2E}\cos\varphi+ix_n\sqrt{2E}\sin\varphi)\right.$$
$$\left.\times\hat{\Psi}(k_1,\cdots k_{n-2},\sqrt{2E}\cos\varphi,\sqrt{2E}\sin\varphi)\right|^2=(2\pi)^{-1}\int_W dx_{n-1}\,dx_n\int_{\mathbf{R}^{n-2}}dk_1\cdots dk_{n-2}$$
$$\times\int_0^\infty dE\left|\int_0^{2\pi}d\varphi\exp(ix_{n-1}\sqrt{2E}\cos\varphi+ix_n\sqrt{2E}\sin\varphi)\hat{\Psi}(k_1,\cdots,k_{n-2},\sqrt{2E}\right.$$
$$\left.\times\cos\varphi,\sqrt{2E}\sin\varphi)\right|^2\leqslant\int_W dx_{n-1}\,dx_n\int_{\mathbf{R}^{n-2}}dk_1\cdots dk_{n-2}\int_0^\infty dE\int_0^{2\pi}d\varphi\,|\hat{\Psi}(k_1,\cdots k_{n-2},\sqrt{2E}$$
$$\times\cos\varphi,\sqrt{2E}\sin\varphi)|^2=\int_W dx_{n-1}\,dx_n\int_{\mathbf{R}^n}dk\,|\hat{\Psi}(k)|^2=\lambda_2(W)\|\Psi\|^2. \tag{B1}$$

As can be seen, Parseval's identity has been applied twice, and the Schwarz inequality produced the $\leqslant$ sign.

To check that the inequality (4.3) holds for an arbitrary $\Psi\in L^2(\mathbf{R}^n)$, we take a sequence $\Psi^{(N)}\in L^2(\mathbf{R}^n)$ such that $\hat{\Psi}^N\in\mathscr{C}_0^\infty(\mathbf{R}^n)$ and $\Psi^{(N)}\to\Psi$ in the L^2 norm. Using the Fatou lemma we have

$$\int_{\mathbf{R}}dt\int_\Omega dx|\Psi_t(x)|^2=\int_{\mathbf{R}}dt\lim_{N\to\infty}\int_\Omega dx|\Psi_t^{(N)}(x)|^2$$
$$\leqslant\liminf_{N\to\infty}\int_{\mathbf{R}}dt\int_\Omega dx|\Psi_t^{(N)}(x)|^2\leqslant\lim_{N\to\infty}\lambda_2(W)\|\Psi^{(N)}\|^2=\lambda_2(W)\|\Psi\|^2. \tag{B2}$$

□

Proof of Theorem 4.3: Let $\Psi\in L^2(\mathbf{R})$ be such that $\hat{\Psi}\in L^2(\mathbf{R})\cap L^1(\mathbf{R})$ and that $\hat{\Psi}(k)=0$ for $k\in(-\epsilon,\epsilon)$, with some $\epsilon>0$. Then

$$\tau(\Omega,-\infty,\infty;\Psi)=\int_{\mathbf{R}}dt\int_\Omega dx|\Psi_t(x)|^2=\int_{\mathbf{R}}dt\int_\Omega dx\left|(2\pi)^{-1}\int_{\mathbf{R}}dk\exp\left(ikx-it\left(\frac{k^2}{2}\right)\right)\hat{\Psi}(k)\right|^2$$
$$=\int_\Omega dx\int_{\mathbf{R}}dt\left|(2\pi)^{-1}\int_0^\infty dE\exp(-itE)(2E)^{-1/2}[\exp(i\sqrt{2E}\,x)\hat{\Psi}(\sqrt{2E})\right.$$
$$\left.+\exp(-i\sqrt{2E}\,x)\hat{\Psi}(-\sqrt{2E})]\right|^2=\int_\Omega dx\int_0^\infty dE(2E)^{-1}|\exp(i\sqrt{2E}x)\hat{\Psi}\sqrt{2E})$$
$$+\exp(-i\sqrt{2E}x)\hat{\Psi}(-\sqrt{2E})|^2=\int_\Omega dx\int_0^\infty dk\,k^{-1}|\exp(ikx)\hat{\Psi}(k)+\exp(-ikx)\hat{\Psi}(-k)|^2. \tag{B3}$$

Due to the factor k^{-1} it is easy to see that Theorem 4.3 is true for $t_1 = -\infty$, $t_2 = \infty$. Its validity for semibounded intervals (t_1,t_2) can be easily deduced from the time inversion invariance of the Schrödinger equation and the fact that $\tau(\Omega, -\infty,\infty;\Psi) = \tau(\Omega, -\infty,t_0;\Psi) + \tau(\Omega,t_0,\infty;\Psi)$. □

[1]R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948).
[2]E. Nelson, *Quantum Fluctuations* (Princeton U. P., Princeton, 1985).
[3]G. Cavalleri and G. Spavieri, Nuovo Cimento B **95**, 194 (1986), and references therein.
[4]H. Ekstein and A. J. F. Siegert, Ann. Phys. (NY) **68**, 509 (1971).
[5]E. B. Davies, Helv. Phys. Acta **48**, 365 (1975).
[6]A. J. Baz, Sov. J. Nucl. Phys. **4**, 182 (1967); **5**, 161 (1967).
[7]B. J. Verhaar, A. M. Schulte, and J. De Kam, Physica A **91**, 119 (1978).
[8]Ph. A. Martin, Acta Phys. Austriaca, Suppl. XXIII, 157 (1981), and references therein.
[9]D. Bolle and T. A. Osborn, J. Math. Phys. **20**, 1121 (1979), and references therein.
[10]T. Kato, *Perturbation Theory for Linear Operators* (Springer, New York, 1966), Chap. VI.
[11]H. Dym and H. P. McKean, *Fourier Series and Integrals* (Academic, New York, 1972), Chap. 3.
[12]R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain* (AMS, Providence, RI, 1934); Theorems V and XII.
[13]B. Misra and E. C. G. Sudershan, J. Math. Phys. **18**, 756 (1977).
[14]The simplest way to prove the spreading property is to apply an asymptotic formula for $\exp(-itH_0)$, see M. Reed and B. Simon, *Methods of Modern Mathematical Physics II, Fourier Analysis, Self-Adjointness* (Academic, New York, 1975).
[15]This strengthens less detailed results known in the literature. E.g., Ref. 16 proves that for $\Omega \cap \Omega_1 = \emptyset$ and any $\epsilon > 0$ there is a $t \in (0,\epsilon)$ with $E_q(\Omega)\Psi_t \neq 0$. The statement about the measure of the set (4.2) is crucial for the problem of the sojourn time operator.
[16]G. C. Hegerfeldt and S. N. M. Ruijsenaars, Phys. Rev. D **22**, 377 (1980).

An infinite number of hidden variables in hyper-Kähler metrics

K. Takasaki
Research Institute for Mathematical Sciences, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto-shi 606, Japan

(Received 13 September 1988; accepted for publication 22 February 1989)

Three types of hidden variables, both independent and dependent, are shown to underlie the hyper-Kähler geometry in a complexified setting. The variables satisfy an infinite set of differential equations ("hierarchy"), just as in the case of most nonlinear integrable systems. The notion of the Plebanski key functions [J. Math. Phys. **16**, 2395 (1975)] is extended to this hierarchy to give an analog of the notion of the "τ function." Two examples of special solutions, which are reminiscent of several solution techniques in the theory of nonlinear integrable systems, are presented for illustration.

I. INTRODUCTION

A significant property of so-called "nonlinear integrable systems" is the existence of various "hidden variables." For both theoretical, and historical reasons, it appears appropriate to classify the variables into three categories. The first category is comprised of independent variables which play the role of "time variables" in underlying dynamical flows. This fact was first discovered by Gardner *et al.*[1] in the case of the Korteweg–de Vries equation. Gardner *et al.* pointed out that the KdV equation is an infinite-dimensional Hamiltonian system with an infinite number of first integrals; the first integrals can generate commuting Hamiltonian flows, with each flow having its own "time variable." Lax[2] gave another interpretation of this fact from a more general standpoint without relying on the Hamiltonian picture. The essence of Lax's method is to rewrite the nonlinear system in question into the so-called Lax form, the combatibility condition of a linear system; this opened a way to the introduction of the second category of hidden variables, which are solutions ("wavefunctions") of the linear system. The discovery of the third category came from a seemingly very different direction through the work of Hitora,[3] who found the variables in the course of the development of his original method, the "bilinearization technique."

Now, it is widely recognized that these three categories of hidden variables are closely linked. Besides, mathematical structures become most clear at the level of the "hierarchy," i.e., the totality of differential equations satisfied by these hidden variables, rather than the original form of equations such as the original KdV equation. This fact lies at the heart of recent progress in the theory of nonlinear integrable systems; Hirota's independent variable, which is also called the "τ function,"[4,5] has come to play the most central role.

We now attempt to show a similar structure in hyper-Kähler metrics. This paper is organized as follows. Section II is a brief review of some basic notions on (complexified) hyper-Kähler metrics along the lines of Plebanski,[6] Boyer and Plebanski,[7] and Gindikin,[8] which is closely related with the Penrose twistor method.[9] In Sec. III, we shall introduce a series of hidden independent variables which play the role of the "time variables" in the sense mentioned above. The original field equations for hyper-Kähler metrics can be then extended into a "hierarchy." In Sec. IV we shall extend the notion of the Plebanski "second key function" to that hierarchy. In Secs. V and VI we shall illustrate these abstract constructions in the case of special solutions, which are reminiscent of several solution techniques in the theory of nonlinear integrable systems. Section VII presents a treatment of the Plebanski "first key function " in our context, introducing a larger set of hidden variables. Our conclusion is given in Sec. VIII.

II. BASIC NOTIONS CONCERNING "COMPLEX" HYPER-KÄHLER METRICS

In the standard differential-geometric theory (cf. Ref. 10) hyper-Kähler metrics are understood in the context of Riemannian geometry on a real $4r$-dimensonal ($r \geqslant 1$) manifold M; however, this is not very suited for our purpose. What is more convenient is to reformulate the hyper-Kähler property in the language of "complex metrics" (cf. Ref. 11), namely, nondegenerate complex analytic bilinear forms on the holomorphic tangent bundle of a $4r$-dimensional manifold X ("heaven" in the terminology of Plebanski[6]).

According to the customary setting of twistor theory, let us start from a complex metric written as

$$g = \tfrac{1}{2}\epsilon_{AB}\epsilon_{\alpha\beta}e^{A\alpha}e^{B\beta}, \tag{1}$$

where $e^{A\alpha}$, $A = 1,\ldots, r$, $\alpha = 1, 2$ are a collection of linearly independent differential one-forms ("vielbein") and ϵ_{AB} and $\epsilon_{\alpha\beta}$ are the ordinary symplectic forms in $2r$- and 2-dimensions normalized as $\epsilon_{12} = -\epsilon_{21} = \cdots = \epsilon_{2r-12r} = -\epsilon_{2r2r-1} = 1$, where the other components have been put to zero. We also apply the Einstein summation convention to symplectic indices. Writing a metric in the form (1) is not unique; there remains "local gauge freedom" of the symplectic rotations $e^{A\alpha} \to h^{A}_{B}e^{B\beta}k^{\alpha}_{\beta}$, $h = (h^{A}_{B})$, and $k = (k^{\alpha}_{\beta})$ which take values in $\mathrm{Sp}(r,\mathbb{C})$ and $\mathrm{Sp}(1,\mathbb{C})$. According to the work of Plebanski,[6] Boyer and Plebanski,[7] and Gindikin[8] the hyper-Kähler property, after an appropriate $\mathrm{Sp}(1,\mathbb{C})$ gauge transformation as above, reduces to the exterior differential equations

$$d\omega^{\alpha\beta} = 0, \quad \omega^{\alpha\beta} := \tfrac{1}{2}\epsilon_{AB}e^{A\alpha}\wedge e^{B\beta}. \tag{2}$$

 0022-2488/89/071515-07$02.50

Equations (1) and (2) (note the symmetry $\omega^{\alpha\beta}=\omega^{\beta\alpha}$) can be gathered into

$$d\omega(\lambda)=0,\quad \omega(\lambda):=\tfrac{1}{2}\epsilon_{AB}(e^{A1}+\lambda e^{A2})\wedge(e^{B1}+\lambda e^{B2}), \tag{3}$$

where λ is a new parameter with values in $\mathbb{P}^1$ (the "spectral parameter" in the terminology of the theory of nonlinear integrable systems) understood to be constant under the total differential $d=d_X$ on X, i.e., $d\lambda=0$.

The two-form $\omega(\lambda)$ (what Gindikin[8] calls a "bundle of two-forms") is degenerate, but of constant rank: $\omega(\lambda)^{\wedge r}\neq 0$ and $\omega(\lambda)^{\wedge(r+1)}=0$ [where $\omega(\lambda)^{\wedge k}$ denotes the k th exterior power]. Therefore, a theorem of Darboux, under the exterior differential equation (3), asserts that there exist $2r$ functions $u^A(\lambda)$ also depending on λ such that

$$\omega(\lambda)=\tfrac{1}{2}\epsilon_{AB}\,du^A(\lambda)\wedge du^B(\lambda). \tag{4}$$

To be precise, the Darboux theorem is of local nature and ensures that at any point of $X\times\mathbb{P}^1$ there are such functions $u^A(\lambda)$, but only in a neighborhood of that point. For the moment let us assume that X is replaced by a small open subset, so that the domain of definition of $u^A(\lambda)$ with respect to space-time variables covers the whole X.

In a neighborhood of $\lambda=\infty$, in particular, one can choose $u^A(\lambda)$ to have a Laurent expansion as

$$u^A(\lambda)=\sum_{n=-\infty}^{1}u_n^A\lambda^n,\quad u_n^A=\text{functions on } X. \tag{5}$$

Evidently $u^A(\lambda)$ satisfy

$$(\epsilon_{AB}du^A(\lambda)\wedge du^B(\lambda))_-=0, \tag{6}$$

where $(\ )_-$ denotes the projection onto the part of the negative powers of λ; likewise, we shall use the notation $(\ )_+$ in an opposite sense

$$\left(\sum a_n\lambda^n\right)_+:=\sum_{n\geq 0}a_n\lambda^n,\quad \left(\sum a_n\lambda^n\right)_-:=\sum_{n<0}a_n\lambda^n. \tag{7}$$

One can readily show the relation

$$du_0^1\wedge\cdots\wedge du_0^{2r}\wedge du_1^1\wedge\cdots\wedge du_1^{2r}=e^{11}\wedge\cdots\wedge e^{2r1}\wedge e^{12}\wedge\cdots\wedge e^{2r2}\neq 0;$$

hence (u_0^A,u_1^A) can play the role of local coordinates on X.

From the exterior differential equations (6) above $e_{AB}u^A_{-1}\,du_0^B+\epsilon_{AB}u^A_{-2}\,du_1^B$ becomes, in particular, a closed form. Therefore, at least locally, there is a potential Θ such that

$$d\Theta=\epsilon_{AB}u^A_{-1}\,du_0^B+\epsilon_{AB}u^A_{-2}\,du_1^B. \tag{8}$$

Equation (8) is nothing less than the Plebanski second key function.[6]

If one applies the Darboux theorem in a neighborhood of $\lambda=0$, it then follows that there are the Laurent series

$$\hat{u}^A(\lambda)=\sum_{n=0}^{\infty}\hat{u}_n^A\lambda^n,\quad \hat{u}_n^A=\text{functions on } X \tag{9}$$

that satisfy

$$\omega(\lambda)=\tfrac{1}{2}\epsilon_{AB}\,d\hat{u}^A(\lambda)\wedge d\hat{u}^B(\lambda). \tag{10}$$

From Eqs. (9) and (10) one can further introduce the two potentials $\hat{\Theta}$ and Ω through

$$d\hat{\Theta}=\epsilon_{AB}\hat{u}_2^A\,d\hat{u}_1^B+\epsilon_{AB}\hat{u}_3^A\,d\hat{u}_0^B, \tag{11}$$

$$d\Omega=-\epsilon_{AB}u_0^A\,du_1^B+\epsilon_{AB}\hat{u}_1^A\,d\hat{u}_0^B. \tag{12}$$

The potential $\hat{\Theta}$ is another kind of second key function and Ω is the first key function which is the counterpart of the Kähler potential.

The functions $u^A(\lambda)$ and $\hat{u}^A(\lambda)$ are closely related to the Penrose twistor method (cf. Ref. 9): These functions show a prototype of what we shall argue later; $u^A(\lambda)$, $\hat{u}^A(\lambda)$, Θ, $\hat{\Theta}$, and Ω are all hidden variables defined as solutions of some differential equations whose integrability is ensured by the basic exterior differential equations (2). According to the classification mentioned in Sec. I, the first two variables above (or their Laurent coefficients) may be thought of as the "second category" and the other three variables may be thought of as the "third category." In the following we mostly focus on $u^A(\lambda)$, Θ, and the relevant hidden variables of the "first category." The other variables require a much more complicated treatment (cf. Sec. VII).

III. THE HYPER-KÄHLER HIERARCHY

We now introduce hidden variables that belong to the "first category." In Sec. II the Laurent coefficients u_n^A were limited to $n\leq 1$. Let us insert the remaining coefficients u_n^A $(n\geq 2)$ into $u^A(\lambda)$ as

$$u^A(\lambda)=\sum_{n=-\infty}^{\infty}u_n^A\lambda^n \tag{13}$$

and consider the same exterior differential equation (6) above, where u_n^A for $n\geq 0$ (resp., $n<0$) are viewed as independent (resp., dependent) variables. Thus the dependent variables are the same as before, but we now have an infinite number of new independent variables. The following result is fundamental (see the Appendix for a proof).

Proposition: The following systems of equations (14)–(19) are equivalent:

$$(\epsilon_{AB}\,du^A(\lambda)\wedge du^B(\lambda))_-=0; \tag{14}$$

$$du^A(\lambda)=\frac{\partial u^A(\lambda)}{\partial u_0^B}e^B(\lambda), \tag{15a}$$

$$\{u^A(\lambda),u^B(\lambda)\}=e^{AB}, \tag{15b}$$

where

$$e^B(\lambda):=\left(\frac{\partial u_A(\lambda)}{\partial u_{B0}}du^A(\lambda)\right)_+;$$

$$\partial_A^n(\lambda)u^B(\lambda)=0\quad(n\geq 1), \tag{16a}$$

$$\{u^A(\lambda),u^B(\lambda)\}=e^{AB}, \tag{16b}$$

where

$$\partial_A^n(\lambda):=\frac{\partial}{\partial u_n^A}-\lambda\frac{\partial}{\partial u_{n-1}^A}+H(u_{A,-n});$$

$$\frac{\partial u^B(\lambda)}{\partial u_n^A}=\{u^B(\lambda),(\lambda^n u_A(\lambda))_+\}\quad(n\geq 1), \tag{17a}$$

$$\{u^A(\lambda),u^B(\lambda)\}=\epsilon^{AB}; \tag{17b}$$

$$\frac{\partial u_{B,-n}}{\partial u_m^A}-\frac{\partial u_{A,-m}}{\partial u_n^B}+\{u_{A,-m},u_{B,-n}\}=0, \tag{18a}$$

$$\frac{\partial u_{B,-n}}{\partial u^A_{m-1}} = \frac{\partial u_{A,-m}}{\partial u^B_{n-1}} \quad (n,m \geqslant 1)\,; \tag{18b}$$

$$\frac{\partial(\lambda^n u_B(\lambda))_+}{\partial u^A_m} - \frac{\partial(\lambda^m u_A(\lambda))_+}{\partial u^B_n} + \{(\lambda^m u_A(\lambda))_+, (\lambda^n u_B(\lambda))_+\} + \lambda^{n+m}\epsilon_{AB} = 0 \quad (n,m \geqslant 1)\,. \tag{19}$$

Several new notations arise from (14)–(19):

$$\{F_1,F_2\} := \epsilon^{AB}\frac{\partial F_1}{\partial u^A_0}\frac{\partial F_2}{\partial u^B_0} \quad \text{(Poisson bracket)}\,, \tag{20}$$

$$H(F) := \epsilon^{AB}\frac{\partial F}{\partial u^A_0}\frac{\partial}{\partial u^B_0} \quad \text{(Hamiltonian vector field)}\,, \tag{21}$$

$$\xi_A = \epsilon_{AB}\xi^B, \quad \eta^B = \eta_A\epsilon^{AB} \quad \text{(raising/lowering indices)}\,, \tag{22}$$

$$\epsilon^{AB} = \epsilon_{AB}, \quad \text{for } 1 \leqslant A, B \leqslant 2r\,. \tag{23}$$

Let us call the system of differential equations the *hyper-Kähler hierarchy.* Each of the equivalent expressions (15)–(19) has its own interesting meaning (cf. Ref. 12 for the case of the four-dimensional sector): Eq. (15) is a key for relating the present setting with the integrability of a Pfaffian system, which leads to the construction of curved twistor space; Eq. (16) resembles the "linear system" of a more familiar type of nonlinear integrable system; Eq. (17) gives a Hamiltonian form of this hierarchy, where the role of u^A_n $(n \geqslant 0)$ as "time variables" is most clear in this representation; Eq. (18) will be used to introduce a hidden variable of the "third category" in Sec. IV; and Eq. (19) may be thought of as a "zero-curvature representation." It should be noted that the basic Lie algebra structure manifest in Eqs. (14)–(19) is not that of matrix Lie algebras, but of symplectic geometry (such as Hamiltonian vector fields and Poisson brackets).

IV. THE SECOND KEY FUNCTION IN HYPER-KÄHLER HIERARCHY

The notion of the second key function Θ carries over to our hyper-Kähler hierarchy because from the λ^{-1} term of the basic exterior differential equation (6) one obtains the relation

$$d\left(\sum_{n=0}^{\infty} u_{A,-n-1}\, du^A_n\right) = 0\,, \tag{24}$$

which allows one to introduce a function Θ as

$$d\Theta = -\sum_{n=0}^{\infty} u_{A,-n-1}\, du^A_n\,, \tag{25}$$

or, equivalently,

$$u_{A,-n-1} = -\frac{\partial\Theta}{\partial u^A_n} \quad (n \geqslant 0)\,. \tag{26}$$

Note that Θ is unique except for an integration constant:

$$\Theta \to \Theta + \text{const}\,. \tag{27}$$

It is remarkable that the dependent variables u^A_{-n-1} $(n \geqslant 0)$ thus become derivatives of just one function Θ: This is an advantage of introducing the new independent variables u^A_n $(n \geqslant 2)$. One will then expect to rewrite the whole hierarchy into a system with a single unknown function. Let us recall Eq. (18) in the proposition in Sec. III. The second part of (18) is nothing less than the integrability conditions of the equations defining Θ. From the first part of (18) one obtains

$$\frac{\partial^2\Theta}{\partial u^A_{m-1}\,\partial u^B_n} - \frac{\partial^2\Theta}{\partial u^A_m\,\partial u^B_{n-1}} + \left\{\frac{\partial\Theta}{\partial u^A_{m-1}}, \frac{\partial\Theta}{\partial u^B_{n-1}}\right\} = 0 \tag{28}$$

for $m, n \geqslant 1$; (28) are all that Θ has to satisfy. The equations for $m = n = 1$ give a hyper-Kähler version of the Plebanski "second heavenly equations."[6]

The above situation is indeed reminiscent of the role of the τ function in soliton theory: This is a major reason why Θ may be thought of as an analog of the τ-function.

Despite this remarkable similarity, it is at present very difficult to analyze the detailed structure of Θ, mostly because of the lack of such an explicit parametrization as that available in the case of the τ function.[4,5] In this respect it will be of much importance to seek as many examples as possible for which one can compute Θ in a closed form: In Secs. V and VI we present such cases, each of which has an analog in the theory of nonlinear integrable systems.

V. LEGENDRE TRANSFORMATIONS

The first example for computing Θ is derived from a series of discrete transformations of solutions. For notational convenience let us rewrite u^A for odd A as $\{u^a;\ 1 \leqslant a \leqslant r\}$ and u^A for even A as $\{v^a;\ 1 \leqslant a \leqslant r\}$. Thus

$$\tfrac{1}{2}\epsilon_{AB}\, du^A \wedge du^B = du^1 \wedge dv^1 + \cdots + du^r \wedge dv^r\,. \tag{29}$$

The transformations we now construct have a set of discrete parameters $l(a) \in \mathbb{Z}$, $1 \leqslant a \leqslant r$. Given such integers $l = \{l(a)\}$ the transformations send a solution $\{u^a(\lambda), v^a(\lambda)\}$ into $\{u'^a(\lambda), v'^a(\lambda)\}$ as

$$u'^1(\lambda) := \lambda^{l(1)}u^1(\lambda), \ldots, u'^r(\lambda) := \lambda^{l(r)}u^r(\lambda)\,, \tag{30a}$$

$$v'^1(\lambda) := \lambda^{-l(1)}v^1(\lambda), \ldots, v'^r(\lambda) := \lambda^{-l(r)}v^r(\lambda)\,. \tag{30b}$$

For simplicity let us consider the case where

$$l(a) \geqslant 0 \text{ for all } a = 1,\ldots,r\,, \tag{31}$$

although the following results can be readily extended to a general case.

Evidently the transformation (30) above retains the symplectic form and hence, every equation in the hyper-Kähler hierarchy. For this to make sense as a transformation of solutions, however, it must be further ensured that $u'^a_n = u^a_{n-l(a)}$, $v'^a_n = v^a_{n+l(a)}$ $(n \geqslant 0)$ can be chosen as new independent variables. A sufficient condition ("regularity") for this is

$$\det \frac{\partial(\ldots,u^a_{-1},\ldots,u^a_{-l(a)},\ldots)}{\partial(\ldots,v^a_0,\ldots,v^a_{l(a)-1},\ldots)} \neq 0\,, \tag{32}$$

where a inside the Jacobi determinant ranges over $\{1,\ldots,r\}$, but if $l(a) = 0$ the corresponding rows and columns are omitted; thus the determinant becomes of the size $\Sigma_{a=1}^{r} l(a)$.

Let us consider how Θ transforms. One can compute $d\Theta'$, where Θ' denotes the first key function after the transformation, as

$$d\Theta' = \sum_{a=1}^{r} \sum_{n\geqslant 0} (u'^a_{-n-1}\, dv'^a_n - v'^a_{-n-1}\, du'^a_n)$$
$$= \sum_{a=1}^{r} \sum_{n\geqslant 0} (u^a_{-n-1}\, dv^a_n - v^a_{-n-1}\, du^a_n)$$
$$- d\left(\sum_{a=1}^{r} \sum_{n=0}^{l(a)-1} u^a_{-n-1}\, v^a_n\right)$$
$$= d\left(\Theta - \sum_{a=1}^{r} \sum_{n=0}^{l(a)-1} u^a_{-n-1}\, v^a_n\right). \tag{33}$$

Therefore,

$$\Theta' = \Theta - \sum_{a=1}^{r} \sum_{n=0}^{l(a)-1} u^a_{-n-1}\, v^a_n \quad (+\text{const}). \tag{34}$$

Terms with $l(a)=0$ are again understood to be omitted from the rhs. This is a kind of "Legendre transformation."

In the language of Θ and Θ' the above solution procedure may be rephrased as follows. First, define Θ' to be a function of u'^a_n $(n\geqslant 0)$ and v'^a_n $[n\geqslant -l(a)]$ as

$$\Theta' = \Theta - \sum_{a=1}^{r} \sum_{n=0}^{l(a)-1} u'^a_n\, v'^a_{-n-1}, \tag{35}$$

where Θ is regarded as a function of u'^a_n $[n\geqslant l(a)]$ and v'^a_n $[n\geqslant -l(a)]$ under the relation $u'^a_n = u^a_{n-l(a)}$, $v'^a_n = v^a_{n+l(a)}$; then, solve the equations (nonlinear, in general)

$$\frac{\partial\Theta'}{\partial v'^a_n} = \frac{\partial\Theta}{\partial v'^a_n} - u'^a_{n+l(a)} = 0 \quad (-l(a)\leqslant n\leqslant -1) \tag{36}$$

with respect to v'^a_n $[-l(a)\leqslant n\leqslant -1]$ and eliminate them from the above expression of Θ', which gives the final answer. The regularity condition (32) now reads as

$$\det\left(\frac{\partial^2\Theta}{\partial v^a_m\, \partial v^b_n}; 1\leqslant a,b\leqslant r, 0\leqslant m,n\leqslant l(a)-1\right)\neq 0. \tag{37}$$

From a "seed" solution, one can thus obtain a series of new solutions as long as the "regularity" conditions (32) and (37) are satisfied. For example, let us take a hyper-Kähler version of the so-called "complex *pp* wave" as such a seed solution, for which

$$\Theta = \frac{1}{2\pi i}\oint F\left(\sum_{n=0}^{\infty} v^1_n \lambda^n, \ldots, \sum_{n=0}^{\infty} v^r_n \lambda^n, \lambda\right) d\lambda, \tag{38}$$

where F is an arbitrary function of $r+1$ variables. The Legendre transformation with $l(1)=\cdots=l(r)=1$ then gives rise to a class of solutions which are essentially the same as those presented by Hitchin *et al.*[10]

These solutions may be thought of as analogs of the so-called "Atiyah–Ward ansatz" for self-dual gauge fields.[13,14] The point of view of "transformations" as presented above is also parallel to the interpretation of the Atiyah–Ward ansatz discussed by Ueno and Nakamura[15] as "Riemann–Hilbert transformations." The existence of such a series of self-dual (i.e., $r=1$) metrics was conjectured by Ward.[16]

VI. GINDIKIN'S CONSTRUCTION IN HYPER-KÄHLER HIERARCHY

The second example for computing Θ is inspired by a construction of Gindikin.[8] We again use u^a, v^a instead of u^A. The solution is written as

$$u^a(\lambda) = \sum_{n=0}^{\infty} u^a_n \lambda^n + \sum_{i=1}^{I} \frac{f^a_i}{\lambda-\alpha_i}, \tag{39a}$$

$$v^a(\lambda) = \sum_{n=0}^{\infty} v^a_n \lambda^n + \sum_{j=1}^{J} \frac{g^a_j}{\lambda-\beta_j}, \tag{39b}$$

where α_i and β_j are constants such that $\alpha_i\neq\beta_j$ $(\forall i,j)$; f^a_i and g^a_j are functions of u^a_n and v^a_n $(n\geqslant 0)$ defined through

$$f^a_i = F_{i,a}(v(\alpha_i)) \quad (1\leqslant i\leqslant I), \tag{40a}$$

$$g^a_j = G_{j,a}(u(\beta_j)) \quad (1\leqslant j\leqslant J), \tag{40b}$$

where $F_i = F_i(v^1,\ldots,v^r)$ $(1\leqslant i\leqslant I)$ and $G_j = G_j(u^1,\ldots,u^r)$ $(1\leqslant j\leqslant J)$ are arbitrary functions and $F_{i,a}$ and $G_{i,a}$ denote their derivatives $F_{i,a} := \partial F_i/\partial v^a$, $G_{j,a} := \partial G_j/\partial u^a$. We assume some "regularity" condition to ensure the existence of a solution of Eqs. (40a) and (40b).

To check that the above construction gives a solution of the hyper-Kähler hierarchy, let us note that $\Sigma_{a=1}^{r} du^a(\lambda)\wedge dv^a(\lambda)$ does not have poles at any $\lambda\neq\infty$ (so that its Laurent expansion around $\lambda=\infty$ consists of just non-negative powers of λ) if and only if the residues at $\lambda=\alpha_i$ and β_j vanish. This results in

$$\sum_{a=1}^{r} df^a_i \wedge dv^a(\alpha_i)$$
$$= \mathrm{res}_{\lambda=\alpha_i} \sum_{a=1}^{r} du^a(\lambda)\wedge du^a(\lambda) = 0, \tag{41a}$$

$$\sum_{a=1}^{r} du^a(\beta_j)\wedge dg^a_j$$
$$= \mathrm{res}_{\lambda=\beta_j} \sum_{a=1}^{r} du^a(\lambda)\wedge du^a(\lambda) = 0. \tag{41b}$$

It is easy to see that f^a_i and g^a_j defined in Eqs. (40a) and (40b) indeed satisfy Eqs. (41a) and (41b). This construction is very similar to a method found in the theory of nonlinear integrable systems (see, e.g., Ref. 17).

Let us compute the second key function Θ. Since

$$1/(\lambda-\alpha_i) = \lambda^{-1} + \alpha_i\lambda^{-2} + \cdots,$$
$$1/(\lambda-\beta_j) = \lambda^{-1} + \beta_j\lambda^{-2} + \cdots \tag{42}$$

around $\lambda=\infty$, the Laurent coefficients of $u^a(\lambda)$ and $v^a(\lambda)$ can be written as

$$u^a_{-n-1} = \sum_{i=1}^{I} f^a_i a^n_i \quad (n\geqslant 0), \tag{43a}$$

$$v^a_{-n-1} = \sum_{j=1}^{J} g^a_j \beta^n_j \quad (n\geqslant 0). \tag{43b}$$

Then

$$d\Theta = \sum_{a=1}^{r} \sum_{n>0} \left(\sum_{i=1}^{I} f_i^a \alpha_i^n \, dv_n^a - \sum_{j=1}^{J} g_j^a \beta_j^n \, du_n^a \right)$$
$$= \sum_{a=1}^{r} \sum_{i=1}^{I} f_i^a \, d\left[v^a(\alpha_i) - \sum_{j=1}^{J} \frac{g_j^a}{\alpha_i - \beta_j} \right]$$
$$- \sum_{a=1}^{r} \sum_{j=1}^{J} g_j^a \, d\left[u^a(\beta_j) - \sum_{i=1}^{I} \frac{f_i^a}{\beta_j - \alpha_i} \right]$$
$$= d\left[\sum_{i=1}^{I} F_i(v(\alpha_i)) - \sum_{j=1}^{J} G_j(u(\beta_j)) \right.$$
$$\left. - \sum_{a=1}^{r} \sum_{i,j} \frac{f_i^a g_j^a}{\alpha_i - \beta_j} \right]. \tag{44}$$

Therefore,

$$\Theta = \sum_{i=1}^{I} F_i(v(\alpha_i)) - \sum_{j=1}^{J} G_j(u(\beta_j))$$
$$- \sum_{a=1}^{r} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{f_i^a g_j^a}{\alpha_i - \beta_j} \quad (+\text{const}) . \tag{45}$$

The key function thus obtained has a very suggestive form: This allows, for example, the physical interpretation that follows. The first two blocks on the rhs of formula (45) were originally "complex *pp* waves" of a very degenerate form. In each block the principle of "linear superposition" is realized; this is not the case between the two blocks because they are of a different type ("right/left handed"). The terms in the last sum of (45) give a correction caused by their "nonlinear superposition."

It is remarkable that the construction, just as in the case of the Legendre transformations, includes the step of solving a set of finite-dimensional, but mostly *nonlinear* equations. In other words, one is forced to solve "implicit functions." This circumstance is decisively different from traditional nonlinear integrable systems which, in a similar situation, reduce to linear equations. This is also a manifestation of the difference of underlying group structures.

Further, the construction shown above can be extended to the case with multiple poles at α_i and β_j.

VII. THE FIRST KEY FUNCTION AND ANOTHER SET OF HIDDEN VARIABLES

In Secs. II–VI we have focused on the second key function and related hidden variables. The second key function has the advantage of simplifying relevant mathematical structures. However, in possible applications to physics the first key function (i.e., the Kähler potential) plays a far more important role. Let us briefly discuss how to deal with the first key function in our approach.

In the treatment of Ω a basic coordinate system is given by $(u_1^A, \hat{u}_0^A)$, rather than (u_1^A, u_0^A). Here u_1^A and $\hat{u}_0^A$ correspond to z and $\bar{z}$ in the ordinary setting of Kähler geometry: They should therefore be treated on an equal footing. This suggests that we can introduce new coefficients in $u^A(\lambda)$ and $\hat{u}^A(\lambda)$ as

$$u^A(\lambda) = \sum_{n=-\infty}^{\infty} u_n^A \lambda^n , \tag{46a}$$

$$\hat{u}^A(\lambda) = \sum_{n=-\infty}^{\infty} \hat{u}_n^A \lambda^n , \tag{46b}$$

with the exterior differential equation

$$\epsilon_{AB} \, du^A(\lambda) \wedge du^B(\lambda) - \epsilon_{AB} \, d\hat{u}^A(\lambda) \wedge d\hat{u}^B(\lambda) = 0 , \tag{47}$$

where u_n^A $(n \geqslant 1)$ and $\hat{u}_n^A$ $(n \leqslant 0)$ are understood as independent variables and the other coefficients are understood as dependent variables (unknown functions).

Equations (46) and (47) also have several equivalent expressions. One can indeed find their explicit forms (cf. the Appendix), although they become considerably complicated. Here let us just show the counterpart of the "Hamiltonian form" [cf. Eq. (17)] because the hierarchy structure becomes most manifest in that expression. In fact, two different symplectic structures are available; one is given on the (u_1^A) space with the symplectic form $\epsilon_{AB} \, du_1^A \wedge du_1^B/2$ and the other is given on the $(\hat{u}_0^A)$ space with $\epsilon_{AB} \, d\hat{u}_0^A \wedge d\hat{u}_0^B/2$. In the first setting the notion of Hamiltonian vector fields and Poisson brackets are defined as

$$H^{(1)}(F) := \epsilon^{AB} \frac{\partial F}{\partial u_1^A} \frac{\partial}{\partial u_1^B} , \tag{48}$$

$$\{F_1, F_2\}^{(1)} := H^{(1)}(F_1) F_2 , \tag{49}$$

where the superscript "(1)" has been added to distinguish them from those in (20), (21). An equivalent expression of the exterior differential equation (47) is the Hamiltonian system

$$\frac{\partial w_A(\lambda)}{\partial u_n^B} = \{ w_A(\lambda), (\lambda^{n-1} u_B(\lambda))_+ \}^{(1)} \quad (n \geqslant 1) , \tag{50a}$$

$$\frac{\partial w_A(\lambda)}{\partial \hat{u}_n^B} = \{ w_A(\lambda), (\lambda^{n-1} \hat{u}_B(\lambda))_- \}^{(1)} , \quad (n \leqslant 0) \tag{50b}$$

for $w_A(\lambda) = u_A(\lambda), \hat{u}_A(\lambda)$ coupled with the constraints

$$\{ u^A(\lambda), u^B(\lambda) \}^{(1)} = \lambda^2 \epsilon^{AB} , \tag{50c}$$

$$\{ \hat{u}^A(\lambda), \hat{u}^B(\lambda) \}^{(1)} = \lambda^2 \epsilon^{AB} . \tag{50d}$$

Now the notion of the first key function Ω can be extended to the new hierarchy (50) as

$$d\Omega = - \sum_{n \geqslant 1} \epsilon_{AB} u_{-n+1}^A \, du_n^B + \sum_{n \leqslant 0} \epsilon_{AB} \hat{u}_{-n+1}^A \hat{u}_n^B . \tag{51}$$

The closedness of the rhs of (51) follows from the λ term of the exterior differential equation (47). The unknown functions are thus written in terms of a single one, with which the hierarchy (50) takes the following form:

$$\frac{\partial^2 \Omega}{\partial u_m^A \, \partial u_{n-1}^B} - \frac{\partial^2 \Omega}{\partial u_{m-1}^A \, \partial u_n^B} + \left\{ \frac{\partial \Omega}{\partial u_{m-1}^A}, \frac{\partial \Omega}{\partial u_{n-1}^B} \right\}^{(1)} = 0 , \tag{52a}$$

$$\frac{\partial^2 \Omega}{\partial u_m^A \, \partial \hat{u}_{n-1}^B} - \frac{\partial^2 \Omega}{\partial u_{m-1}^A \, \partial \hat{u}_n^B} - \left\{ \frac{\partial \Omega}{\partial u_{m-1}^A}, \frac{\partial \Omega}{\partial \hat{u}_{n-1}^B} \right\}^{(1)} = 0 , \tag{52b}$$

$$\frac{\partial^2 \Omega}{\partial \hat{u}_m^A \, \partial \hat{u}_{n-1}^B} - \frac{\partial^2 \Omega}{\partial \hat{u}_{m-1}^A \, \partial \hat{u}_n^B} + \left\{ \frac{\partial \Omega}{\partial \hat{u}_{m-1}^A}, \frac{\partial \Omega}{\partial \hat{u}_{n-1}^B} \right\}^{(1)} = 0 , \tag{52c}$$

$$\frac{\partial^2\Omega}{\partial\hat{u}_0^A\,\partial u_n^B}-\left\{\frac{\partial\Omega}{\partial\hat{u}_o^A},\frac{\partial\Omega}{\partial u_{n-1}^B}\right\}^{(1)}=0\,, \tag{52d}$$

$$\frac{\partial^2\Omega}{\partial\hat{u}_0^A\,\partial\hat{u}_n^B}-\left\{\frac{\partial\Omega}{\partial\hat{u}_0^A},\frac{\partial\Omega}{\partial\hat{u}_{n-1}^A}\right\}^{(1)}=0\,, \tag{52e}$$

$$\left\{\frac{\partial\Omega}{\partial\hat{u}_0^A},\frac{\partial\Omega}{\partial\hat{u}_0^B}\right\}^{(1)}=\epsilon_{AB}\,, \tag{52f}$$

where n and m range over all the possible values of integers [e.g., m, $n\geqslant 2$ in Eq (52a)]. Equation (52f) is nothing less than a hyper-Kähler version of the "first heavenly equation" of Plebanski.[6]

The special solutions presented in Secs. V and VI can be extended to the hierarchy (52); Ω also has a similar form.

The second key functions Θ and $\hat{\Theta}$ are also meaningful in the above setting: One can redefine them as

$$d\Theta=\sum_{n\geqslant 0}\epsilon_{AB}u^A_{-n-1}u^B_n-\sum_{n\leqslant -1}\epsilon_{AB}\hat{u}^A_{-n-1}\,d\hat{u}^B_n\,, \tag{53}$$

$$d\hat{\Theta}=-\sum_{n\geqslant 2}\epsilon_{AB}u^A_{-n+3}u^B_n+\sum_{n\leqslant 1}\epsilon_{AB}\hat{u}^A_{-n+3}\,d\hat{u}^B_n\,, \tag{54}$$

where $\{u^A_n\ (n\geqslant 0),\ \hat{u}^A_n\ (n\leqslant -1)\}$ and $\{u^A_n\ (n\geqslant 2),\ \hat{u}^A_n\ (n\leqslant 1)\}$, respectively, are understood as independent variables for each case. In view of relations (51), (53), and (54) one will find that there is evidently no substantial difference among the key functions Ω, Θ, and $\hat{\Theta}$, with each connected with the other through a shift of the integer indices.

VIII. CONCLUSION

Hyper-Kähler metrics (in a complexified setting) thus have three types of hidden variables, each of which has a counterpart in the theory of nonlinear integrable systems. With this analogy the notion the "hyper-Kähler hierarchy" is introduced, which is an analog of hierarchy structures of nonlinear integrable systems such as the "KP hierarchy," etc. In this setting the Plebanski key functions (both first and second) indeed play a "key" role as a "generating function" of all unknown functions of the system. This seems to explain an ultimate meaning of the key functions which was hidden, or, at least, unclear in the original framework of Plebanski[6] and Boyer and Plebanski.[7] As illustrated in two examples of special solutions, the key functions are useful not only for the analysis of mathematical concepts, but also for the study of special solutions.

Another role to be played by these hidden variables (in particular, the dependent variables) can be found in the analysis of "hidden symmetries" of the system. This is indeed the case for other various nonlinear integrable systems; it turns out that the hyper-Kähler case also has a large set of hidden symmetries. This fact was first pointed out by Boyer and Plebanski[7] as a structure of "nonlinear superposition." A crucial point that distinguishes the hyper-Kähler case is the difference of an underlying group structure; the Lie algebra of hidden symmetries is the loop algebra of infinitesimal canonical transformations (Hamiltonian vector fields). This explains why Hamiltonian vector fields and Poisson brackets occur in every aspect of the relevant differential equations. A detailed analysis of hidden symmetries will be reported in a subsequent paper.

Although the physical meaning of these hidden variables is less clear from the present context, one can imagine that they might emerge in some crucial part of applications such as in a detailed analysis of the renormalization structure of hyper-Kähler sigma models.[18]

ACKNOWLEDGMENTS

The author is sincerely grateful to Dr. Izumi Ojima and Hiroaki Kanno for fruitful discussions.

APPENDIX: PROOF OF THE PROPOSITION

In what follows we give a proof of the proposition presented in Sec. II. We will then discuss how to extend the proof into the setting of Sec. VII.

Equivalence of (14) and (15): Since the Poisson commutation relation $\{u^A,u^B\}=\epsilon^{AB}$ means that $(\partial u^A/\partial u_0^B)$ takes values in $\mathrm{Sp}(r,\mathbb{C})$, (14) readily follows from (15). To prove the converse, let us consider the contraction of the two-form $\frac{1}{2}\epsilon_{AB}du^A\wedge du^B$ with the vector fields $\partial/\partial u_0^C$ and $\partial/\partial u_0^D$, which can be written as

$$\begin{aligned}\frac{1}{2}\epsilon_{AB}\,du^A\wedge du^B\left(\frac{\partial}{\partial u_0^C},\frac{\partial}{\partial u_0^D}\right)&=\epsilon_{AB}\frac{\partial u^A}{\partial u_0^C}\frac{\partial u^B}{\partial u_0^D}\\&=\epsilon_{CD}+(\text{negative powers of }\lambda)\,;\end{aligned} \tag{A1}$$

however, this should be made up of non-negative powers of λ because of the assumption. The rhs of (A1) is thereby equal to ϵ_{CD}; therefore, $(\partial u^A/\partial u_0^B)$ is symplectic. The Poisson commutation relation in (15) thus follows. This also implies that the one-forms

$$e^A(\lambda):=\frac{\partial u_B(\lambda)}{\partial u_{A0}}\,du^B(\lambda) \tag{A2}$$

satisfy the relation

$$du^A(\lambda)=\frac{\partial u^A(\lambda)}{\partial u_0^B}\,e^B(\lambda) \tag{A3}$$

[$(\partial u_B/\partial u_{A0})$ gives the inverse of $(\partial u^A/\partial u_0^B)$]. On the other hand, from the definition,

$$e^A(\lambda)=\tfrac{1}{2}\iota_{\partial/\partial u_A}(\epsilon_{BC}\,du^B(\lambda)\wedge du^C(\lambda))\,, \tag{A4}$$

where ι_∂, in general, stands for the left contraction ("inner derivative") with a vector field ∂. From (15) the rhs does not contain negative powers of λ, i.e., $(e^A(\lambda))_-=0$. Thus (15) $\Rightarrow$ (14).

Equivalence of (15)–(17): The equivalence of (15)–(17) is due to the equivalence of two expressions of the theorem of Frobenius, i.e., those with one-forms and those with vector fields. A cotangent frame ("vielbein") $\{e^a\}$ of linearly independent one-forms on a manifold determines, in general, a "dual" tangent frame $\{\partial_a\}$ of linearly independent vector fields through the normalization relation

$$\langle e^a,\partial_b\rangle(:=\iota_{\partial_b}(e^a))=\delta^a_b\,. \tag{A5}$$

This correspondence is bilateral and under relation (A5) the total derivative of any function can be written as

$$df = \partial_\alpha f \cdot e^\alpha . \tag{A6}$$

From (A6), in particular, for any index subset I,

$$df = \sum_{\alpha \in I} \partial_\alpha f \cdot e^\alpha \Leftrightarrow \partial_\alpha f = 0 \quad (\alpha \in I^c) , \tag{A7}$$

where I^c denotes the complement of I in the whole index set of coordinates. Bearing the general framework given above in mind, let us consider the cotangent frame $\{e^A(\lambda), e_n^A(\lambda)\ (1 \leqslant A \leqslant 2r;\ 1 \leqslant n < \infty)\}$, where $e_n^A(\lambda) = \Sigma_{k=0}^{\infty} du_{n+k}^A \lambda^k$, in the space of the independent variables $\{u_n^A;\ n = 0, 1, \ldots\}$. Precisely, $e^A(\lambda)$ are written as

$$e^A(\lambda) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{\partial u_{B,-k}}{\partial u_{A0}} \lambda^{n-k} du_n^B . \tag{A8}$$

The dual tangent frame of (A8) is $\{\partial/\partial u_0^A, \partial_A^n(\lambda)\ (1 \leqslant A \leqslant 2r;\ 1 \leqslant n)\}$; the equivalence of (15) and (16) thus follows. If, in turn, one takes the cotangent frame $\{e^A(\lambda), du_n^A\ (1 \leqslant A \leqslant 2r;\ 1 \leqslant n)\}$, the dual is given by $\{\partial/\partial u_0^A, \partial/\partial u_n^A + H((\lambda^n u_A(\lambda))_+)\ (1 \leqslant A \leqslant 2r;\ 1 \leqslant n)\}$. The equivalence with (17) follows from this observation.

Equivalence of (17) and (19): This part of the proof requires somewhat strange ideas, as follows. Let us first derive (19) from (17). Replacing $A \leftrightarrow B$ and $m \leftrightarrow n$ in (17a), taking the difference with the original form, and also using the Poisson commutation relation (17b), one obtains

$$\frac{\partial(\lambda^n u_B)_+}{\partial u_m^A} - \frac{\partial(\lambda^m u_A)_+}{\partial u_n^B} + \{(\lambda^m u_A)_+, (\lambda^n u_B)_+\} + \lambda^{m+n}\epsilon_{AB} + \frac{\partial(\lambda^n u_B)_-}{\partial u_m^A} - \frac{\partial(\lambda^m u_A)_-}{\partial u_n^B} + \{(\lambda^m u_A)_-, (\lambda^n u_B)_-\} = 0 . \tag{A9}$$

The lhs of (A9) consists of two parts, where the first four terms are made up of non-negative powers of λ and the others are made up of negative ones. Each of these two parts should thereby vanish, leading to (19).

The converse requires another trick. We multiply Eq. (19) with λ^{-m-n} and consider the limit as $m, n \to \infty$. Each term in the result can be evaluated as follows:

$$\lambda^{-m-n} \frac{\partial(\lambda^n u_B)_+}{\partial u_m^A} = \epsilon_{BA} + O(\lambda^{-m-1}) , \tag{A10a}$$

$$\lambda^{-m-n} \frac{\partial(\lambda^m u_A)_+}{\partial u_n^B} = \epsilon_{AB} + O(\lambda^{-n-1}) , \tag{A10b}$$

$$\lambda^{-m-n}\{(\lambda^n u_A)_+, (\lambda^m u_B)_+\} = \{u_A, u_B\} + O(\lambda^{-\min(m,n)-1}) , \tag{A10c}$$

$$\lambda^{-m-n} \cdot \lambda^{m+n}\epsilon_{AB} = \epsilon_{AB} , \tag{A10d}$$

where $O(\lambda^{-k})$, in general, stands for a linear combination of $\lambda^{-k}, \lambda^{-k-1}, \ldots$. From relations (A10),

$$\{u_A, u_B\} = \epsilon_{AB} + O(\lambda^{-\min(m,n)-1}); \tag{A11}$$

therefore, in the limit as $m, n \to \infty$ one obtains (17b). To derive (17a), we now multiply (19) with λ^{-n} and evaluate each term in much the same way as above. Then

$$\frac{\partial u_B}{\partial u_m^A} + \{(\lambda^m u_A)_+, u_B\} = O(\lambda^{m-n}) . \tag{A12}$$

The limit of (A12) as $n \to \infty$ gives (17a), completing the proof of the equivalence of (17) and (19).

Equivalence of (18) and (19): One can check the equivalence of (18) and (19) by simply forming appropriate linear combinations of equations on each system. We omit the details.

The proposition is thus proved.

Extending these arguments to the setting of Sec. VII is rather straightforward, although somewhat complicated. The first step is to rewrite the exterior differential equation (47) into the following system:

$$\frac{\partial u_B(\lambda)}{\partial u_{A1}} du^B(\lambda) = \frac{\partial \hat{u}_B(\lambda)}{\partial \hat{u}_{A1}} d\hat{u}^B(\lambda) , \tag{A13a}$$

$$\{u^A(\lambda), u^B(\lambda)\}^{(1)} = \lambda^2 \epsilon^{AB} , \tag{A13b}$$

$$\{\hat{u}^A(\lambda), \hat{u}^B(\lambda)\}^{(1)} = \lambda^2 \epsilon^{AB} . \tag{A13c}$$

Let us define the one-forms $e^{(1)A}(\lambda)$ with the lhs and rhs of the first equations: Unlike $e^A(\lambda)$ they include all integer powers of λ, but one can construct a system of vector fields perpendicular to them anyway. A choice is $\{\partial/\partial u_n^A + H^{(1)}((\lambda^{n-1} u_A(\lambda))_+)\ (n \geqslant 1),\ \partial/\partial \hat{u}_n^A + H^{(1)}((\lambda^{n-1}\hat{u}_A(\lambda))_-)\ (n \leqslant 0)\}$. These vector fields give rise to linear systems for $u^A(\lambda)$ and $\hat{u}^A(\lambda)$ [which correspond to Eqs. (16)]: Rewriting them as in Eqs. (17) and (18) can be done in much the same way as above.

[1] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Phys. Rev. Lett. **19**, 1095 (1967).
[2] P. D. Lax, Commun. Pure. Appl. Math. **21**, 467 (1968).
[3] R. Hirota, in *Bäcklund Transformations, the Inverse Scattering Method, Solitons, and Their Applications,* edited by R. Miura, Lecture Notes in Mathematics 515 (Springer, New York 1976).
[4] M. Sato and Y. Sato, in *Proccedings of the U. S.–Japan Seminar on Nonlinear Partical Differential Equations in Applied Science,* Tokyo, 1982, edited by P. D. Lax, H. Fujita, and G. Strang (North-Holland, Amsterdam and Kinokuniya, Tokyo, 1982).
[5] E. Date, M. Kashiwara, M. Jimbo, and T. Miwa, in *Nonlinear Integrable Systems—Classical Theory and Quantum Theory,* Kyoto, 1981, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983).
[6] J. F. Plebanski, J. Math. Phys. **16**, 2395 (1975).
[7] C. P. Boyer and J. F. Plebanski, J. Math. Phys. **26**, 229 (1985).
[8] S. G. Gindikin, Funct. Anal. Appl. **20**, 238 (1986).
[9] R. Penrose, Gen. Rel. Grav. **7**, 31 (1976).
[10] N. J. Hitchin, A. Kahlhede, U. Lindström, and M. Roček, Commun. Math. Phys. **108**, 535 (1987).
[11] C. LeBrun, Trans. AMS **278**, 209 (1983).
[12] K. Takasaki, Publ. RIMS Kyoto Univ. **22**, 949 (1986).
[13] M. F. Atiyah and R. S. Ward, Commun. Math. Phys. **55**, 117 (1977).
[14] E. T. Corrigan, D. B. Fairlie, R. G. Yates, and P. Goddard, Commun. Math. Phys. **58**, 223 (1978).
[15] K. Ueno and Y. Nakakmura, Publ. RIMS Kyoto Univ. **19**, 943 (1983).
[16] R. S. Ward, in *Complex Manifold Technique in Theoretical Physics,* edited by D. E. Lerner and P. D. Sommers (Pitman, London, 1978).
[17] J. Harnad, Y. Saint-Aubin, and S. Shnider, Commun. Math. Phys. **93**, 33 (1984).
[18] L. Alvarez-Gaumé and D. Z. Freedman, Commun. Math. Phys. **80**, 443 (1981).

On the time evolution operator for time-dependent quadratic Hamiltonians

Francisco M. Fernández
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), División Química Teórica, Sucursal 4, Casilla de Correo 16, (1900) La Plata, Argentina

(Received 30 August 1988; accepted for publication 15 February 1989)

The Schrödinger equation with a time-dependent quadratic Hamiltonian is investigated. The time-evolution operator is written as a product of exponential operators determined by the Heisenberg equations of motion. This product operator is shown to be global in the occupation number representation when the Hamiltonian is Hermitian. The success of some physical applications of the product-form representation is explained.

I. INTRODUCTION

There has recently been great interest in the Schrödinger equation with a Hamiltonian that can be written as a linear combination of operators which span a finite-dimensional Lie algebra.[1,2] Time-dependent quadratic Hamiltonians, for instance, prove to be useful models in optics[3,4] and in studying collisional energy transfer.[5,6]

In such cases the time-evolution operator is most frequently written as a product of simple exponential operators (the so-called uncoupling theorem) because it offers many advantages[1,2,4–6] with respect to the exponential form.[7] However, even the product form so widely used has not, in general, been proved to be global; i.e., valid for all time values. It has been known since long ago that this representation is global for all solvable Lie algebras and for any real 2×2 system of equations.[8,9] In other cases the product-form solution has only been proved to exist in a neighborhood of the initial time.[8,9]

It has been claimed[4,9] that the uncoupling theorem is global for the split three-dimensional simple Lie algebra, but results seem to depend on the operator basis.[8] It is worth mentioning that the condition that the matrix ξ in Refs. 8 and 9 is invertible is not enough to ensure that the uncoupling theorem is global.[8] For this reason it is still necessary to determine the operator basis in every case so that the product form is global.

This paper addresses the above mentioned question in the case of the Schrödinger equation with a general time-dependent quadratic Hamiltonian.

The main results are obtained in Sec. II for the occupation number representation. Two illustrative examples are considered in Sec. III. Further comments and conclusions are found in Sec. IV.

II. TIME-EVOLUTION OPERATOR FOR QUADRATIC TIME-DEPENDENT HAMILTONIANS

Let $\mathbb{C}^{N\times M}$ be the set of all complex $N\times M$ matrices. A general time-dependent quadratic Hamiltonian operator has the following form:

$$H(t)=\mathbf{a}^\dagger F_1(t)\mathbf{a}+\tfrac{1}{2}[\mathbf{a}^T F_2(t)\mathbf{a}+\mathbf{a}^\dagger F_9(t)\mathbf{a}_+]+F_4^T(t)\mathbf{a}+\mathbf{a}^\dagger F_5(t)+F_6(t)\hat{1}, \tag{1}$$

where F_1, F_2, F_3 ($F_2^T=F_2$, $F_3^T=F_3$) belong to $\mathbb{C}^{N\times N}$, F_4, F_5 belong to $\mathbb{C}^{N\times 1}$, and F_6 is a scalar. $\hat{1}$ is the identity operator and the superscripts T and $\dagger$ stand for transpose and adjoint, respectively. For the sake of simplicity the following notation is used:

$$\mathbf{a}=\begin{pmatrix}a_1\\ a_2\\ \vdots\\ a_N\end{pmatrix},\quad \mathbf{a}^\dagger=(a_1^\dagger a_2^\dagger\cdots a_N^\dagger),\quad a_+=(a^\dagger)^T, \tag{2}$$

where $a_j^\dagger$ and a_k are the creation and annihilation operators, respectively, which obey $[a_j,a_k^\dagger]=\delta_{jk}$ or, in matrix notation,

$$[\mathbf{a},\ \mathbf{a}^\dagger]=I, \tag{3}$$

where I is the $N\times N$ identity matrix. The operator (1) will not be assumed to be Hermitian in order to take into account gain or loss mechanisms, as in the case of the evolution of the optical field in a free-electron laser.[2,4] Nevertheless, $H(t)$ will be called the Hamiltonian operator.

The Schrödinger equation reads

$$\frac{d}{dt}U(t)=-iH(t)U(t),\quad U(0)=\hat{1}, \tag{4}$$

where the time-evolution operator will not be unitary unless $H^\dagger=H$. Since $\{\hat{1}, a_j, a_k^\dagger, a_ja_k^\dagger, a_ja_k, a_j^\dagger a_k^\dagger, j,k=1,2,\ldots,N\}$ generates a $(2N^2+3N+1)$-dimensional Lie algebra L, the time-evolution operator can be written[1,2,8,9]

$$U=\prod_{j=1}^{6}U_j,\quad U_j=\exp(X_j), \tag{5a}$$

where

$$X_1=\mathbf{a}^\dagger G_1(t)\mathbf{a},\quad X_2=\mathbf{a}^T G_2(t)\mathbf{a},\quad X_3=\mathbf{a}^\dagger G_3(t)\mathbf{a}_+, \tag{5b}$$

$$X_4=G_4(t)\mathbf{a},\quad X_5=\mathbf{a}^\dagger G_5(t),\quad X_6=G_6(t)\hat{1},$$

G_1, G_2, G_3 ($G_2^T=G_2$, $G_3^T=G_3$), belong to $\mathbb{C}^{N\times N}$, G_4 and G_5 belong to $\mathbb{C}^{N\times 1}$, and G_6 is a scalar. The initial conditions are $G_j(0)=0, j=1,2,\ldots,6$.

Nonlinear first-order differential equations for the G_j's can be obtained from the equality[1,2,4–6,8,9]

$$H(t)=i\left[\frac{d}{dt}U(t)\right]U(t)^{-1}. \tag{6}$$

However, it is much simpler to proceed as discussed by Kolsrud,[9] Fernández and Castro,[10] and Fernández[11] for the coordinate and momentum representation.

 0022-2488/89/071522-03$02.50

To this end we define the time-dependent operators

$$\mathbf{a}(t) = U^{-1}\mathbf{a}U, \quad \mathbf{a}_+(t) = U^{-1}\mathbf{a}_+U, \tag{7}$$

which satisfy the following equations of motion:

$$\frac{d}{dt}\mathbf{a}(t) = iU^{-1}[H, \mathbf{a}]U,$$

$$\frac{d}{dt}\mathbf{a}_+(t) = iU^{-1}[H, \mathbf{a}_+]U. \tag{8}$$

Their solutions can be written

$$\mathbf{a}(t) = A_{10}(t) + A_{11}(t)\mathbf{a} + A_{12}(t)\mathbf{a}_+,$$

$$\mathbf{a}_+(t) = A_{20}(t) + A_{21}(t)\mathbf{a} + A_{22}(t)\mathbf{a}_+, \tag{9}$$

where A_{10}, A_{20} belong to $\mathbb{C}^{N\times 1}$ and A_{11}, A_{12}, A_{21}, A_{22} belong to $\mathbb{C}^{N\times N}$. Since $\mathbf{a}(0) = \mathbf{a}$ and $\mathbf{a}_+(0) = \mathbf{a}_+$, then $A_{10}(0) = A_{20}(0) = \mathbf{0}$, $A_{12}(0) = A_{21}(0) = \mathbf{0}$, and $A_{11}(0) = A_{22}(0) = I$.

A straightforward calculation using Eqs. (8) and (9) shows that

$$\frac{d}{dt}A_{1j} = -i(F_1(t)A_{1j} + F_3(t)A_{2j} + F_5(t)\delta_{j0}),$$

$$j = 0,1,2, \tag{10}$$

$$\frac{d}{dt}A_{2j} = i(F_1^T(t)A_{2j} + F_2(t)A_{1j} + F_4(t)\delta_{j0}).$$

Besides, since $[\mathbf{a}(t), \mathbf{a}^\dagger(t)] = I$ it is found that

$$A_{11}A_{22}^T - A_{12}A_{21}^T = I, \tag{11a}$$

$$A_{11}A_{12}^T - A_{12}A_{11}^T = \mathbf{0}, \tag{11b}$$

$$A_{21}A_{22}^T - A_{22}A_{21}^T = \mathbf{0}. \tag{11c}$$

On introducing Eqs. (5) into Eqs. (7) and using the well-known Baker–Hausdorff formula[8,9]

$$e^x y e^{-x} = \sum_{j=0}^{\infty}\frac{y_j}{j!},$$

$$y_j = [x, y_{j-1}], \quad j = 1,2,\ldots, \quad y_0 = y, \tag{12}$$

it is not difficult to prove that

$$\exp(G_1) = A_{11}, \quad G_2 = -A_{11}^T A_{21}, \quad G_3 = A_{11}^{-1}A_{12}, \tag{13a}$$

$$G_4 = A_{21}^T A_{10} - A_{11}^T A_{20}, \quad G_5 = A_{22}^T A_{10} - A_{12}^T A_{20}. \tag{13b}$$

In order to obtain Eqs. (13b) use has also been made of Eqs. (11). The function G_6 is easily obtained from Eq. (6). The only contribution to the coefficient of $\hat{1}$ comes from a term of the form

$$\cdots U_4^{-1}\frac{d}{dt}X_5U_4\cdots + \frac{d}{dt}G_6\hat{1} = F_6\hat{1} + \cdots .$$

Therefore

$$\frac{d}{dt}G_6 = F_6 - G_4\frac{d}{dt}G_5. \tag{13c}$$

Since in most physical problems the functions of time in $H(t)$ are analytic,[4–6] the solutions of the equations of motion (10) will also be analytic. Therefore, according to Eqs. (13) the product operator (5) exists for those t values for which $\det A_{11} \neq 0$. Furthermore, since $A_{11}(0) = I$ there is always a neighborhood of $t = 0$ where such a condition takes place. Clearly, the problem of finding the interval of validity of the product operator (5) has been greatly simplified. The main result of this paper is given in the theorem below.

Theorem: If H is Hermitian then the product operator (5) is global.

Proof: If $H^\dagger = H$ then $A_{10}^* = A_{20}$, $A_{11}^* = A_{22}$, $A_{12}^* = A_{21}$, and Eq. (11a) becomes

$$A_{11}A_{11}^\dagger - A_{12}A_{12}^\dagger = I. \tag{14}$$

Since the eigenvalues of $A_{12}A_{12}^\dagger$ are larger than or equal to zero, then the eigenvalues of $A_{11}A_{11}^\dagger$ will be larger than or equal to unity. Therefore

$$\det(A_{11}^\dagger A_{11}) = |\det A_{11}|^2 \geqslant 1, \tag{15}$$

for all t values.

The properties of the product operator depend largely on the operator basis used. For instance, the result given above does not hold when the coordinate and momentum operators are chosen to be the basis of the Lie algebra L.[10–12] This fact will be explained in more detail in the next section.

III. EXAMPLES

In order to illustrate the main result of the previous section we consider two exactly solvable one-dimensional problems. Since $N = 1$, the A_{jk}'s in Eqs. (9)–(11) are no longer matrices but scalar-valued complex functions of time. The first example is given by the operator

$$H(t) = \omega(a^\dagger a + \tfrac{1}{2}) + c/2[e^{2i\omega t}a^2 + e^{-2i\omega t}(a^\dagger)^2], \tag{16}$$

where ω and c are complex numbers. The operators $(a^\dagger a + aa^\dagger)/2$, $a^2/2$, and $(a^\dagger)^2/2$ span a realization of the split three-dimensional Lie algebra.

The equations of motion (10) can be exactly solved and the result is

$$A_{10} = A_{20} = 0, \quad A_{11} = e^{-i\omega t}\cosh(ct),$$
$$A_{12} = -ie^{-i\omega t}\sinh(ct),$$
$$A_{21} = ie^{i\omega t}\sinh(ct), \quad A_{22} = e^{i\omega t}\cosh(ct). \tag{17}$$

Clearly, if H is Hermitian then ω and c are real numbers and $|A_{11}| \geqslant 1$ in agreement with the theorem of the previous section. On the other hand, if c is purely imaginary ($c = i|c|$), then A_{11} vanishes for $t = (k + \frac{1}{2})\pi/|c|$, where $k = 0, \pm 1, \pm 2, \ldots$. Therefore, when H is not Hermitian, as in the case of the free-electron laser,[4] one has to be very careful in using the product form for the time-evolution operator.

In this case $U(t)$ can also be written

$$U(t) = \exp[-i\omega t(a^\dagger a + \tfrac{1}{2})]$$
$$\times\exp[-(i/2)ct\{a^2 + (a^\dagger)^2\}], \tag{18}$$

which certainly holds for all t values without restriction on ω and c.

The next example is the Hermitian operator

$$H(t) = \tfrac{1}{2}(e^{2t}p^2 + e^{-2t}q^2), \tag{19}$$

where $p = -i\,d/dq$. In this case $U(t)$ is unitary and can be written in terms of the Hermitian operators $(qp+pq)/2$, $q^2/2$, and $p^2/2$ which span another realization of the split three-dimensional Lie algebra.

The equations of motion for

$$p(t) = U^\dagger p U, \quad q(t) = U^\dagger q U, \tag{20}$$

are

$$\frac{d}{dt}p(t) = -e^{-2t}q(t), \quad \frac{d}{dt}q(t) = e^{2t}p(t). \tag{21}$$

Their solutions are easily found to be

$$\begin{aligned} p(t) &= P_p(t)p + P_q(t)q, \\ q(t) &= Q_p(t)p + Q_q(t)q, \end{aligned} \tag{22}$$

where

$$\begin{aligned} P_p &= (1+t)e^{-t}, \quad P_q = -te^{-t}, \\ Q_p &= te^t, \quad Q_q = (1-t)e^t. \end{aligned} \tag{23}$$

Notice that $P_pQ_q - P_qQ_p = 1$ for all t values.

Finally, the time-evolution operator can be written

$$\begin{aligned} U(t) = &\exp(iP_qq^2/2Q_q)\exp[-i\ln Q_q(qp+pq)/2] \\ &\times\exp(-iQ_pp^2/2Q_q), \end{aligned} \tag{24}$$

which does not hold when $t = 1$.

It can be easily verified that

$$\begin{aligned} U(t) = &\exp[-it(qp+pq)] \\ &\times\exp[-itp^2/2 - itq^2/2 + it(qp+pq)], \end{aligned} \tag{25}$$

also satisfies the Schrödinger equation with the Hamiltonian (19). It is therefore concluded that the form of the time-evolution operator in the coordinate-momentum representation has to be chosen carefully if a global operator is to be obtained. This result is due to the fact that there is no theorem in the coordinate-momentum representation similar to the one shown in the previous section for the occupation number representation.

Since

$$p = 2^{-1/2}i(a^\dagger - a), \quad q = 2^{-1/2}(a^\dagger + a), \tag{26}$$

the operator (19) can be written in the occupation number representation as

$$\begin{aligned} H = &\cosh(2t)(a^\dagger a + \tfrac{1}{2}) \\ &-\tfrac{1}{2}\sinh(2t)[a^2 + (a^\dagger)^2]. \end{aligned} \tag{27}$$

Besides, it follows from Eqs. (22) and (26) that

$$\begin{aligned} A_{11} = A_{22}^* &= \tfrac{1}{2}[P_p + Q_q + i(P_q - Q_p)], \\ A_{12} = A_{21}^* &= \tfrac{1}{2}[Q_p - P_p + i(Q_p + P_q)]. \end{aligned} \tag{28}$$

Therefore, $|A_{11}|^2 = \frac{1}{4}(P_p^2 + Q_q^2 + P_q^2 + Q_p^2 + 2)$ and the product operator in the occupation-number representation exists for all t values in agreement with the results of the previous section.

IV. CONCLUSIONS

There are many advantages in taking into account the quantum-mechanical equations of motion. First, since the time-evolution operator is expressed in terms of the solutions of first-order linear differential equations, numerical calculation, when necessary, is simpler.

Second, analytical results are more easily derived as shown by the theorem in Sec. II, which is of great importance for the many physical applications of the algebra L spanned by the operators $\{\hat{1}, a_j, a_k^\dagger, a_j^\dagger a_k, a_j a_k, a_j^\dagger a_k^\dagger, \{j,k = 1,2,...,N\}$ because it provides a global product-form time-evolution operator. It is hoped that this result will motivate the study of the equations of motion for other algebras.

Third, most of the physical properties of the system, such as matrix elements and transition probabilities, can be written in terms of the solutions of the equations of motion and are independent of the form given to the time-evolution operator.[5,12]

As stated before, the product-form time-evolution operator proves to be useful in studying collisional energy transfer.[5,6] It is interesting to pay attention to the way some of those results were obtained when the algebra was not solvable. Benjamin[6] could integrate the nonlinear equations of motion from $-\infty$ to $+\infty$ because the time-evolution operator was expressed in the occupation number representation. Gazdy and Micha,[5] on the other hand, used the coordinate-momentum representation which, as argued before, was not proved to give rise to a global product-form time-evolution operator. However, they expressed their final results in terms of the solutions of the equations of motion which can always be integrated without difficulty. It is worth mentioning that numerical integration of the nonlinear equations of Ref. 5 did not reveal any singular point in the solutions. However, such a result does not prove that singularities cannot occur in other cases. They were certainly found when using the Schrödinger picture instead of the interaction one.

[1] F. Wolf and H. J. Korsch, Phys. Rev. A **37**, 1934 (1988).
[2] G. Dattoli, M. Richetta, and A. Torre, Phys. Rev. A **37**, 2007 (1988).
[3] *Lie Methods in Optics, Lecture Notes in Physics*, edited by J. Sanchez Mondragón and K. B. Wolf (Spring, Berlin, 1986).
[4] G. Dattoli, A. Renieri, A. Torre, and J. C. Gallardo, Phys. Rev. A **35**, 4175 (1987); F. Ciocci, G. Dattoli, A. Renieri, and A. Torre, Phys. Rep. **141**, 1 (1986).
[5] B. Gazdy and D. A. Micha, J. Chem. Phys. **82**, 4926 (1985).
[6] I. Benjamin, J. Chem. Phys. **85**, 5611 (1986).
[7] W. Magnus, Commun. Pure. Appl. Math. **7**, 649 (1954).
[8] J. Wei and E. Norman, Proc. Am. Math. Soc. **15**, 327 (1963).
[9] J. Wei and E. Norman, J. Math. Phys. **4**, 575 (1963).
[10] M. Kolsrud, Phys. Rev. **104**, 1186 (1954).
[11] F. M. Fernández and E. A. Castro, Phys. Lett. A **125**, 77 (1987).
[12] F. M. Fernández, J. Phys. A **21**, 1357 (1988).

Exact solutions of the Schrödinger equation for nonseparable anharmonic oscillator potentials in two dimensions

D. R. Taylor and P. G. L. Leach
Centre for Nonlinear Studies and Department of Computational and Applied Mathematics, The University of the Witwatersrand, Johannesburg, Wits Post Office, 2050, Republic of South Africa

(Received 9 September 1988; accepted for publication 9 March 1989)

An explicit procedure is given to construct exact closed-form solutions to the time-independent Schrödinger equation in two dimensions $[\nabla^2 + \lambda - w(x,y)]\psi = 0$, where $w(x,y)$ is a polynomial potential of degree greater than two not separable in Cartesian coordinates. Several examples are discussed for which $w(x,y)$ is a sextic polynomial. As has already been seen in studies of the corresponding one-dimensional problem, a complete set of eigenvalues and wave functions is not found. However, these closed-form solutions can be used to check the accuracy and efficiency of numerical algorithms.

I. INTRODUCTION

In quantum mechanics the solution of the time-independent Schrödinger equation

$$(-\tfrac{1}{2}\nabla^2 + V)\psi = E\psi \tag{1.1}$$

(in appropriate units, i.e., $\hbar$ and m taken as unity throughout) is of importance, amongst other problems, in the determination of molecular spectra. Generally, the equation cannot be solved for the energy E and wave function ψ in closed form and we must resort to numerical algorithms. However, from the earliest days of quantum mechanics to now, various models of physical problems that are exactly soluble have been constructed. To mention but a few, we have some simple models of molecules such as ammonia and hydrogen-bonded solids[1] and benzene.[2]

In more recent years there have been a number of studies of closed-form solutions in cases where the potential is some form of anharmonic oscillator. The potentials considered have either been one-dimensional or n-dimensional with S_{n-1} symmetry. The earlier papers[3] tended to be *ad hoc*. More recently, a systematic procedure has been developed for the construction of these wave functions.[4] However, all of the models considered have essentially been for one dimension with $x\in(-\infty,\infty)$ or $r\in(0,\infty)$.

In this paper we explore the systematic construction of closed-form solutions to (1.1) for two-dimensional models with anharmonic potentials that are polynomial in form but without any assumption of symmetry. Although we have in mind the quantum mechanical origin of the problem, it can be restated as the construction of eigensolutions to the two-dimensional eigenvalue problem:

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \lambda - w(x,y)\right]\psi(x,y) = 0 \tag{1.2}$$

on domain $\Re^2$ with the boundary condition $\lim x\to\infty$, $\lim y\to\infty$ $\psi(x,y) = 0$, with the limits being considered independently.

Trivially, the Schrödinger equation (1.1) is always satisfied by a real-valued C^2 function $\bar{\psi}$ for $E = 0$ and $V = \frac{1}{2}(\nabla^2\bar{\psi})/\bar{\psi}$ provided V and $\bar{\psi}$ make physical sense. However, if $\bar{\psi}$ is the only closed-form solution that this potential provides, the result is of little or no practical use. What is useful is a potential for which more than one, preferably many more than one, explicit solution can be obtained. The value of such solutions is that, if the potential can be treated by perturbation methods, the exact solutions provide an excellent check of the perturbation expansion for those states for which they exist. Alternatively, if some numerical algorithm is used that is not a perturbation method, the closed-form solutions again provide a check of the validity and efficiency of the algorithm.

II. PRELIMINARIES

The two-dimensional Schrödinger equation (1.2) can be solved by separation of variables if there exists a coordinate transformation $(x,y)\to(\xi,\eta)$ such that in the new variables the Laplacian and the potential are separable. We note some trivial examples. If $w(x,y) = u(x) + v(y)$, (1.2) is separable in the original Cartesian coordinates. For

$$w(x,y) = u(ax + by) + v(cx + dy)$$

subject to the conditions $ad - bc\neq 0$ and $ac + bd = 0$, under the change of variables

$$\xi = ax + by, \quad \eta = cx + dy,$$

(1.2) becomes

$$\left[(a^2 + b^2)\frac{\partial^2}{\partial\xi^2} + (c^2 + d^2)\frac{\partial^2}{\partial\eta^2} + \lambda - u(\xi) - v(\eta)\right]\psi = 0,$$

which is also separable. If

$$w(x,y) = u(x^2 + y^2) + v(y/x)/(x^2 + y^2),$$

the transformation to plane polar coordinates $r^2 = x^2 + y^2$, $\tan\theta = y/x$ yields the separable equation

$$\left[\frac{1}{r}\frac{\partial}{\partial r}r\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2} + \lambda - \tilde{u}(r) - \frac{1}{r^2}\tilde{v}(\theta)\right]\psi = 0.$$

Finally, if $w(x,y)$ is a homogeneous quadratic form in x and y, there always exists an orthogonal tranformation that diagonalizes $w(x,y)$ and leaves the Laplacian form invariant.

In this paper we are interested in $w(x,y)$, a polynomial of degree greater than two for which no assumptions of sep-

 0022-2488/89/071525-08$02.50

arability under some transformation of coordinates are made. Suppose that

$$\psi = g(x,y)\exp[f(x,y)]. \tag{2.1}$$

Then

$$w-\lambda = \partial_1^2 f + \partial_2^2 f + (\partial_1 f)^2 + (\partial_2 f)^2 + (1/g) \times(\partial_1^2 g + \partial_2^2 g + 2\,\partial_1 g\,\partial_1 f + 2\,\partial_2 g\,\partial_2 f). \tag{2.2}$$

Anharmonic polynomial potentials are obtained using polynomial f and g provided

$$\partial_1^2 g + \partial_2^2 g + 2\,\partial_1 g\,\partial_1 f + 2\,\partial_2 g\,\partial_2 f = gh, \tag{2.3}$$

where h is a polynomial in x and y. For the wave function to be square integrable over $\Re^2$, the leading powers in f must be even with negative coefficients. Let

$$f(x,y) = -\sum_{i=1}^{2m}\sum_{j=1}^{2n} f_{ij}x^i y^j, \quad f_{2m,2n}>0,$$

$$g(x,y) = \sum_{i=0}^{k}\sum_{j=0}^{l} g_{ij}x^i y^j.$$

Then, from (2.3), $h(x,y)$ is of degree at most $2m-2$ in x and $2n-2$ in y, i.e.,

$$h(x,y) = \sum_{i=0}^{2m-2}\sum_{j=0}^{2n-2} h_{ij}x^i y^j.$$

In particular, if $m=n$ and $k=l$, $h(x,y)$ has the form

$$h(x,y) = \sum_{i=0}^{2n-2}\sum_{j=0}^{2n-2} h_{ij}x^i y^j$$

and $g(x,y)$ has the form

$$g(x,y) = \sum_{i=0}^{k}\sum_{j=0}^{k} g_{ij}x^i y^j.$$

Using (2.3) and the expansions above for $f(x,y)$, $g(x,y)$, and $h(x,y)$, the relationships between f_{ij}, g_{ij}, and h_{ij} may be determined.

In practice, we are given the potentials $w(x,y)$ and we seek $f(x,y)$ and $g(x,y)$. An example of this approach (the only one we know of for the two-dimensional problem) is found in the work of Makarewicz.[5] However, here we shall adopt a constructive approach. We choose an $f(x,y)$ and find relations between g_{ij} and h_{ij}, which will give eigenfunctions and eigenvalues for the potential defined by

$$w(x,y)-\lambda = \partial_1^2 f(x,y) + \partial_2^2 f(x,y) + (\partial_1 f(x,y))^2 + (\partial_2 f(x,y))^2 + h(x,y). \tag{2.4}$$

This concludes the basic theory. We now look at some specific examples to see what sorts of results can be obtained.

III. CONSTRUCTION OF SPECIFIC EXAMPLES

In selecting a suitable $f(x,y)$ for the purpose of these examples we take the simplest homogeneous polynomial that is nontrivial, viz.,

$$f(x,y) = -(x^4 + Bx^2y^2 + y^4), \quad B > -2. \tag{3.1}$$

The form of f is representative of all polynomials of the form $-A^4(x^4 + Bx^2y^2 + y^4)$ since the factor A^4 may be scaled away by a similarity transformation. The restriction $B > -2$ ensures that f is negative $\forall x,y \neq 0$ and its exponential tends to zero as $x,y \to \infty$ as is required for bound states.

Note that f, a homogeneous quadratic in x and y, always leads to a potential that is quadratic and the Schrödinger equation is then separable to two one-dimensional oscillators by means of an orthogonal rotation of axes. We also assume g to be homogeneous in its highest power, which we take to be of degree k. Then, Eq. (2.3)

$$\partial_1^2 g + \partial_2^2 g + 2\,\partial_1 g\,\partial_1 f + 2\,\partial_2 g\,\partial_2 f = gh$$

is of degree

$$(k-2)\ \ (k-2)\ \ (k+2)\ \ (k+2)\ \ (k+l),$$

so that l is at most 2. If, furthermore, we restrict our attention to wave functions that are either even or odd in x and y, h takes the form

$$h(x,y) = \alpha + \beta x^2 + \gamma xy + \delta y^2, \tag{3.2}$$

where γ is zero for wave functions even in both x and y. We now proceed to construct the potential $w(x,y)$ and associated eigenvalues λ for various values of k. We shall see that an increase in the degree of g increases the constraint placed on the admissible value of B in f. We discuss a number of typical results below. In all cases, $f(x,y)$ is given by (3.1) and the form of $h(x,y)$ by (3.2).

A. Even $g(x,y)$

Suppose that g has the form

$$g(x,y) = a + bx^2 + dy^2.$$

Then, after collecting the coefficients of like powers, (2.3) becomes

$$(2b+2d-a\alpha) + (-a\beta - b\alpha)x^2 + (-a\gamma)xy + (-a\delta - d\alpha)y^2 + (-16-\beta)bx^4 + (-b\gamma)x^3y + (-8bB - 8dB - b\delta - d\beta) \times x^2y^2 + (-d\gamma)xy^3 + (-16-\delta)dy^4 = 0.$$

To maintain the assumed quadratic nature of g it is evident from the coefficients of x^3y and xy^3 that $\gamma = 0$. Equating each remaining coefficient to zero gives the overdetermined system of equations

$$\begin{bmatrix} \alpha & -2 & -2 \\ \beta & \alpha & 0 \\ \delta & 0 & \alpha \\ 0 & 16+\beta & 0 \\ 0 & 0 & 16+\delta \\ 0 & 8B+\delta & 8B+\beta \end{bmatrix} \begin{bmatrix} a \\ b \\ d \end{bmatrix} = 0. \tag{3.3}$$

From rows 4 and 5 of Eq. (3.3), it is evident that $\beta = \delta = -16$. From row 5 of Eq. (3.3) either $B=2$ or $d=-b$ or both. Treating rows 1–3 of Eq. (3.3) as an eigenvalue problem, we find the following eigenvalues and eigenvectors

$$\alpha = 0, \qquad \alpha = 8, \qquad \alpha = -8,$$

$$\mathbf{u}_0 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} k_1, \quad \mathbf{u}_8 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} k_2, \quad \mathbf{u}_{-8} = \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix} k_3,$$

where k_1, k_2, and k_3 are parameters the values of which are determined from the normalization requirement $\|\psi\|^2 = 1$. We note that only the first eigenvector is consistent with B

arbitrary and $d = -b$. So we obtain only the single eigenfunction corresponding to $\lambda = 0$ for the potential

$$w(x,y) = 16x^6 + 4B(B+4)x^2y^2(x^2+y^2) + 16y^4 - (28+2B)(x^2+y^2).$$

It is

$$\psi(x,y) = k_1(x^2-y^2)\exp[-(x^4+Bx^2y^2+y^4)].$$

For the case $B = 2$, the potential is

$$w(x,y) = 16(x^2+y^2)^3 - 32(x^2+y^2)$$

and the eigenvalues and eigenfunctions are

$\lambda = -8$, $\psi_{-8} = k_2(1+2(x^2+y^2))\exp[-(x^2+y^2)^2]$,

$\lambda = 0$, $\psi_0 = k_1(x^2-y^2)\exp[-(x^2+y^2)^2]$,

$\lambda = 8$, $\psi_8 = k_3(1-2(x^2+y^2))\exp[-(x^2+y^2)^2]$.

The latter result corresponds to a potential that is separable in plane polar coordinates r,θ. The terms ψ_8 and ψ_{-8} correspond to zero angular quantum number, m and ψ_0 to $m = \pm 2$.

B. Odd $g(x,y)$

1. Cubic

With f and h as for AI, let

$$g = ax + by + cx^3 + dx^2y + exy^2 + fy^3.$$

Substituting f, g, and h into (2.3), collecting coefficients of like powers and equating the coefficients of independent powers to zero, we have the following overdetermined system of equations:

$$\begin{bmatrix}
\alpha & & -6 & & -2 & \\
 & \alpha & & -2 & & -6 \\
\beta+8 & & \alpha & & & \\
\gamma & \beta+4B & & \alpha & & \\
\delta+4B & \gamma & & & \alpha & \\
 & \delta+8 & & & & \alpha \\
 & & \beta+24 & & & \\
 & & \gamma & \beta+16+4B & & \\
 & & \delta+12B & \gamma & \beta+8+8B & \\
 & & & \delta+8+8B & \gamma & \beta+12B \\
 & & & & \delta+16+4B & \gamma \\
 & & & & & \delta+24
\end{bmatrix} \mathbf{g} = 0, \tag{3.4}$$

where $\mathbf{g}^T = (a,b,c,d,e,f)$.

From rows 7 and 12 of Eq. (3.4) either $c = f \equiv 0$ or $\delta = \beta = -24$. If c and f were identically zero, rows 8 and 11 of Eq. (3.4) would require $d = e \equiv 0$ or $\beta = \delta = -16-4B$. Since we wish to consider a g, which is a general cubic, we assume $\delta = \beta = -24$. This reduces (3.4) to

$$\begin{bmatrix}
\alpha & & -6 & & -2 & \\
 & \alpha & & -2 & & -6 \\
-16 & & \alpha & & & \\
\gamma & 4B-24 & & \alpha & & \\
4B-24 & \gamma & & & \alpha & \\
 & -16 & & & & \alpha \\
 & & \gamma & 4B-8 & & \\
 & & 12B-24 & \gamma & 8B-16 & \\
 & & & 8B-16 & \gamma & 12B-24 \\
 & & & & 4B-8 & \gamma
\end{bmatrix} \mathbf{g} = 0. \tag{3.5}$$

This equation is of form

$$\begin{bmatrix} A & B \\ 0 & D \end{bmatrix}\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = 0,$$

where $\mathbf{u}^T = [a,b]$ and $\mathbf{v}^T = [c,d,e,f]$. We note that the mixed term coefficient γ in the even-g case was necessarily zero whereas the odd case produces a double eigenvalue problem involving the constant term and mixed term coefficient, α and γ, of $h(x,y)$.

We first solve $D\mathbf{v} = 0$. Setting $\gamma = K[4(B-2)]$ and factoring out $4(B-2)$, this becomes

$$\begin{bmatrix} K & 1 & & \\ 3 & K & 2 & \\ & 2 & K & 3 \\ & & 1 & K \end{bmatrix}\mathbf{v} = 0, \tag{3.6}$$

which is the eigenvalue problem for a centrosymmetric matrix. The characteristic equation for (3.6) is

$$C_P(K) = K^4 - 10K + 9 = (K^2 - 1)(K^2 - 9) = 0$$

and the eigenvalues and associated eigenvectors are

$$\gamma = 4(B-2), \qquad \gamma = -4(B-2)$$

$$\mathbf{u}_{+1} = k_1 \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}, \quad \mathbf{u}_{-1} = k_2 \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix},$$

$$\gamma = 12(B-2), \qquad \gamma = -12(B-2),$$

$$\mathbf{u}_{+3} = k_3 \begin{bmatrix} 1 \\ -3 \\ 3 \\ -1 \end{bmatrix} \quad \mathbf{u}_{-3} = k_4 \begin{bmatrix} 1 \\ 3 \\ 3 \\ 1 \end{bmatrix},$$

where k_i, $i = 1,4$ are arbitrary constants.

We now turn to the first of the set of equations, viz., $A\mathbf{u} + B\mathbf{v} = \mathbf{0}$, to seek consistent solutions for each eigenvalue γ. The equation is

$$\begin{bmatrix} \alpha & & -6 & & -2 & \\ & \alpha & & -2 & & -6 \\ -16 & & \alpha & & & \\ \gamma & 4(B-6) & & \alpha & & \\ 4(B-6) & \gamma & & & \alpha & \\ & -16 & & & & \alpha \end{bmatrix} \mathbf{g} = \mathbf{0}. \tag{3.7}$$

Before substituting for the particular values of γ we perform the following reductions. Rows 3 and 4 of Eq. (3.7) involve nonzero coefficients of a and b, respectively, resulting in

$$a = \alpha c/16, \quad b = \alpha f/16. \tag{3.8}$$

Rows 4 and 5 of Eq. (3.7) become

$$\begin{bmatrix} \gamma & 4(B-6) \\ 4(B-6) & \gamma \end{bmatrix} \begin{bmatrix} c\alpha/16 \\ f\alpha/16 \end{bmatrix} + \alpha \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} d \\ e \end{bmatrix} = \mathbf{0}. \tag{3.9}$$

On substitution for a and b from (3.8), rows 1 and 2 of Eq. (3.7) become

$$\begin{bmatrix} d \\ e \end{bmatrix} = \frac{1}{16} \begin{bmatrix} 0 & (\tfrac{1}{2}\alpha^2 - 48) \\ (\tfrac{1}{2}\alpha^2 - 48) & 0 \end{bmatrix} \begin{bmatrix} c \\ f \end{bmatrix}. \tag{3.10}$$

Substitution of (3.10) into (3.9) yields

$$\frac{\alpha}{16} \begin{bmatrix} \gamma & 4(B-6) \\ 4(B-6) & \gamma \end{bmatrix} \begin{bmatrix} c \\ f \end{bmatrix} + \frac{\alpha}{16} \begin{bmatrix} 0 & (\tfrac{1}{2}\alpha^2 - 4\gamma) \\ (\tfrac{1}{2}\alpha^2 - 48) & 0 \end{bmatrix} \begin{bmatrix} c \\ f \end{bmatrix} = \mathbf{0},$$

from which it follows that $\alpha = 0$ or

$$\begin{bmatrix} \gamma & (\tfrac{1}{2}\alpha^2 - 48 + 4(B-6)) \\ (\tfrac{1}{2}\alpha^2 - 48 + 4(B-6)) & \gamma \end{bmatrix} \times \begin{bmatrix} c \\ f \end{bmatrix} = \mathbf{0}. \tag{3.11}$$

From $D\mathbf{v} = \mathbf{0}$, c and f are uniquely determined with respect to a specific γ and substitution into the matrix equation (3.11) gives two equations in α. It is evident from the symmetry in c and f that the two equations are identical.

We find the following cases:

(a) $\gamma = 4(B-2)$, $c = k_1$, $f = k_1$;
$\alpha = \pm 4\sqrt{10 - B}$;
$a_+ = \pm k_1(\sqrt{10 - B}/4)$, $b_+ = \pm k_1(\sqrt{10 - B}/4)$;

(b) $\gamma = -4(B-2)$, $c = k_2$, $f = -k_2$;
$\alpha = \pm 4\sqrt{10 - B}$;
$a_- = \pm k_2(\sqrt{10 - B}/4)$, $b_- = \mp k_2(\sqrt{10 - B}/4)$;

(c) $\gamma = 12(B-2)$, $c = k_3$, $f = -k_3$,
$\alpha = \pm 4\sqrt{B + 6}$,
$a_+ = \pm k_3(\sqrt{B + 6}/4)$, $b_+ = \mp k_3(\sqrt{B + 6}/4)$;

(d) $\gamma = -12(B-2)$, $c = k_4$, $f = k_4$,
$\alpha = \pm 4\sqrt{B + 6}$,
$a_- = \pm k_4(\sqrt{B + 6}/4)$, $b_- = \pm k_4(\sqrt{B + 6}/4)$.

There are four possible potentials and for each we have three eigenvalues and eigenfunctions. Putting

$$W(x,y) = 16x^6 + 4B(B+4)x^2y^2(x^2 + y^2) + 16y^6 - (28 + 2B)(x^2 + y^2),$$

for

$$w_1(x,y) = W(x,y) - 24x^2 + 4(B-2)xy - 24y^2$$

we have

$$\lambda = -4\sqrt{10 - B}, \quad \psi(x,y) = k_{11}(x+y)\{\sqrt{10 - B}/4 + (x-y)^2\}\exp[f(x,y)],$$
$$\lambda = 0, \quad \psi(x,y) = k_{12}(x+y)(x-y)^2 \exp[f(x,y)],$$
$$\lambda = 4\sqrt{10 - B}, \quad \psi(x,y) = k_{13}(x+y)\{-\sqrt{10 - B}/4 + (x-y)^2\}\exp[f(x,y)],$$

for

$$w_2(x,y) = W(x,y) - 24x^2 - 4(B-2)xy - 24y^2,$$

we have

$$\lambda = -4\sqrt{10 - B}, \quad \psi(x,y) = k_{21}(x-y)\{\sqrt{10 - B}/4 + (x+y)^2\}\exp[f(x,y)],$$
$$\lambda = 0, \quad \psi(x,y) = k_{22}(x-y)(x+y)^2 \exp[f(x,y)],$$
$$\lambda = 4\sqrt{10 - B}, \quad \psi(x,y) = k_{23}(x-y)\{-\sqrt{10 - B}/4 + (x+y)^2\}\exp[f(x,y)],$$

for

$$w_3(x,y) = W(x,y) - 24x^2 + 12(B-2)xy - 24y^2,$$

we have

$$\lambda = -4\sqrt{B+6}, \quad \psi(x,y) = k_{31}(x-y)\{\sqrt{B+6}/4 + (x-y)^2\}\exp[f(x,y)],$$

$$\lambda = 0, \quad \psi(x,y) = k_{32}(x-y)^3\exp[f(x,y)],$$

$$\lambda = 4\sqrt{B+6}, \quad \psi(x,y) = k_{33}(x-y)\{-\sqrt{B+6}/4 + (x-y)^2\}\exp[f(x,y)],$$

and for

$$w_4(x,y) = W(x,y) - 24x^2 - 12(B-2)xy - 24y^2,$$

we have

$$\lambda = -4\sqrt{B+6}, \quad \psi(x,y) = k_{41}(x+y)\{\sqrt{B+6}/4 + (x+y)^2\}\exp[f(x,y)],$$

$$\lambda = 0, \quad \psi(x,y) = k_{42}(x+y)^3\exp[f(x,y)],$$

$$\lambda = 4\sqrt{B+6}, \quad \psi(x,y) = k_{43}(x+y)\{-\sqrt{B+6}/4 + (x+y)^2\}\exp[f(x,y)].$$

Note that w_1 and w_2, w_3 and w_4 are essentially the same since one is obtained from the other by a rotation of axes through an angle $\pi/2$. Furthermore, for the special case $B = 2$, we obtain a degeneracy in the eigenvalues and the two wave functions.

2. Quintic

With f and h as for Secs. III A 1 and III B 1 let

$$g = ax + by + cx^3 + dx^2y + exy^2 + fy^3 + ix^5 + jx^4y + kx^3y^2 + lx^2y^3 + mxy^4 + ny^5.$$

Substituting f, g, and h into (2.3) and proceeding as for Sec. III B 1 we obtain the overdetermined system of equations

$$\begin{bmatrix}
\alpha & & -6 & & -2 & \\
 & \alpha & & -2 & & -6 \\
(\beta+8) & & \alpha & & & \\
\alpha & (\beta+4B) & & \alpha & & \\
(\delta+4B) & \gamma & & & \alpha & \\
 & (\delta+8) & & & & \alpha \\
 & & (\beta+24) & & & \\
 & & \gamma & (\beta+4B+16) & & \\
 & & (\delta+12B) & \gamma & (\beta+\delta B+8) & \\
 & & & (\delta+8B+8) & \gamma & (\beta+12B) \\
 & & & & (\delta+4B+16) & \gamma \\
 & & & c & & (\delta+24)
\end{bmatrix}\mathbf{u}$$

$$+\begin{bmatrix}
-20 & -2 & & & \\
 & -12 & -6 & & \\
 & -6 & & -12 & \\
 & & -2 & & -20 \\
\alpha & & & & \\
 & \alpha & & & \\
 & & \alpha & & \\
 & & & \alpha & \\
 & & & & \alpha
\end{bmatrix}\mathbf{v} = \mathbf{0},$$

$$\begin{bmatrix}
(\beta+40) & & & & & \\
\gamma & (\beta+4B+32) & & & & \\
(\delta+20B) & \gamma & (\delta B+24) & & & \\
 & (\delta+16B+8) & \gamma & (\beta+12B+16) & & \\
 & & (\delta+12B+16) & \gamma & (\beta+16B+8) & \\
 & & & (\delta+8B+24) & \gamma & (\beta+20B) \\
 & & & & (\delta+4B+32) & \gamma \\
 & & & & & (\delta+40)
\end{bmatrix}\mathbf{v} = \mathbf{0}, \qquad (3.12)$$

where $\mathbf{u}^T = [a,b,c,d,e,f]$ and $\mathbf{v}^T = [i,j,k,l,m,n]$.

By analysis of rows 13 and 20 of Eq. (3.12), either $i = n \equiv 0$ or $\delta = \beta - 40$. The former choice will imply that $j = m \equiv 0$ or $\delta = \beta = -4(B+8)$. Putting $j = m = 0$ would then require $k = l \equiv 0$ or $\delta = \beta = -8(B+3)$. Since we wish to consider a g that is a general quintic, we take $\delta = \beta = -40$.

$$\begin{bmatrix}
\alpha & & -6 & & -2 & \\
& \alpha & & -2 & & -6 \\
-32 & & \alpha & & & \\
\alpha & 4B-40 & & \alpha & & \\
4B-40 & \gamma & & & \alpha & \\
& -32 & & & & \alpha \\
& & -16 & & & \\
& & \gamma & 4B-24 & & \\
& & 12B-40 & \gamma & 8B-32 & \\
& & & 8B-32 & \gamma & 12B-40 \\
& & & & 4B-24 & \gamma \\
& & & & & -16
\end{bmatrix} \mathbf{u} = \mathbf{0},$$

$$\begin{bmatrix}
-20 & & -2 & & & \\
& -12 & & -6 & & \\
& & -6 & & -12 & \\
& & & -2 & & -20 \\
\alpha & & & & & \\
& \alpha & & & & \\
& & \alpha & & & \\
& & & \alpha & & \\
& & & & \alpha & \\
& & & & & \alpha
\end{bmatrix} \mathbf{v} = \mathbf{0}, \tag{3.13}$$

$$\begin{bmatrix}
\gamma & 4B-8 & & & & \\
20B-40 & \gamma & 8B-16 & & & \\
& 16B-32 & \gamma & 12B-24 & & \\
& & 12B-24 & \gamma & 16B-32 & \\
& & & 8B-16 & \gamma & 20B-40 \\
& & & & 4B-8 & \gamma
\end{bmatrix} \mathbf{v} = \mathbf{0}.$$

As in Sec. III B 1 this equation is of the form

$$\begin{bmatrix} A & B \\ 0 & D \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \mathbf{0}.$$

The eigenvalues and associated eigenvectors for $D\mathbf{v} = \mathbf{0}$ are

$$\gamma = 4(B-2), \qquad \gamma = -4(B-2),$$

$$\mathbf{u}_{+1} = k_1 \begin{bmatrix} 1 \\ -1 \\ -2 \\ 2 \\ 1 \\ -1 \end{bmatrix}, \qquad \mathbf{u}_{-1} = k_2 \begin{bmatrix} 1 \\ 1 \\ -2 \\ -2 \\ 1 \\ 1 \end{bmatrix},$$

$$\gamma = 12(B-2), \qquad \gamma = -12(B-2)$$

$$\mathbf{u}_{+3} = k_3 \begin{bmatrix} 1 \\ -3 \\ 2 \\ 2 \\ -3 \\ 1 \end{bmatrix}, \qquad \mathbf{u}_{-3} = k_4 \begin{bmatrix} 1 \\ 3 \\ 2 \\ -2 \\ -3 \\ -1 \end{bmatrix},$$

$$\gamma = 20(B-2), \qquad \gamma = -20(B-2),$$

$$\mathbf{u}_{+5} = k_5 \begin{bmatrix} 1 \\ -5 \\ 10 \\ -10 \\ 5 \\ -1 \end{bmatrix}, \quad \mathbf{u}_{-5} = k_6 \begin{bmatrix} 1 \\ 5 \\ 10 \\ 10 \\ 5 \\ 1 \end{bmatrix},$$

where k_i, $i = 1,6$ are arbitrary constants.

We now return to the top set of equations, viz., $A\mathbf{u} + B\mathbf{v} = \mathbf{0}$ to seek consistent solutions pertaining to each eigenvalue γ. Manipulation of the equation leads to the set of consistency equations for each γ_i, $i = 1,6$,

$$\begin{aligned}
&\gamma = \pm 4(B-2);\\
&\quad \alpha^4 - 512\alpha^2 + 16\,384 = 0,\\
&\quad \alpha(B-6)[\alpha^2/256 - 2] = 0,\\
&\quad \alpha(B-6)[\alpha^2/256 - 1] = 0,
\end{aligned} \tag{3.14}$$

$$\begin{aligned}
&\gamma = \pm -12(B-2);\\
&\quad \alpha^4/1024 + (B-22)\alpha^2/32 - 12(B-6) = 0,\\
&\quad \alpha(B-6)[\alpha^2/64 - 6] = 0,\\
&\quad \alpha[\alpha^2 - 128(12B-62)/(58-14)] = 0,
\end{aligned} \tag{3.15}$$

$$\begin{aligned}
&\gamma = \pm -20(B-2);\\
&\quad \alpha^4/1024 - (B+26)\alpha^2/32 + 20(B-6) = 0,\\
&\quad \alpha(B-6)[\alpha^2/256 - 23/4] = 0,\\
&\quad \alpha[\alpha^2 - 218(21B+2)/(3B-2)] = 0.
\end{aligned} \tag{3.16}$$

By examination, a consistent solution depends on the value of B. If we place $B = 6$ we obtain the following consistent, real solutions:

$$\alpha = \pm 2^4\sqrt{1+\sqrt{3}}, \tag{3.14'}$$

$$\alpha = 0, \tag{3.15'}$$

$$\alpha = 0. \tag{3.16'}$$

We find that there are six possible potentials (cf. Sec. III B 1) and for each we have only one eigenvalue and eigenfunction, except for $\gamma = \pm 4(B-2)$. Putting

$$\begin{aligned}
W(x,y) &= 16x^2 + 240x^2y^2(x^2+y^2)\\
&\quad + 16y^6 - 40(x^2+y^2),\\
f(x,y) &= -(x^4 + 6x^2y^2 + y^4),
\end{aligned}$$

for

$$w_1(x,y) = W(x,y) - 40x^2 + 16xy - 40y^2, \tag{3.17}$$

we have

$$\lambda = -2^4\sqrt{1+\sqrt{3}},$$

$$\begin{aligned}
\psi_1(x,y) &= k_{11}(x-y)\left[\sqrt{3}/2 + \sqrt{1+\sqrt{3}}\right.\\
&\quad \times (x^2 + 4(\sqrt{3}-1)xy + y^2) + (x-y)^2(x+y)^2]\\
&\quad \times \exp[f(x,y)],
\end{aligned} \tag{3.18}$$

$$\lambda = 2^4\sqrt{1+\sqrt{3}},$$

$$\begin{aligned}
\psi_2(x,y) &= k_{12}(x+y)\left[\sqrt{3}/2 - \sqrt{1+\sqrt{3}}\right.\\
&\quad \times (x^2 + 4(\sqrt{3}-1)xy + y^2) + (x-y)^2(x+y)^2]\\
&\quad \times \exp[f(x,y)].
\end{aligned} \tag{3.19}$$

For

$$w_2(x,y) = W(x,y) - 40x^2 - 16xy - 40y^2, \tag{3.20}$$

we have

$$\lambda = -2^4\sqrt{1+\sqrt{3}},$$

$$\begin{aligned}
\psi_3(x,y) &= k_{21}(x-y)\left[\sqrt{3}/2 + \sqrt{1+\sqrt{3}}\right.\\
&\quad \times (x^2 - 4(\sqrt{3}-1)xy + y^2) + (x-y)^2(x+y)^2]\\
&\quad \times \exp[f(x,y)],
\end{aligned} \tag{3.21}$$

$$\lambda = 2^4\sqrt{1+\sqrt{3}},$$

$$\begin{aligned}
\psi_4(x,y) &= k_{22}(x+y)\left[\sqrt{3}/2 - \sqrt{1+\sqrt{3}}\right.\\
&\quad \times (x^2 - 4(\sqrt{3}-1)xy + y^2) + (x-y)^2(x+y)^2]\\
&\quad \times \exp[f(x,y)].
\end{aligned} \tag{3.22}$$

For

$$w_3(x,y) = W(x,y) - 40x^2 + 48xy - 40y^2, \tag{3.23}$$

we have

$$\begin{aligned}
&\lambda = 0,\\
&\psi_5(x,y) = k_3(x+y)[-\tfrac{3}{4} + (x-y)^4]\exp[f(x,y)].
\end{aligned} \tag{3.24}$$

For

$$w_4(x,y) = W(x,y) - 40x^2 - 48xy - 40y^2, \tag{3.25}$$

we have

$$\begin{aligned}
&\lambda = 0,\\
&\psi_6(x,y) = k_4(x-y)[-\tfrac{3}{4} + (x+y)^4]\exp[f(x,y)].
\end{aligned} \tag{3.26}$$

For

$$w_5(x,y) = W(x,y) - 40x^2 + 80xy - 40y^2, \tag{3.27}$$

we have

$$\begin{aligned}
&\lambda = 0,\\
&\psi_7(x,y) = k_5(x-y)[-\tfrac{5}{4} + (x-y)^4]\exp[f(x,y)].
\end{aligned} \tag{3.28}$$

For

$$w_6(x,y) = W(x,y) - 40x^2 - 80xy - 40y^2, \tag{3.29}$$

we have

$$\begin{aligned}
&\lambda = 0,\\
&\psi_8(x,y) = k_6(x+y)[-\tfrac{5}{4} + (x+y)^4]\exp[f(x,y)].
\end{aligned} \tag{3.30}$$

We observe that (3.20)–(3.22) are just (3.17)–(3.19) rotated by $\pi/2$. For each, we obtain two eigenfunctions. The eigenstate ψ_1 (3.18) is zero only along the line $y = x$. The eigenstate ψ_2 (3.19) is zero also along the curve described by

$$\sqrt{3}/2 - \sqrt{1+\sqrt{3}}(x^2 + 4(\sqrt{3}-1)xy + y^2) + (x+y)^2(x-y)^2 = 0.$$

For (3.20) the results are the same with $x \to y$ and $y \to -x$ (rotation by $\pi/2$).

The potentials (3.25) and (3.29) and the associated eigenfunctions are just the potentials (3.23) and (3.27) and their associated eigenfunctions rotated by $\pi/2$. The eigenfunction (3.25) is zero along $y = -x$ and $y = x \pm (\sqrt{5}/2)^{1/2}$.

IV. CONCLUSION

In this paper we have described a general procedure for the construction of closed-form wave functions for a class of two-dimensional, nonseparable, polynomial anharmonic oscillators. This has extended the already known results for one-dimensional and n-dimensional polynomial anharmonic oscillators with S_{n-1} symmetry. We then considered a number of examples in which the potential was sextic to demonstrate how the constructive procedure works in practice. Writing the wave function as

$$\psi(x,y) = g(x,y)\exp[f(x,y)],$$

we found that, for $g(x,y)$ of a low degree, multiple eigenstates were obtained whereas for $g(x,y)$ of higher degree (see Sec. III B 2) generically only a single eigenstate was obtained for each permissible potential. As such, these results are not of much use in themselves as complete sets of eigenvalues are required. However, they do provide a very effective check of the efficiency of numerical algorithms (say finite differences or perturbation expansions), which could be used to provide the complete set.

We note that in the examples of $g(x,y)$ even in x and y and that of Sec. III A 1 we found $h(x,y)$ to be even. When $g(x,y)$ was odd in x and y (Secs. III B 1 and III B 2), the parameter B in $f(x,y)$ was required to satisfy constraints in addition to the requirement that $f(x,y)$ be negative definite (i.e., $B > -2$). For the case of Sec. III B 1, the constraint was $-6 \leqslant B \leqslant 10$ so that physically acceptable values of B were confined to the interval $(-2, 10]$. For the case of Sec. III B 2, in which the degree of $g(x,y)$ was higher than in Sec. III B 1, the constraint was more specific, viz., $B = 6$. It may well be that $g(x,y)$ of even higher degree will not permit a physically acceptable value of B or even any value of B at all for which a consistent solution to the matrix equation corresponding to (3.13) exists.

It is evident that for the two odd cases considered (Secs. III B 1 and III B 2) and the form of $f(x,y)$ adopted, the matrix equation to be satisfied has the form

$$\begin{bmatrix} A & B \\ 0 & D \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \mathbf{0},$$

where $\mathbf{g}^T = (\mathbf{u}^T, \mathbf{v}^T)$. It is apparent that the matrix D is centrosymmetric with eigenvalues $\gamma_i = 4K_i(B-2)$, $i = 1, n(n$ even), where $K_i = \pm 1, \pm 3, \ldots, \pm m$ (m odd). This obviates the necessity to determine the entire matrix of coefficients. We merely need the truncated matrix of coefficients of polynomial terms up to k, assuming $g(x,y)$ an odd polynomial of degree k to give the coefficients in $A\mathbf{u} + B\mathbf{v} = \mathbf{0}$. The various γ_i's are already known. It could be of interest to continue further with $g(x,y)$ (odd) of higher degree to ascertain whether closed-form consistent solutions do exist.

[1]See, for example, D. M. Dennison and G. E. Uhenbeck, Phys. Rev. **41**, 313 (1932); F. T. Wall and G. Glockler, J. Chem. Phys. **5**, 314 (1937); R. L. Somorjai and D. F. Hornig, *ibid.* **36**, 1980 (1962).

[2]See, for example, H. Hartmann, Theoret. Chim. Acta (Berlin) **24**, 201 (1972); C. C. Gerry, Phys. Lett. A **118**, 445 (1986); H. Hornburger, K-D. Reinsch, and M. Melzig, Z. Phys. D **5**, 151 (1987).

[3]G. P. Flessas, Phys. Lett. A **27**, 289 (1979); A **81**, 17 (1981); J. Phys. A **14**, L209 (1981); Phys. Lett. A **78**, 19 (1980).

[4]P. G. L. Leach, Physica D **17**, 331 (1985); M. Blecher and P. G. L. Leach, Proceedings of the 13th SANUM conference, Umhlanga Rocks, South Africa (University of Natal, Durban, 1987), pp. 5–17; P. G. L. Leach, J. Math. Phys. **25**, 2974 (1984).

[5]J. Makarewicz, J. Phys. A **16**, L553 (1983).

N-body quantum scattering theory in two Hilbert spaces. V. Computation strategy

Colston Chandler
Department of Physics and Astronomy, University of New Mexico, Albuquerque, New Mexico 87131
A. G. Gibson
Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131

(Received 2 March 1988; accepted for publication 15 March 1989)

In this paper the development of our previously published theory of approximations for the Chandler–Gibson (CG) equations is continued. In particular, our approximation theory is rigorously brought to the point where N-particle scattering calculations can begin. This is accomplished by mapping the CG operator equations into a function equation form, where the unknowns belong to a new (third!) computational Hilbert space $\mathscr{L}$. This mapping is facilitated by rescaling the Jacobi momentum variables for the relative free motion of the asymptotic clusters so that surfaces of constant kinetic energy are hyperspheres. The input terms to the resulting equations are expanded in a basis on the surface of the kinetic energy hypersphere. This leads to a system of infinitely many coupled one-dimensional integral equations with the kinetic energy as the continuous variable. The half-on-shell variant of these equations is then transformed to a K-matrix form. Our approximations result from truncating this system to a finite number of equations, which is equivalent to using a finite basis approximation of the original input terms. The basis sets could be hyperspherical harmonics, but the use of hyperspherical spline functions is also proposed. Our method generalizes the well-known method of partial waves for channels with two clusters, and it accommodates breakup channels in a straightforward way.

I. INTRODUCTION

In previous publications[1–5] we have used a two Hilbert space formulation of scattering theory to establish a new theoretical approach to solving nonrelativistic multichannel quantum scattering problems. In particular, we have derived a new system of N-particle integral equations, which, in contrast to the Faddeev–Yakubovskiĭ equations, has the same general form for any number $N \geqslant 2$ of particles. We have proved existence, uniqueness, and stability results for these equations, and we have developed an approximation scheme that preserves the unitarity of the approximate scattering operators and converges to the exact scattering operator. We have also studied the real energy limits of these equations in the appropriate operator topologies. The equations and approximations of this theory are cast within the framework of operators that act either on the asymptotic Hilbert space $\mathscr{H}$ or on the full N-particle Hilbert space $\mathscr{H}_N$.

In the present paper we develop a computational strategy for solving our equations, which, we believe, makes feasible the practical calculation of N-body scattering cross sections where the number of clusters (fragments) in the partitions of the N particles remains small. For example, we are initially interested in obtaining solutions for two cluster elastic and rearrangement problems and for three cluster breakup problems. In the present considerations we are ignoring particle spin and long-range forces between particles. We also do not discuss the effects of the exchange of identical particles, but these effects can be easily incorporated into the strategy of this paper by using the methods of Refs. 6 and 7.

The goal of this paper is to bring our theory rigorously to the point where N-particle scattering calculations can begin. This requires the mapping of our previous *operator equation* results to a more concrete *function equation* form by using certain injection operators ρ and ρ^*. The resulting equations are then cast into a half-on-shell $\mathscr{K}$*-matrix equation* form, which is most suitable for calculations.

The input terms to the $\mathscr{K}$-matrix equation are complicated multidimensional integrals, but we show how the evaluation of these integrals can be reduced to the expansion of certain "potential" and "overlap" functions in approximate basis sets on the surface of the partition *kinetic energy hyperspheres*. These basis sets could be *hyperspherical harmonics*,[8] but, in an effort to obtain superior convergence for breakup calculations, we are currently proposing to use hyperspherical spline functions. These functions fit data with a minimum of curvature between the data points. Two-dimensional spherical spline functions have been successfully used to fit geophysical and meteorological data on the surface of the earth.[9,10] Work is in progress on generalizing these spline functions to n-dimensional hyperspheres.

In Sec. II we recall the definitions of exact and approximate scattering systems, which were introduced in Refs. 1–4. We also state the assumptions that we are making about these systems.

We define partitions of N particles into disjoint clusters in Sec. III. We then define *kinetic energy hyperspherical coordinates*, which turn out to be the most useful coordinates for our method of approximation. We derive the Jacobian of the transformation from clustered Jacobi momentum co-

 0022-2488/89/071533-12$02.50

ordinates to kinetic energy hyperspherical coordinates.

In Sec. IV we define a *computation* space $\mathscr{L}$ to be the direct sum of copies of $\mathscr{L}^2[0,\infty)$. This *third* Hilbert space is suitable for making calculations. We also define the mapping ρ, which maps the asymptotic space $\mathscr{H}$ onto $\mathscr{L}$ and its adjoint ρ^*, which maps $\mathscr{L}$ onto $\mathscr{H}$. We then prove several properties about these mappings.

Section V is devoted to the construction of the asymptotic approximation space $\mathscr{H}^\pi$ and the projection operator Π, which projects $\mathscr{H}$ onto this space. We prove that these operators satisfy the hypotheses required in our general approximation theory.[3,4]

The derivation of our "transition matrix" $\mathscr{M}$-operator equations is contained in Sec. VI. We put this equation half on-shell and then derive the $\mathscr{K}$-matrix form of our integral equations.

In Sec. VII we discuss the evaluation of the input terms in our $\mathscr{K}$-matrix equation. We prove a theorem that establishes that these integrals can be calculated using expansions in approximate bases on the surfaces of kinetic energy hyperspheres.

We show in Sec. VIII how the N-particle *scattering amplitude* can be calculated from the solution of our $\mathscr{K}$-matrix equation.

In Sec. IX we summarize our computational method in five steps. We believe that this summary illustrates the feasibility of our computational strategy.

We plan to publish in future papers the results of test calculations using the computation method developed in this manuscript.

II. SCATTERING SYSTEMS

A. Exact scattering systems

In previous papers,[1-5] we have shown that the N-body *scattering operator* S: $\mathscr{H}\to\mathscr{H}$ is obtainable from the *transition operator* $T(z)$: $\mathscr{D}(H)\subset\mathscr{H}\to\mathscr{H}$,

$$T(z)\equiv(z-H)J^*R_N(z)V. \tag{2.1}$$

Here $H=\oplus_A H_A$ is the *asymptotic Hamiltonian* acting on the *asymptotic Hilbert space* $\mathscr{H}$, H_N is the *total Hamiltonian* acting on the *N-body Hilbert space* $\mathscr{H}_N$, $R_N(z)\equiv(z-H_N)^{-1}$, J: $\mathscr{H}\to\mathscr{H}_N$ is a bounded *injection operator* with adjoint J^*: $\mathscr{H}_N\to\mathscr{H}$ and $V\equiv H_NJ-JH$. Here $\mathscr{H}$ is assumed to have a decomposition $\mathscr{H}=\mathscr{H}^a\oplus\mathscr{H}^b$ that reduces H, and the orthogonal projection of $\mathscr{H}$ onto $\mathscr{H}^a$ is denoted by I^a. The sextuple

$$\mathfrak{S}\equiv\{\mathscr{H}_N,H_N,\mathscr{H},H,J,I^a\} \tag{2.2}$$

then characterizes the *exact scattering system*.

We assume that the exact scattering system $\mathfrak{S}$ satisfies Assumption A of Ref. 4. Also, since the scattering operator is zero on $\mathscr{H}^b$, it is only the operator $I^aT(z)I^a$ that is needed (Ref. 4, Theorem 2.1). Consequently, in order to simplify the presentation of the concepts in the following, we will assume that $\mathscr{H}=\mathscr{H}^a$ and $I^a=I$, the identity operator on $\mathscr{H}$.

Assumption A: In the notation of the present paper (with $\mathscr{H}=\mathscr{H}^a$), the exact scattering system $\mathfrak{S}=\{\mathscr{H}_N,H_N,\mathscr{H},H,J,I\}$ is said to satisfy Assumption A if the following five statements are true.

(A1) $\mathscr{H}_N$ is a separable Hilbert space, and H_N: $\mathscr{D}(H_N)\subset\mathscr{H}_N\to\mathscr{H}_N$ is a self-adjoint operator that is bounded from below.

(A2) $\mathscr{H}$ is a separable Hilbert space, and H: $\mathscr{D}(H)\subset\mathscr{H}\to\mathscr{H}$ is a self-adjoint operator that is bounded from below. Its spectral family is denoted by $E(\lambda)$. I is the identity operator on $\mathscr{H}$. In addition, H has only absolutely continuous spectrum consisting of a half-line.

(A3) J: $\mathscr{H}\to\mathscr{H}_N$ is a bounded linear operator. J maps $\mathscr{D}(H)$ into $\mathscr{D}(H_N)$, and J^*, the adjoint of J, maps $\mathscr{D}(H_N)$ into $\mathscr{D}(H)$. The operator JJ^*: $\mathscr{H}_N\to\mathscr{H}_N$ has a bounded inverse.

(A4) The operators V: $\mathscr{D}(V)\subset\mathscr{H}\to\mathscr{H}_N$ and V^*: $\mathscr{D}(V^*)\subset\mathscr{H}_N\to\mathscr{H}$,

$$V\equiv H_NJ-JH \quad\text{and}\quad V^*\equiv J^*H_N-HJ^*, \tag{2.3}$$

satisfy $V\ll H$ and $V^*\ll H_N$, respectively, where $K\ll L$ means that K is infinitesimally small with respect to L.[3,11]

(A5) The wave operators $\Omega^\pm:\mathscr{H}\to\mathscr{H}_N$, defined by

$$\Omega^\pm\equiv s\text{-}\lim_{t\to\pm\infty}e^{iH_Nt}Je^{-iHt}, \tag{2.4}$$

exist and satisfy

$$\Omega^{\pm*}\Omega^\pm=I. \tag{2.5}$$

The exact transition operator $T(z)$, defined in Eq. (2.1), is obtained from the M *operator* $M(z)$: $\mathscr{D}(H)\subset\mathscr{H}\to\mathscr{H}$,

$$M(z)\equiv(z-H)J^*(JJ^*)^{-1}R_N(z)V \tag{2.6}$$

by the identity

$$T(z)=(z-H)J^*JR(z)M(z), \tag{2.7}$$

where $R(z)\equiv(z-H)^{-1}$. Furthermore, $M(z)$ is a solution of the M equation

$$M(z)=J^*V+[J^*VR(z)-(J^*J-I)]M(z). \tag{2.8}$$

The operators $T(z)$ and $M(z)$ have domain and range in the direct sum asymptotic space $\mathscr{H}=\oplus_A\mathscr{H}_A$, where A is a *partition* (called a clustering in Ref. 1) of the N particles into n_A disjoint *clusters* (called fragments in Ref. 1), and $\mathscr{H}_A$ is a *partition Hilbert space.* We denote the *partition matrix elements* of $M(z)$ [$T(z)$ by $M_{BA}(z)$ $T_{BA}(z)$]. These operators map $\mathscr{D}(H_A)\subset\mathscr{H}_A$ into $\mathscr{H}_B$, where H_A is a *partition Hamiltonian* operator. In terms of partition matrix elements, Eq. (2.8) may be rewritten in the form

$$M_{BA}(z)=P_B\bar{V}_AP_A+\sum_C P_B[\bar{V}_CR_C(z)-\bar{\delta}_{BC}]P_CM_{CA}(z), \tag{2.9}$$

where $R_A(z)\equiv(z-H_A)^{-1}$, $\bar{V}_A\equiv H_N-H_A$, P_A is the orthogonal projection of $\mathscr{H}_N$ onto $\mathscr{H}_A$, *and* $\bar{\delta}_{BC}\equiv1-\delta_{BC}$, with δ_{BC} the Kronecker delta.

B. Approximate scattering systems

Associated with the exact scattering system $\mathfrak{S}$ are *approximate scattering systems,*

$$\mathcal{S}(\Pi)=\{\mathscr{H}_\pi,H_\pi,\mathscr{H}^\pi,H,J^\pi,\Pi\},\tag{2.10}$$

where $\Pi:\mathscr{H}\to\mathscr{H}^\pi\equiv\Pi\mathscr{H}$ is an orthogonal projection operator that commutes with H. Here $J^\pi\equiv J\Pi$, and $\mathscr{H}_\pi$ is the closure of the range of J^π. In addition, H_π is the *approximate total Hamiltonian*, defined by

$$H_\pi\equiv P_\pi H_N P_\pi,\tag{2.11}$$

where P_π is the orthogonal projection of $\mathscr{H}_N$ onto $\mathscr{H}_\pi$.

The *approximate scattering operator* $S^\pi:\mathscr{H}^\pi\to\mathscr{H}^\pi$ is obtainable from the *approximate transition operator*,

$$T^\pi(z)\equiv(z-H)J^{\pi *}R_\pi(z)V^\pi,\tag{2.12}$$

or the *approximate M* operator,

$$M^\pi(z)\equiv(z-H)J^{\pi *}(J^\pi J^{\pi *})^{-1}R_\pi(z)V^\pi,\tag{2.13}$$

where $R_\pi(z)\equiv(z-H_\pi)^{-1}$ and $V^\pi\equiv H_\pi J^\pi-J^\pi H$ (Ref. 4, Theorems 3.3 and 3.9). Furthermore, $M^\pi(z)$ is a solution of the equation

$$M^\pi(z)=J^{\pi *}V^\pi+[J^{\pi *}V^\pi R(z)-(J^{\pi *}J^\pi-\Pi)]M^\pi(z).$$

Our goal is to construct an operator Π and then solve Eq. (2.14) for $M^\pi(z)$. Now Π is a direct sum, $\Pi=\oplus_A\Pi_A$, of *partition projection operators* Π_A: $\mathscr{H}_A\subset\mathscr{H}_N\to\mathscr{H}_A^\pi\equiv\Pi_A\mathscr{H}_A$, and it suffices to construct the operators Π_A for each partition A. Denoting the partition matrix elements of $M^\pi(z)$ by $M^\pi_{BA}(z)$, Eq. (2.14) may be rewritten in the form

$$M^\pi_{BA}(z)=\Pi_B\bar{V}_A\Pi_A+\sum_C\Pi_B[\bar{V}_CR_C(z)-\bar{\delta}_{BC}]\Pi_CM^\pi_{CA}(z).\tag{2.15}$$

Assumption Π. In the notation of the present paper (with $\mathscr{H}=\mathscr{H}^a$), the approximate scattering system $\mathcal{S}(\Pi)$ is said to satisfy Assumption Π if the following statements are true.

(Π0) The exact scattering system $\mathcal{S}$ satisfies Assumptions (A1)–(A4).

(Π1) The orthogonal projection $\Pi:\mathscr{H}\to\mathscr{H}^\pi$ maps $\mathscr{D}(H)$ into $\mathscr{D}(H)\cap\mathscr{H}^\pi$.

(Π2) The operator Π commutes with H on $\mathscr{D}(H)$.

(Π3) The operator $J^\pi J^{\pi *}$ has a bounded inverse on $\mathscr{H}_\pi$.

III. KINETIC ENERGY HYPERSPHERES

Let A denote a partition of N particles ($N\geqslant 2$) moving in n-dimensional space ($n\geqslant 1$) into n_A disjoint clusters. Let the momenta of the particles be given in terms of *clustered Jacobi momenta* $(\mathbf{p}_A,\mathbf{q}_A)$, where $\mathbf{p}_A=(\mathbf{p}_A^{(1)},\ldots,\mathbf{p}_A^{(N-n_A)})$ is the collection of $n(N-n_A)$ momenta *internal* to the clusters and $\mathbf{q}_A=(\mathbf{q}_A^{(1)},\ldots,\mathbf{q}_A^{(n_A-1)})$ is the collection of $n(n_A-1)$ momenta *external* to (or between) the clusters with the total momentum of the entire system removed (cf. Ref. 11, Sec. XI.5 of Vol. III and Ref. 12, Chapters 14 and 15). Here $\mathbf{p}_A^{(m)}$ and $\mathbf{q}_A^{(m)}$ are n-dimensional internal and external momentum vectors, respectively. A *channel* α is a specification of a partition A and a bound state for each cluster in A. The *channel subspace* $\mathscr{H}_\alpha$ is then a tensor product space,

$$\mathscr{H}_\alpha=\hat{\mathscr{H}}_\alpha\otimes\mathscr{H}_A^0,\tag{3.1}$$

where $\hat{\mathscr{H}}_\alpha\subset\mathscr{L}^2(\mathbb{R}^{n(N-n_A)})$ is the one-dimensional subspace consisting of multiples of the product of the normalized cluster bound state wave functions, and $\mathscr{H}_A^0\equiv\mathscr{L}^2(\mathbb{R}^{n(n_A-1)})$. The *partition subspace* $\mathscr{H}_A$ is the closed linear span of the orthogonal subspaces $\mathscr{H}_\alpha$. The *partition Hamiltonian* H_A, with the center of mass removed, decomposes into a sum,

$$H_A=\hat{H}_A+H_A^0.\tag{3.2}$$

If $\psi_\alpha=\hat{\phi}_\alpha\otimes\psi_A^0\in\mathscr{H}_\alpha=\hat{\mathscr{H}}_\alpha\otimes\mathscr{H}_A^0$, then

$$(H_A\psi_\alpha)(\mathbf{p}_A,\mathbf{q}_A)=[\epsilon_\alpha+T_A(\mathbf{q}_A)]\hat{\phi}_\alpha(\mathbf{p}_A)\psi_A^0(\mathbf{q}_A),\tag{3.3}$$

where ϵ_α is the threshold energy (a sum of cluster eigenvalues) for channel α and

$$T_A(\mathbf{q}_A)\equiv\sum_{m=1}^{n_A-1}\frac{q_A^{(m)^2}}{2\mu_A^{(m)}},\tag{3.4}$$

the kinetic energy of relative motion of the clusters in partition A. Here the $\mu_A^{(m)}$ are the reduced masses associated with the Jacobi variables $\mathbf{q}_A$, and $q_A^{(m)^2}\equiv|\mathbf{q}_A^{(m)}|^2$. (Recall that $q_A^{(m)^2}$ is the negative of an n-dimensional Laplacian operator in coordinate space.)

For a fixed kinetic energy, the right-hand side of Eq. (3.4) is a *hyperellipse* in $n(n_A-1)$ dimensions. In order to work with *hyperspheres*, we change scale by defining scaled momentum vectors,

$$\mathbf{k}_A^{(m)}\equiv\mathbf{q}_A^{(m)}/\sqrt{2\mu_A^{(m)}}\tag{3.5}$$

in $\mathbb{R}^n$ and $\mathbf{k}_A\equiv(\mathbf{k}_A^{(1)},\ldots,\mathbf{k}_A^{(n_A-1)})$ in $\mathbb{R}^{n(n_A-1)}$. We next express $\mathbf{k}_A$ in hyperspherical coordinates $(\hat{k}_A,k_A)$ by letting $k_A\equiv\|\mathbf{k}_A\|$ and letting $\hat{k}_A$ denote the $n(n_A-1)-1$ angular variables. We finally let $\lambda\equiv k_A^2$. Since

$$\lambda=T_A(\mathbf{q}_A(\hat{k}_A,\lambda)),\tag{3.6}$$

the coordinates $(\hat{k}_A,\lambda)$ define *kinetic energy hyperspheres* in $n(n_A-1)$ dimensions. We denote the $n(n_A-1)-1$-dimensional surface of the *unit* kinetic energy hypersphere by Γ_A.

Lemma 3.1: The Jacobian of the transformation from the external coordinates $\mathbf{q}_A$ to the kinetic energy hyperspherical coordinates $(\hat{k}_A,\lambda)$ is

$$\frac{\partial\mathbf{q}_A}{\partial(\hat{k}_A,\lambda)}=j_A\lambda^{[n(n_A-1)-2]/2}=\nu_A^2(\lambda),\tag{3.7}$$

where

$$j_A\equiv\frac{1}{2}\prod_{m=1}^{n_A-1}(2\mu_A^{(m)})^{n/2}\tag{3.8}$$

and

$$\nu_A(\lambda)\equiv j_A^{1/2}\lambda^{[n(n_A-1)-2]/4}.\tag{3.9}$$

Proof: By the multiplication property of Jacobians,

$$\begin{aligned}\frac{\partial\mathbf{q}_A}{\partial(\hat{k}_A,\lambda)}&=\frac{\partial\mathbf{q}_A}{\partial\mathbf{k}_A}\frac{\partial\mathbf{k}_A}{\partial(\hat{k}_A,k_A)}\frac{\partial(\hat{k}_A,k_A)}{\partial(\hat{k}_A,\lambda)}\\&=(2j_A)(k_A^{n(n_A-1)-1})(2\lambda^{1/2})^{-1}\\&=j_A\lambda^{[n(n_A-1)-2]/2}.\end{aligned}\tag{3.10}$$

□

The inverse transformation from the $(\hat{k}_A,\lambda)$ coordinates to the $\mathbf{q}_A$ coordinates is obtained from

$$\mathbf{q}_A = (\sqrt{2\mu_A^{(1)}\lambda}\,\hat{k}_A^{(1)},\ldots,\sqrt{2\mu_A^{(n_A-1)}\lambda}\,\hat{k}_A^{(n_A-1)}). \quad (3.11)$$

The Jacobian of this inverse transformation is the reciprocal of the Jacobian in Eq. (3.7).

The Hilbert space $\mathscr{H}_A^0 = \mathscr{L}^2(\mathbb{R}^{n(n_A-1)})$ in Eq. (3.1) is isomorphic to the tensor product space,

$$\mathscr{H}_A^0 = \mathscr{L}^2(\Gamma_A)\otimes\mathscr{L}^2(\mathbb{R}^+), \quad (3.12)$$

where $\mathscr{L}^2(\Gamma_A)$ is the Hilbert space of square integrable functions $\chi(\hat{k}_a)$ on Γ_A and $\mathscr{L}^2(\mathbb{R}^{(+)})$ is the Hilbert space of square integrable functions $f(\lambda)$ of the kinetic energy $\lambda\in\mathbb{R}^+\equiv[0,\infty)$.

Combining Eqs. (3.1) and (3.12) we obtain the decomposition

$$\mathscr{H}_\alpha = \hat{\mathscr{H}}_\alpha\otimes\mathscr{L}^2(\Gamma_A)\otimes\mathscr{L}^2(\mathbb{R}^+), \quad (3.13)$$

for the channel subspace $\mathscr{H}_\alpha$.

The channel subspaces $\mathscr{H}_\alpha$ in Eq. (3.13) are orthogonal subspaces of the partition subspace $\mathscr{H}_A$. Therefore we may sum them to obtain the decomposition

$$\mathscr{H}_A = \sum_{\alpha\in A}[\hat{\mathscr{H}}_\alpha\otimes\mathscr{L}^2(\Gamma_A)\otimes\mathscr{L}^2(\mathbb{R}^+)]. \quad (3.14)$$

The asymptotic Hilbert space $\mathscr{H}$ may then be written in the dissected form,

$$\mathscr{H} = \bigoplus_A\sum_{\alpha\in A}[\hat{\mathscr{H}}_A\otimes\mathscr{L}^2(\Gamma_A)\otimes\mathscr{L}^2(\mathbb{R}^+)]. \quad (3.15)$$

IV. THE INJECTION OPERATORS ρ

Let $\{\hat{\phi}_\alpha(\mathbf{p}_A)\}$ be a complete orthonormal set of bound state wave functions in $\hat{\mathscr{H}}_A$ and let $\{g_m(\lambda)\}$ be a complete orthonormal basis in $\mathscr{L}^2(\mathbb{R}^+)$. For each bound state $\alpha\in A$ let $\{\chi_{\alpha i}(\hat{k}_A)\}$, $i=1,2,\ldots$, be some countably infinite linearly independent set that is uniformly bounded and forms a Schauder basis (Ref. 13, pp. 277–280) for $\mathscr{L}^2(\Gamma_A)$.

Remark 4.1: For reasons of flexibility, we do not always assume that all of the functions $\chi_{\alpha i}(\hat{k}_A)$ are orthogonal. If, for example, they are chosen to be spherical harmonics, then they are orthogonal. However, if they are chosen to be spline functions, then they may not be orthogonal. Therefore, for *computational* purposes, we wish to allow the possibility that at least a finite number of the $\chi_{\alpha i}(\hat{k}_A)$ are not orthogonal. For *theoretical* purposes, however, we can assume that all except possibly a finite number of the $\chi_{\alpha i}(\hat{k}_A)$ have been orthonormalized using, for example, the Gram–Schmidt process. The resulting set $\{\chi_{\alpha i}(\hat{k}_A)\}$ is then guaranteed to be a Schauder basis for $\mathscr{L}^2(\Gamma_A)$ (Ref. 13).

By Eqs. (3.14) and (3.7), any vector $\psi_A\in\mathscr{H}_A$ has an expansion in terms of these basis sets of the form

$$\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda) = \nu_A^{-1}(\lambda)\sum_{\alpha,i,m}c_{\alpha im}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)g_m(\lambda). \quad (4.1)$$

Summing with respect to m and letting $f_{\alpha i}(\lambda)\equiv\Sigma_m c_{\alpha im}g_m(\lambda)$, yields

$$\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda) = \nu_A^{-1}(\lambda)\sum_{\alpha,i}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda). \quad (4.2)$$

According to Eq. (3.15), any vector $\Psi\in\mathscr{H}=\oplus_A\mathscr{H}_A$ may then be represented by an expansion of the form

$$\Psi(\mathbf{p}_A,\hat{k}_A,\lambda) = \bigoplus_A\nu_A^{-1}(\lambda)\sum_{\alpha,i}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda). \quad (4.3)$$

In order to simplify the notation, the indices α, i (or α', i', etc.) will be used only for partition A, the indices β, j will be used only for partition B, and the indices γ, k will be used only for partition C. The dependence of $f_{\alpha i}$, $\hat{\phi}_\alpha$, $\chi_{\alpha i}$, etc., on A is then implicit and need not be explicitly shown. Also, when no confusion will arise, we will use the symbols $\|\cdot\|$ and $(\cdot,\cdot)$ to denote an $\mathscr{L}^2$ norm and an inner product, respectively, on an $\mathscr{L}^2$ Hilbert space. The vectors will make it clear which space is referred to.

The coefficient functions $f_{\alpha i}(\lambda)$ in Eqs. (4.2) and (4.3) belong to $\mathscr{L}^2(\mathbb{R}^+)$ and satisfy $\|f_{\alpha i}\|^2=\Sigma_m|c_{\alpha im}|^2<\infty$. We define injection operators $\rho_{\alpha i}:\mathscr{H}_A\to\mathscr{L}^2(\mathbb{R}^+)$ for $\psi_A\in\mathscr{H}_A$ by

$$(\rho_{\alpha i}\psi_A)(\lambda) \equiv \nu_A(\lambda)\int d\mathbf{p}_A\,d\hat{k}_A\,\hat{\phi}_\alpha^*(\mathbf{p}_A)\chi_{\alpha i}^*(\hat{k}_A)\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda). \quad (4.4)$$

Here $\hat{\phi}_\alpha^*$ and $\chi_{\alpha i}^*$ are the complex conjugates of $\hat{\phi}_\alpha$ and $\chi_{\alpha i}$, respectively. In order to handle all indices α, i and partitions A simultaneously, we take an external direct sum and define the *partition computation space* $\mathscr{L}_A$ by

$$\mathscr{L}_A \equiv \bigoplus_{\alpha,i\in A}\mathscr{L}^2(\mathbb{R}^+), \quad (4.5)$$

and the *computation space* $\mathscr{L}$ by

$$\mathscr{L} \equiv \bigoplus_A\mathscr{L}_A = \bigoplus_{\alpha,i\in A}\mathscr{L}^2(\mathbb{R}^+). \quad (4.6)$$

The $f_A=\oplus_{\alpha,i}f_{\alpha i}\in\mathscr{L}_A$ if $\Sigma_{\alpha,i}\|f_{\alpha i}\|^2<\infty$, and $f=\oplus_{\alpha,i,A}f_{\alpha i}\in\mathscr{L}$ if $\Sigma_{\alpha,i,A}\|f_{\alpha i}\|^2<\infty$. Communication from $\mathscr{H}_A$ to $\mathscr{L}_A$ is then provided by the injection operator $\rho_A:\mathscr{H}_A\to\mathscr{L}_A$, defined for $\psi_A\in\mathscr{H}_A$, by

$$(\rho_A\psi_A)(\lambda)\equiv\Big(\bigoplus_{\alpha,i\in A}\rho_{\alpha i}\psi_A\Big)(\lambda) = \bigoplus_{\alpha,i\in A}\nu_A(\lambda)\int d\mathbf{p}_A\,d\hat{k}_A \times\hat{\phi}_\alpha^*(\mathbf{p}_A)\chi_{\alpha i}^*(\hat{k}_A)\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda). \quad (4.7)$$

Communication from $\mathscr{H}$ to $\mathscr{L}$ is accomplished by the injection operator $\rho:\mathscr{H}\to\mathscr{L}$, defined for $\Psi=\oplus\psi_A\in\mathscr{H}$, by

$$(\rho\Psi)(\lambda)\equiv\left(\bigoplus_A \rho_A\psi_A\right)(\lambda)$$

$$=\bigoplus_{\alpha,i,A}\nu_A(\lambda)\int d\mathbf{p}_A\,d\hat{k}_A\,\hat{\phi}^*_\alpha(\mathbf{p}_A)\times\chi^*_{\alpha i}(\hat{k}_A)\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda). \tag{4.8}$$

The fact that the ranges of $\rho_{\alpha i}$, ρ_A, and ρ are contained in $\mathscr{L}^2(\mathbb{R}^+)$, $\mathscr{L}_A$, and $\mathscr{L}$, respectively, will follow from Lemma 4.4 below.

Communication in the reverse direction is provided by the adjoint operators. The following lemma provides formulas for these adjoints.

Lemma 4.2: The adjoint of $\rho_{\alpha i}$ is the injection operator $\rho^*_{\alpha i}\colon \mathscr{L}^2(\mathbb{R}^+)\to\mathscr{H}_\alpha$, defined for $f_{\alpha i}\in\mathscr{L}^2(\mathbb{R}^+)$, by

$$(\rho^*_{\alpha i}f_{\alpha i})(\mathbf{p}_A,\hat{k}_A,\lambda)\equiv\nu_A^{-1}(\lambda)\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda). \tag{4.9}$$

The adjoint of ρ_A is the injection operator $\rho^*_A\colon \mathscr{L}_A\to\mathscr{H}_A$, defined for $f_A=\oplus f_{\alpha i}\in\mathscr{L}_A$, by

$$(\rho^*_A f_A)(\mathbf{p}_A,\hat{k}_A,\lambda)\equiv\nu_A^{-1}(\lambda)\sum_{\alpha,i}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda). \tag{4.10}$$

The adjoint of ρ is the injection operator $\rho^*\colon\mathscr{L}\to\mathscr{H}$ defined for $f=\oplus_{\alpha,i,A}f_{\alpha i}\in\mathscr{L}$ by

$$(\rho^* f)\equiv\left(\bigoplus_A\rho^*_A f_A\right)(\mathbf{p}_A,\hat{k}_A,\lambda)$$

$$=\bigoplus_A\nu_A^{-1}(\lambda)\sum_{\alpha,i}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda). \tag{4.11}$$

Proof: Let ρ be defined by Eq. (4.8), and let $\Psi\in\mathscr{H}$ and $f=\oplus_{\alpha,i,A}f_{\alpha i}\in\mathscr{L}$. Then

$$(\rho\Psi,f)_{\mathscr{L}}=\sum_{\alpha,i,A}(\rho_{\alpha i}\psi_A,f_{\alpha i})_{\mathscr{L}^2(\mathbb{R}^+)}$$

$$=\sum_{\alpha,i,A}\int_0^\infty d\lambda\nu_A(\lambda)\int d\mathbf{p}_A\,d\hat{k}_A\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)\psi^*_A(\mathbf{p}_A,\hat{k}_A,\lambda)f_{\alpha i}(\lambda)$$

$$=\sum_A\int d\mathbf{p}_A\,d\hat{k}_A d\lambda\nu_A^2(\lambda)\psi^*_A(\mathbf{p}_A,\hat{k}_A,\lambda)\nu_A^{-1}(\lambda)\sum_{\alpha,i}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda)$$

$$=\sum_A(\psi_A,\rho^*_A(f_A))_{\mathscr{H}_A}=(\Psi,\rho^*f)_{\mathscr{H}}, \tag{4.12}$$

with ρ^* defined by Eq. (4.11). The interchange in the order of integration is justified by the Tonelli and Fubini theorems and the interchange in the order of summation and integration is justified by the Lebesgue dominated convergence theorem.[14] This proves Eq. (4.11). The proof of Eq. (4.9) is obtained by omitting the summations over α, i, and A in Eq. (4.12). Equation (4.10) is obtained by omitting the summations over A in Eq. (4.12). □

Remark 4.3: When expressed in clustered Jacobi momentum coordinates $(\mathbf{p}_A,\mathbf{q}_A)$ the injection operators ρ and ρ^* become

$$(\rho\Psi)(\lambda)=\bigoplus_{\alpha,i,A}\int d\mathbf{p}_A\,d\mathbf{q}_A\delta(\lambda-T_A(\mathbf{q}_A))\times\hat{\phi}^*_\alpha(\mathbf{p}_A)X^*_{\alpha i}(\mathbf{q}_A)\psi_A(\mathbf{p}_A,\mathbf{q}_A) \tag{4.13}$$

and

$$(\rho^*f)(\mathbf{p}_A,\mathbf{q}_A)=\bigoplus_A\sum_{\alpha,i}\hat{\phi}_\alpha(\mathbf{p}_A)X_{\alpha i}(\mathbf{q}_A)f_{\alpha i}(T_A(\mathbf{q}_A)), \tag{4.14}$$

respectively, where δ is the Dirac delta function and

$$X_{\alpha i}(\mathbf{q}_A)\equiv\chi_{\alpha i}(\hat{k}_A(\mathbf{q}_A))j_A^{-1/2}[T_A(\mathbf{q}_A)]^{-[n(n_A-1)-2]/4}. \tag{4.15}$$

The definition of $\rho_{\alpha i}$ in Ref. 15 corresponds to Eq. (4.13) with $X_{\alpha i}(\mathbf{q}_A)$ denoted by $\chi_{\alpha i}(\mathbf{q}_A)$ and $T_A(\mathbf{q}_A)$ replaced by $T_A(\mathbf{q}_A)+\epsilon_\alpha$.

In general, we denote the inner product of $\chi_{\alpha i}$ with $\chi_{\alpha' i'}$ by $\omega_{\alpha' i',\alpha i}$. That is,

$$\omega_{\alpha' i',\alpha i}\equiv(\chi_{\alpha' i'},\chi_{\alpha i})=\int_{\Gamma_A}d\hat{k}_A\,\chi^*_{\alpha' i'}(\hat{k}_A)\chi_{\alpha i}(\hat{k}_A). \tag{4.16}$$

Let ω_A denote the (countably infinite-dimensional) block diagonal matrix with elements $\delta_{\alpha'\alpha}\omega_{\alpha' i',\alpha i}$. That is,

$$\omega_A\equiv[\delta_{\alpha'\alpha}\omega_{\alpha' i',\alpha i}]. \tag{4.17}$$

We assume that the matrices ω_A are bounded and we denote the operator norm of ω_A by $\|\omega_A\|$. This assumption will, for example, be satisfied if all but a finite number of the $\chi_{\alpha i}(\hat{k}_A)$ functions are orthonormalized (cf. Remark 4.1). Finally, we define ω to be the block diagonal matrix with entries ω_A on the main diagonal and zeros elsewhere. That is,

$$\omega\equiv\begin{bmatrix}\omega_A & 0 & \cdots\\ 0 & \omega_B & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}. \tag{4.18}$$

We denote the operator norm of the bounded matrix ω by $\|\omega\|$.

If we view $f_A(\lambda)$ as a column vector with components $f_{\alpha i}(\lambda)$, then $\omega_A f_A$ is defined by matrix multiplication and $\omega_A\colon\mathscr{L}_A\to\mathscr{L}_A$. If we view $f(\lambda)$ as a column vector with components $f_A(\lambda)$, then ωf is defined by matrix multiplication and $\omega\colon\mathscr{L}\to\mathscr{L}$. The linear independence of the sets $\{\hat{\phi}_A(\mathbf{p}_A)\}$ and $\{\chi_{\alpha i}(\hat{k}_A)\}$ assures that the matrices ω_A are

nonsingular for all partitions A and, hence, that ω is nonsingular. We denote the inverse matrices by ω_A^{-1} and ω^{-1}, respectively.

Lemma 4.4: The operator ρ is a bounded bijective (one to one and onto) mapping from $\mathscr{H}$ to $\mathscr{L}$. The adjoint operator ρ^* is a bounded bijective mapping from $\mathscr{L}$ to $\mathscr{H}$.

Proof: Since $\rho = \oplus_A \rho_A$, it suffices to show that ρ_A is a bounded bijective mapping from $\mathscr{H}_A$ to $\mathscr{L}_A$ for each A. Let $\psi_A \in \mathscr{H}_A$ be given by Eq. (4.2) and define $\rho_A\psi_A$ by Eq. (4.7). Then

$$(\rho_A\psi_A)(\lambda) = \bigoplus_{\alpha',i'} \int d\mathbf{p}_A\, d\hat{k}_A \hat{\phi}^*_{\alpha'}(\mathbf{p}_A)\chi^*_{\alpha' i'}(\hat{k}_A) \times \sum_{\alpha,i} \hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A) f_{\alpha i}(\lambda)$$

$$= \bigoplus_{\alpha',i'} \sum_{\alpha,i} \delta_{\alpha'\alpha}\omega_{\alpha' i',\alpha i} f_{\alpha i}(\lambda) = \omega_A f_A(\lambda), \tag{4.19}$$

where the interchange in the order of summation and integration is valid by the Lebesgue dominated convergence theorem. It follows that the $\mathscr{L}_A$ norm of $\rho_A\psi_A$ satisfies

$$\|\rho_A\psi_A\| = \|\omega_A f_A\| \leqslant \|\omega_A\|\|f_A\| < \infty, \tag{4.20}$$

which shows that ρ_A is bounded and $\rho_A\psi_A \in \mathscr{L}_A$. Suppose that $(\rho_A\psi_A)(\lambda) = 0$. Then

$$f_A(\lambda) = (\omega_A^{-1}\omega_A f_A)(\lambda) = (\omega_A^{-1}\rho_A\psi_A)(\lambda) = 0, \tag{4.21}$$

and the representation of ψ_A in Eq. (4.2) implies that $\psi_A = 0$. Since ρ_A is a linear operator, it follows that ρ_A is injective (one to one). In order to show that ρ_A is surjective (onto), let $f_A(\lambda)$ be an arbitrary function in $\mathscr{L}_A$. Let $(\omega_A^{-1}f_A)_{\alpha i}$ denote the αith component of $\omega_A^{-1}f_A$. Then

$$\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda) \equiv \nu_A^{-1}(\lambda)\sum_{\alpha,i} \hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)(\omega_A^{-1}f_A)_{\alpha i} \tag{4.22}$$

is in $\mathscr{H}_A$ and $(\rho_A\psi_A)(\lambda) = f_A(\lambda)$. It follows that ρ_A and, hence, ρ is bijective.

Since $\rho^* = \oplus_A \rho_A^*$, it suffices to show that ρ_A^* is a bounded bijective mapping from $\mathscr{L}_A$ to $\mathscr{H}_A$ for each A. Let $f_A = \oplus f_{\alpha i} \in \mathscr{L}_A$ and define $\rho_A^* f_A$ by Eq. (4.10). Let $c_A \equiv \sup|\omega_{\alpha i,\alpha i}|$. Then

$$\|\rho_A^* f_A\|^2 \leqslant \sum_{\alpha,i} \int d\mathbf{p}_A\, d\hat{k}_A\, d\lambda\, |\hat{\phi}_\alpha(\mathbf{p}_A)|^2 |\chi_{\alpha i}(\hat{k}_A)|^2 |f_{\alpha i}(\lambda)|^2$$

$$\leqslant c_A \sum_{\alpha,i} \int d\lambda\, |f_{\alpha i}(\lambda)|^2 = c_A \sum_{\alpha,i} \|f_{\alpha i}\|^2 = c_A\|f_A\|^2 < \infty, \tag{4.23}$$

which implies that ρ_A^* is bounded and $\rho_A^* f_A \in \mathscr{H}_A$. Let the ψ_A in Eq. (4.2) be an arbitrary vector in $\mathscr{H}_A$ and let $f_A \equiv \oplus_{\alpha i} f_{\alpha i}$, then $\rho_A^* f_A = \psi_A$ and ρ_A^* is surjective. Finally, if $(\rho_A^* f_A)(\mathbf{p}_A,\hat{k}_A,\lambda) = 0$, then the uniqueness of the coefficients $f_{\alpha i}(\lambda)$ in the Schauder basis expansion in Eq. (4.10) implies that $f_{\alpha i}(\lambda) = 0$ for all α, i. Thus $f_A(\lambda) = 0$. This, and the linearity of ρ_A^*, proves that ρ_A^* is injective and, hence, bijective. □

The following theorem gives a matrix representation of the operators $\rho\rho^*$ and $(\rho\rho^*)^{-1}$, and establishes that $\rho^*(\rho\rho^*)^{-1}\rho$ is the identity operator on $\mathscr{H}$.

Theorem 4.5: A matrix representation of the operator $\rho\rho^*$: $\mathscr{L} \to \mathscr{L}$ is

$$\rho\rho^* = \bigoplus_A \rho_A\rho_A^* = \bigoplus_A \omega_A\mathscr{I}_A = \omega\mathscr{I}, \tag{4.24}$$

where $\mathscr{I}$ is the identity operator on $\mathscr{L}$ and $\mathscr{I}_A$ is the identity operator on $\mathscr{L}_A$. That is, for and $f = \oplus_A f_A$,

$$(\rho\rho^* f)(\lambda) = \bigoplus_A (\rho_A\rho_A^* f_A)(\lambda) = \bigoplus_A \omega_A f_A(\lambda) = \omega f(\lambda). \tag{4.25}$$

A matrix representation of the operator $(\rho\rho^*)^{-1}$: $\mathscr{L} \to \mathscr{L}$ is

$$(\rho\rho^*)^{-1} = \bigoplus_A (\rho_A\rho_A^*)^{-1} = \bigoplus_A \omega_A^{-1}\mathscr{I}_A = \omega^{-1}\mathscr{I}. \tag{4.26}$$

Furthermore, the identity

$$\rho^*(\rho\rho^*)^{-1}\rho = I \tag{4.27}$$

holds, where I is the identity operator on $\mathscr{H}$.

Proof: By Eqs. (4.8), (4.11), and (4.19),

$$(\rho\rho^* f)(\lambda) = \bigoplus_A (\rho_A\rho_A^* f_A)(\lambda) = \bigoplus_A (\omega_A f_A)(\lambda) = \omega f(\lambda). \tag{4.28}$$

Since $\rho\rho^*$ is a bijective mapping by Lemma 4.4, the inverse operator $(\rho\rho^*)^{-1}$ exists and is given by Eq. (4.26).

Let $\Psi = \oplus_A \psi_A$ with $\psi_A \in \mathscr{H}_A$ given by Eq. (4.2) be an arbitrary vector in $\mathscr{H}$. Using Eqs. (4.8), (4.11), (4.19), (4.26), and (4.10), we obtain

$$\rho^*(\rho\rho^*)^{-1}\rho\Psi = \bigoplus_A \rho_A^*(\rho_A\rho_A^*)^{-1}\rho_A\psi_A$$

$$= \bigoplus_A \rho_A^*\omega_A^{-1}\omega_A f_A$$

$$= \bigoplus_A \psi_A = \Psi. \tag{4.29}$$

This completes the proof of the theorem. □

A consequence of Eqs. (3.3), (3.6), and (4.9) is that

$$(H_A\rho_{\alpha i}^* f_{\alpha i})(\mathbf{p}_A,\hat{k}_A,\lambda) = (\epsilon_\alpha + \lambda)\nu_A^{-1}(\lambda)\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda). \tag{4.30}$$

By forming sums of powers of H_A and then limits of these sums, we establish that

$$(h(H_A)\rho_{\alpha i}^* f_{\alpha i})(\mathbf{p}_A,\hat{k}_A,\lambda) = h(\epsilon_\alpha+\lambda)\nu_A^{-1}(\lambda)\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda), \quad (4.31)$$

for any continuous function h. This *functional calculus* may be generalized to obtain a matrix representation for operators of the form $h(H)\rho^*$. However, we are mainly interested in the case when $h(H)$ is the resolvent operator $R(z)\equiv(z-H)^{-1}$ for Im $z\neq 0$, and we will exhibit the generalization only for this case. Let $\mathscr{R}_\alpha(z)$: $\mathscr{L}^2(\mathbb{R}^+)\to\mathscr{L}^2(\mathbb{R}^+)$ be defined for $f_{\alpha i}(\lambda)\in\mathscr{L}^2(\mathbb{R}^+)$ by

$$(\mathscr{R}_\alpha(z)f_{\alpha i})(\lambda)\equiv(z-\epsilon_\alpha-\lambda)^{-1}f_{\alpha i}(\lambda) \quad (4.32)$$

and let

$$\mathscr{R}_A(z)\equiv[\mathscr{R}_\alpha(z)\delta_{\alpha' i',\alpha i}] \quad (4.33)$$

denote the diagonal matrix with diagonal elements $\mathscr{R}_\alpha(z)$. Let $\mathscr{R}(z)$ denote the diagonal matrix

$$\mathscr{R}(z)\equiv\begin{bmatrix}\mathscr{R}_A(z) & 0 & \cdots\\ 0 & \mathscr{R}_B(z) & \cdots\\ \vdots & \vdots & \ddots\end{bmatrix}. \quad (4.34)$$

We then have the following theorem.

Theorem 4.6: The operator $\rho R(z)\rho^*$: $\mathscr{L}\to\mathscr{L}$ has the matrix representations

$$\rho R(z)\rho^* = \omega\mathscr{R}(z)\mathscr{I} = \mathscr{R}(z)\omega\mathscr{I} \quad (4.35)$$

for Im $z\neq 0$. That is, for any $f(\lambda)\in\mathscr{L}$,

$$(\rho R(z)\rho^* f)(\lambda) = \omega\mathscr{R}(z)f(\lambda) = \mathscr{R}(z)\omega f(\lambda). \quad (4.36)$$

Proof: Since $R(z)=\oplus_A R_A(z)$, Eqs. (4.8), (4.11), (4.28)–(4.30), and (4.24) yield

$$\begin{aligned}(\rho R(z)\rho^* f)(\lambda) &= \bigoplus_A(\rho_A R_A(z)\rho_A^* f_A)(\lambda)\\ &= \bigoplus_A \rho_A\nu_A^{-1}(\lambda)\sum_{\alpha,i}\mathscr{R}_\alpha(z)\\ &\quad\times\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda)\\ &= \bigoplus_A(\rho_A\rho_A^*\mathscr{R}_A(z)f_A)(\lambda)\\ &= (\rho\rho^*\mathscr{R}(z)f)(\lambda)\\ &= \omega\mathscr{R}(z)f(\lambda). \end{aligned} \quad (4.37)$$

This proves the first of Eqs. (4.36). The second of Eqs. (4.36) then follows immediately, since the diagonal matrix $\mathscr{R}(z)$ commutes with ω. □

V. CONSTRUCTION OF THE PROJECTION OPERATOR Π

The infinite sum in Eq. (4.2) can be well approximated by a finite number of judiciously chosen terms. Let us suppose that we have chosen n_α terms from the set $\{\hat{\phi}_\alpha(\mathbf{p}_A)\}$ and, for each bound state α, we have chosen $n_{\alpha i}$ terms from the set $\{\chi_{\alpha i}(\hat{k}_A)\}$. Let $A(\Pi)$ denote the finite subset of indices α, $i\in A$ remaining after the truncation for partition A. The closed linear span of all vectors $\psi_A\in\mathscr{H}_A$ of the form

$$\psi_A(\mathbf{p}_A,\hat{k}_A,\lambda)=\nu_A^{-1}(\lambda)\sum_{\alpha,i\in A(\Pi)}\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A)f_{\alpha i}(\lambda) \quad (5.1)$$

is a subspace of $\mathscr{H}_A$, which we denote by $\mathscr{H}_A^\pi$. The direct sum $\mathscr{H}^\pi\equiv\oplus_A\mathscr{H}_A^\pi$ of the approximation subspaces $\mathscr{H}_A^\pi$ over all partitions A is then a closed *asymptotic approximation subspace* of $\mathscr{H}$. We remark that some of the subspaces $\mathscr{H}_A^\pi$ may be empty if the partition A is not of interest in a particular scattering process.

Let Π_A denote the orthogonal projection of $\mathscr{H}_A$ onto $\mathscr{H}_A^\pi$, and let $\Pi\equiv\oplus_A\Pi_A$ denote the orthogonal projection of $\mathscr{H}$ onto $\mathscr{H}^\pi$. It follows that if a vector $\psi_A\in\mathscr{H}_A$ has the expansion given in Eq. (4.2), then $\Pi_A\psi_A$ has the expansion given in Eq. (5.1).

Associated with the subspaces $\mathscr{H}_A^\pi$ are the *truncated partition computation subspaces* $\mathscr{L}_A^\pi$, defined by

$$\mathscr{L}_A^\pi\equiv\bigoplus_{\alpha,i\in A(\Pi)}\mathscr{L}^2(\mathbb{R}^+), \quad (5.2)$$

and the *truncated computation subspace* $\mathscr{L}^\pi$, defined by

$$\mathscr{L}^\pi\equiv\bigoplus_A\mathscr{L}_A^\pi. \quad (5.3)$$

We define *truncated injection operators* ρ^π: $\mathscr{H}\to\mathscr{L}^\pi$ and $\rho^{\pi*}$: $\mathscr{L}\to\mathscr{H}^\pi$ as follows. First, let

$$\rho_A^\pi\equiv\rho_A\Pi_A \quad\text{and}\quad \rho_A^{\pi*}\equiv\Pi_A\rho_A^* \quad (5.4)$$

and then let

$$\rho^\pi\equiv\rho\Pi=\bigoplus_A\rho_A^\pi \quad\text{and}\quad \rho^{\pi*}\equiv\Pi\rho^*=\bigoplus_A\rho_A^{\pi*}. \quad (5.5)$$

We also let ω_A^π denote the $n_{\alpha i}\times n_{\alpha i}$-dimensional block diagonal matrices defined by Eq. (4.17) with α, i restricted to be in $A(\pi)$. We finally define ω^π to be the block diagonal matrix with entries ω_A^π on the main diagonal and zeros elsewhere [cf. Eq. (4.18)].

Lemma 5.1: The operator ρ^π is a bounded bijective mapping from $\mathscr{H}^\pi$ to $\mathscr{L}^\pi$. The adjoint operator $\rho^{\pi*}$ is a bounded bijective mapping from $\mathscr{L}^\pi$ to $\mathscr{H}^\pi$.

Proof: The proof is a replication of the proof of Lemma 4.4 with the indices α, i and α', i' restricted to belong to $A(\Pi)$. □

Theorem 5.2: A matrix representation of the operator $\rho^\pi\rho^{\pi*}$: $\mathscr{L}^\pi\to\mathscr{L}^\pi$ is

$$\rho^\pi\rho^{\pi*}=\bigoplus_A\rho_A^\pi\rho_A^{\pi*}=\bigoplus_A\omega_A^\pi\mathscr{I}_A^\pi=\omega^\pi\mathscr{I}^\pi, \quad (5.6)$$

where $\mathscr{I}^\pi$ is the identity operator on $\mathscr{L}^\pi$ and $\mathscr{I}_A^\pi$ is the identity operator on $\mathscr{L}_A^\pi$. A matrix representation of the operator $(\rho^*\rho^{\pi*})^{-1}$: $\mathscr{L}^\pi\to\mathscr{L}^\pi$ is

$$\begin{aligned}(\rho^\pi\rho^{\pi*})^{-1} &= \bigoplus_A(\rho_A^\pi\rho_A^{\pi*})^{-1}\\ &= \bigoplus_A(\omega_A^\pi)^{-1}\mathscr{I}_A^\pi=(\omega^\pi)^{-1}\mathscr{I}^\pi.\end{aligned} \quad (5.7)$$

Proof: The proof is a replication of the proof of the corresponding results in Theorem 4.5. □

Theorem 5.3: The orthogonal projection operators Π_A:

$\mathscr{H}_A \to \mathscr{H}_A^\pi$ and $\Pi: \mathscr{H} \to \mathscr{H}^\pi$ are given by the identities

$$\Pi_A = \rho_A^{\pi *}(\rho_A^\pi \rho_A^{\pi *})^{-1}\rho_A^\pi = \rho_A^{\pi *}(\omega_A^\pi)^{-1}\rho_A^\pi \tag{5.8}$$

and

$$\Pi = \rho^{\pi *}(\rho^\pi \rho^{\pi *})^{-1}\rho^\pi = \rho^{\pi *}(\omega^\pi)^{-1}\rho^\pi, \tag{5.9}$$

respectively.

Proof: Let $\hat{\Pi}_A \equiv \rho_A^{\pi *}(\rho_A^\pi \rho_A^{\pi *})^{-1}\rho_A^\pi$. Then clearly, $\hat{\Pi}_A^* = \hat{\Pi}_A$ and $\hat{\Pi}_A^2 = \hat{\Pi}_A$. It follows that $\hat{\Pi}_A$ is an orthogonal projection operator. It remains to show that the range of Π_A is equal to the range of $\hat{\Pi}_A$. The definition of $\hat{\Pi}_A$ and the second equation in Eq. (5.4) imply that $\mathrm{Ran}(\hat{\Pi}_A) \subset \mathrm{Ran}(\Pi_A)$. Suppose that $\psi_A \in \mathrm{Ran}(\Pi_A)$. Then ψ_A has an expansion of the form of Eq. (5.1). Then Eq. (4.19) with $\alpha,i,\alpha',i' \in \mathscr{A}(\Pi)$ gives

$$(\rho_A^\pi \psi_A)(\lambda) = \omega_A^\pi f_A(\lambda), \tag{5.10}$$

where $f_A = \oplus_{\alpha,i \in \mathscr{A}(\Pi)} f_{\alpha i}$. Multiplication of Eq. (5.10) by $\rho_A^{\pi *}(\rho_A^\pi \rho_A^{\pi *})^{-1}$ and using the second of Eqs. (5.7), Eq. (4.10), and the second of Eqs. (5.4), yields that $\hat{\Pi}_A \psi_A = \psi_A$ and, hence, that $\mathrm{Ran}(\Pi_A) \subset \mathrm{Ran}(\hat{\Pi}_A)$. It follows that $\Pi_A = \hat{\Pi}_A$. This proves the first of Eqs. (5.8). The second equation in Eq. (5.8) is a consequence of Theorem 5.2.

Eqs. (5.9) follow from Eqs. (5.8) by forming the direct sum over all partitions A. □

It remains to show that our definition of the projection operator Π leads to an approximate scattering system $\mathfrak{S}(\Pi)$, which satisfies Assumption Π. This is necessary in order to invoke the results of previous papers in this series. The following theorem provides a verification of Assumption Π.

Theorem 5.4: Suppose that the exact scattering system $\mathfrak{S}$ satisfies Assumptions (A1)–(A4). Let $\Pi = \oplus_A \Pi_A$ be the orthogonal projection of $\mathscr{H} = \oplus_A \mathscr{H}_A$ onto $\mathscr{H}^\pi = \oplus_A \mathscr{H}_A^\pi$, where $\mathscr{H}_A^\pi$ is the closed linear span of vectors $\psi_A \in \mathscr{H}_A$ having the form of Eq. (5.1) when expressed in kinetic energy hyperspherical coordinates $(\mathbf{p}_A, \hat{k}_A, \lambda)$. Suppose that $\Pi_B \Pi_A$ for $B \neq A$ are all compact operators. Then the approximate scattering system $\mathfrak{S}(\Pi)$ satisfies Assumption Π.

Proof: Let $\Psi = \oplus_A \psi_A \in \mathscr{D}(H) \subset \mathscr{H}$. Then $\psi_A \in \mathscr{D}(H_A) \subset \mathscr{H}_A$ and

$$\Pi_A \psi_A = \nu_A^{-1}(\lambda) \sum_{\alpha,i \in \mathscr{A}(\Pi)} \hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A) f_{\alpha i}(\lambda). \tag{5.11}$$

Since the summation in Eq. (5.11) is finite, it follows from Eqs. (3.3) and (3.6) that $\Pi_A \psi_A \in \mathscr{D}(H_A) \cap \mathscr{H}_A^\pi$, and

$$H_A \Pi_A \psi_A = \nu_A^{-1}(\lambda) \sum_{i,i \in \mathscr{A}(\Pi)} (\epsilon_\alpha + \lambda)\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A) f_{\alpha i}(\lambda). \tag{5.12}$$

On the other hand, if ψ_A given by Eq. (4.2), belongs to $\mathscr{D}(H_A)$, then

$$H_A \psi_A = \nu_A^{-1}(\lambda) \sum_{\alpha,i} (\epsilon_\alpha + \lambda)\hat{\phi}_\alpha(\mathbf{p}_A)\chi_{\alpha i}(\hat{k}_A) f_{\alpha i}(\lambda), \tag{5.13}$$

with the summation converging in the norm of $\mathscr{H}_A$. Therefore $\Pi_A H_A \psi_A$ is also equal to the right-hand side of Eq. (5.12). This shows that Π_A maps $\mathscr{D}(H_A)$ into $\mathscr{D}(H_A) \cap \mathscr{H}_A^\pi$ and that Π_A commutes with H_A on $\mathscr{D}(H_A)$. It follows that

$$H\Pi\Psi = \bigoplus_A H_A \Pi_A \psi_A = \bigoplus_A \Pi_A H_A \psi_A = \Pi H \Psi, \tag{5.14}$$

which proves statements (Π1) and (Π2). Since $\Pi_B \Pi_A$ for $B \neq A$ are all compact operators, statement (Π3) follows from Ref. 3, Theorem 3.4 and Remark 3.5. □

Remark 5.5: The hypothesis in Theorem 5.4 that $\Pi_B \Pi_A$ are compact operators for all $B \neq A$ can be shown to be satisfied in most practical situations (cf. Ref. 15, Theorem 13). The lengthy proof will be given in a subsequent paper.

VI. INTEGRAL EQUATIONS

The operator form of the approximate M equations [Eq. (2.14) or (2.15)] is not directly useful for computations. In this section we derive computationally useful *integral equations* by mapping to and from the computation space $\mathscr{L}$ with the injection operators ρ and ρ^*.

Equation (2.14) is an approximate version of the exact M equation in Eq. (2.8). In the following manipulations there are also exact and approximate versions of all the equations. Notationally, the approximate versions are obtained from the corresponding exact versions by adding superscripts π to the appropriate operators or functions. Therefore we will present only the exact versions of the equations in this section.

All of the operators (J^*V), $(J^*J - I)$, $M(z)$, and $R(z)$, appearing in Eq. (2.8), map (a subspace of) the direct sum asymptotic Hilbert space $\mathscr{H}$ into $\mathscr{H}$. We define corresponding matrix versions of these operators that map (a subspace of) the direct sum computation space $\mathscr{L}$ into $\mathscr{L}$ as follows. Let

$$\mathscr{B} \equiv (\rho\rho^*)^{-1}\rho J^* V \rho^*, \tag{6.1}$$

$$\mathscr{C} \equiv (\rho\rho^*)^{-1}\rho(J^*J - I)\rho^*, \tag{6.2}$$

and

$$\mathscr{M}(z) \equiv (\rho\rho^*)^{-1}\rho M(z)\rho^*. \tag{6.3}$$

The corresponding operator $\mathscr{R}(z)$ has already been defined in Eq. (4.34). By Theorems 4.5 and 4.6 it satisfies

$$\mathscr{R}(z) = (\rho\rho^*)^{-1}\rho R(z)\rho^*. \tag{6.4}$$

We multiply Eq. (2.8) on the left-hand side by $(\rho\rho^*)^{-1}\rho$, on the right-hand side by ρ^*, and insert $I = \rho^*(\rho\rho^*)^{-1}\rho$ between V and $R(z)$ and also at the left of $M(z)$ in the right side of the equation. Then Eqs. (6.1)–(6.4) give

$$\mathscr{M}(z) = \mathscr{B} + [\mathscr{B}\mathscr{R}(z) - \mathscr{C}]\mathscr{M}(z). \tag{6.5}$$

Definition: The *kernel* of an operator $\mathscr{A}: \mathscr{L} \to \mathscr{L}$ is defined to be the function $\mathscr{A}(\lambda,u)$, which maps $f(\mu) \in \mathscr{L}$ into $(Af)(\lambda) \in \mathscr{L}$ via the formula

$$(\mathscr{A}f)(\lambda) = \int_0^\infty d\mu \mathscr{A}(\lambda,\mu) f(\mu). \tag{6.6}$$

The kernel of the diagonal matrix $\mathscr{R}(\lambda,z)$, defined in Eq. (4.34), is the matrix

kernel of $\mathscr{R}(z)$

$$\equiv\delta(\lambda-\mu)\mathscr{R}(\mu,z)$$
$$\equiv\delta(\lambda-\mu)[\delta_{BA}\,\delta_{\alpha' i',\alpha i}(z-\epsilon_\alpha-\mu)^{-1}]. \tag{6.7}$$

We assume that the operators $\mathscr{B}$, $\mathscr{C}$, and $\mathscr{M}(z)$, defined in Eqs. (6.1)–(6.3), are integral operators and can be represented by kernels. This assumption will be verified in a subsequent paper. We denote these kernels by $\mathscr{B}(\lambda,\mu)$, $\mathscr{C}(\lambda,\mu)$, and $\mathscr{M}(\lambda,\mu,z)$, respectively. The kernel form of Eq. (6.5) is then

$$\mathscr{M}(\lambda,\mu,z)=\mathscr{B}(\lambda,\mu)+\int_0^\infty d\eta[\mathscr{B}(\lambda,\eta)\mathscr{R}(\eta,z) - \mathscr{C}(\lambda,\eta)]\mathscr{M}(\eta,\mu,z). \tag{6.8}$$

We may view Eq. (6.8) as a matrix-valued function equation on $\mathscr{L}=\oplus_{\alpha,i,A}\mathscr{L}^2(\mathbb{R}^+)$. The elements of these matrices, therefore, satisfy the system of equations

$$\mathscr{M}_{\beta j,\alpha i}(\lambda,\mu,z) = \mathscr{B}_{\beta j,\alpha i}(\lambda,\mu)+\sum_C\sum_{\gamma,k\in C}\int_0^\infty d\eta\left[\frac{\mathscr{B}_{\beta j,\gamma k}(\lambda,\eta)}{z-\epsilon_\gamma-\eta} - \mathscr{C}_{\beta j,\gamma k}(\lambda,\eta)\right]\mathscr{M}_{\gamma k,\alpha i}(\eta,\mu,z), \tag{6.9}$$

where $\alpha,i\in A$ and $\beta,j\in B$.

The computational features of Eq. (6.9) can be improved by putting it into $\mathscr{K}$-*matrix* form. We first put Eq. (6.9) half-on-shell. Letting E denote the total energy, set $z=E+i\epsilon$ in Eq. (6.9), and then take the limit as $\epsilon\to0^+$. Noting that $\mu=E-\epsilon_\alpha$, we define $\mathscr{M}_{\beta j,\gamma k}(\lambda,\mu)$: $\mathscr{L}^2(\mathbb{R}^+)\to\mathscr{L}^2(\mathbb{R}^+)$ by the following limits:

$$\mathscr{M}_{\beta j,\gamma k}(\lambda,\mu)\equiv\lim_{\epsilon\to0^+}\mathscr{M}_{\beta j,\gamma k}(\lambda,\mu+\epsilon_\alpha-\epsilon_\gamma,\mu+\epsilon_\alpha+i\epsilon) = \mathscr{M}_{\beta j,\gamma k}(\lambda,E-\epsilon_\gamma,E+i0), \tag{6.10}$$

which we assume exist. We also define $\mathscr{M}(\lambda,\mu)$ to be the matrix with elements $\mathscr{M}_{\beta j,\gamma k}(\lambda,\mu)$. We next introduce the Cauchy principal value integral $\mathcal{f}$ via the identity (Ref. 11, Vol. I,p. 137)

$$\lim_{\epsilon\to0^+}\int_0^\infty d\eta\frac{\mathscr{B}_{\beta j,\gamma k}(\lambda,\eta)}{\mu+\epsilon_\alpha+i\epsilon-\epsilon_\gamma-\eta}\mathscr{M}_{\gamma k,\alpha i}(\eta,\mu) = \mathcal{f}\int_0^\infty d\eta\frac{\mathscr{B}_{\beta j,\gamma k}(\lambda,\eta)}{\mu+\epsilon_\alpha-\epsilon_\gamma-\eta}\mathscr{M}_{\gamma k,\alpha i}(\eta,\mu) - i\pi\mathscr{B}_{\beta j,\gamma k}(\lambda,\mu+\epsilon_\alpha-\epsilon_\gamma)\times\mathscr{M}_{\gamma k,\alpha i}(\mu+\epsilon_\alpha-\epsilon_\gamma,\mu). \tag{6.11}$$

Define $\mathscr{M}(\mu)$ to be the fully on-shell Kernel matrix with matrix elements

$$\mathscr{M}_{\beta j,\gamma k}(\mu)\equiv\mathscr{M}_{\beta j,\gamma k}(\mu+\epsilon_\alpha-\epsilon_\beta,\mu) = \mathscr{M}_{\beta j,\gamma k}(E-\epsilon_\beta,E-\epsilon_\gamma,E+i0), \tag{6.12}$$

and define the real matrix-valued Kernel function $\mathscr{K}(\lambda,\mu)$ by

$$\mathscr{K}(\lambda,\mu)\equiv\mathscr{M}(\lambda,\mu)[\mathscr{I}-i\pi\mathscr{M}(\mu)]^{-1}. \tag{6.13}$$

Equation (6.8) may then be written in the following $\mathscr{K}$-matrix form:

$$\mathscr{K}(\lambda,\mu)=\mathscr{B}(\lambda,\mu)+\mathcal{f}\int_0^\infty d\eta[\mathscr{B}(\lambda,\eta)\mathscr{R}(\eta,E) - \mathscr{C}(\lambda,\eta)]\mathscr{K}(\eta,\mu). \tag{6.14}$$

The matrix elements $\mathscr{K}_{\beta j,\alpha i}(\lambda,\mu)$ thus satisfy the system of equations

$$\mathscr{K}_{\beta j,\alpha i}(\lambda,\mu) = \mathscr{B}_{\beta j,\alpha i}(\lambda,\mu)+\sum_C\sum_{\gamma,k\in C}\mathcal{f}\int_0^\infty d\eta\left[\frac{\mathscr{B}_{\beta j,\gamma k}(\lambda,\eta)}{\mu+\epsilon_\alpha-\epsilon_\gamma-\eta} - \mathscr{C}_{\beta j,\gamma k}(\lambda,\eta)\right]\mathscr{K}_{\gamma k,\alpha i}(\eta,\mu), \tag{6.15}$$

where $\alpha,i\in A$ and $\beta,j\in B$.

The system of equations in Eq. (6.15) is to be solved for the matrix elements $\mathscr{K}_{\beta j,\alpha i}(\lambda,\mu)$. The half-on-shell solution matrix $\mathscr{M}(\lambda,\mu)$ is then recovered from the matrix $\mathscr{K}(\lambda,\mu)$ via the identity

$$\mathscr{M}(\lambda,\mu)=\mathscr{K}(\lambda,\mu)[\mathscr{I}+i\pi\mathscr{K}(\mu)]^{-1}, \tag{6.16}$$

where $\mathscr{K}(\mu)$ is the on-shell kernel matrix with elements

$$\mathscr{K}_{\beta j,\gamma k}(\mu)\equiv\mathscr{K}_{\beta j,\gamma k}(\mu+\epsilon_\alpha-\epsilon_\beta,\mu) = \mathscr{K}_{\beta j,\gamma k}(E-\epsilon_\beta,E-\epsilon_\alpha). \tag{6.17}$$

Remark 6.1: We remark that the operators $\mathscr{B}$, $\mathscr{C}$, $\mathscr{M}(z)$, and $\mathscr{R}(z)$, defined in Eqs. (6.1)–(6.4), could be defined in alternative ways. In particular, these operators could have been defined to have an additional factor $(\rho\rho^*)^{-1}$ on their right-hand sides. This would cause the right-hand side of Eq. (6.8) to have an additional factor of ω at the left of $\mathscr{M}(\eta,\mu,z)$. In order to keep the kernel of the integral equation as simple as possible, we have not included these additional factors $(\rho\rho^*)^{-1}$. An interesting third alternative is to define the operators in Eqs. (6.1)–(6.4) with the factor $(\rho\rho^*)^{-1/2}$ on both the left and right sides. Whether or not this is desirable will depend on computational efficiency, and we postpone further discussion of it until more computational experience has been obtained.

In order to solve Eq. (6.14), we must first evaluate the input terms $\mathscr{B}(\lambda,\mu)$ and $\mathscr{C}(\lambda,\mu)$.

VII. EVALUATION OF $\mathscr{B}(\lambda,\mu)$ AND $\mathscr{C}(\lambda,\mu)$

Recall that $\mathscr{B}(\lambda,\mu)$ is the kernel of the operator $\mathscr{B}$ defined in Eq. (6.1), and $\mathscr{C}(\lambda,\mu)$ is the kernel of the operator $\mathscr{C}$ defined in Eq. (6.2). In this section we derive more detailed formulas for these kernel functions that facilitate practical calculations.

Let P_α denote the orthogonal projection operator of $\mathscr{H}_N$ onto $\mathscr{H}_\alpha$. For $\psi_N\in\mathscr{H}_N$, the vector $P_\alpha\psi_N\in\mathscr{H}_\alpha$ is thus given by

$$(P_\alpha\psi_N)(\mathbf{p}_A,\mathbf{q}_A)=\hat\phi_\alpha(\mathbf{p}_A)\int d\mathbf{p}'_A\,\hat\phi^*_\alpha(\mathbf{p}'_A)\psi_N(\mathbf{p}'_A,\mathbf{q}_A). \tag{7.1}$$

The operators J^*V: $\mathscr{H}\to\mathscr{H}$ and (J^*J-I): $\mathscr{H}\to\mathscr{H}$ have expansions

$$J^*V=\bigoplus_B\sum_\beta\sum_{\alpha,A}P_\beta\overline{V}_AP_\alpha \tag{7.2}$$

and

$$J^*J - I = \bigoplus_B \sum_\beta \sum_{\alpha,A} \bar{\delta}_{\beta\alpha} P_\beta P_\alpha, \tag{7.3}$$

respectively, where $\overline{V}_A \equiv H_N - H_A$ and $\bar{\delta}_{\beta\alpha} \equiv 1 - \delta_{\beta\alpha}$ with $\delta_{\beta\alpha}$ the Kronecker delta.

We assume that the bound state wave functions in $\hat{\phi}_\alpha$ and the potentials in $\overline{V}_A$ are known. The first step is to express $P_\beta \overline{V}_A P_\alpha$ and $\bar{\delta}_{\beta\alpha} P_\beta P_\alpha$ in kinetic energy hyperspherical coordinates. Let

$$B_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) \equiv \text{kernel of } P_\beta \overline{V}_A P_\alpha \tag{7.4}$$

and

$$C_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) \equiv \text{kernel of } \bar{\delta}_{\beta\alpha} P_\beta P_\alpha. \tag{7.5}$$

The operators $P_\beta \overline{V}_A P_\alpha$: $\mathscr{H}_\alpha \to \mathscr{H}_\beta$ and $\bar{\delta}_{\beta\alpha} P_\beta P_\alpha$: $\mathscr{H}_\alpha \to \mathscr{H}_\beta$ are, hence, given for $\psi_\alpha \in \mathscr{H}_\alpha$ by

$$\begin{aligned}(P_\beta \overline{V}_A P_\alpha \psi_\alpha)(\mathbf{p}_B,\hat{k}_B,\lambda) &= \int d\mathbf{p}_A\, d\hat{k}_A\, d\mu\, \nu_A^2(\mu) B_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) \\ &\quad \times \psi_\alpha(\mathbf{p}_A,\hat{k}_A,\mu)\end{aligned} \tag{7.6}$$

and

$$\begin{aligned}(\bar{\delta}_{\beta\alpha} P_\beta P_\alpha \psi_\alpha)(\mathbf{p}_B,\hat{k}_B,\lambda) &= \int d\mathbf{p}_A\, d\hat{k}_A\, d\mu\, \nu_A^2(\mu) C_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) \\ &\quad \times \psi_\alpha(\mathbf{p}_A,\hat{k}_A,\mu),\end{aligned} \tag{7.7}$$

respectively, where $\nu_A^2(\mu)$ is the Jacobian of the transformation from the $(\mathbf{p}_A \mathbf{q}_A)$ coordinates to the $(\mathbf{p}_A,\hat{k}_A,\mu)$ coordinates.

The matrix-valued kernel function $\mathscr{B}(\lambda,\mu)$ may, therefore, be expressed by the formula

$$\begin{aligned}\mathscr{B}(\lambda,\mu) = \omega^{-1} \bigoplus_{\beta j,B} \sum_{\alpha,i,A} &\nu_B(\lambda)\nu_A(\mu) \\ &\times \int d\mathbf{p}_B\, d\hat{k}_B\, d\mathbf{p}_A\, d\hat{k}_A \hat{\phi}_\beta^*(\mathbf{p}_B) \chi_{\beta j}^*(\hat{k}_B) \\ &\times B_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) \hat{\phi}_\alpha(\mathbf{p}_A) \chi_{\alpha i}(\hat{k}_A);\end{aligned} \tag{7.8}$$

and $\mathscr{C}(\lambda,\mu)$ may be expressed by the corresponding formula with $B_{\beta\alpha}$ replaced by $\mathscr{C}_{\beta\alpha}$. The following theorem provides alternative useful formulas for $\mathscr{B}(\lambda,\mu)$ and $\mathscr{C}(\lambda,\mu)$.

Theorem 7.1: Suppose that the kernel functions $B_{\beta\alpha}$ and $C_{\beta\alpha}$, defined in Eqs. (7.4) and (7.5), respectively, can be expanded in the basis sets $\{\chi_{\alpha i}(\hat{k}_A)\}$ and $\{\chi_{\beta j}(\hat{k}_B)\}$ as

$$\begin{aligned}B_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) &= \hat{\phi}_\beta(\mathbf{p}_\beta) \left\{ \sum_{ij} \chi_{\beta j}(\hat{k}_B) \widetilde{\mathscr{B}}_{\beta j,\alpha i}(\lambda,\mu) \chi_{\alpha i}^*(\hat{k}_A) \right\} \hat{\phi}_\alpha^*(\mathbf{p}_A)\end{aligned} \tag{7.9}$$

and

$$\begin{aligned}C_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu) &= \hat{\phi}_\beta(\mathbf{p}_\beta) \left\{ \sum_{ij} \chi_{\beta j}(\hat{k}_B) \widetilde{\mathscr{C}}_{\beta j,\alpha i}(\lambda,\mu) \chi_{\alpha i}^*(\hat{k}_A) \right\} \hat{\phi}_\alpha^*(\mathbf{p}_A),\end{aligned} \tag{7.10}$$

thereby obtaining the *Γ_A-sphere to Γ_B-sphere change of basis matrices*,

$$\widetilde{\mathscr{B}}_{BA}(\lambda,\mu) \equiv [\widetilde{\mathscr{B}}_{\beta j,\alpha i}(\lambda,\mu)] \tag{7.11}$$

and

$$\widetilde{\mathscr{C}}_{BA}(\lambda,\mu) \equiv [\widetilde{\mathscr{C}}_{\beta j,\alpha i}(\lambda,\mu)]. \tag{7.12}$$

Then

$$\mathscr{B}(\lambda,\mu) = \bigoplus_B \sum_A \nu_B(\lambda)\nu_A(\mu) \widetilde{\mathscr{B}}_{BA}(\lambda,\mu)\omega_A, \tag{7.13}$$

and

$$\mathscr{C}(\lambda,\mu) = \bigoplus_B \sum_A \nu_B(\lambda)\nu_A(\mu) \widetilde{\mathscr{C}}_{BA}(\lambda,\mu)\omega_A. \tag{7.14}$$

Proof: Let Eq. (7.9), with primes on the subscripts i and j, be substituted into Eq. (7.8). The $d\mathbf{p}_B\, d\hat{k}_B\, d\mathbf{p}_A\, d\hat{k}_A$ integrations may then be evaluated to yield

$$\begin{aligned}\mathscr{B}(\lambda,\mu) &= \omega^{-1} \bigoplus_{\beta j,B} \sum_{\alpha,i,A} \nu_B(\lambda)\nu_A(\mu) \sum_{i',j} \omega_{\beta j,\beta j'} \\ &\quad \times \widetilde{\mathscr{B}}_{\beta j',\alpha i'}(\lambda,\mu) \omega_{\alpha i',\alpha i} \\ &= \bigoplus_B \omega_B^{-1} \sum_A \nu_B(\lambda)\nu_A(\mu)\omega_B \widetilde{\mathscr{B}}_{BA}(\lambda,\mu)\omega_A,\end{aligned} \tag{7.15}$$

which is equivalent to Eq. (7.13). The proof of Eq. (7.14) is similar. □

Remark 7.2: The important consequence of Theorem 7.1 is that complicated integrals of the form of Eq. (7.8) do not have to be computed by numerical integration. They may be evaluated by computing the expansions in Eqs. (7.9) and (7.10) and then using the identities in Eqs. (7.13) and (7.14).

VIII. THE SCATTERING AMPLITUDE

In this section we show how the scattering amplitude can be calculated from the solution $\mathscr{K}(\mu)$ of Eq. (6.14).

We first solve Eq. (6.3) for $M(z)$ by multiplying on the left-hand side by ρ^* and on the right-hand side by $(\rho\rho^*)^{-1}\rho$. This gives

$$M(z) = \rho^* \mathscr{M}(z)(\rho\rho^*)^{-1}\rho. \tag{8.1}$$

Thus for $\Psi = \oplus_A \psi_A \in \mathscr{D}(H) \subset \mathscr{H}$,

$$\begin{aligned}M(z)\Psi &= \bigoplus_B \sum_A M_{BA}(z)\psi_A \\ &= \bigoplus_B \sum_A \rho_B^* \mathscr{M}_{BA}(z)\omega_A^{-1}\rho_A\psi_A.\end{aligned} \tag{8.2}$$

Assuming that the off-shell operator $M_{BA}(z)$ is an integral operator with kernel $M_{BA}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu;z)$ then

$$\begin{aligned}(M_{BA}(z)\psi_A)(\mathbf{p}_B,\hat{k}_B,\lambda) &= \int d\mathbf{p}_A\, d\hat{k}_A\, d\mu\, \nu_A^2(\mu) M_{BA}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu;z) \\ &\quad \times \psi_A(\mathbf{p}_A,\hat{k}_A,\mu).\end{aligned} \tag{8.3}$$

On the other hand, Eqs. (4.7) and (4.10) yield

$$(\rho_B^* M_{BA}(z)\omega_A^{-1}\rho_A\psi_A)(\mathbf{p}_B,\hat{k}_B,\lambda)$$
$$= \nu_B^{-1}(\lambda)\sum_{\alpha,i,\beta,j}\hat{\phi}_\beta(\mathbf{p}_B)\chi_{\beta j}(\hat{k}_B)$$
$$\times\int d\mu\,\mathcal{N}_{\beta j,\alpha i}(\lambda,\mu,z)\nu_A(\mu)$$
$$\times\int d\mathbf{p}_A\,d\hat{k}_A\,\hat{\phi}_\alpha^*(\mathbf{p}_A)\chi_{\alpha i}^*(\hat{k}_A)\psi_A(\mathbf{p}_A,\hat{k}_A,\mu), \quad (8.4)$$

where $\mathcal{N}_{\beta j,\alpha i}(\lambda,\mu,z)$ are the matrix elements of the kernel of the operator $\mathcal{M}_{BA}(z)\omega_A^{-1}$. By comparing Eqs. (8.3) and (8.4), we conclude that

$$M_{BA}(\mathbf{p}_B,\hat{k}_B,\lambda;\mathbf{p}_A,\hat{k}_A,\mu;z)$$
$$= \nu_B^{-1}(\lambda)\nu_A^{-1}(\mu)\sum_{\alpha,i,\beta,j}\hat{\phi}_\beta(\mathbf{p}_B)\chi_{\beta j}(\hat{k}_B)\mathcal{N}_{\beta j,\alpha i}(\lambda,\mu,z)$$
$$\times\hat{\phi}_\alpha^*(\mathbf{p}_A)\chi_{\alpha i}^*(\hat{k}_A). \quad (8.5)$$

In order to obtain the *exact scattering amplitude* $F\equiv S-I$, we would multiply Eq. (8.1) on the left by $(z-H)J^*JR(z)$ and use Eq. (2.4) to obtain transition operator analogs of Eqs. (8.2)–(8.5) and then invoke Theorem 1 of Ref. 1.

The *approximate scattering amplitude* $F^\pi\equiv S^\pi-\Pi$ can, however, usually be obtained directly from $M^\pi(z)$. [Here, and in the following, a superscript π indicates that the operators are approximate operators acting on the truncated asymptotic space $\mathscr{H}^\pi=\oplus_A\mathscr{H}_A^\pi$ defined in Eq. (5.1).] In particular, suppose that the approximate scattering system $\mathfrak{S}(\Pi)$ satisfies the hypotheses of Theorem 3.9 of Ref. 4. Since the limits in Eq. (6.10) are assumed to exist, the approximate version of Eq. (8.5) can be put on shell by letting $\lambda+\epsilon_\beta=\mu+\epsilon_\alpha$, $z=\mu+\epsilon_\alpha+i\epsilon$ and taking the limit as $\epsilon\to 0^+$. Let

$$\mathcal{N}^\pi(\mu)=[\mathcal{N}^\pi_{\beta j,\alpha i}(\mu)]\equiv\mathcal{M}^\pi(\mu)(\omega^\pi)^{-1}, \quad (8.6)$$

where the on-shell matrix elements $\mathcal{N}^\pi_{\beta j,\alpha i}(\mu)$ are the limits as $\epsilon\to 0^+$ of $\mathcal{N}^\pi_{\beta j,\alpha i}(\mu+\epsilon_\alpha-\epsilon_\beta,\mu,\mu+\epsilon_\alpha+i\epsilon)$. It then follows from Ref. 4, Theorem 3.9 and Eq. (8.5) that the kernel of F^π is given by

$$\text{kernel of } F^\pi=\bigoplus_B\sum_A\sum_{\substack{\alpha\in A(\Pi)\\ \beta\in B(\Pi)}}\delta(\lambda+\epsilon_\beta-\mu-\epsilon_\alpha)$$
$$\times F^\pi_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\mu;\mathbf{p}_A,\hat{k}_A,\mu), \quad (8.7)$$

where

$$F^\pi_{\beta\alpha}(\mathbf{p}_B,\hat{k}_B,\mu;\mathbf{p}_A,\hat{k}_A,\mu)$$
$$= -2\pi i\nu_B^{-1}(\mu)\nu_A^{-1}(\mu)\sum_{\substack{i\in A(\Pi)\\ j\in B(\Pi)}}\hat{\phi}_\beta(\mathbf{p}_B)\chi_{\beta j}(\hat{k}_B)\mathcal{N}^\pi_{\beta j,\alpha i}(\mu)$$
$$\times\hat{\phi}_\alpha^*(\mathbf{p}_A)\chi_{\alpha i}^*(\hat{k}_A), \quad (8.8)$$

are the kernels of the approximate scattering amplitudes for scattering from partition A to partition B.

The operator $\mathcal{N}^\pi(\mu)$, defined in Eq. (8.6), can be calculated by the following procedure. Multiply the approximate version of Eq. (6.16) on the right-hand side by $(\omega^\pi)^{-1}$ and let $\lambda=\mu+\epsilon_\alpha-\epsilon_\beta$. This gives

$$\mathcal{N}^\pi(\mu)=\mathcal{K}^\pi(\mu)[\mathcal{I}^\pi+i\pi\mathcal{K}^\pi(\mu)]^{-1}(\omega^\pi)^{-1}$$
$$=\mathcal{K}^\pi(\mu)\{\omega^\pi[\mathcal{I}^\pi+i\pi\mathcal{K}^\pi(\mu)]\}^{-1}. \quad (8.9)$$

Consequently, the matrix $\mathcal{N}^\pi(\mu)$ can be efficiently computed from the solution matrix $\mathcal{K}^\pi(\mu)$ of the approximate version of Eq. (6.14) by solving the following equation for $\mathcal{N}^\pi(\mu)$:

$$\mathcal{N}^\pi(\mu)\{\omega^\pi[\mathcal{I}^\pi+i\pi\mathcal{K}^\pi(\mu)]\}=\mathcal{K}^\pi(\mu). \quad (8.10)$$

Remark 8.1: Theorem 3.9 of Ref. 4 assumes that $J^{\pi*}J^\pi-\Pi$ is a compact operator and that $VE(\Delta)\Pi$ is a trace class operator for finite intervals Δ. The lengthy proof that these assumptions are satisfied will be contained in a subsequent paper (cf. Ref. 15, Theorems 13 and 14).

IX. SUMMARY OF THE COMPUTATIONAL METHOD

The computation strategy developed in this paper for calculating an approximation to the scattering amplitude is summarized in the following five step procedure.

(1) For each partition A, select the n_α bound states $\alpha\in A$ to be used in the approximation. The bound state wave functions and potentials are usually given in either position or momentum coordinates. The first step is to express the kernels $B_{\beta\alpha}$ and $C_{\beta\alpha}$ of the channel matrix elements of the "potential" J^*V and the "overlap" term J^*J-I in kinetic energy hyperspherical coordinates [cf. Eqs. (7.1)–(7.7)].

(2) For each bound state α choose some finite set of $n_{\alpha i}$ linearly independent basis functions $\{\chi_{\alpha i}(\hat{k}_A)\}$ on the surface Γ_A of the unit kinetic energy hypersphere. Let $A(\Pi)$ denote the set of $n_\alpha n_{\alpha i}$ indices $\alpha,i\in A$ used in the approximation. Compute the inner products $\omega^\pi_{\alpha i',\alpha i}$ in Eq. (4.16) of the $\chi_{\alpha i}$ basis functions and thus obtain the block diagonal matrices ω_A^π in Eq. (4.17).

(3) Expand the kernel functions $B_{\beta\alpha}$ and $C_{\beta\alpha}$ in the approximate basis sets $\{\chi_{\alpha i}(\hat{k}_A)\}$, $\alpha,i\in A(\Pi)$, and $\{\chi_{\beta j}(\hat{k}_B)\}$, $\beta,j\in B(\Pi)$, to obtain the Γ_A-sphere to Γ_B-sphere change of basis matrices $\mathscr{B}^\pi_{BA}(\lambda,\mu)$ and $\mathscr{C}^\pi_{BA}(\lambda,\mu)$ [cf. Eqs. (7.9)–(7.12)]. Then, using the identities in Eqs. (7.13) and (7.14), compute the matrices $\mathscr{B}^\pi(\lambda,\mu)$ and $\mathscr{C}^\pi(\lambda,\mu)$.

(4) Solve the half-on-shell system of $\mathscr{K}$-matrix equations in Eq. (6.15) to obtain the on-shell approximate $\mathscr{K}$ matrix $\mathscr{K}^\pi(\mu)$.

(5) Solve Eq. (8.10) for the on-shell "transition matrix" $\mathcal{N}^\pi(\mu)$ and then use the identities in Eqs. (8.7) and (8.8) to calculate the kernel of the approximate scattering amplitude F^π.

Remark 9.1: (i) For given bound state wave functions and potentials the coordinate transformations in step (1) need to be performed only once.

(ii) Two possible choices of the basis functions $\{\chi_{\alpha i}(\hat{k}_A)\}$ are the following. The first is

$$\chi_{\alpha i}(\hat{k}_A)=Y_l^m(\hat{k}_A), \quad (9.1)$$

the hyperspherical harmonics (Ref. 8), where $i=(l,m)$ is now a double index. In this case

$$\omega_{\alpha i',\alpha i}=\delta_{\alpha i',\alpha i}, \quad (9.2)$$

since the hyperspherical harmonics are orthonormal. These

basis functions are known to work well when dim $\Gamma_A = 2$. However, for dim $\Gamma_A > 2$ the series in Eq. (8.8) may converge very slowly, making it necessary to use a large number of basis functions (see, however, Ref. 17).

A second possible choice of the basis functions $\{\chi_{\alpha i}(\hat{k}_A)\}$ is spline functions defined on the surface Γ_A of the unit kinetic energy hypersphere. Such functions have been defined for dim $\Gamma_A = 2$ by Wahba.[10] One example is

$$\chi_{\alpha i}(\hat{k}_A) \equiv \frac{1}{2\pi}\left[\ln\left(1+\sqrt{\frac{2}{1-\cos(\hat{k}_A\cdot\hat{k}_{Ai})}}\right)-1\right], \tag{9.3}$$

where $\{\hat{k}_{Ai}\}$ is some set of points on Γ_A. Other examples with more smoothness are given in Ref. 10. These functions are not orthogonal, but they are linearly independent and

$$\omega_{\alpha i',\alpha i} = \frac{1}{4\pi}\sum_{l=1}^{\infty}|\xi_l|^2(2l+1)P_l(\hat{k}_{Ai'}\cdot\hat{k}_{Ai}), \tag{9.4}$$

where ξ_l are known constants and $P_l(z)$ are the Legendre polynomials. Spline functions of this type have been successfully used to interpolate meteorological and geomagnetic data.[9,10] Work is in progress on generalizing these spline functions to dim $\Gamma_A > 2$.

(iii) For given wave functions, potentials, and approximate basis sets $\{\chi_{\alpha i}(\hat{k}_A)\}$, $\alpha, i \in A(\Pi)$, the expansions in step (3) can be calculated and stored for repeated use in steps (4) and (5). These expansions can be done with readily available linear algebra software packages such as the subroutines in LINPACK.

(iv) Computer programs to solve the system of equations in Eq. (6.15) have already been developed and tested.[16]

(v) If $\mathscr{N}^{\pi *}(\mu)$ denotes the complex conjugate transpose of the matrix $\mathscr{N}^{\pi}(\mu)$, then Eq. (8.10) is equivalent to the system of equations,

$$\{[\mathscr{I}^{\pi} - i\pi\mathscr{K}^{\pi T}(\mu)]\omega^{\pi T}\}\mathscr{N}^{\pi *}(\mu) = \mathscr{K}^{\pi T}(\mu), \tag{9.5}$$

where a superscript T denotes the transpose of a matrix. Equation (9.5) can be solved using standard linear algebra programs.

ACKNOWLEDGMENTS

We have benefited greatly from discussions about this work with B. Bertram, F. Coester, Z. Kuruoglu, G. W. Pletsch, I. Sloan, H. Taijeron, and numerous other colleagues.

We also gratefully acknowledge the financial support of Grants No. PHY-8603342 and No. INT-8400053 from the National Science Foundation. C. Chandler thanks the U. S.–German Fulbright Commission, the Minna–James–Heinemann–Stiftung in collaboration with the NATO Senior Scientists Programme, the German Academic Exchange Service (DAAD), and the University of Bonn for support during the beginning phases of this work.

[1] C. Chandler and A. G. Gibson, "N-Body Quantum Scattering Theory in Two Hilbert Spaces. I. The Basic Equations," J. Math Phys. **18**, 2336 (1977).

[2] C. Chandler and A. G. Gibson, "N-Body Quantum Scattering Theory in Two Hilbert Spaces. II. Some Asymptotic Limits," J. Math. Phys. **19**, 1610 (1978).

[3] C. Chandler and A. G. Gibson, "N-Body Quantum Scattering Theory in Two Hilbert Spaces. III. Theory of Approximations," J. Funct. Anal. **52**, 80 (1983).

[4] C. Chandler and A. G. Gibson, "N-Body Quantum Scattering Theory in Two Hilbert Spaces. IV. Approximate Equations," J. Math. Phys. **25**, 1841 (1984).

[5] C. Chandler and A. G. Gibson, "Transition from Time-Dependent to Time-Independent Multichannel Quantum Scattering Theory," J. Math. Phys. **14**, 1328 (1973).

[6] Gy. Bencze and C. Chandler, "On the Treatment of Exchange Effects in Direct Reactions," Phys. Lett. B **154**, 347 (1985).

[7] Gy. Bencze, C. Chandler, A. G. Gibson, and G. W. Pletsch, "Algebraic Problems in Multiparticle Scattering Theory," University of New Mexico preprint, 1987.

[8] C. Müller, "Spherical Harmonics," *Lecture Notes in Mathematics* (Springer, Berlin, 1966), Vol. 17.

[9] L. Shure, R. L. Parker, and G. E. Backus, "Harmonic Splines for Geomagnetic Modelling," Phys. Earth Planet. Inter. **28**, 215 (1982).

[10] G. Wahba, "Spline Interpolation and Smoothing on the Sphere," SIAM J. Sci. Stat. Comput. **2**, 5 (1981).

[11] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol. I. Functional Analysis and Vol. III. Scattering Theory* (Academic, New York, 1972 and 1979).

[12] W. O. Amrein, J. M. Jauch, and K. B. Sinha, *Scattering Theory in Quantum Mechanics* (Benjamin, Reading, MA, 1977).

[13] I. Stakgold, *Green's Functions and Boundary Value Problems* (Wiley, New York, 1979).

[14] H. L. Royden, *Real Analysis* (Macmillan, New York, 1963).

[15] C. Chandler and A. G. Gibson, "A Two-Hilbert-Space Formulation of Multichannel Scattering Theory," in *Mathematical Methods and Applications of Scattering Theory*, edited by J. A. DeSanto, A. W. Sáenz, and W. W. Zachary (Springer, Berlin, 1980), pp. 134–148.

[16] B. Bertram and A. G. Gibson, "Comparison of Solution Methods for Integral Equations with Cauchy Singularities," in *Integral Methods in Science and Engineering*, edited by F. R. Payne, C. C. Corduneanu, A. Haji-Sheikh, and T. Huang (Hemisphere, Washington, DC, 1986), pp. 99–105.

[17] M. Fabre de la Ripelle, H. Fiedeldey, and S. A. Sofianos, "An Integro-Differential Equation for Few- and Many-Body Systems," University of South Africa preprint, 1987.

Gupta–Bleuler quantization of free massless Lagrangian gauge fields of arbitrary helicity: The bosonic case

Noel A. Doughty[a] and Richard A. Arnold[b]
Department of Physics, University of Canterbury, Christchurch 1, New Zealand

(Received 28 June 1988; accepted for publication 19 October 1988)

The Gupta–Bleuler canonical quantization procedure for free massless gauge fields is applied to the Lagrangian potentials of arbitrary integer helicity. The supplementary conditions required to limit the particle states to the physical subspace of positive-definite metric and positive energy are explicitly set out using procedures that are uniform for all spins, bosonic and fermionic. The well-known lower-spin results (spins 1 and 2), where some of the arbitrary spin features are vacuous, arise as special cases.

I. INTRODUCTION

Supersymmetric string field theories[1] are currently viewed as serious candidates for a unified theory of all four interactions in which massless particles corresponding not only to the photon and to the Yang–Mills bosons but also to the graviton appear automatically. The excitations of some string theories[2] correspond to an infinite sequence of particle states of increasing spin. These results have led to renewed interest in higher spin fields and particles.[2,3]

Particles of arbitrary spin were first considered by Dirac[4] in 1936, and then by Fierz and Pauli[5–7] in 1939 and 1940, de Wet[8] in 1940, and Gårding[9] in 1945. The calculations of Fierz and Pauli and of de Wet both confirmed the existence of only two independent solutions for the wave equations of a massless field of arbitrary spin, apparently independently of any consideration of the representation theory of the Poincaré group[10–12] also first presented in 1939. Poincaré irrep fields of a given arbitrary spin may be constructed from Lorentz irreps in a variety of ways.[11–13] The systematic use of Lagrangian potentials can be traced back to the program initated by Fierz and Pauli.[6] They established that certain pathologies in the interaction of fields are avoided by determining the forms of the free fields that are derivable from an action principle and carrying out their coupling in the Lagrangian formulation.

The field equations and source constraints of Lagrangian potentials of increasingly high spin ($s = 2,\frac{3}{2},3,\ldots$) are well known to involve features and notational difficulties not present in the lower-spin cases, and these can seriously hinder an extension of their analysis to arbitrarily high spin. Textbook treatments of arbitrary spin were included in Corson[14] in 1953 and Umezawa[15] in 1956. Fronsdal[16] systematically developed the higher-spin projection operators in 1958. These projection operators were used by Chang[17] in 1967 to explicitly construct nonlocal Lagrangians up to spin $\frac{7}{2}$. These were then made local by the systematic introduction of auxiliary variables. His results agreed with the local field equations of earlier calculations.[6,18,19] Chang also determined the (anti)commutation relations and Green's functions for all spins. Chang's results are incomplete for higher spins; some of the features not considered are the tracelessness of the gauge functions, the differences between the true source constraint and one that is simply divergence-free, the vanishing double trace on bosonic potentials, and the vanishing triple γ-trace on the fermionic potentials. Lurié[20] gave a textbook discussion in 1968 of the Bargmann–Wigner field equations[11,12,21] and the Rarita–Schwinger representation of massive arbitrary spin fields. Schwinger discussed the wave equations, actions, and source constraints of arbitrary spin fields in his 1970 textbook.[22]

Despite the difficulties caused by some of the abovementioned features, the Lagrangian formulation of the massive fields of arbitrary spin was completed by Singh and Hagen[23] in 1974, and the massless Lagrangians were constructed by Fronsdal[24] and Fang and Fronsdal[25] in 1978. These two sets of papers are the culmination for free-field higher spin of the program that started with Fierz and Pauli.[6]

In 1979 Curtright[26] derived the Lagrangians for massless fields of arbitrary integer and half-odd-integer spin together in a pair of spins $(s,s+\frac{1}{2})$ using a simple and brief supersymmetry argument from postulated gauge invariance. Subsequent analyses by Freedman and de Wit[27] in 1979–80 established a systematic formulation for the higher-spin massless equations using a hierarchy of generalized Christoffel symbols with simple gauge properties. The highest of these quantities played the role of a generalized gauge-invariant Weyl (vacuum Riemann) tensor for spin s. In 1983 Oakley[28] described a direct method for obtaining Lagrangians for any helicity using explicitly indexed Weyl ($SL_{2,C}$) spinors.

Burgers[29] and Berends, Burgers, and van Dam[3] extended and exploited the systematics of massless higher spin in 1985. Doughty, Wiltshire, and Collins[30] developed primarily matrix methods of handling most of the spinor indexing of the arbitrary spin field strengths and potentials, and showed that the details of the gauge properties of the Lagrangian formulation of the massless Poincaré irreps of arbitrary spin may be obtained by integration of the almost trivial non-Lagrangian field strengths based on Lorentz irreps of unmixed spin. The gauge transformations, the traceless-

[a] On leave at Faculty of Mathematical Studies and Department of Electronics and Computer Science, University of Southampton, SO9 5NH, England.
[b] Present address: Institute of Astronomy, Madingley Road, Cambridge CB3 OHA, England.

 0022-2488/89/071545-09$02.50

ness of the gauge functions, the potential symmetries, and the vanishing double and triple traces of the higher-spin ($\geqslant\frac{7}{2}$) potentials all follow naturally from the integration process.

The principal reason for setting up higher-spin field equations is in the description of particle interactions. Many of the articles referred to above examine the quantization of the Lagrangian fields constructed, although sometimes only for the massive case. In the massless case the standard canonical quantization procedures encounter difficulties (for spin $\geqslant 1$) due to the gauge freedom in the Lagrangian potentials. Fermi's original solution to this problem in 1932 for the Maxwell field[31,32] involved canonical quantization with loss of manifest covariance. A variety of canonical and noncanonical covariant techniques have been developed to handle this situation, one of the most widespread being the Faddeev–Popov ansatz[33–35] used in the context of the path-integral quantization technique of Feynman and Hibbs.[36] One of the covariant canonical techniques is Gupta–Bleuler indefinite-metric quantization, which is well known for the spin 1 (Maxwell)[37–40] and spin 2 (linearized gravitational)[41] cases.

Arnowitt, Deser, and Misner[42] canonically quantized the massless spin 2 field using the Schwinger action principle in 1959 as the first step in a series of papers on the quantization of general relativity. Quantization of the arbitrary spin case was given by Weinberg[13] in 1964–65 in non-Lagrangian form and developed further by Nelson and Good[43] and Hammer *et al.*[44] in 1968. Deser, Trubatch, and Trubatch[45] discussed the quantized massless spin 2 (linearized gravity) field in 1965. Fronsdal[24] used the Lagrangians he constructed to verify, for arbitrary integer spin s, that the only massless quanta transmitted between sources are of helicity $\pm s$. Oakley[46] applied an axiomatic noncanonical quantization method in 1984 to the arbitrary helicity massless fields described as Weyl spinors.

Quantization based on path-integral techniques is now extensively used with gauge fields,[35] especially in curved spacetime[47] and supergravity[48] where new features such as ghost fields assume considerable importance. Nevertheless, such techniques are equivalent to canonical procedures for the arbitrary spin fields in flat space-time of concern here. The canonical procedures are closely related to the standard techniques for massive fields and the well-known quantization procedures for lower-spin gauge fields.[40] The physical interpretation of the canonical technique is clear and immediate.

We therefore apply here the Gupta–Bleuler method of covariant canonical quantization involving an indefinite-metric[49,50] state space to the arbitrary integer helicity gauge fields. We set out, explicitly and uniformly for all spins, the supplementary conditions on the particle states that take into account the gauge freedom and limit the physical states to the subspace of positive-definite metric and positive energy. We determine the operator combinations that lead to cancellation of pure gauge contributions and the explicit forms of the equal-time and covariant commutation relations in such a way that the well-known lower-spin results (spins 1 and 2) are reproduced as special cases not containing all the features of the arbitrary spin field.

In Sec. II we summarize the arbitary integer spin Lagrangian wave equations, as well as their gauge invariance and covariant partial gauge fixing. We carry out the quantization of the corresponding quantum field operator in Sec. III, and we apply the Gupta–Bleuler technique to isolate the physical states in Sec. IV. In Sec. V we confirm the particle content for nonzero spin s to be the two helicity modes $\pm s$ expected for a massless particle,[11] and we discuss our results in Sec. VI. Our notation and conventions are set out in the Appendix.

The corresponding fermionic results will be presented by Arnold and Doughty.[51] Some of the material is developed in more detail by Arnold,[52] and the principal results of each paper have been reported briefly by Doughty and Arnold.[53]

II. BOSONIC FIELD EQUATIONS AND GAUGE CONDITIONS

The standard free Lagrangian potential of integer spins s with maximal gauge freedom is a completely symmetric rank s Fierz–Pauli tensor $\phi_{\mu_1\cdots\mu_s}$ with zero double trace

$$\phi^{\lambda\rho}{}_{\lambda\rho\mu_5\cdots\mu_s}=0 \quad (\text{spin}\geqslant 4)\,. \tag{1}$$

This potential becomes a massless irrep of the Poincaré group by virtue of satisfying second-order field equations[3,24,27,30] of the form

$$U_{\mu_1\cdots\mu_s}\equiv W_{\mu_1\cdots\mu_s}-\frac{1}{2}\sum_{\mu 2}\eta_{\mu_1\mu_2}W^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}=0\,, \tag{2}$$

where

$$W_{\mu_1\cdots\mu_s}\equiv\Box\,\phi_{\mu_1\cdots\mu_s}+\sum_{\mu 2}\partial_{\mu_1}\partial_{\mu_2}\phi^{\lambda}{}_{\lambda\mu_3\cdots\mu_s} -\partial^{\lambda}\sum_{\mu 1}\partial_{\mu_1}\phi_{\lambda\mu_2\cdots\mu_s}\,. \tag{3}$$

Following de Wit and Freedman,[27] we use $\Sigma_{\mu 1}$ to denote a symmetrized sum of s terms over a set of indices $\{\mu_1,\mu_2\cdots\mu_s\}$ already symmetrized on $\{\mu_2\cdots\mu_s\}$, while $\Sigma_{\mu 2}$ is a symmetrized sum of $\frac{1}{2}s(s+1)$ terms over the independent permutations of the set of indices $\{\mu_1\mu_2,\mu_3\cdots\mu_s\}$ in which the two sets $\{\mu_1\mu_2\}$ and $\{\mu_3\cdots\mu_s\}$ are separately symmetrized. For example, $\Sigma_{\mu 1}\,\partial_{\mu_1}\alpha_{\mu_2\cdots\mu_s}=s\,\partial_{(\mu_1}\alpha_{\mu_2\cdots\mu_s)}$.

The above equations are invariant under a gauge transformation

$$\delta\phi_{\mu_1\cdots\mu_s}=\sum_{\mu 1}\partial_{\mu_1}\alpha_{\mu_2\cdots\mu_s}(x)\,, \tag{4}$$

where the gauge function $\alpha_{\mu_2\cdots\mu_s}(x)$ is a rank-$(s-1)$ completely symmetric tensor of zero trace, $\alpha^{\lambda}{}_{\lambda\mu_4\cdots\mu_s}=0$. The field equations are derivable[27] from the Lagrangian density

$$\mathscr{L}=(-1)^{s+1}\left[\tfrac{1}{2}\phi_{\mu_1\cdots\mu_s}W^{\mu_1\cdots\mu_s} -\tfrac{1}{8}s(s-1)\phi^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}W^{\rho}{}_{\rho}{}^{\mu_3\cdots\mu_s}\right]. \tag{5}$$

A permissible and natural covariant gauge condition is

$$F_{\mu_2\cdots\mu_s}\equiv\partial^{\lambda}\phi_{\lambda\mu_2\cdots\mu_s}-\frac{1}{2}\sum_{\mu 1}\partial_{\mu_2}\phi^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}=0\,, \tag{6}$$

in which gauge the field equations reduce to

$$U_{\mu_1\cdots\mu_s}=\Box\left(\phi_{\mu_1\cdots\mu_s}-\frac{1}{2}\sum_{\mu 2}\eta_{\mu_1\mu_2}\phi^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}\right)=0\,. \tag{7}$$

This gauge corresponds to the Lorentz, de Donder, and harmonic conditions of the spin 1 (Maxwell), spin 2 (linearized gravity), and general relativistic field equations, respectively.[54]

These partially gauge-fixed equations have residual gauge invariance given by Eq. (4) provided the traceless gauge functions are harmonic, $\Box\alpha_{\mu_2\cdots\mu_s}=0$. Corresponding to a familiar procedure for the spin 2 case,[54] we make a field redefinition to a barred potential,

$$\bar{\phi}_{\mu_1\cdots\mu_s}\equiv\phi_{\mu_1\cdots\mu_s}-\frac{1}{2}\sum_{\mu 2}\eta_{\mu_1\mu_2}\phi^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}\quad(\text{for } s\geqslant 2)\,, \tag{8}$$

for which the partially gauge-fixed field equation is the d'Alembertian wave equation,

$$\Box\bar{\phi}_{\mu_1\cdots\mu_s}=0\,. \tag{9}$$

The original potential is given in terms of the barred potential by

$$\phi_{\mu_1\cdots\mu_s}\equiv\bar{\phi}_{\mu_1\cdots\mu_s}-\frac{1}{2(s-1)}\sum_{\mu 2}\eta_{\mu_1\mu_2}\bar{\phi}^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}\quad(\text{for } s\geqslant 2)\,, \tag{10}$$

and the gauge constraint is expressed as

$$F_{\mu_2\cdots\mu_s}\equiv\partial^{\lambda}\bar{\phi}_{\lambda\mu_2\cdots\mu_s}-\frac{1}{2(s-1)}\partial^{\lambda}\sum_{\mu 2}\eta_{\mu_2\mu_3}\bar{\phi}^{\rho}{}_{\rho\lambda\mu_4\cdots\mu_s}=0\,. \tag{11}$$

The residual gauge transformations are

$$\delta\bar{\phi}_{\mu_1\cdots\mu_s}=\sum_{\mu 1}\partial_{\mu_1}\alpha_{\mu_2\cdots\mu_s}-\partial^{\lambda}\sum_{\mu 2}\eta_{\mu_1\mu_2}\alpha_{\lambda\mu_3\cdots\mu_s}\,, \tag{12}$$

where the gauge functions are traceless and satisfy $\Box\alpha_{\mu_2\cdots\mu_s}=0$.

Although we may express the Lagrangian of Eq. (5) in terms of the barred field variable, it is important to realize that the Lagrangian field is still the unbarred field ϕ. That is to say, the field equations are derived from the Lagrangian by variation with respect to ϕ and not $\bar{\phi}$. However, if we impose the gauge conditions of Eq. (11) on the *Lagrangian*, instead of just on the field equations derivable from it, then the Lagrangian can be written

$$\mathcal{L}=\tfrac{1}{2}(-1)^{s+1}\phi^{\mu_1\cdots\mu_s}\Box\bar{\phi}_{\mu_1\cdots\mu_s}\,, \tag{13}$$

which is equivalent (after discarding some surface terms) to

$$\mathcal{L}=\tfrac{1}{2}(-1)^{s}(\partial^{\lambda}\bar{\phi}^{\mu_1\cdots\mu_s}\partial_{\lambda}\bar{\phi}_{\mu_1\cdots\mu_s}-\tfrac{1}{4}s\,\partial^{\lambda}\bar{\phi}^{\rho}{}_{\rho}{}^{\mu_3\cdots\mu_s}\partial_{\lambda}\bar{\phi}^{\sigma}{}_{\sigma\mu_3\cdots\mu_s})\,. \tag{14}$$

Under such conditions $\bar{\phi}$ does behave as a Lagrangian field. Henceforth we will use $\bar{\phi}_{\mu_1\cdots\mu_s}$ as the Lagrangian field and assume that the gauge conditions are satisfied.

We introduce another field variable $\bar{\chi}$ defined as the trace of the barred field,

$$\bar{\chi}_{\mu_3\cdots\mu_s}\equiv\bar{\phi}^{\lambda}{}_{\lambda\mu_3\cdots\mu_s}\quad(\text{for } s\geqslant 2)\,, \tag{15}$$

which will itself be traceless:

$$\bar{\chi}^{\lambda}{}_{\lambda\mu_5\cdots\mu_s}=0\quad(\text{for } s\geqslant 4)\,. \tag{16}$$

If we treat $\bar{\phi}$ and $\bar{\chi}$ as independent Lagrangian fields, Eq. (15) can be regarded as a constraint imposed on $\bar{\phi}$. The benefits of this procedure are immediate. By taking the trace of the gauge conditions of Eq. (11), we find that they split into the two independent sets of conditions

$$\partial^{\lambda}\bar{\phi}_{\lambda\mu_2\cdots\mu_s}=0\quad\text{and}\quad\partial^{\lambda}\bar{\chi}_{\lambda\mu_4\cdots\mu_s}=0\,. \tag{17}$$

The equations of motion likewise split into

$$\Box\bar{\phi}^{\mu_1\cdots\mu_s}=0\quad\text{and}\quad\Box\bar{\chi}^{\mu_3\cdots\mu_s}=0\,, \tag{18}$$

which are derived from the (partially gauge-fixed) Lagrangian

$$\mathcal{L}=\tfrac{1}{2}(-1)^{s}(\partial^{\lambda}\bar{\phi}^{\mu_1\cdots\mu_s}\partial_{\lambda}\bar{\phi}_{\mu_1\cdots\mu_s}-\tfrac{1}{4}s\,\partial^{\lambda}\bar{\chi}^{\mu_3\cdots\mu_s}\partial_{\lambda}\bar{\chi}_{\mu_3\cdots\mu_s})\,. \tag{19}$$

We abbreviate $\mu_1,\cdots,\mu_s$ and $\mu_3,\cdots,\mu_s$ as $\underline{\mu}$ and $\bar{\underline{\mu}}$, respectively, to rewrite the field equations in the form

$$\Box\bar{\phi}_{\underline{\mu}}=0\quad\text{and}\quad\Box\bar{\chi}_{\bar{\underline{\mu}}}=0\,, \tag{20}$$

with Lagrangian density

$$\mathcal{L}=\tfrac{1}{2}(-1)^{s}(\partial^{\lambda}\bar{\phi}^{\underline{\mu}}\partial_{\lambda}\bar{\phi}_{\underline{\mu}}-\tfrac{1}{4}s\,\partial^{\lambda}\bar{\chi}^{\bar{\underline{\mu}}}\partial_{\lambda}\bar{\chi}_{\bar{\underline{\mu}}})\,. \tag{21}$$

When functional derivatives are taken with respect to symmetric tensors such as $\phi_{\underline{\mu}}$ we must take care to consider the symmetries on the indices. For example, functional differentiation of $h_{\lambda\rho}h^{\lambda\rho}$ with respect to $h_{\mu\nu}$ without taking its symmetry into account gives

$$\frac{\partial}{\partial h_{\mu\nu}}(h_{\lambda\rho}h^{\lambda\rho})=2h^{\mu\nu},\quad\text{for all }\mu\text{ and }\nu\,, \tag{22}$$

while taking the symmetry into account gives

$$\frac{\partial}{\partial h_{\mu\nu}}(h_{\lambda\rho}h^{\lambda\rho})=\begin{cases}2h^{\mu\nu}, & \text{if } \mu=\nu,\\ 4h^{\mu\nu}, & \text{if } \mu\neq\nu,\end{cases} \tag{23}$$

since the off-diagonal terms combine.

Taking the symmetry on the s indices of $\underline{\mu}$ into account leads to

$$\frac{\partial}{\partial\phi_{\underline{\mu}}}(\phi_{\underline{\nu}}\phi^{\underline{\nu}})=2\binom{s}{\underline{\mu}}\phi^{\underline{\mu}}\,. \tag{24}$$

The quantity $\binom{s}{\underline{\mu}}$ is the number of different ways s objects of four different kinds can be ordered (with $n^0_{\underline{\mu}}+n^1_{\underline{\mu}}+n^2_{\underline{\mu}}+n^3_{\underline{\mu}}=s$) and is given by

$$\binom{s}{\underline{\mu}}\equiv\frac{s!}{n^0_{\underline{\mu}}!\,n^1_{\underline{\mu}}!\,n^2_{\underline{\mu}}!\,n^3_{\underline{\mu}}!}\,. \tag{25}$$

In his paper on the quantization of the Fierz–Pauli field, Gupta[41] adopts the convention that symmetries *shall be respected* whenever derivatives are taken. We shall adopt the opposite convention that whenever functional derivatives are taken with respect to symmetric tensors, the symmetries on the indices *will be ignored*. The final results are unaffected, but we shall then be able to avoid the introduction of noncovariant quantities such as $\binom{s}{\underline{\mu}}$ into the commutators.

III. QUANTIZATION

We expand the quantum field operators in the usual way in terms of classical plane-wave solutions to Eqs. (20) to give

$$\bar{\phi}_{\underline{\mu}}(x) = \int d\tilde{k} \sum_{\underline{\lambda}} \epsilon_{\mu_1}^{(\lambda_1)}(\mathbf{k}) \cdots \epsilon_{\mu_s}^{(\lambda_s)}(\mathbf{k}) \times [a^{(\underline{\lambda})}(\mathbf{k}) e^{-ik\cdot x} + a^{(\underline{\lambda})\dagger}(\mathbf{k}) e^{ik\cdot x}] , \tag{26}$$

and

$$\bar{\chi}_{\underline{\bar{\mu}}}(x) = \frac{2}{\sqrt{s}} \int d\tilde{k} \sum_{\underline{\bar{\lambda}}} \epsilon_{\mu_3}^{(\lambda_3)}(\mathbf{k}) \cdots \epsilon_{\mu_s}^{(\lambda_s)}(\mathbf{k}) \times [b^{(\underline{\bar{\lambda}})}(\mathbf{k}) e^{-ik\cdot x} + b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k}) e^{ik\cdot x}] , \tag{27}$$

where $d\tilde{k}$ is the Lorent-invariant bosonic phase space element,[40] and $\{\epsilon_{\mu}^{(\lambda)}\}$ is the spin 1 polarization basis (see the Appendix). The coefficients $a^{(\underline{\lambda})}(\mathbf{k})$ and $b^{(\underline{\bar{\lambda}})}(\mathbf{k})$ of the positive-frequency terms will turn out to be the annihilation operators of two types of spin s quanta. We denote Hermitian conjugation by † and the $2/\sqrt{s}$ factor in the expansion of $\bar{\chi}$ is a convenient normalization that will simplify later equations.

We canonically quantize by imposing the equal-time commutators

$$[\bar{\phi}_{\underline{\mu}}(t,\mathbf{x}), \pi_{\underline{\nu}}^{(\bar{\phi})}(t,\mathbf{y})] = \eta_{\underline{\mu\nu}} i\delta^3(\mathbf{x}-\mathbf{y}) , \tag{28}$$

$$[\bar{\chi}_{\underline{\bar{\mu}}}(t,\mathbf{x}), \pi_{\underline{\bar{\nu}}}^{(\bar{\chi})}(t,\mathbf{y})] = \eta_{\underline{\bar{\mu}\bar{\nu}}} i\delta^3(\mathbf{x}-\mathbf{y}) . \tag{29}$$

All other commutators are zero, while $\eta_{\underline{\mu\nu}}$ and $\eta_{\underline{\bar{\mu}\bar{\nu}}}$ are covariant rank-s and rank-$(s-2)$ Minkowski tensors defined in the Appendix. One may presume that we should completely fix the gauge prior to quantization as the best way to ensure no nonphysical contributions. However, this has the disadvantage of not being a covariant process. Although the use of equal-time commutators already makes the procedure noncovariant, this is not a problem since the covariant commutators follow directly. We could also avoid the equal-time commutators entirely and use the covariant commutators from the start. To maintain the maximum degree of covariance we have used only partial gauge-fixing by the covariant condition (6).

The momenta conjugate to $\bar{\phi}$ and $\bar{\chi}$ are

$$\pi_{\underline{\mu}}^{(\bar{\phi})}(x) = (-1)^s \bar{\phi}_{\underline{\mu}}(x) \quad \text{and} \quad \pi_{\underline{\bar{\mu}}}^{(\bar{\chi})}(x) = (-1)^{s+1}(s/4)\bar{\chi}_{\underline{\bar{\mu}}}(x) . \tag{30}$$

The equal-time commutator may be reexpressed as

$$[\bar{\phi}_{\underline{\mu}}(t,\mathbf{x}), \dot{\bar{\phi}}_{\nu}(t,\mathbf{y})] = (-1)^s \eta_{\underline{\mu\nu}} i\delta^3(\mathbf{x}-\mathbf{y}) , \tag{31}$$

$$[\bar{\chi}_{\underline{\bar{\mu}}}(t,\mathbf{x}), \dot{\bar{\chi}}_{\bar{\nu}}(t,\mathbf{y})] = (-1)^{s+1}(4/s)\eta_{\underline{\bar{\mu}\bar{\nu}}} i\delta^3(\mathbf{x}-\mathbf{y}) , \tag{32}$$

and the coefficients $a^{(\underline{\lambda})}(\mathbf{k})$ and $b^{(\underline{\bar{\lambda}})}(\mathbf{k})$ therefore satisfy

$$[a^{(\underline{\lambda})}(\mathbf{k}), a^{(\underline{\lambda}')\dagger}(k')] = (-1)^s \eta^{(\underline{\lambda\lambda}')} 2k_0 (2\pi)^3 \delta^3(\mathbf{k}-\mathbf{k}') , \tag{33}$$

$$[b^{(\underline{\bar{\lambda}})}(\mathbf{k}), b^{(\underline{\bar{\lambda}}')\dagger}(\mathbf{k}')] = (-1)^{s+1} \eta^{(\underline{\bar{\lambda}\bar{\lambda}}')} 2k_0 (2\pi)^3 \delta^3(\mathbf{k}-\mathbf{k}') , \tag{34}$$

with all other commutators being zero. We can now compute the covariant commutators to obtain

$$[\bar{\phi}_{\underline{\mu}}(x), \bar{\phi}_{\underline{\nu}}(y)] = (-1)^s \eta_{\underline{\mu\nu}} i\Delta(x-y) , \tag{35}$$

$$[\bar{\chi}_{\underline{\bar{\mu}}}(x), \bar{\chi}_{\underline{\bar{\nu}}}(y)] = (-1)^{s+1}(4/s)\eta_{\underline{\bar{\mu}\bar{\nu}}} i\Delta(x-y) , \tag{36}$$

where $\Delta(x-y)$ is the Schwinger Δ-function[40] defined by

$$i\Delta(x) = \int d\tilde{k}(e^{-ik\cdot x} - e^{ik\cdot x}) .$$

We may use the unsymmetrized (nonangular-momentum-conserving) canonical energy-momentum tensor

$$\breve{T}_{\lambda\rho} = (-1)^s (\partial_\lambda \bar{\phi}^{\underline{\mu}} \partial_\rho \bar{\phi}_{\underline{\mu}} - \tfrac{1}{4} s \partial_\lambda \bar{\chi}^{\underline{\bar{\mu}}} \partial_\rho \bar{\chi}_{\underline{\bar{\mu}}} - \tfrac{1}{2} \eta_{\lambda\rho} \partial^\sigma \bar{\phi}^{\underline{\mu}} \partial_\sigma \bar{\phi}_{\underline{\mu}} + \tfrac{1}{8} s \eta_{\lambda\rho} \partial^\sigma \bar{\chi}^{\underline{\bar{\mu}}} \partial_\sigma \bar{\chi}_{\underline{\bar{\mu}}}) ,$$

derived from the partially gauge-constrained Lagrangian of Eq. (21), to determine the four-momentum to be

$$P^\mu = (-1)^s \int d\tilde{k}\, k^\mu \times \Big(\sum_{\underline{\lambda\lambda}'} \eta^{(\underline{\lambda\lambda}')} a^{(\underline{\lambda})\dagger} a^{(\underline{\lambda})}(\mathbf{k}) - \sum_{\underline{\bar{\lambda}\bar{\lambda}}'} \eta^{(\underline{\bar{\lambda}\bar{\lambda}}')} b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k}) b^{(\underline{\bar{\lambda}})}(\mathbf{k}) \Big) . \tag{37}$$

This four-momentum has been normal ordered by the boson prescription to ensure that the ground state $|0\rangle$ has finite energy. That is, all creation operators are commuted to stand to the left of annihilation operators. The factors of $(-)^s \eta^{(\underline{\lambda\lambda}')}$ and $(-)^s \eta^{(\underline{\bar{\lambda}\bar{\lambda}}')}$ show, however, that in its present form the energy operator is not positive-definite.

The commutators of the four-momentum with the a and b operators are as follows:

$$[P_\mu, a^{(\underline{\lambda})}(\mathbf{k})] = -k_\mu a^{(\underline{\lambda})}(\mathbf{k}) ,$$

$$[P_\mu, a^{(\underline{\lambda})\dagger}(\mathbf{k})] = +k_\mu a^{(\underline{\lambda})\dagger}(\mathbf{k}) ,$$

$$[P_\mu, b^{(\underline{\bar{\lambda}})}(\mathbf{k})] = -k_\mu b^{(\underline{\bar{\lambda}})}(\mathbf{k}) ,$$

$$[P_\mu, b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k})] = +k_\mu b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k}) .$$

These commutators imply that the operators $a^{(\underline{\lambda})}(\mathbf{k})$ and $b^{(\underline{\bar{\lambda}})}(\mathbf{k})$ *annihilate* and the operators $a^{(\underline{\lambda})\dagger}(\mathbf{k})$ and $b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k})$ *create* null four-momentum $k = \{k^\mu\} = \{|\mathbf{k}|, \mathbf{k}\}$. Each operator $a^{(\underline{\lambda})}(\mathbf{k})$ and $b^{(\underline{\bar{\lambda}})}(\mathbf{k})$ with a distinct sets of indices (bearing in mind that the operators are symmetric on their sets of indices) creates or annihilates a distinct type of quantum.

The creation operators do not create normalizable states. Instead we construct wave packets with the operators

$$a_f^{(\underline{\lambda})\dagger} \equiv \int d\tilde{k} f(\mathbf{k}) a^{(\underline{\lambda})\dagger}(\mathbf{k})$$

and

$$b_f^{(\underline{\bar{\lambda}})\dagger} \equiv \int d\tilde{k} f(\mathbf{k}) b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k}) , \tag{38}$$

where $\int d\tilde{k}\, |f(\mathbf{k})|^2 = 1$. These wave-packet operators satisfy the commutators

$$[a_f^{(\underline{\lambda})}, a_g^{(\underline{\lambda}')\dagger}] \approx (-1)^s \eta^{(\underline{\lambda\lambda}')} \delta_{fg}$$

and

$$[b_f^{(\bar{\lambda})},b_g^{(\bar{\lambda}')\dagger}]\approx(-1)^{s+1}\eta^{(\bar{\lambda}\bar{\lambda}')}\delta_{fg}\,, \tag{39}$$

all other commutators being zero. The nature of these commutators here requires some explanation. The precise value of the first commutator is

$$[a_f,a_g^\dagger]=\int d\tilde{k}\,f(\mathbf{k})g^*(\mathbf{k})\,.$$

When the peaks of $f(\mathbf{k})$ and $g(\mathbf{k})$ are well separated in momentum space, then the integral $\int d\tilde{k}\,f(\mathbf{k})g^*(\mathbf{k})$ is zero. If the peaks are coincident, then the integral is unity, and if the peaks are close to each other, the integral has a value between 0 and 1. We will assume that whenever we need to use these commutators, the wave packets we use will either be identical or well separated. This assumption is embodied in the use of the delta function δ_{fg}, which is zero if $f(\mathbf{k})$ and $g(\mathbf{k})$ are well separated and 1 if they coincide.

The propagator for the quanta of the boson fields $\bar{\phi}_{\underline{\mu}}$ and $\bar{\chi}_{\underline{\bar{\mu}}}$ are Green's functions for the field equations (20). Evaluation of the time-ordered products

$$\mathbf{T}\bar{\phi}_{\underline{\mu}}(x)\bar{\phi}_{\underline{\nu}}(y)\quad\text{and}\quad\mathbf{T}\bar{\chi}_{\underline{\bar{\mu}}}(x)\bar{\chi}_{\underline{\bar{\nu}}}(y)$$

give the vacuum expectation values of

$$\langle 0|\mathbf{T}\bar{\phi}_{\underline{\mu}}(x)\bar{\phi}_{\underline{\nu}}(y)|0\rangle=(-1)^s\eta_{\underline{\mu\nu}}i\Delta_F(x-y) \tag{40}$$

and

$$\langle 0|\mathbf{T}\bar{\chi}_{\underline{\bar{\mu}}}(x)\bar{\chi}_{\underline{\bar{\nu}}}(y)|0\rangle=(-1)^s(s/4)\eta_{\underline{\bar{\mu}\bar{\nu}}}i\Delta_F(x-y)\,, \tag{41}$$

where $\Delta_F(x-y)$ is the Feynmann propagator.[40] This implies that

$$G^{(\bar{\phi})}_{\underline{\mu\nu}}(x-y)=i(-1)^s\langle 0|\mathbf{T}\bar{\phi}_{\underline{\mu}}(x)\bar{\phi}_{\underline{\nu}}(y)|0\rangle\,, \tag{42}$$

$$G^{(\bar{\chi})}_{\underline{\bar{\mu}\bar{\nu}}}(x-y)=i(-1)^s\langle 0|\mathbf{T}\bar{\chi}_{\underline{\bar{\mu}}}(x)\bar{\chi}_{\underline{\bar{\nu}}}(y)|0\rangle\,, \tag{43}$$

are Green's functions for the respective field equations.

We build up the Fock multiparticle state space from the vacuum state $|0\rangle$, which we normalize, $\langle 0|0\rangle=1$, demand that it have zero four-momentum and spin, $P^\mu|0\rangle=0$ and $S^{\mu\nu}|0\rangle=0$, and require it to satisfy

$$a^{(\underline{\lambda})}(\mathbf{k})|0\rangle=0\quad\text{and}\quad b^{(\bar{\underline{\lambda}})}(\mathbf{k})|0\rangle=0\,. \tag{44}$$

The creation operators all mutually commute, and they therefore create bosonic states.

We define number operators for the two different classes of quanta, a and b, as well as a total number operator:

$$N_a=(-1)^s\sum_{\underline{\lambda}}\sum_{\underline{\lambda}'}\eta^{(\lambda\lambda')}\int d\tilde{k}\,a^{(\underline{\lambda})\dagger}(\mathbf{k})a^{(\underline{\lambda})}(\mathbf{k})$$

$$=\int d\tilde{k}\,\mathcal{N}_a(\mathbf{k})\,, \tag{45}$$

$$N_b=(-1)^{s+1}\sum_{\bar{\underline{\lambda}}}\sum_{\bar{\underline{\lambda}}'}\eta^{(\bar{\lambda}\bar{\lambda}')}\int d\tilde{k}\,b^{(\bar{\underline{\lambda}})\dagger}(\mathbf{k})b_{(\bar{\underline{\lambda}})}(\mathbf{k})$$

$$=\int d\tilde{k}\,\mathcal{N}_b(\mathbf{k})\,, \tag{46}$$

$$N=N_a+N_b=\int d\tilde{k}\,\mathcal{N}(\mathbf{k})=\int d\tilde{k}\,(\mathcal{N}_a(\mathbf{k})+\mathcal{N}_b(\mathbf{k}))\,. \tag{47}$$

These number operators commute with the Hamiltonian. The number of quanta $n=\langle\psi|N|\psi\rangle$ is therefore conserved in time and well-defined.

The state space has an indefinite metric since the one-particle state $a_f^{(\underline{\lambda})\dagger}|0\rangle$, for example, has norm

$$\langle 0|a_f^{(\underline{\lambda})}a_f^{(\underline{\lambda})}|0\rangle=(-1)^s\eta^{(\underline{\lambda}\underline{\lambda})}\gtrless 0\,.$$

As in the well-known lower-spin cases, we can, however, recover a positive-definite norm for *physical* states, which form a subspace of the full state space, by application of supplementary conditions as we shall show in the next section. When this is done, the creation operators creating physical states will be those for which $(-1)^s\eta^{(\underline{\lambda}\underline{\lambda})}>0$.

IV. GUPTA–BLEULER SUPPLEMENTARY CONDITIONS

In 1950 Dirac[55] discussed the generalization of Hamiltonian dynamics to deal with quantization of constrained systems, and this material was developed further in a text[56] in 1964. For the supplementary conditions in Eqs. (1), (16), and (17) not to interfere with the commutation relations on the fields, we must follow the example of the well-known lower-spin cases and apply them using the Gupta–Bleuler technique.[37–41] We demand that if a state $|\psi\rangle$ is to be a physical state, then the following conditions must hold:

$$(\partial^\lambda\bar{\phi}_{\lambda\mu_2\cdots\mu_s})^+|\psi\rangle=0,\quad\text{for } s\geqslant 1\,, \tag{48}$$

$$(\partial^\lambda\bar{\chi}_{\lambda\mu_4\cdots\mu_s})^+|\psi\rangle=0,\quad\text{for } s\geqslant 3\,, \tag{49}$$

$$(\bar{\phi}^\lambda{}_{\lambda\mu_3\cdots\mu_s}-\bar{\chi}_{\mu_3\cdots\mu_s})^+|\psi\rangle=0,\quad\text{for } s\geqslant 2\,, \tag{50}$$

$$(\bar{\phi}^{\lambda\rho}{}_{\lambda\rho\mu_5\cdots\mu_s})^+|\psi\rangle=0,\quad\text{for } s\geqslant 4\,, \tag{51}$$

$$(\bar{\chi}^\lambda{}_{\lambda\mu_5\cdots\mu_s})^+|\psi\rangle=0,\quad\text{for } s\geqslant 4\,, \tag{52}$$

where the $^+$ superscript denotes the positive frequency part. The states satisfying these conditions will form a subspace of the full state space and will have positive-definite norms.

These conditions respectively imply that operators $a^{(\underline{\lambda})}(\mathbf{k})$ and $b^{(\bar{\underline{\lambda}})}(\mathbf{k})$ satisfy the following conditions:

$$[a^{(0\lambda_2\cdots\lambda_s)}(\mathbf{k})-a^{(3\lambda_2\cdots\lambda_s)}(\mathbf{k})]|\psi\rangle=0,\quad\text{for } s\geqslant 1, \tag{53}$$

$$[b^{(0\lambda_4\cdots\lambda_s)}(\mathbf{k})-b^{(3\lambda_4\cdots\lambda_s)}(\mathbf{k})]|\psi\rangle=0,\quad\text{for } s\geqslant 3, \tag{54}$$

$$[a^{(11\lambda_3\cdots\lambda_s)}(\mathbf{k})+a^{(22\lambda_3\cdots\lambda_s)}(\mathbf{k})+(2/\sqrt{s})b^{(\lambda_3\cdots\lambda_s)}(\mathbf{k})]|\psi\rangle$$
$$=0,\quad\text{for } s\geqslant 2, \tag{55}$$

$$[a^{(1111\lambda_5\cdots\lambda_s)}(\mathbf{k})+2a^{(1122\lambda_5\cdots\lambda_s)}(\mathbf{k})+a^{(2222\lambda_5\cdots\lambda_s)}(\mathbf{k})]|\psi\rangle$$
$$=0,\quad\text{for } s\geqslant 4, \tag{56}$$

$$[b^{(11\lambda_5\cdots\lambda_s)}(\mathbf{k})+b^{(22\lambda_5\cdots\lambda_s)}(\mathbf{k})]|\psi\rangle=0,\quad\text{for } s\geqslant 4. \tag{57}$$

We now compute $\mathcal{N}_a(\mathbf{k})$ by expanding the sums over the different indices λ_i to obtain

$$\langle\psi|\mathcal{N}_a(\mathbf{k})|\psi\rangle$$
$$=(-1)^s\sum_{\underline{\lambda}}\sum_{\underline{\lambda}'}\eta^{(\lambda_1\lambda_1')}\cdots\eta^{(\lambda_s\lambda_s')}$$
$$\times\langle\psi|a^{(\lambda_1\cdots\lambda_s)\dagger}(\mathbf{k})a^{(\lambda_1\cdots\lambda_s)}(\mathbf{k})|\psi\rangle$$
$$=(-1)^s\sum_{\lambda_2\cdots\lambda_s}\sum_{\lambda_2'\cdots\lambda_s'}\eta^{(\lambda_2\lambda_2')}\cdots\eta^{(\lambda_s\lambda_s')}$$
$$\times\langle\psi|a^{(0\lambda_2\cdots\lambda_s)\dagger}(\mathbf{k})a^{(0\lambda_2\cdots\lambda_s)}(\mathbf{k})$$
$$-a^{(1\lambda_2\cdots\lambda_s)\dagger}(\mathbf{k})a^{(1\lambda_2\cdots\lambda_s)}(\mathbf{k})$$

$$- a^{(2\lambda_2\cdots\lambda_s)\dagger}(\mathbf{k})a^{(2\lambda_2\cdots\lambda_s)}(\mathbf{k})$$
$$- a^{(3\lambda_2\cdots\lambda_s)\dagger}(\mathbf{k})a^{(3\lambda_2\cdots\lambda_s)}(\mathbf{k})|\psi\rangle,$$

which, with the use of Eq. (53), yields

$$\langle\psi|\mathcal{N}_a(\mathbf{k})|\psi\rangle$$
$$= (-1)^s \sum_{\lambda_2\cdots\lambda_s} \sum_{\lambda_2'\cdots\lambda_s'} \eta^{(\lambda_2\lambda_2')}\cdots\eta^{(\lambda_s\lambda_s')}(-1)$$
$$\times\langle\psi|a^{(1\lambda_2\cdots\lambda_s)\dagger}(\mathbf{k})a^{(1\lambda_2\cdots\lambda_s)}(\mathbf{k}) + a^{(2\lambda_2\cdots\lambda_s)\dagger}(\mathbf{k})$$
$$\times a^{(2\lambda_2\cdots\lambda_s)}(\mathbf{k})|\psi\rangle.$$

All the sums can be expanded and simplified in this way, and induction yields, for all s,

$$\langle\psi|\mathcal{N}_a(\mathbf{k})|\psi\rangle = \sum_{t=0}^{s}\binom{s}{t}\langle\psi|a^{(t \text{ ones and } (s-t) \text{ twos})\dagger}(\mathbf{k})$$
$$\times a^{(t \text{ ones and } (s-t) \text{ twos})}(\mathbf{k})|\psi\rangle.$$

The norm $\langle\psi|\mathcal{N}_b(\mathbf{k})|\psi\rangle$ may be similarly simplified using Eq. (54) and introducing the notation

$$a^{[t]}(\mathbf{k}) \equiv a^{(t \text{ ones and } (s-t) \text{ twos})}(\mathbf{k}), \quad 0\leqslant t\leqslant s; \tag{58}$$

similarly,

$$b^{[t]}(\mathbf{k}) \equiv b^{(t \text{ ones and } (s-t-2) \text{ twos})}(\mathbf{k}), \quad 0\leqslant t\leqslant s-2, \tag{59}$$

gives

$$\langle\psi|\mathcal{N}_a(\mathbf{k})|\psi\rangle = \sum_{t=0}^{s}\binom{s}{t}\langle\psi|a^{[t]\dagger}(\mathbf{k})a^{[t]}(\mathbf{k})|\psi\rangle, \tag{60}$$

$$\langle\psi|\mathcal{N}_b(\mathbf{k})|\psi\rangle = -\sum_{t=0}^{s-2}\binom{s-2}{t}\langle\psi|b^{[t]\dagger}(\mathbf{k})b^{[t]}(\mathbf{k})|\psi\rangle. \tag{61}$$

The remaining supplementary conditions, Eqs. (55)–(57), become

$$[a^{[t+2]}(\mathbf{k}) + a^{[t]}(\mathbf{k}) + (2/\sqrt{s})b^{[t]}(\mathbf{k})]|\psi\rangle = 0,$$
$$\text{for } 0\leqslant t\leqslant s-2 \text{ and } s\geqslant 2, \tag{62}$$
$$[a^{[t+4]}(\mathbf{k}) + 2a^{[t+2]}(\mathbf{k}) + a^{[t]}(\mathbf{k})]|\psi\rangle = 0,$$
$$\text{for } 0\leqslant t\leqslant s-4 \quad \text{and } s\geqslant 4, \tag{63}$$
$$[b^{[t+2]}(\mathbf{k}) + b^{[t]}(\mathbf{k})]|\psi\rangle = 0,$$
$$\text{for } 0\leqslant t\leqslant s-4 \text{ and } s\geqslant 4. \tag{64}$$

Equation (63), which applies when $s\geqslant 4$, allows any operator $a^{[t]}(\mathbf{k})$ with an even index acting on a physical state $|\psi\rangle$ to be expressed in terms of the two lowest even index operators $a^{[0]}(\mathbf{k})|\psi\rangle$ and $a^{[2]}(\mathbf{k})|\psi\rangle$. Likewise any operator $a^{[t]}(\mathbf{k})$ with an odd index acting on $|\psi\rangle$ can be expressed in terms of $a^{[1]}(\mathbf{k})|\psi\rangle$ and $a^{[3]}(\mathbf{k})|\psi\rangle$. Induction on Eq. (63) gives

$$a^{[t]}(\mathbf{k})|\psi\rangle = (-1)^p([p+1]a^{[t-2p]}(\mathbf{k})$$
$$+ pa^{[t-2p-2]}(\mathbf{k}))|\psi\rangle, \quad \text{for integer } p\geqslant(t/2). \tag{65}$$

From Eq. (65) we deduce the reduction formulas

$$\langle\psi|a^{[2n]\dagger}(\mathbf{k})a^{[2n]}(\mathbf{k})|\psi\rangle$$
$$= \langle\psi|n^2a^{[2]\dagger}(\mathbf{k})a^{[2]}(\mathbf{k})$$
$$+ (n-1)^2a^{[0]\dagger}(\mathbf{k})a^{[0]}(\mathbf{k})$$
$$+ n(n-1)(a^{[0]\dagger}(\mathbf{k})a^{[2]}(\mathbf{k})$$
$$+ a^{[2]\dagger}(\mathbf{k})a^{[0]}(\mathbf{k}))|\psi\rangle, \tag{66}$$

and

$$\langle\psi|a^{[2n+1]\dagger}(\mathbf{k})a^{[2n+1]}(\mathbf{k})|\psi\rangle$$
$$= \langle\psi|n^2a^{[3]\dagger}(\mathbf{k})a^{[3]}(\mathbf{k})$$
$$+ (n-1)^2a^{[1]\dagger}(\mathbf{k})a^{[1]}(\mathbf{k})$$
$$+ n(n-1)(a^{[1]\dagger}(\mathbf{k})a^{[3]}(\mathbf{k})$$
$$+ a^{[3]\dagger}(\mathbf{k})a^{[1]}(\mathbf{k}))|\psi\rangle, \tag{67}$$

where n is a non-negative integer.

Equation (64) may also be used to construct reduction formulas for the b-operators showing that each even-indexed $b^{[t]}(\mathbf{k})$ acting on a physical state $|\psi\rangle$ can be expressed in terms of $b^{[0]}(\mathbf{k})|\psi\rangle$ and each odd-indexed operator in terms of $b^{[1]}(\mathbf{k})|\psi\rangle$.

Induction leads to the two reduction relations for b-operators:

$$b^{[2n]}(\mathbf{k})|\psi\rangle = (-1)^n b^{[0]}(\mathbf{k})|\psi\rangle$$

and

$$b^{[2n+1]}(\mathbf{k})|\psi\rangle = (-1)^n b^{[1]}(\mathbf{k})|\psi\rangle, \tag{68}$$

where n is a non-negative integer.

The application of Eqs. (66)–(68) simplifies the expression for $\langle\psi|\mathcal{N}(\mathbf{k})|\psi\rangle$ to

$$\langle\psi|\mathcal{N}(\mathbf{k})|\psi\rangle = 2^{s-4}\langle\psi|\tfrac{1}{2}s(s+1)a^{[2]\dagger}(\mathbf{k})a^{[2]}(\mathbf{k})$$
$$+ (\tfrac{1}{2}s^2 - \tfrac{7}{2}s + 8)a^{[0]\dagger}(\mathbf{k})a^{[0]}(\mathbf{k})$$
$$+ \tfrac{1}{2}s(s-3)(a^{[0]\dagger}(\mathbf{k})a^{[2]}(\mathbf{k})$$
$$+ a^{[2]\dagger}(\mathbf{k})a^{[0]}(\mathbf{k})) - 2b^{[0]\dagger}(\mathbf{k})b^{[0]}(\mathbf{k})|\psi\rangle$$
$$+ 2^{s-4}\langle\psi|(\tfrac{1}{2}s^2 - \tfrac{3}{2}s + 2)a^{[3]\dagger}(\mathbf{k})a^{[3]}(\mathbf{k})$$
$$+ (\tfrac{1}{2}s^2 - \tfrac{11}{2}s + 18)a^{[1]\dagger}(\mathbf{k})a^{[1]}(\mathbf{k})$$
$$+ (\tfrac{1}{2}s^2 - \tfrac{7}{2}s + 6)(a^{[1]\dagger}(\mathbf{k})a^{[3]}(\mathbf{k})$$
$$+ a^{[3]\dagger}(\mathbf{k})a^{[1]}(\mathbf{k})) - 2b^{[1]\dagger}(\mathbf{k})b^{[1]}(\mathbf{k})|\psi\rangle, \tag{69}$$

for $s\geqslant 3$.

The single remaining supplementary condition in Eq. (62) relates the b-operators to the a-operators when they are both acting on physical states. Applying Eq. (62), for $s\geqslant 3$, to Eq. (69) yields

$$\langle\psi|\mathcal{N}(\mathbf{k})|\psi\rangle = 2^{s-5}\langle\psi|(sa^{[2]\dagger}(\mathbf{k})$$
$$+ [s-4]a^{[0]\dagger}(\mathbf{k}))(sa^{[2]}(\mathbf{k})$$
$$+ [s-4]a^{[0]}(\mathbf{k})) + ([s-2]a^{[3]\dagger}(\mathbf{k})$$
$$+ [s-6]a^{[1]\dagger}(k))([s-2]a^{[3]}(\mathbf{k})$$
$$+ [s-6]a^{[1]}(\mathbf{k}))|\psi\rangle. \tag{70}$$

The foregoing treatment of the arbitrary spin case involves in general all four of the operators $a^{[0]}(\mathbf{k})$, $a^{[1]}(\mathbf{k})$, $a^{[2]}(\mathbf{k})$, and $a^{[3]}(\mathbf{k})$. For spins less than three, not all of these operators are present. For spin 2 we find the following expression:

$$\langle\psi|\mathcal{N}(\mathbf{k})|\psi\rangle$$
$$= \tfrac{1}{2}\langle\psi|(a^{[0]\dagger}(\mathbf{k}) - a^{[2]\dagger}(\mathbf{k}))(a^{[0]}(\mathbf{k}) - a^{[2]}(\mathbf{k}))|\psi\rangle \quad (71)$$
$$+ 2\langle\psi|a^{[1]\dagger}(\mathbf{k})a^{[1]}(\mathbf{k})|\psi\rangle \quad (72)$$

The form of Eq. (70) enables us to define two new operators,

$$\bar{a}_1(\mathbf{k}) \equiv 2^{s/2-3}(sa^{[2]}(\mathbf{k}) + [s-4]a^{[0]}(\mathbf{k}) + i[s-2]a^{[3]}(\mathbf{k}) + i[s-6]a^{[1]}(\mathbf{k})), \quad (73)$$

$$\bar{a}_2(\mathbf{k}) \equiv 2^{s/2-3}(sa^{[2]}(\mathbf{k}) + [s-4]a^{[0]}(\mathbf{k}) - i[s-2]a^{[3]}(\mathbf{k}) - i[s-6]a^{[1]}(\mathbf{k})), \quad (74)$$

for $s \geqslant 3$, while for $s = 2$ we define

$$\bar{a}_1(\mathbf{k}) = a^{[1]}(\mathbf{k}) - \tfrac{1}{2}i(a^{[0]}(\mathbf{k}) - a^{[2]}(\mathbf{k})), \quad (75)$$

$$\bar{a}_2(\mathbf{k}) = a^{[1]}(\mathbf{k}) + \tfrac{1}{2}i(a^{[0]}(\mathbf{k}) - a^{[2]}(\mathbf{k})). \quad (76)$$

In terms of these operators we have

$$\langle\psi|\mathcal{N}(\mathbf{k})|\psi\rangle = \langle\psi|\bar{a}_1^\dagger(\mathbf{k})\bar{a}_1(\mathbf{k}) + \bar{a}_2^\dagger(\mathbf{k})\bar{a}_2(\mathbf{k})|\psi\rangle \quad (s\neq 0) \quad (77)$$

and $\langle\psi|\mathcal{N}(\mathbf{k})|\psi\rangle = \langle\psi|a^{[0]\dagger}(\mathbf{k})a^{[0]}(\mathbf{k})|\psi\rangle$ for $s = 0$. The spin 1 and spin 2 results reproduce those of Gupta.[37,41] Equation (77) demonstrates that for nonzero spin each field has two independent quanta, the operators for which have been explicitly displayed. These quanta correspond to the two degrees of freedom of the massless field, one for each of the two polarization states of a classical transverse wave. We shall shortly show that these states are the pure helicity states $\lambda = \pm s$ of the quantum field.

The use of the reduction formulas of Eqs. (66)–(68), along with the equations relating a- and b-operators in Eq. (62), means that the $\bar{a}_1(\mathbf{k})$ and $\bar{a}_2(\mathbf{k})$ do not involve the complete sets of operators $\{a^{[t]}(\mathbf{k})\}$ and $\{b^{[t]}(\mathbf{k})\}$. In consequence, $a^{[t]}(\mathbf{k})$ and $b^{[t]}(\mathbf{k})$ do not commute in the expected fashion for operators that correspond to independent quanta.

V. SPIN OF THE QUANTA

To verify that the quanta created and annihilated by the operators $\bar{a}_1(\mathbf{k})$ and $\bar{a}_2(\mathbf{k})$ carry pure spin s we consider a wave along the x^3 axis and examine the expectation value of the (12) component of the spin angular momentum generator $S^{\mu\nu}$.

First, we compute the canonical spin angular momentum density

$$\mathring{\mathcal{S}}^{\lambda\rho\sigma} = -i\frac{\partial\mathcal{L}}{\partial(\partial_\sigma\bar{\phi}_{\underline{\mu}})}S^{\lambda\rho}_{(s)\underline{\mu\nu}}\bar{\phi}^{\underline{\nu}} - i\frac{\partial\mathcal{L}}{\partial(\partial_\sigma\bar{\chi}_{\underline{\bar{\mu}}})}S^{\lambda\rho}_{(s-2)\underline{\bar{\mu}\bar{\nu}}}\bar{\chi}^{\underline{\bar{\nu}}}, \quad (78)$$

where $S^{\lambda\rho}_{(s)\underline{\mu\nu}}$ is the infinitesimal spin angular momentum generator for spin s given by

$$S_{(s)}{}^{\lambda\rho}{}_{\mu\nu} = \sum_{\mu 1}\eta_{\mu_1\nu_1}\cdots\sum_{\mu 1}\eta_{\mu_{s-1}\nu_{s-1}}S^{\lambda\rho}_{(1)\mu_s\nu_s}$$
$$= i\sum_{\mu 1}\eta_{\mu_1\nu_1}\cdots\sum_{\mu 1}\eta_{\mu_{s-1}\nu_{s-1}}(\delta^\lambda{}_{\mu_s}\delta^\rho{}_{\nu_s} - \delta^\rho{}_{\mu_s}\delta^\lambda{}_{\nu_s}), \quad (79)$$

while $S^{\lambda\rho}_{(s-2)\underline{\bar{\mu}\bar{\nu}}}$ is the corresponding spin $(s-2)$ generator used for $\bar{\chi}_{\underline{\bar{\mu}}}$ and symmetrized over the appropriate index range. The value of the canonical spin density is then

$$\mathring{\mathcal{S}}^{\lambda\rho\sigma} = 2s(-1)^s(\partial^\sigma\bar{\phi}^{\mu_2\cdots\mu_s[\lambda}\bar{\phi}^{\rho]}{}_{\mu_2\cdots\mu_s} - \tfrac{1}{2}s(s-2)(-1)^s(\partial^\sigma\bar{\chi}^{\mu_4\cdots\mu_s[\lambda}\bar{\chi}^{\rho]}{}_{\mu_4\cdots\mu_s}. \quad (80)$$

The spin angular momentum generator is then $S^{\lambda\rho} = S_1^{\lambda\rho} + S_2^{\lambda\rho}$, where

$$S_1^{\lambda\rho} = \int d^3x : 2s(-1)^s\dot{\bar{\phi}}^{\mu_2\cdots\mu_s[\lambda}\bar{\phi}^{\rho]}{}_{\mu_2\cdots\mu_s} :, \quad (81)$$

$$S_2^{\lambda\rho} = \int d^3x : \tfrac{1}{2}s(s-2)(-1)^{s+1}\dot{\bar{\chi}}^{\mu_4\cdots\mu_s[\lambda}\bar{\chi}^{\rho]}{}_{\mu_4\cdots\mu_s} : . \quad (82)$$

Substituting the plane-wave expansions of $\bar{\phi}$ and $\bar{\chi}$ into Eqs. (81) and (82) yields

$$S_1^{\lambda\rho} = \int d\tilde{k}\,\mathcal{S}_1^{\lambda\rho}(\mathbf{k}) \quad \text{and} \quad S_2^{\lambda\rho} = \int d\tilde{k}\,\mathcal{S}_2^{\lambda\rho}(\mathbf{k}),$$

where

$$\mathring{\mathcal{S}}_{1\lambda\rho}(\mathbf{k}) = -\frac{1}{2}is(-1)^s\sum_{\underline{\lambda}}\sum_{\underline{\lambda}'}2\epsilon^{(\lambda_1)}_{[\lambda}(\mathbf{k})\epsilon^{(\lambda'_1)}_{\rho]}(\mathbf{k})\eta^{(\lambda_2\lambda'_2)}\cdots\eta^{(\lambda_s\lambda'_s)}(a^{(\underline{\lambda}')\dagger}(\mathbf{k})a^{(\underline{\lambda})}(\mathbf{k}) - a^{(\underline{\lambda})\dagger}(\mathbf{k})a^{(\underline{\lambda}')}(\mathbf{k})),$$

$$\mathring{\mathcal{S}}_{2\lambda\rho}(\mathbf{k}) = \frac{1}{2}i(s-2)(-1)^s\sum_{\underline{\bar{\lambda}}}\sum_{\underline{\bar{\lambda}}'}2\epsilon^{(\lambda_3)}_{[\lambda}(\mathbf{k})\epsilon^{(\lambda'_3)}_{\rho]}(\mathbf{k})\eta^{(\lambda_4\lambda'_4)}\cdots\eta^{(\lambda_s\lambda'_s)}(b^{(\underline{\bar{\lambda}}')\dagger}(\mathbf{k})b^{(\underline{\bar{\lambda}})}(\mathbf{k}) - b^{(\underline{\bar{\lambda}})\dagger}(\mathbf{k})b^{(\underline{\bar{\lambda}}')}(\mathbf{k})).$$

For a plane wave along the x^3 axis, we find that the helicity λ of the wave is given by

$$\lambda = -\int d\tilde{k}\,\langle\psi|(\mathring{\mathcal{S}}_1^{12}(\mathbf{k}) + \mathring{\mathcal{S}}_2^{12}(\mathbf{k}))|\psi\rangle.$$

In order to evaluate this expression we need to again invoke the supplementary conditions and reduction formulas of the previous section. When this is done, we find that we can write, for $s \geqslant 3$,

$$\langle\psi|\mathring{\mathcal{S}}^{12}(\mathbf{k})|\psi\rangle = -s\langle\psi|(\bar{a}_1^\dagger(\mathbf{k})\bar{a}_1(\mathbf{k}) - \bar{a}_2^\dagger(\mathbf{k})\bar{a}_2(\mathbf{k}))|\psi\rangle = -s\langle\psi|\mathcal{N}_1(\mathbf{k}) - \mathcal{N}_2(\mathbf{k})|\psi\rangle,$$

or

$$\lambda = -\langle\psi|S^{12}|\psi\rangle = s\langle\psi|(N_1 - N_2)|\psi\rangle = s(n_1 - n_2). \quad (83)$$

This result implies that $\bar{a}_1$ and $\bar{a}_2$ quanta carry helicities $\lambda = s$ and $\lambda = -s$, respectively. This conclusion extends to spins $s = 0$, 1, and 2, although the two helicity states coincide for a spin zero particle.

VI. DISCUSSION

No consistent interactions have been discovered[3,27] in flat space-time for fields of spin higher than 2, although a

gravitational interaction for such particles is presented in Vasil'ev and Fradkin.[57] One could argue, however, that the greater degree of consistency apparent in some supersymmetric string theories, compared with particle field theory, is an indication that consistent interaction may involve an essentially infinite number of particles of increasing spin corresponding to the excitations of the string. Indeed there are some indications[3] that no couplings of higher-spin fields are possible without the participation of an infinite number of such fields.

Calculations involving arbitrary spin fields increase rapidly in complexity with increasing spin, especially if each case is performed separately without taking into account the regularity that must link all the equations owing to their common origin as irreps of the Poincaré group.[11] Furthermore, as the spin increases, new features occur that are vacuous for the lower-spin cases. The first bosonic field to include apparently all the features of the arbitrary spin case, such as the zero double trace of the potentials, is spin 4. If consistent interaction of higher-spin fields does involve an infinite number or effectively very large number of participating fields, it is imperative that the theory be developed in a very systematic and general way.

Once all the new general features not appearing at lower spin are known and taken into account, and provided one uses a systematic notation suited to the arbitrary spin case, most calculations involving noninteracting arbitrary helicity potentials become considerably more tractable both classically and at the quantum level.

We have explicitly set out here one example of such a calculation, the canonical quantization of the arbitrary integer helicity gauge fields. We have used the higher-spin systematics of Freedman and de Wit and the very simple and physically transparent Gupta–Bleuler technique of indefinite metric to maintain covariance. The supplementary conditions imposed on the subspace of physical states provides not only a positive-definite metric for the Fock multiparticle space but also ensure the positivity of the energy required for the quantum stability of the vacuum. A first step in any interacting theory is the availability of a complete analysis of the free fields and a tractable compact notation for manipulating the quantities involved. Our calculations, like a number of others mentioned earlier, show that the high symmetry of the potentials and other highly systematic properties of Poincaré irreps permit free field calculations to be performed at arbitrarily high spin with results that accord in every way with the familiar lower-spin results.

APPENDIX

We use a timelike convention for the Minkowski matrix:

$$\eta = \{\eta^{\mu\nu}\} = \{\mathrm{diag}(1,-1,-1,-1)\}.$$

We define a covariant ranks-s Minkowski metric tensor $\eta_{\underline{\mu\nu}}$ by

$$\eta_{\underline{\mu\nu}} = \frac{1}{s!}\sum_{\underline{\mu}1}\eta_{\mu_1\nu_1}\sum_{\underline{\mu}1}\eta_{\mu_2\nu_2}\cdots\sum_{\underline{\mu}1}\eta_{\mu_s\nu_s} = \begin{cases}(-1)^{s-n^0_{\underline{\mu}}}\binom{s}{\underline{\mu}}^{-1}, & \text{if } \underline{\mu} \text{ is a permutation of } \underline{\nu},\\ 0, & \text{if } \underline{\mu} \text{ is not a permutation of } \underline{\nu},\end{cases} \tag{A1}$$

and a similar quantity $\eta_{\underline{\bar{\mu}\bar{\nu}}}$ of rank $(s-2)$.

These tensors act as raising and lowering operators on whole sets of indices. For example,

$$\eta_{\mu\nu}\phi^{\underline{\nu}} = \frac{1}{s!}\sum_{\underline{\mu}1}\eta_{\mu_1\nu_1}\cdots\sum_{\underline{\mu}1}\eta_{\mu_s\nu_s}\phi^{\nu_1\cdots\nu_s} = \eta_{\mu_1\nu_1}\cdots\eta_{\mu_s\nu_s}\phi^{\nu_1\cdots\nu_s} = \phi_{\mu_1\cdots\mu_s} = \phi_{\underline{\mu}}. \tag{A2}$$

We construct a spin 1 polarization basis of four-vectors $\{\epsilon^{(\lambda)}_{\mu}(\mathbf{k})\}$ $(\lambda = 0,1,2,3)$ as follows.

(i) $\epsilon^{(0)}_{\mu}(\mathbf{k}) \equiv n_{\mu}$ is normalized, timelike $(n^2 = 1)$, and future-pointing $(n_0 > 0)$.

(ii) $\epsilon^{(3)}_{\mu}(\mathbf{k})$ is normalized, spacelike $(\epsilon^{(3)}\cdot\epsilon^{(3)} = -1)$, orthogonal to n_{μ}, and in the plane of k_{μ} and n_{μ}, and is thus given by $\epsilon^{(3)}_{\mu}(\mathbf{k}) = (k_{\mu} - n_{\mu}k\cdot n)/k\cdot n$.

(iii) $\epsilon^{(1)}_{\mu}(\mathbf{k})$ and $\epsilon^{(2)}_{\mu}(\mathbf{k})$ are a pair of orthonormal, spacelike vectors that are orthogonal to k and n [and therefore also to $\epsilon^{(3)}_{\mu}(\mathbf{k})$].

As a result we have the following orthonormality and completeness relations:

$$\epsilon^{(\lambda)}(\mathbf{k})\cdot\epsilon^{(\lambda')}(\mathbf{k}) = \eta^{(\lambda\lambda')}$$

and

$$\sum_{\lambda=0}^{3}\eta^{(\lambda\lambda)}\epsilon^{(\lambda)}_{\mu}(\mathbf{k})\epsilon^{(\lambda)}_{\nu}(\mathbf{k}) = \eta_{\mu\nu}, \tag{A3}$$

where $\eta^{(\lambda\lambda')} = \mathrm{diag}(1,-1,-1,-1)$ in which the round brackets indicate that λ and λ' are *not* space-time indices and are *not* automatically summed over.

If we take the propagation direction as axis 3, we have $k = \{k^{\mu}\} = k^0(1,0,0,1)$, and a suitable polarization basis is $\{\epsilon^{(\lambda)}_{\mu}\} = \{\delta^{\lambda}_{\mu}\}$, namely,

$$\{\epsilon^0_{\mu}\} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \{\epsilon^1_{\mu}\} = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \{\epsilon^2_{\mu}\} = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \quad \{\epsilon^3_{\mu}\} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}. \tag{A4}$$

We can construct a corresponding product basis $\epsilon^{(\lambda_1)}\otimes\epsilon^{(\lambda_2)}\otimes\cdots\otimes\epsilon^{(\lambda_s)}$ for rank-s tensors.

For handling whole sets of polarization indices we similarly define

$$\eta^{(\underline{\lambda}\underline{\lambda}')} = \frac{1}{s!}\sum_{\underline{\lambda}1}\eta^{(\lambda_1\lambda'_1)}\sum_{\underline{\lambda}1}\eta^{(\lambda_2\lambda'_2)}\cdots\sum_{\underline{\lambda}1}\eta^{(\lambda_s\lambda'_s)} = \begin{cases}(-1)^{s-n^0_{\underline{\lambda}}}\binom{s}{\underline{\lambda}}^{-1}, & \text{if } \underline{\lambda} \text{ is a permutation of } \underline{\lambda}',\\ 0, & \text{if } \underline{\lambda} \text{ is not a permutation of } \underline{\lambda},\end{cases} \tag{A5}$$

and similarly for rank $(s-2)$.

[1] M. B. Green, Nature **314**, 409 (1985); M. B. Green and J. H. Schwarz, Phys. Lett. B **149**, 117 (1984); D. J. Gross, J. A. Harvey, E. Martinec, and R. Rohm, Phys. Rev. Lett. **54**, 502 (1985).

[2] G. B. West, Nucl. Phys. B **277**, 102 (1986).

[3]F. A. Berends, G. J. H. Burgers, and H. van Dam, Z. Phys. C. **24**, 247 (1984); Nucl. Phys. B **260**, 295 (1985a); **271**, 429 (1985b).
[4]P. A. M. Dirac, Proc. R. Soc. London Ser. A **155**, 447 (1936).
[5]M. Fierz, Hevl. Phys. Acta **12**, 3 (1939); **13**, 45 (1940).
[6]M. Fierz and W. Pauli, Proc. R. Soc. London Ser. A **173**, 211 (1939).
[7]W. Pauli and M. Fierz, Helv. Phys. Acta **12**, 297 (1939).
[8]J. S. de Wet, Phys. Rev. **58**, 236 (1940).
[9]L. Gårding, Proc. Cambridge Philos. Soc. **41**, 49 (1945).
[10]E. P. Wigner, Ann. Math. **40**, 149 (1939).
[11]V. Bargmann and E. P. Wigner, Proc. Natl. Acad. Sci. USA **34**, 211 (1948).
[12]A. O. Barut and R. Rączka, *Theory of Group Representations and Applications* (Polish Scientific, Warsaw, 1980).
[13]S. Weinberg, Phys. Rev. B **133**, 1318 (1964a); **4**, 882 (1964b); **4**, 1049 (1964c); **138**, 988 (1965).
[14]E. M. Corson, *Introduction to Tensors, Spinors, and Relativistic Wave Equations* (Blackie, Glasgow, 1953).
[15]H. Umezawa, *Quantum Field Theory* (North-Holland, Amsterdam, 1956).
[16]C. Fronsdal, Nuovo Cimento Suppl. **9**, 416 (1958).
[17]S.-J. Chang, Phys. Rev. **161**, 1308 (1967).
[18]W. Rarita and J. Schwinger, Phys. Rev. **60**, 61 (1941).
[19]A. Kawakami and S. Kamefuchi, Nuovo Cimento **48**, 239 (1967).
[20]D. Lurié, *Particles and Fields* (Interscience, New York, 1968).
[21]M. A. Rodriguez and M. Lorente, Nuovo Cimento A **83**, 249 (1984).
[22]J. Schwinger, *Particles, Sources and Fields* (Addison-Wesley, Reading, MA, 1970).
[23]L. P. S. Singh and C. R. Hagen, Phys. Rev. D **9**, 898,910 (1974).
[24]C. Fronsdal, Phys. Rev. D **18**, 3624 (1978).
[25]J. Fang and C. Fronsdal, Phys. Rev. D **18**, 3630 (1978).
[26]T. Curtwright, Phys. Rev. B **85**, 219 (1979).
[27]D. Z. Freedman, "Systematics of higher spin gauge fields," in *Supergravity*, edited by P. van Nieuwenhuizen and D. Z. Freedman (North-Holland, Amsterdam, 1979), pp. 263–268; B. de Wit and D. Z. Freedman, Phys. Rev. D **21**, 358 (1980).
[28]C. G. Oakley, "A direct method for obtaining Lagrangians for any helicity," University of Oxford, Department of Theoretical Physics, preprint 26/83 (private communcation, January 1987).
[29]G. J. H. Burgers, Ph.D. thesis, Rijksuniversiteit te Leiden, 1985.
[30]N. A. Doughty and D. L. Wiltshire, J. Phys. A: Math. Gen. **19**, 3727 (1986); N. A. Doughty and G. P. Collins, J. Math. Phys. **27**, 1639 (1986a); J. Phys. A: Math. Gen. **19**, L887 (1986b); G. P. Collins and N. A. Doughty, J. Math. Phys. **28**, 448 (1987); D. L. Wiltshire, M.Sc. thesis, University of Canterbury, Christchurch, New Zealand, 1983; G. P. Collins, M.Sc. thesis, University of Canterbury, Christchurch, New Zealand, 1985.
[31]E. Fermi, Rev. Mod. Phys. **4**, 87 (1932).
[32]J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).
[33]L. D. Faddeev and V. N. Popov, Phys. Lett. B **25**, 29 (1967).
[34]L. H. Ryder, *Quantum Field Theory* (Cambridge U. P., Cambridge, England, 1985).
[35]D. Bailin and A. Love, *Introduction to Gauge Field Theory* (Hilger, Bristol, 1986).
[36]R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).
[37]S. N. Gupta, Proc. Phys. Soc. London Sect. A **63**, 681 (1950); **64**, 850 (1951).
[38]K. Bleuler, Helv. Phys. Acta **23**, 567 (1950).
[39]N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Wiley, New York, 1980), 3rd ed.; J. M. Jauch and F. Rohrlich, *The Theory of Photons and Electrons* (Addison–Wesley, Reading, MA, 1955); F. Mandl and G. Shaw, *Quantum Field Theory* (Wiley, New York, 1984); S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory* (Harper and Row, New York, 1961).
[40]C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).
[41]S. N. Gupta, Proc. Phys. Soc. London Sect. A **65**, 161 (1952).
[42]R. Arnowitt and S. Deser, Phys. Rev. **113**, 745 (1959).
[43]T. J. Nelson and R. H. Good, Rev. Mod. Phys. **40**, 508 (1968).
[44]C. L. Hammer, S. C. McDonald, and D. L. Pursey, Phys. Rev. **171**, 1349 (1968).
[45]S. Deser, J. Trubatch, and S. Trubatch, Nuovo Cimento **39**, 1159 (1965).
[46]C. G. Oakley, Ph.D. thesis, Trinity College, University of Oxford, 1984.
[47]N. D. Birrell and P. C. W. Davies, *Quantum Fields in Curved Space* (Cambridge U. P., Cambridge, England, 1982).
[48]A. Das and D. W. Freedman, Nucl. Phys. B **114**, 271 (1976).
[49]P. A. M. Dirac, Proc. R. Soc. London Ser. A **180**, 1 (1942); Comm. Dublin Institute for Advanced Studies A No. 1 (1943); *Principles of Quantum Mechanics* (Oxford U. P., Oxford, 1947).
[50]W. Pauli, Rev. Mod. Phys. **15**, 175 (1943).
[51]R. A. Arnold and N. A. Doughty, "Gupta–Bleuler quantization of massless free Lagrangian gauge fields of arbitrary helicity: The fermionic case" (unpublished).
[52]R. A. Arnold, M.Sc. thesis, University of Canterbury, Christchurch, New Zealand, 1987.
[53]N. A. Doughty and R. A. Arnold, "Quantization of free massless Lagrangian gauge fields of arbitrary helicity," in *Proceedings of the International Symposium on Spacetime Symmetries*, University of Maryland, 24–28 May, edited by Y. S. Kim and W. W. Zachary (Elsevier, Amsterdam, 1988).
[54]C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[55]P. A. M. Dirac, Can. J. Math. **2**, 129 (1950).
[56]P. A. M. Dirac, *Generalized Hamiltonian Dynamics* (Belfer Graduate School of Science, New York, 1964).
[57]M. A. Vasil'ev and E. S. Fradkin, Sov. Phys. JETP Lett. **44**, 622 (1986).

Equivalence of generic equation method and the phenomenological model for linear transport problems in a two-state random scattering medium

D. C. Sahni
Theoretical Physics Division, Bhabha Atomic Research Centre, Bombay, India

(Received 3 November 1988; accepted for publication 15 February 1989)

Linear particle transport problems in a two-state random Markov medium are considered. Explicit equations are constructed that determine the expected particle density in a scattering medium. It is shown that the generic equation method is equivalent to the phenomenological model if the free flights of the particle are uncorrelated.

I. INTRODUCTION

The problem of particle transport in a two-state, static random (Markov) medium has been the subject of some recent publication.[1–9] Two different methods have been suggested. The generic equation approach[1,2] takes the linear transport equation

$$\frac{\partial\psi}{\partial t} + v\cdot\nabla_r\psi + v\sigma_t(r,v,t)\psi(r,v,t) = Q(r,v,t) + \int v'\sigma_{pr}(r,v'\to v,t)\psi(r,v',t)dv' \qquad (1)$$

as the starting point. Here the total macroscopic cross section σ_t, the production cross section σ_{pr} due to scattering and/or fission, and the particle source Q are known only in a statistical sense. Thus at each space point r and instant of time t σ_t, σ_{pr} assume one of the two values corresponding to the associated fluid A or B. One is then interested in the ensemble average $\bar{\psi}$ of the particle density 'ψ'. The problem was rigorously solved only for a purely capturing, homogeneous Markov medium.

In the second method, termed as the phenomenological model,[4,8] one considers two coupled transport like equations

$$\begin{bmatrix} \{\partial/\partial t + v\cdot\nabla + v\sigma_{tA} + v\lambda_A\} & -v\lambda_B \\ -v\lambda_A & \{\partial/\partial t + v\cdot\nabla + v\sigma_{tB} + v\lambda_B\} \end{bmatrix} \begin{bmatrix} \psi_A(r,v,t) \\ \psi_B(r,v,t) \end{bmatrix} = \begin{bmatrix} \int v'\sigma_{prA}(r,v'\to v,t)dv' & 0 \\ 0 & \int v'\sigma_{prB}(r,v'\to v,t)dv' \end{bmatrix} \begin{bmatrix} \psi_A(r,v',t) \\ \psi_B(r,v',t) \end{bmatrix} + \begin{bmatrix} Q_A(r,v,t) \\ Q_B(r,v,t) \end{bmatrix}, \qquad (2)$$

for the average particle densities ψ_A and ψ_B at the point (r,v,t) depending upon if the medium at (r,t) is of type A or B. It has been shown[9] that the two approaches are equivalent if the fluids are purely capturing. For a scattering medium, Pomraning[9] has shown that for a simple rod model, where the particles are constrained to travel along a line and suffer only either forward or backward collisions, the phenomenological model gives quite different results compared to the generic equation method. He has thus concluded that the two models are equivalent only for a purely capturing medium and that it is not possible to define a joint Markov process in general.

Equation (1) describes a physical process. The particles emitted by the source at some point r suffer a number of scattering collisions before being absorbed by the medium at some other point r'. Between two collisions the particles travel in straight lines called free-flights and come across alternating packets of fluids A and B of random thickness. In this paper we will show that if these free-flights are uncorrelated (its precise definition is given in Sec. II) the two approaches are equivalent.

We note that the generic equation method is equivalent to taking a snapshot of the random medium, solving the transport problem for the particular realization of the medium and then taking a phase average over all possible realizations. In a purely capturing medium the particles undergo only one free-flight from their emission by the source to capture by the medium. The changes in the random medium, as seen by the particle during its first (and the last) free-flight and the collision suffered by it can be combined into a joint Markov process. Thus the two methods yield identical results. Even in a scattering medium a joint Markov process can be defined for the first free-flight. During the subsequent free-flights, following scattering collisions, the particle encounters the same (particular) medium realization if it retraces its earlier path or a part thereof. In the rod model, a particle, on being scattered backward (forward scattering is of no consequence), retraces its earlier trajectory in the reverse direction. It therefore encounters alternating fluid packets (medium components) of exactly the same thickness as seen in the previous flight. It is therefore not surprising that for the rod model it is impossible to define a joint Markov process coupling the particle transport and the random changes in the medium.

The situation is not very different if we consider a layered system consisting of alternating slabs (infinite in transverse directions) of random thickness. This is the case if the slab geometry transport equation is used as the starting point (generic equation). Again the location of all the planes of discontinuity in the medium are known *a priori* to the concerned particle from its previous free-flight and it is not possible to define a joint Markov process.

 0022-2488/89/071554-06$02.50

If, however, we consider the particle transport process in a concrete shield and model it as a random medium to account for many different and irregular (in effect random) shapes and size of the gravel, the situation is quite different. Here one could not use the slab geometry transport equation as the starting point even if the transverse dimensions of the shield were infinite. The cross sections at any r being random functions are *not* constant in a transverse plane for any particular medium realization. Thus the general transport equation does not reduce to its slab geometry analog for any particular realization; rather the slab geometry equation is applicable only for the average quantities. Moreover in any realistic scattering model, e.g., isotropic scattering, the particle is highly unlikely to be thrown back along its previous trajectory. Thus the medium encountered by the particle after scattering is quite different from that seen in the previous flight except for a small cone of directions along the backward direction, depending upon how irregular is the shape of the gravel. It is therefore more appropriate to assume that every free-flight is independent (uncorrelated) of all previous flights, at least as a first approximation. Along each of them it is possible to define a joint Markov process just as it is done for the (only) free-flight in a purely capturing medium.

In order to simplify the calculations we will consider only the stationary, one-speed transport equation with isotropic scattering and sources in a convex region V with no incoming particles from outside. We will also assume that the source $Q(r)$ is known and that the total cross sections, σ_A and σ_B, are uniform. We will also assume that only the fluid A scatters (and may absorb) particles while the fluid B is purely capturing. Further, the scattering cross section σ_{sA} is also uniform. In Sec. II we construct an expression for the average total particle flux (or density as $v=1$) $\langle\psi(r)\rangle$ following Eq. (1), and using the above assumption of uncorrelated free-flight. In Sec. III we solve the same problem using a phenomenological model and show that the two expressions are identical.

II. GENERIC EQUATION METHOD

Consider the stationary, one-speed transport equation with isotropic scattering and sources ($v=1$),

$$\Omega\cdot\nabla_r\psi+\sigma_t(r)\psi(r,\Omega) = \frac{1}{(4\pi)}\left[Q(r)+\sigma_s(r)\int\psi(r,\Omega')d\Omega'\right]. \tag{3}$$

If there are no incident particles from outside the region V of interest, Eq. (3) is equivalent to the integral equation

$$\varphi(r)=\int_V[\sigma_s(r')\varphi(r')+Q(r')]\frac{\{\exp[-\tau(r,r')]\}}{4\pi|r-r'|^2}dr', \tag{4}$$

where the total flux $\varphi(r)$ is given by

$$\varphi(r)=\int\psi(r,\Omega)d\Omega, \tag{5}$$

and the optical distance $\tau(r,r')$ between the points r and r' is given by

$$\tau(r,r')=\int_0^{|r-r'|}\sigma_t\left(r'+\epsilon\frac{r-r'}{|r-r'|}\right)ds. \tag{6}$$

We assume that $Q(r)$ is a known source while the cross section $\sigma_t(r)$ [and $\sigma_s(r)$] is a two-state random function of r. It takes the values σ_A or σ_B. The scattering cross section σ_{sB} is assumed to be zero while σ_{sA} has a finite value. In addition to $\varphi(r)$ we will need the scattering collision density $\chi(r)$ defined by the relation

$$\chi(r)=\sigma_s(r)\varphi(r). \tag{7}$$

$\chi(r)$ satisfies the integral equation

$$\chi(r)=\sigma_s\int_V[\chi(r')+Q(r')]\frac{\{\exp[-\tau(r,r')]\}}{4\pi|r-r'|^2}dr'. \tag{8}$$

We now wish to solve Eqs. (4) and (8) and then take the ensemble average $\langle\varphi(r)\rangle$ and $\langle\chi(r)\rangle$. The solution is given by the Neumann series

$$\begin{aligned}\varphi(r)=&\int_V Q(r')\frac{\{\exp[-\tau(r,r')]\}}{4\pi|r-r'|^2}dr'+\sum_{n=1}^{\infty}\int_V\sigma_s(r_1)\\&\times\frac{\{\exp[-\tau(r,r_1)]\}}{4\pi|r-r_1|^2}\\&\times dr_1\int_V\sigma_s(r_2)\frac{\{\exp[-\tau(r_1,r_2)]\}}{4\pi|r_1-r_2|^2}\\&\times dr_2\cdots\int_V Q(r')\frac{\{\exp[-\tau(r_n,r')]\}}{4\pi|r_n-r'|^2}dr',\end{aligned} \tag{9}$$

and

$$\begin{aligned}\chi(r)=&\sigma_s(r)\int_V Q(r')\frac{\{\exp[-\tau(r,r')]\}}{4\pi|r-r'|^2}dr'\\&+\sigma_s(r)\sum_{n=2}^{\infty}\int_V\sigma_s(r_1)\frac{\{\exp[-\tau(r,r_1)]\}}{4\pi|r-r_1|^2}dr_1\\&\times\int_V\sigma_s(r_2)\frac{\{\exp[-\tau(r_1,r_2)]\}}{4\pi|r_1-r_2|^2}dr_2\cdots\sigma_s(r_{n-1})\\&\times\int_V Q(r_n)\frac{\{\exp[-\tau(r_n,r_{n-1})]\}}{4\pi|r_n-r_{n-1}|^2}dr_n.\end{aligned} \tag{10}$$

We now have to take the ensemble average of the rhs of Eqs. (9) and (10). Consider first Eq. (10). It is clear that this average is the sum of averages of various terms on the rhs. Since σ_{sB} has been taken as zero, contribution of the "n" fold integral is finite only if the points $(r,r_1,r_2,\ldots,r_{n-1})$ lie in the fluid "A" while the fluid at r_n can be of either type. Likewise in the first term on the rhs of (10) the point r must have fluid A while there is no restriction at the point r'. We now invoke the assumption of uncorrelated free-flights to write the ensemble average

$$\begin{aligned}&\langle\sigma_s(r)\sigma_s(r_1)\sigma_s(r_2)\cdots\sigma_s(r_{n-1})\exp\{-\tau(r,r_1)\}\\&\quad\times\exp\{-\tau(r_1,r_2)\}\cdots\exp\{-\tau(r_n,r_{n-1})\}\rangle\\&\quad=p_A\sigma_{sA}^n\langle\exp\{-\tau(r,r_1)\}|r,r_1\in A\rangle\\&\qquad\times\langle\exp\{-\tau(r_1,r_2)\}|r_1,r_2\in A\rangle\cdots\\&\qquad\times\langle\exp\{-\tau(r_n,r_{n-1})\}|r_{n-1}\in A\rangle.\end{aligned} \tag{11}$$

Equation (11) is a precise statement of uncorrelated free-flights. It states that the ensemble average of the product of "n" exponential factors is the product of the average of the

individual exponentials. The factor p_A accounts for the point r to lie in the medium A. The expression $\langle \exp\{-\tau(r,r_1)\} \times |r,r_1 \in A\rangle$ denotes the average of the exponential factor when both of the points r,r_1 lie in the medium A. This is given by

$$\langle \exp\{-\tau(r,r_1)\}|r,r_1\in A\rangle = \int_0^R \exp(-\tau)p_{AA}(\tau,R)d\tau, \quad R=|r-r_1|,$$
$$= [D_2\exp(-\nu_2|r-r_1|) - D_1\exp(-\nu_1|r-r_1|)]. \tag{12}$$

Here $p_{AA}(\tau,R)d\tau$ denotes the joint probability that the optical distance between the points (r,r_1) lies between τ and $\tau+d\tau$ and the point r_1 lies in the fluid 'A', *given* that the point r lies in the fluid "A." The geometric distance between the points is denoted as R. We note that $p_{AA}(\tau,R)$ and similar probabilities are given by Eqs. (45)–(48) of Ref. 1 and are not normalized to unity. Thus the expression $\langle \exp\{-\tau(r,r_1)\}|r,r_1\in A\rangle$ is strictly not an "average value" in the conventional sense, rather the first part of Eq. (12) provides its definition.

Similarly we have for the expression

$$\langle \exp\{-\tau(r_n,r_{n-1})\}|r_{n-1}\in A\rangle = \int_0^R \exp(-\tau)[p_{AA}(\tau,R)+p_{AB}(\tau,R)]d\tau,$$
$$= [C_2\exp(-\nu_2|r_n-r_{n-1}|) - C_1\exp(-\nu_1|r_n-r_{n-1}|)]$$
$$R=|(r_n-r_{n-1})|. \tag{13}$$

The derivation of Eqs. (12) and (13) is indicated in the Appendix. We will use these relations in addition to the one derived by Levermore *et al.*[1]

$$\langle \exp(-\tau(r,r'))\rangle = [E_2\exp(-\nu_2|r-r'|) - E_1\exp(-\nu_1|r-r'|)]. \tag{14}$$

In Eqs. (12)–(14), ν_1 and ν_2 are the roots of the quadratic equation

$$(\sigma_A+\lambda_A-\nu)(\sigma_B+\lambda_B-\nu)-\lambda_A\lambda_B=0, \tag{15}$$

where λ_A (or λ_B) measures the transition probability from the state A to B (from B to A) in an infinitesimal distance 'ds' as $\lambda_A\,ds$ ($\lambda_B\,ds$). The coefficients D_1,D_2,E_1,E_2 and C_1,C_2 are given by

$$D_1=\frac{\nu_2-(\sigma_A+\lambda_A)}{\nu_1-\nu_2}; \qquad D_2=\frac{\nu_1-(\sigma_A+\lambda_A)}{\nu_1-\nu_2},$$
$$C_1=\frac{\nu_2-\sigma_A}{\nu_1-\nu_2}; \qquad C_2=\frac{\nu_1-\sigma_A}{\nu_1-\nu_2}, \tag{16}$$
$$E_1=\frac{\nu_2-\bar\sigma}{\nu_1-\nu_2}; \qquad E_2=\frac{\nu_1-\bar\sigma}{\nu_1-\nu_2},$$

where the average cross section $\bar\sigma$ is defined as

$$\bar\sigma = p_A\sigma_A + p_B\sigma_B = \frac{\sigma_A\lambda_B+\sigma_B\lambda_A}{\lambda_A+\lambda_B}. \tag{17}$$

Substituting from Eqs. (11)–(13) in Eq. (10) we get

$$\langle\chi(r)\rangle = p_A\sigma_{sA}\int_V \frac{Q(r')dr'}{4\pi|r-r'|^2}[C_2\exp(-\nu_2|r-r'|) - C_1\exp(-\nu_1|r-r'|)]$$
$$+\sum_{n=2}^{\infty} p_A\sigma_{sA}^n\int_V\frac{dr_1}{4\pi|r-r_1|^2}\int_V\frac{dr_2}{4\pi|r_1-r_2|^2}\cdots\int_V Q(r_n)\frac{dr_n}{4\pi|r_n-r_{n-1}|^2}$$
$$\times\left[\prod_{i=1}^{n-1}\{D_2\exp(-\nu_2|r_i-r_{i-1}|) - D_1\exp(-\nu_1|r_i-r_{i-1}|)\}\right]$$
$$\times[C_2\exp(-\nu_2|r_n-r_{n-1}|) - C_1\exp(-\nu_1|r_n-r_{n-1}|)], \tag{18}$$

where r_0 is same as r. The Neumann series (18) can be summed very easily and we find that the ensemble average $\langle\chi(r)\rangle$ satisfies the integral equation

$$\langle\chi(r)\rangle = p_A\sigma_{sA}\int_V\frac{Q(r')dr'}{4\pi|r-r'|^2}[C_2\exp(-\nu_2|r-r'|) - C_1\exp(-\nu_1|r-r'|)] + \sigma_{sA}\int_V\frac{\langle\chi(r')\rangle dr'}{4\pi|r-r'|^2}$$
$$\times[D_2\exp(-\nu_2|r-r'|) - D_1\exp(-\nu_1|r-r'|)]. \tag{19}$$

Taking the ensemble average of Eq. (9) and proceeding as above we get

$$\langle\varphi(r)\rangle = \int_V\frac{Q(r')dr'}{4\pi|r-r'|^2}[E_2\exp(-\nu_2|r-r'|) - E_1\exp(-\nu_1|r-r'|)] + \sigma_{sA}\int_V\frac{\langle\chi(r')\rangle dr'}{4\pi|r-r'|^2}$$
$$\times[C_2\exp(-\nu_2|r-r'|) - C_1\exp(-\nu_1|r-r'|)]. \tag{20}$$

Equations (19) and (20) determine the "ensemble average scattering collision density" and the "ensemble average flux," respectively. We will now show that the phenomenological model also leads to the same results.

III. PHENOMENOLOGICAL MODEL

The stationary, one-speed form of matrix equations (2) with isotropic scattering (and $\sigma_{sB}=0$) and a known isotropic source $Q(r)$ reads as ($v=1$)

$$\begin{bmatrix}(\Omega\cdot\nabla+\sigma_A+\lambda_A) & -\lambda_B\\ -\lambda_A & (\Omega\cdot\nabla+\sigma_B+\lambda_B)\end{bmatrix}\begin{bmatrix}\psi_A(r,\Omega)\\ \psi_B(r,\Omega)\end{bmatrix}$$

$$=\frac{1}{4\pi}\begin{bmatrix}\sigma_{sA} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\int\psi_A(r,\Omega')d\Omega'\\ \int\psi_B(r,\Omega')d\Omega'\end{bmatrix}+\frac{Q(r)}{4\pi}\begin{bmatrix}p_A\\ p_B\end{bmatrix}, \quad (21)$$

where p_A and p_B are the probabilities for the medium to be in the state A or B. These are given in terms of λ coefficients by the expression

$$p_A=\frac{\lambda_B}{\lambda_A+\lambda_B}, \quad p_B=\frac{\lambda_A}{\lambda_A+\lambda_B}. \quad (22)$$

The expected angular flux $\langle\psi(r,\Omega)\rangle$ is the sum of the quantities $\psi_A(r,\Omega)$ and $\psi_B(r,\Omega)$. We first cast Eq. (21) in a form with a symmetric matrix on the lhs. Thus we have

$$\Omega\cdot\nabla\begin{bmatrix}\sqrt{\lambda_A}\psi_A\\ \sqrt{\lambda_B}\psi_B\end{bmatrix}+\begin{bmatrix}\sigma_A+\lambda_A & -\sqrt{\lambda_A\lambda_B}\\ -\sqrt{\lambda_A\lambda_B} & \sigma_B+\lambda_B\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_A}\psi_A\\ \sqrt{\lambda_B}\psi_B\end{bmatrix}$$

$$=\frac{1}{4\pi}\begin{bmatrix}\sigma_{sA} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_A}\int\psi_A\,d\Omega'\\ \sqrt{\lambda_B}\int\psi_B\,d\Omega'\end{bmatrix}+\frac{Q(r)}{4\pi}\frac{\sqrt{\lambda_A\lambda_B}}{\lambda_A+\lambda_B}\begin{bmatrix}\sqrt{\lambda_A}\\ \sqrt{\lambda_B}\end{bmatrix}. \quad (23)$$

The 2×2 real symmetric matrix Y, involving σ_A, σ_B, etc. on the lhs can be diagonalized a by real orthogonal matrix T. Let us therefore assume that for some real θ, we have

$$\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\sigma_A+\lambda_A & -\sqrt{\lambda_A\lambda_B}\\ -\sqrt{\lambda_A\lambda_B} & \sigma_B+\lambda_B\end{bmatrix}\times\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}=\begin{bmatrix}\nu_1 & 0\\ 0 & \nu_2\end{bmatrix},$$

$$\mathsf{T}^{-1}\mathsf{Y}\mathsf{T}=\Lambda \quad (24)$$

where Λ is the diagonal matrix of the eigenvalues defined in Eq. (15).

We now define a column vector f by the relation

$$f=\mathsf{T}^{-1}(\sqrt{\lambda}\,\psi), \quad \text{i.e.,} \quad \begin{bmatrix}f_A\\ f_B\end{bmatrix}=\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_A}\psi_A\\ \sqrt{\lambda_B}\psi_B\end{bmatrix}. \quad (25)$$

Premultiplying Eq. (23) by the matrix T^{-1} we obtain

$$\Omega\cdot\nabla\begin{bmatrix}f_A\\ f_B\end{bmatrix}+\begin{bmatrix}\nu_1 & 0\\ 0 & \nu_2\end{bmatrix}\begin{bmatrix}f_A\\ f_B\end{bmatrix}=\frac{1}{4\pi}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\sigma_{sA} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\int f_A\,d\Omega'\\ \int f_B\,d\Omega'\end{bmatrix}$$

$$+\frac{Q(r)}{4\pi}\frac{\sqrt{\lambda_A\lambda_B}}{\lambda_A+\lambda_B}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta\cos\theta\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_A}\\ \sqrt{\lambda_B}\end{bmatrix}. \quad (26)$$

Equation (26) can be cast into an integral form. Denoting $\int f_A(r,\Omega')d\Omega'$ and $\int f_B(r,\Omega')d\Omega'$ by $h_A(r)$ and $h_B(r)$, respectively, we have

$$\begin{bmatrix}f_A(r,\Omega)\\ f_B(r,\Omega)\end{bmatrix}=\int_V\frac{dr'}{4\pi|r-r'|^2}\delta\left(\Omega-\frac{r-r'}{|r-r'|}\right)\begin{pmatrix}\exp(-\nu_1|r-r'|) & 0\\ 0 & \exp(-\nu_2|r-r'|)\end{pmatrix}$$

$$\times\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\left\{\sigma_{sA}\begin{bmatrix}h_A(r')\cos\theta-h_B(r')\sin\theta\\ 0\end{bmatrix}+Q(r')\frac{\sqrt{\lambda_A}\lambda_B}{\lambda_A+\lambda_B}\begin{bmatrix}\sqrt{\lambda_A}\\ \sqrt{\lambda_B}\end{bmatrix}\right\}. \quad (27)$$

Integrating this equation with respect to Ω and transforming back to the original variables, we have for the total flux $\varphi_A(r)$ and $\varphi_B(r)$ defined by the relations,

$$\varphi_A(r)=\int\psi_A(r,\Omega)d\Omega; \quad \varphi_B(r)=\int\psi_B(r,\Omega)d\Omega, \quad (28)$$

the integral equations

$$\begin{bmatrix}\varphi_A(r)\\ \varphi_B(r)\end{bmatrix}=\int_V\frac{dr'}{4\pi|r-r'|^2}\begin{bmatrix}1/\sqrt{\lambda_A} & 0\\ 0 & 1/\sqrt{\lambda_B}\end{bmatrix}\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\exp(-\nu_1|r-r'|) & 0\\ 0 & \exp(-\nu_2|r-r'|)\end{bmatrix}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}$$

$$\times\begin{bmatrix}\sigma_{sA} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_A} & 0\\ 0 & \sqrt{\lambda_B}\end{bmatrix}\begin{bmatrix}\varphi_A(r')\\ \varphi_B(r')\end{bmatrix}+\int_V\frac{dr'}{4\pi|r-r'|^2}Q(r')\frac{\sqrt{\lambda_A\lambda_B}}{\lambda_A+\lambda_B}\begin{bmatrix}1/\sqrt{\lambda_A} & 0\\ 0 & 1/\sqrt{\lambda_B}\end{bmatrix}\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}$$

$$\times\begin{bmatrix}\exp(-\nu_1|r-r'|) & 0\\ 0 & \exp(-\nu_2|r-r'|)\end{bmatrix}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_A}\\ \sqrt{\lambda_B}\end{bmatrix}. \quad (29)$$

Premultiplying Eq. (29) by the matrix $\begin{bmatrix}\sigma_{sA} & 0\\ 0 & 0\end{bmatrix}\times\begin{bmatrix}\sqrt{\lambda_A} & 0\\ 0 & \sqrt{\lambda_B}\end{bmatrix}$ we have an equation for the flux $\varphi_A(r)$ given by

$$\sigma_{sA}\sqrt{\lambda_A}\varphi_A(r)=\sigma_{sA}\int_V\frac{\sigma_{sA}\sqrt{\lambda_A}\varphi_A(r')dr'}{4\pi|r-r'|^2}\left[(\sin^2\theta)\exp(-\nu_2|r-r'|)+(\cos^2\theta)\exp(-\nu_1|r-r'|)\right]$$
$$+\sigma_{sA}\frac{\lambda_A\lambda_B}{\lambda_A+\lambda_B}\int_V\frac{Q(r')dr'}{4\pi|r-r'|^2}\left[(\sin\theta)\left\{\frac{\sin\theta}{\sqrt{\lambda_A}}-\frac{\cos\theta}{\sqrt{\lambda_B}}\right\}\exp(-\nu_2|r-r'|)\right.$$
$$\left.+(\cos\theta)\left\{\frac{\cos\theta}{\sqrt{\lambda_A}}+\frac{\sin\theta}{\sqrt{\lambda_B}}\right\}\exp(-\nu_1|r-r'|)\right]. \tag{30}$$

Adding the components $\varphi_A(r)$ and $\varphi_B(r)$ of Eq. (29) we have

$$\langle\varphi(r)\rangle=\varphi_A(r)+\varphi_B(r)$$
$$=\int_V\frac{\sigma_{sA}\sqrt{\lambda_A}\varphi_A(r')dr'}{4\pi|r-r'|^2}\left[(\sin\theta)\left\{\frac{\sin\theta}{\sqrt{\lambda_A}}-\frac{\cos\theta}{\sqrt{\lambda_B}}\right\}\exp(-\nu_2|r-r'|)+(\cos\theta)\left\{\frac{\cos\theta}{\sqrt{\lambda_A}}+\frac{\sin\theta}{\sqrt{\lambda_B}}\right\}\exp(-\nu_1|r-r'|)\right]$$
$$+\frac{\lambda_A\lambda_B}{\lambda_A+\lambda_B}\int_V\frac{Q(r')dr'}{4\pi|r-r'|^2}\left[\left\{\frac{\sin\theta}{\sqrt{\lambda_A}}-\frac{\cos\theta}{\sqrt{\lambda_B}}\right\}^2\exp(-\nu_2|r-r'|)+\left\{\frac{\cos\theta}{\sqrt{\lambda_A}}+\frac{\sin\theta}{\sqrt{\lambda_B}}\right\}^2\exp(-\nu_1|r-r'|)\right]. \tag{31}$$

On comparing Eqs. (30) and (31) with Eqs. (19) and (20), we see that they are of the same form. In fact we can identify $\langle\chi(r)\rangle$ as $\sigma_{sA}\varphi(r)$. In order that these equations are identical we need show that the coefficients occurring in them are the same. Thus we must have

$$\cos^2\theta=-D_1=-\frac{\nu_2-(\sigma_A+\lambda_A)}{\nu_1-\nu_2},$$
$$\sin^2\theta=D_2=\frac{\nu_1-(\sigma_A+\lambda_A)}{\nu_1-\nu_2},$$
$$\sqrt{\lambda_A}\left\{\frac{\cos\theta}{\sqrt{\lambda_A}}+\frac{\sin\theta}{\sqrt{\lambda_B}}\right\}\cos\theta=-C_1=-\frac{\nu_2-\sigma_A}{\nu_1-\nu_2},$$
$$\sqrt{\lambda_A}\left\{\frac{\sin\theta}{\sqrt{\lambda_B}}-\frac{\cos\theta}{\sqrt{\lambda_B}}\right\}\sin\theta=C_2=\frac{\nu_1-\sigma_A}{\nu_1-\nu_2},$$
$$\frac{\lambda_A\lambda_B}{\lambda_A+\lambda_B}\left\{\frac{\cos\theta}{\sqrt{\lambda_A}}+\frac{\sin\theta}{\sqrt{\lambda_B}}\right\}^2=-E_1=-\frac{\nu_2-\bar{\sigma}}{\nu_1-\nu_2},$$
$$\frac{\lambda_A\lambda_B}{\lambda_A+\lambda_B}\left\{\frac{\sin\theta}{\sqrt{\lambda_B}}-\frac{\cos\theta}{\sqrt{\lambda_B}}\right\}^2=E_2=\frac{\nu_1-\bar{\sigma}}{\nu_1-\nu_2}. \tag{32}$$

The explicit form of the diagonalizing matrix T can now be introduced. This matrix is given by the eigenvectors of the matrix Y. In fact we have

$$\sin\theta=\frac{\Delta/\{\sqrt{\lambda_A\lambda_B}\}}{\sqrt{\{1+\Delta^2/\lambda_A\lambda_B\}}};$$
$$\cos\theta=\frac{1}{\sqrt{\{1+\Delta^2/\lambda_A\lambda_B\}}}, \tag{33}$$

where Δ is given by

$$\Delta=\sigma_A+\lambda_A-\nu_1=\nu_2-(\sigma_B+\lambda_B). \tag{34}$$

Here ν_1 and ν_2 are the roots of the quadratic equation (15) and hence we have

$$\nu_1+\nu_2=(\sigma_A+\lambda_A)+(\sigma_B+\lambda_B),$$
$$\nu_1\nu_2=(\sigma_A+\lambda_A)(\sigma_B+\lambda_B)-\lambda_A\lambda_B. \tag{35}$$

We can also show that

$$\nu_1-\nu_2=[\{(\sigma_A+\lambda_A)-(\sigma_B+\lambda_B)\}^2+4\lambda_A\lambda_B]^{1/2}. \tag{36}$$

Using (34)–(36) it can be easily shown that

$$\frac{\Delta/(\lambda_A\lambda_B)}{\{1+\Delta^2/\lambda_A\lambda_B\}}=\frac{1}{\nu_2-\nu_1}. \tag{37}$$

With the help of Eq. (37) it can be shown that the relations (32) are in fact identities. This completes the proof of the equivalence of the two approaches as stated above.

ACKNOWLEDGMENTS

I wish to thank my colleagues Dr. S. V. G. Menon, Dr. V. Kumar, and Dr. S. D. Paranjape for many fruitful discussions.

APPENDIX: EVALUATION OF AVERAGE EXPONENTIAL FACTORS

We wish to evaluate the expectation value of the exponential factors $\langle\exp\{-\tau(r,r')\}|r\in A\rangle$ and $\langle\exp\{-\tau(r,r')\}|r,r'\in A\rangle$, when one or both of the end points are located in the fluid "A." Here τ is the optical distance between the space points r and r'. Let R denote the geometric distance between them. If β denotes the distance in the fluid B, we have

$$\tau=\sigma_B\beta+\sigma_A(R-\beta). \tag{A1}$$

In order to evaluate the expectation value of the exponential factors we need first to compute the probabilities for the optical distance to lie between τ and $\tau+d\tau$, $p_{AA}(\tau,R)d\tau$ and $p_{AB}(\tau,R)d\tau$, when (1) both of the end points lie in the fluid A and (2) when the starting point lies in "A" while the other end point lies in "B," respectively. The transformation from the variable τ to β and back is governed by the relation

$$p_{AA}(\tau,R)d\tau=p_{AA}(\beta,R)d\beta;$$
$$p_{AB}(\tau,R)d\tau=p_{AB}(\beta,R)d\beta. \tag{A2}$$

The expressions for these probabilities can be derived by the methods indicated by Lindley.[10] Thus we have

$$p_{AA}(\beta,R)d\beta$$
$$=\exp(-\lambda_AR)\delta(\beta)d\beta$$
$$+\sum_{n=1}^{\infty}d\beta\int_\beta^R\exp\{-\lambda_A(R-\gamma)\}$$
$$\times h_n(\beta)\rho_n(\gamma-\beta)d\gamma. \tag{A3}$$

The first term on the rhs corresponds to the case when the entire medium between r and r' lies in a fluid of A type. The nth term in the sum gives the contribution to p_{AA} when there are exactly "n" transitions from A to B and an equal number (n) from B to A. The last transition (from B to A) occurs at a distance γ from the starting point. The factor $\exp\{-\lambda_A(R-\gamma)\}$ gives the probability for no further transitions in the interval (γ,R), and $h_n(\beta)d\beta$ gives the probability for the sum of 'n' intervals in the state B (following transitions from state A to B) to lie between β and $\beta+d\beta$. Likewise $\rho_n(\gamma-\beta)d\gamma$ is the probability for the sum of 'n' intervals in the state A to be between $\gamma-\beta$ and $\gamma-\beta+d\gamma$. The expressions for the $h_n(\beta)$ and $\rho_n(\gamma-\beta)$ are given by Lindley,[10]

$$h_n(\beta)=\lambda_B^n\frac{\beta^{n-1}}{(n-1)!}\exp(-\lambda_B\beta),$$
$$\rho_n(\tau)=\lambda_A^n\frac{\gamma^{n-1}}{(n-1)!}\exp(-\lambda_A\gamma). \tag{A4}$$

Substituting Eq. (A4) in (A3), summing over n, and changing back to τ, we have

$$\begin{aligned}p_{AA}(\tau,R)&=\exp(-\lambda_A R)\delta(\tau-\sigma_A R)\\&\quad+\exp[-\{\lambda_A(R-\beta)+\lambda_B\beta\}]\\&\quad\times\sqrt{\lambda_A\lambda_B(R-\beta)/\beta}\\&\quad\times I_1\{2\sqrt{\lambda_A\lambda_B(R-\beta)\beta}\}.\end{aligned} \tag{A5}$$

The exponential factor $\langle\exp\{-\tau(r,r')\}|r,r'\in A\rangle$ is then given by

$$\langle\exp\{-\tau(r,r')\}|r,\ r'\in A\rangle=\int_{\sigma_B R}^{\sigma_A R}\exp(-\tau)p_{AA}(\tau,R)d\tau, \tag{A6}$$

where we have assumed without any loss of generality that $\sigma_A>\sigma_B$. The integral can be evaluated by taking its Laplace transform:

$$\int_0^\infty\exp(-kR)dR\int_{\sigma_B R}^{\sigma_A R}\exp(-\tau)p_{AA}(\tau,R)d\tau=\frac{\lambda_B+\sigma_B+k}{(\lambda_B+\sigma_B+k)(\lambda_A+\sigma_A+k)-\lambda_A\lambda_B}. \tag{A7}$$

Inverting the Laplace transform, we have

$$\begin{aligned}\int_{\sigma_B R}^{\sigma_A R}\exp(-\tau)p_{AA}(\tau,R)d\tau&=\frac{\lambda_B+\sigma_B-\nu_2}{\nu_1-\nu_2}\exp(-\nu_2 R)\\&\quad-\frac{\lambda_B+\sigma_B-\nu_{12}}{\nu_1-\nu_2}\exp(-\nu_1 R).\end{aligned} \tag{A8}$$

This expression reduces to Eq. (12) of text if we use Eq. (35) for the sum of the roots ν_1 and ν_2. Similarly we have for the probability $p_{AB}(\tau,R)$,

$$\begin{aligned}p_{AB}(\tau,R)&=\exp[-\{\lambda_A(R-\beta)+\lambda_B\beta\}]\lambda_A\\&\quad\times I_0\{2\sqrt{\lambda_A\lambda_B(R-\beta)\beta}\},\end{aligned} \tag{A9}$$

and the other exponential factor can again be evaluated by its Laplace transform. The results are stated in the text.

[1] C. D. Levermore, G. C. Pomraining, O. L. Sanzo, and J. Wong, J. Math. Phys. **27**, 2526 (1986).
[2] G. C. Pomraning, Transport Theory Statist. Phys. **15**, 773 (1986).
[3] C. D. Levermore, "Transport in a random medium with anisotropy" (under preparation 1987).
[4] C. D. Levermore, G. C. Pomraning, and J. Wong, J. Math. Phys. **29**, 995 (1988).
[5] D. Vanderhaegen, J. Quant. Spectrosc. Radiative Transfer **36**, 557 (1986).
[6] D. Vanderhaegen and C. Deutsch, "Linear radiation transport in a randomly distributed binary mixture—an exact treatment for a scattering case." J. Statis. Phys. (to be published).
[7] D. C. Sahni, "An application of reactor noise techniques to neutron transport problems in a random medium" (communicated to Ann. Nucl. Energy, 1988).
[8] G. C. Pomraning, C. D. Levermore, and J. Wong, Wing Conference on Transport Theory (1988).
[9] G. C. Pomraning, J. Quant. Spectrosc. Radiative Transfer **40**, 479 (1988).
[10] D. V. Lindley, *Introduction to Probability and Statistics from a Beyesian Viewpoint* (Cambridge U.P., Cambridge, 1980), Pt. 1.

A static cylindrically symmetric solution for perfect fluid in general relativity

W. Davidson[a)]
Mathematics Department, University of Otago, Dunedin, New Zealand

(Received 6 December 1988; accepted for publication 22 March 1989)

A solution of Einstein's equations for a stationary cylindrically symmetric perfect fluid is presented. The extent of the fluid is radially infinite but the proper density μ and pressure p are physically well behaved everywhere. On the axis μ and p are finite and positive. As the radial coordinate ρ increases μ and p decrease monotonically to zero at infinity. The ratio $p/\mu(<1)$ also steadily decreases. It is shown that the regularity condition at the axis is satisfied. The solution is algebraically general of Petrov type I.

I. INTRODUCTION

There are rather few known nonvacuum solutions to Einstein's equations for the case of cylindrical symmetry, even when the space-time is stationary. For such a case the main successes[1-3] have been for a perfect fluid obeying a γ law, that is, when the proper density μ and pressure p have the relation $p = \text{const}\,\mu$. In the present paper a new one-parameter solution is given for perfect fluid of infinite extent and physically realistic characteristics. The ratio p/μ is not constant in space but is monotonically decreasing between reasonable limits. The solution, of Petrov type I, is without singularity.

II. FIELD EQUATIONS

The metric for stationary, cylindrically symmetric space-time may be taken in the form[4]

$$ds^2 = e^{2k-2U}(d\rho^2 + dz^2) + W^2 e^{-2U} d\phi^2 - e^{2U} dt^2, \tag{1}$$

where k, W, and U are functions of the radial coordinate ρ only. The metric has three Killing vectors ∂_z, ∂_ϕ, and ∂_t, corresponding to a group G_3 on T_3. Assuming a perfect fluid the Einstein equations are

$$K_0 p e^{2k-2U} = k'W'W^{-1} - U'^2, \tag{2}$$

$$= W''W^{-1} + U'^2 - k'W'W^{-1}, \tag{3}$$

$$= k'' + U'^2, \tag{4}$$

$$K_0 \mu e^{2k-2U} = 2U'' + 2U'W'W^{-1} - k'' - W''W^{-1} - U'^2, \tag{5}$$

where a prime indicates differentiation with respect to ρ.

If we choose to make U the independent variable rather than ρ, then the equations take the form (a dot representing d/dU)

$$K_0 p e^{2k-2U} = (k\dot{W}W^{-1} - 1)(\dot{\rho})^{-2}, \tag{6}$$

$$= (\ddot{W}W^{-1} - \dot{W}W^{-1}\ddot{\rho}(\dot{\rho})^{-1} - k\dot{W}W^{-1} + 1)(\dot{\rho})^{-2}, \tag{7}$$

$$= (\ddot{k} - \dot{k}\ddot{\rho}(\dot{\rho})^{-1} + 1)(\dot{\rho})^{-2}, \tag{8}$$

$$K_0 \mu e^{2k-2U} = ((\dot{k} + \dot{W}W^{-1} - 2)\ddot{\rho}(\dot{\rho})^{-1} + 2\dot{W}W^{-1} - \ddot{k} - \ddot{W}W^{-1} - 1)(\dot{\rho})^{-2}. \tag{9}$$

[a)] Permanent address: 25 Paddock Close, Edwinstowe, Notts, NG21 9LP, England.

III. A ONE-PARAMETER SOLUTION

It may be verified that a solution is given by the equations

$$e^k = e^{U/3}(81 - Ae^{-20U/3})^{1/2}, \tag{10}$$

$$W = e^{3U}(1 - Ae^{-20U/3})^{1/2}, \tag{11}$$

$$p = Be^{-34U/3}(9 + Ae^{-20U/3})^{-4}, \tag{12}$$

$$\mu = \tfrac{1}{3}Be^{-34U/3}(9 + Ae^{-20U/3})^{-5} \times (279 - 49Ae^{-20U/3}), \tag{13}$$

where A and B are positive constants connected by the relation

$$K_0 B = (3^{11/10} \times 10^3) A^{19/10}. \tag{14}$$

In the resulting metric,

$$ds^2 = e^{-4U/3}(81 - Ae^{-20U/3})(d\rho^2 + dz^2) + e^{4U}(1 - Ae^{-20U/3})d\phi^2 - e^{2U} dt^2, \tag{15}$$

the coordinate ρ is expressed in terms of the function U by the integral

$$\rho(U) = -\frac{3}{2}\int_{\infty}^{\lambda(U)} \sigma^2(\sigma - 3)^{-29/20} \times (3\sigma + 1)^{-21/20}(8\sigma + 3)^{-1}\, d\sigma, \tag{16}$$

where

$$\lambda(U) = \tfrac{1}{3}(9 + Ae^{-20U/3})(1 - Ae^{-20U/3})^{-1}. \tag{17}$$

The following ranges correspond:

$$0 \leqslant \rho(U) < \infty, \quad \infty > \lambda(U) > 3, \quad U_0 \leqslant U < \infty. \tag{18}$$

By (16)–(18) the axis of symmetry ($\rho = 0$) corresponds to $\lambda = \infty$, $U = U_0$, and $Ae^{-20U_0/3} = 1$. From (17) we have the relation

$$\frac{d\lambda}{dU} = -\tfrac{2}{3}(\lambda - 3)(3\lambda + 1), \tag{19}$$

so that

$$\frac{d\rho}{dU} = \lambda^2(\lambda - 3)^{-9/20}(3\lambda + 1)^{-1/20}(8\lambda + 3)^{-1} > 0. \tag{20}$$

It follows that ρ increases as a monotonic function of U. Also, $\rho \to \infty$ as $U \to \infty$ [while $\lambda(U) \downarrow 3$ by (17)]. In fact, (16) shows that if U_1 is very large and $U > U_1$ then the increment in ρ is

 0022-2488/89/071560-02$02.50

$$\rho(U)-\rho(U_1)\sim(\text{const})(\sigma-3)^{-9/20}\Big|_{\lambda(U_1)}^{\lambda(U)}\to\infty,$$
$$\text{as } \lambda\to 3. \tag{21}$$

The radial proper distance to coordinate $\rho(U)$, occurring at $U(\rho)$, is

$$l(\rho(U))=\int_{U_0}^{U(\rho)} e^{-2x/3}(81-Ae^{-20x/3})^{1/2}\frac{d\rho}{dx}\,dx, \tag{22}$$

x being the running value of U. Using the σ of (16) as the running value of λ, we may write the integral in the form

$$l(\rho(U))=\int_{\infty}^{\lambda(U)}\{e^{-2x/3}(81-Ae^{-20x/3})^{1/2}\}\frac{d\rho}{d\sigma}\,d\sigma, \tag{23}$$

where inside the curly brackets we have to substitute for x in terms of σ by means of (17).

Suppose $\rho\sim 0$ $[\lambda(U)\to\infty]$; then $l(\rho(U))$ is given by

$$l(\rho(U))\sim(\text{const})\sigma^{-1/2}\Big|_{\infty}^{\lambda(U)}\sim(\text{const})(\lambda(U))^{-1/2}. \tag{24}$$

Now the circumference $C(l)$ of a small circle in the plane $z=$ const centered on the axis and of radius $l(\rho(U))$, is, by the metric (15),

$$C(l)\sim(\text{const})(1-Ae^{-20U/3})^{1/2}\sim(\text{const})(\lambda(U))^{-1/2}. \tag{25}$$

It follows that

$$C(l)/l\to\text{const}, \quad \text{as } \rho\to 0, \tag{26}$$

so that the condition for elementary flatness at the axis is satisfied. Alternatively, by taking $x^i=\rho$, z, ϕ, and t, for $i=1$, 2, 3, and 4, respectively, and setting

$$\eta^i=(0,0,1,0), \quad X=\eta^i\eta_i, \tag{27}$$

we can show that the regularity condition,

$$X^{,i}X_{,i}/X\to\text{const}, \quad \text{as } \rho\to 0, \tag{28}$$

is satisfied. That is, by suitable scaling (depending on A) ϕ can be made the usual polar angle.

Suppose $\rho(U)\to\infty$ $[\lambda(U)\to 3]$; in this case from (17) we find

$$e^{-2U/3}\sim(\text{const})(\lambda-3)^{1/10}. \tag{29}$$

Hence (23) provides, by (16) and (29), for U_1 large and $U>U_1$,

$$l(\rho(U))-l(\rho(U_1))\sim(\text{const})(\sigma-3)^{-7/20}\Big|_{\lambda(U_1)}^{\lambda(U)}\to\infty,$$
$$\text{as } \lambda(U)\to 3. \tag{30}$$

It follows that the fluid cylinder is radially infinite.

The solution for $0\leqslant\rho<\infty$ $(U_0\leqslant U<\infty)$ is without singularity. Straightforward calculations show that the solution is algebraically general of Petrov type I.

IV. PHYSICAL PROPERTIES

From (12), (13), and (17) we find that

$$\begin{aligned} &p>0, \quad \frac{dp}{dU}<0,\\ &\mu>0, \quad \frac{d\mu}{dU}<0, \end{aligned} \quad \text{for } 0\leqslant\rho<\infty\,(U_0\leqslant U<\infty). \tag{31}$$

At the axis $\rho=0$, setting $U=U_0$, $Ae^{-20U_0/3}=1$, one obtains $p=p_0$, $\mu=\mu_0$, where

$$\begin{aligned} p_0&=10^{-4}Be^{-34U_0/3},\\ \mu_0&=\tfrac{23}{3}\times 10^{-4}Be^{-34U_0/3}. \end{aligned} \tag{32}$$

From these values on the axis p, μ decrease monotonically to zero as $\rho\to\infty$. For the ratio p/μ we obtain

$$p/\mu=3(9+Ae^{-20U/3})(279-49Ae^{-20U/3})^{-1}, \tag{33}$$

which is also steadily decreasing and falls in the range

$$\tfrac{3}{23}\geqslant p/\mu>\tfrac{3}{31}. \tag{34}$$

The velocity of sound in the fluid, $a=(dp/dU/d\mu/dU)^{1/2}$, also takes reasonable values, varying in the range ~ 0.41 at the axis to ~ 0.31 at infinity (velocity of light $c=1$).

The mass per unit length of the fluid cylinder is finite.

V. CONCLUSION

The cylindrically symmetric static one-parameter solution to Einstein's equations is physically well behaved and without singularity in the (spatially infinite) range $0\leqslant\rho<\infty$.

[1] A. B. Evans, J. Phys. A **10**, 1303 (1977).
[2] A. F. da F. Teixera and I. Wolk, Nuovo Cimento B **41**, 387 (1977).
[3] D. Kramer, Class. Quantum Grav. **5**, 393 (1988).
[4] D. Kramer, H. Stephani, E. Herlt, and M. MacCallum, *Exact Solutions of Einstein's Field Equations* (Cambridge U. P., Cambridge, 1980), Chaps. 17 and 20.

Multiple-soliton solutions of Einstein's equations

A. Economou and D. Tsoubelis
Department of Physics, Division of Theoretical Physics, GR 451 10 Ioannina, The University of Ioannina, Greece

(Received 11 August 1988; accepted for publication 22 February 1989)

Using the Belinsky–Zakharov generating technique and a flat metric as a seed, two- and four-soliton solutions of the Einstein vacuum equations for the cases of stationary axisymmetric, cylindrically symmetric, or plane symmetric gravitational fields are considered. Three- and five-parameter classes of exact solutions are obtained, some of which are new.

I. INTRODUCTION

Among the techniques developed in recent years for the generation of new solutions of the Einstein vacuum (and stiff matter or Einstein–Maxwell) equations from simpler known ones the inverse scattering method (ISM) of Belinsky and Zakharov[1] (BZ) has turned out to be one of the most fertile. Because it applies in all cases where the space-time manifold admits a pair of commuting non-null Killing vectors, the BZ technique has already been used for the construction of a large number of exact solutions representing stationary axisymmetric, cylindrically symmetric, and plane symmetric gravitational fields.[2–7]

In the present paper we consider the "2- and 2×2 soliton" solutions of the Einstein vacuum equations, which can be constructed using the BZ technique and a simple (flat) Kasner metric as the known or "seed" solution. This type of diagonal seed was first used by BZ in illustrative examples of the application of their method in the original papers cited above. Where we differ from BZ and other authors who have used the Kasner or other diagonal seed metrics is in treating the three cases mentioned below in a unified manner. Thus following the lead of Letelier,[3] we first develop a simplified version of the BZ formulas for the product metric coefficients for the general "N-soliton" solution in terms of determinants of $N\times N$ matrices. Subsequently, we specialize our results to the two-soliton case and use the method of Tomitatsu[8] to make them applicable to the four-soliton double poles or the two$\times$two-soliton case as well. The application of these results to the simple Kasner metric mentioned above allows us to accomplish the following objectives.

First, we rederive several important solutions of the Einstein vacuum equations recently discovered by other methods. These include the Chandrasekhar–Xanthopoulos[9] colliding plane-waves solution; the cosmic string plus gravitational waves solution of Xanthopoulos[10]; and the five-parameter family of stationary axisymmetric metrics discovered by Kinnersley and Chitre,[11] which generalized the $\delta=2$ solution of Tomimatsu and Sato.[12] There are two points gained by this rederivation. On one hand, the interrelations between the above solutions are clearly brought out and a frame of their classification is established. On the other hand, the advantage of the BZ technique of giving all the components of the product metric tensor by algebraic means is made explicit. Thus the BZ technique allows us to construct the Kinnersley–Chitre[11] metric completely, while the method by which this solution was arrived at originally allowed for the construction of the corresponding Ernst potential only.

The second objective accomplished in the applications part of this paper consists of generating several classes of new solutions, including (i) one-parameter generalizations of the Chandrasekhar–Xanthopoulos[9] and Xanthopoulos[10] metrics, and (ii) six families of pentaparametric solutions (the Kinnersley–Chitre metric being one of them)—two in each of the stationary axisymmetric, cylindrically symmetric, and plane symmetric groups of space-times.

The plan of our paper is as follows. In Sec. II, we present an outline of the BZ solution-generating method. In Sec. III, we give a set of determinantal formulas for the product metric coefficients which hold in the general N-soliton case when the seed metric is diagonal. On the basis of the above formulas, the two-soliton product metric coefficients are constructed in Sec. IV in terms of the two "pole trajectory" functions and a pair of arbitrary real or complex parameters. Section IV also covers the method by which the formulas given in Sec. III are made applicable in the "double-poles" case. Last, in Sec. V, we present the new solutions of the Einstein equations that can be obtained as two- and two$\times$two-soliton products of the application of the results of Sec. IV on the Kasner seed case.

II. THE BZ SOLITON TECHNIQUE

The metric of the space-times under consideration can be written in the form

$$ds^2=f(x^3,x^4)[-\epsilon(dx^4)^2+(dx^3)^2]+g_{ab}(x^3,x^4)dx^a\,dx^b, \tag{2.1}$$

where a, b run from 1 to 2 and $\epsilon=\pm1$.

Introducing the coordinates ζ, η, and the function α via

$$\zeta\equiv x^3+\sqrt{\epsilon}x^4,\quad \eta=x^3-\sqrt{\epsilon}x^4, \tag{2.2}$$

and

$$\alpha^2=\det(g)\equiv\det(g_{ab}), \tag{2.3}$$

respectively, one can write the Einstein vacuum equations for the metric (2.1) in the form

$$(\alpha,_\zeta g^{-1})_{,\eta}+(\alpha g_{,\eta}g^{-1})_{,\zeta}=0, \tag{2.4}$$

$$(\ln f)_{,\zeta}=(\ln\alpha)_{,\zeta\zeta}/(\ln\alpha)_{,\zeta}+\mathrm{Tr}(\alpha g_{,\zeta}g^{-1})^2/4\alpha\,\alpha_{,\zeta}, \tag{2.5a}$$

$$(\ln f)_{,\eta}=(\ln\alpha)_{,\eta\eta}/(\ln\alpha)_{,\eta}+\mathrm{Tr}(\alpha g_{,\eta}g^{-1})^2/4\alpha\,\alpha_{,\eta}, \tag{2.5b}$$

 0022-2488/89/071562-08$02.50

where $(\ )_x \equiv \partial_x(\ ) \equiv \partial(\ )/\partial x$. In particular, the trace of Eq. (2.3) reads as

$$\alpha_{,\zeta\eta} = 0 , \tag{2.6}$$

whose general solution can be given in the form

$$\alpha(\zeta,\eta) = a(\zeta) + b(\eta) , \tag{2.7}$$

where a, b are arbitrary functions of the indicated arguments.

In the BZ approach, Eq. (2.4), which is the integrability condition of Eqs. (2.5), is replaced by the coupled "Schrödinger equations"

$$D_\zeta\psi \equiv \left(\partial_\zeta + \frac{2\alpha_{,\zeta}\lambda}{\alpha-\lambda}\partial_\lambda\right)\psi = \frac{\alpha g_{,\zeta} g^{-1}}{(\alpha-\lambda)}\psi , \tag{2.8a}$$

$$D_\eta\psi \equiv \left(\partial_\eta + \frac{2\alpha_{,\eta}\lambda}{\alpha-\lambda}\partial_\lambda\right)\psi = \frac{\alpha g_{,\eta} g^{-1}}{(\alpha-\lambda)}\psi \tag{2.8b}$$

for the 2×2 matrix "wavefunction" ψ, which depends on the complex "spectral parameter" λ in such a way that

$$\lim_{\lambda\to 0}\psi(\zeta,\eta,\lambda) = g(\zeta,\eta) . \tag{2.9}$$

Suppose now that $\psi^{(0)}$ is a known solution of Eqs. (2.8) which corresponds to $g^{(0)}$ in the sense of Eq. (2.9). Then, as BZ have shown, the ansatz

$$\psi = X\psi^{(0)} , \tag{2.10}$$

together with the assumption that the "scattering matrix" $X(\zeta,\eta,\lambda)$ has only simple poles in the complex λ plane, leads to a new solution of Eqs. (2.8) which, thanks to Eq. (2.9), determines a new solution of Eq. (2.4).

The poles μ_k, $k = 1,\ldots,N$ of the scattering matrix X are found to be coordiante dependent. Specifically, the pole trajectories are given by

$$\mu_k = (w_k - \beta) + [(w_k-\beta)^2 - \alpha^2]^{1/2} , \tag{2.11}$$

where w_k are arbitrary constants and

$$\beta = a(\zeta) - b(\eta) \tag{2.12}$$

is the harmonic conjugate of the function α defined by Eq. (2.7).

It turns out that the functions $\mu_k(\zeta,\eta)$, along with the matrix $\psi^{(0)}$, determine the new metric (f,g) produced from the seed $(f^{(0)}, g^{(0)})$ completely and only via algebraic manipulations. This is made explicit in the final BZ formulas, which read as

$$f = cf^{(0)}\alpha^{-N^2/2}\left|\prod_{k=1}^{N}\mu_k\right|^{N+1} \times\left[\prod_{\substack{k,l=1\\ k>l}}^{N}(\mu_k-\mu_l)^2\right]^{-1}\det(\Gamma) , \tag{2.13a}$$

$$g_{ab} = \left[\prod_{k=1}^{N}|\mu_k/\alpha|\right] \times\left[g^{(0)}_{ab} - \sum_{k,l=1}^{N}\frac{\Gamma^{-1}_{kl}N^{(k)}_a N^{(l)}_b}{\mu_k\mu_l}\right] , \tag{2.13b}$$

where

$$\Gamma_{kl} \equiv (\mu_k\mu_l - \alpha^2)^{-1} n^{(k)}_a g^{(0)}_{ab} n^{(l)}_b , \tag{2.14a}$$

$$N^{(k)}_a \equiv g^{(0)}_{ab} n^{(k)}_b , \tag{2.14b}$$

$$n^{(k)}_a \equiv m^{(k)}_b M^{(k)}_{ba} , \quad m^{(k)}_b \text{ arbitrary constants}, \tag{2.14c}$$

$$M^{(k)} \equiv [\psi^{(0)}(\eta,\zeta;\mu_k)]^{-1} , \tag{2.14d}$$

and c is an arbitrary real constant.

In Eqs. (2.14) the summation convention holds for the indices a, b, while in Eq. (2.13a) the term in square brackets should be replaced by unity when $N = 1$.

III. SOLUTIONS DERIVABLE FROM A DIAGONAL SEED

It is clear from Eqs. (2.14) that the key item in the construction of the new solution (f,g) of Einstein's vacuum equations starting from the "seed metric" $(f^{(0)},g^{(0)})$ is the set of N matrices $\{\psi^{(0)}(\zeta,\eta;\mu_k)\}$: These are obtained by integrating the system of equations (2.8) along the pole trajectories. However, when $g^{(0)}$ is not diagonal, the system (2.8) is generally very difficult to integrate. Therefore, one usually adopts the simplifying assumption that $g^{(0)}$ is a diagonal matrix, in which case $\psi^{(0)}$ can be assumed to be diagonal as well. Integrating the trace of Eqs. (2.8) along the pole trajectories μ_k we obtain

$$\det(\psi^{(0)}) = 2w_k\mu_k , \tag{3.1}$$

which allows us to write the matrix $\psi^{(0)}$ in the form

$$\psi^{(0)}(\zeta,\eta;\mu_k) = \mathrm{diag}(\psi_k, 2w_k\mu_k\psi_k^{-1}) , \tag{3.2}$$

where the function ψ_k satisfies

$$(\ln\psi_k)_{,\zeta} = [\alpha/(\alpha-\mu_k)](\ln g^{(0)}_{11})_{,\zeta} , \tag{3.3a}$$

$$(\ln\psi_k)_{,\eta} = [\alpha/(\alpha+\mu_k)](\ln g^{(0)}_{11})_{,\eta} . \tag{3.3b}$$

The diagonality assumption for $g^{(0)}$ simplifies not only the procedure that leads to the matrix $\psi^{(0)}$, but the algebraic system of Eqs. (2.13) and (2.14) as well. Thus substituting Eq. (3.2) into Eqs. (2.13), one finds that

$$g_{11} = \left|\prod_k\sigma_k\right|\left\{1 - \sum_{k,l}\left(\frac{S_kS_l}{\sigma_k\sigma_l}\right)\Delta^{-1}_{kl}\right\} g^{(0)}_{11} , \tag{3.4a}$$

$$g_{22} = \left|\prod_k\sigma_k\right|\left\{1 - \sum_{k,l}\left(\frac{1}{\sigma_k\sigma_l}\right)\Delta^{-1}_{kl}\right\} g^{(0)}_{22} , \tag{3.4b}$$

$$g_{12} = -\left(\frac{\alpha}{\sqrt{\epsilon}}\right)\left|\prod_k\sigma_k\right|\sum_{k,l}\left(\frac{S_k}{\sigma_k\sigma_l}\right)\Delta^{-1}_{kl} , \tag{3.4c}$$

where

$$\sigma_k \equiv (\mu_k/\alpha) , \tag{3.5a}$$

$$S_k \equiv q_k g^{(0)}_{11}\psi_k^{-2}\sigma_k , \quad q_k \text{ arbitrary constants}, \tag{3.5b}$$

$$\Delta_{k,l} \equiv (S_kS_{l+1})/(\sigma_k\sigma_l - 1) , \tag{3.5c}$$

and all the sums and products run from 1–N. Finally, the identities

$$\det(\gamma_k\delta_l + \Delta_{kl}) = \left(1 + \sum_{k,l}\gamma_k\delta_l\Delta^{-1}_{kl}\right)\det(\Delta) , \tag{3.6a}$$

$$\det(\gamma_k\delta_l\Delta_{kl}) = \left(\prod_k\gamma_k\right)^2\det(\Delta) \tag{3.6b}$$

allow us to write Eq. (3.4) in the determinantal form (to be compared with the third paper listed in Ref. 3)

$$g_{11} = \left|\prod_{k=1}^{N}\sigma_k\right|\frac{L_{(-1)}}{L_{(0)}}g^0_{11} , \tag{3.7a}$$

$$g_{22}=\epsilon^N\left|\prod_{k=1}^{N}\sigma_k^{-1}\right|\frac{L_{(+1)}}{L_{(0)}}g_{22}^{(0)}, \tag{3.7b}$$

$$g_{12}=\frac{\alpha}{\sqrt{\epsilon}}\left|\prod_{k=1}^{N}\sigma_k\right|\frac{L}{L_{(0)}}, \tag{3.7c}$$

where

$$L_{(\delta)}\equiv\det\Delta_{(\delta)}\ ,\quad \Delta_{(\delta)kl}\equiv\left[\frac{(\sigma_k\sigma_l)^{\delta}S_kS_l+1}{\sigma_k\sigma_l-1}\right], \tag{3.8a}$$

with $\delta=0,\ \pm 1$ and

$$L\equiv\det(\Delta_{(0)kl})-\det[\Delta_{(0)kl}+(S_k/\sigma_k\sigma_l)]\ . \tag{3.8b}$$

Similarly, the metric coefficient f can be written in the form

$$f=c\frac{\alpha^{-N^2/2}\left|\Pi_{k=1}^{N}\sigma_k\right|^N}{\left[\Pi_{\substack{k,l=1\\k>l}}^{N}(\sigma_k-\sigma_l)^2\right]\left[\Pi_{k=1}^{N}S_k\right]}L_{(0)}f^{(0)}. \tag{3.9}$$

Before concluding this section, let us note that the product metric, given by Eqs. (3.7)–(3.9), depends on $2N$ parameters—the arbitrary constants w_k and q_k. The w_k determine the pole trajectories μ_k via Eq. (2.9) and are incorporated in the function σ_k according to Eq. (3.5a). The q_k appear as multiplicative constants in the functions S_k defined by Eq. (3.5b). Thus one can make the functions S_k vanish by setting all the q_k's equal to zero. According to Eq. (3.8b) this choice leads to a diagonal product metric and therefore, it represents the easiest application of the BZ formulas. A diagonal product metric is also obtained if we let all the (q_k^{-1})'s go to zero. Such diagonal N-soliton metrics were given and studied by Carr and Verdaguer for the case where the Kasner cosmological model serves as seed.[5] In the general nondiagonal case, the functions S_k depend on the ψ_k's; the latter are obtained by integrating Eqs. (3.3). Assuming that the diagonal seed metric is written in the form

$$g^{(0)}=\mathrm{diag}(\alpha/\sqrt{\epsilon})(\epsilon e^{\phi},e^{-\phi}), \tag{3.10}$$

one can obtain the S_k functions more directly from the expression

$$\ln S_k=-\int\frac{(\ln\sigma_k)_{,\xi}}{(\ln\alpha)_{,\xi}}\phi_{,\xi}\,d\xi+\frac{(\ln\sigma_k)_{,\eta}}{(\ln\alpha)_{,\eta}}\phi_{,\eta}\,d\eta, \tag{3.11}$$

which results by combining Eqs. (3.3), (3.5a), and (3.5b). Still, specific applications of the BZ technique can be carried to completion only by assuming simple expressions for the function ϕ, which determines the g_{ab} part of the seed metric.

IV. TWO- AND TWO×TWO-SOLITON SOLUTIONS

A. A pair of simple poles

Let us now assume that the scattering matrix X has only two simple poles in the complex λ plane, located at w_1 and w_2, respectively. Then Eqs. (3.7)–(3.9) and simple algebra give the following expressions for the general two-soliton solution derivable from a diagonal seed:

$$g_{11}=[(\sigma_1\sigma_2-1)^2(\sigma_1S_2-\sigma_2S_1)^2+(\sigma_1-\sigma_2)^2\times(\sigma_1\sigma_2+S_1S_2)^2]\,Zg_{11}^{(0)}, \tag{4.1a}$$

$$g_{22}=[(\sigma_1\sigma_2-1)^2(\sigma_1S_1-\sigma_2S_2)^2+(\sigma_1-\sigma_2)^2\times(1+\sigma_1\sigma_2S_1S_2)^2]\,Zg_{22}^{(0)}, \tag{4.1b}$$

$$g_{12}=2(w_2-w_1)\sigma_1\sigma_2[(\sigma_1S_1-\sigma_2S_2)(\sigma_1\sigma_2+S_1S_2)+(\sigma_1S_2-\sigma_2S_1)(1+\sigma_1\sigma_2S_1S_2)]\,Z, \tag{4.1c}$$

$$f=\mathrm{const}[\sigma_1\sigma_2S_1S_2(\sigma_1^2-1)(\sigma_2^2-1)Z]^{-1}f^{(0)}, \tag{4.1d}$$

where

$$Z^{-1}\equiv|\sigma_1\sigma_2|[(\sigma_1-\sigma_2)^2(1+S_1S_2)^2+(\sigma_1\sigma_2-1)^2(S_1-S_2)^2]\ . \tag{4.2}$$

In deriving Eqs. (4.1c) and (4.1d) we made use of the identity

$$\alpha(\sigma_i-\sigma_j)(\sigma_i\sigma_j-1)=2(w_i-w_j)\sigma_i\sigma_j\ , \tag{4.3}$$

which follows from Eq. (2.11), and constant quantities were absorbed in the multiplicative constant c of Eq. (3.9).

At this point, let it be noted that according to Eqs. (2.11) and (3.5a),

$$\sigma_k^{(+)}\sigma_k^{(-)}=1\ \ (\text{no sum over } k)\ , \tag{4.4}$$

where $\sigma_k^{(\pm)}$ denote the σ_k functions corresponding to the $(\pm)$ choice of sign in Eq. (2.11). Similarly, Eq. (3.11) implies the relation

$$S_k^{(+)}S_k^{(1)}=D_k\ , \tag{4.5}$$

with D_k an arbitrary constant, between the $S_k^{(\pm)}$ functions that correspond to the $\sigma_k^{(\pm)}$'s, as in Eq. (3.5b). When relations (4.4) and (4.5) are taken into account, it is an easy matter to verify that the rhs of Eqs. (4.1) are invariant under the transformation

$$\{\sigma_j^{(+)},S_j^{(+)}\}\to\{\sigma_j^{(-)}=1/\sigma_j^{(+)},S_j^{(-)}=-1/S_j^{(+)}\}\ . \tag{4.6}$$

Therefore, any choice of sign in Eq. (2.11) implies no loss of generality.

On the other hand, the poles w_1, w_2 must both be real or complex conjugate. Therefore, there is no loss of generality if we take

$$w_1=-w_2=w,\quad \text{when } w_k\in\mathbb{R}, \tag{4.7a}$$

$$w_1=w_2=iw,\quad \text{when } w_k\in\mathbb{C}, \tag{4.7b}$$

since this choice implies no more than a translation in the $(\alpha/\sqrt{\epsilon},\beta)$ plane. Correspondingly, the pole trajectories can be chosen to be

$$\mu_1=(w-\beta)+[(w-\beta)^2-\alpha^2]^{1/2},\qquad \mu_2=-(w+\beta)+[(w+\beta)^2-\alpha^2]^{1/2}, \tag{4.8}$$

when $w_k\in\mathbb{R}$ and

$$\mu_1=\mu_2=(iw-\beta)+[(w-\beta)^2-\alpha^2]^{1/2} \tag{4.9}$$

where $w_k\in\mathbb{C}$.

The expressions obtained thus far for the N-soliton solution are given in terms of the real-valued harmonic conjugate functions $(\alpha/\sqrt{\epsilon},\beta)$ which can be retained as the coordinate system replacing the original (x^3,x^4) or (ξ,η) system. However, in the two-soliton case, it turns out to be much more convenient to introduce the coordinates (x,y) defined by

$$\beta=wxy\ ,$$

$$\frac{\alpha}{\sqrt{\epsilon}}=\begin{cases}w[\epsilon(1-x^2)(1-y^2)]^{1/2}, & \text{when } w_k\in\mathbb{R}\ ,\\ w[\epsilon(1+x^2)(y^2-1)]^{1/2}, & \text{when } w_k\in\mathbb{C}\ .\end{cases} \tag{4.10}$$

Substituting Eqs. (4.8) and (4.10) into (3.10a), we find that

$$\sigma_1 = (1+x)(1-y)/[(1-x^2)(1-y^2)]^{1/2}, \qquad \sigma_2 = (x-1)(1-y)/[(1-x^2)(1-y^2)]^{1/2} \tag{4.11}$$

when $w_k \in \mathbb{R}$. In order to obtain the corresponding σ_k's when $w_k \in \mathbb{C}$ one can simply use the mapping

$$(x, y; w) \to (-ix, y; iw): (\text{real poles case}) \to (\text{complex poles case}), \tag{4.12}$$

which is implicit in Eq. (4.10).

B. A pair of double poles

The BZ formulas for the N-soliton solution are immediately applicable only when the N poles are distinct. Thus when two or more of the w_k's coincide one has to turn to the use of limiting procedures in order to find the appropriate version of the above formulas.

Consider, for example, the case where both w_1 and w_2 of Sec. IV A are double poles: By this we mean the case where the scattering matrix has four simple poles and we let $(w_3, w_4) \to (w_1, w_2)$. We then turn to Eqs. (3.7)–(3.9) in order to obtain the coefficients (f, g) of the new metric. Note, however, that if we let $(\sigma_3, \sigma_4) \to (\sigma_1, \sigma_2)$ and $(S_3, S_4) \to (S_1, S_2)$ as $(w_3, w_4) \to (w_1, w_2)$, then the four-soliton [4×4] matrix $\Delta_{(\delta)ij}$ will have pairs of equal rows and columns. As a result, the L functions will vanish, making Eqs. (3.7) and (3.9) inapplicable.

Therefore, let us consider the alternative[8] where $\sigma_1 = \sigma_1^{(+)}$, $\sigma_2 = \sigma_2^{(+)}$ and as $(w_3, w_4) \to (w_1, w_2)$:

$$(\sigma_3, \sigma_4) \to (\sigma_1^{(-)}, \sigma_2^{(-)})$$

and

$$(S_3, S_4) \to (S_1^{(-)} = D_1/S_1^{(+)}, S_2^{(-)} = D_2/S_2^{(+)}), \tag{4.13}$$

where D_i's are arbitrary constants. The problem that now arises for the [1,3], [2,4] elements of the symmetric [4×4] matrices $\Delta_{(\delta)ij}$ is overcome by letting the D_i's go to -1. Therefore, let us consider

$$D_i = -1 + \xi_i (w_{i+2} - w_i), \tag{4.14}$$

where $i = 1,2$ and ξ_i are arbitrary constants. Then using Eqs. (3.15a), (4.8), (4.9), and (4.13) we find that the limit of $\Delta_{(\delta)ij}$ as $(w_3, w_4) \to (w_1, w_2)$ is the symmetric matrix $E_{(\delta)ij}$, where

$$E_{(\delta)kl} = \Delta_{(\delta)kl}, \tag{4.15a}$$

$$E_{(\delta)k+2,l+2} = -(\sigma_k \sigma_l)^{1-\delta} (S_k S_l)^{-1} \Delta_{(\delta)kl}, \tag{4.15b}$$

$$E_{(\delta)13} = -\left[\frac{\partial(\ln \sigma_1)}{\partial w_1}\right]^{-1} \left[\xi_1 + \frac{\partial(\ln S_1)}{\partial w_1}\right] - \delta, \tag{4.15c}$$

$$E_{(\delta)14} = \sigma_2^{1-\delta} S_2^{-1} (\sigma_2 - \sigma_1)^{-1} (\sigma_1^{\delta} S_1 - \sigma_2^{\delta} S_2), \tag{4.15d}$$

$$E_{(\delta)23} = \sigma_1^{1-\delta} S_1^{-1} (\sigma_2 - \sigma_1)^{-1} (\sigma_1^{\delta} S_1 - \sigma_2^{\delta} S_2), \tag{4.15e}$$

$$E_{(\delta)24} = -\left[\frac{\partial(\ln \sigma_2)}{\partial w_2}\right]^{-1} \left[\xi_2 + \frac{\partial(\ln S_2)}{\partial w_2}\right] - \delta, \tag{4.15f}$$

where k, l run from 1–2. Moreover, from Eqs. (2.9) and (3.5a) it easily follows that

$$\frac{\partial(\ln \sigma_k)}{\partial w_k} = \frac{2\sigma_k}{\alpha(\sigma_k^2 - 1)}. \tag{4.16}$$

Thus the equality of the pairs of rows or columns of $\Delta_{(\delta)ij}$ in the limit $(w_3, w_4) \to (w_1, w_2)$ has been broken and, once the functions S_k have been determined, Eqs. (4.15) and (4.16) provide the elements necessary for calculating the L functions of the four-soliton formulas.

V. APPLICATIONS

A. A pair of simple poles

As a first application of the results obtained in Sec. IV, let us consider the two-soliton solutions that can be derived from the metric

$$ds^2 = -\epsilon(dx^4)^2 + (dx^3)^2 + \epsilon(dx^1)^2 + (\alpha(x^3, x^4)\sqrt{\epsilon})^2 (dx^2)^2, \tag{5.1}$$

where $(\alpha/\sqrt{\epsilon})$ is any real solution of Eq. (2.6).

The metric (5.1) results from taking

$$\phi = -\ln(\alpha/\sqrt{\epsilon}) \tag{5.2}$$

in Eq. (3.10), a choice that is in accord with the vacuum field equations (2.4). Substitution of Eq. (5.2) into (3.11) leads to the relation

$$S_k = Q_k \sigma_k, \quad Q_k \text{ arbitrary constants}, \tag{5.3}$$

which makes explicit the simplicity of expression (5.2) for ϕ from the standpoint of the soliton technique.

By choosing the value of ϵ and the specific form of α one specifies the "gauge." Thus we will distinguish the following cases.

(i) For the axisymmetric gauge,

$$\epsilon = -1, \quad (x^1, x^2, x^3, x^4) = (t, \varphi, z, \rho), \quad \alpha = i\rho, \quad \beta = z. \tag{5.4}$$

In this case Eq. (5.1) becomes

$$ds^2 = -dt^2 + d\rho^2 + \rho^2 d\varphi^2 + dz^2, \tag{5.5}$$

which makes the range and meaning of the coordinates evident.

(ii) For the plane symmetric gauge,

$$\epsilon = 1, \quad (x^1, x^2, x^3, x^4) = (x', y', z', t'), \quad \alpha = t', \quad \beta = z'. \tag{5.6}$$

Now, the flat metric in Eq. (5.1) can be considered to represent a Kasner or Bianchi type I universe.

(iii) For the cylindrically symmetric gauge,

$$\epsilon = 1, \quad (x^1, x^2, x^3, x^4) = (z, \varphi, \rho, t), \quad \alpha = \rho, \quad \beta = t. \tag{5.7}$$

As in the axisymmetric gauge, in this case the seed metric is also the Minkowski metric in cylindrical coordinates. However, the product metric will be different. In the present case the two-soliton solution will preserve the cylindrical symmetry of the Minkowsi space-time, while in the axisymmetric gauge it is the stationary axially symmetric character of the original metric that will be preserved.

Returning to Eq. (5.3), let us note that the resulting metric will be real provided that Q_1 and Q_2 are chosen to be

real or complex conjugate when the poles w_1 and w_2 are real or complex conjugate, respectively. In the real poles case, let us introduce the parameters p, q, and l via the relations

$$\frac{1}{q} = -\frac{Q_1+Q_2}{1+Q_1Q_2}, \quad \frac{p}{q} = \frac{Q_1-Q_2}{1+Q_1Q_2}, \quad \frac{l}{q} = \frac{Q_1Q_2-1}{Q_1Q_2+1}. \tag{5.8}$$

Then

$$q^2+p^2 = 1+l^2. \tag{5.9}$$

In order to cover the complex poles case, all we need do is let $p \to ip$ in Eq. (5.8); then (5.9) is replaced by

$$q^2-p^2 = 1+l^2. \tag{5.10}$$

1. Complex conjugate poles

Let us now substitute Eq. (5.3) into Eq. (4.1) and consider the case of complex w_k first. Taking into account the $p \to ip$ version of Eq. (5.8) we find that the product metric can be written in the form

$$ds^2 = f[-\epsilon(dx^4)^2+(dx^3)^2] + \epsilon(Y/X)[dx^1-\omega\,dx^2]^2 + (\alpha^2/\epsilon)(X/Y)(dx^2)^2, \tag{5.11}$$

where

$$f = \text{const}[X/(x^2+y^2)], \tag{5.12a}$$

$$X = (l-qy)^2+(1-px)^2, \tag{5.12b}$$

$$Y = q^2(y^2-1)+p^2(x^2+1), \tag{5.12c}$$

$$\left(\frac{\omega}{2w}\right) = -\frac{q(y^2-1)(p+x)+p(x^2+1)(q-ly)}{q^2(y^2-1)p^2(x^2+1)} = \frac{1}{p}\frac{q(y^2-1)(1-px+l^2-lq)+lp^2(y-1)(x^2+1)}{q^2(y^2-1)+p^2(x^2+1)} - \frac{(q-l)}{p}. \tag{5.12d}$$

Depending on the gauge, the line element given by Eqs. (5.11) and (5.12) represents the following three classes of space-times.

(i) The $a^2 > m^2$ Kerr–NUT (Newman–Unti–Tamburino) metrices. This can be made explicit by first gauging away the constant $2p^{-1}(q-l)$ in the second version of Eq. (5.12d) by letting $x^1 \to x^1 + 2p^{-1}(q-l)x^2$ and then choosing the axisymmetric gauge.

Inverting Eq. (4.10) we obtain

$$2wx = r_+ + r_-, \quad 2iwy = r_+ - r_-, \tag{5.13}$$

where

$$r_\pm = [(z+iw)^2+\rho^2]^{1/2}. \tag{5.14}$$

Equation (4.10) also gives the restriction $|y|<1$ for the range of the y coordinate. The asymptotic form of the metric shows that one must choose the ratio (w/p) such that

$$w = -mp, \tag{5.15}$$

where m is the mass parameter, while the arbitrary constant figuring in Eq. (5.12a) must be taken to be equal to p^{-2}. Substituting Eq. (5.15) into (5.10) one obtains

$$w = (a^2-m^2-b^2)^{1/2}, \tag{5.16}$$

where

$$a \equiv qm, \quad b \equiv lm. \tag{5.17}$$

Finally, by introducing the coordinates (r,θ) via the relations

$$wx = r-m, \quad y = \cos\theta \tag{5.18}$$

one obtains the Boyer–Linquist form of the Kerr–NUT metric, whereby a and b are seen to stand for the angular momentum and NUT parameter, respectively.

The derivation of the $a^2 > m^2$ Kerr–NUT solution along the lines described above was obtained by BZ as one of the first applications of their ISM.[1]

(ii) Gravitational solitons propagating in a Kasner universe. When the plane symmetric gauge is chosen, Eq. (4.10) gives

$$2wx = r_+ + r_-, \quad 2iwy = r_+ - r_-, \quad r_\pm \equiv [(z \pm iw)^2 - t^2]^{1/2} \tag{5.19}$$

and the metric given by Eqs. (5.11) and (5.12) represents a pair of gravitational solitons propagating along opposite directions of the z axis. The solitons converge if the Kasner background is collapsing, i.e., for $t \in (-\infty, 0)$, or diverge if the universe is expanding, i.e., for $t \in (0, \infty)$. This family of solutions was also first obtained by BZ.[1]

(iii) Cylindrical gravitational waves reflecting off the symmetry axis. In the cylindrically symmetric gauge Eq. (4.10) gives

$$2wx = r_+ + r_-, \quad 2iwy = r_+ - r_-, \quad r_\pm \equiv [(t+iw)^2 - \rho^2]^{1/2}. \tag{5.20}$$

As in the axially symmetric case, the metric is easily regularized on the axis by gauging away the constant term in the second version of Eq. (5.12d). As shown by Economou and Tsoubelis,[7] the solution that is obtained in this fashion represents a solitary gravitational wave which, having started from $\rho \to \infty$ at $t \to -\infty$, reaches near the symmetry axis $\rho = 0$ and reflects from it at $t = 0$.

In the present case one can choose the arbitrary constant in the expression for the metric coeffficient f to be different from p^{-2}. Then the axis region is characterized by an angle deficit and the solution can be interpreted as a gravitational wave interacting with a cosmic string which occupies the axis of symmetry.

The $l = 0$ subclass of cylindrically symmetric solutions given by Eqs. (5.11) and (5.12) was first obtained by Xanthopoulos[10] using a nonsolitonic technique. In fact, the whole class of solutions under consideration retains the Petrov type

D character of the Kerr–NUT metric and therefore, must be a member of the Kinnersley[13] family of solutions.

On the other hand, Letelier[3] has obtained a family of cylindrically symmetric solutions using the BZ ISM and a diagonal seed. Because the final expressions are very complicated and depend heavily on the gauge functions, it would have been very hard for us to check if the class of metrics presented above is contained in the Letelier family of solutions.

2. Real poles

As noted in Sec. IV A, the metric coefficients for the $w_k \in \mathbb{R}$ case are obtained from the ones corresponding to $w_k \in \mathbb{C}$ by the substitution $(x,w,p) \to (ix, -iw, -ip)$. Thus when $w_k \in \mathbb{R}$, Eq. (4.10) and its inverse read as

$$\alpha/\sqrt{\epsilon} = w[\epsilon(1-x^2)(1-y^2)]^{1/2}, \quad \beta = wxy \tag{5.21}$$

and

$$\begin{aligned} &2wx = r_+ + r_-, \quad 2wy = r_+ - r_-, \\ &r_\pm \equiv [(\beta \pm w)^2 - \alpha^2]^{1/2}, \end{aligned} \tag{5.22}$$

respectively. Similarly, Eq. (5.12) becomes, in this case,

$$f = \text{const}[X/(x^2-y^2)], \tag{5.23a}$$

$$X = (l-qy)^2 + (1-px)^2, \tag{5.23b}$$

$$Y = p^2(x^2-1) + q^2(y^2-1), \tag{5.23c}$$

$$\begin{aligned} (\omega/2w) = &-(pY)^{-1}[q(1-px+l^2-lq)(1-y^2) \\ &+ lp^2(1-y)(1-x^2)] - p^{-1}(q-l). \end{aligned} \tag{5.23d}$$

The line element given by Eqs. (5.11) and (5.23) corresponds to the following classes of space-times.

(i) The Kerr–NUT $m^2 > a^2$ solutions. This class of solutions is obtained in the axisymmetric gauge ($\epsilon = -1$). Let

$$\alpha/i = \rho = w[(x^2-1)(1-y^2)]^{1/2}, \quad \beta = z = wxy \tag{5.24}$$

and

$$w = (m^2 - a^2 + b^2)^{1/2}, \tag{5.25}$$

with a, b as in Eq. (5.17). Choosing the arbitrary constant that figures in Eq. (5.23a) to be equal to p^{-2} again, we have in Eqs. (5.11) and (5.23) the Kerr–NUT $m^2 > a^2$ metrics either in the Weyl normal coordinates (ρ,z) or the prolate spheroidal coordinates (x,y). In the latter case the coordinate patch consists of the strip $x \in (1,\infty)$, $y \in (-1,1)$. In terms of the Boyer–Lindquist coordinates (r,θ) defined by Eq. (5.18), this strip corresponds to the region $r > r_1 \equiv m + w$, i.e., to that part of space-time that lies outside the event horizon.

(ii) Colliding plane waves. In the plane symmetric gauge Eq. (4.10) becomes

$$\alpha = t = w[(1-x^2)(1-y^2)]^{1/2}, \quad \beta = z = wxy. \tag{5.26}$$

Thus the metric given by Eqs. (5.11) and (5.23) is real in those regions of the (x,y) plane where either $|x| < 1$ and $|y| < 1$ or $|x| > 1$ and $|y| > 1$. On the other hand, according to Eq. (5.23a) these regions are bisected by straight lines along which the metric coefficient f is singular. This implies that having chosen the metric in any one of the above regions, one must determine a well-defined process of continuing it beyond the boundaries. For example, let $w = -|w|$ in Eq. (5.26) and consider the interior of the triangle defined by the points (0,0), (1,1), and $(-1,1)$ of the (x,y) plane. The corresponding region in the (t,z) plane is bounded by the lines $t = 0$ and $t = -|w| \pm z$. As shown by Chandrasekhar and Xanthopoulos,[9] if one assumes that the space-time metric in this triangular region is the one defined by Eqs. (5.11) and (5.23), with $l = 0$, then for $t < 0$, one can extend it beyond the lines $t = -|w| \pm z$ using the Khan–Penrose[14] technique. The resulting solution represents gravitational plane waves which collide at $t = -|w|$ and Eqs. (5.11) and (5.23) give the metric in the region of the waves' interaction.

Exactly in the same way, one can extend the general $l \neq 0$ metric and verify that the resulting metric again represents collision of gravitational plane waves. We note here that this metric can be obtained from the $l = 0$ case by an application of the hyperbolic version of Ehlers transformation. In fact, Ernst–Garcia–Hauser[15] (EGH) have recently obtained new solutions by applying this transformation to some known colliding wave metrics, including the Chadrasekhar–Xanthopoulos,[9] Nutku–Halil,[16] and Ferrari–Ibanez–Bruni[17] solutions, which can all be generated from an appropriate Kasner seed metric using the BZ soliton technique. However, as pointed out by Letelier[3] and manifest in our case, this transformation is built into the BZ method and if one considers the general solutions one immediately covers the EGH generalizations.

(iii) Cylindrical waves. In the cylindrically symmetric gauge Eqs. (5.21) gives

$$\alpha = \rho = w[(1-x^2)(1-y^2)]^{1/2}, \quad \beta = t = wxy. \tag{5.27}$$

Again one is restricted to regions where either $|x| < 1$ and $|y| < 1$ or $|x| > 1$ and $|y| > 1$. In terms of the (t,ρ) coordinates, the inverse of Eq. (5.27), which reads as

$$\begin{aligned} &2wx = r_+ + r_-, \quad 2wy = r_+ - r_-, \\ &r_\pm \equiv [(t \pm w)^2 - \rho^2]^{1/2}, \end{aligned} \tag{5.28}$$

shows that the solution given by Eqs. (5.11) and (5.23) is valid only in the three disconnected regions I, II, and III bounded by the symmetry axis $\rho = 0$ and the lines $t = |w| + \rho$, $t = \pm|w| \pm \rho$, and $t = -|w| - \rho$, respectively. In each of the regions I–III we have a metric representing cylindrical gravitational waves since the metric is time-dependent and cylindrically symmetric. However, one has to determine the fashion in which the metric extends beyond the $|x| = |y|$ lines before one has a clear picture of the physical interpretation of the line element given by Eqs. (5.11) and (5.23) in the cylindrically symmetric gauge. Work by the present authors regarding this point is under progress.

B. A pair of double poles

Starting with the same seed metric that was used in Sec. V A, a whole family of new solutions is obtained by simply assuming that the poles w_1 and w_2 are now double poles. This follows from the fact that in this case the results of Sec. IV B apply, whereby two more parameters enter the picture, namely ξ_1 and ξ_2. Just as the Q_k's of Eq. (5.3), these parameters must be chosen to be either real or complex conjugate when the poles (w_1,w_2) are real or complex conjugate, re-

spectively, because otherwise the product metric coefficients will not be real.

Let it be noted that according to Eqs. (4.15), the elements of the pertinent $[4\times4]$ matrices $\{E_{(\delta)}\}$ have already been expressed in terms of the known functions (σ_1,σ_2) and the four parameters (Q_1,Q_2,ξ_1,ξ_2). Therefore, the calculation of the metric coefficients corresponding to the four-soliton case at hand is a matter of straightforward, if tedious, algebra. Since the resulting expressions for the intermediate L functions are very lengthy, we restrict ourselves to presenting only the product metric coefficients. As in the simple poles case presented in Sec. V A, the double-poles' solutions split into two branches corresponding to the w_k's being real or complex conjugate, respectively, as follows.

1. Complex conjugate poles

When $w_k\in\mathbb{C}$, the product metric is given by Eq. (5.11) where, now,

$$f=\text{const}[X/(x^2+y^2)^4]\,, \tag{5.29a}$$

$$Y=E^2+D^2\,, \tag{5.29b}$$

$$X=F^2+G^2\,, \tag{5.29c}$$

$$(\omega/4w)=(FH+GR)/Y+\text{const}\,, \tag{5.29d}$$

$$E\equiv p^2(x^2+1)^2-q^2(y^2-1)^2 - (r^2+s^2)(1+p^2)(x^2+y^2)^2\,, \tag{5.29e}$$

$$D\equiv 2[(x^2+1)(y^2-1)]^{1/2}\{-pq(x^2+y^2)+(y^2-x^2) \times(qr-lps)-2zy(qs+lpr)\}\,, \tag{5.29f}$$

$$F\equiv(x^2+1)(1-px)^2-(y^2-1)(l-qy)^2 - (1+p^2)(x^2+y^2)\{1+(r^2+s^2)(x^2+y^2) + 2(rx+sy)\}\,, \tag{5.29g}$$

$$G\equiv 2(x^2+1)\{[(p+r)x+sy](l-qy) + p(ry-sx)(q-ly)\} + 2(y^2-1)\{[(q+ls)y+lrx](1-px) + q(ry-sx)(p+x)\}\,, \tag{5.29h}$$

$$H\equiv q(p+x)(y^4-1)-p(q-ly)(x^4-1) - (x^2+y^2)\{[r(q-ly)+sp(qy-l)](x^2+1) + [qr(1-px)-ls(x+p)](y^2-1)\}\,, \tag{5.29i}$$

$$R\equiv py(x^4-1)-qlx(y^4-1) + 2yx[p^2(x^2+1)+q^2(y^2-1)] + (1+p^2)(x^2+y^2)[ry(x^2+1)+sx(y^2-1)]\,, \tag{5.29j}$$

and the real parameters r and s stand for the combinations

$$r=(w/2)(\xi_1+\xi_2)\,, \tag{5.30a}$$

$$s=(w/2i)(\xi_1-\xi_2)\,. \tag{5.30b}$$

Except for very particular choices for the values of the parameters involved, the metric coefficients given by Eqs. (5.29) share with their two-soliton analogs the same behavior on the (t,ρ) or (x,y) plane; therefore, the physical interpretation of the latter as described in Sec. V A *1* applies here as well. However, in the present case, the corresponding space-time structure is much richer than the one found in the two-soliton solutions. The solutions belonging to the axisymmetric gauge, for example, give the analog of the Kerr $a^2>m^2$ metric for the Kinnersley–Chitre[11] class of stationary axially symmetric space-times discussed below. However, a detailed analysis of the above solutions is required before an exact physical interpretaion is put forward: Since the same is true for the solutions that follow, it should be obvious that such an analysis cannot be presented in the context of the present paper.

2. Real poles

The metric coefficients of the four-soliton solution that results from the choice $w_k\in\mathbb{R}$ can be obtained from those corresponding to the case $w_k\in\mathbb{C}$ by the substitution

$$(x,w,p,r,s)\to(ix,-iw,-ip,-ir,-s)\,. \tag{5.31}$$

Equation (5.31) is a consequence of the pertinent formulas and a simple extension of the analogous result obtained in the two-soliton case. However, since no complete list of these coefficients has been published thus far, we prefer to give them here explicitly. Again, the line element has the form given by Eq. (5.11) where, now,

$$f=\text{const}[X/(x^2-y^2)^4]\,, \tag{5.32a}$$

$$Y=E^2-D^2\,, \tag{5.32b}$$

$$X=F^2+G^2\,, \tag{5.32c}$$

$$(\omega/4w)=-(FH+GR)/Y+\text{const}\,, \tag{5.32d}$$

$$E\equiv -p^2(x^2-1)^2-q^2(y^2-1)^2 + (r^2-s^2)(1-p^2)(x^2-y^2)^2\,, \tag{5.32e}$$

$$D\equiv 2[(x^2-1)(1-y^2)]^{1/2}\{pq(x^2-y^2)+(y^2+x^2) \times(qr+lps)-2xy(qs+lpr)\}\,, \tag{5.32f}$$

$$F\equiv -(x^2-1)(1-px)^2-(y^2-1) \times(l-qy)^2+(1-p^2)(x^2-y^2) \times\{1+(r^2-s^2)(x^2-y^2)+2(rx-sy)\}\,, \tag{5.32g}$$

$$G\equiv -2(x^2-1)\{[(p+r)x-sy] \times(l-qy)-p(ry-sx)(q-ly)\} + 2(y^2-1)\{[(q-ls)y+lrx](1-px) + q(ry-sx)(x-p)\}\,, \tag{5.32h}$$

$$H\equiv q(x-p)(y^4-1)+p(q-ly)(x^4-1) + (x^2-y^2)\{[r(q-ly)-sp(qy-l)](x^2-1) + [qr(1-px)-ls(x-p)](1-y^2)\}\,, \tag{5.32i}$$

$$R\equiv -py(x^4-1)-qlx(y^4-1) + 2yx[p^2(x^2-1)+q^2(y^2-1)] - (1-p^2)(x^2-y^2)[ry(x^2-1)-sx(y^2-1)]\,. \tag{5.32j}$$

Depending on the gauge, the following classes of solutions can be distinguished.

(i) In the axisymmetric gauge Eqs. (5.32) give the Kinnersley–Chitre[11] class of metrics: The latter represents a

two-parameter generalization of the $\delta = 2$ Tomimatsu–Sato[12] class and was discovered using the symmetry transformations that leave the field equations invariant. Our method of deriving the same class of solutions verifies the Tomimatsu conjecture[8] that the Kinnersley–Chitre[11] metrics should be the product of letting the two poles that appear in the derivation of the $m^2 > a^2$ Kerr metric via the BZ technique to become double. Moreover, our method makes explicit a particular advantage of the BZ technique over the one used by Kinnersley and Chitre. This consists of the fact that the BZ method leads directly to all the components of the product metric, while the Kinnersley–Chitre method leads to the Ernst potential, which implies that some integrations must be performed before the metric is specified completely.

(ii) In the cylindrically symmetric gauge Eqs. (5.32) represent "cylindrical waves."

(iii) In the plane symmetric gauge Eqs. (5.32) represent "interacting plane waves." The necessity of the quotation marks derives, in cases (i) and (ii) from the fact that the corresponding solutions are valid in disconnected spacetime regions that are separated from each other by the null hypersurfaces $x^2 = y^2$ along which f is singular. Therefore, no claim to a concrete physical interpretation can be substantiated before any one of the above regions is appropriately extended. Given that Chandrasekhar and Xanthopoulos[9] have already shown that such an extension is not possible in the plane-waves version of the Tomimatsu–Sato[12] solution using the well-known Khan–Penrose technique, it seems that the five-parameters family presented above cannot fare better.

ACKNOWLEDGMENT

We are indebted to Professor Basilis C. Xanthopoulos for a critical reading of the manuscript and suggestions for its improvement.

[1]V. A. Belinsky and V. E. Zakharov, Zh. Eksp. Teor. Fiz. **75**, 1935 (1978) [Sov. Phys. JETP **48**, 985 (1978)]; Zh. Eksp. Teor. Fiz. **77**, 3 (1979) [Sov. Phys. JETP **50**, 1 (1979)].
[2]R. T. Jantzen, Nuovo Cimento **59**, 287 (1980); Gen. Rel. Grav. **15**, 115 (1983).
[3]P. S. Letelier, J. Math. Phys. **25**, 2675 (1984); J. Math. Phys. **26**, 467 (1985); Class. Quant. Grav. **2**, 419 (1985); Nuovo Cimento **97**, 1 (1987) and references therein.
[4]P. S. Letelier and S. R. Oliveira, J. Math Phys. **28**, 165 (1987).
[5]B. J. Carr and E. Verdaguer, Phys. Rev. D **28**, 2995 (1983).
[6]J. Cespedes and E. Verdaguer, Phys. Rev. D **36**, 2259 (1987).
[7]A. Economou and D. Tsoubelis, Phys. Rev. D **38**, 498 (1988).
[8]A. Tomimatsu, Prog. Theor. Phys. Lett. **63**, 1054 (1980).
[9]S. Chadrasekhar and B. C. Xanthopoulos, Proc. R. Soc. London Ser. A **408**, 175 (1986).
[10]B. C. Xanthopoulos, Phys. Lett. B **178**, 163 (1986); Phys. Rev. D **34**, 3608 (1986).
[11]W. Kinnersley and D. M. Chitre, Phys. Rev. Lett. **40**, 1608 (1978); J. Math. Phys. **19**, 2037 (1978).
[12]A. Tomimatsu and H. Sato, Phys. Rev. Lett. **29**, 1344 (1972); Prog. Theor. Phys. **50**, 95 (1973).
[13]W. Kinnersley, J. Math. Phys. **10**, 1195 (1986).
[14]K. A. Khan and R. Penrose, Nature **229**, 185 (1971).
[15]F. J. Ernst, A. Garcia, and I. Hauser, J. Math. Phys. **28**, 2155 (1987).
[16]Y. Nutku and M. Halil, Phys. Rev. Let. **39**, 1379 (1977).
[17]V. Ferrari, J. Ibanez, and M. Bruni, Phys. Rev. D **36**, 1053 (1987); Phys. Lett. A **122**, 459 (1987).

Higher-dimensional unification by isometry

Y. M. Cho[a)]
Department of Physics, Yale University, New Haven, Connecticut 06511
D. S. Kimm[b)]
Department of Physics, Seoul National University, Seoul 151, Korea

(Received 17 November 1987; accepted for publication 25 January 1989)

A unified theory based on a homogeneous fiber bundle $Q(M,G/H)$ is discussed in detail. In spite of the fact that the theory retains the full G-gauge invariance, the physical gauge group K is shown to be $K = H^*/(H \cap H^*)$, where H^* is the centralizer of H in G. A principal fiber bundle $P(M,G,H)$ is also constructed by introducing an additional left action H on $P(M,G)$ that commutes with the right action G, and a unified theory based on $P(M,G,H)$ is discussed. It is shown that the theory based on $Q(M,G/H)$ is, in fact, the H projection of the Einstein–Hilbert action from $P(M,G,H)$, with the identification $Q(M,G/H) = P(M,G,H)/H$.

I. INTRODUCTION

It has long been recognized that[1,2] the gauge theory and gravitation could be unified into an Einstein–Hilbert action in a higher-dimensional space which unifies the four-dimensional space-time with an n-dimensional internal space. In this unification the gauge symmetry emerges from the isometry of the unified metric. In a prototype unification[2,3] where the Killing vector fields of the isometry G are linearly independent, the isometry makes the unified space a principal fiber bundle[4] $P(M,G)$ with the space-time M as the base manifold and G as the structure group. In this case the isometry becomes the physical gauge symmetry of the unified theory.

In the general case when the Killing vector fields of the isometry G are *not* linearly independent, the unified space has the structure of a homogeneous fiber bundle $Q(M,G/H)$.[5] In this case questions arise of how the isometry restricts the metric and the curvature on Q, how the Einstein–Hilbert action on Q can be reduced to a unified action on M, and what is the resulting gauge symmetry of the theory. These questions have been investigated recently.[5,6] In this paper we discuss them in more detail and compare with other attempts in the literature.[7,8] We show that, even though the isometry makes the theory gauge invariant under G, it restricts the physical gauge group (the holonomy group) K to be $K = H^*/(H \cap H^*)$, where H^* is the commutant subgroup (the centralizer) of H in G.

Other purposes of the paper are to construct a fiber bundle, which we denote by $P(M,G,H)$, by introducing a left action[5,9] H on $P(M,G)$ that commutes with the right action G, and to discuss the unified theory based on $P(M,G,H)$. The theory is interesting in its own right. But perhaps more importantly the left isometry provides us with a better understanding of the geometry of $P(M,G)$ and $Q(M,G/H)$. First, it gives a natural homomorphism between $P(M,G)$ and $Q(M,G/H)$, because $P(M,G,H)$ can also be identified as a principal fiber bundle $P(Q,H)$ with $Q(M,G/H) = P/H$. So it provides an alternative method to obtain the unified theory based on $Q(M,G/H)$. In fact, we show that theory based on $Q(M,G/H)$ is nothing but the H projection of the one obtained from $P(M,G,H)$. Another motivation behind the left isometry is that it gives us a natural tool to study the non-Abelian monopoles[9] of $P(M,G)$. The left isometry demands the holonomy group of the connection on $P(M,G)$ to be H^*. But when the second homotopy $\pi_2(G/H)$ defined by the left isometry becomes nonzero, the gauge potential becomes dual, capable of describing both electric and magnetic charges of H^*. In fact, choosing H to be Cartan's subgroup of G (in which case H^* coincides with H), one can describe all possible magnetic charges of $P(M,G)$. Furthermore this observation, together with the fact that the connection space forms an affine space, allows us to express the most general gauge potential on $P(M,G)$ as the sum of a dual potential of $H^* = H$ and a gauge-covariant vector field which has no neutral component (the valence potential). With this gauge-independent decomposition of the potential into the dual part and the valence part, one can construct the most general nontrivial non-Abelian gauge theory.[10,11]

An attractive aspect of our unification is that it provides a simple and consistent method of dimensional reduction. A central issue in any (supersymmetric or not) higher-dimensional unification is how to reduce the theory to a four-dimensional effective theory. So far a popular method of dimensional reduction has been the zero-mode approximation[12] of the harmonic expansion, obtained after a spontaneous compactification[13] of the internal space. Unfortunately this approximation bears many undesirable features: a logical ambiguity on the definition of the zero-modes due to the possibility of a spontaneous symmetry breaking among them,[14] the consistency problem,[15,16] the problem of quantum stability,[17] and others. Our approach provides an alternative method of dimensional reduction free of these undesirable features. In our case the dimensional reduction is obtained by the isometry or, in general, by the right invariance[18] when the matter fields are present, which automatically reduces the higher-dimensional fields to a *finite* number of physical modes whose internal space dependence is completely fixed. So there is no need of a spontaneous compactification[14] or a harmonic expansion. More importantly, as long as the isometry remains rigid against quantum fluctuations, the geometry will not only exclude any higher modes but also precludes any intrinsic inconsistency.

[a)] On leave from Department of Physics, Seoul National University, Seoul 151, Korea.
[b)] Present address: Department of Physics, Kangnung National University, Kangnung 210, Korea.

 0022-2488/89/071570-09$02.50

The paper is organized as follows. In Sec. II we give a brief review of the higher-dimensional unification based on the principal fiber bundle $P(M,G)$ for later convenience. In Sec. III the unified theory based on $Q(M,G/H)$ is discussed in detail. In Sec. IV we discuss the consistency problem of a dimensional reduction and compare our method with others. In Sec. V we construct a principal fiber bundle $P(M,G,H)$ by introducing a left isometry H on $P(M,G)$. In Sec. VI we discuss a unified theory based on $P(M,G,H)$, and show that the theory based on $Q(M,G/H)$ is the H projection of the one obtained from $P(M,G,H)$. Finally in the last section we compare our unification with others and discuss physical implications of our results.

II. UNIFIED THEORY ON $P(M,G)$: A BRIEF REVIEW

We start from a unified space P that has the following properties.[2,3]

(i) P is a $(4+n)$-dimensional metric manifold with the metric g_{AB}.

(ii) There exist n linearly independent Killing vector fields ξ_a $(a=1,2,...,n)$ which form an isometry group G with the following commutation relations:

$$\mathcal{L}_{\xi_a} g_{AB}=0, \quad [\xi_a,\xi_b]=(1/\kappa) f_{ab}{}^{c}\xi_c , \tag{1}$$

where $\mathcal{L}_{\xi_a}$ is the Lie derivative along ξ_a, and κ is a scale parameter. We further require G to be unimodular for the reason that will become clear in the following.

(iii) The integral manifold of the Killing vectors is a metric submanifold, i.e., the metric

$$\phi_{ab}=g_{AB}\xi_a^{A}\xi_b^{B} \tag{2}$$

is invertible.[19] Notice that since the Killing vectors define an n-dimensional involutive distribution[4] on P, they admit a unique maximal integral manifold by virtue of the Frobenius theorem.

Now, let M be P/G and π be the canonical projection from P to M. Then one may view the unified space P as a principal fiber bundle $P(M,G)$ and the Killing vector fields as the fundamental vector fields that generate the right action[4] of G on P. We will identify M as the physical space-time and the vertical fiber (the integral manifold of the Killing fields) as the internal space.

The existence of the metric g_{AB} allows us to define the horizontal subspace H_p of the tangent space $T_p(P)$ at each $p\in P$ as the horizontal complement of the Killing vectors with respect to the metric. By virtue of the Killing symmetry H_p will be invariant under the right action G. Now, let U be an open neighborhood of $x\in M$, ∂_μ $(\mu=1,2,3,4)$ a local coordinate basis on U, and D_μ the horizontal lift of ∂_μ on $\pi^{-1}(U)$. Clearly $D_\mu \otimes \xi_a$ serves a basis on $\pi^{-1}(U)$. In this horizontal lift basis[2] the metric g_{AB} should become block diagonal:

$$g_{AB}=\left(\begin{array}{c|c} g_{\mu\nu} & 0 \\ \hline 0 & \phi_{ab} \end{array}\right). \tag{3}$$

In the following we will identify $g_{\mu\nu}$ as the physical metric on M up to a conformal transformation.

Let σ be a local cross section[4] in $\pi^{-1}(U)$ [i.e., a smooth mapping from $x\in U$ to $\sigma(x)\in\pi^{-1}(U)$ such that $\pi(\sigma(x))=x$], and let $\partial_\mu \otimes \xi_a$ be the local direct product basis[2] on $\pi^{-1}(U)\simeq U\times G$ defined by this local trivialization. Clearly this basis has the following commutation relations:

$$\begin{aligned} &[\partial_\mu,\partial_\nu]=0, \\ &[\partial_\mu,\xi_a]=0, \\ &[\xi_a,\xi_b]=(1/\kappa) f_{ab}{}^{c}\xi_c . \end{aligned} \tag{4}$$

Now, from the concept of the horizontality the definition of the gauge potential follows. Since the $\mathcal{G}$-valued ($\mathcal{G}$ is the Lie algebra of G) connection one-form on P defined by H_p is nothing but the dual one-form ω^a of ξ_a, one can define the gauge potential $A_\mu{}^a$ by[2]

$$A_\mu^{\,a}=(1/e\kappa)\omega_A^a\,\partial_\mu^A , \tag{5}$$

where e is a coupling constant. So the choice of a local cross section amounts to the choice of a local gauge. From (5) one has

$$D_\mu=\partial_\mu - e\kappa A_\mu{}^{a}\xi_a , \tag{6}$$

so that g_{AB} has the following expression in the local direct product basis[2]:

$$g_{AB}=\left(\begin{array}{c|c} g_{\mu\nu}+e^2\kappa^2\phi_{ab}A_\mu{}^{a}A_\nu{}^{b} & e\kappa A_\mu{}^{a}\phi_{ab} \\ \hline e\kappa\phi_{ab}A_\nu{}^{b} & \phi_{ab} \end{array}\right). \tag{7}$$

From (4) one obtains the following commutation relations for the horizontal lift basis:

$$\begin{aligned} &[D_\mu,D_\nu]=-e\kappa F_{\mu\nu}{}^{a}\xi_a , \\ &[D_\mu,\xi_a]=0, \\ &[\xi_a,\xi_b]=(1/\kappa) f_{ab}{}^{c}\xi_c , \end{aligned} \tag{8}$$

where $F_{\mu\nu}{}^{a}$ is the field strength of the potential $A_\mu{}^{a}$,

$$F_{\mu\nu}{}^{c}=\partial_\mu A_\nu{}^{c}-\partial_\nu A_\mu{}^{c}+e f_{ab}{}^{c}A_\mu{}^{a}A_\nu{}^{b} .$$

This implies that the horizontal subspace H_p can be integrable if and only if the field strength vanishes.

Notice that the isometry (1) requires the metric g_{AB} to be *right invariant*,[3,18] and determines its internal space dependence completely:

$$\begin{aligned} &\partial_a g_{\mu\nu}=0, \\ &\partial_a A_\mu{}^{c}=-(1/\kappa) f_{ab}{}^{c}A_\mu{}^{b}, \\ &\partial_a \phi_{bc}=(1/\kappa) f_{ab}{}^{d}\phi_{dc}+(1/\kappa) f_{ac}{}^{d}\phi_{bd} , \end{aligned} \tag{9}$$

where we have put $\partial_a=\xi_a$. Now one can calculate the scalar curvature R of P. Assuming no torsion one finds[2]

$$\begin{aligned} R_P=R_M+R_G+(e^2\kappa^2/4)\phi_{ab}F_{\mu\nu}{}^{a}F_{\mu\nu}{}^{b} \\ +\tfrac{1}{4}\phi^{ab}\phi^{cd}[(D_\mu\phi_{ac})(D_\nu\phi_{bd})+(D_\mu\phi_{ab})(D_\mu\phi_{cd})] \\ +\nabla_\mu(\phi^{ab}D_\mu\phi_{ab}) , \end{aligned} \tag{10}$$

where R_M and R_G are the scalar curvature of M and G, and ∇_μ is the gauge and generally covariant derivative. Notice that R_p has no fiber dependence, which is a direct consequence of the isometry (1).

To obtain the unified interaction, we start from the Einstein–Hilbert action on P,

$$I_P=-\frac{1}{16\pi G_0}\int \sqrt{g}\sqrt{\phi}(R_P+\Lambda)d^4x\,d^nG , \tag{11}$$

where G_0 and Λ are the $(4+n)$-dimensional gravitational and cosmological constants, $d^n G$ is the right-invariant Harr measure on G, and

$$g = |\det g_{\mu\nu}|, \quad \phi = |\det \phi_{ab}| .$$

Now, when G is unimodular one has

$$\partial_a \phi = \phi \times \phi^{bc} \partial_a \phi_{bc} = (2/\kappa) \phi f_{ab}{}^{b} = 0$$

so that ϕ becomes explicitly independent of the internal fiber. In this case the dimensional reduction amounts to performing the trivial integration over the fiber, after which one obtains the four-dimensional Lagrangian

$$L_P = -(\mu/16\pi G_0)\sqrt{g}\sqrt{\phi}(R_P + \Lambda) , \tag{12}$$

where μ is the right-invariant volume of G. Then, identifying G_0/μ as the four-dimensional gravitational constant G and requiring

$$e^2\kappa^2/16\pi G = 1 , \tag{13}$$

one obtains the desired unification.[2,3]

One can simplify the Lagrangian (12) further by putting $\phi_{ab} = \phi^{1/n}\rho_{ab}$ ($|\det \rho_{ab}| = 1$). Removing a total divergence one finds

$$L = -\frac{1}{16\pi G}\sqrt{g}\sqrt{\phi}\Big[R_M + R_G + 4\pi G \phi^{1/n}\rho_{ab}F_{\mu\nu}{}^{a}F_{\mu\nu}{}^{b} - \frac{n-1}{4n}\frac{(\partial_\mu \phi)^2}{\phi^2} + \frac{1}{4}\rho^{ab}\rho^{cd}(D_\mu \rho_{ac})(D_\mu \rho_{bd}) + \lambda(|\det \rho_{ab}| - 1)\Big] ,$$

where λ is introduced as a Lagrange multiplier. But now the Lagrangian appears *unstable* due to the negative kinetic term of the ϕ field. This defect, however, is superficial and can easily be removed by reparametrizing the fields. To see this we make the conformal transformation

$$g_{\mu\nu} \to \sqrt{\phi} g_{\mu\nu}$$

and find L is given, in terms of the *new* metric, by[18] (up to a total divergence)

$$L = -\frac{1}{16\pi G}\sqrt{g}\Big[R_M - \exp\Big(-\sqrt{\frac{n+2}{n}}\sigma\Big)\hat{R}_G + \exp\Big(-\sqrt{\frac{n}{n+2}}\sigma\Big)\Lambda + 4\pi G \exp\Big(\sqrt{\frac{n+2}{n}}\sigma\Big)\rho_{ab}F_{\mu\nu}{}^{a}F_{\mu\nu}{}^{b} + \frac{1}{2}(\partial_\mu \sigma)^2 + \tfrac{1}{4}(D_\mu \rho^{ab})(D_\mu \rho_{ab}) + \lambda(|\det \rho_{ab}| - 1)\Big] , \tag{14}$$

where $\hat{R}_G = R_G(\rho_{ab})$ and we have introduced the dilaton field σ by

$$\sigma = \tfrac{1}{2}\sqrt{(n+2)/n}\,\log \phi .$$

This suggests that one should treat the new metric, but not the old one, as the physical space-time metric. There are three important aspects of the unified Lagrangian worth mentioning. First the gravitational coupling (i.e., the Newton's coupling) G_N of the Kaluza–Klein gauge field is given by

$$G_N = Ge^{-c\sigma} ,$$

where $c = -\sqrt{(n+2)/n}$. In general, in the presence of matter fields one can show that[18] the value of the constant c depends on what kind of matter field one has, but the modification of the gravitational coupling by the dilaton is a generic feature of the higher-dimensional unification. Second, the dilaton acquires a nontrivial potential determined by $\hat{R}_G$ and Λ. Finally the dynamics of the internal metric ρ_{ab} is described by a generalized nonlinear sigma model, with the minimal gauge coupling to $A_\mu{}^a$ and the self-interaction potential specified by $\hat{R}_G$.

III. $Q(M,G/H)$

Notice that on $P(M,G)$ the isometry G acts freely on P, which restricts the internal space to be isomorphic to G itself. In general, however, one would like the internal space to be a homogeneous space G/H on which the isometry G acts effectively but not necessarily freely. Assuming that the isometry acts transitively on the internal space, this would be the most general type of isometry one can impose on the unified space. Under this circumstance the unified space becomes a homogeneous fiber bundle $Q(M,G/H)$ rather than a principal one. The problem then is to find how the isometry reduces the metric on Q down to four-dimensional fields, and what is the resulting unified theory. In this section we resolve this problem.[5,17]

Let Q be the $(4+m)$-dimensional unified space ($m = \dim G/H$) which admits an n-dimensional isometry G with the Killing fields h_a ($a = 1,2,...,n$):

$$\mathcal{L}_{h_a} g_{AB} = 0 , \quad [h_a, h_b] = (1/\kappa) f_{ab}{}^{c} h_c . \tag{15}$$

Also let π be the projection of Q to $M = Q/G$, U a neighborhood of $x \in M$, and σ a local cross section on $\pi^{-1}(U)$. With this local trivialization we introduce local coordinates $(x^\mu, y^{\underline{a}})$ ($a = 1,2,...,m$), which are the direct product of the space-time coordinates x^μ of M and the internal coordinates $y^{\underline{a}}$ of G/H. Then in the basis $\partial_\mu \otimes \partial_{\underline{a}}$ the metric can always be put into the following form:

$$g_{AB} = \left(\begin{array}{c|c} g_{\mu\nu} + e^2\kappa^2 g_{\underline{ab}} B_\mu{}^{\underline{a}} B_\nu{}^{\underline{b}} & e\kappa B_\mu{}^{\underline{a}} g_{\underline{ab}} \\ \hline e\kappa g_{\underline{ab}} B_\nu{}^{\underline{b}} & g_{\underline{ab}} \end{array}\right). \tag{16}$$

Now with

$$h_a = \partial_a = h_a{}^{\underline{b}}\partial_{\underline{b}} ,$$
$$[h_a, \partial_{\underline{b}}] = F_{a\underline{b}}{}^{\underline{c}} h_{\underline{c}} = -(\partial_{\underline{b}} h_a{}^{\underline{b}})\partial_{\underline{c}} , \tag{17}$$

one finds the following expression for the Killing condition (15):

$$\partial_a \phi_{\underline{bc}} = F_{a\underline{b}}{}^{\underline{d}}\phi_{\underline{dc}} + F_{a\underline{c}}{}^{\underline{d}}\phi_{\underline{bd}} ,$$
$$\partial_a B_\mu{}^{\underline{c}} = -F_{a\underline{b}}{}^{\underline{c}} B_\mu{}^{\underline{b}} , \quad \partial_a g_{\mu\nu} = 0 . \tag{18}$$

This is the generalization of the Killing condition (9) to $Q(M,G/H)$.

To keep the analogy between $P(M,G)$ and $Q(M,G/H)$ whenever possible, it is very useful to introduce the "dual one-form" $\phi^a = \phi_{\underline{b}}{}^{a}\, dy^{\underline{b}}$ of the Killing fields h_a by

$$\phi_{\underline{a}}{}^{c} h_c{}^{\underline{b}} = \delta_{\underline{a}}{}^{\underline{b}} , \quad \partial_a \phi_{\underline{b}}{}^{c} = -(1/\kappa) f_{ab}{}^{c}\phi_{\underline{b}}{}^{b} + F_{a\underline{b}}{}^{\underline{c}}\phi_{\underline{c}}{}^{c} . \tag{19}$$

The existence and uniqueness of such a dual one-form will be

proved in Sec. V. Now with $F_{\underline{ab}}{}^{\underline{c}} = \phi_{\underline{a}}{}^{a} F_{ab}{}^{\underline{c}}$, (18) can be written as

$$\partial_{\underline{a}} \phi_{\underline{bc}} = F_{\underline{ab}}{}^{\underline{d}} \phi_{\underline{dc}} + F_{\underline{ac}}{}^{\underline{d}} \phi_{\underline{bd}} ,$$
$$\partial_{\underline{a}} B_{\mu}{}^{\underline{c}} = -F_{\underline{ab}}{}^{\underline{c}} B_{\mu}{}^{\underline{b}}, \quad \partial_{\underline{a}} g_{\mu\nu} = 0 .$$

Notice that the first equality tells us that $F_{\underline{ab}}{}^{\underline{c}}$ is a metric connection on G/H. However, it is *not* Riemannian because it has a nonvanishing torsion $t_{\underline{ab}}{}^{\underline{c}}$,

$$t_{\underline{ab}}{}^{\underline{c}} = F_{\underline{ab}}{}^{\underline{c}} - F_{\underline{ba}}{}^{\underline{c}} = (1/\kappa) f_{ab}{}^{c} \phi_{\underline{a}}{}^{a} \phi_{\underline{b}}{}^{b} h_{c}{}^{\underline{c}} . \tag{20}$$

The torsion-free Riemannian connection $\Gamma_{\underline{ab}}{}^{\underline{c}}$ is given by

$$\Gamma_{\underline{ab}}{}^{\underline{c}} = F_{\underline{ab}}{}^{\underline{c}} - C_{\underline{ab}}{}^{\underline{c}} ,$$

where

$$C_{\underline{ab}}{}^{\underline{c}} = \tfrac{1}{2}(t_{\underline{ab}}{}^{\underline{c}} + t^{\underline{c}}{}_{\cdot\underline{ab}} + t^{\underline{c}}{}_{\cdot\underline{ba}})$$

is the contortion.

To find the curvature of the potential $B_{\mu}{}^{\underline{a}}$ let us define a horizontal basis D_{μ} by

$$D_{\mu} = \partial_{\mu} - e\kappa B_{\mu}{}^{\underline{a}} \partial_{\underline{a}} . \tag{21}$$

Clearly the metric (16) becomes block diagonal in the basis $(D_{\mu} \otimes \partial_{\underline{a}})$,

$$g_{AB} = \left(\begin{array}{c|c} g_{\mu\nu} & 0 \\ \hline 0 & g_{\underline{ab}} \end{array}\right) . \tag{22}$$

Now, in analogy with (8) we obtain

$$[D_{\mu}, D_{\nu}] = -e\kappa G_{\mu\nu}{}^{\underline{a}} \partial_{\underline{a}} ,$$
$$[\partial_{\underline{a}}, D_{\mu}] = F_{\underline{ab}}{}^{\underline{c}} B_{\mu}{}^{\underline{b}}, \quad [\partial_{\underline{a}}, \partial_{\underline{b}}] = 0 , \tag{23}$$

where the field strength $G_{\mu\nu}{}^{\underline{a}}$ of the potential $B_{\mu}{}^{\underline{a}}$ is given by

$$G_{\mu\nu}{}^{\underline{c}} = \partial_{\mu} B_{\nu}{}^{\underline{c}} - \partial_{\nu} B_{\mu}{}^{\underline{c}} + e\kappa t_{\underline{ab}}{}^{\underline{c}} B_{\mu}{}^{\underline{a}} B_{\nu}{}^{\underline{b}} .$$

Notice that the torsion determines the self-interaction of the potential.

To make the above geometry of $Q(M, G/H)$ more transparent let us define the "covariant" potential $B_{\mu}{}^{a}$ by

$$B_{\mu}{}^{a} = B_{\mu}{}^{\underline{a}} \phi_{\underline{a}}{}^{a} . \tag{24}$$

From the definition it follows that

$$G_{\mu\nu}{}^{\underline{a}} = G_{\mu\nu}{}^{a} h_{a}{}^{\underline{a}} ,$$

where $G_{\mu\nu}{}^{a}$ is the canonical field strength of $B_{\mu}{}^{a}$:

$$G_{\mu\nu}{}^{c} = \partial_{\mu} B_{\nu}{}^{c} - \partial_{\nu} B_{\mu}{}^{c} + e f_{ab}{}^{c} B_{\mu}{}^{a} B_{\nu}{}^{b} .$$

Now, in terms of the covariant potential the Killing condition (18) has the following familiar form:

$$\partial_{a} B_{\mu}{}^{c} = -(1/\kappa) f_{ab}{}^{c} B_{\mu}{}^{b}, \quad \partial_{a} G_{\mu\nu}{}^{c} = -(1/\kappa) f_{ab}{}^{c} G_{\mu\nu}{}^{b} . \tag{25}$$

Similarly one may introduce the "covariant" metric h_{ab} of G/H by

$$h_{ab} = g_{\underline{ab}} h_{a}{}^{\underline{a}} h_{b}{}^{\underline{b}}, \quad h^{ab} = g^{\underline{ab}} \phi_{\underline{a}}{}^{a} \phi_{\underline{b}}{}^{b} \tag{26}$$

and find the following Killing condition:

$$\partial_{a} h_{bc} = (1/\kappa) f_{ab}{}^{d} h_{dc} + (1/\kappa) f_{ac}{}^{d} h_{bd} ,$$
$$\partial_{a} h^{bc} = -(1/\kappa) f_{ad}{}^{b} h^{dc} - (1/\kappa) f_{ad}{}^{c} h^{bd} . \tag{27}$$

Thus, in terms of the covariant fields, the Killing condition on $Q(M, G/H)$ has exactly the same expression as the condition (9) on $P(M, G)$. However, this appearance is misleading because these "covariant" fields do *not* always represent the physical degrees of freedom. This must be obvious because, first of all, h_{ab} is singular as an $(n \times n)$ matrix. In fact the matrix $h_{a}{}^{b}$, defined by

$$h_{a}{}^{b} = h_{ac} h^{bc} = h_{a}{}^{\underline{a}} \phi_{\underline{a}}{}^{b} , \tag{28}$$

forms the projection operator from $\mathcal{G}$ to $\mathcal{G}/\mathcal{H}$ ($\mathcal{H}$ is the Lie algebra of H). Moreover one has

$$h_{ab} = h_{a}{}^{c} h_{b}{}^{d} h_{cd}, \quad B_{\mu}{}^{a} = h_{b}{}^{a} B_{\mu}{}^{b} , \tag{29}$$

so that both h_{ab} and $B_{\mu}{}^{a}$ can have only G/H degrees of freedom. This point will become important soon.

To construct the unified action one must calculate the scalar curvature R on Q. Assuming no torsion we find

$$R_{Q} = R_{M} + R_{G/H} + (e^2\kappa^2/4) h_{ab} G_{\mu\nu}{}^{a} G_{\mu\nu}{}^{b}$$
$$+ \tfrac{1}{4} h^{ab} h^{cd} [(D_{\mu} h_{ac})(D_{\mu} h_{bd})$$
$$+ (D_{\mu} h_{ab})(D_{\mu} h_{cd})]$$
$$+ \nabla_{\mu} (h^{ab} D_{\mu} h_{ab}) , \tag{30}$$

where $D_{\mu} = \partial_{\mu} - e\kappa B_{\mu}{}^{a} \partial_{a}$ is the gauge-covariant derivative defined by (21), and ∇_{μ} is the gauge and generally covariant derivative. Notice that R_{Q} is *explicitly* G invariant. The fact that R_{Q} should have a G-invariant expression is obvious from the isometry (15). To obtain the above result, however, one has to do the calculation in the basis (23) first, and then express R_{Q} in the G-invariant form. To calculate $R_{G/H}$ one can make use of the identity

$$R_{\underline{abc}}{}^{\underline{d}} \phi_{\underline{d}}{}^{d} = -(\nabla_{\underline{a}} \nabla_{\underline{b}} - \nabla_{\underline{b}} \nabla_{\underline{a}}) \phi_{\underline{c}}{}^{d}$$

and find

$$R_{\underline{ab}} = \tfrac{1}{2} \phi_{\underline{a}}{}^{a} \phi_{\underline{b}}{}^{b} f_{ac}{}^{d} f_{bd}{}^{c} + \tfrac{1}{2} t_{\underline{ac}}{}^{\underline{d}} t_{\underline{bc}}{}^{\underline{d}}$$
$$- \tfrac{1}{4} t_{\underline{cda}} t_{\underline{cdb}} + \tfrac{1}{2} (t_{\underline{cab}} + t_{\underline{cba}}) t_{\underline{cd}}{}^{\underline{d}} .$$

So, when G is unimodular, one obtains

$$R_{G/H} = (1/2\kappa^2) f_{ab}{}^{d} f_{cd}{}^{b} h^{ac} + (1/4\kappa^2) f_{ab}{}^{e} f_{cd}{}^{f} h^{ac} h^{bd} h_{ef} . \tag{31}$$

The result (30) looks exactly the same as (10), except here R_{G} is replaced by $R_{G/H}$. However, one has to keep in mind that the "covariant" fields in R_{Q} do not represent the physical degrees of freedom.

To determine the *physical* content of R_{Q}, notice first that the internal metric h_{ab} must be ad(H) invariant. This is so because any G-invariant metric on G/H has to be ad(H) invariant.[4] One can make this ad(H) invariance explicit by choosing a cross section σ_{0} in $\pi^{-1}(U)$ on which the isotropy subgroup of G becomes exactly H. Indeed on σ_{0} one has $h_{a} = 0$ when h_{a} belongs to $\mathcal{H}$, so that one finds

$$\partial_{a} h_{bc} = (1/\kappa) f_{ab}{}^{d} h_{dc} + (1/\kappa) f_{ac}{}^{d} h_{bd} = 0$$

when ∂_{a} belongs to $\mathcal{H}$. This proves the ad(H) invariance of h_{ab}. By the same token (25) tells us that $B_{\mu}{}^{a}$ (and $G_{\mu\nu}{}^{a}$) must also be ad(H) invariant. This together with (29) means that the physical gauge group K must be $K = H^{*}/(H \cap H^{*})$, where H^{*} is the centralizer (the commutant) of H in G. In other words, in terms of the covariant potential $B_{\mu}{}^{a}$ the holonomy group has to be K, but not G. Notice that K can also be expressed as $K = N/H$, where N is the normalizer of H in G.

The dimensional reduction from $Q(M, G/H)$ can be ob-

tained exactly the same way as before. One starts from the Einstein–Hilbert action on Q,

$$I_Q = -\frac{1}{16\pi G_0}\int \sqrt{g_M}\,\sqrt{g_{G/H}}\,(R_Q+\Lambda)d^4x\,d^my\,, \quad (32)$$

and notices that

$$\sqrt{g_{G/H}(x,y)}\;d^my = \sqrt{g_{G/H}(\sigma_0(x),0)}\,d^m\mu_{G/H}$$
$$= \sqrt{h(x)}\,d^m\mu_{G/H}\,,$$

where $d^m\mu_{G/H}$ is the G-invariant measure on G/H and

$$h(x) = |\det g_{ab}(\sigma_0(x))| = |\det h_{\underline{ab}}(\sigma_0(x))|\,.$$

So after the fiber integration one obtains the following four-dimensional Lagrangian (up to a total divergence):

$$L = \frac{\mu_{G/H}}{16\pi G_0}\sqrt{g_M}\,\sqrt{h}\Big[R_M + R_{G/H} + \frac{1}{4}\frac{e^2\kappa^2}{4}h_{\underline{ab}}G_{\mu\nu}{}^{\underline{a}}G_{\mu\nu}{}^{\underline{b}}$$
$$+\frac{1}{4}h^{\underline{ab}}h^{\underline{cd}}[(D_\mu h_{\underline{ac}})(D_\mu h_{\underline{bd}})$$
$$-(D_\mu h_{\underline{ab}})(D_\mu h_{\underline{cd}})] + \Lambda\Big]\,, \quad (33)$$

where now the gauge group is restricted to K. There are two things to be noticed here. First, when H becomes the identity subgroup, the Lagrangian becomes exactly identical to (12), the one we obtained from $P(M,G)$. Second, the Lagrangian is explicitly invariant under K, because the choice of the cross section σ_0 still leaves the K-gauge degrees of freedom completely arbitrary. However, at a first glance the above dimensional reduction appears to depend on the choice of a cross section. Now we prove the σ independence of the dimensional reduction. To do this notice that under an infinitesimal change of the cross section from $\sigma_0(x)$ to $\sigma(x)$ generated by $\delta y^{\underline{a}}(x)$, one has

$$\delta h(x) = \delta y^{\underline{a}}\partial_{\underline{a}}h(x)|_{\sigma_0} = (2/\kappa)h(x)\delta y^{\underline{a}}f_{\underline{ab}}{}^{\underline{b}}$$
$$= (2/\kappa)h(x)\delta y^{\underline{a}}f_{\underline{ab}}{}^{b}\,,$$

where the last equality follows from (19). Clearly this (together with the fiber independence of R_Q) tells us that, when G is unimodular, the reduction procedure is independent of the choice of a cross section.

IV. CONSISTENCY OF DIMENSIONAL REDUCTION

A central issue in any (supersymmetric or not) higher-dimensional unification is how to perceive the extra dimension. On this issue there are two different points of view. So far the popular view has been to treat the full $(4+n)$-dimensional space as physical, as is done in supersymmetric Kaluza–Klein unification.[12] Here the dimensional reduction is regarded as a low-energy approximation of the full theory, which one obtains by keeping only the "zero-modes" of the harmonic expansion after a spontaneous compactification[13] of the internal space. The alternative view is to treat only the four-dimensional space as physical, which one can do by imposing an exact isometry[3,6] as we did in this paper. In this view the dimensional reduction is not an approximation but an inevitable consequence of the isometry. No matter how one regards the dimensional reduction, however, a logical consistency requires that the resulting four-dimensional theory should remain compatible with the higher-dimensional theory. A minimum requirement of the consistency[15,16] is that the solutions of the four-dimensional effective theory must remain solutions of the higher-dimensional equations of motion obtained before the dimensional reduction. We will call a dimensional reduction procedure consistent if it satisfies this criterion.

One can easily show that the dimensional reduction by isometry described in the previous sections is consistent. In fact the four-dimensional equations of motion obtained from (14) or (33) become exactly identical to the higher-dimensional Einstein equations on $P(M,G)$ or $Q(M,G/H)$. The equivalence follows from the fact that when G is unimodular the action integrals before and after the dimensional reduction become equivalent to each other, as far as the variation of the action integral is concerned. This is so because when G is unimodular, the integral over the fiber does not involve averaging out the fiber dependence of the fields. So one can obtain the Euler–Lagrange equations either before the fiber integration or after, with the same result. Thus in our case the consistency is built in by the geometry. However, it must be emphasized that the consistency is guaranteed *only if G is unimodular.* In fact, one can easily show by constructing explicit examples[20] that the isometry alone is not sufficient to guarantee the consistency of the dimensional reduction.

Now we wish to make a few comments. First, when the matter fields are present the isometry of the metric should be generalized to the right invariance[18] (or the invariance under the right action of G) of *all* fields, including the matter fields. The right invariance will then determine the fiber dependence of fields uniquely and give us a consistent dimensional reduction. Another point is that, to apply our dimensional reduction method, we need to specify not just the internal space G/H, but both G and H. This is so because a given homogeneous space may admit more than one transitively acting group. For instance, S^7 topology has four transitively acting groups,[16] and can be identified as one of the following: $SO(8)/SO(7)$, $SO(7)/G_2$, $SU(4)/SU(3)$, or $SO(5)/SO(3)$. So for the 11-dimensional supergravity the physical gauge group K resulting from our dimensional reduction method could be either identity, $U(1)$, or $SU(2)$, depending on which G/H one chooses. In general, for a given internal space one should choose the smallest isometry G to obtain the largest K. Finally, we emphasize that our dimensional reduction does *not* require the internal space to be compact, because the reduction is not based on a harmonic expansion. In fact in our approach one can easily construct a well-defined unified theory with a noncompact internal space.[14] In this respect we remark that our dimensional reduction is more general than the one proposed by some authors[8] recently. In their reduction the compactness of G/H has been a prerequisite for a consistent dimensional reduction. We do not require this. In contrast we require the unimodularity of the isometry as a necessary condition for a consistent dimensional reduction. This requirement has been absent in their reduction.

At this point it is perhaps instructive to compare our dimensional reduction method in more detail with the popular one widely accepted in supersymmetric Kaluza–Klein

unification.[12] In this approach the dimensional reduction is regarded as a "zero-modes" approximation of the full theory which one obtains after a spontaneous compactification[13] of the internal space. The justification for this approximation is that when the internal space is compactified by a Planck scale, all the massive modes can safely be neglected in the low-energy limit. Unfortunately the matter is more complicated[17] and the approximation faces serious problems. First of all, it is not a simple matter to determine what are the "zero-modes" of the theory exactly. The zero-modes of the harmonic expansion do not necessarily become the massless modes because the physical (the four-dimensional) mass of the zero-modes can be determined only after one studies the possibility of a spontaneous symmetry breaking among them. On this problem the popular zero-modes ansatz does not help either. In fact the zero-modes ansatz, which identifies the isometry of the vacuum internal metric to be the physical gauge group of the four-dimensional effective theory, has a critical defect of its own.[14] This can be seen from our analysis of the previous section which tells us that, when the dimension of the isometry G is larger than that of the internal space, the gauge potential defined by the zero-modes ansatz should become *linearly dependent*. As a result not all the G-gauge degrees of freedom can become physical. In fact, one can argue that this is the origin of the consistency problem[15] of the zero-modes ansatz. To avoid this difficulty recently some authors have proposed the so-called "K invariance" of the zero-modes,[16] which, when applied to the 11-dimensional supergravity, apparently gives the same physical gauge group as our reduction method. In spite of this apparent similarity, however, it is impossible to miss the fundamental difference between the two approaches. To illustrate this point let us consider the case when the gauge group becomes SU(2). In this case they start from SO(8) as the vacuum isometry, but require the zero-modes to be singlets under the SO(5) subgroup (the K invariance) to obtain SU(2) as the physical gauge group.[16] Evidently this SU(2) is the subgroup of SO(8) that commutes with SO(5). In contrast, in our case SU(2) is obtained by identifying the internal space as SO(5)/SO(3), but here as the subgroup of SO(5) that commutes with SO(3). Furthermore in the scalar sector they seem to identify the scalars [the most general SO(5)-invariant metric on S^7] as SO(5) singlets. But in our case the scalars that describe the most general SO(5)-invariant metric on SO(5)/SO(3) are certainly *not* singlets of SO(5). They become singlets only under the SO(3) subgroup [the ad(H) invariance] of SO(5). What is more, in our case all the physical degrees of freedom are determined without ever mentioning SO(8). Of course, the difference between the two methods goes far beyond this. In their case the reduction is possible only after a spontaneous compactification of the internal space, which makes some of their "zero-modes" extremely heavy. Although this does not cause a problem for the consistency of the dimensional reduction it certainly makes the physical validity of the zero-modes approximation questionable.[14] In our case this problem of validity does not arise because our dimensional reduction does not involve any approximation.

But perhaps a more serious problem with the zero-modes approximation lies in its quantum instability.[17] This problem arises because, no matter what zero-modes one starts with, there is no way to keep them from interacting with the "higher-modes" in the high-energy limit. The question then is how does one know whether the nature of the four-dimensional effective theory (the zero-modes and their interaction) will remain unchanged when the quantum fluctuation turns on the interaction with the higher-modes. This is really the consistency problem *at the quantum level*, which could be potentially more serious than the consistency problem at the classical level that we have discussed above.

V. LEFT ISOMETRY

Now we go back to $P(M,G)$ of Sec. II and introduce another isometry[5,9] H on P with the Killing vector fields m_i ($i = 1, 2, ..., k$; $k = \dim H$),

$$\mathcal{L}_{m_i} g_{AB} = 0 , \tag{34}$$

which has the following properties.

(i) They are linearly independent, and internal:

$$m_i = m_i^a \xi_a .$$

(ii) They commute with the right isometry G but formally form a subgroup H of G,

$$[m_i,\xi_a] = 0, \quad [m_i,m_j] = -(1/\kappa) f_{ij}^{(H)k} m_k . \tag{35}$$

To make sure that H is independent of G we further require that $\mathcal{H}$ does not contain the center of $\mathcal{G}$.

(iii) The integral manifold of the Killing fields is a metric submanifold, or the metric

$$g_{ij} = g_{AB} m_i^A m_j^B \tag{36}$$

is invertible. We will call this manifold the H fiber.

The above isometry makes each fiber $\pi^{-1}(x)$ of $P(M,G)$ a principle fiber bundle $G(G/H,H)$ of its own. To see this notice that (35) implies that each m_i^a forms an adjoint representation of G so that locally one may always find a cross section σ_H in $\pi^{-1}(U)$ on which m_i becomes exactly identical to ξ_i. In this local trivialization m_i may be regarded as the right translation of $\xi_i(\sigma_H(x))$ on the fiber. More precisely with the local parametrization of $p \in \pi^{-1}(x)$ by $p = (x,a)$ ($x \in M$, $a \in G$) one has

$$m_i(x,a) = R_{a*} \xi_i(x,e) ,$$

where R_a is the right multiplication of a, and e is the identity element of G. So m_i generates a *left* action H on $\pi^{-1}(x)$. From this one may view $\pi^{-1}(x)$ as a principal fiber bundle $G(G/H,H)$, but this time with the structure group H acting on the left. We will denote the $P(M,G)$ that has the additional H structure by $P(M,G,H)$.

Notice that not all the subgroups H may be qualified to describe a left isometry. First H must admit a bi-invariant metric. But, more importantly, G/H must be *reductive* since the metric ϕ_{ab} on $\pi^{-1}(x)$ should be able to define a G-invariant connection on $G(G/H,H)$. In other words $\mathcal{G}$ must have the following reductive decomposition[4]:

$$\mathcal{G} = \mathcal{H} + \mathcal{M} \text{ (direct sum)}, \quad \text{ad}(H)\mathcal{M} = \mathcal{M} .$$

This is the necessary and sufficient condition for $G(G/H,H)$ to admit a G-invariant H connection.

With $Q = P(M,G,H)/H$, P may be regarded as a principal fiber bundle $P(Q,H)$, with the structure group acting *on the left*. Moreover the quotient space Q may be regarded as a homogeneous fiber bundle $Q(M,G/H)$ on which G acts effectively on the right. But here it is important to make a distinction between $Q(M,G/H)$ and the associated bundle[4] $E(M,G/H,G,P)$ of $P(M,G)$ that can be obtained by projecting out the right-isometry subgroup H_R of G from P. Although E is diffiomorphic to Q, notice that for E the group G acts on the left on the standard fiber G/H, but on Q it still acts on the right. More significantly $E = P/H_R$ can always be obtained from $P(M,G)$ without the introduction of the left isometry. The fact that this difference is not a matter of semantics will become obvious in the following.

The left isometry allows us to introduce a horizontal subspace $\tilde{H}_P$ at each $p \in P(Q,H)$ which is horizontal with respect to the H fiber. The corresponding G-invariant connection one-form θ^i is given by the dual one-form of m_i:

$$\theta^i{}_A = g^{ij} g_{AB} m_j^B = g^{ij} \phi_{ab} \omega_A^a m_j^b = \theta^i{}_a \omega_A^a , \tag{37}$$

where ω^a is the connection one-form of $P(M,G)$. Now the projection operator of the tangent space $T(P)$ that projects out the horizontal component is given by

$$\hat{h}_A{}^B = \delta_A{}^B - \theta_A{}^i m_i{}^B = \delta_A{}^B - \hat{k}_A{}^B ,$$

where $\hat{k}_A{}^B$ is the projection operator for the vertical H component. Since the projection operates within the basis ξ_a of $\pi^{-1}(x)$, one may also express the projection operator by

$$\hat{h}_a{}^b = \delta_a{}^b - \theta_a{}^i m_i{}^b = \delta_a{}^b - \hat{k}_a{}^b . \tag{38}$$

From this one has the following decomposition of ξ_a and ω^a:

$$\begin{aligned} \xi_a &= \hat{h}_a{}^b \xi_b + \hat{k}_a{}^b \xi_b = \hat{h}_a + \theta^i{}_a m_i , \\ \omega^a &= \hat{h}_b{}^a \omega^b + \hat{k}_b{}^a \omega^b = \hat{\phi}^a + m_i^a \theta^i , \end{aligned} \tag{39}$$

where $\hat{h}_a$ and $\hat{\phi}^a$ are the horizontal components of ξ_a and ω^a. Notice that they (as well as ξ_a and ω^a) are left invariant,

$$\mathcal{L}_{m_i} \hat{h}_a = 0 , \quad \mathcal{L}_{m_i} \hat{\phi}^a = 0 , \tag{40}$$

which follows directly from (34).

Any tensor field on $P(Q,H)$ that is invariant under H (and horizontal with respect to H) may be projected down to a tensor field on Q. Conversely any tensor field on Q has a unique horizontal lift on $P(Q,H)$ with the above H invariance. To make this one-to-one correspondence more explicit let $\partial_\mu \otimes \partial_{\underline{a}}$ be the coordinate basis introduced in Sec. III, $\hat{h}_{\underline{a}}$ the horizontal lift of $\partial_{\underline{a}}$ to $P(Q,H)$, and $\hat{\phi}^{\underline{a}}$ the dual one-form of $\hat{h}_{\underline{a}}$. Then instead of ξ_a one may use $h_{\underline{a}} \otimes m_i$ as a basis on $\pi^{-1}(x)$. Now one can put

$$\begin{aligned} &\hat{h}_a = \hat{h}_a{}^{\underline{b}} \hat{h}_{\underline{b}} , \quad \hat{\phi}^a = \hat{\phi}_{\underline{b}}{}^a \hat{\phi}^{\underline{b}} , \\ &\hat{\phi}_{\underline{a}}{}^{\underline{c}} \hat{h}_{\underline{c}}{}^{\underline{b}} = \delta_{\underline{a}}{}^{\underline{b}} , \quad \hat{h}_a{}^{\underline{c}} \hat{\phi}_{\underline{c}}{}^b = \hat{h}_a{}^b , \end{aligned} \tag{41}$$

and obtain the following commutation relations:

$$\begin{aligned} &[\hat{h}_{\underline{a}}, \hat{h}_{\underline{b}}] = -(1/\kappa) f_{ab}{}^c \hat{\phi}_{\underline{a}}{}^a \hat{\phi}_{\underline{b}}{}^b \theta_c{}^i m_i , \\ &[\hat{h}_{\underline{a}}, m_j] = 0 , \quad [m_i, m_j] = -(1/\kappa) f_{ij}^{(H)k} m_k . \end{aligned} \tag{42}$$

In this basis the metric ϕ_{ab} becomes block diagonal,

$$\phi_{ab} = \left(\begin{array}{c|c} \hat{g}_{\underline{ab}} & 0 \\ \hline 0 & g_{ij} \end{array} \right) ,$$

and the right invariance of ϕ_{ab} can be expressed by

$$\partial_a \hat{g}_{\underline{bc}} = \hat{F}_{\underline{ab}}{}^{\underline{d}} \hat{g}_{\underline{dc}} + \hat{F}_{\underline{ac}}{}^{\underline{d}} \hat{g}_{\underline{bd}} , \quad \partial_a g_{ij} = 0 , \tag{43}$$

where $\hat{F}_{\underline{ab}}{}^{\underline{c}}$ is defined by

$$[\hat{\xi}_a, \hat{h}_{\underline{b}}] = \hat{F}_{\underline{ab}}{}^{\underline{c}} \hat{h}_{\underline{c}} . \tag{44}$$

Actually this equality also follows from the right invariance of ϕ_{ab}. Now, notice that since $\hat{h}_a{}^{\underline{b}}$ and $\hat{\phi}_{\underline{a}}{}^b$ have no H dependence they may also be regarded as functions on $Q(M,G/H)$. In fact, after the projection on Q, $\hat{h}_a$ should be identified as the generators h_a of G on Q so that $\hat{h}_a{}^b$ becomes $h_a{}^b$. More significantly one can now recognize the dual one-form ϕ^a of h_a defined by (19) as the H projection of $\hat{\phi}^a$. Indeed (19) follows from (44). This proves the existence and uniqueness of ϕ^a on $Q(M,G/H)$, when G/H is reductive. At this point the parallel between $P(M,G,H)$ and $Q(M,G/H)$ becomes unmistakable. For instance, the covariant fields $B_\mu{}^a$ and h_{ab} on Q are nothing more than the horizontal components of $A_\mu{}^a$ and ϕ_{ab} on $P(Q,H)$. We close this section by pointing out that this kind of natural homomorphism does not exist between $P(M,G)$ and its associated bundle $E(M,G/H,G,P)$.

VI. UNIFICATION FROM $P(M,G,H)$

The unified theory based on $P(M,G,H)$ can easily be obtained from the action integral (11) of $P(M,G)$ by imposing the left isometry H on it. The left isometry requires the metric ϕ_{ab} to be ad(H) invariant. So the scalar curvature R_G on $\pi^{-1}(x)$ can be expressed as the following sum of two intrinsic curvatures of H and G/H, and an extrinsic part that comes from the nontrivial embedding of G/H into G:

$$R'_G = R_H + R_{G/H} + R_E , \tag{45}$$

where

$$\begin{aligned} R_H &= \tfrac{1}{4} g^{ik} f_{ij}{}^l f_{kl}{}^j , \\ R_{G/H} &= \tfrac{1}{2} h^{ab} f_{ac}{}^d f_{bd}{}^c + \tfrac{1}{4} h^{ab} h^{cd} h_{ef} f_{ac}{}^e f_{bd}{}^f , \\ R_E &= \tfrac{1}{4} g_{ij} h^{ab} h^{cd} \theta^i{}_e \theta^j{}_f f_{ac}{}^e f_{bd}{}^f . \end{aligned}$$

Notice that here h_{ab} represents the horizontal component of ϕ_{ab}. As for the gauge potential the left isometry requires that[5,9]

$$D_\mu m_i^a = 0 . \tag{46}$$

This, together with (8), means that *the only nonvanishing components of $F_{\mu\nu}^a$ must be those of the little group H^* which leaves m_i^a invariant*. So H^* must be the commutant subgroup of H in G. Mathematically this means that[4] the holonomy group of the potential $A_\mu{}^a$ of $P(M,G)$ must become H^*, even though it has the apparent G-gauge degrees of freedom. In other words, the left isometry requires the connection on $P(M,G)$ to be reducible to a connection on a principal bundle $P^*(M,H^*)$.

With the above remarks the Einstein–Hilbert action on $P(M,G,H)$ can be written as

$$I = -\frac{1}{16\pi G_0} \int \sqrt{g_M} \sqrt{gh} \, (R'_P + \Lambda) d^4x \, d^nG , \tag{47}$$

where g_M and h are the same as before, $g = |\det g_{ij}|$, and R'_P is given by

$$R'_P = R_M + R'_G + \frac{e^2\kappa^2}{4}(h_{ab} + g_{ij}\,\theta^i_a\,\theta^i_b)F_{\mu\nu}{}^a F_{\mu\nu}{}^b$$
$$+ \frac{1}{4} h^{ab}h^{cd}[(D_\mu h_{ac})(D_\mu h_{bd})$$
$$- (D_\mu h_{ab})(D_\mu h_{cd})]$$
$$+ \frac{1}{2} g_{ij} h^{ab}(D_\mu \theta^i_a)(D_\mu \theta^j_b)$$
$$- \frac{k-1}{4k}\frac{(\partial_\mu g)^2}{g^2} - \frac{1}{2}\frac{\partial_\mu g}{g}\frac{\partial_\mu h}{h}.$$

Notice that although the action integral is expressed in an explicitly G-invariant form, one has to keep in mind that the *physical* gauge symmetry is restricted to the holonomy group H^*. Now, the dimensional reduction can easily be performed as before, and the unimodularity of G guarantees the consistency of the dimensional reduction.

It must become clear that, upon the H projection, the unified action (47) is reduced to the action integral (32) on $Q(M,G/H)$. This is so because under the projection g_{ij} and θ^i_a vanish, and H^* is reduced to $K = H^*/(H \cap H^*)$. In this sense the left isometry provides an alternative way to obtain the unified theory based on $Q(M,G/H)$. However, the above unified theory on $P(M,G,H)$ is interesting in its own right. First, it has a remarkably suggestive form in that the field θ^i_a could play an important role in a spontaneous symmetry breaking as a Higgs field. But, more importantly, the left isometry can be implemented in such a way that it adds a nontrivial topological structure to the theory.[9] To understand this notice that the Killing vector fields $m^a_i(x)$, regarded as a mapping from an S^2 of M to the homogeneous space G/H, determines the second homotopy $\pi_2(G/H)$ of G/H. So when the left isometry has an isolated singularity inside S^2, $\pi_2(G/H)$ becomes nonzero. In this case the gauge potential defined by (46) must necessarily contain a magnetic flux of a non-Abelian magnetic charge of H^*. This means that when the potential is reduced to $P^*(M,H^*)$, it will develop a string singularity in M and thus make $P^*(M,H^*)$ nontrivial. So the theory will effectively describe *a nontrivial non-Abelian gauge theory* of H^*. A particularly interesting case is obtained when H becomes Cartan's subgroup of G, in which case H^* coincides with H. In this case the topology of the theory can be chosen in such a way that the gauge field contains both electric and magnetic components of H^*, and becomes capable of describing all possible non-Abelian monopoles of $P(M,G)$. The resulting theory becomes *a dual gauge theory* of $H^* = H$. For this reason the left isometry is sometimes called the magnetic symmetry.[5,9]

VII. DISCUSSION

In this paper we have discussed a set of unified theories that can be obtained imposing an isometry G to the unified metric. The isometry provides a natural method of dimensional reduction guaranteed to be consistent when G is unimodular. When matter fields are present the isometry can easily be generalized to include the matter fields. The discussion in Sec. III was based on the existence of the one-form ϕ^a defined by (19). In this paper we were able to prove the existence and uniqueness of the one-form only when G/H becomes reductive. But we suppose that this will not pose a serious restriction from the physical point of view.

Recently some authors[7,8] have proposed a method to construct a unified theory based on $E(M,G/H)$ by identifying it as an associated bundle of $P(M,N/H)$, where N is the normalizer of H in G. Aside from the obvious discrepancy between their action integral on $E(M,G/H)$ and ours on $Q(M,G/H)$, which originates from their use of an incorrect volume element on E, their result is very similar to our result of Sec. III. Indeed as a manifold our $Q(M,G/H)$ can be viewed as identical to their $E(M,G/H)$. In spite of this similarity, however, we wish to emphasize that the difference, especially in the way we obtain the unified action, is also evident. To see the difference notice that on E they introduce *two actions* separately[7]: the right action (the "global" symmetry) G and the gauge action (the "local" symmetry) $K = N/H$ which commutes with (and thus is independent of) G. But in our case there is *only one right action* G on Q, and the gauge symmetry K is obtained as a subgroup of G. In fact, our K is obtained as the holonomy group of the isometry G. Besides, the natural homomorphism between $P(M,G)$ and $Q(M,G/H)$ that we have established in Sec. V does not come easily between $P(M,G)$ and $E(M,G/H)$. This discrepancy is obvious when H becomes the identity subgroup. In this limit Q becomes $P(M,G)$ itself with only one right action G, but E becomes an associated bundle $E(M,G,G,P)$ of P on which they still have two actions, the "global" G and the "local" G, which act independently from the right and from the left. The difference goes beyond this. In their case they require the existence of a compactifying ground state solution as a prerequisite[8] for a consistent dimensional reduction. In our reduction we find no reason why one must make such a requirement. In fact, one can easily show that[14] a perfectly consistent dimensional reduction is possible with a noncompact G/H. We emphasize, however, that one must require G to be a unimodular as a necessary (and sufficient) condition for the consistency of the dimensional reduction.

The introduction of $P(M,G,H)$ in Sec. V provides a general method of prolongation and reduction of a gauge symmetry. One can obtain the reduction of the physical gauge symmetry from $P(M,G)$ by making the left isometry H larger and larger starting from the identity subgroup. Conversely one may obtain the prolongation starting from $H = G$ and making H smaller and smaller. A potentially interesting case of the prolongation is *the extended gauge theory*. One can obtain this as a most general nontrivial non-Abelian gauge theory[10,11] by adding a valence potential[10] (i.e., a gauge-covariant vector field that has no neutral component) to the dual gauge theory of $H^* = H$. Since the connection space (the space of gauge potentials) forms an affine space, a most general gauge potential of $P(M,G)$ can be expressed as the sum of a gauge-covariant vector field that has no H^* component (the valence potential) and the dual potential of $H^* = H$. The result is the extended gauge theory in which the gauge potential of G is decomposed into the valence part and the dual part in a gauge-independent way. This prolongation allows us to have an alternative, completely uncon-

ventional and nevertheless physically very interesting, interpretation of a non-Abelian gauge symmetry.[10,11]

ACKNOWLEDGMENTS

One of us (Y.M.C.) thanks T. Appelquist, A. Chodos, and V. Moncrief for their kind hospitality during his visit at the Yale University.

This work is supported in part by U. S. Department of Energy Contract No. DE-AC02-76ERO 3075, and by the Korean Ministry of Education, Korean Science and Engineering Foundation, and Daewoo Foundation.

[1]T. Kaluza, Sitzungsber Preuss. Akad. Wiss. **1921**, 966; O. Klein, Z. Phys. **37**, 895 (1926); B. deWitt, in *Relativity, Groups, and Topology* (Gordon and Breach, New York, 1965); R. Kerner, Ann. Inst. H. Poincaré **9**, 143 (1968); A. Trautmann, Rep. Math. Phys. **1**, 29 (1970).

[2]Y. M. Cho, J. Math. Phys. **16**, 2029 (1975); Y. M. Cho and P. G. O. Freund, Phys. Rev. D **12**, 1711 (1975); L. N. Chang, K. Macrae, and F. Mansouri, *ibid.* **13**, 235 (1976).

[3]Y. M. Cho and P. S. Jang, Phys. Rev. D **12**, 3189 (1975); F. Mansouri and L. Witten, Phys. Lett. B **127**, 341 (1983).

[4]See, e.g., S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, Vols. I & II (Interscience, New York, 1969).

[5]Y. M. Cho, CERN report TH-3414, 1982; see also Y. M. Cho, in *Proceedings of XIIIth International Colloquium on Group Theoretical Methods in Physics*, edited by W. Zachary (World Scientific, Singapore, 1984).

[6]Y. M. Cho, Phys. Lett. B **186**, 38 (1987).

[7]R. Coquereaux and A. Jadczyk, Commun. Math. Phys. **80**, 79 (1983).

[8]R. Coquereaux and A. Jadczyk, Nucl. Phys. B **276**, 617 (1986).

[9]Y. M. Cho, Phys. Rev. Lett. **44**, 1115 (1980); Phys. Rev. D **21**, 1080 (1980).

[10]Y. M. Cho, Phys. Rev. Lett. **46**, 302 (1981); Phys. Rev. D **23**, 2415 (1981).

[11]Y. M. Cho, Phys. Lett. B **167**, 432 (1986).

[12]E. Witten, Nucl. Phys. B **186**, 412 (1981); A. Salam and J. Strathdee, Ann. Phys. (NY) **141**, 316 (1982); M. J. Duff, Nucl. Phys. B **219**, 389 (1983).

[13]P. G. O. Freund and M. Rubin, Phys. Lett. B **97**, 233 (1980); F. Englert, *ibid.* **119**, 339 (1982).

[14]Y. M. Cho, Phys. Rev. Lett. **55**, 2932 (1985).

[15]M. J. Duff, B. Nilsson, C. Pope, and N. Warner, Phys. Lett. B **149**, 90 (1984).

[16]M. J. Duff and C. Pope, Nucl. Phys. B **255**, 355 (1985).

[17]Y. M. Cho, in *Proceedings of XIVth International Colloquium on Group Theoretical Methods in Physics*, edited by Y. M. Cho (World Scientific, Singapore, 1986); B. deWitt and H. Nicolai, Nucl. Phys. B **281**, 211 (1987).

[18]Y. M. Cho, Phys. Rev. D **35**, 2628 (1987).

[19]Here we use the index notation: The capital indices A and B indicate the type of tensors, and the lower case indices μ,ν,a,b are used to label them.

[20]M. A. H. MacCallum and A. H. Taub, Commun. Math. Phys. **25**, 173 (1972).

Geometric model for gravitational and electroweak interactions

M. Rosenbaum, J. C. D'Olivo, and E. Nahmad-Achar
Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Apartado Postal 70-543, México, D. F. 04510

R. Bautista and J. Muciño
Instituto de Mathemáticas, Universidad Nacional Autónoma de México, México, D. F. 04510

(Received 4 August 1988; accepted for publication 8 February 1989)

A fiber bundle treatment for Kaluza–Klein-type geometric unification of gravitation with the bosonic sector of the standard electroweak theory is presented. The most general G-invariant quadratic Lagrangian is constructed explicitly, and it is shown that the Higgs field sector, including the symmetry-breaking potential, arise naturally from torsion in the fiber through an adequate choice of its transformation properties.

I. INTRODUCTION

This paper is part of a program intended to study spontaneously compactified solutions of gravitation—Yang–Mills–Higgs systems, where the Higgs scalar fields originate from the torsion and acquire dynamics through the introduction in the Lagrangian of quadratic terms in the curvature tensor.

A comprehensive geometrical treatment for Kaluza–Klein-type unification of gauge fields and gravitation has been developed by Cho.[1] The inclusion of torsion in this principal fiber bundle (PFB) formalism as a source of the Higgs fields was considered by Katanayev and Volovich.[2] Rosenbaum and Ryan[3] have applied the approach of Cremmer and Scherk[4] to study spontaneously compactified solutions to the field equations resulting from the most general quadratic Lagrangian that can be constructed from the curvature and torsion in the PFB. They showed that for $SO(3)$ as a characteristic group, and a Gauss–Bonnet combination of the quadratic terms in the curvature, the compactified solutions that were obtained also led to direct predictions on the size of the dimensionless coupling constant of the Yang–Mills fields remarkably close to the value of the coupling constant for the $SU(2)$ factor in the $SU(2)\times U(1)$ electroweak model. Since $SU(2)$ is a covering group of $SO(3)$, it is reasonable to expect that some of the salient features of the model of Rosenbaum and Ryan should be preserved when extending it to the structure group $SU(2)\times U(1)$ and, in particular, it is worthwhile to test if the agreement in the value of the coupling constant mentioned above is still preserved. This study is presently being completed and it will be the subject of a forthcoming paper.

Here we shall concentrate on the development of the appropriate fiber bundle formalism for $SU(2)\times U(1)$. This, we believe, is by itself an interesting result. First, because we arrive at a general G-invariant Kaluza–Klein-type Lagrangian that unifies geometrically the bosonic part of the electroweak model with gravitation, and second, because of the inherent mathematical elegance of the resulting theory.

As we will show, our construction cannot be based on a simple extension of the ideas contained in Refs. 2 and 3, where an $\mathscr{A}d$-invariant Lagrangian was obtained rather directly by allowing the torsion components (which generate the Higgs fields) to transform in the same way as the gauge field tensor, i.e., in accordance to the adjoint representation of the group. In the case of the direct product group $SU(2)\times U(1)$, since $U(1)$ is Abelian, such an assumption, which is equivalent to the seemingly most natural requirement of right invariance of the torsion, would lead to the loss of important dynamical information on the real Higgs field which is associated with the generator of $U(1)$.

In the phenomenological approach to the electroweak model this problem is resolved, of course, by having the group act on the gauge fields in accordance with the adjoint representation, while the scalar fields are required, in an *ad hoc* fashion, to transform as a complex spinor doublet under $SU(2)$.

To arrive at this result from a geometrical point of view requires a generalization of the law of transformation of the torsion as well as a very careful choice of connections and representation of base vectors for our frame bundle. But once the proper choice is made all the terms in the bosonic part of the gravitation–electroweak Lagrangian follow unequivocally and in a geometrically unified manner from the curvature and torsion of the bundle.

It is important to stress that in its present stage our theory does not consider fermionic fields, and it must be regarded so far as an $SU(2)\times U(1)$ gauge theory coupled only to gravitation and a complex doublet of Higgs fields, where these fields, as well as the quartic scalar potential and the negative mass term required for spontaneous symmetry breaking, originate from torsion. Further remarks on the possibility of also including the fermion and Yukawa-type Lagrangians within the framework of our formalism, which is needed to complete the electroweak model, will be given in Sec. V.

This paper is organized as follows. Section II is dedicated to the construction of the several principal fiber bundles needed for our theory. In Sec. III we derive the different components of the curvature and torsion tensors on the bundle of frames and use the results to obtain a G-invariant Lagrangian which is, therefore, well defined on the base manifold. Section IV contains a procedure for relating the dynamical form of the Higgs Lagrangian in our formalism with the form which is commonly used by field theorists and particle physicists.

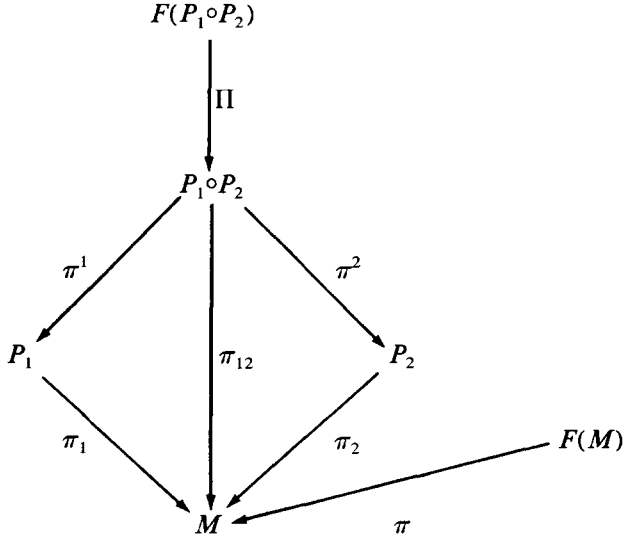
As far as notation is concerned, we will consistently use the following ranges for our indices: latin lower case letters

 0022-2488/89/071579-12$02.50

from the middle of the alphabet will have the range $1\leqslant i, j, k,\ldots,\leqslant n$, greek lower case letters will have the range $1\leqslant\alpha,\beta,\gamma,\ldots,\leqslant 3$, upper case latin letters from the beginning of the alphabet will have the range $1\leqslant A, B, C,\ldots,\leqslant 4$, and lower case latin letters from the beginning of the alphabet will cover the full range $1\leqslant a, b, c, d,\ldots,\leqslant n+4$. The spaces to which these indices refer will be self-evident from the text. With respect to sign conventions, we follow those of Landau–Lifshitz,[5] i.e., the signature of the base manifold metric is $\mathrm{sgn}(g_{ij}) = (+, -,\ldots,-)$; the Riemann tensor is defined by $\xi^{a}{}_{;bc} - \xi^{a}{}_{;cb} = R^{a}{}_{dbc}\xi^{d}$, and the Ricci tensor by $R_{cd} = R^{a}{}_{cad}$. The Einstein equations then take the form $R_{ab} - \frac{1}{2}Rg_{ab} = 8\pi T_{ab}$ with $T_{00}\geqslant 0$.

II. THE BUNDLE FRAMEWORK FOR SU(2)×U(1)

As mentioned in the Introduction, the most adequate framework for a Kaluza–Klein theory that naturally unifies gravitation with the gauge and scalar fields is the principal fiber bundle formalism. In the specific case of $\mathrm{SU}(2)\times\mathrm{U}(1)$, the theory requires five different PFB's that are interrelated according to the diagram



For the description of these constructions we shall rely closely on the notation used by Bleecker.[6] Thus M denotes an n-dimensional oriented manifold, which we take to be space-time and which acts as the base space of the following PFB's:

(1) $\pi: F(M)\to M$, is the orthonormal frame bundle of M with group $O(r,s)$. For $u\in F(M)_x$ and the usual basis $\{e_i\}$, $i=1,\ldots,n$ of $\mathfrak{R}^n$, we choose an orthonormal frame at $x\in U\subset M$ by means of the linear isomorphism $u: \mathfrak{R}^n\to T_xM$, i.e., $u(e_i)=\overline{E}_i$, $i=1,\ldots,n$, are orthonormal vector fields with respect to the metric g on M, defined in a neighborhood U of $x=\pi(u)$ in such a way that the local section σ: $U\to F(M)$ determined by $\overline{E}_1,\ldots,\overline{E}_n$ is tangent to the horizontal subspace of $T_{\sigma(x)}F(M)$ relative to the connection $\theta(g)$. Consequently $\theta(g)(\sigma^*\overline{E}_i) = (\sigma^*\theta(g))(\overline{E}_i) = \bar{\theta}(g)(\overline{E}_i) = 0$ at x.

The curvature of the connection $\theta(g)\in\Lambda^1(F(M), \mathcal{O}(r,s))$ is given by

$$\Omega^{\theta(g)} \equiv D^{\theta(g)}\theta(g) = d\theta(g) + \tfrac{1}{2}[\theta(g),\theta(g)]\in\overline{\Lambda}^2(F(M),\mathcal{O}(r,s)). \quad (2.1)$$

Note that, since $u^{-1}(\pi_*\sigma_*\overline{E}_i) = e_i$ and $\theta(g)(\sigma_*\overline{E}_i) = 0$, the vectors $\sigma_*\overline{E}_i\in T_u(F(M))$ are standard horizontal vectors relative to $\theta(g)$. Thus $\Omega^{\theta(g)}(\sigma_*\overline{E}_i,\sigma_*\overline{E}_j)(e_k)$ is the image of $e_k\in\mathfrak{R}^n$. We can therefore write

$$\Omega^{\theta(g)}(\sigma_*\overline{E}_i,\sigma_*\overline{E}_j)(e_k) = R^{h}{}_{kij}(\sigma(x))e_h. \quad (2.2)$$

Also note that for $X_u\in T_uF(M)$ we can define the canonical one-form $\varphi_M\in\overline{\Lambda}^1(F(M),\mathfrak{R}^n)$ by

$$\varphi_M(X_u) = u^{-1}(\pi_*(X_u))\in\mathfrak{R}^n. \quad (2.3)$$

In terms of this canonical one-form the torsion two-form $\Theta^{\theta(g)}$ is given by

$$\Theta^{\theta(g)} = D^{\theta(g)}\varphi_M = d\varphi_M + \theta(g)\,\dot{\wedge}\,\varphi_M\in\overline{\Lambda}^2(F(M),\mathfrak{R}^n), \quad (2.4)$$

where the quantity $\theta(g)\,\dot{\wedge}\,\varphi_M$ is defined by

$$(\theta(g)\,\dot{\wedge}\,\varphi_M)(X_u,Y_u) = \theta(g)(X_u)\cdot\varphi_M(Y_u) - \theta(g)(Y_u)\cdot\varphi_M(X_u), \quad (2.5)$$

and the "dot" operation denotes the left action of $O(r,s)$ on $\mathfrak{R}^n$.

Furthermore, since $\Theta^{\theta(g)}(\sigma_*\overline{E}_i,\sigma_*\overline{E}_j)\in\mathfrak{R}^n$, we can write

$$\Theta^{\theta(g)}(\sigma_*\overline{E}_i,\sigma_*\overline{E}_j) = S^{k}{}_{ij}(\sigma(x))e_k. \quad (2.6)$$

If we now let $\overline{\varphi}^{i}_{M}$ be the one-forms dual to $\overline{E}_i$, i.e., $\overline{\varphi}^{i}_{M}(\overline{E}_j) = \delta^i_j$, and we set $X_u = \sigma_*\overline{E}_i$ in (2.3), then

$$\varphi_M(\sigma_*\overline{E}_i) = (\sigma^*\varphi_M)(\overline{E}_i) = e_i = (\overline{\varphi}^{1}_{M}(\overline{E}_i),\ldots,\overline{\varphi}^{n}_{M}(\overline{E}_i)). \quad (2.7)$$

Thus

$$\sigma^*\varphi_M = \overline{\varphi}_M = \overline{\varphi}^{i}_{M}e_i. \quad (2.8)$$

The pullback with the local section σ of the canonical one-forms allows us to relate the curvature and torsion tensors in $F(M)$, as given by (2.2) and (2.6), to the corresponding tensors in $T_x(M)$. Indeed, acting with σ^* on (2.1) we get

$$(\overline{\Omega}^{\bar{\theta}(g)})^{h}{}_{k} \equiv (\sigma^*\Omega^{\theta(g)})^{h}{}_{k} = D^{\bar{\theta}(g)}\bar{\theta}^{h}{}_{k}(g) = \tfrac{1}{2}\underline{R}^{h}{}_{kij}(x)\overline{\varphi}^{i}_{M}\wedge\overline{\varphi}^{j}_{M} \quad (2.9)$$

or

$$\overline{\Omega}^{\bar{\theta}(g)}(\overline{E}_i,\overline{E}_j) = \underline{R}^{h}{}_{kij}(x)e_h\otimes\hat{e}^k, \quad (2.10)$$

where $\hat{e}^k$ is the dual of e_k.

On the other hand, from (2.2) we get

$$\overline{\Omega}^{\bar{\theta}(g)}(\overline{E}_i,\overline{E}_j) = R^{h}{}_{kij}(\sigma(x))e_h\otimes\hat{e}^k. \quad (2.11)$$

Consequently, we have the following lemma.

Lemma 1: Let $R^{h}{}_{kij}$ be the components of an L tensor in $C(F(M),T^{1,3})$ (according to the definitions in Ref. 6), and let $R^{h}{}_{kij}\in\mathcal{T}^{1,3}$ be the components of the curvature tensor of (M,g) relative to the orthonormal fields $\{\overline{E}_i\}$, then

$$R^{h}{}_{kij}(\sigma(x)) = \underline{R}^{h}{}_{kij}(x). \quad (2.12)$$

Proceeding in a similar fashion with the torsion, we get, from (2.4) and (2.8),

$$\overline{\Theta}^{\bar{\theta}(g)} = D^{\bar{\theta}(g)}\overline{\varphi}_M = d\overline{\varphi}_M + \bar{\theta}(g)\,\dot{\wedge}\,\overline{\varphi}_M = \tfrac{1}{2}\underline{S}^{i}{}_{jk}(x)e_i\otimes(\overline{\varphi}^{j}_{M}\wedge\overline{\varphi}^{k}_{M})\in\overline{\Lambda}^2(M,\mathfrak{R}^n), \quad (2.13)$$

while (2.6) yields

$$\overline{\Theta}^{\bar{\partial}(g)}(\overline{E}_j,\overline{E}_k) = S^i{}_{jk}(\sigma(x))e_i. \tag{2.14}$$

Therefore, we also have the following lemma.

Lemma 2: Let $S^i{}_{jk}$ be the components of an L tensor in $C(F(M), T^{1,2})$, and let $\underline{S}^i{}_{jk} \in \mathcal{T}^{1,2}$ be the components of the torsion tensor of (M, g) relative to the orthonormal fields $\{\overline{E}_i\}$, then

$$S^i{}_{jk}(\sigma(x)) = \underline{S}^i{}_{jk}(x). \tag{2.15}$$

In our calculations in the following sections, we will be frequently using the isomorphisms implied by Eqs. (2.12) and (2.15).

(2) π_1: $P_1 \to M$, is a PFB with group $G_1 = \mathrm{SU}(2)$ and connection $\omega_1 \in \Lambda^1(P_1, \mathcal{G}_1)$, where $\mathcal{G}_1$ is the Lie algebra of G_1.

The curvature of the connection ω_1 is

$$\Omega_1 \equiv D^{\omega_1}\omega_1 \equiv d\omega_1 + \tfrac{1}{2}[\omega_1,\omega_1] \in \overline{\Lambda}^2(P_1, \mathcal{G}_1). \tag{2.16}$$

Note that if l_α $(\alpha = 1,2,3)$ is a basis for $\mathcal{G}_1$, we can write $\omega_1 = \omega_1^\alpha l_\alpha$, and

$$\Omega_1 = (d\omega_1^\alpha + \tfrac{1}{2}c^\alpha{}_{\beta\gamma}\omega_1^\beta \wedge \omega_1^\gamma) l_\alpha, \tag{2.17}$$

where $c^\alpha{}_{\beta\gamma}$ are the structure constants of G_1.

Moreover, if we let $E_1^{(1)},\ldots,E_n^{(1)}$ be an orthonormal basis of the horizontal subspace of $T_{p_1}P_1$ relative to ω_1, such that $\pi_{1*}E_i^{(1)} = \overline{E}_i$, and we also let $\overline{\varphi}^1_{(1)},\ldots,\overline{\varphi}^n_{(1)}$ be the one-forms dual to $E_1^{(1)},\ldots,E_n^{(1)}$, then we can also write

$$\Omega_1 = \tfrac{1}{2}(\Omega_1)^\alpha{}_{ij} l_\alpha \otimes (\overline{\varphi}^i_{(1)} \wedge \overline{\varphi}^j_{(1)}), \tag{2.18}$$

where $(\Omega_1)^\alpha{}_{ij}$ is a real function in $\pi_1{}^{-1}(U)$.

(3) π_2: $P_2 \to M$ is a PFB with group $G_2 = \mathrm{U}(1)$ and connection $\omega_2 \in \Lambda^1(P_2, \mathcal{G}_2)$, where $\mathcal{G}_2$ is the Lie algebra of G_2.

Since U(1) is Abelian, the curvature of the connection ω_2 is

$$\Omega_2 \equiv D^{\omega_2}\omega_2 = d\omega_2 \in \overline{\Lambda}^2(P_2, \mathcal{G}_2). \tag{2.19}$$

Taking $i = \sqrt{-1}$ as the basis for $\mathcal{G}_2$, we can write $\omega_2 = (-i\omega_2)i$, so that

$$(-i\Omega_2) = d(-i\omega_2) \in \overline{\Lambda}^2(P_2, \mathfrak{R}). \tag{2.20}$$

As an orthonormal basis of the horizontal subspace of $T_{p_2}P_2$ relative to ω_2, we take $\{E_i^{(2)}\}$, $i = 1,\ldots,n$, and the corresponding dual one-forms $\{\overline{\varphi}^i_{(2)}\}$ as above. We also require that $\pi_{2*}E_i^{(2)} = \overline{E}_i$. Thus

$$\Omega_2 = \tfrac{1}{2}(\Omega_2)_{ij} i \otimes \overline{\varphi}^i_{(2)} \wedge \overline{\varphi}^j_{(2)}, \tag{2.21}$$

where $(\Omega_2)_{ij}$ is a real function defined on $\pi_2{}^{-1}(U)$.

(4) π_{12}: $P_1 \circ P_2 \to M$ is the PFB with group $\mathrm{SU}(2) \times \mathrm{U}(1)$, obtained by splicing the bundles π_i: $P_i \to M$. In this way we have that $P_1 \circ P_2 = \{(p_1,p_2) \in P_1 \times P_2 | \pi_1(p_1) = \pi_2(p_2)\}$, and that $\pi_{12}(p_1,p_2) = \pi_1(p_1) = \pi_2(p_2)$. Also, for $(g_1,g_2) \in G_1 \times G_2$ and $(p_1,p_2) \in P_1 \circ P_2$ we define the right action of the product group by $(p_1,p_2)(g_1,g_2) = (p_1g_1, p_2g_2)$. The connections ω_1 and ω_2 may be used to construct a connection for $P_1 \circ P_2$. To this end note first that corresponding to the projections π^i: $P_1 \circ P_2 \to P_i$ given by $\pi^i(p_1,p_2) = p_i$, $i = 1,2$, it can be shown that π^1: $P_1 \circ P_2 \to P_1$ and π^2: $P_1 \circ P_2 \to P_2$ are also PFB's with characteristic groups $\{1\} \times \mathrm{U}(1) \cong \mathrm{U}(1)$ and $\mathrm{SU}(2) \times \{1\} \cong \mathrm{SU}(2)$, respectively. Moreover, since the differential $\pi^1{}_*$ maps tangent vectors in $T(P_1 \circ P_2)$ onto $T(P_1)$, we have that for $X_{(p_1,p_2)} \in T(P_1 \circ P_2)$, $\pi^{1*}\omega_1(X_{(p_1,p_2)}) = \omega_1(\pi^1{}_* X_{(p_1,p_2)}) = \omega_1(X_{p_1})$. However, ω_1 vanishes on horizontal vectors, so $\pi^1{}_* X_{(p_1,p_2)}$ has to be vertical on the fiber on which ω_1 acts, i.e., the pullback $\mathring{\omega}_1 \equiv \pi^{1*}\omega_1$ is a connection for π^2: $P_1 \circ P_2 \to P_2$. Similarly, $\mathring{\omega}_2 \equiv \pi^{2*}\omega_2$ is a connection for π^1: $P_1 \circ P_2 \to P_1$.

It is now a simple matter to show that $\pi^{1*}\omega_1 \oplus \pi^{2*}\omega_2$ is a connection for the spliced bundle π_{12}: $P_1 \circ P_2 \to M$. One only needs to prove that, given $X = X_1 \oplus X_2$ where $X_i \in \mathcal{G}_i$ (the Lie algebra of G_i, $i = 1,2$) and the fundamental vector field $(X_1 \oplus X_2)^*$ defined by

$$\begin{aligned}(X_1 \oplus X_2)^* &= \frac{d}{dt}((p_1,p_2)(\exp tX_1, \exp tX_2))|_{t=0} \\ &= X_1{}^* \oplus X_2{}^*,\end{aligned} \tag{2.22}$$

one gets

$$(\mathring{\omega}_1 \oplus \mathring{\omega}_2)(X_1{}^* \oplus X_2{}^*) = X_1 \oplus X_2, \tag{2.23}$$

and also

$$\begin{aligned}&(\mathring{\omega}_1 \oplus \mathring{\omega}_2)(R_{(g_1,g_2)*}(X_1 \oplus X_2)^*) \\ &\quad = \mathscr{A}\mathfrak{d}_{(g_1,g_2)^{-1}}(\mathring{\omega}_1 \oplus \mathring{\omega}_2)((X_1 \oplus X_2)^*).\end{aligned} \tag{2.24}$$

We introduce additional structure on $P_1 \circ P_2$ by defining a nondegenerate bundle metric h as follows.

Let k_1 and k_2 be $\mathscr{A}\mathfrak{d}$-invariant metrics on $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, and set

$$h = \pi_{12}^* g + k_1\mathring{\omega}_1 \oplus k_2\mathring{\omega}_2. \tag{2.25}$$

Note that for $X, Y \in T_{(p_1,p_2)}(P_1 \circ P_2)$, we have

$$\begin{aligned}h(X,Y) = g(\pi_{12*}X, \pi_{12*}Y) &+ k_1(\mathring{\omega}_1(X), \mathring{\omega}_1(Y)) \\ &+ k_2(\mathring{\omega}_2(X), \mathring{\omega}_2(Y)).\end{aligned} \tag{2.26}$$

It is easy to verify that for all $(g_1,g_2) \in G_1 \times G_2$, the right action $R_{(g_1,g_2)}$: $P_1 \circ P_2 \to P_1 \circ P_2$ on the fibers is an isometry of $(P_1 \circ P_2, h)$.

Relative to this metric, an orthonormal frame at $(p_1,p_2) \in P_1 \circ P_2$ is given by

$$\mathring{E}_1,\ldots,\mathring{E}_n,\mathring{E}_{n+1},\ldots,\mathring{E}_{n+4}.$$

Here, $\mathring{E}_1,\ldots,\mathring{E}_n$ [defined on $\pi_{12}{}^{-1}(U) \in P_1 \circ P_2$] are horizontal lifts of the orthonormal basis $\overline{E}_1,\ldots,\overline{E}_n$ on (M,g) such that $\pi_{12*}\mathring{E}_i = \overline{E}_i$ and $(\mathring{\omega}_1 \oplus \mathring{\omega}_2)(\mathring{E}_i) = 0$, while $\mathring{E}_{n+\alpha} = l_\alpha^* \oplus 0$ $(\alpha = 1,2,3)$ and $\mathring{E}_{n+4} = 0 \oplus l_4^*$ are fundamental vertical fields on $P_1 \circ P_2$, i.e.,

$$\mathring{\omega}_1(l_\alpha^* \oplus 0) = l_\alpha \in \mathcal{G}_1, \quad \mathring{\omega}_2(0 \oplus l_4^*) = l_4 \in \mathcal{G}_2. \tag{2.27}$$

Furthermore, l_1,l_2,l_3,l_4 are chosen so that they constitute an orthonormal basis of $\mathcal{G}_1 \oplus \mathcal{G}_2$ relative to $k_1 \oplus k_2$.

Consequently,

$$\begin{aligned}&h_{ij} = h(\mathring{E}_i,\mathring{E}_j) = g(\overline{E}_i,\overline{E}_j) = g_{ij} = \pm\delta_{ij}, \\ &\qquad i,j = 1,\ldots,n; \\ &h_{\alpha\beta} = h(\mathring{E}_{n+\alpha},\mathring{E}_{n+\beta}) = k_1(l_\alpha,l_\beta) = (k_1)_{\alpha\beta} = \delta_{\alpha\beta}, \\ &\qquad \alpha,\beta = 1,2,3; \\ &h_{44} = h(\mathring{E}_{n+4},\mathring{E}_{n+4}) = k_2(l_4,l_4) = 1,\end{aligned} \tag{2.28}$$

and all other cross terms in the bundle metric components vanish.

If we set $l_\alpha = -(i/2)\sigma_\alpha$, where the σ_α's obey the Pauli

algebra, then the SU(2)-manifold metric k_1 is given explicitly in terms of the corresponding structure constants by

$$(k_1)_{\alpha\beta} = -\tfrac{1}{2}c^{\lambda}{}_{\alpha\gamma}c^{\gamma}{}_{\beta\lambda} = -\tfrac{1}{2}\epsilon_{\lambda\alpha\gamma}\epsilon_{\gamma\beta\lambda} = \delta_{\alpha\beta}, \quad (2.29)$$

where $\epsilon_{\alpha\beta\gamma}$ is the usual Levi-Civita symbol.

Also, recalling that for the infinitesimal generator of U(1) we have $l_4 = i$, it follows that $k_2(l_4,l_4) = k_2(i,i) = 1$.

With this particular choice of an orthonormal basis on a neighborhood of (p_1,p_2), the calculations in the following sections will simplify considerably.

The curvature $\Omega^{\mathring{\omega}_1\oplus\mathring{\omega}_2}\in\overline{\Lambda}^2(P_1\circ P_2,\mathscr{G}_1\oplus\mathscr{G}_2)$ of $\mathring{\omega}_1\oplus\mathring{\omega}_2$ is given by

$$\begin{aligned}\Omega^{\mathring{\omega}_1\oplus\mathring{\omega}_2} &\equiv D^{\mathring{\omega}_1\oplus\mathring{\omega}_2}(\mathring{\omega}_1\oplus\mathring{\omega}_2)\\ &= d(\mathring{\omega}_1\oplus\mathring{\omega}_2) + \tfrac{1}{2}[\mathring{\omega}_1\oplus\mathring{\omega}_2,\mathring{\omega}_1\oplus\mathring{\omega}_2]\\ &= d\mathring{\omega}_1 + \tfrac{1}{2}[\mathring{\omega}_1,\mathring{\omega}_1]\oplus d\mathring{\omega}_2\\ &= \pi^{1*}(\Omega_1{}^{\omega_1})\oplus\pi^{2*}(\Omega_2{}^{\omega_2})\in\overline{\Lambda}^2(P_1\circ P_2,\mathscr{G}_1)\\ &\quad\oplus\overline{\Lambda}^2(P_1\circ P_2,\mathscr{G}_2). \end{aligned}\quad (2.30)$$

This result is a particular case of a theorem for spliced bundles which states that for any $\alpha\in\overline{\Lambda}^k(P_1\circ P_2,\mathscr{G}_1\oplus\mathscr{G}_2)$ and projections $\overline{\mathscr{G}}_i\colon \overline{\Lambda}^k(P_1\circ P_2,\mathscr{G}_1\oplus\mathscr{G}_2)\to\overline{\Lambda}^k(P_1\circ P_2,\mathscr{G}_i)$ induced by the projections $\mathscr{G}_1\oplus\mathscr{G}_2\to\mathscr{G}_i$, there is a unique form $\alpha_i\in\overline{\Lambda}^k(P_i,\mathscr{G}_i)$ such that $\pi^{i*}\alpha_i = \overline{\mathscr{G}}_i(\alpha)$, and $\alpha = \overline{\mathscr{G}}_1(\alpha)\oplus\overline{\mathscr{G}}_2(\alpha) = \pi^{1*}(\alpha_1)\oplus\pi^{2*}(\alpha_2)$.

Moreover, if we let $\mathring{\overline{\varphi}}^1,\ldots,\mathring{\overline{\varphi}}^{n+4}$ be one-forms dual to $\mathring{E}_1,\ldots,\mathring{E}_{n+4}$, and recall that $\Omega^{\mathring{\omega}_1\oplus\mathring{\omega}_2}$ vanishes on vertical vectors, we can write [making use of (2.18) and (2.21)]

$$\begin{aligned}\Omega^{\mathring{\omega}_1\oplus\mathring{\omega}_2} &= (\mathring{\Omega}_1)^{\alpha}{}_{ij}(l_\alpha)\otimes(\pi^{1*}\overline{\varphi}^{\,i}_{(1)})\wedge(\pi^{1*}\overline{\varphi}^{\,j}_{(1)})\\ &\quad\oplus(\mathring{\Omega}_2)_{ij}(i)\otimes(\pi^{2*}\overline{\varphi}^{\,i}_{(2)})\wedge(\pi^{2*}\overline{\varphi}^{\,j}_{(2)}),\end{aligned}\quad (2.31)$$

where

$$(\mathring{\Omega}_1)^{\alpha}{}_{ij} = (\Omega_1)^{\alpha}{}_{ij}\circ\pi^1, \quad (\mathring{\Omega}_2)_{ij} = (\Omega_2)_{ij}\circ\pi^2.$$

Note that $\pi^1_*\mathring{E}_i = E_i^{(1)}$ and $\pi^2_*\mathring{E}_i = E_i^{(2)}$. In fact,

$$(\mathring{\omega}_1\oplus\mathring{\omega}_2)(\mathring{E}_i) = \mathring{\omega}_1(\mathring{E}_i)\oplus\mathring{\omega}_2(\mathring{E}_i) = 0, \quad (2.32)$$

so $\mathring{\omega}_1(\mathring{E}_i) = 0$ and $\mathring{\omega}_2(\mathring{E}_i) = 0$. Moreover, $\pi_{1*}\pi^{1}_{*}\mathring{E}_i = (\pi_{12})_*\mathring{E}_i = \overline{E}_i$ and $\omega_1(\pi^1_*\mathring{E}_i) = (\pi^1)^*\omega_1(\mathring{E}_i) = \mathring{\omega}_1(\mathring{E}_i) = 0$. Therefore $\pi^1_*\mathring{E}_i$ is horizontal in $T(P_1)$ and projects onto $\overline{E}_i$; hence $\pi^1_*\mathring{E}_i = E_i^{(1)}$. In a similar way $\pi^2_*\mathring{E}_i = E_i^{(2)}$.

From the above it follows that

$$\begin{aligned}\delta^i_j = \overline{\varphi}^{\,i}_{(1)}(E_j^{(1)}) &= \overline{\varphi}^{\,i}_{(1)}(\pi^1_*\mathring{E}_j) = (\pi^{1*}\overline{\varphi}^{\,i}_{(1)})(\mathring{E}_j)\\ &= (\pi^{2*}\overline{\varphi}^{\,i}_{(2)})(\mathring{E}_j),\end{aligned}$$

i.e.,

$$\pi^{1*}\overline{\varphi}^{\,i}_{(1)} = \pi^{2*}\overline{\varphi}^{\,i}_{(2)} = \mathring{\overline{\varphi}}^{\,i}. \quad (2.33)$$

Consequently (2.31) becomes

$$\Omega^{\mathring{\omega}_1\oplus\mathring{\omega}_2} = [(\mathring{\Omega}_1)^{\alpha}{}_{ij}l_\alpha\oplus(\mathring{\Omega}_2)_{ij}i]\otimes\mathring{\overline{\varphi}}^{\,i}\wedge\mathring{\overline{\varphi}}^{\,j}. \quad (2.34)$$

Note also that (2.27) implies

$$\mathring{\omega}_1^{\alpha} = \mathring{\overline{\varphi}}^{\,n+\alpha}, \quad (-i\mathring{\omega}_2) = \mathring{\overline{\varphi}}^{\,n+4}. \quad (2.35)$$

The one other construction that appears in our diagram is the orthonormal bundle of frames $\Pi\colon F(P_1\circ P_2)\to P_1\circ P_2$, for which the manifold $P_1\circ P_2$, which has just been described, acts as a base manifold.

If we let $\theta(h)\in\Lambda^1(F(P_1\circ P_2),\mathscr{O}(r+4,s))$ denote a general connection on $F(P_1\circ P_2)$, we can now choose the vectors $\mathring{E}_1,\ldots,\mathring{E}_{n+4}$ as an orthonormal frame for the horizontal subspace $T_{(p_1,p_2)}F(P_1\circ P_2)$ relative to $\theta(h)$.

Furthermore, if $\Omega = D^{\theta(h)}\theta(h)\in\overline{\Lambda}^2(F(P_1\circ P_2),\mathscr{O}(r+4,s))$ is the curvature of $\theta(h)$, $\tilde{\sigma}\colon \Pi^{-1}(U)\to F(P_1\circ P_2)$ is a local section determined by the above orthonormal fields, and $\overline{e}_a$ are standard horizontal vectors on $F(P_1\circ P_2)$ associated with $e_a\in\Re^{n+4}$, then

$$\Omega^{\theta(h)}(\overline{e}_c,\overline{e}_d)(e_b) = R^{a}{}_{bcd}(\tilde{\sigma}(p_1,p_2))e_a. \quad (2.36)$$

Note on the other hand that

$$\Pi_*\tilde{\sigma}_*\mathring{E}_c = \mathring{E}_c = \Pi_*\overline{e}_c. \quad (2.37)$$

Thus $\overline{e}_c$ and $\tilde{\sigma}_*\mathring{E}_c$ differ at most by a vertical vector on $T_{\tilde{\sigma}(p_1,p_2)}F(P_1\circ P_2)$, but $\Omega^{\theta(g)}$ vanishes on vertical vectors, so

$$\begin{aligned}\Omega^{\theta(h)}(\overline{e}_c,\overline{e}_d) &= \Omega^{\theta(h)}(\tilde{\sigma}_*\mathring{E}_c,\tilde{\sigma}_*\mathring{E}_d)\\ &= (d\bar{\theta}(h) + \bar{\theta}(h)\,\dot{\wedge}\,\bar{\theta}(h))(\mathring{E}_c,\mathring{E}_d)\\ &= \Omega^{\bar{\theta}(h)}(\mathring{E}_c,\mathring{E}_d).\end{aligned}\quad (2.38)$$

If we now write

$$(\Omega^{\bar{\theta}(h)})^{a}{}_{b} = \tfrac{1}{2}\mathscr{R}^{a}{}_{bcd}(p_1,p_2)\mathring{\overline{\varphi}}^{\,c}\wedge\mathring{\overline{\varphi}}^{\,d}, \quad (2.39)$$

or

$$\Omega^{\bar{\theta}(h)} = \tfrac{1}{2}\mathscr{R}^{a}{}_{bcd}(p_1,p_2)e_a\otimes\hat{e}^b\otimes(\mathring{\overline{\varphi}}^{\,c}\wedge\mathring{\overline{\varphi}}^{\,d}), \quad (2.40)$$

and make use of (2.36), we get

$$R^{a}{}_{bcd}(\tilde{\sigma}(p_1,p_2)) = \mathscr{R}^{a}{}_{bcd}(p_1,p_2). \quad (2.41)$$

This last expression is the equivalent result in $\Pi\colon F(P_1\circ P_2)\to(P_1\circ P_2)$ to our previous Lemma 1.

In analogy to (2.7) and (2.8) we can use the local section $\tilde{\sigma}$ to define canonical one-forms $\mathring{\varphi}\in\overline{\Lambda}^1(F(P_1\circ P_2),\Re^{n+4})$ such that

$$\tilde{\sigma}^*\mathring{\varphi} = \mathring{\overline{\varphi}}^{\,a}e_a. \quad (2.42)$$

Corresponding to $\mathring{\varphi}$, we have that the torsion two-form $\mathring{\Theta}\in\overline{\Lambda}^2(F(P_1\circ P_2),\Re^{n+4})$ of $\theta(h)$ is given by

$$\mathring{\Theta}^{\theta(h)}\equiv D^{\theta(h)}\mathring{\varphi} = d\mathring{\varphi} + \theta(h)\,\dot{\wedge}\,\mathring{\varphi}. \quad (2.43)$$

Since $\mathring{\Theta}^{\theta(h)}$ is $\Re^{n+4}$ valued, we can write

$$\mathring{\Theta}^{\theta(h)}(\tilde{\sigma}_*\mathring{E}_c,\tilde{\sigma}_*\mathring{E}_d) = S^{a}{}_{cd}(\tilde{\sigma}(p_1,p_2))e_a. \quad (2.44)$$

However, we also have

$$\begin{aligned}\mathring{\Theta}^{\theta(h)}(\tilde{\sigma}_*\mathring{E}_c,\tilde{\sigma}_*\mathring{E}_d) &= (\tilde{\sigma}^*\mathring{\Theta}^{\theta(h)})(\mathring{E}_c,\mathring{E}_d)\\ &\equiv\mathring{\overline{\Theta}}^{\bar{\theta}(h)}(\mathring{E}_c,\mathring{E}_d)\\ &= (d\mathring{\overline{\varphi}} + \bar{\theta}(h)\,\dot{\wedge}\,\mathring{\overline{\varphi}})(\mathring{E}_c,\mathring{E}_d)\\ &= [\tfrac{1}{2}\mathscr{S}^{a}{}_{bf}(p_1,p_2)e_a\\ &\quad\otimes(\mathring{\overline{\varphi}}^{\,b}\wedge\mathring{\overline{\varphi}}^{\,f})](\mathring{E}_c,\mathring{E}_d),\end{aligned}\quad (2.45)$$

where $\bar{\theta}(h)\in\Lambda^1(P_1\circ P_2,\mathscr{O}(r+4,s))$ is the pullback with $\tilde{\sigma}$ of the connection $\theta(h)$. Comparing (2.44) and (2.45) we get

$$\mathscr{S}^{a}{}_{cd}(p_1,p_2) = S^{a}{}_{cd}(\tilde{\sigma}(p_1,p_2)), \quad (2.46)$$

which is the analog of Lemma (2.15) derived before.

As we mentioned in the Introduction, a judicious choice on the transformation properties of the torsion tensor is needed in order that the Higgs fields, which will originate

from the torsion itself, couple correctly with the Yang–Mills fields.

For this purpose let $g\in SU(2)\times U(1)$ and define the linear transformation $t_0(g)$: $\mathcal{G}_1\oplus\mathcal{G}_2\to\mathcal{G}_1\otimes\mathcal{G}_2$, by

$$t_0(g)V=\rho(g)\circ\mathcal{A}\mathfrak{d}_g V. \tag{2.47}$$

Here $V\in\mathcal{G}_1\oplus\mathcal{G}_2$ and $\rho(g)$ is a 4×4 real-matrix representation of g. The application of $\rho(g)$ on the basis element of $\mathcal{G}_1\oplus\mathcal{G}_2$ is given by

$$\rho(g)\circ\mathring{l}_A\equiv(\rho(g))_A{}^B\mathring{l}_B,\quad A,B=1,\ldots,4, \tag{2.48}$$

where $\mathring{l}_\alpha=(l_\alpha\oplus0)$, $\alpha=1,2,3$, and $\mathring{l}_4=(0\oplus i)$.

It is easy to verify that the following matrices are appropriate real linear representations of the infinitesimal generators of $SU(2)\times U(1)$:

$$\rho(\mathring{l}_1)=-\frac{1}{2}\begin{pmatrix}0&0&0&1\\0&0&-1&0\\0&1&0&0\\-1&0&0&0\end{pmatrix}=-\frac{1}{2}\sigma_1\otimes i\sigma_2, \tag{2.49a}$$

$$\rho(\mathring{l}_2)=\frac{1}{2}\begin{pmatrix}0&0&-1&0\\0&0&0&-1\\1&0&0&0\\0&1&0&0\end{pmatrix}=-\frac{1}{2}i\sigma_2\otimes I_2, \tag{2.49b}$$

$$\rho(\mathring{l}_3)=-\frac{1}{2}\begin{pmatrix}0&1&0&0\\-1&0&0&0\\0&0&0&-1\\0&0&1&0\end{pmatrix}=-\frac{1}{2}\sigma_3\otimes i\sigma_2, \tag{2.49c}$$

$$\rho(\mathring{l}_4)=-\frac{1}{2}\begin{pmatrix}0&1&0&0\\-1&0&0&0\\0&0&0&1\\0&0&-1&0\end{pmatrix}=-\frac{1}{2}I_2\otimes i\sigma_2. \tag{2.49d}$$

Given $t_0(g)$ we can define a transformation $t(g)$: $T(P_1\circ P_2)\to T(P_1\circ P_2)$ in the following way: Let X be a vector field in $T(P_1\circ P_2)$, and at each $(p_1,p_2)\in P_1\circ P_2$ let $X_H(p_1,p_2)$ and $X_V(p_1,p_2)$ be the horizontal and vertical components of $X(p_1,p_2)$. Write $X_V(p_1,p_2)=V^*_{p_1,p_2}$ for some $V\in\mathcal{G}_1\oplus\mathcal{G}_2$. We then require that

$$t_{(p_1,p_2)}(g)(X_H(p_1,p_2))=X_H(p_1,p_2), \tag{2.50a}$$

and

$$t_{(p_1,p_2)}(g)(X_V(p_1,p_2))=(t_0(g)V)^*_{(p_1,p_2)}. \tag{2.50b}$$

Finally, we also have the isomorphism $\hat{t}(g)$: $TF(P_1\circ P_2)\to TF(P_1\circ P_2)$, which is in turn induced by $t(g)$ according to the commutative diagram

$$\begin{array}{ccc} TF(P_1\circ P_2) & \xrightarrow{\hat{t}(g)} & TF(P_1\circ P_2)\\ \downarrow\Pi_* & & \Pi_*\downarrow\\ T(P_1\circ P_2) & \xrightarrow{t(g)} & T(P_1\circ P_2)\end{array}.$$

Hence

$$\Pi_*(\hat{t}(g)\tilde{X})=t(g)(\Pi_*\tilde{X}),\quad\forall\tilde{X}\in TF(P_1\circ P_2). \tag{2.51}$$

Making use of the definitions (2.47)–(2.51), and letting $\hat{R}_g$ denote the right action diffeomorphism $\hat{R}_g$: $F(P_1\circ P_2)\to F(P_1\circ P_2)$, we now prescribe the following transformation for torsion:

$$\hat{R}_g{}^*\mathring{\Theta}(\tilde{X},\tilde{Y})=\hat{t}(g^{-1})\circ\mathring{\Theta}\tilde{X},\tilde{Y})=\mathring{\Theta}(\hat{t}(g^{-1})\tilde{X},\hat{t}(g^{-1})\tilde{Y}). \tag{2.52}$$

In order to calculate $\hat{R}_g{}^*\mathring{\Theta}(\tilde{\sigma}_*X,\tilde{\sigma}_*Y)$, with $X,Y\in T(P_1\circ P_2)$, note first that we can identify $\mathbb{R}^{n+4}$ with $\mathbb{R}^n\oplus\mathcal{G}_1\oplus\mathcal{G}_2$ by means of a vector space isomorphism

$$(e_i,<\ ,\ >)\to(e_i,<\ ,\ >),\quad i=1,\ldots,n,$$
$$(e_{n+A},<\ ,\ >)\to(\mathring{l}_A,k_1\oplus k_2),$$

where $\{e_a,\ a=1,\ldots,n+4\}$ is the canonical basis of $\mathbb{R}^{n+4}$, $<\ ,\ >$ denotes the standard scalar product in $\mathbb{R}^{n+4}$, and $\mathring{l}_A$, $A=1,\ldots,4$ are the basis elements of the Lie algebra for $\mathcal{G}_1\oplus\mathcal{G}_2$ which we introduced earlier. This in turn allows us to define an application

$$a\colon SU(2)\times U(1)\to O(\mathbb{R}^n\oplus\mathcal{G}_1\oplus\mathcal{G}_2)\cong O(\mathbb{R}^{n+4}),$$

such that

$$a(g)=\begin{pmatrix}Id_{\mathbb{R}^n}&0\\0&\mathcal{A}\mathfrak{d}_g\end{pmatrix}. \tag{2.53}$$

It then follows that

$$R_{a(g^{-1})}(\tilde{\sigma}\circ R_g)=\hat{R}_g(\tilde{\sigma}). \tag{2.54}$$

In fact, acting with the left side of the above expression on e_i gives

$$\begin{aligned}&R_{a(g^{-1})}(\tilde{\sigma}\circ R_g)(p_1,p_2)(e_i)\\&=\tilde{\sigma}_{(p_1,p_2)g}\circ a(g^{-1})(e_i)=\tilde{\sigma}_{(p_1,p_2)g}(e_i)\\&=\mathring{E}_i((p_1,p_2)g)=(R_{g^*}\mathring{E}_i)_{(p_1,p_2)}\\&=R_{g^*}\tilde{\sigma}_{(p_1,p_2)}(e_i)=\hat{R}_g(\tilde{\sigma})(p_1,p_2)(e_i),\end{aligned}$$

and acting with the left side of (2.54) on e_{n+A} yields

$$\begin{aligned}&R_{a(g^{-1})}(\tilde{\sigma}\circ R_g)(p_1,p_2)(e_{n+A})\\&=\tilde{\sigma}_{(p_1,p_2)g}\circ(\mathcal{A}\mathfrak{d}_{g^{-1}}l_A)=(\mathcal{A}\mathfrak{d}_{g^{-1}}l_A)^*_{(p_1,p_2)g}\\&=R_{g^*}(l_A)^*_{(p_1,p_2)}=R_{g^*}\tilde{\sigma}_{(p_1,p_2)}(e_{n+a})\\&=\hat{R}_g(\tilde{\sigma})(p_1,p_2)(e_{n+A}).\end{aligned}$$

Equation (2.54) then follows, and therefore

$$\begin{aligned}&\hat{R}_g{}^*\mathring{\Theta}(\tilde{\sigma}_*X,\tilde{\sigma}_*Y)\\&\equiv\mathring{\Theta}(\hat{R}_{g^*}\tilde{\sigma}_*X,\hat{R}_{g^*}\tilde{\sigma}_*Y)\\&=\mathring{\Theta}(R_{a(g^{-1})_*}\tilde{\sigma}_*R_{g^*}X,R_{a(g^{-1})_*}\tilde{\sigma}_*R_{g^*}Y)\\&=a(g)\cdot\mathring{\Theta}(R_{g^*}X,R_{g^*}Y).\end{aligned} \tag{2.55}$$

Consequently (2.52) is equivalent to

$$\mathring{\Theta}(R_{g^*}X,R_{g^*}Y) = a(g^{-1})\cdot\mathring{\Theta}(\hat{t}(g^{-1})\tilde{\sigma}_*X,\hat{t}(g^{-1})\tilde{\sigma}_*Y). \tag{2.56}$$

Furthermore, $\hat{t}(g^{-1})\tilde{\sigma}_*X$ and $\tilde{\sigma}_*t(g^{-1})X$ are both in $T_{(p_1,p_2)}F(P_1\circ P_2)$, and

$$\Pi_*\hat{t}(g^{-1})\tilde{\sigma}_*X = t(g^{-1})\Pi_*\tilde{\sigma}_*X$$
$$= t(g^{-1})X = \Pi_*\sigma_*t(g^{-1})X.$$

Thus $\hat{t}(g^{-1})\tilde{\sigma}_*X$ and $\tilde{\sigma}_*t(g^{-1})X$ differ at most by a vertical vector. However, since $\mathring{\Theta}$ vanishes on vertical vectors, we may write (2.56) as

$$\mathring{\Theta}_{(p_1,p_2)g}(R_{g^*}X,R_{g^*}Y)$$
$$= a(g^{-1})\cdot\mathring{\Theta}_{(p_1,p_2)}(t(g^{-1})X,t(g^{-1})Y). \tag{2.57}$$

We can carry our analysis of the transformation properties of the torsion further by recalling that the difference between any two connections vanishes on vertical vectors. Consequently, if we write $\theta(h) = \tau(h) + \theta(h)_{\mathrm{LC}}$, where $\theta(h)_{\mathrm{LC}}$ is the unique Levi-Civita connection and $\tau(h)\in\bar{\Lambda}^1(F(P_1\circ P_2),\mathcal{O}(r+4,s))$, we have

$$\mathring{\Theta}^{\theta(h)} = d\mathring{\varphi} + (\theta(h)_{\mathrm{LC}} + \tau(h))\wedge\mathring{\varphi}$$
$$= \mathring{\Theta}^{\theta(h)_{\mathrm{LC}}} + \tau(h)\wedge\mathring{\varphi}$$
$$= \tau(h)\wedge\mathring{\varphi} \tag{2.58}$$

(since the torsion from the Levi-Civita connection vanishes).

If we now let $\tilde{\sigma}^*\tau(h) = \bar{\tau}(h)$, and observe that

$$(\bar{\tau}\wedge\mathring{\bar{\varphi}})(X,Y) = \bar{\tau}(X)\cdot\mathring{\bar{\varphi}}(Y) - \bar{\tau}(Y)\cdot\mathring{\bar{\varphi}}(X),$$

we obtain from (2.56) the result

$$\bar{\tau}_{(p_1,p_2)g}(R_{g^*}X)\cdot\mathring{\bar{\varphi}}_{(p_1,p_2)g}(R_{g^*}Y)$$
$$- \bar{\tau}_{(p_1,p_2)g}(R_{g^*}Y)\cdot\mathring{\bar{\varphi}}_{(p_1,p_2)g}(R_{g^*}X)$$
$$= a(g^{-1})\cdot\bar{\tau}_{(p_1,p_2)}(t(g^{-1})X)\cdot\mathring{\bar{\varphi}}_{(p_1,p_2)}(t(g^{-1})Y)$$
$$- a(g^{-1})\cdot\bar{\tau}_{(p_1,p_2)}(t(g^{-1})Y)\cdot\mathring{\bar{\varphi}}_{(p_1,p_2)}(t(g^{-1})X). \tag{2.59}$$

Furthermore, since $\bar{\tau}\in\Lambda^1(P_1\circ P_2,\mathcal{O}(r+4,s))$, we can write [comparing with (2.45)]

$$\bar{\tau}_{(p_1,p_2)} = \tfrac{1}{2}\mathscr{S}^a{}_{bc}(p_1,p_2)\mathring{\bar{\varphi}}^b{}_{(p_1,p_2)}e_a\otimes\hat{e}^c. \tag{2.60}$$

Substituting (2.60) in (2.59) and making use of (2.50), we get the following.

(i) For $X = \mathring{E}_i$, $Y = \mathring{E}_j$ (X,Y both horizontal)

$$\mathscr{S}^a{}_{ij}((p_1,p_2)g) = a(g^{-1})^a{}_b\mathscr{S}^b{}_{ij}(p_1,p_2), \tag{2.61}$$

with $a(g^{-1})^a{}_b = \hat{e}^a(a(g^{-1})(e_b))$.

(ii) For $X = \mathring{E}_i$, $Y = \mathring{E}_{n+A}$ (X horizontal and Y vertical),

$$\mathscr{S}^a{}_{i\,n+A}((p_1,p_2)g) = (\rho(g^{-1}))_A{}^B a(g^{-1})^a{}_b\mathscr{S}^b{}_{i\,n+B}(p_1,p_2). \tag{2.62}$$

(iii) For $X = \mathring{E}_{n+A}$, $Y = \mathring{E}_{n+B}$ (X,Y both vertical),

$$\mathscr{S}^a{}_{n+A\;n+B}((p_1,p_2)g)$$
$$= (\rho(g^{-1}))_A{}^C(\rho(g^{-1}))_B{}^D a(g^{-1})^a{}_b$$
$$\times\mathscr{S}^b{}_{n+C\;n+D}(p_1,p_2). \tag{2.63}$$

To be mathematically more precise, in (2.58) we should actually write $\theta(h) = \tau_1(h) + \tau_2(h) + (\theta(h))_{\mathrm{LC}}$. So that $\mathring{\Theta}^{\theta(h)} = \tau_1(h)\wedge\mathring{\varphi} + \tau_2(h)\wedge\mathring{\varphi} = \mathring{\Theta}^{\theta_1(h)} + \mathring{\Theta}^{\theta_2(h)}$, where $\theta_i(h) = \tau_i(h) + (\theta(h))_{\mathrm{LC}}$, $i = 1,2$. We would then have that $\mathring{\Theta}^{\theta_1(h)}$ transforms according to (2.52), and for $\mathring{\Theta}_2 = \mathring{\Theta}^{\theta_2(h)}$ we have

$$\hat{R}_g^*\mathring{\Theta}_2(X,Y) = a(g)\mathring{\Theta}_2(\hat{t}_0(g^{-1})X,\hat{t}_0(g^{-1})Y)$$

with $t_0(g) = \mathscr{A}\mathfrak{d}_g$. Observe, however, that $\mathscr{S}^a{}_{bc} = \overline{\mathscr{S}}^a{}_{bc} + \overline{\mathscr{S}}^a{}_{bc}$, where $\overline{\mathscr{S}}^a{}_{bc}$ and $\overline{\overline{\mathscr{S}}}^a{}_{bc}$ are the associated tensors of $\mathring{\Theta}^{\theta_1(h)}$ and $\mathring{\Theta}^{\theta_2(h)}$, respectively. The first one transforms according to (2.60)–(2.63), while the second one has to be constant on each fiber.

With these basic definitions and results we are now ready to compute the components relative to $\mathring{E}_1,\ldots,\mathring{E}_n$, $\mathring{E}_{n+1},\ldots,\mathring{E}_{n+4}$ of the curvature and torsion tensors for the metric h on $P_1\circ P_2$. This we shall do in the following section.

III. THE UNIFIED LAGRANGIAN

Recall that the components of the Riemann tensor on $P_1\circ P_2$ are related to the connection one-forms $\bar{\theta}(h) = \tilde{\sigma}^*\theta(h)$ by means of (2.38) and (2.39). In matrix notation these equations lead to

$$\tfrac{1}{2}\mathscr{R}^a{}_{bcd}(p_1,p_2)\mathring{\bar{\varphi}}^c\wedge\mathring{\bar{\varphi}}^d = d\bar{\theta}(h)^a{}_b + \bar{\theta}(h)^a{}_e\wedge\bar{\theta}(h)^e{}_b. \tag{3.1}$$

Therefore, in order to evaluate $\mathscr{R}^a{}_{bcd}$ we need first to calculate the various matrix terms $\bar{\theta}(h)^a{}_b$ for $1\leqslant a,b\leqslant n+4$ relative to the choice of orthonormal basis described in the preceding section. To do this we make use of (2.43) and (2.13) to write

$$d\mathring{\bar{\varphi}}^a = \mathring{\bar{\Theta}}^a - \bar{\theta}(h)^a{}_b\wedge\mathring{\bar{\varphi}}^b, \tag{3.2}$$

$$d\bar{\varphi}^i{}_M = \bar{\Theta}^i - \bar{\theta}(g)^i{}_j\wedge\bar{\varphi}^j{}_M. \tag{3.3}$$

Moreover, since $\bar{\varphi}^i{}_M(\mathring{\bar{E}}_j) = \delta^i_j = \bar{\varphi}^i{}_M(\pi_{12*}\mathring{E}_j)$, we have that $\pi_{12}{}^*\bar{\varphi}^i{}_M = \mathring{\bar{\varphi}}^i$. Consequently, pulling back (3.3) with π_{12} yields

$$d\mathring{\bar{\varphi}}^i = \mathring{\bar{\Theta}}^i(g) - \pi_{12}{}^*\bar{\theta}(g)^i{}_j\wedge\mathring{\bar{\varphi}}^j. \tag{3.4}$$

Note parenthetically that this last expression implies

$$-\mathring{\bar{\varphi}}^i([\mathring{E}_j,\mathring{E}_k]) = \underset{\sim}{S}{}^i{}_{jk}(x). \tag{3.5}$$

Thus if we impose the restriction of vanishing torsion on the base space M, i.e.,

$$\underset{\sim}{S}{}^i{}_{jk} = 0, \tag{3.6}$$

then it immediately follows that the commutator of the basis vectors $\mathring{E}_i$, $i = 1,\ldots,n$, has to be vertical. We shall use this result later on for deriving the form of the covariant derivative of the Higgs fields.

Now, from (2.35) and (3.2) we have

$$d(-i\mathring{\omega}_2) = \mathring{\bar{\Theta}}^{n+4} - \bar{\theta}(h)^{n+4}{}_b\wedge\mathring{\bar{\varphi}}^b,$$
$$d\mathring{\omega}_1^\alpha = \mathring{\bar{\Theta}}^{n+\alpha} - \bar{\theta}(h)^{n+\alpha}{}_b\wedge\mathring{\bar{\varphi}}^b.$$

Substituting on the left side of these equations the pullback with π^2 and π^1, respectively, of (2.20) and (2.17), gives

$$\tfrac{1}{2}(\mathring{\Omega}_2)_{ij}\mathring{\bar{\varphi}}^i\wedge\mathring{\bar{\varphi}}^j = \mathring{\bar{\Theta}}^{n+4} - \bar{\theta}(h)^{n+4}{}_b\wedge\mathring{\bar{\varphi}}^b, \tag{3.7}$$

$$\tfrac{1}{2}(\mathring{\Omega}_1)^\alpha{}_{ij}\mathring{\bar{\varphi}}^i\wedge\mathring{\bar{\varphi}}^j - \tfrac{1}{2}c^\alpha{}_{\beta\gamma}\mathring{\bar{\varphi}}^\beta\wedge\mathring{\bar{\varphi}}^\gamma$$
$$= \mathring{\bar{\Theta}}^{n+\alpha} - \bar{\theta}(h)^{n+\alpha}{}_b\wedge\mathring{\bar{\varphi}}^b. \tag{3.8}$$

Furthermore, since $\bar{\theta}(h)$ is $\mathcal{O}(r+4,s)$-valued, the matrix elements $\bar{\theta}(h)^a{}_b$ must satisfy the constraint

$$\bar{\theta}(h)_{ab} + \bar{\theta}(h)_{ba} = 0. \tag{3.9}$$

This condition is fulfilled if we require that

$$\bar{\theta}(h)^{n+A}{}_{n+B} = -(\rho(\mathring{l}_C))^A{}_B \mathring{\bar{\varphi}}^{n+C}, \tag{3.10}$$

where the matrices $\rho(\mathring{l}_C)$ are the infinitesimal generators of $SU(2)\times U(1)$ which we explicitly displayed in (2.49a)–(2.49d).

It is important to note here the fact that (3.10) does not fully specify the connection. The remaining freedom is clearly manifest in the equations that relate some of the connection coefficients to undefined components of the torsion tensor.

Since in the end we want the Higgs fields to originate geometrically from torsion, we make the additional assumption that our connection is semisymmetric.[7] It has been shown (cf. Theorems 1 and 2 of Ref. 8) that this assumption is tantamount to essentially taking only the first two terms in a unique decomposition for the torsion tensor. Introducing further terms resulting from a spin-tensor H (described in the paper referred to above) provides a way to generalize our results in the sense that additional fields appear, whose physical meaning remains to be determined, and would make it also possible to investigate nonmetric theories within the framework of our formalism.

In the context of the semisymmetry assumption we have that

$$\begin{aligned}\mathcal{S}^{n+\alpha}{}_{n+4\,i} &= \mathcal{S}^{n+4}{}_{n+\alpha\,i} = \mathcal{S}^{i}{}_{n+4\,n+\alpha} \\ &= \mathcal{S}^{i}{}_{n+\alpha\,n+\beta} = 0,\end{aligned} \tag{3.11}$$

and since by our argument following Eq. (2.63) it is reasonable to assume that each of the connections $\theta_1(h)$ and $\theta_2(h)$ in the decomposition $\mathring{\Theta}^{\theta(h)} = \mathring{\Theta}^{\theta_1(h)} + \mathring{\Theta}^{\theta_2(h)}$ are to be semisymmetric, it follows that the torsion components $\overline{\mathcal{S}}^{n+\alpha}{}_{n+\beta\,i}$, $\overline{\mathcal{S}}^{n+4}{}_{n+4\,i}$, $\overline{\overline{\mathcal{S}}}^{n+\alpha}{}_{n+\beta\,i}$, $\overline{\overline{\mathcal{S}}}^{n+4}{}_{n+4\,i}$, must be proportional to quantities of the form $\delta^{n+\alpha}_{n+\beta}\Phi_i$ and $\delta^{n+4}_{n+4}\Phi_i$, respectively, where Φ_i is an additional vector field. Using (2.62) we obtain that if $\overline{\mathcal{S}}^{n+\alpha}{}_{n+\beta\,i} = c\delta^\alpha_\beta\Phi_i \neq 0$, then $c\delta^\beta_\alpha\Phi_i = \rho(g^{-1})_\alpha{}^\lambda a(g^{-1})^\beta{}_\sigma\delta^\sigma_\lambda\Phi_i c$, from where it follows that $\rho(g^{-1})_\alpha{}^\lambda a(g^{-1})^\beta{}_\lambda = \delta^\beta_\alpha$, i.e., $\rho(g^{-1}) = a(g)$, which is impossible because G is a non-Abelian group. Thus we have $\overline{\mathcal{S}}^{n+\alpha}{}_{n+\beta\,i} = 0$ and, as a consequence, $\mathcal{S}^{n+\alpha}{}_{n+\beta i} = \overline{\overline{\mathcal{S}}}^{n+\alpha}{}_{n+\beta}$ must be proportional to $\delta^\alpha_\beta\Phi_i$ with Φ_i constant on each fiber. This vector field could be considered if one wished to generalize our present results. However, for the purposes stated above, we choose to set Φ_i equal to zero in view of the fact that it is obviously not a Higgs field, and that we can use the remaining freedom that we still have in selecting our connection.

Using (3.10) and (3.11) in (3.7) and evaluating on $(\mathring{E}_{n+4},\mathring{E}_{n+4})$, $(\mathring{E}_{n+4},\mathring{E}_{n+\gamma})$, $(\mathring{E}_{n+4},\mathring{E}_i)$, $(\mathring{E}_{n+\alpha},\mathring{E}_i)$, $(\mathring{E}_{n+\alpha},\mathring{E}_{n+\beta})$, and $(\mathring{E}_i,\mathring{E}_j)$, yields, respectively,

$$\begin{aligned}&\text{(a) } \mathcal{S}^{n+4}{}_{n+4\,n+4} = 0, \\ &\text{(b) } \mathcal{S}^{n+4}{}_{n+4\,n+\gamma} = -(\rho(\mathring{l}_4))^4{}_\gamma,\end{aligned} \tag{3.12}$$

$$\begin{aligned}&\text{(a) } \bar{\theta}(h)^{n+4}{}_i(\mathring{E}_{n+4}) = 0, \\ &\text{(b) } \bar{\theta}(h)^{n+4}{}_i(\mathring{E}_{n+\alpha}) = 0,\end{aligned} \tag{3.13}$$

$$\mathcal{S}^{n+4}{}_{n+\alpha\,n+\beta} = (\rho(\mathring{\ell}_\beta))^4{}_\alpha - (\rho(\mathring{\ell}_\alpha))^4{}_\beta, \tag{3.14}$$

$$(\mathring{\Omega}_2)_{ij} = \mathcal{S}^{n+4}{}_{ij} - \bar{\theta}(h)^{n+4}{}_j(\mathring{E}_i) + \bar{\theta}(h)^{n+4}_i(\mathring{E}_j). \tag{3.15}$$

Making use of (3.13) and (3.15) it immediately follows that

$$\bar{\theta}(h)^{n+4}{}_i = \tfrac{1}{2}[(\mathring{\Omega}_2)_{ij} - \mathcal{S}^{n+4}{}_{ij}]\mathring{\bar{\varphi}}^j - \mathcal{S}_{ij}{}^{n+4}\mathring{\bar{\varphi}}^j, \tag{3.16a}$$

with

$$\mathcal{S}_{ij}{}^{n+4} = \mathcal{S}_{ji}{}^{n+4}. \tag{3.16b}$$

Also, because of (3.9),

$$\bar{\theta}(h)^i{}_{n+4} = -\tfrac{1}{2}[(\mathring{\Omega}_2)^i{}_j - \mathcal{S}_{n+4}{}^i{}_j]\mathring{\bar{\varphi}}^j + \mathcal{S}_j{}^i{}_{n+4}\mathring{\bar{\varphi}}^j. \tag{3.17}$$

Similarly, evaluating (3.8) on $(\mathring{E}_{n+\gamma},\mathring{E}_{n+4})$, $(\mathring{E}_{n+4},\mathring{E}_{n+4})$, $(\mathring{E}_{n+4},\mathring{E}_i)$, $(\mathring{E}_{n+\beta},\mathring{E}_i)$, $(\mathring{E}_{n+\beta},\mathring{E}_{n+\gamma})$, and $(\mathring{E}_i,\mathring{E}_j)$, we get, respectively,

$$\begin{aligned}&\text{(a) } \mathcal{S}^{n+\alpha}{}_{n+\gamma\,n+4} = -(\rho(\mathring{l}_\gamma))^\alpha{}_4 + (\rho(\mathring{l}_4))^\alpha{}_\gamma, \\ &\text{(b) } \mathcal{S}^{n+\alpha}{}_{n+4\,n+4} = 0,\end{aligned} \tag{3.18}$$

$$\begin{aligned}&\text{(a) } \bar{\theta}(h)^{n+\alpha}{}_i(\mathring{E}_{n+4}) = 0, \\ &\text{(b) } \bar{\theta}(h)^{n+\alpha}{}_i(\mathring{E}_{n+\beta}) = 0,\end{aligned} \tag{3.19}$$

$$\mathcal{S}^{n+\alpha}{}_{n+\beta\,n+\gamma} = (\rho(\mathring{l}_\gamma))^\alpha{}_\beta - (\rho(\mathring{l}_\beta))^\alpha{}_\gamma - c^\alpha{}_{\beta\gamma}, \tag{3.20}$$

$$(\mathring{\Omega}_1)^\alpha{}_{ij} = \mathcal{S}^{n+\alpha}{}_{ij} - \bar{\theta}(h)^{n+\alpha}{}_j(\mathring{E}_i) + \bar{\theta}(h)^{n+\alpha}{}_i(\mathring{E}_j). \tag{3.21}$$

It is obvious from (3.19) and (3.21) that

$$\bar{\theta}(h)^{n+\alpha}{}_i = \tfrac{1}{2}[(\mathring{\Omega}_1)^\alpha{}_{ij} - \mathcal{S}^{n+\alpha}{}_{ij}]\mathring{\bar{\varphi}}^j - \mathcal{S}_{ij}{}^{n+\alpha}\mathring{\bar{\varphi}}^j, \tag{3.22a}$$

with

$$\mathcal{S}_{ij}{}^{n+\alpha} = \mathcal{S}_{ji}{}^{n+\alpha}. \tag{3.22b}$$

Moreover, from (3.9) we also have

$$\bar{\theta}(h)^i{}_{n+\alpha} = -\tfrac{1}{2}[(\mathring{\Omega}_1)_\alpha{}^i{}_j - \mathcal{S}_{n+\alpha}{}^i{}_j]\mathring{\bar{\varphi}}^j + \mathcal{S}^i{}_{jn+\alpha}\mathring{\bar{\varphi}}^j. \tag{3.23}$$

The remaining expressions that we need for the connection coefficients are obtained by noting that (3.2) also implies that

$$\begin{aligned}d\mathring{\bar{\varphi}}^i = {}& \mathcal{S}^i{}_{jn+A}\mathring{\bar{\varphi}}^j \wedge \mathring{\bar{\varphi}}^{n+A} \\ &- \bar{\theta}(h)^i{}_j \wedge \mathring{\bar{\varphi}}^j - \bar{\theta}(h)^i{}_{n+A} \wedge \mathring{\bar{\varphi}}^{n+A}.\end{aligned} \tag{3.24}$$

Thus evaluating on $(\mathring{E}_j, \mathring{E}_{n+4})$ results in

$$\bar{\theta}(h)^i{}_j(\mathring{E}_{n+4}) = -\tfrac{1}{2}[(\mathring{\Omega}_2)^i{}_j - \mathcal{S}_{n+4}{}^i{}_j]. \tag{3.25}$$

Finally, note that substituting in (3.24) the values for $\bar{\theta}(h)^i{}_{n+\alpha}$ and $\bar{\theta}(h)^i{}_{n+4}$ given by (3.23) and (3.17), and equating the result to (3.4), yields

$$\begin{aligned}\bar{\theta}(h)^i{}_j = {}& \pi^*_{12}\bar{\theta}(g)^i{}_j - \tfrac{1}{2}[(\mathring{\Omega}_1)_\alpha{}^i{}_j - \mathcal{S}_{n+\alpha}{}^i{}_j]\mathring{\varphi}^{n+\alpha} \\ &- \tfrac{1}{2}[(\mathring{\Omega}_2)^i{}_j - \mathcal{S}_{n+4}{}^i{}_j]\mathring{\bar{\varphi}}^{n+4}.\end{aligned} \tag{3.26}$$

As we mentioned previously, the Higgs fields originate directly from torsion by assuming that the connection $\bar{\theta}(h)$ is semisymmetric. With this in mind, we make the following additional ansatz on the torsion:

$$\mathscr{S}_{ij}{}^{n+A} = (1/n)g_{ij}\Phi^A, \tag{3.27}$$

and

$$\mathscr{S}^{n+\alpha}{}_{ij} = (\mathring{\Omega}_1)^{\alpha}{}_{ij}, \quad \mathscr{S}^{n+4}{}_{ij} = (\mathring{\Omega}_2)_{ij}. \tag{3.28}$$

Note that with these last assumptions the connection matrices $\bar{\theta}(h)^a{}_b$, as given by (3.9), (3.10), (3.16), (3.22), and (3.26), are uniquely specified. Also, as we will show in Sec. IV, the four scalar fields Φ^A introduced in (3.27) can be identified with the real Higgs fields.

We now have all the ingredients that are needed to evaluate from (3.1) the components $\mathscr{R}^a{}_{bcd}$ of the curvature tensor of the metric h on $P_1 \circ P_2$ relative to our orthonormal basis $\mathring{E}_1,\ldots,\mathring{E}_{n+4}$. Since the calculation, although lengthy, is fairly straightforward, we only state the final results here:

$$\begin{aligned}
&\mathscr{R}^{n+\alpha}{}_{n+\beta\, n+\mu\, n+\nu} = 2\epsilon_{\gamma\mu\nu}(\rho(\mathring{l}_\gamma))^{\alpha}{}_{\beta},\\
&\mathscr{R}^{n+\alpha}{}_{n+\beta\, ij} = -(\rho(\mathring{l}_\gamma))^{\alpha}{}_{\beta}(\mathring{\Omega}_1)^{\gamma}{}_{ij} - (\rho(\mathring{l}_4))^{\alpha}{}_{\beta}(\mathring{\Omega}_2)_{ij},\\
&\mathscr{R}^{n+4}{}_{n+\alpha\, n+\mu\, n+\nu} = 2\epsilon_{\beta\mu\nu}(\rho(\mathring{l}_\beta))^{4}{}_{\alpha},\\
&\mathscr{R}^{n+4}{}_{n+\alpha\, ij} = -(\rho(\mathring{l}_\beta))^{4}{}_{\alpha}(\mathring{\Omega}_1)^{\beta}{}_{ij} - (\rho(\mathring{l}_4))^{4}{}_{\alpha}(\mathring{\Omega}_2)_{ij},
\end{aligned} \tag{3.29}$$

$$\begin{aligned}
&\mathscr{R}^{n+4}{}_{ikj} = (1/n)g_{ik}\mathring{E}_j[\Phi^4] - (1/n)g_{ij}\mathring{E}_k[\Phi^4],\\
&\mathscr{R}^{n+\alpha}{}_{ijk} = (1/n)g_{ik}\mathring{E}_j[\Phi^\alpha] - (1/n)g_{ij}\mathring{E}_k[\Phi^\alpha],\\
&\mathscr{R}^{i}{}_{jkm} = \underline{R}^{i}{}_{jkm} + (1/n)[\delta^i_m g_{jk} - \delta^i_k g_{jm}]\Phi_A\Phi^A,
\end{aligned}$$

and all other components vanish. In the above expressions terms of the form $\mathring{E}_i[\Phi^A] = d\Phi^A(\mathring{E}_i)$ denote directional derivatives. In Sec. IV we will show that this directional derivative is a covariant derivative, i.e.,

$$\mathring{E}_i[\Phi^A] = d\Phi^A(\mathring{E}_i) \equiv D_i\Phi^A. \tag{3.30}$$

Moreover, we will also establish the relation between the covariant derivatives of our four real-scalar fields Φ_A and the covariant derivative of the Higgs complex spin doublet as it commonly appears in the electroweak model.

For the construction of the Lagrangian density we also need the nonvanishing components of the Ricci tensor on $P_1 \circ P_2$ as well as the Ricci scalar. These follow directly from (3.29) and are given by

$$\begin{aligned}
&\mathscr{R}_{jm} = \underline{R}_{jm} + [(1-n)/n^2]g_{jm}\Phi_A\Phi^A,\\
&\mathscr{R}_{n+4\, n+\alpha} = 2\epsilon_{\alpha\gamma\beta}(\rho(\mathring{l}_\gamma))^{\beta}{}_4,\\
&\mathscr{R}_{n+4\, i} = [(1-n)/n]\mathring{E}_i[\Phi_4],\\
&\mathscr{R}_{n+\alpha\, n+\beta} = 2\epsilon_{\lambda\gamma\beta}(\rho(\mathring{l}_\lambda))^{\gamma}{}_{\alpha},\\
&\mathscr{R}_{n+\alpha\, i} = [(1-n)/n]\mathring{E}_i[\Phi_\alpha],\\
&\mathscr{R} = \underline{R} + [(1-n)/n]\Phi^A\Phi_A + 2.
\end{aligned} \tag{3.31}$$

General Lagrangian density: We construct the most general G-invariant Lagrangian density on $P_1 \circ P_2$ up to quadratic terms in the Riemann, Ricci, and torsion tensors as well as in the Ricci scalar by adding up all the G-invariant terms that can be obtained from Eqs. (3.14), (3.18), (3.20), (3.27)–(3.31). The result is

$$\begin{aligned}
\mathscr{L} = \frac{\sqrt{|g|}}{V_I}\Big\{&\alpha_0(\underline{R} - \frac{n-1}{n}\Phi^A\Phi_A + 2) + \alpha_1\Big[\frac{(n-1)^2}{n^2}(\Phi^A\Phi_A)^2 - \frac{2(n-1)}{n}\underline{R}\Phi_A\Phi^A - 4\frac{(n-1)}{n}\Phi^A\Phi_A + (\underline{R}+2)^2\Big]\\
&+ \alpha_2\Big[\underline{R}_{ijkm}\underline{R}^{ijkm} - \frac{4}{n^2}\underline{R}(\Phi_A\Phi^A) + \frac{2}{n^3}(n-1)(\Phi_A\Phi^A)^2\Big] - \alpha_3(\mathring{\Omega}_1)^{\gamma}{}_{ij}(\mathring{\Omega}_1)_{\gamma}{}^{ij} - \alpha_4(\mathring{\Omega}_2)_{ij}(\mathring{\Omega}_2)^{ij} + \alpha_5(D^i\Phi^A)(D_i\Phi_A)\\
&+ \alpha_6\Big[\underline{R}_{ij}\underline{R}^{ij} - \frac{2}{n^2}(n-1)\underline{R}(\Phi_A\Phi^A) + \frac{(n-1)^2}{n^3}(\Phi_A\Phi^A)^2\Big] + \alpha_7\frac{n-1}{n}\Phi_A\Phi^A + K\Big\},
\end{aligned} \tag{3.32}$$

where V_I is the volume of the n–4 compact "internal" coordinates of the base manifold, and K is a constant that contributes to the cosmological constant.

Before proceeding with the proper dimensioning and physical interpretation of the different terms and parameters in the above Lagrangian density, we show explicitly that all the entries in (3.32) are indeed G invariant. Clearly the terms containing the several contractions of the Riemann tensor are G invariant since, according to (2.9) or (2.10), the components $\underline{R}^{h}{}_{ijk}$ are defined on M and are therefore independent of the choice of point on the fiber. The quantities $(\mathring{\Omega}_1)_{\alpha ij}(\mathring{\Omega}_1)^{\alpha ij}$ and $(\mathring{\Omega}_2)_{ij}(\mathring{\Omega}_2)^{ij}$ are also G invariant because

$$\begin{aligned}
(\mathring{\Omega}_1)_{\alpha ij}(\mathring{\Omega}_1)^{\alpha ij} &= g^{ik}g^{jl}k_1((\Omega_1)(p_1)(E_i^{(1)},E_j^{(1)}),(\Omega_1)(p_1)(E_k^{(1)},E_l^{(1)}))\\
&= g^{ik}g^{jl}k_1(\mathscr{A}\mathfrak{d}_{g_1^{-1}}(\Omega_1)(p_1)(E_i^{(1)},E_j^{(1)}),\mathscr{A}\mathfrak{d}_{g_1^{-1}}(\Omega_1)(p_1)(E_k^{(1)},E_l^{(1)}))\\
&= g^{ik}g^{jl}k_1((\Omega_1)(p_1g_1)(R_{g_1*}E_i^{(1)},R_{g_1*}E_j^{(1)}),(\Omega_1)(p_1g_1)(R_{g_1*}E_k^{(1)},R_{g_1*}E_l^{(1)}))\\
&= g^{ik}g^{jl}k_1((\Omega_1)(p_1g_1)((E_i^{(1)})p_1g_1,(E_j^{(1)})p_1g_1),(\Omega_1)(p_1g_1)((E_k^{(1)})p_1g_1,(E_l^{(1)})p_1g_1)),
\end{aligned} \tag{3.33}$$

i.e., $k_1(\Omega_1,\Omega_1)$ is well defined on M since it is independent of the choice of p_1. An even simpler argument applies to $k_2(\Omega_2,\Omega_2)$ since in this case the group is Abelian.

To prove that $\Phi_A\Phi^A$ is G-invariant we need the transformation properties of the fields Φ^A. These follow readily from (2.62) and (3.27). We thus have

$$\Phi_A((p_1,p_2)g) = \rho(g^{-1})_A{}^B\Phi_B(p_1,p_2), \tag{3.34}$$

and since the matrices $\rho(g)$ are orthogonal [cf. Eqs. (2.49) for the infinitesimal generators], it immediately follows that

$$(\Phi_A\Phi^A)_{(p_1,p_2)g} = (\Phi_A\Phi^A)_{(p_1,p_2)}.$$

Finally, since $D_i\Phi_A \equiv \mathring{E}_i[\Phi_A]$, it is obvious that the term $(D_i\Phi_A)(D^i\Phi^A)$ is also independent of the point in the fiber where it is evaluated.

In summary, the Lagrangian density (3.32) is a well-

defined function on the base manifold M, and we can write an action by integrating it over a volume element μ_g on M determined by g and the orientation of M, i.e.,

$$I=\int_U \mathscr{L}\mu_g, \tag{3.35}$$

where U is an open subset of M with compact closure.

We now turn to the physical interpretation of the curvatures Ω_1 and Ω_2. Recall that by (2.17) and (2.20) we have

$$(\Omega_1)^{\alpha}{}_{ij}=(d\omega_1^{\alpha})(E_i^{(1)},E_j^{(1)})+\epsilon_{\alpha\beta\gamma}\omega_1^{\beta}(E_i^{(1)})\omega_1^{\gamma}(E_j^{(1)}), \tag{2.17'}$$

$$(\Omega_2)_{ij}=d(-i\omega_2)(E_i^{(2)},E_j^{(2)}). \tag{2.20'}$$

If we let $(\sigma_1)_u$ and $(\sigma_2)_u$ be local sections $(\sigma_1)_u$: $M\to P_1$, $(\sigma_2)_u$: $M\to P_2$, such that $(\sigma_1)_{u*}\overline{E}_i\in T_{p_1}P_1$ and $(\sigma_2)_{u*}\overline{E}_i\in T_{p_2}P_2$, and if we further choose the orthonormal basis at each $x\in U\subset M$ to be a coordinate basis $\overline{E}_i=\partial_i$, we then have

$$\begin{aligned}(\Omega_1)^{\alpha}{}_{ij}&=\partial_i((\sigma_1)_u^*\omega_1^{\alpha}(\partial_j))-\partial_j((\sigma_1)_u^*\omega_1^{\alpha}(\partial_i))\\&\quad+\epsilon_{\alpha\beta\gamma}((\sigma_1)_u^*\omega_1^{\beta}(\partial_i))((\sigma_1)_u^*\omega_1^{\gamma}(\partial_j))\\&=g(\partial_i W^{\alpha}{}_j-\partial_j W^{\alpha}{}_i+g\epsilon_{\alpha\beta\gamma}W^{\beta}{}_iW^{\gamma}{}_j),\end{aligned} \tag{3.36}$$

$$\begin{aligned}(\Omega_2)_{ij}&=\partial_i((\sigma_2)_u^*(-i\omega_2)(\partial_j))-\partial_j((\sigma_2)_u^*(-i\omega_2)(\partial_i))\\&=g'(\partial_iB_j-\partial_jB_i).\end{aligned} \tag{3.37}$$

Here we have used the definitions

$$gW^{\alpha}{}_j\equiv((\sigma_1)_u^*\omega_1^{\alpha})(\partial_j),\quad g'B_j\equiv((\sigma_2)_u^*(-i\omega_2))(\partial_j), \tag{3.38}$$

and g, g' denote the dimensionless coupling constants for the SU(2) and U(1) factors, respectively.

Hence

$$F^{\alpha}{}_{ij}\equiv(1/g)(\Omega_1)^{\alpha}{}_{ij}=\partial_iW^{\alpha}{}_j-\partial_jW^{\alpha}{}_i+g\epsilon_{\alpha\beta\gamma}W^{\beta}{}_iW^{\gamma}{}_j, \tag{3.39}$$

$$F_{ij}\equiv(1/g')(\Omega_2)_{ij}=\partial_iB_j-\partial_jB_i, \tag{3.40}$$

are the field tensors for the SU(2) and U(1) vector bosons, respectively.

To conclude this section we have only to properly dimension and interpret the parameters that occur in (3.32) in order to bring it into the usual form of Einstein–Cartan gravity coupled to the Yang–Mills and Higgs fields for the electroweak model.

For this purpose, assume that $\hbar=c=1$ so that the action integral (3.35) is dimensionless. This in turn implies that the Lagrangian density has to have units of $(\text{length})^{-4}$. Since all our quantities in (3.32) are so far dimensionless, we need to introduce appropriate powers of a mass scale factor τ [in units of $(\text{length})^{-1}$] into each of the terms. Thus the Riemann and gauge field tensors have to be multiplied by τ^2, while D_i and Φ_A require a factor of τ. We will use, however, the same notation for the newly dimensioned quantities as there is no risk of confusion.

Therefore, after combining terms our action becomes

$$\begin{aligned}I=\frac{1}{V_I}\int\sqrt{|g|}\Big\{&-\kappa\underline{R}+\alpha_1\underline{R}^2+\alpha_2\underline{R}_{ijkm}\underline{R}^{ijkm}+\alpha_6\underline{R}_{ij}\underline{R}^{ij}\\&-\frac{1}{4}F^{\alpha}{}_{ij}F_{\alpha}{}^{ij}-\frac{1}{4}F_{ij}F^{ij}+\frac{1}{2}(D_i\Phi_A)(D^i\Phi^A)\\&+\frac{m^2}{2}\Phi_A\Phi^A-\frac{\lambda}{4}(\Phi_A\Phi^A)^2+\frac{n}{2(n-1)}\lambda\underline{R}\Phi_A\Phi^A\\&-\kappa\Lambda\Big\}d^nx,\end{aligned} \tag{3.41}$$

where we have made the following obvious identifications in order to fix the physical parameters:

$$(\alpha_0+4\alpha_1)\tau^2=-\kappa$$

(the proportionality factor in the Einstein–Hilbert Lagrangian), (3.42)

$$2[(n-1)/n](\alpha_7\tau^2+\kappa)=m^2>0$$

(square of the mass parameter associated with the Higgs field), (3.43)

$$4[(1-n)/n^3][\alpha_1n(n-1)+2\alpha_2+\alpha_6(n-1)]=\lambda>0$$

(coupling constant of the self-interaction term of the scalar field). (3.44)

Λ = the cosmological constant. (3.45)

Also, in order to normalize the free Lagrangians of the Yang–Mills and the Higgs fields to their customary values, we have set

$$\alpha_3g^2=\alpha_4g'^2=\tfrac{1}{4}, \tag{3.46}$$

$$\alpha_5=\tfrac{1}{2}. \tag{3.47}$$

Note that (3.46) provides a relation between the parameters α_3 and α_4 and the Weinberg angle. Indeed,

$$\tan\theta_W=g'/g=\sqrt{\alpha_3/\alpha_4}, \tag{3.48}$$

so we see that the deviation of the Weinberg angle from $\pi/4$ measures the relative extent by which the SU(2) and U(1) sectors of the theory deviate from an Einstein–Cartan model in which torsion only occurs implicitly in the curvature terms. For this latter case, α_3 would equal α_4.

IV. COVARIANT DERIVATIVE OF THE HIGGS SPIN COMPLEX DOUBLET

In the preceding section we defined the operator D_i acting on the scalar fields Φ_A as their directional derivative [Eq. (3.30)]. Here we want to obtain an explicit expression for this differential operator that will allow us to relate the Lagrangian of the scalar fields, as given in (3.41), to the form in which it usually appears in the electroweak model.

In order to accomplish this, we first locally trivialize the fiber bundle π_{12}: $P_1\circ P_2\to M$ (i.e., we choose a gauge) by taking the local section $\sigma_{g_1\times g_2}(x)$, $x\in U\subset M$, defined as the set of points $(p_1,p_2)=(x,g_1\times g_2)$ with fixed g_1 and g_2.

Since $\sigma_{g_1\times g_2}(x)$ is a submanifold of $P_1\circ P_2$ which is diffeomorphic to U, the basis vectors $\mathring{\partial}_i\equiv\sigma_{g_1\times g_2*}\overline{E}_i=\sigma_{g_1\times g_2*}\partial_i$ of the tangent space of $\sigma_{g_1\times g_2}(x)$ form a closed Lie algebra. We can therefore take as a new basis in $\pi_{12}^{-1}(U)$ the local external direct sum of $\{\mathring{\partial}_i\}$ and $\{\mathring{E}_{n+A}\equiv l_A^*, A=1,\ldots,4\}$.

In terms of this basis we can write

$$\mathring{E}_i = \mathring{\partial}_i - gW^{\alpha}{}_i \mathring{E}_{n+\alpha} - g'B_i \mathring{E}_{n+4}, \tag{4.1}$$

where $W^{\alpha}{}_i$, B_i are the gauge potentials defined in (3.38) with $(\sigma_i)_u = \pi^i \circ \sigma_{g_1 \times g_2}$.

Note the (4.1), together with (2.34) and (3.38), implies

$$\begin{aligned}\mathring{\varphi}^{n+A}(\mathring{E}_i) &= \delta^{n+A}_{n+\beta}\mathring{\omega}_1^{\beta}(\mathring{\partial}_i) + \delta^{n+A}_{n+4}(-i\mathring{\omega}_2)(\mathring{\partial}_i) \\ &\quad - gW^{\beta}{}_i \mathring{\varphi}^{n+A}(\mathring{E}_{n+\beta}) - g'B_i \mathring{\varphi}^{n+A}(\mathring{E}_{n+4}) \\ &= 0,\end{aligned} \tag{4.2}$$

as required by the definition of $\mathring{\varphi}^{n+A}$. Also note that, since

$$g\mathring{E}_{n+\alpha}[W^{\beta}{}_j] = \pounds_{(\ell^*_{\alpha} \oplus 0)} \mathring{\omega}_1^{\beta}(\mathring{\partial}_j)$$

and

$$\begin{aligned}\pounds_{(l^*_{\alpha}\oplus 0)}[\mathring{\omega}_1^{\beta}(\mathring{E}_{n+\gamma})] = 0 &= (\pounds_{(l^*_{\alpha}\oplus 0)}\mathring{\omega}_1^{\beta})(\mathring{E}_{n+\gamma}) \\ &\quad + \mathring{\omega}_1^{\beta}([\mathring{E}_{n+\alpha}, \mathring{E}_{n+\gamma}]),\end{aligned}$$

it follows that

$$\mathring{E}_{n+\alpha}[W^{\beta}{}_j] = -\epsilon_{\beta\alpha\gamma} W^{\gamma}{}_j. \tag{4.3}$$

By the same line of reasoning we find [since U(1) is Abelian]

$$\mathring{E}_{n+4}[B_i] = 0. \tag{4.4}$$

It is now a simple matter to verify that (4.1) leads to the correct expression for the commutator $[\mathring{E}_i, \mathring{E}_j]$. Indeed, making use of (4.3) and (4.4), we obtain

$$[\mathring{E}_i, \mathring{E}_j] = -((\sigma_1)^*_u \Omega_1)^{\alpha}{}_{ij} \mathring{E}_{n+\alpha} - ((\sigma_2)^*_u \Omega_2)_{ij} \mathring{E}_{n+4}. \tag{4.5}$$

Thus the commutator is vertical, as required by (3.5) and (3.6). Furthermore, from (2.30) we have

$$d\mathring{\varphi}^{n+\alpha} = \pi^{1*}(\Omega_1) - \tfrac{1}{2}\epsilon_{\alpha\beta\gamma}\mathring{\varphi}^{n+\beta} \wedge \mathring{\varphi}^{n+\gamma}, \tag{4.6}$$

$$d\mathring{\varphi}^{n+4} = \pi^{2*}(-i\Omega_2), \tag{4.7}$$

and evaluating these two expressions on $(\mathring{E}_i, \mathring{E}_j)$, we get

$$-\mathring{\varphi}^{n+\alpha}([\mathring{E}_i, \mathring{E}_j]) = ((\sigma_1)^*_u \Omega_1)^{\alpha}{}_{ij}, \tag{4.8a}$$

$$-\mathring{\varphi}^{n+4}([\mathring{E}_i, \mathring{E}_j]) = ((\sigma_2)^*_u \Omega_2)_{ij}, \tag{4.8b}$$

respectively. But (4.8a) and (4.8b) are the same as what we derive from applying $\mathring{\varphi}^{n+A}$ to (4.5). Consequently, the expression (4.1) for $\mathring{E}_i$ in terms of the external direct sum basis is consistent with our previous results.

By virtue of (4.1) the directional derivatives of our scalar fields become

$$\begin{aligned}D_i\Phi_A \equiv \mathring{E}_i[\Phi_A] &= \mathring{\partial}_i\Phi_A - gW^{\alpha}{}_i \mathring{E}_{n+\alpha}[\Phi_A] \\ &\quad - g'B_i \mathring{E}_{n+4}[\Phi_A].\end{aligned} \tag{4.9}$$

Thus we now need to evaluate the quantities $\mathring{E}_{n+\alpha}[\Phi_A]$ and $E_{n+4}[\Phi_A]$. These follow directly by noting that

$$\begin{aligned}(\mathring{E}_{n+B}[\Phi_A])_{(p_1,p_2)} &= (\pounds_{\mathring{E}_{n+B}}\Phi_A)_{(p_1,p_2)} \\ &= \lim_{t\to 0}(1/t)[\Phi_A((p_1,p_2)g(t)) \\ &\quad - \Phi_A(p_1,p_2)],\end{aligned} \tag{4.10}$$

and making use of (3.34). We get

$$\mathring{E}_{n+B}[\Phi_A] = -(\rho(\mathring{l}_B))_A{}^C \Phi_C, \tag{4.11}$$

where $(\rho(\mathring{l}_B))_A{}^C$ are the matrices given in (2.49).

Substituting (4.11) into (4.9) and operating explicitly with the representation given in (2.49), we arrive at the following matrix expressions for the directional derivatives:

$$D_i\Phi = \mathring{\partial}_i\Phi - g\mathbf{W}_i\Phi + g'B_i\rho(\mathring{l}_4)\Phi, \tag{4.12}$$

where

$$\Phi = \begin{pmatrix}\Phi_1\\ \Phi_2\\ \Phi_3\\ \Phi_4\end{pmatrix}, \quad \mathbf{W}_i = \frac{1}{2}\begin{pmatrix} 0 & W_i^3 & W_i^2 & W_i^1 \\ -W_i^3 & 0 & -W_i^1 & W_i^2 \\ -W_i^2 & W_i^1 & 0 & -W_i^3 \\ -W_i^1 & -W_i^2 & W_i^3 & 0\end{pmatrix}, \tag{4.13}$$

and $\rho(\mathring{l}_4)$ is defined in (2.49d).

Since the torsion is a real tensor, the model calls naturally for the real representation of the Higgs fields that we have been using, but in order to cast the Lagrangian in the *usual* form (i.e., the way it most commonly appears in the literature of the standard model), we make the following transformation:

$$\begin{pmatrix}\phi\\ \phi^*\end{pmatrix} = U\Phi, \tag{4.14}$$

where

$$U = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & -i & 0 & 0\\ 0 & 0 & 1 & -i\\ 1 & i & 0 & 0\\ 0 & 0 & 1 & i\end{pmatrix} \tag{4.15}$$

is a unitary matrix. The quantity ϕ in (4.14) is the complex doublet scalar field of the standard electroweak model and is related to our real scalar fields by means of

$$\phi \equiv \begin{pmatrix}\varphi^+\\ \varphi^0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}\Phi_1 - i\Phi_2\\ \Phi_3 - i\Phi_4\end{pmatrix}, \tag{4.16}$$

while ϕ^* stands for its ordinary complex conjugate.

From (4.12) it is a simple matter to verify that

$$U(D_i\Phi) = UD_iU^\dagger U\Phi$$
$$= \begin{pmatrix} \mathscr{D}_i\phi \\ (\mathscr{D}_i\phi)^* \end{pmatrix}, \tag{4.17}$$

where

$$\mathscr{D}_i\phi = \mathring{\partial}_i\phi - (i/2)gW^\alpha{}_i\sigma_\alpha\phi - (i/2)g'B_i\phi \tag{4.18}$$

corresponds to the covariant derivative of the standard model, and the 2×2 matrices σ_α are the usual Pauli matrices. Using (4.14), and remembering that the components of Φ are real fields, we obtain

$$\Phi_A\Phi^A = \Phi^\dagger\Phi = (U\Phi)^\dagger(U\Phi)$$
$$= \phi^\dagger\phi + (\phi^\dagger\phi)^* = 2\phi^\dagger\phi. \tag{4.19}$$

Similarly, (4.17) gives

$$D^i\Phi^A D_i\Phi_A = (D^i\Phi)^\dagger(D_i\Phi) = 2(\mathscr{D}^i\phi)^\dagger(\mathscr{D}_i\phi). \tag{4.20}$$

Finally, substituting (4.19) and (4.20) into (3.41) yields the following form for our action integral:

$$I = \frac{1}{V_I}\int\sqrt{|g|}\Big\{ -\kappa\underline{R} + \alpha_1\underline{R}^2 + \alpha_2\underline{R}_{ijkm}\underline{R}^{ijkm} + \alpha_6\underline{R}_{ij}\underline{R}^{ij} - \frac{1}{4}F^\alpha{}_{ij}F_\alpha{}^{ij} - \frac{1}{4}F_{ij}F^{ij} + (\mathscr{D}_i\phi)^\dagger(\mathscr{D}^i\phi) + m^2\phi^\dagger\phi$$
$$- \lambda(\phi^\dagger\phi)^2 + \frac{n}{n-1}\lambda\underline{R}\phi^\dagger\phi - \kappa\Lambda\Big\}d^nx. \tag{4.21}$$

Note that in the action (4.21) the complex Higgs doublet still has four degrees of freedom, which in turn implies the existence of spurious Goldstone bosons. To eliminate these unphysical states one may still resort to a unitary gauge choice (although it is not even certain that local unitary gauges exist about every $x\in M$) such that

$$\phi = \begin{pmatrix} 0 \\ [\rho(x) + \varphi_0]/\sqrt{2} \end{pmatrix}, \tag{4.22}$$

where $\rho(x)$ denotes the remaining massive Higgs boson and φ_0 is the vacuum value of the scalar field.

An interesting point to note in (4.21) is the appearance of an extra curvature dependent "mass" term

$$[4n/(n-1)]\lambda\underline{R}\phi^\dagger\phi,$$

with its coefficient determined by the dimension of the base manifold M. This term will play an important role in the compactification analysis to be implemented in a forthcoming paper.

V. CONCLUSIONS

We have developed a formalism based on fiber bundle structures, which makes possible a geometric unification of the Yang–Mills and Higgs field sector of the standard electroweak model with gravitation. The theory requires a non-Levi-Civita connection on the bundle of frames and the ensuing torsion on the frame acts as a source for the scalar field Lagrangian, including the symmetry breaking potential.

In order to give torsion a dynamical character, the theory has to include terms quadratic in the fiber-bundle curvature. Quadratic Lagrangians in the curvature are however of interest, both because they appear naturally in the low energy limit of superstring theory, and also because through compactification of the extra dimensions of the base manifold the solutions of the modified field equations suggest a possible means of predicting values for the coupling constants of the theory.

The mathematical structures are necessarily more complicated than those used in the literature. First, because of the need to include spliced bundles in order to accommodate the direct product of two groups; and second, because the characteristic group of the spliced bundle is not semisimple, which requires in turn a careful choice of the transformation properties for the different components of the torsion, as opposed to a mere action of the adjoint representation of the group (as done in previous works). These new requirements on torsion seem to be essential for more realistic models such as the one considered here, as well as others which would include the SU(3) color gauge fields.

As pointed out in the Introduction, our theory does not yet encompass the fermionic fields needed to obtain the fermion and Yukawa Lagrangians which would complete the description of the electroweak interactions coupled to gravitation. One could, of course, resort to the phenomenological approach found in some of the literature[9] on modern Kaluza–Klein theories, where fermions are included by means of an additional Lagrangian term of the generic form

$$(\det e_i{}^A)\bar{\psi}e^i{}_A\Gamma^A D_i\psi, \tag{5.1}$$

where $e_i{}^A$ is a vielbein, $i = 1,\ldots,n$ are general coordinate in-

dices of the base manifold, Γ^A are the generators of the Clifford algebra relative to the standard inner product in $\mathfrak{R}^n$ with signature $(+,-,\ldots,-)$, and $D_i\psi$ are spinor connections.

Note that by allowing the spinor connections in (5.1) to contain torsion again, a Yukawa-type scalar–spinor interaction may be obtained without having to insert it in an *ad hoc* manner. This approach for introducing fermions is, however, rather unsatisfactory from a unification goal point of view, first, because the fermionic terms in the Lagrangian do not derive from a "pure" Einstein–Hilbert action principle, and second, because in addition to having to put in the terms of the form (5.1) by construction, the assignments of the left and right-handed fermions to multiplets of SU(2) in the present state of the electroweak model must rely heavily on experimental data. Furthermore, the standard model would have to be extended in order to determine the values of the Yukawa coupling constants. Attempts to resolve these drawbacks have led to a variety of alternative theories of supergravity, including a combination of these with Kaluza–Klein theories, as well as to the ongoing massive effort in superstring theory.

One should not rule out the possibility of an altogether different conception on the structure of the space-time manifold in order to achieve a theory of grand unification, such as the one implied in the twistor program. Work along this line of research is presently being pursued by our group, which might lead to an adequate incorporation of fermions within the framework of our theory, consisting essentially on the use of supertwistors for the frame bundle of the base manifold in the construction of fiber bundle spaces.

Regardless of such aspects of a more speculative nature, the results presented here suggest that torsion, in addition to its already acquired importance in supergravity theories, may also play a determinant role as a geometric source of the Higgs fields required for the symmetry breaking process in gauge theories, independently of which theory will ultimately prove to be the right one.

[1] Y. M. Cho, J. Math. Phys. **16**, 2029 (1975).
[2] M. O. Katanayev and I. V. Volovich, Phys. Lett. B **156**, 327 (1985).
[3] M. Rosenbaum and M. Ryan, Phys. Rev. D **37**, 2920 (1988).
[4] E. Cremmer and J. Scherk, Nucl. Phys. B **108**, 409 (1976).
[5] L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields* (Pergamon, New York, 1983), 4th ed.
[6] D. Bleecker, *Gauge Theory and Variational Principles* (Addison–Wesley, Reading, MA, 1981).
[7] J. A. Schouten, *Ricci Calculus* (Springer, Berlin, 1954).
[8] See papers by K. S. Viswanathan and by M. J. Duff in *An Introduction to Kaluza–Klein Theories*, edited by H. C. Lee (World Scientific, Singapore, 1984).
[9] C. P. Luher and M. Rosenbaum, J. Math. Phys. **15**, 1120 (1974).

von Neumann lattices and magnetic orbitals in finite phase space

J. Zak
Department of Physics, Technion—Israel Institute of Technology, Haifa 32000, Israel

(Received 29 November 1988; accepted for publication 22 March 1989)

It is proven that in the finite phase space of dimensionality M, the number of independent von Neumann states is $M - r$, where r is the number of distinct zeros of the kq function. When applied to magnetic orbitals, this leads to a linear dependence between them for each zero of the corresponding kq function. Different interesting identities follow from the linear dependence of von Neumann states.

I. INTRODUCTION

A general theory of finite phase space quantum mechanics was developed by Schwinger.[1] It was later applied to dynamical systems,[2] to the magnetic field problem,[3] and to the Weyl–Heisenberg group.[4] Finite phase space is achieved by applying boundary conditions, both to the wave function ψ and its Fourier transform F_ψ. Thus, for one degree of freedom, x and p (the coordinate and the momentum) form the phase plane and the boundary conditions assume the form[3]

$$\psi(x + Mc) = \psi(x), \tag{1}$$

$$F_\psi(p + \hbar(2\pi/c)) = F_\psi(p), \tag{2}$$

where M is an integer and c a constant. In finite phase space both the coordinate and momentum are quantized and assume the following discrete values:

$$\begin{aligned} &x = sc, \quad s = 1,...,M, \\ &p = \hbar(2\pi/Mc)t, \quad t = 1,...,M. \end{aligned} \tag{3}$$

More precisely, x and p are no longer operators in the space where the boundary conditions (1) and (2) hold. These operators are replaced by the exponentials $\exp(ix(2\pi/Mc))$ and $\exp((i/\hbar)pc)$ whose eigenvalues are determined by x and p, respectively, in relation (3). In a finite phase plane the maximal number of independent states is M. Correspondingly, also, a von Neumann lattice[5] cannot contain more than M independent states. Given a state $|v\rangle$ in a finite phase plane, one creates a von Neumann lattice $|v_{mn}\rangle$ in complete analogy with the infinite case.[6] For this we choose a constant,

$$a = M_1 c, \quad M = M_1 M_2, \tag{4}$$

and define the shift operator,

$$D(\alpha_{mn}) = (-1)^{mn} \exp(i(2\pi/a)mx)\exp(-(i/\hbar)pna), \tag{5}$$

where $m = 1,...,M_1$ and $n = 1,...,M_2$. The reason m and n assume a finite number of values is because

$$\exp(i(2\pi/a)M_1 x) = \exp(-(i/\hbar)pM_2 a) = 1.$$

With these definitions, the von Neumann lattice $|v_{mn}\rangle$ in the finite phase plane assumes the form [$\langle x|v\rangle$ and $\langle p|v\rangle$ satisfy the conditions (1) and (2), respectively];

$$|v_{mn}\rangle = D(\alpha_{mn})|v\rangle, \tag{6}$$

where $D(\alpha_{mn})$ is defined in relation (5). There are M states in the set (6). One of the questions we shall address in this paper is how many independent states there are among the set in relation (6). It is well known that in the infinite case the von Neumann set is complete.[7–9] This is, however, not the case for finite phase space. Thus it is shown in this paper that, in general, the number of independent states in the von Neumann set (6) is smaller than M. This means that, in general, the set in (6) is incomplete (in the infinite case, it is overcomplete[7–9]). In investigating the completeness of the von Neumann set (6), we shall use the kq representation[10] in finite phase space. The reason for this is that there is a connection between the number of zeros of the kq function and the overcompleteness of von Neumann sets.[11,12] In this paper it is shown that the number of independent states in the finite phase space von Neumann set [relation (6)] equals $M - r$, where r is the number of distinct zeros of $\langle k,q|v\rangle$, the state $|v\rangle$ in the kq representation.

Another subject discussed in this paper is the connection between von Neumann lattices and electronic states in a magnetic field. This connection originates from the fact that the commuting magnetic translations can be identified with the shift operators [relation (5)] in phase space. By using this identification we show in this paper that the number of independent orbitals is, in general, smaller than the number of commuting magnetic translations.

The paper is organized in the following way. In Sec. II the finite phase space kq representation is discussed. A connection is established between $C(k,q)$ (the wave function in the kq representation) in infinite phase space and $C^{(f)}(k,q)$ in finite phase space [the superscript f will be used for denoting states in the space with the boundary conditions (1) and (2)]. In Sec. III, von Neumann lattices are derived in finite phase space and the role of the zeros of kq functions is investigated. A theorem proven about the number of independent states in a von Neumann set. In Sec. IV this theorem is applied to magnetic orbitals. Section V contains a number of conclusions.

II. THE kq REPRESENTATION IN FINITE PHASE SPACE

For constructing a kq representation we choose a constant [as in Eq. (4)] and look for eigenfunctions of the basic operators $\exp[ix(2\pi/a)]$ and $\exp[(i/\hbar)pa]$. In the x representations, these eigenfunctions are[4,10]

$$\psi_{kq}(x) = \frac{1}{\sqrt{M_2}} \sum_{s=1}^{M_2} \exp(iksa)\Delta(x - q - sa), \tag{7}$$

where $\Delta(x)$ is unity when x is a multiple of Mc and is zero

 0022-2488/89/071591-06$02.50

otherwise. Here k and q in relation (7) assume the following values:

$$k = (2\pi/Mc)g, \quad g = 1,...,M_2,$$
$$q = hc, \quad h = 1,...,M_1. \tag{8}$$

Correspondingly, the kq function, $C^{(f)}(k,q)$, is (the subscript f denotes the fact that the function is in the finite phase plane)

$$C^{(f)}(k,q) = \frac{1}{\sqrt{M_2}} \sum_{s=1}^{M_2} \exp(iksa)\psi^{(f)}(q - sa). \tag{9}$$

An advantage of working with the kq function follows from the simplicity of the action on it with the basic operators,

$$\exp(ix(2\pi/a))C^{(f)}(k,q) = \exp(iq(2\pi/a))C^{(f)}(k,q),$$
$$\exp((i/\hbar)pa)C^{(f)}(k,q) = \exp(ika)C^{(f)}(k,q). \tag{10}$$

This will be used in the next section for the von Neumann lattices.

Finite phase space wave functions $\psi^{(f)}(x)$ can be defined by starting with functions $\psi(x)$ in infinite phase space and by making them satisfy the boundary conditions (1) and (2). What we are going to show is that despite the fact that $\psi^{(f)}(x)$ and $\psi(x)$ are very different functions, their kq functions, $C^{(f)}(k,q)$ and $C(k,q)$, differ only by a constant factor. Given a function $\psi(x)$, one can symmetry adapt it to the conditions (1) and (2). This is achieved by writing the double infinite sum

$$\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \exp\left(i\frac{2\pi}{c}mx\right) \exp\left(\frac{i}{\hbar}pnMc\right)\psi(x) = c \sum_{m=-\infty}^{\infty} \delta(x - mc) \sum_{n=-\infty}^{\infty} \psi(x + nMc), \tag{11}$$

where the formula was used,[13]

$$\sum_{m=-\infty}^{\infty} \exp\left(i\frac{2\pi}{c}mx\right) = c \sum_{m=-\infty}^{\infty} \delta(x - mc). \tag{12}$$

The sum of the δ function in relation (11) quantizes the x coordinate and makes it assume the values in relation (3). One can avoid using the δ functions in relation (11) by assuming that x takes on discrete values, as in relation (3). For one degree of freedom we shall define a finite phase space function by the following formula:

$$\psi^{(f)}(x) = A \sum_{j=-\infty}^{\infty} \psi(x + jMc), \tag{13}$$

where x assumes the values in relation (3) and where A is a normalization constant. It is assumed that the infinite sum in relation (13) exists. Clearly, $\psi^{(f)}(x)$ satisfies relation (1). It is easy to check that $\psi^{(f)}(x)$ also satisfied relation (2), provided x is discrete and assumes the values in relation (3) [because then $\exp(ix(2\pi/c)) = 1$]. Alternatively, one could start with the Fourier transform $F_\psi(p)$ and define the sum

$$F_\psi^{(f)}(p) = B \sum_{j'=-\infty}^{\infty} F_\psi\left(p + j'\hbar\frac{2\pi}{c}\right), \tag{14}$$

where B is a normalization constant. By definition, $F_\psi^{(f)}(p)$ satisfies relation (2). Again, it also satisfies relation (1), provided p is discrete and assumes the values in relation (3) [because then $\exp((i/\hbar)pMc) = 1$]. As was already mentioned, functions in the finite phase space [satisfying relations (1) and (2)] depend on discrete arguments, x or p, as given by relation (3). Since F_ψ is the Fourier transform of ψ, the normalization constants in relations (13) and (14) are not independent. One can show that $B = (A/c)\sqrt{2\pi\hbar/M}$. By using formula (12), one finds

$$\psi^{(f)}(x) = \frac{A\sqrt{2\pi\hbar}}{Mc} \sum_{j=-\infty}^{\infty} \exp\left(i\frac{2\pi}{Mc}jx\right) F_\psi\left(\hbar\frac{2\pi}{Ms}j\right), \tag{15}$$

$$F_\psi^{(f)}(p) = \frac{A}{\sqrt{M}} \sum_{j=-\infty}^{\infty} \exp\left(-\frac{i}{\hbar}pjc\right)\psi(jc). \tag{16}$$

Finally, $F_\psi^{(f)}(p)$ and $\psi^{(f)}(x)$ are Fourier transforms of one another (as it should be). Thus

$$\psi^{(f)}(x) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} \exp\left(i\frac{2\pi}{Mc}jx\right) F_\psi^{(f)}\left(\hbar\frac{2\pi}{Mc}j\right). \tag{17}$$

From the above formulas it is obvious that $\psi(x)$ and $\psi^{(f)}(x)$ are completely different functions [the same is true for $F_\psi(p)$ and $F_\psi^{(f)}(p)$]. Thus, for the ground state of a harmonic oscillator,

$$\psi_0(x) = (1/\pi\lambda^2)^{1/4}\exp(-x^2/2\lambda^2), \quad \lambda^2 = \hbar/m\omega, \tag{18}$$

where m is the mass and ω the cyclic frequency. A simple calculation gives [by using formula (13)]

$$\psi_0^{(f)}(x) = A\frac{\sqrt{2\pi^{1/2}\lambda}}{Mc}\vartheta_3\left(\frac{\pi x}{Mc}\middle| i\frac{2\pi\lambda^2}{M^2c^2}\right), \tag{19}$$

where $\vartheta_3(z|\tau)$ is a theta function.[14] Similarly,

$$F_0^{(f)}(p) = \frac{A}{\sqrt{\pi^{1/2}\lambda M}}\vartheta_3\left(\frac{pc}{2\hbar}\middle| i\frac{c^2}{2\pi\lambda^2}\right). \tag{20}$$

This shows that if a function in finite phase space is constructed by the above formulas, then, in general, $\psi^{(f)}(x)$ and $\psi(x)$ (and the same is true for $F_\psi^{(f)}$ and F_ψ) are different functions. This is, however, not so for their kq functions. It turns out that $C^{(f)}(k,q)$ and $C(k,q)$ differ only by a constant factor. This is a direct consequence of the definition of these functions. Let us prove it. We have

$$\begin{aligned} C^{(f)}(k,q) &= \frac{1}{\sqrt{M_2}} \sum_{s=1}^{M_2} \exp(iksa)\psi^{(f)}(q - sa) \\ &= \frac{A}{\sqrt{M_2}} \sum_{s=1}^{M_2} \sum_{j=-\infty}^{\infty} \exp(iksa)\psi(q - sa - jMc) \\ &= \frac{A}{\sqrt{M_2}} \sum_{n=-\infty}^{\infty} \exp(ikna)\psi(q - na) \\ &\equiv \frac{A}{\sqrt{M_2}}\left(\frac{2\pi}{a}\right)^{1/2} C(k,q). \end{aligned} \tag{21}$$

On the left-hand side $C^{(f)}(h,q)$ is defined on discrete values only [relation (8)] and equality (21) is therefore meaningful for these discrete values only. In deriving the equality (21) we have explicitly used the discreteness of k [see relation (3)] in replacing the sums over s and j by a single one

over n. This interesting property of the kq functions [relation (21)] will be of much use in the von Neumann lattices in the next section. In particular, it follows from relation (21) that $C^{(f)}(k,q)$ and $C(k,q)$ have zeros at the same values of k and q. Thus, for the ground state of the harmonic oscillator [relation (18)], the kq function is[11]

$$C_0(k,q)=\left(\frac{a}{2\pi}\frac{1}{\lambda\sqrt{\pi}}\right)^{1/2}\exp\left(-\frac{q^2}{2\lambda^2}\right)\times\vartheta_3\left(\frac{ka}{2}-i\frac{qa}{2\lambda^2}\middle| i\frac{a^2}{2\pi\lambda^2}\right), \tag{22}$$

with a single zero at $k=\pi/a$, $q=a/2$. Also, it follows that $C_0^{(f)}(k,q)$ has a single zero at $k=\pi/a$ and $q=a/2$.

III. von NEUMANN LATTICES

We shall now investigate the problem of the number of independent states in the von Neumann set [relation (6)] for finite phase space. It is convenient to write this set in the kq representation. We have, by using relations (5) and (10),

$$\langle k,q|v_{mn}\rangle=(-1)^{mn}\exp(i(2\pi/a)qm-ikan)\langle k,q|v\rangle. \tag{23}$$

Let us assume that $\langle k_0,q_0|v\rangle=0$ (k_0,q_0 is a zero of the kq function $\langle k,q|v\rangle$). It is then easy to check that

$$\sum_{m=1}^{M_1}\sum_{n=1}^{M_2}(-1)^{mn}\exp\left(-i\frac{2\pi}{a}q_0m+ik_0an\right)\times D(\alpha_{mn})\langle k,q|v\rangle=0. \tag{24}$$

For proving this equality we rewrite it, by using Eq. (23),

$$\sum_{m=1}^{M_1}\sum_{n=1}^{M_2}\exp\left(i\frac{2\pi}{a}(q-q_0)m-ia(k-k_0)n\right)\times\langle k,q|v\rangle=0. \tag{25}$$

Equation (25) consists of two factors: the double sum and the function $\langle k,q|v\rangle$. The double sum does not vanish only when $k=k_0$ and $q=q_0$ (it is zero otherwise). However, at the point (k_0,q_0) we have $\langle k_0,q_0|v\rangle=0$. This proves relation (25) and, equally, relation (24). Because of the importance of relation (24), let us rewrite it in a more general form, without specifying the representation. We have

$$\sum_{m=1}^{M_1}\sum_{n=1}^{M_2}(-1)^{mn}\exp\left(-i\frac{2\pi}{a}q_0m+ik_0an\right)\times D(\alpha_{mn})|v\rangle=0, \tag{26}$$

for any state $|v\rangle$ whose kq function vanishes at k_0,q_0, $\langle k_0,q_0|v\rangle=0$. Since relation (26) represents a linear dependence between the states of the von Neumann set [relation (6)], we have therefore proven the following statement. For each zero $\langle k_0,q_0|v\rangle$ of the kq function there is a linear dependence [relation (26)] of the von Neumann set [relation (6)]. A consequence of this statement is that the von Neumann set (6) contains not more than $M-r$ independent states, where r is the number of distinct zeros of $\langle k,q|v\rangle$. One can also prove that the set (6) contains not less than $M-r$ independent states. For this consider the kq functions, $C(k,q)$, that vanish at all the r zeros of $\langle k,q|v\rangle$. These $C(k,q)$ form a $M-r$-dimensional space. Let us show that the von Neumann set is complete with respect to such functions $C(k,q)$. For this it is sufficient to show that from the orthogonality of $C(k,q)$ to the von Neumann set (6), it follows that $C(k,q)$ vanishes. The assumption that $C(k,q)$ is orthogonal to the von Neumann set (6) means

$$\sum_{k,q}(-1)^{mn}\exp\left(-i\frac{2\pi}{a}qm+iakn\right)\langle v|k,q\rangle C(k,q)=0. \tag{27}$$

From here it follows that

$$\langle v|k,q\rangle C(k,q)=0 \tag{28}$$

and consequently $C(k,q)=0$ at all points where $\langle v|k,q\rangle$ does not vanish. The vanishing of $C(k,q)$ proves that the von Neumann set (6) is complete with respect to functions in the $M-r$ dimensional space. From here it follows that there are at least $M-r$ independent functions in the set (6). We have therefore proven the following theorem.

Theorem: The number of independent functions in the von Neumann set (6) equals $M-r$, where M is the dimensionality of the finite phase space and r is the number of distinct zeros of $\langle k,q|v\rangle$.

Before looking at the consequences of the theorem let us first demonstrate it on the example of the ground state [relations (18) and (22)] of the harmonic oscillator. In this case $\langle k,q|v\rangle=C_0^{(f)}(k,q)$. As was mentioned above, $C_0^{(f)}(k,q)$ has a single zero at $k=\pi/a$, $q=a/2$. Bearing in mind that $a=M_1c$ and that k and q assume the values in relation (7), it is clear that M_1 and M_2 have to be even for $k=\pi/a$, $q=a/2$ to appear in relation (7). Therefore, let us assume that M_1 and M_2 are even and write relation (6) in the x representation [by using relations (5) and (19)],

$$\sum_{n=1}^{M_2}\sum_{m=1}^{M_1}(-1)^{m+n}\exp\left(i\frac{2\pi}{a}mx\right)\times\vartheta_3\left(\frac{\pi x}{Mc}-\frac{\pi na}{Mc}\middle| i\frac{2\pi\lambda^2}{M^2c^2}\right)=0. \tag{29}$$

The summation over m can be performed and leads to the result that x has to equal an odd multiple of $a/2$, $x=(a/2)(2s+1)$. For this value of x the summation on n in relation (29) gives

$$\sum_{n=1}^{M_2}(-1)^n\vartheta_3\left(\frac{\pi x}{Mc}-\frac{\pi na}{Mc}\middle| i\frac{2\pi\lambda^2}{M^2c^2}\right)=0,\qquad x=(a/2)(2s+1). \tag{30}$$

This is an interesting identity for theta functions.

Now we are going to look at the consequences of the above theorem. If $\langle k,q|v\rangle$ has no zeros, then the von Neumann set (6) is complete. Thus, in the above example (for the ground state of a harmonic oscillator), when either M_1 or M_2 are odd, $C_0^{(f)}(k,q)$ has no zero [$k=\pi/a$ and $q=a/2$ are not among the allowed values of k and q in relation (3)], the set in relation (6) with $\langle k,q|v\rangle=C_0^{(f)}(k,q)$ is complete (it contains M independent states). Let us point out that in infinite phase space every continuous $C(k,q)$ has at least one zero. As we have just seen, this is not necessarily the case in finite phase space and it might very well happen that $C^{(f)}(k,q)$ has no zeros. Then, the von Neumann set built from such a state is complete.

When $\langle k,q|v\rangle$ has no zeros, it is easy to build a biorthogonal set[6,7] to the von Neumann set. By definition, the biorthogonal set $|\tilde{v}_{mn}\rangle$ satisfies the condition

$$\langle \tilde{v}_{mn}|v_{m'n'}\rangle = \delta_{mm'}\delta_{nn'}. \tag{31}$$

In the kq representation, $\langle k,q|\tilde{v}_{mn}\rangle$ assumes a very simple form,[6]

$$\langle k,q|\tilde{v}_{mn}\rangle = \frac{(-1)^{mn}}{M}\exp\left(-iq\frac{2\pi}{a}m + ikan\right)\frac{1}{\langle v|k,q\rangle}. \tag{32}$$

From relation (23) it is obvious that this is the biorthogonal set. An arbitrary function, $C^{(f)}(k,q)$, can be expanded in the complete von Neumann set $|v_{mn}\rangle$. We have

$$C^{(f)}(k,q) = \sum_{m,n} A_{mn}|v_{mn}\rangle, \tag{33}$$

where the expansion coefficients are found according to the formula

$$A_{mn} = \sum_{k,q} C^{(f)}(k,q)\langle \tilde{v}_{mn}|k,q\rangle. \tag{34}$$

For the sake of comparison with infinite phase space it is of interest to consider the case when $C^{(f)}(k,q) = \langle k,q|v\rangle$ in relation (34). Then $A_{00} = 1$, and all other coefficients in relation (34) are zero. This result is very different from the one in the infinite phase space, where in such a case all the coefficients $A_{mn} \neq 0$.

On the other hand, when $\langle k,q|v\rangle$ has a number of zeros, there is a linear relationship between the members of the von Neumann set [relation (26)] for each zero of the function $\langle k,q|v\rangle$. In the x representation (or p representation) these relationships are not trivial at all. One of them was given in relation (30) for the ground state of a harmonic oscillator. For the first excited state of a harmonic oscillator, $C_1^{(f)}(k,q)$ has three zeros[11]: $k = q = 0$, $k = 0$, $q = a/2$, and $k = \pi/a$, $q = 0$. Correspondingly, there will be three linear relationships between the states in the von Neumann set. These are interesting identities between theta functions that are obtained as a side product from the linear dependence between the members of the von Neumann set.

IV. LOCALIZED MAGNETIC ORBITALS IN FINITE-PHASE SPACE

It is of interest to apply the theorem of Sec. III to magnetic orbitals. As is well known, there is a connection between von Neumann lattices and localized magnetic orbitals.[12,15,16] In this section we investigate how the linear relationships [relation (26)] between members of the von Neumann set apply to the problem of a Bloch electron in a magnetic field. For simplicity we consider an electron in a magnetic field $\mathbf{H}\|z$ when the motion is in the xy plane. The Hamiltonian for this problem is (we use the symmetric gauge for the vector $\mathbf{A} = \frac{1}{2}\mathbf{H}\times\mathbf{r}$)

$$H = [\mathbf{p} + (e/2c)\mathbf{H}\times\mathbf{r}]^2/2m. \tag{35}$$

In this case it is convenient to work with the canonical coordinates[11] ($\lambda_H^2 = \hbar c/eH$),

$$\begin{aligned}
&\overline{P} = (p_x - \hbar y/2\lambda_H^2), \quad \hbar\overline{Q} = (p_y + \hbar x/2\lambda_H^2)\lambda_H^2,\\
&[\overline{Q},\overline{P}] = -i\hbar, \quad \hbar Q = (p_x + \hbar y/2\lambda_H^2)\lambda_H^2,\\
&P = p_y - \hbar x/2\lambda_H^2, \quad [Q,P] = -i\hbar.
\end{aligned} \tag{36}$$

The Hamiltonian depends on $\overline{Q}$ and $\overline{P}$ only, while Q and P are constants of motion.[17] Since the latter do not appear in the Hamiltonian, we can apply to them the finite phase space boundary conditions [relations (1) and (2)]. Correspondingly, all the formulas that were developed for the xp degree of freedom will now hold for Q and P. Given a wave function $\phi(\overline{Q},Q)$, its xy transform, $\psi(x,y)$, is given by the following unitary transformation[11] [$\phi(\overline{Q},Q)$ and $\psi(x,y)$ represent the same state in two different representations]:

$$\begin{aligned}
\psi(x,y) = \frac{1}{2\pi\lambda_H^2}\int_{-\infty}^{\infty}\int d\overline{Q}\,dQ\\
\times\exp\left[-\frac{i}{2\lambda_H^2}(xy + 2\overline{Q}Q - 2xQ - 2y\overline{Q})\right]\\
\times\phi(\overline{Q},Q)
\end{aligned} \tag{37}$$

As was mentioned above, in the $\overline{Q}Q$ representation boundary conditions apply only to the QP degree of freedom. In defining a finite phase space function, $\phi^{(f)}(\overline{Q},Q)$, we shall use formula (11). Correspondingly, for $\psi^{(f)}(x,y)$ we have

$$\begin{aligned}
&\psi^{(f)}(x,y)\\
&= D\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\exp\left(i\frac{2\pi}{c}mQ\right)\\
&\quad\times\exp\left(\frac{i}{\hbar}PnMc\right)\psi(x,y),
\end{aligned} \tag{38}$$

where the operators Q, P are given in relation (36).

For being able to apply relation (26) to magnetic orbitals, let us consider a product function,

$$\phi(\overline{Q},Q) = \phi_1(\overline{Q})\phi_2(Q). \tag{39}$$

Then relation (26) will apply to $\phi_2(Q)$ (QP is the degree of freedom to which boundary conditions are applied). We shall assume that the elementary magnetic translations lead to shifts by c [see relations (1) and (2)] in both x and y directions. For this, the rationality condition on the magnetic field is assumed to be

$$Na^2 = 2\pi\lambda_H^2, \quad M = M_1M_2 = NM_1^2. \tag{40}$$

With these notations, relation (26), for the product function in relation (39), will become (we write the relation in the xy representation)

$$\begin{aligned}
&\sum_{m=1}^{M_1}\sum_{n=1}^{NM_1}\exp\left(-i\frac{2\pi}{a}q_0m + iak_0n\right)\exp\left(\frac{i}{\hbar}\pi_{cx}mNa\right)\\
&\quad\times\exp\left(-\frac{i}{\hbar}\pi_{cy}na\right)\psi^{(f)}(x,y)\\
&= \sum_{m=1}^{M_1}\sum_{n=1}^{NM_1}(-1)^{mn}\\
&\quad\times\exp\left(-i\frac{2\pi}{a}q_0m + iak_0n + i\frac{\pi}{a}ym + i\frac{\pi}{Na}xn\right)\\
&\quad\times\psi^{(f)}(x + mNa, y - na) = 0,
\end{aligned} \tag{41}$$

where (q_0,k_0) is a zero of the $\phi_2(Q)$ in the kq representation

and π_{cx}, π_{cy} ($\pi_{cx}=\hbar Q/\lambda_H^2, \pi_{cy}=P$) are the infinitesimal magnetic translations.[10,17] Relation (41) shows that the localized magnetic orbitals $\psi^{(f)}(x+mNa, y-na)$ are not independent, and that there is a linear relation between them that holds for every zero of the kq function of $\phi_2(Q)$.

As an example let us consider, for the product function in relation (39), the ground state of the Hamiltonian (35) in the form of a Dingle function.[11] This means that both $\phi_1(\bar{Q})$ and $\phi_2(Q)$ are ground states of a one-dimensional harmonic oscillator. From relations (37) and (39), it follows[11] that

$$\psi(x,y)=(1/\lambda_H\sqrt{2\pi})\exp(-(x^2+y^2)/4\lambda_H^2). \quad (42)$$

By using the definition in relation (38), a simple calculation gives

$$\psi^{(f)}(x,y)=D\exp\left(-\frac{x^2+y^2}{4\lambda_H^2}\right)\vartheta_3\left(\frac{\pi y}{2c}+i\frac{\pi x}{2c}\,\middle|\, i\frac{M}{2}\right)\times\vartheta_3\left(\frac{\pi x}{2c}-i\frac{\pi y}{2c}\,\middle|\, i\frac{M}{2}\right), \quad (43)$$

where $\vartheta_3(z|\tau)$ is the Jacoby theta function[14] and D a normalization constant. For writing formula (41) we notice that the kq function for $\phi_2(Q)$ (ground state of a harmonic oscillator) has a zero at $k_0=\pi/a$ and $q_0=a/2$. Therefore from relations (41) and (43) we have

$$\sum_{n=1}^{M_1}\sum_{n=1}^{NM_1}(-1)^{nm+m+n}\exp\left(i\frac{\pi}{a}ym+i\frac{\pi}{Na}xn\right)\times\exp\left[-\frac{(x+mNa)^2+(y-na)^2}{4\lambda_H^2}\right]\vartheta_3\left(\frac{\pi}{2c}(y-na)+\frac{i\pi}{2c}(x+mNa)\,\middle|\, i\frac{M}{2}\right)\vartheta_3\left[\frac{\pi}{2c}(x+mNa)-\frac{i\pi}{2c}(y-na)i\frac{M}{2}\right]=0. \quad (44)$$

This is an identity that holds for all values of x and y. For the particular case of $M=4$ and $M_1=M_2=2$, relation (44) becomes [$z=(\pi/2c)y+i(\pi/2c)x$]

$$\vartheta_3(z|2i)\vartheta_3(iz|2i)-\vartheta_3(z|2i)\vartheta_2(iz|2i)-\vartheta_2(z|2i)\vartheta_3(iz|2i)-\vartheta_2(z|2i)\vartheta_2(iz|2i)=0. \quad (45)$$

In this identity we have two kinds of theta functions, ϑ_3 and ϑ_2. It is easy to check its validity for different x and y values. The identity (45) [or, more generally, (44)] between theta functions does not seem to appear in textbooks.

What we have shown in this section is that magnetic orbitals induced by commuting magnetic translations are, in general, not independent. Thus relation (41) gives a linear dependence between them for the product function in relation (39). It should be pointed out that a product function in the $\bar{Q}Q$ representation will not necessarily lead to a product function in the xy representation. The result in relation (42) is a very special case and, in general, $\psi(x,y)$ is not a product function (a function of x multiplied by a function of y) when $\psi(\bar{Q}Q)$ in relation (37) is given by relation (39).

V. CONCLUSIONS

We have shown in this paper that von Neumann lattices in finite phase space have completeness properties that are very sensitive to the boundary conditions. The reason for this is that the completeness depends on the number of zeros in the kq function. As was shown in the example of the harmonic oscillator, the zero $k_0=\pi/a$, $q_0=a/2$ can be removed by an appropriate choice of boundary conditions. When a zero of the kq function is present at (k_0,q_0), this leads to an identity [relation (26)]. In the x representation such identities can be built in the following way. One starts with an arbitrary function $\psi(x)$ and, by using formula (13), one builds $\psi^{(f)}(x)$. From relation (26) it then follows that for every zero of $C^{(f)}(k,q)$ [the kq function of $\psi^{(f)}(x)$], one obtains the following identity [see the example in relation (30)]:

$$\sum_{n=1}^{M_2}\exp(ik_0an)\psi^{(f)}(x-na)=0, \,|\, x=q_0 \text{ modulo } a. \quad (46)$$

This means that the functions $\psi^{(f)}(x-na)$ for $n=1,\ldots,M_2$ are not linearly independent!

We would like to point out that in infinite phase space the relations (24)–(26) have the meaning of distributions, and in the kq representation one has[6]

$$\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\exp\left[i\frac{2\pi}{a}(q-q_0)m-ia(k-k_0)n\right]\langle k,q|v\rangle=\sum_{m=-\infty}^{\infty}\delta(q-q_0-ma)\times\sum_{n=-\infty}^{\infty}\delta\left(k-k_0-\frac{2\pi}{a}n\right)\langle k,q|v\rangle. \quad (47)$$

On the other hand, in finite phase space these relations become purely algebraic identities [relations (24)–(26) and (46)]. It is interesting to point out that for the magnetic orbitals in the $\bar{Q}Q$ representation relation (11) leads also to a distribution. However, when transforming to the xy representation, there is an integration involved [relation (37)] that cancels the distribution and one obtains relation (38). Then, by using relation (26), an algebraic identity [relation (41)] is obtained by connecting magnetic orbitals at different lattice points. This means that the magnetic orbitals in relation (41) are not linearly independent. One can show that also in the infinite phase space (without magnetic boundary conditions) there are linear relationships between magnetic orbitals. One such linear relationship is given in Ref. 16. A more general result can be proven that for each zero of the kq function of $\phi_2(Q)$ [relation (39)], there is a linear dependence between magnetic orbitals. In this aspect, magnetic orbitals are very different from localized orbitals in the absence of a magnetic field.[18]

[1] J. Schwinger, Proc. Natl. Acad. Sci. **46**, 570 (1960).
[2] J. H. Hannay and M. V. Berry, Physica D **1**, 267 (1980).
[3] J. Zak, Phys. Lett. A **116**, 195 (1986).
[4] J. Zak, Phys. Rev. B **39**, 694 (1989).
[5] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton U.P., Princeton, NJ, 1955).
[6] M. Boon and J. Zak, Phys. Rev. B **18**, 6744 (1978).
[7] A. M. Perelomov, Theor. Math. Phys. **6**, 156 (1971) [Russian: Teor. Mat. Fiz. **6**, 213 (1971)].

[8]V. Bargmann, P. Butera, L. Girandello, and J. R. Klauder, Rep. Math. Phys. **2**, 221 (1971).
[9]H. Bacry, A. Grossmann, and J. Zak, Phys. Rev. B **12**, 1118 (1975).
[10]J. Zak, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull (Academic, New York, 1972), Vol. 27.
[11]M. Boon, in *Lectures Notes in Physics*, Vol. 50 (Springer, Berlin, 1976).
[12]M. Boon, J. Zak, and I. J. Zucker, J. Math. Phys. **24**, 316 (1983).
[13]M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions* (Cambridge U. P., Cambridge, 1960).
[14]E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge U. P., Cambridge, 1950).
[15]I. Dana and J. Zak, Phys. Rev. B **28**, 811 (1983).
[16]D. J. Thouless, J. Phys. C: Solid State Phys. **17**, L325 (1984).
[17]M. H. Johnson and B. A. Lippman, Phys. Rev. **76**, 828 (1949).
[18]This subject will be discussed in a future publication.

Schrödinger representation in Euclidean quantum field theory

U. Semmler
Universität Bonn, Physikalisches Institut, Nussallee 12, D-5300 Bonn 1, West Germany

(Received 24 June 1988; accepted for publication 18 January 1989)

The Schrödinger representation of a Euclidean quantum field is constructed nonperturbatively in a new way by defining the probability amplitude $\psi[u,t]$ as a limit of a functional integral depending on two parameters. It is shown in which sense $\psi[u,t]$ solves the formal Schrödinger equation. Finally, the role of the potential is investigated.

I. INTRODUCTION

In quantum field theory three descriptions of the dynamics are known: the Heisenberg, the Schrödinger, and the interaction representation. The theory of quantum fields is usually treated in the Heisenberg picture.[1-3] The Schrödinger picture came into disrepute, because it has been considered to be nonrenormalizable (as the interaction picture is), see Ref. 3.

Let us recall two contributions to the Schrödinger representation of quantum fields: Symanzik has computed the renormalization of a special Schrödinger amplitude by using the perturbation expansion of a formal functional integral (see Ref. 4). But his renormalized field operator $\phi(x,\tau)$ diverges for τ approaching the boundary of the time interval: His renormalization procedure violates the boundary conditions imposed by the Schrödinger representation.

In constructive quantum field theory the quantum fields are represented by stochastic fields and the corresponding process is realized in $\mathcal{S}'(\mathbb{R}^d)'$ (the dual of the Schwartz space) for special interactions. The Schrödinger equation is thereby a stochastic differential equation (see Refs. 5–7).

In the first approach the functional integral is used as a formal tool to derive the perturbation expansion, in the second we have to integrate over tempered distributions.

In the present work we define the Schrödinger amplitude as a functional integral over a set of continuous functions. The following formal considerations show the underlying ideas.

In the Schrödinger picture and in the diagonal representation of the field operator ϕ one is interested in the temporal evolution of the amplitude $\langle u|\psi(t)\rangle = :\psi[u,t]$ ($|u\rangle$ is the generalized eigenstate of ϕ). The Schrödinger amplitude $\psi[u,t]$ represents the probability amplitude that the quantum field assumes at time t the classical field u.

In the following we define the field functional $\psi[u,t]$ as a functional integral depending on the initial functional $\psi[u,t_a]$. To obtain a well-defined measure we use the Euclidean quantum field theory (EQFT), i.e., we replace t by $-it$. Our starting point is the Feynman–Kac formula, the representation by a functional integral of a solution to the diffusion equation. The direct generalization of the Feynman–Kac formula to Euclidean quantum field theory is the formal object

$$\int_{C((t_a,t)\times\mathbb{R}^3)} d\varphi \exp\left(-\int_{\mathbb{R}^3} dx \int_{t_a}^{t} ds\, L(\varphi,\partial\varphi)\right),$$

L = Lagrange function.

Unfortunately a precise definition of this object is unknown, nevertheless some physicists work with such objects. We define it by a limit of a well-defined functional integral and discuss the consequences of this definition. The advantage of this approach to Euclidean quantum field theory (i.e., the direct definition of the states) is that we do not have to define the Hilbert space and the Hamilton operator. These problems can be treated after the definition of ψ.

In Sec. II a short introduction to the notation of an abstract Wiener space is given and we express the solution to the diffusion equation with a quadratic potential by a functional integral. In Sec. III we define the field functional as a limit of a functional integral. By showing that ψ solves formally the Schrödinger equation we verify that ψ is in fact the correct physical object. By doing this we obtain a correlation between the field functional and the regularization of the formal Schrödinger equation. Section IV is devoted to the computation of the kernel of the evolution operator. We will see that we cannot compute the kernel, but we shall define an equivalent object.

In Sec. V we verify by explicit calculations that the theory is well defined for the vanishing and quadratic potential. We present a proposal for general potential. The renormalization in first order and the comparison with the work of Symanzik[4] is done in Ref. 8 (see also Sec. V C).

II. DIFFUSION AND ABSTRACT WIENER SPACE

We use the well-known theory of diffusion (see Refs. 9–11) as a starting point for the treatment of the Euclidean quantum field theory. The theory of diffusion can be reformulated in terms of abstract Wiener spaces (AWS). An AWS is a triple (i, H, B). Thereby H is a real, separable Hilbert space and B a real, separable Banach space, which is the completion of H under the norm of B. Here i denotes the inclusion map of H into B. The norm of B is measurable on H. The main point in the definition of an AWS is the measurability of the B norm (see Ref. 12, Definition 4.4). An AWS defines in a canonical way a Gaussian measure p_α (for each $\alpha>0$) on the Borel σ field (generated by the open sets) of B (see Refs. 12–16).

The solution to the diffusion equation

$$\frac{\partial}{\partial t}\psi(x,t) = \left[\frac{\alpha}{2}\frac{\partial^2}{\partial x^2} - \frac{\omega^2}{2\alpha}x^2 - V(x)\right]\psi(x,t),$$
$$\alpha>0, \quad \omega>0,$$

with $\psi(x,t_a) = :\psi_A(x)$ is the functional integral

 0022-2488/89/071597-09$02.50

$$\psi(x,t) = (\cosh(\omega(t-t_a)))^{-1/2}\exp\left(-\frac{\omega^2}{2\alpha}(t-t_a)x^2\right)$$
$$\times\int_{C[t_a,t]_-} dp_\alpha^-(y)\left[\exp\left(-\frac{\omega^2}{\alpha}x\int_{t_a}^t ds\, y(s)\right)\right.$$
$$\left.\times\exp\left(-\int_{t_a}^t ds\, V(y(s)+x)\right)\psi_A(y(t)+x)\right]$$
$$=\int_{\mathbf{R}} dz\,\psi_A(z)M(x,z)\int_{C[t_a,t]_0} dp_\alpha^0(y)$$
$$\times\exp\left(-\int_{t_a}^t ds\, V(y(s)+q(s))\right), \qquad (2.1)$$

where q is the solution to $\ddot{q}-\omega^2 q=0$, $q(t_a)=z$, $q(t)=x$,

$$M(x,z) := \left(\frac{\omega}{2\pi\alpha\sinh(\omega(t-t_a))}\right)^{1/2}$$
$$\times\exp\left(-\frac{\omega}{2\alpha}(x^2+z^2)\coth(\omega(t-t_a))\right.$$
$$\left.+\frac{\omega}{\alpha}\frac{xz}{\sinh(\omega(t-t_a))}\right).$$

Equation (2.1) is proved in Ref. 8 by using the Trotter product formula (see Refs. 17 and 18). The corresponding AWS's are $(i, W^{1,2}(t_a,t)_*, C[t_a,t]_*)$. Here $W^{1,2}(t_a,t)_-$ is the closed subspace of the Sobolev space $W^{1,2}(t_a,t)$ containing functions $f(\tau)$ which are vanishing at $\tau=t$ (for $*=0$ also at $\tau=t_a$). The B norm is equal to the sup norm. The inner product of H is equal to the free part of the Lagrange function

$$\langle f,g\rangle_{W^{1,2}(t_a,t)_*} = \omega^2\int_{t_a}^t d\tau f(\tau)g(\tau)$$
$$+\int_{t_a}^t d\tau \dot{f}(\tau)\dot{g}(\tau).$$

M multiplicated with $\exp(-(\omega/2)(t-t_a))$ coincides with the Mehler kernel (see Ref. 5).

III. SCHRÖDINGER REPRESENTATION OF EUCLIDEAN QUANTUM FIELD THEORY

To construct the Schrödinger representation of EQFT we are looking for a representation by a functional integral of $\psi[u,t]$. In the following we consider scalar fields with the Euclidean action $(m>0, \alpha>0)$

$$\frac{1}{\hbar}W[\varphi] := \frac{1}{\hbar}\int_{t_a}^t dx_0\int_{\mathbf{R}^d} dx\, L(\varphi,\partial\varphi,x)$$
$$= -\frac{1}{2\alpha\hbar}\int_{\mathbf{R}^d} dx\int_{t_a}^t ds\left[m^2\varphi(x,s)^2\right.$$
$$\left.+\sum_{i=0}^d\left(\frac{\partial}{\partial x_i}\varphi(x,s)\right)^2\right]$$
$$-\frac{1}{\hbar}\int_{\mathbf{R}^d} dx\int_{t_a}^t ds\, V(\varphi(x,s)). \qquad (3.1)$$

To generalize Eq. (2.1) to EQFT, we have to perform three steps: (i) definition of a measure in $C(G)$ with $G\subset\mathbb{R}^n$ open, in particular in $C((t_a,t)\times\mathbb{R}^d)$, (ii) making an ansatz for $\psi[u,t]$, (iii) determination of the functional differential equation which ψ solves (i.e., the functional Schrödinger equation).

A. Construction of a measure in $C(\bar{G})$

In this subsection we define an AWS (i, H, B) with $B=C(\bar{G})$ where

$$C(\bar{G}) := \{f: G\to\mathbb{R} : f \text{ is bounded and uniformly continuous in } G\},$$
$$G\subset\mathbb{R}^n,\ G \text{ open},$$
$$|f|_{C(\bar{G})} := \sup_{x\in G}|f(x)|. \qquad (3.2)$$

According to the theory of diffusion where the inner product of the Sobolev space $W^{1,2}(t_a,t)_-$ is equal to the free part of the Euclidean action, we demand in EQFT that the inner product of H coincides with the free part of the Euclidean action (3.1), but $(i, W^{1,2}(G), C(\bar{G}))$ is not an AWS. This is an unpleasant fact; however, the following question arises of course: Which Sobolev space is contained in $C(\bar{G})$? The *Sobolev Imbedding theorem* gives the answer (see Ref. 19, Theorem 5.4).

Theorem 3.1: If $G\subset\mathbb{R}^n$, G is open and if G has the strong local Lipschitz property (see Ref. 19, Definition 4.5), then there exists the following continuous imbedding: $W^{r,2}(G)\to C(\bar{G})$, where $r=1$ for $n=1$, $r=2$ for $n=2, 3$ and $r=3$ for $n=4$.

We define the Sobolev space $W^{r,2}(G)$ as usual (see Ref. 19), with two little modifications, however: we are using only real-valued functions and the following inner product:

$$\langle u,v\rangle_{W^{r,2}(G)} := \sum_{0\le|\alpha|\le r} a_\alpha\int_G dx\, D^\alpha u(x)D^\alpha v(x), \qquad (3.3)$$

with

$$a_\alpha := \begin{cases} m^2>0, & \text{for } |\alpha|=0, \quad r\geqslant 1,\\ 1, & \text{for } |\alpha|=1, \quad r\geqslant 1,\\ \epsilon_1>0, & \text{for } |\alpha|=2, \quad r\geqslant 2,\\ \epsilon_2>0, & \text{for } |\alpha|=3, \quad r=3.\end{cases}$$

In the case of $r=3$ we define the quantities ϵ_i to be C^∞ functions of the parameter $\epsilon\in\mathbb{R}_+$ with $\lim_{\epsilon\to 0}\epsilon_i(\epsilon)=0$. We choose $\epsilon_1=\epsilon$ for $r=2$. The inner product of $W^{r,2}(G)$ is for $\epsilon=0$ equal to the free part of the Euclidean action.

If $\bar{G}$ is compact, we have the following theorem, proved by Dudley in Ref. 20 (see also Refs. 13 and 21).

Theorem 3.2: If $G\subset\mathbb{R}^n$, G open, $\bar{G}$ compact, and G has the strong local Lipschitz property, then $(i, W^{r,2}(G), C(\bar{G}))$ is an AWS.

To define the Schrödinger representation of EQFT we are interested in functional integrals over $C(\Omega)$ [respectively, $C(\Omega_N)$] with

$$\Omega_N := (t_a,t)\times(-N,N)^d, \quad N\in\mathbb{N}, \quad t_a<t,$$
$$\Omega := \Omega_\infty = (t_a,t)\times\mathbb{R}^d.$$

By verifying the definition we see that

Ω and Ω_N have the strong local Lipschitz property for $-\infty\leqslant t_a<t\leqslant\infty$. (3.4)

To be more precise we have to define a functional integral over $C(\Omega)_-$ and $C(\Omega)_0$ [see Eq. (2.1)]. Therefore, we have to impose boundary values for Sobolev functions. Because of Theorem 3.1 we can do this for $N \leqslant \infty$:

$$W^{r,2}(\Omega_N)_-:$$
$$= \{f \in W^{r,2}(\Omega): f(x,s) = 0 \text{ for all } (x,s) \in \Omega \setminus \overline{\Omega}_N$$
$$\text{and } f(x,t_a) = 0 \text{ for all } x \in \mathbb{R}^d\},$$
$$W^{r,2}(\Omega_N)_0 := \{f \in W^{r,2}(\Omega_N)_-: f(x,t) = 0 \text{ for all } x \in \mathbb{R}^d\}.$$

We define the corresponding $C(\Omega_N)_*$ spaces (for $N \leqslant \infty$) to be the completion of $W^{r,2}(\Omega_N)$ with respect to the norm of $C(\overline{\Omega})$. A function $f \in W^{r,2}(\Omega)$ vanishes at infinity [$f(x) \to 0$ for $|x| \to \infty$]. This can be proven by using Ref. 19, Theorem 3.18. Therefore, functions of $C(\Omega)_*$ also vanish at infinity.

Remark 3.3: Let $W^{r,2}(G) \subset C(\overline{G})$ be as in Theorem 3.1. Because of the continuity of the imbedding (i.e., $|x|_B \leqslant c|x|_H$ for all $x \in H$) we have the inclusion $C(\overline{G})' \subset W^{r,2}(G)$. The delta function $\delta_x \in C(\overline{G})'$ [defined by $\delta_x(f) = f(x)$ for all $f \in C(\overline{G})$] can be expressed therefore by a function $\delta_x(\cdot) \in H$. This function obeys

$$\langle \delta_x f \rangle_{W^{r,2}(G)} = f(x) \text{ for all } f \in W^{r,2}(G),$$

i.e., the function δ_x obeys

$$P(D)\delta_x := (m^2 - \Delta + \epsilon_1 \Delta^2 - \epsilon_2 \Delta^3)\delta_x(z) = \delta(x - z)$$

in the distributional sense. The function δ_x is therefore the Green's function of the operator $P(D)$ and gives us physically a uniquely defined regularization of the Green's function of $(-\Delta + m^2)$. Because the operator $P(D)$ is an elliptic, partial differential operator, we obtain by using Ref. 22, Corollary 4.1.2

$$\delta_x \in C^\infty(G \setminus \{x\}) \cap C(\overline{G}).$$

δ_x determines uniquely $W^{r,2}(G)$; it is a "reproducing kernel" (see Refs. 13 and 23). For the imbedding constant c it can be proved that $c = \sup_{x \in G} \delta_x(x)$. [In Ref. 8 the functions $\delta_x^* \in W^{r,2}(\Omega)_*$ are computed.]

B. Definition of the field functional: First part

To define an ansatz for the field functional $\psi[u,t]$ we generalize the first functional integral in (2.1). To avoid at this stage difficulties with the generalization of the factor $\cosh(\omega(t - t_a))$, which can be interpreted as $\det(-d^2/dt^2 + \omega^2)/\det(-d^2/dt^2)$, we generalize "obviously" $\varphi(x,t) := (\cosh(\omega(t - t_a)))^{1/2}\psi(x,t)$ by the following definition.

Definition 3.4:

$$\varphi^{\epsilon,N}[u,t]:$$
$$= \exp\left(-\frac{1}{2\alpha}(t - t_a)|u|_W^2\right)$$
$$\times \int_{C(\Omega_N)_-} dp_\alpha^-(y) \exp\left(-\frac{1}{\alpha}\langle \xi, y \rangle\right)$$
$$\times \exp\left(-\int_{t_a}^{t} ds\, V[\pi_s(y) + u]\right) \varphi_A[\pi_t(y) + u],$$

with $u \in W^{r,2}((- N,N)^d) :=$ the closed linear subspace of $W^{r,2}(\mathbb{R}^d)$ which contains the functions vanishing in $\mathbb{R}^d \setminus [-N,N]^d$. The measure p_α^- is that of the AWS $(i, W^{r,2}(\Omega_N)_-, C(\Omega_N)_-)$. [If (i, H, B) is an AWS, then $(i, H_0, \overline{H_0}^{\|_B})$ is an AWS for a closed subspace H_0 of H.]

We choose the potential V and the initial functional φ_A such that the functional integral exists. The element ξ of $W^{r,2}(\Omega_N)_-$ is defined by

$$\langle \xi, y \rangle_W := \sum_{0 < |\alpha| < r} a_\alpha \int_\Omega dx\, ds\, D^\alpha y(x,s) D^\alpha u(x),$$
$$\text{for } y \in W^{r,2}(\Omega_N)_-$$

(more about ξ in Appendix A). The function $\langle h, \cdot \rangle_H$ with $h \in H$ can be extended canonically almost everywhere on B; this is valid for an arbitrary AWS (i,H,B) (see Ref. 12). Finally π_s is defined by

$$\pi_s: C(\Omega)_- \to C(\mathbb{R}^d)_0, \quad \pi_s(y)(x) := y(x,s),$$

where $C(\mathbb{R}^d)_0$ is equal to the set of continuous functions from $\mathbb{R}^d$ into $\mathbb{R}$ which vanish at infinity.

Remark 3.5: Originally we wanted integrals over $C(\Omega)_-$ instead over $C(\Omega_N)_-$. If we use functions that are measurable with respect to the Borel σ field of $C(\Omega)_-$, we define

$$\int_{C(\Omega)_-} dp_\alpha^-(x) f(x) := \lim_{N \to \infty} \int_{C(\Omega_N)_-} dp_\alpha^-(x) f(x),$$

if the right-hand side exists. Call such functions *integrable over* $C(\Omega)_-$. We demand in addition to Definition 3.4 that the integral is integrable over $C(\Omega)_-$. Therefore, we define

$$\varphi^\epsilon[u,t] := \lim_{N \to \infty} \varphi^{\epsilon,N}[\pi_N(u), t],$$

with $u \in W^{r,2}(\mathbb{R}^d)$ and π_N the orthogonal projection onto $W^{r,2}((- N,N)^d)$. An easy example for an integrable function over $C(\Omega)_-$ is a bounded function of the form $f(y_1(x),\ldots,y_n(x))$, where the $y_i \in (C(\Omega)_-)'$ are linearly independent.

Remark 3.6: At this stage it is not clear how to define $\varphi^{\epsilon,N}$ for $u \in C((- N,N)^d)_0$ [the completion of $W^{r,2}((- N,N)^d)$ under the $|\cdot|_c$ norm]. For later use we substitute in Definition 3.4 $\chi_{W^{r,2}((- N,N)^d)}(u) \cdot |u|^2_{W^{r,2}((- N,N)^d)}$ for $|u|_W^2$ and $\chi_{W^{r,2}((- N,N)^d)}\xi(u)$ for $\xi(u)$.

The measure p_α^- does not exist for $\epsilon = 0$, but the investigation of $\lim_{\epsilon \to 0} \varphi^\epsilon$ makes sense. This limit is bounded for a bounded integrand.

C. Functional Schrödinger equation and field functional

We call here in an abuse of language the generalized diffusion equation, which the functional φ^ϵ satisfies, the functional Schrödinger equation. With the ansatz for φ^ϵ we have a solution, but not the differential equation for it. Physically we do not accept φ^ϵ to be the correct object, if it is not in some formal way the solution to the formal Schrödinger equation (see Ref. 24)

$$\frac{\partial}{\partial t}\varphi = \frac{\alpha}{2}\int_{\mathbb{R}^d} dx \frac{\delta^2 \varphi}{\delta u(x)^2} - V[u]\varphi + K\varphi \tag{3.5}$$

with a correction term K standing for the neglected factor $\det(-\Delta + m^2)/\det(-\Delta)$. Besides the difficulties caused by the derivatives at the same point, we have to determine

the norm, with respect to the derivative is defined. In an infinite-dimensional Banach space B the usual derivative defined by the norm of B has some bad properties: For example there exists no differentiable partition of unity, but an H-C^1 differentiable one (see Refs. 12 and 25). We use this kind of differential, because the functional $\varphi^{\epsilon,N}$ with the substitution of Remark (3.6) is $W^{r,2}((-N,N)^d)$ differentiable on $C((-N,N)^d)_0$. We regularize therefore

$$\int_{\mathbb{R}^d} dx \frac{\delta^2\varphi}{\delta u(x)^2}$$

by

$$\lim_{N\to\infty} \sum_{i=1}^{L} \frac{\delta^2\varphi^{\epsilon,N}}{\delta u^2}(g_i^N,g_i^N) =: \sum_{i=1}^{L} \frac{\delta^2\varphi^{\epsilon}}{\delta u^2}(g_i,g_i), \quad (3.6)$$

with

$$L\in\mathbb{N}, \quad \frac{\delta}{\delta u} := W^{r,2}((-N,N)^d) \text{ differential}, \quad g_i\in\mathscr{S}(\mathbb{R}^d),$$

$$g_i^N := \pi_N(g_i), \quad \{g_i\}_1^\infty = \text{orthonormal basis of } L^2(\mathbb{R}^d).$$

Formally we obtain of course

$$\lim_{L\to\infty} \sum_{i=1}^{L} \frac{\delta^2\varphi^{\epsilon}}{\delta u^2}(g_i,g_i) = \int_{\mathbb{R}^d} dx \frac{\delta^2\varphi^{\epsilon}}{\delta u(x)^2}.$$

Taking into account these definitions we derive the following functional differential equation for $\varphi^{\epsilon,N}$ with $u\in W^{r,2}((-N,N)^d)$.

Theorem 3.7:

$$\frac{\partial}{\partial t}\varphi^{\epsilon,N}[u,t] - \frac{\alpha}{2}\sum_{i=1}^{L}\frac{\delta^2\varphi^{\epsilon,N}}{\delta u\,\delta u}(g_i^N,g_i^N) + \varphi^{\epsilon,N}[u,t]\left\{-\frac{1}{2\alpha}|u|^2_{W^{r,2}((-N,N)^d)} - V[u] + K\right\} = I_1 + I_2 + I_3$$

with (see Appendix A)

$$K = K(L,N,m^2,\epsilon,t-t_a):$$

$$= -\frac{1}{2}\sum_{i=1}^{L}\int_{\mathbb{R}^d} dk\, Fg_i^N(k)Fg_i^N(-k)D_\tau|_{t_a}D_s|_{t_a}\Delta_\tau^-(k,s)$$

$$\overset{\epsilon=0}{=} \frac{1}{2}\sum_{i=1}^{L}\int_{\mathbb{R}^d} dk\,\left[Fg_i^N(k)Fg_i^N(-k)\sqrt{k^2+m^2}\right.$$

$$\left.\times\tanh(\sqrt{k^2+m^2}(t-t_a))\right].$$

The expressions for I_i are given in Appendix B. To prove Theorem 3.7 we use the following theorems and apply several times the formula of integration by parts (see Ref. 12). Thereby we assume such a potential V and an initial value φ_A that these operations are valid.

Theorem 3.8: Let $G = G(1) := (t_a,t_a+1)\times(-N,N)^d$, $\phi_s: \mathbb{R}\times\mathbb{R}^d\to\mathbb{R}\times\mathbb{R}^d$, $(\tau,x)\to(t_a+s(\tau-t_a),x)$, $s>0$, $G(s) := \phi_s(G)$, $H(s) := W^{r,2}(G(s))$, $B(s) := C(\overline{G(s)})$, $y\to f(y,s)$ be differentiable with respect to s for each $y\in B(s)$, $f(\cdot,s)$ be $H(s)$-C^2 differentiable and $f(y,s)$, $\partial_s f(y,s)$ and $(\delta^2/\delta y^2)f(y,s)$ be continuous bounded functions (with respect to y) for each s.

If the $L^1(B(s))$ limit of

$$\lim_{K\to\infty}\sum_{l=1}^{K}\frac{\delta^2 f}{\delta y^2}\left[\left(\frac{d}{ds}e_l,e_l\right)+\left(e_l,\frac{d}{ds}e_l\right)\right] =: \int_{G(s)} dz_1\int_{G(s)} dz_2 \frac{\delta^2 f}{\delta y(z_1)\delta y(z_2)}\frac{\partial}{\partial s}\delta^N_{z_1}(z_2)$$

exists [see Appendix (A1)], then we have

$$\frac{d}{ds}\int_{B(s)} dp_\alpha(y)f(y,s) = \int_{B(s)} dp_\alpha(y)\frac{\partial}{\partial s}f(y,s) + \frac{\alpha}{2}\int_{B(s)} dp_\alpha(y)\int_{G(s)} dz_1\int_{G(s)} dz_2 \times\frac{\delta^2 f}{\delta y(z_1)\delta y(z_2)}\frac{\partial}{\partial s}\delta^N_{z_1}(z_2).$$

Proof: $\langle\delta_x^N, f\rangle_W := f(x)$ for all $f\in H(s)$. We use a basis $\{\bar e_i(s)\}_1^\infty$ of $H(1)$ with

$$\delta_{ij} = \sum_\alpha \beta_\alpha \int_G dx\, D^\alpha \bar e_i D^\alpha \bar e_j$$

and $\beta_\alpha := a_\alpha\cdot s\cdot s^{-2\alpha_0}$, $\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_d)$. Then $\{e_i := \bar e_i\circ\phi_s^{-1}\}_1^\infty$ is an orthonormal basis of $H(s)$ and we obtain

$$\frac{d}{ds}\int_{B(s)} dp_\alpha(y)f(y,s) = \lim_{k\to\infty}(2\pi\alpha)^{-k/2}\int_{\mathbb{R}^k} dz\frac{\partial}{\partial s}f\left(\sum_1^k z_ie_i,s\right)e^{-(1/2\alpha)z^2} + \alpha\lim_{k\to\infty}(2\pi\alpha)^{-k/2}\int_{\mathbb{R}^k} dz\left[e^{-(1/2\alpha)z^2}\times\sum_{l=1}^{k}\frac{\delta^2 f}{\delta y^2}\Big|_{(\Sigma z_ie_i,s)}\left(\frac{d}{ds}e_l,e_l\right)\right].$$

The second theorem is concerned with "differentiating under the integral sign." This theorem, a generalization of a theorem in Ref. 26 (Chap. XIV, §4), completes the proof of Theorem 3.7. Let the potential V and the initial functional φ_A be, such that the functional integrals in I_i are bounded, then we have (use Appendix A)

$$\lim_{\epsilon\to 0}\lim_{L\to\infty}\lim_{N\to\infty} I_i = 0. \quad (3.7)$$

By applying these limits we obtain from Theorem 3.7 the modified (by the K term) Schrödinger equation. It is t_a dependent because of the K term. Let us therefore define (see Remark 3.5)

$$\psi^{\epsilon,L,N}[u,t]: = \exp\left(\int_{t_a}^{t} ds\, K(L,N,\epsilon,m^2,s-t_a)\right)\varphi^{\epsilon,N,L}[\pi_N(u),t], \quad (3.8)$$

with $\varphi^{\epsilon,N,L} := \varphi^{\epsilon,N} =$ Definition 3.4, where $\varphi_A[u]$ is chosen to be L dependent. The limit $N\to\infty$ exists (see remark 3.5 and Theorem 3.7). Why this L-dependent initial functional φ_A^L? For $\epsilon = 0$ and $N = \infty$ we see that $\exp(-\int_{ta}^{t} dt\, K)$ goes to zero with $L\to\infty$. The definition (3.8) is therefore meaningless unless we choose $\varphi^{\epsilon,N}$ to be L dependent for each potential V. This can be done only if the initial functional φ_A is L dependent. In fact, we shall see in Sec. V by explicit calculations that we have to define the ground state for the vanishing potential to be L dependent. The regularization of

the variational derivative by L therefore causes an L-dependent definition of $\varphi^{\epsilon,N}$. Taking this all together we have derived the *functional Schrödinger equation*

$$0=\lim_{\epsilon\to 0}\lim_{L\to\infty}\lim_{N\to\infty}\left\{\frac{\partial}{\partial t}\psi^{\epsilon,L,N}+H^{\epsilon,L,N}\psi^{\epsilon,L,N}\right\} \tag{3.9}$$

with the *Hamilton operator*

$$(H^{\epsilon,L,N}\psi)[u]:=-\frac{\alpha}{2}\sum_{i=1}^{L}\frac{\delta^2\psi}{\delta u\,\delta u}(g_i^N,g_i^N)+\left[V[u]+\frac{1}{2\alpha}|u|^2_{W^{r,2}((-N,N)^d)}\right]\psi. \tag{3.10}$$

IV. REPRESENTATION OF THE KERNEL

In Sec. III we have generalized the first functional integral of (2.1); let us now generalize the second. Formally we have to consider

$$\psi[u,t]=\langle u|e^{-(t-t_a)H}|\psi_A\rangle=\int_{C(\mathbb{R}^d)_0}dv\langle u|e^{-(t-t_a)H}|v\rangle\psi_A[v]. \tag{4.1}$$

For $t=t_a$ the quantity

$$\langle u|\exp(-(t-t_a)H)|v\rangle$$

yields a functional delta function. This is questionable, but a representation in the form of

$$\varphi^{\epsilon}[u,t]=\int_{C(\mathbb{R}^d)_0}d\hat{p}_\alpha(v)\varphi_A[v]K(u,t;v,t_a) \tag{4.2}$$

makes sense. We call such a representation the *representation of the kernel* and call K the *kernel*. If the kernel K is given, we can compute the time evolution of each initial functional by Eq. (4.2). To obtain a representation of the kernel, we need a measure on $C(\mathbb{R}^d)_0$.

A. Construction of a measure on $C(\mathbb{R}^d)_0$

We are looking for an AWS $(i,H_N,C((-N,N)^d)_0)$ with a Hilbert space H_N which is determined by $W^{r,2}(\Omega_N)_-$. A canonical map from $C(\Omega_N)_-$ into $C((-N,N)^d)_0$ is the map π_t.

Lemma 4.1:

$$\pi_t(W^{r,2}(\Omega)_-)=W^{r-1/2,2}(\mathbb{R}^d).$$

For a proof see Ref. 19 (Remark 7.50). The Sobolev space $W^{s,2}(\mathbb{R}^d)$ for $s\in\mathbb{R}$, $s>0$ is defined by (see Ref. 19)

$$W^{s,2}(\mathbb{R}^d):=\left\{u\in L^2(\mathbb{R}^d):\int_{\mathbb{R}^d}dk(1+|k|^2)^s|Fu(k)|^2<\infty\right\}.$$

Decomposing $W^{r,2}(\Omega)_-=:W^{r,2}(\Omega_0)\oplus W^{r,2}(\Omega)_0^\perp$ we get an isomorphism of $W^{r,2}(\Omega)_0^\perp\simeq W^{r-1/2,2}(\mathbb{R}^d)$ (as sets). The functions of $W^{r,2}(\Omega)_0^\perp$ are uniquely determined by their boundary values $\pi_t(u)$. We have the inclusion $W^{r-1/2,2}(\mathbb{R}^d)\subset C(\mathbb{R}^d)_0$, but $W^{r-1/2,2}((-N,N)^d)$ is not the appropriate candidate for H_N. We need a new inner product on $W^{r-1/2,2}(\mathbb{R}^d)$ such that the above isomorphism is an isometry of Hilbert spaces.

Definition 4.2:

$$H_\infty:=\left\{f\in L^2(\mathbb{R}^d):\int_{\mathbb{R}^d}dk\,|Ff(k)|^2\frac{1}{\Delta_t^-(k,t)}<\infty\right\}.$$

Lemma 4.3: H_∞ is a real separable Hilbert space with

$$\langle u,v\rangle_{H_\infty}:=\int_{\mathbb{R}^d}dk\frac{Fu(k)^*Fv(k)}{\Delta_t^-(k,t)},$$

$$H_\infty=W^{r-1/2,2}(\mathbb{R}^d)\subset C(\mathbb{R}^d)_0,$$

$$W^{r,2}(\Omega)_0^\perp\xrightarrow{\pi_t}H_\infty,\ \pi_t \text{ is an isometry.}$$

Proof: For $u,v\in W^{r,2}(\Omega)_0^\perp$ we have

$$\langle u,v\rangle_{W^{r,2}(\Omega)_-}=\int_{\mathbb{R}^d}dk\frac{Fu(k,t)^*Fv(k,t)}{\Delta_t^-(k,t)}.$$

Therefore, $W^{r-1/2,2}\subset H_\infty$. We get $\supset$ by estimating $\Delta_t^-(k,t)^{-1}$ (see Appendix A and Ref. 8).

Therefore the desired AWS is constructed.

Theorem 4.4: $(i,H_N,C((-N,N)^d)_0)$ is an AWS (with measure $\hat{p}_\alpha$) with

$$H_N:=\{f\in H_\infty: f(x)=0 \text{ for } x\in\mathbb{R}^d\setminus(-N,N)^d\}.$$

Proof: $(i,W^{r,2}(\Omega_N)_0^\perp,C^\perp)$ is an AWS with $C^\perp:=$ completion of $W^{r,2}(\Omega_N)_0^\perp$ under the norm of $C(\Omega)_-$. The map π_t extends to an isomorphism from $C^\perp$ to $C((-N,N)^d)_0$. We have

$$|u|_{C((-N,N)^d)_0}\le|\pi_t^{-1}(u)|_{C^\perp}\le C|u|_{C((-N,N)^d)_0},$$

with

$$C:=\sup_y\left\{\int_{\mathbb{R}^d}dx|[(1-2\epsilon_1\Delta+3\epsilon_2\Delta^2)\partial_s|_t-(\epsilon_1-3\epsilon_2\Delta)\partial_s^3|_t+\epsilon_2\partial_s^5|_t]\delta_y^0(x,s)|\right\}$$

[see Appendix (A4)]. This yields the theorem and we get the formula

$$\int_{C^\perp}dp_\alpha(x)f(x)=\int_{C((-N,N)^d)_0}d\hat{p}_\alpha(x)f(\pi_t^{-1}(x)). \tag{4.3}$$

The above constructed measure is the same as in the diffusion theory, there we have [see (2.1), $\delta_t^-(t):=(1/\omega)\tanh(\omega(t-t_a))$]

$$d\hat{p}_\alpha(x)=dz(2\pi\alpha\delta_t^-(t))^{-1/2}\exp[-(1/2\alpha)z^2/\delta_t^-(t)].$$

The Green's function $\hat{\delta}_y\in H_\infty$ (uniquely defined by $\langle\hat{\delta}_y,f\rangle_{H_\infty}=f(y)$ for all $f\in H_\infty$) is equal to $\hat{\delta}_y(x)=\delta_{y,t}^-(x,t)$.

B. Representation of the kernel

We derive the following theorem by using the formula (4.3), the translation theorem (see Ref. 12), Appendix A, Fubini's theorem, and the fact that $(i,W^{r,2}(\Omega_N)_0,C(\Omega_N)_0)$ is an AWS (with measure p_α^0).

Theorem 4.5:

$$\varphi^{\epsilon,N}[u,t]=\exp\left(-\frac{1}{2\alpha}|u|^2_{H_N}\right)\int_{C((-N,N)^d)_0}d\hat{p}_\alpha(v)\times\left[\exp\left(-\frac{1}{\alpha}\langle\pi_t(\xi)-u,v\rangle_{H_N}\right)\varphi_A[v]\right.$$

$$\times \int_{C(\Omega_N)_0} dp_\alpha^0(y)$$
$$\times \exp\left(-\int_{t_a}^{t} ds V[\pi_s(y+y_0)] \right) \Bigg]$$

with

$$y_0(x,s) := (\pi_t^{-1}(v))(x,s) + (\pi_t^{-1}(u))(x,t-(s-t_a)),$$
$$y_0(x,t_a) = u(x),$$
$$y_0(x,t) = v(x)$$

(more about y_0 in Appendix A).

Theorem 4.5 gives us the desired representation of the kernel, but not the kernel $\langle u|e^{-(t-t_a)H}|v\rangle$. A part of the time dependence is incorporated in the measure $\hat{p}_\alpha$, which approaches the Dirac measure for $t \to t_a$, i.e., $\lim_{t\to t_a} \int d\hat{p}_\alpha(v) f(v) = f(0)$.

C. Properties of the kernel

The kernel $K(v,t;u,t_a)$ defined by

$$K(v,t;u,t_a):$$
$$= \int_{C(\Omega)_0} dp_\alpha^0(y) \exp\left(-\int_{t_a}^{t} ds\, V[\pi_s(y+y_0)] \right) \quad (4.4)$$

has the following properties:

$$K(u,t;v,t_a) = K(v,t;u,t_a) \geqslant 0, \quad (4.5)$$

$$\lim_{t\to t_a} K(u,t;v,t_a) = 1.$$

Proof: The first part is a consequence of the equality

$$\delta_{y,s}^0(x,\tau) = \delta_{y,t-(s-t_a)}^0(x,t-(\tau-t_a)).$$

We prove here the second part only for bounded potentials. We prove in the same way as in Sec. III.

Theorem 4.6:

$$\frac{\partial}{\partial t} K - \frac{\alpha}{2} \sum_{i=1}^{L} \frac{\delta^2 K}{\delta u\, \delta u}(g_i^N, g_i^N)$$
$$+ \frac{\delta}{\delta u} K(f) + V[u]\cdot K = I_4 + I_5$$

with

$$Ff(k) := \frac{Fu(k)}{\Delta_t^-(k,t)} + Fv(k) D_s|_{t_a} D_\tau|_t \Delta_s^0(k,\tau),$$

$$\lim_{\epsilon\to 0} \lim_{L\to\infty} \lim_{N\to\infty} I_i = 0$$

(I_4, I_5 are given in Appendix B).

The kernel can formally be interpreted as

$$K(u,t;v,t_a)\text{“}= \text{”}\frac{\langle u|e^{-H(t-t_a)}|v\rangle}{\langle u|e^{-H_0(t-t_a)}|v\rangle},$$

with H = Hamilton operator of the quantum field, H_0 = Hamilton operator of the free quantum field.

V. THE ROLE OF THE POTENTIAL

A. The potential $V=0$

We expect the ground state functional for $V=0$ to be (see Ref. 1, Chap. 7e)

$\varphi_G[v]$ = normalization

$$\cdot \exp\left(-\frac{1}{2\alpha} \int_{\mathbb{R}^d} dk\, |Fv(k)|^2 \sqrt{m^2+k^2} \right). \quad (5.1)$$

This functional, however, does not exist for an arbitrary $u \in C(\mathbb{R}^d)_0$ and we cannot define it nonzero $\hat{p}_\alpha$ almost everywhere (as we will see later). Remembering (3.8) we choose for φ_ϵ an L-dependent regularization

$$\varphi_G^L[u] := \exp\left(-\frac{1}{2\alpha} \int_{\mathbb{R}^{2d}} dx\, dy\, u(x) K_L(x,y) u(y) \right)$$

with

$$\lim_{L\to\infty} \varphi_G^L[u] = \varphi_G[u] \text{ for } u \in W^{1/2,2}(\mathbb{R}^d).$$

This is solved for example by

$$K_L(x,y) := \sum_{i=1}^{L} \sum_{j=1}^{L} g_i(x)$$
$$\times \left\{ \int_{\mathbb{R}^d} dk Fg_i(k)^* \sqrt{m^2+k^2}\, Fg_j(k) \right\} g_j(y).$$

By using the destruction operator $a_L(f)$, defined by $a_L(f)(\varphi_G^L) = 0$ and the corresponding creation operator, we derive that φ_G^L solves the functional Schrödinger equation with the ground state energy

$$E_0 := \frac{1}{2} \sum_{i=1}^{L} \int_{\mathbb{R}^d} dk\, Fg_i^N(k) Fg_i^N(-k) \sqrt{m^2+k^2}.$$

Beyond that we obtain the Klein–Gordon equation for the one-particle states.

Let us now discuss the L dependence of the ground state: By calculations done with the functional integral

$$\varphi^{\epsilon,L,N} = \text{Definition 3.4 with } \varphi_A = \varphi_G^L \text{ and } V = 0$$

we derive Lemma 5.1 which is the explicit verification that φ_G^L solves the functional Schrödinger equation (see Appendix C).

Lemma 5.1:

$$0 = \lim_{\epsilon\to 0} \lim_{L\to\infty} \lim_{N\to\infty} \left\{ \exp\left(-\int_{t_a}^{t} dt\, K \right) \varphi^{\epsilon,N,L}[u,t] \right.$$
$$\left. - \exp[-(t-t_a)E_0] \varphi_G^L[u] \right\}.$$

From Lemma (5.1) we conclude

$$\varphi^{\epsilon,L,N}[u,t] \overset{L\to\infty}{\to} \exp\left(\int_{t_a}^{t} dt\, K - (t-t_a)E_0 \right) \varphi_G^L[u] \overset{L\to\infty}{\to} 0$$

by using (C2). Therefore, φ_G = (5.1) cannot be defined almost everywhere on $C(\mathbb{R}^d)_0$ except zero.

B. The potential $V=\phi^2$

Physically the quadratic potential

$$V[u] = \frac{M^2}{2\alpha} \int_{\mathbb{R}^d} dx\, u(x)^2 \quad (5.2)$$

should be equivalent to the replacement $m^2 \to m^2 + M^2$. This is valid by the functional Schrödinger equation, but on the level of functional integrals we have to modify (5.2), if we use it for an arbitrary $u \in C(\mathbb{R}^d)_0$,

$$V_L[u]:=\frac{M^2}{2\alpha}\sum_{i=1}^{L}\int_{\mathbb{R}^d}dx\,g_i(x)u(x)$$
$$\times\int_{\mathbb{R}^d}dy\,g_i(y)u(y)\overset{L\to\infty}{\to}V[u]\,. \tag{5.3}$$

Why this L dependence? There are two arguments: (a) The functional Schrödinger equation does not change. (b) With this potential we derive Lemma 5.2.

Lemma 5.2:

$$0=\lim_{\epsilon\to 0}\lim_{L\to\infty}\lim_{N\to\infty}\left\{\exp\left(-\int_{t_a}^{t}dt\,K\right)\varphi^{\epsilon,N}[u,t]_v\right.$$
$$-\exp\left(\int_{t_a}^{t}ds\,K(L,N,\epsilon,m^2+M^2,s-t_a)\right)$$
$$\left.\times\,\varphi^{\epsilon,N}[u,t]_{M^2}\right\},$$

with

$\varphi^{\epsilon,N}[u,t]_v:=\varphi^{\epsilon,N}[u,t]$ with $V=$ (5.3),

$\varphi^{\epsilon,N}[u,t]_{M^2}:=\varphi^{\epsilon,N}[u,t]$ with $V=0$

and m^2 replaced by m^2+M^2.

This can be proven by introducing an $L^2(t_a,t)$ basis and by using an assumption similar to (C1) and by using the formulas (1.421) and (1.431) of Ref. 27 (see Ref. 8).

At least for the quadratic potential we are forced by the regularization of the formal Schrödinger equation to regularize the potential by L.

C. General potential

How should we proceed for a general potential $V[u]$? If $V[u]$ does not exist for an arbitrary $u\in C(\mathbb{R}^d)_0$, we have to modify it. An obvious modification is

$$V^L[u]\overset{L\to\infty}{\to}V[u]$$

such that $V^L[u]$ exists for each $u\in C(\mathbb{R}^d)_0$, because the Schrödinger equation does not change. We can choose for example,

$$V^L[u]=\int_{\mathbb{R}^d}dx\,h_L(x)u(x)^n \tag{5.4}$$

with a function $h_L\in L^1(\mathbb{R}^d)$ and $h_L(x)\overset{L\to\infty}{\to}1$ for all x. By using this regularization the renormalization of the ϕ_4^4 theory can be computed. Thereby we have to commute the limit $\epsilon\to 0$ with the Taylor expansion of the kernel with respect to the coupling constant (in general this is wrong). Neglecting this problem, the renormalization in first order is computed in Ref. 8 and it is shown that the divergences arising in the procedure of Symanzik[4] can be removed.

Remembering the explicit calculations of the foregoing sections we can use another modification. If we replace in the Schrödinger equation and in $\varphi^{\epsilon,N}[u,t]$ the function u by $u_L:=\sum_{i=1}^{L}g_i\langle g_i,u\rangle$, we obtain for the potential V and the initial functional φ_A the modifications

$$V[\cdot]\to V\left[\sum_{i=1}^{L}g_i\langle g_i,\cdot\rangle_{L^2}\right]=:V^L[\cdot]\,,$$

$$\varphi_A\to\varphi_A\left[\sum_{i=1}^{L}g_i\langle g_i,\cdot\rangle_{L^2}\right]=:\varphi_A^L[\cdot]\,.$$

This gives a connection to lattice theories, if we introduce a $L^2(t_a,t)$ basis for a discretization of

$$\int_{t_a}^{t}ds\,V[\pi_s(y)+y_0]\,.$$

Using the theorem of dominated convergence this yields a representation of the kernel (4.4) by a limit of finite-dimensional integrals.

VI. CONCLUSION

We have constructed the physical states $\psi[u,t]$ by a limit of well-defined functional integrals

$$\psi[u,t]:=\lim_{\epsilon\to 0}\lim_{L\to\infty}\lim_{N\to\infty}\psi^{\epsilon,L,N}[u,t],\quad u\in C(\mathbb{R}^d)_0\,.$$

Thereby the two main ideas have been:

(a) Definition of a measure on $C(\overline{G})$ for an open $G\subset\mathbb{R}^n$ by using a Lagrange function with additional derivatives of order $k\geqslant 2$. For $\epsilon=0$ these additional derivatives are vanishing.

(b) Regularization of the formal Schrödinger equation by (3.6). This L-dependent regularization has influenced the definition of physical states.

The Hilbert space can be defined probably only for $\psi^{\epsilon,L,N}$ with an L-dependent measure on $C(\mathbb{R}^d)_0$.

For further investigations one has to compute the theory for a general potential by using the proposal of Sec. V C and one has to discuss the influence of the potential on the L dependence of $\psi^{\epsilon,L,N}$. The usual renormalization procedure seems to be not the appropriate solution to this problem.

Finally an interesting and important generalization of our construction is the Schrödinger representation of quantum gauge theories. The integration theory in the infinite-dimensional Banach manifold is defined by Kuo in Ref. 28.

ACKNOWLEDGMENTS

The author wishes to thank Dr. K. Meetz and Dr. H. Schenk for many helpful discussions.

APPENDIX A

(A1) Let $W^{r,2}(G)\subset C(\overline{G})$ as in Theorem 3.1 then (see Remark 3.3)

$$\delta_x(y)=\sum_{n=1}^{\infty}e_n(x)e_n(y)\quad\text{for all } x,y\in G\,,$$

with an orthonormal basis $\{e_n\}_1^\infty$ of $W^{r,2}(G)$.

Using

$$(P(k)-\partial_s D_s)\Delta_\tau^{(\tilde{0})}(k,s)=\delta(s-\tau)$$

with

$$P(k):=m^2+k^2+\epsilon_1k^4+\epsilon_2k^6$$

and

$$D_s:=(1+2\epsilon_1k^2+3\epsilon_2k^4)\partial_s$$
$$-(\epsilon_1+3\epsilon_2k^2)\partial_s^3+\epsilon_2\partial_s^5\,,$$

$$(2\pi)^{-d/2}\int_{\mathbf{R}^d} dx\, e^{-ikx}\delta^*_{y,s}(x,\tau)$$
$$=:(2\pi)^{-d/2}e^{-iky}\Delta^*_s(k,\tau)\,,$$

and symmetry relations for $\Delta^0_\tau(k,s)$, one derives (see Ref. 8).

$$\text{(A2)}\quad \frac{1}{\Delta_t^-(k,t)} = -D_s|_t D_\tau|_t \Delta^0_\tau(k,s),$$

$$\text{(A3)}\quad (-\pi_t^{-1}(u)-\tilde{\xi})(x,s)+u(x)$$
$$=\pi_t^{-1}(u)(x,t-(s-t_a))\,,$$
$$2\langle \pi_t(\xi),u\rangle_{H_\infty}+|\tilde{\xi}|^2_{W^{r,2}(\Omega)_0}=(t-t_a)|u|^2_{W^{r,2}(\mathbf{R}^d)}$$

with

$$\tilde{\xi}:=\pi_0(\xi),\quad \pi_0\colon W^{r,2}(\Omega)_-\to W^{r,2}(\Omega)_0$$

orthogonal projection.

(A4) Let $u\in W^{r,2}(\Omega)_0^\perp$, then we have

$$P(D)u=0\,,$$

with

$$u(x,t_a)=0\,,$$
$$\{(\epsilon_1-3\epsilon_2\Delta)\partial_s^2-\epsilon_2\partial_s^4\}u(x,s)|_{\substack{s=t\\ s=t_a}}=0\,,$$
$$\{\epsilon_2\partial_s^3\}u(x,s)|_{\substack{s=t\\ s=t_a}}=0\,,$$
$$u\in C^\infty(\Omega)\cap C(\bar{\Omega})\,,$$
$$Fu(k,s)=-Fu(k,t)D_\tau|_t\Delta^0_\tau(k,s)\,.$$

(A5) Let ξ be defined as in Definition 3.4, then we have

$$F\xi(k,s)=Fu(k)(1-D_\tau|_{t_a}\Delta_\tau^-(k,s))\,,$$
$$\langle \xi,\xi\rangle_{W^{r,2}(\Omega)_-}=\int_{\mathbf{R}^d} dk\,|Fu(k)|^2\{P(k)(t-t_a)$$
$$+D_\tau|_{t_a}D_s|_{t_a}\Delta_\tau^-(k,s)\}\,.$$

(A6) Let y_0 be defined as in Theorem 4.5, then we have for $u,v\in H_\infty$

$$P(D)y_0=0$$

with

$$y_0(x,t_a)=u(x),\quad y_0(x,t)=v(x),$$
$$\{(\epsilon_1-3\epsilon_2\Delta)\partial_s^2-\epsilon_2\partial_s^4\}|_{s=t}\,y_0(x,s)=0,$$
$$\epsilon_2\partial_s^3|_{\substack{s=t\\ s=t_a}}y_0(x,s)=0\,,$$
$$Fy_0(k,s)=Fu(k)D_\tau|_{t_a}\Delta^0_\tau(k,s)-Fv(k)D_\tau|_{t_a}\Delta^0_\tau(k,s).$$

APPENDIX B

$$I_1:=\varphi^{\epsilon,N}[u,t]\cdot\left\{\frac{1}{2\alpha}\frac{\partial}{\partial t}\langle\xi,\xi\rangle_W-\frac{1}{2\alpha}(t-t_a)^2\sum_{i=1}^L\langle u,g_i^N\rangle_W\langle u,g_i^N\rangle_W\right.$$
$$\left.+\frac{1}{2\alpha}(t-t_a)\sum_{i=1}^L\langle u,g_i^N\rangle_W\frac{\delta\langle\xi,\xi\rangle_W}{\delta u}(g_i^N)-\frac{1}{8\alpha}\sum_{i=1}^L\frac{\delta\langle\xi,\xi\rangle_W}{\delta u}(g_i^N)\frac{\delta\langle\xi,\xi\rangle_W}{\delta u}(g_i^N)\right\},$$

$$I_2:=\exp\left(-\frac{1}{2\alpha}(t-t_a)|u|^2_W\right)\int_{C(\Omega_N)_-}dp_\alpha^-(y)\exp\left(-\frac{1}{\alpha}\langle\xi,y\rangle\right)$$
$$\times\int_{\Omega_N}dz\,ds\frac{\delta}{\delta y(z,s)}\left[\varphi_A[\pi_t(y)+u]\exp\left(-\int_{t_a}^t ds\,V[\pi_s(y)+u]\right)\right]$$
$$\times\left\{-\frac{\partial}{\partial t}\xi(z,s)-\frac{\partial}{\partial s}\Big|_s\xi(z,s)-(t-t_a)\sum_{i=1}^L\langle u,g_i^N\rangle_W\int_{\mathbf{R}^d}dy\left(\frac{\delta\xi(z,s)}{\delta u(y)}-\delta(y-z)g_i^N(y)\right)\right.$$
$$\left.-\frac{1}{2}\sum_{i=1}^L\int_{\mathbf{R}^2}dx\,dy\left(\delta(x-z)-\frac{\delta\xi(z,s)}{\delta u(x)}\right)\frac{\delta\langle\xi,\xi\rangle}{\delta u(y)}g_i^N(x)g_i^N(y)\right\},$$

$$I_3:=\exp\left(-\frac{1}{2\alpha}(t-t_a)|u|^2_W\right)\int_{C(\Omega_N)_-}dp_\alpha^-(y)\exp\left(-\frac{1}{\alpha}\langle\xi,y\rangle\right)$$
$$\times\int_{\Omega_N}dz_1\,ds_1\int_{\Omega_N}dz_2\,ds_2\frac{\delta}{\delta y(z_1,s_1)\delta y(z_2,s_2)}\left[\varphi_A[\pi_t(y)+u]\exp\left(-\int_{t_a}^t ds\,V[\pi_s(y)+u]\right)\right]$$
$$\times\frac{\alpha}{2}\left\{\frac{\partial}{\partial t}\delta_{z_1,s_1}(z_2,s_2)+\frac{\partial}{\partial s}\Big|_{s_1}\delta_{z_1,s}(z_2,s_2)+\frac{\partial}{\partial s}\Big|_{s_2}\delta_{z_1,s_1}(z_2,s)\right.$$
$$\left.-\sum_{i=1}^L\int_{\mathbf{R}^{2d}}dx\,dy\left(\frac{\delta\xi(z_1,s_1)}{\delta u(x)}-\delta(x-z_1)\right)\left(\frac{\delta\xi(z_2,s_2)}{\delta u(y)}-\delta(y-z_2)\right)g_i^N(x)g_i^N(y)\right\},$$

$$I_4:=\int_{C(\Omega_N)_0}dp_\alpha^0(y)\int_{\Omega_N}dz\frac{\delta}{\delta y(z)}\left[\exp\left(-\int_{t_a}^t ds\,V[\pi_s(y+y_0)]\right)\right]\left\{\frac{\partial}{\partial t}y_0(z)+\frac{\delta y_0(z)}{\delta u}(f)\right\},$$

$$I_5 := \frac{\alpha}{2}\int_{C(\Omega_N)_0} dp^0_\alpha(y)\int_{\Omega_N} dz_1\int_{\Omega_N} dz_2 \frac{\delta^2}{\delta y(z_1)\delta y(z_2)}\left[\exp\left(-\int_{t_a}^{t} ds\, V[\pi_s(y+y_0)]\right)\right]$$
$$\times\left\{\frac{\partial}{\partial t}\delta^0_{z_1}(z_2) - \int_{\Omega_N} dx\int_{\Omega_N} dy \frac{\delta y_0(z_1)}{\delta u(x)}\frac{\delta y_0(z_2)}{\delta u(y)}\sum_{i=1}^{L} g_i^N(x) g_i^N(y)\right\}.$$

APPENDIX C

The main part in the proof of Lemma 5.1 is the calculation of

$$\det(Id + zP_L HP_L GP_L) \quad \text{for } 0<z<\|P_L HP_L GP_L\|^{-1}$$

by using the formula

$$\det(Id+A) = \exp\left[\mathrm{Tr}\left(\sum_{k=1}^{\infty}(-1)^{k+1}k^{-1}z^k A^k\right)\right]$$

(See Ref. 29) with a bounded self-adjoint operator $\hat{F} := H\circ G : L^2(\mathbb{R}^d)\to L^2(\mathbb{R}^d)$ of norm 1 and

$$F(\hat{F}(f))(k) = \tanh(\sqrt{k^2+m^2}(t-t_a))Ff(k) \quad \text{for } \epsilon = 0$$

and

$$P_L := \sum_{i=1}^{L} g_i \langle g_i, \cdot\rangle_{L^2(\mathbb{R}^d)}.$$

If we assume that

$$\lim_{L\to\infty}\mathrm{Tr}\left\{P_L\left(\sum_{k=1}^{\infty}\frac{1}{k}(-1)^{k+1}z^k[\hat{F}^k - (P_L HP_L GP_L)^k]\right)P_L\right\} = 0, \tag{C1}$$

then we have in the limit $L\to\infty$,

$$\det(Id + zP_L HP_L GP_L)$$
$$\simeq \exp\left(-\frac{1}{2}\sum_{i=1}^{L}\int_{\mathbb{R}^d} dk\, Fg_i(k)Fg_i(-k) \times \ln[1+z\tanh(\sqrt{k^2+m^2}(t-t_a))]\right)$$
$$\overset{z=1}{=} \exp\left(\int_{t_a}^{t} dt\, K|_{\epsilon=0} - E_0(t-t_a)\right). \tag{C2}$$

[1] S. S. Schweber, *An Introduction to Relativistic Quantum Field Theory*, 2nd ed. (Harper and Row, New York, 1962).
[2] P. Ramond, *Field Theory* (Benjamin–Cummings, Reading, MA, 1981).
[3] N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Interscience, New York, 1959).
[4] K. Symanzik, "Schrödinger representation and Casimir effect in renormalizable quantum field theory," Nucl. Phys. B **190**, 1 (1981).
[5] J. Glimm and A. Jaffe, *Quantum Physics* (Springer, New York, 1981).
[6] S. Albeverio and R. Høegh-Krohn, "Diffusion fields, quantum fields, and fields with values in Lie groups," in *Stochastic Analysis and Applications*, edited by M. A. Pinsky (Dekker, New York, 1984).
[7] S. Albeverio *et al.*, "Local relativistic invariant flows for quantum fields," Commun. Math. Phys. **90**, 329 (1983).
[8] U. Semmler, "Zur Schrödinger-Darstellung der euklidischen Quantenfeldtheorie," Ph.D. thesis, Universität Bonn, March 1987.
[9] K. D. Ellworthy, *Stochastic Differential Equations on Manifolds* (Cambridge U.P., Cambridge, 1982).
[10] B. Simon, *Functional Integration and Quantum Physics* (Academic, New York, 1979).
[11] M. Freidlin, *Functional Integration and Partial Differential Equations* (Princeton U.P., Princeton, NJ, 1985).
[12] H.-H. Kuo, "Gaussian measures in Banach spaces," Heidelberg LNM 463, 1975.
[13] P. Baxendale, "Gaussian measures on function spaces," Am. J. Math. **98**, 891 (1976).
[14] H. Sato, "Gaussian measures on a Banach space and abstract Wiener measure," Nagoya Math. J. **36**, 65 (1969).
[15] K. R. Parthasarathy, *Probability Measures on Metric Spaces* (Academic, New York, 1967).
[16] R. Ramer, "On nonlinear transformations of Gaussian measures," J. Funct. Anal. **15**, 166 (1974).
[17] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1975), Vol. 2.
[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* (Academic, New York, 1972), Vol. 1.
[19] R. A. Adams, *Sobolev Spaces* (Academic, New York, 1975).
[20] R. M. Dudley, "Sample functions of the Gaussian process," Ann. Probab. **1**, 66 (1973).
[21] P. Baxendale, "Wiener processes on manifolds of maps," Proc. R. Soc. Edinburgh Sec. A **87**, 127 (1980).
[22] L. Hörmander, *Linear Partial Differential Operators*, 3rd ed. (Springer, Berlin, 1969).
[23] K. R. Parthasarathy and K. Schmidt, "Positive definite kernels, continuous tensor products, and central limit theorems of probability theory," Berlin LNM 272, 1972.
[24] I. M. Gelfand and A. M. Yaglom, "Integration in functional spaces and its applications in quantum physics," J. Math. Phys. **1**, 48 (1960).
[25] M. A. Piech, "The exterior algebra for Wiemann manifolds, J. Funct. Anal. **28**, 279 (1978).
[26] S. Lang, *Real analysis* (Addison Wesley, Reading, MA, 1969).
[27] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980).
[28] H.-H. Kuo, "Integration theory on infinite-dimensional manifolds," Trans. Am. Math. Soc. **159**, 57 (1971).
[29] B. Simon, *Trace Ideals and Their Applications* (Cambridge U.P., Cambridge, 1979).

The abundant symmetry structure of hierarchies of nonlinear equations obtained by reciprocal links

Sandra Carillo
Dipartimento di M. M. M. per le Sc. Appl., Universita di Roma, Roma, Italy
Benno Fuchssteiner
Universität Paderborn, D-4790 Paderborn, West Germany

(Received 4 October 1988; accepted for publication 25 January 1989)

Explicit computation for a Kawamoto-type equation shows that there is a rich associated symmetry structure for four separate hierarchies of nonlinear integrodifferential equations. Contrary to the general belief that symmetry groups for nonlinear evolution equations in 1 + 1 dimensions have to be Abelian, it is shown that, in this case, the symmetry group is *non*commutative. Its semisimple part is isomorphic to the affine Lie algebra $A_1^{(1)}$ associated to sl(2,C). In two of the additional hierarchies that were found, an explicit dependence of the independent variable occurs. Surprisingly, the generic invariance for the Kawamoto-type equation obtained in Rogers and Carillo [Phys. Scr. **36**, 865 (1987)] via a reciprocal link to the Möbius invariance of the singularity equation of the Kaup–Kupershmidt (KK) equation only holds for one of the additional hierarchies of symmetry groups. Thus the generic invariance is not a universal property for the complete symmetry group of equations obtained by reciprocal links. In addition to these results, the bi-Hamiltonian formulation of the hierarchy is given. A direct Bäcklund transformation between the (KK) hierarchy and the hierarchy of singularity equation for the Caudrey–Dodd–Gibbon–Sawada–Kotera equation is exhibited: This shows that the abundant symmetry structure found for the Kawamoto equation must exist for all fifth-order equations, which are known to be completely integrable since these equations are connected either by Bäcklund transformations or reciprocal links. It is shown that similar results must hold for all hierarchies emerging out of singularity hierarchies via reciprocal links. Furthermore, general aspects of the results are discussed.

I. RESULTS FOR THE KAWAMOTO EQUATION

Based on the fundamental work on reciprocal transformations by Rogers *et al.*[1,2] we are able to exhibit many surprising properties for the symmetry group of the Kawamoto-type equation

$$\rho_t = K_0(\rho) = 10\rho^4\rho_{xx}\rho_{xxx} + 5\rho^4\rho_x\rho_{xxxx} + \rho^5\rho_{xxxxx}. \quad (1.1)$$

Equation (1.1) was related[3] by a Bäcklund transformation to the Kawamoto equation[4]

$$\rho_t = K(\rho) = \tfrac{5}{2}\rho^4\rho_{xx}\rho_{xxx} + 5\rho^4\rho_x\rho_{xxxx} + \tfrac{15}{4}\rho^3\rho_x^2\rho_{xxx} + \rho^5\rho_{xxxxx}. \quad (1.2)$$

It will be shown that Eq. (1.1) has the recursion operator

$$\Phi(\rho) = \rho^2 D\bar{J}(u)\bar{\Theta}(u)D^{-1}\rho^{-2}, \quad (1.3)$$

where

$$u = \rho\rho_{xx} - \frac{1}{2}\rho_x^2, \quad D = \frac{d}{dx} \quad (1.4)$$

and where $\bar{\Theta}$ and $\bar{J}$ are the operators

$$\bar{\Theta}(u) = \rho D\rho D\rho u_x + 3puu_x \quad (1.5)$$

and

$$\bar{J}(u) = \rho D\rho D\rho u_x + 3(\rho uu_x + \rho Du^2) + 2[\rho D\rho DuD^{-1}u\rho^{-1} + D^{-1}uDuD\rho u] + 8[u^2D^{-1}u\rho^{-1} + D^{-1}u^3\rho^{-1}], \quad (1.6)$$

respectively. These operators can be found in Ref. 3. The operator (1.3) is proved (see Sec. III) to be hereditary. Recall that being hereditary means (see Ref. 5) that with respect to the vector field Lie algebra

$$\Phi^2[A,B] + [\Phi A,\Phi B] = \Phi\{[\Phi A,B] + [A,\Phi B]\} \quad (1.7)$$

for the arbitrary vector fields A and B. The vector field Lie algebra is defined via the variational derivative in the following way:

$$[A(\rho),B(\rho)] = \frac{\partial}{\partial\epsilon}\bigg|_{\epsilon=0} \times\{B(\rho + \epsilon A(\rho)) - A(\rho + \epsilon B(\rho))\}. \quad (1.8)$$

An operator Φ that is a recursion operator for K means that

$$\Phi[K,A] = [K,\Phi A] \quad \text{for all } A. \quad (1.9)$$

In other words, Φ has to be invariant with respect to K. Being hereditary implies that the property of being a recursion operator is inherited from K to ΦK, i.e., if (1.9) holds then

$$\Phi[\Phi K,A] = [\Phi K,\Phi A] \quad \text{for all } A. \quad (1.10)$$

For application of the result (1.10) with respect to Eq. (1.1) we begin with the vector fields

$$T_0(\rho) = \rho_x, \quad (1.11)$$

$$R_0(\rho) = x\rho_x - \rho, \quad (1.12)$$

$$N_0(\rho) = \tfrac{1}{2}x^2\rho_x - x\rho. \quad (1.13)$$

It is easily verified that the fields (1.11)–(1.13) commute

 0022-2488/89/071606-08$02.50

with the $K_0(\rho)$ given by Eq. (1.1), i.e.,

$$[K_0,V]=0 \tag{1.14}$$

for V in any of these fields. Via application of the fact that Φ from Eq. (1.3) is a recursion operator for $K_0(\rho)$, we find that the sequences

$$T_{n+1}(\rho)=\Phi(\rho)T_n(\rho), \tag{1.15}$$

$$R_{n+1}(\rho)=\Phi(\rho)R_n(\rho), \tag{1.16}$$

$$N_{n+1}(\rho)=\Phi(\rho)N_n(\rho), \tag{1.17}$$

$$K_{n+1}(\rho)=\Phi(\rho)K_n(\rho) \tag{1.18}$$

commute with $K_0(\rho)$, i.e. these vector fields are infinitesimal generators of one-parameter symmetry groups for the Kawamoto-type equation (1.1).

Written explicitly, these vector fields look somewhat complicated. Thus, already,

$$\rho_t=T_1(\rho)$$

is of tenth order in ρ.

For example, since R_0 and T_0 do not commute we have found a nonlinear equation in one independent space variable having a huge *non*commutative symmetry group. Because of translational invariance, Φ is a recursion operator for T_0 as well. Therefore, elementary application of hereditariness (see Ref. 5) shows that for all n, m we have

$$[K_n,K_m]=[K_n,T_m]=[T_n,T_m]=0 \tag{1.19}$$

since K_0, T_0 commute. Hence, any member of the two hierarchies given by the K_m and the T_n defines a nonlinear equation itself having an infinite series of symmetry generators. In addition, Φ is a recursion operator for all the K_n and T_m. Since none of the base members R_0, N_0 commutes with T_0 none of the R_n, N_n can commute with any of the T_n (formal application of the hereditariness of Φ^{-1}; see Ref. 5).

Thus the symmetry group of

$$\rho_t=K_0(\rho)$$

is of a much larger size than that of

$$\rho_t=T_1(\rho).$$

By formal arguments (or by a lengthy direct computation) one can show that the Φ is also a recursion operator for the base members $R_0(\rho)$ and $N_0(\rho)$. Thus the sequences $R_n(\rho)$ and $N_n(\rho)$ also constitute commuting hierarchies. Hence Φ is a recursion operator for the vector fields defined in (1.15)–(1.18).

The Lie derivatives L_V for the tensors Φ and the vector fields K are defined in the following way (see Ref. 6):

$$L_V K=[V,K], \tag{1.20}$$

$$(L_V\Phi)K=L_V(\Phi K)-\Phi(L_V K). \tag{1.21}$$

Thus Φ is a recursion operator for K if and only if $L_K\Phi=0$. Now, using the invariance of Φ with respect to all vector fields encountered thus far we find the general formula for commutators by using the product rule for Lie derivatives. Take any two of the base members given by K_0 and (1.11)–(1.13), say V and W; then for

$$V_n=\Phi^n V, \quad W_m=\Phi^m W \tag{1.22}$$

we find by invariance of Φ that

$$[V_n,W_m]=\Phi^{n+m}[V,W]. \tag{1.23}$$

Equation (1.23) is seen from

$$\begin{aligned}[V_n,W_m]&=L_{V_n}W_m=L_{V_n}\Phi^m W=\Phi^m L_{V_n}W\\&=-\Phi^m L_W V_n=-\Phi^m\Phi^n L_W V\\&=\Phi^{m+n}[V,W].\end{aligned}$$

Hence the Lie algebra of the symmetry group of (1.1) is completely known by computing the commutators of the base members: These commutators are

$$[T_0,R_0]=-T_0, \quad [T_0,N_0]=-R_0, \tag{1.24}$$

$$[R_0,N_0]=N_0, \tag{1.25}$$

$$[K_0,T_0]=[K_0,R_0]=[K_0,N_0]=0. \tag{1.26}$$

The commutation relations (1.24)–(1.26) show that the Lie algebras spanned by $\{R_0,N_0,T_0\}$ and $\{R_0,N_0,T_0,K_0\}$ are isomorphic to sl(2 C) and gl(2,C), respectively. If we also consider application of the inverse of Φ then formula (1.23) shows that the Lie algebras generated by $\{R_n,N_n,T_n|n\in Z_0\}$ and $\{R_n,N_n,T_n,K_0|n\in Z_0\}$ are the affine algebras associated to sl(2,C) and gl(2,C), respectively: This follows because by (1.23) Φ acts as if it were a multiplication by a formal variable. Hence the Lie algebra of the symmetry group is a Kac–Moody algebra and its semisimple part is, up to isomorphism, the loop algebra $A_1^{(1)}$. (In the physics literature the affine algebras associated to Lie algebras that are not semisimple are also called Kac–Moody algebras.)

One of the remarkable properties of the Kawamoto-type equation is that it is invariant under the transformation (see Ref. 3)

$$\rho(x)\to\bar\rho(\bar x)=[(cx+d)^2/(ad-bc)]\rho(x) \tag{1.27}$$

$$x\to\bar x=(ax+b)/(cx+d). \tag{1.28}$$

Such an invariance was first observed for the Harry Dym equation in 2 + 1 dimensions.[2] Since the operator Φ is also invariant under the transformation (1.27) and (1.28) certainly all members of the hierarchy

$$\{K_n(\rho)|n\in N_0\}$$

are again invariant. It seems interesting to ask whether the additional symmetries we exhibit in this paper are also invariant under this transformation. It turns out that none of these additional symmetries is invariant under this Möbius-type transformation. Therefore, it is quite clear that these symmetries cannot be found by a reciprocal link to a Möbius invariant singularity equation. However, the transformation (1.27) and (1.28) provides the key step for the construction of the additional symmetries.

We conclude this section by mentioning some further results about the hierarchies given by the symmetry group generators of the Kawamoto-type equation.

First, the hereditary operator Φ admits a symplectic–implectic factorization (see Refs. 7 and 8)

$$\Phi(\rho)=\hat\Theta(\rho)\hat J(\rho), \tag{1.29}$$

where the operator

$$\hat\Theta(\rho)=\rho^2 D\Theta_1(u,\rho)^{-1}\rho^2 D \tag{1.30}$$

is implectic and where

$$\hat J(\rho)=D^{-1}\rho^{-2}\Theta_1(u,p)J_1(u,p)\Theta_1(u,\rho)D^{-1}\rho^{-2} \tag{1.31}$$

is symplectic. Here u is the quantity given in (1.4) and $\Theta_1(u,\rho)$ and $J_1(u,\rho)$ are

$$\Theta_1(u,\rho) = (\rho D)^3 + u\rho D + \rho D u, \tag{1.32}$$

$$\begin{aligned} J_1(u,\rho) = {} & (\rho D)^3 + 2((\rho D)^2 u(\rho D)^{-1} \\ & + (\rho D)^{-1} u(\rho D)^2) \\ & + 8(u^2(\rho D)^{-1} + (\rho D)^{-1} u^2) \\ & + 3(u\rho d + \rho D u). \end{aligned} \tag{1.33}$$

Since the product of the operators (1.32) and (1.33) is hereditary these operators are compatible (in the sense of Gelfand–Dorfmann[9]). Now $K_0(\rho)$ can be written as

$$K_0(\rho) = \hat{\Theta}(\rho) G_0(\rho), \tag{1.34}$$

with

$$G_0(\rho) = \rho^{-1} D^{-1} \rho^{-2} \Theta_1(u,\rho)\{\rho(\rho u_x)_x + \tfrac{1}{4}u^2\} \tag{1.35}$$

the gradient of the scalar quantity

$$G_0(\rho) = \nabla P_0(\rho) = \nabla \int_0^1 \left(\int_{-\infty}^{+\infty} G_0(\lambda\rho) dx \right) d\lambda. \tag{1.36}$$

Here, the gradient ∇P is defined in the usual way:

$$\langle \nabla P(\rho), F(\rho) \rangle := \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon = 0} P(\rho + \epsilon F(\rho)). \tag{1.37}$$

However, the bilinear form $\langle\ ,\ \rangle$, representing the duality between tangent and cotangent space, differs from the well-known cases insofar as there is an additional multiplication by ρ^{-1}:

$$\langle G(\rho), F(\rho) \rangle := \int_{-\infty}^{+\infty} G(\rho) F(\rho) \rho^{-1} dx; \tag{1.38}$$

the reason for this will become clear in Sec. III. From the hereditary structure and the symplectic–implectic factorization we know that any member of the hierarchy has a Hamiltonian formulation

$$\rho_t = K_n(\rho) = \hat{\Theta}(\rho) G_n(\rho), \tag{1.39}$$

where the G_n are gradients of the quantities P_n constructed analogously to (1.36). Hence all the P_n are conserved quantities for any of the flows

$$\rho_t = K_n(\rho). \tag{1.40}$$

As usual, $\hat{\Theta}(\rho)$ can be used to define suitable Poisson brackets between scalar fields:

$$\{P_1, P_2\}_\Theta = \langle \nabla P_1, \hat{\Theta}(\rho) \nabla P_2 \rangle. \tag{1.41}$$

The conserved quantities are in involution with respect to these Poisson brackets. We would like to mention that additional Poisson brackets (having the same involutory property for the G_n) can be constructed via the implectic operators

$$\hat{\Theta}_{n+1}(\rho) = \Phi(\rho) \hat{\Theta}_n(\rho), \quad \hat{\Theta}_0 = \hat{\Theta}. \tag{1.42}$$

This is a well-known consequence of the compatible symplectic–implectic factorization (see Refs. 7 and 8). In addition to the symmetries we found, there are, as usual, time-dependent symmetry group generators coming out of the scaling symmetry (the elementary master symmetry see Ref. 10):

$$M_0(\rho) = x\rho_x. \tag{1.43}$$

We have

$$L_{M_0} \Phi = 10\Phi \tag{1.44}$$

and

$$L_{M_0} K_0 = 5K_0, \tag{1.45}$$

$$L_{M_0} T_0 = T_0, \tag{1.46}$$

$$L_{M_0} R_0 = -R_0, \tag{1.47}$$

$$L_{M_0} N_0 = -2N_0. \tag{1.48}$$

From (1.44)–(1.48) we easily conclude that for any M_n defined by

$$M_{n+1} = \Phi M_n \tag{1.49}$$

we have

$$[V, [V, M_n]] = 0 \tag{1.50}$$

whenever V is any member of the symmetry group generators we have found. Thus for the time-dependent vector fields defined by

$$M_n(t) = \exp(-L_V t) M_n,\ L_V = \text{Lie derivative w.r.t. } V \tag{1.51}$$

we know that the Taylor series of the exponential function truncates after first order in t. Thus the $M_n(t)$ are linear in t and fulfill

$$\frac{\partial}{\partial t} M_n(t) = [M_n(t), V], \tag{1.52}$$

which is just the definition of a time-dependent symmetry generator for

$$\rho_t = V(\rho).$$

Whether or not there are time-dependent symmetries that are polynomials of higher order in t is not yet completely known.

Master symmetries, and the time-dependent symmetries they generate, may not be of direct physical interest; however, they are highly interesting for structural investigations of nonlinear systems for the following reasons.

(i) Often group invariant solutions corresponding to time-dependent symmetry groups are of special interest. For example, similarity solutions are of this kind.

(ii) Master symmetries and time-dependent symmetries provide simple recursion schemes for the generation of symmetry groups: This method works beautifully even in cases where the hereditary approach via local operators fails. Such is the case for the Benjamin–Ono equation,[10,11] the Kadomtsev–Petviashvili equation,[10,12] the Landau–Lifshitz equation,[13] and several quantum mechanical spin chains.[14] Even in cases where the hereditary approach is successful the master symmetries give a more efficient computation tool and in those cases where the hereditary approach fails, a direct approach to the angle variables is given by the master symmetries.[15]

(iii) Master symmetries and time-dependent symmetries are compatible with respect to changes of evolution parameters given by coordinate transformations in the space of independent variables.[16] This leads to the construction of new completely integrable systems.

(iv) Although noncommutative, the algebraic structure of the time-dependent symmmetry group generators is quite often much more transparent than that of the algebra of symmetry generators. Thus in most cases, this algebra is finitely generated. Furthermore, it can happen (see below) that for two different systems the algebras of time-dependent symmetry group generators are isomorphic, whereas those of time-independent symmetry group generators are not isomorphic.

Let us illustrate points (iv) and (iii) in the case under consideration.

In reference to point (iv), since maps of the form $\exp(-tL_V)$ are Lie algebra isomorphisms, formula (1.51) implies that for any evolution equation the algebras of master symmetries and time-dependent symmetries are isomorphic. From the considerations above we know this algebra for the Kawamoto-type equation (1.1): It is the algebra generated by recursive application of Φ to the algebra given by the linear span of

$$\{K_0, M_0, R_0, T_0, N_0\}. \tag{1.53}$$

However, the set (1.53) also gives the master symmetries for all equations of the form

$$\rho_t = V(\rho), \tag{1.54}$$

where V is a symmetry group generator of (1.1). The only difference is that the R_0, T_0, N_0 that were symmetry group generators for (1.1) now may become generators of time-dependent symmetry group generators for (1.54). This is a simple consequence of

$$[V,[V,F]] = 0 \tag{1.55}$$

if F is any of the elements of (1.53). Hence, for all equations of the form (1.54) the algebras of master symmetries are equal. Now, by application of the isomorphism (1.51) we see that for all equations (1.54) the algebras of time-dependent symmetries are isomorphic. Because of (1.44) we easily find that these algebras are isomorphic to the affine algebra associated to (1.53). Further, (1.53) is isomorphic to the algebra $\mathrm{st}(3,\mathbb{C})$, the algebra of traceless upper triangular 3×3 matrices. Hence, all the algebras under consideration of time-dependent symmetry group generators are isomorphic to the affine algebra $A(\mathrm{st}(3,\mathbb{C}))$.

In reference to point (iii), until now, in $1+1$ dimensions (variables x and t) the generators of time-dependent symmetries were of such a form that there were equal powers in x and t. This was a consequence of the Abelian structure of the symmetry group. Here we have, as a result of the non-Abelian structure, the case where there are hierarchies of symmetries explicitly depending on x and independent of t. This allows us to do the following (see Ref. 16): interchange the variables x and t; then (1.1) becomes a fifth-order equation in t:

$$\rho_x = K(\rho) = 10\rho^4\rho_{tt}\rho_{ttt} + 5\rho^4\rho_t\rho_{tttt} + \rho^5\rho_{ttttt}. \tag{1.56}$$

Equation (1.56) is formally written as an evolution equation by introducing four new components; then for this equation the original time-independent symmetry for (1.1), which depended on x, becomes a genuine master symmetry for (1.56) which generates recursively the symmetry group and which is now independent of the new independent variable x.

Now we go back to the original equation (1.1), where we can obtain the symmetry group out of the group found for (1.56). This has a most interesting aspect, namely that via this procedure, out of one *single* nontrivial x-dependent symmetry group generator one can recursively compute the whole symmetry group. Hence, all the details of the hereditary structure must be hidden in this single x-dependent symmetry generator. Another interesting aspect seems to us that there are systems in $1+1$ dimensions where master symmetries exist which do not come out of breaking the obvious symmetry of translation invariance. This is again connected to the non-Abelian structure of the symmetry groups under consideration. This non-Abelian structure in itself is interesting, especially in light of the claims in the work of Tu.[17]

Most of the results mentioned are (more or less) consequences of the hereditariness of Φ. Thus we have to prove this property. This is not so easy for the following reason: Φ is an operator with 14 terms of up to 10th order in ρ. For the hereditariness (see Ref. 14) the variational derivative of Φ in the direction of some arbitrary ΦV has to be computed and it has to be checked whether or not a certain combination (having roughly two times the number of terms as this variational derivative) equals zero. However, after carrying out all differentiations one discovers that (as a result of the product rule) the variational derivative of Φ has more than 1 000 000 terms. Thus we have to check whether or not an integrodifferential operator of this size is equal to zero. This is not an easy task, especially since some of these terms only cancel after a sophisticated application of integration by parts. Of course, this certainly cannot be done by hand—maybe computers can do it. Indeed, we[18] have developed computer programs based on MAPLE for these computations. However, without further simplification even these cannot handle this Φ, because the usual swapspace (of about 40 MB) of a sophisticated workstation is eaten up by Φ in short time. Thus we either have to develop more sophisticated computer programs or we have to look for different means to prove the hereditariness of Φ. (In the meantime Oevel[19] has restructured some of our program packages. In fact, by these new programs it can be proved directly that Φ is hereditary. However, the necessary CPU time is still enormous.) We decided to seek the latter; this will be done—among other things—in the subsequent sections.

II. THE SINGULARITY EQUATION FOR THE CAUDREY-DODD-GIBBON-SAWADA-KOTERA (CDGSK) equation

We start with the Kaup–Kupershmidt (KK) equation (see Refs. 22–24)

$$u_t = (u_{xxxx} + 10uu_{xx} + \tfrac{15}{2}u_x^2 + \tfrac{20}{3}u^3)_x. \tag{2.1}$$

The recursion operator for Eq. (2.1) is well known (see Ref. 23):

$$\Phi(u) = \Theta(u)J(u), \tag{2.2}$$

where

$$\Theta(u) = D^3 + uD + Du, \tag{2.3}$$

$$J(u) = D^3 + 2(D^2uD^{-1} + D^{-1}uD^2) + 8(u^2D^{-1} + D^{-1}u^2) + 3(uD + Du). \tag{2.4}$$

A rigorous proof that Φ is hereditary can be shown given by using computer algebra (see Ref. 18).

We introduce variable transformations. Let us first recall the elementary transformation formulas.[7,25] If, implicitly,

$$B(u,s)=0 \quad \text{(Bäcklund transformation)} \tag{2.5}$$

defines, at least locally, a diffeomorphism between the u and s manifold, then

$$\Pi = -B_s^{-1}B_u \quad \text{(transformation operator)} \tag{2.6}$$

defines a Lie algebra isomorphism from the u to the s vector fields. Here B_s and B_u denote the partial variational derivatives with respect to s and u. Properties such as hereditary, implectic, and symplectic are Lie algebra properties; hence, we obtain the transformation formulas

$$\tilde{\Theta}(s)=\Pi\Theta(u)\Pi^{+} \quad \text{(implectic operators)}, \tag{2.7}$$

$$\tilde{J}(s)=(\Pi^{+})^{-1}J(u)\Pi^{-1} \quad \text{(symplectic operators)}, \tag{2.8}$$

$$\tilde{\Phi}(s)=\Pi\Phi(u)\Pi^{-1} \quad \text{(hereditary operators)}, \tag{2.9}$$

where, of course, the variable u has to be expressed in terms of s by use of (2.5). Let us consider the concrete Bäcklund transformation:

$$u=(s_x/s)_x-\tfrac{1}{2}(s_x/s)^2=-2\sqrt{s}(1/\sqrt{s})_{xx}. \tag{2.10}$$

Then one verifies that

$$B_sDs=-\Theta(u), \quad \Pi=Ds\Theta(u)^{-1}, \tag{2.11}$$

where $\Theta(u)$ is the operator (2.3). Hence, we find that the operator

$$\tilde{\Phi}(s)=\tilde{\Theta}(s)\tilde{J}(s)=\Pi\Phi(u)\Pi^{-1}=DsJ(u)\Theta(u)s^{-1}D^{-1} \tag{2.12}$$

is hereditary and that $\tilde{\Theta}(s)\tilde{J}(s)$ gives its implectic–symplectic factorization, where $\tilde{\Theta}(s)$ and $\tilde{J}(s)$ are given by (2.8) and (2.7), respectively.

The two base members[23] of the KK series

$$u_t=u_x, \tag{2.13}$$

$$u_t=\Theta(u)\{u_{xx}+4u^2\} \tag{2.14}$$

are transformed into

$$s_t=\Gamma_0(s):=\Pi u(x)=s_x, \tag{2.15}$$

$$s_t=\Gamma_1(s):=\Pi\Theta(u)\{u_{xx}+4u^2\}=(su_{xx}+4su^2)_x. \tag{2.16}$$

Using $\tilde{\Phi}(s)=\tilde{\Theta}(s)\tilde{J}(s)$ we have found that the flows

$$s_t=\Gamma_n(s), \tag{2.17}$$

with

$$\Gamma_{n+2}(s)=\tilde{\Theta}(s)\tilde{J}(s)\Gamma_n(s), \tag{2.18}$$

commute. Now, taking the explicit form of Φ we may rewrite recursion (2.17), (2.18) by introducing

$$H_n(u)=s^{-1}D^{-1}\Gamma_n(s) \tag{2.19}$$

as

$$s_t=[sH_n(u)]_x \tag{2.20}$$

and

$$H_{n+2}(u)=J(u)\Theta(u)H_n(u), \tag{2.21}$$

where u and s are related by (2.10). Equation (2.20) is, apart from a rescaling of time, the equation that describes the potentials of the singularity manifolds of the CDGSK equation,[20,21] as was found by Weiss[24] (see, also, Ref. 23). Thus the main statement of this section is that the flows of the singularity hierarchy of the CDGSK equation commute: This statement is completely independent of whether or not these equations actually do or do not describe singularities.

Of course, now the conserved quantities, the Hamiltonian formulation, and the Poisson brackets can be constructed out of the symplectic–implectic factorization in the same way as was done in Sec. I for the Kawamoto equation.

III. THE RECIPROCAL LINK

We introduce the following transformations between the s manifold and a manifold of the functions $\rho=\rho(\bar{x})$:

$$\bar{x}:=D^{-1}s(x), \tag{3.1}$$

$$\rho(\bar{x})=\rho(\bar{x}(s,x)):=s(x). \tag{3.2}$$

The map (3.1) and (3.2) acts on the product of the function and independent variable. Since the map assigns locally to each s some ρ we may treat it as a Bäcklund transformation. In particular, after some modification, we may apply formulas (2.7)–(2.9) for the transformation of the symplectic–implectic factorization. The crucial transformation operator Π, which maps small changes of s into small changes of ρ, is easily computed. We consider the perturbation

$$s_\epsilon=s+\epsilon v. \tag{3.3}$$

Hence,

$$\bar{x}_\epsilon=D^{-1}s(x)+\epsilon D^{-1}v(x), \tag{3.4}$$

$$\rho_\epsilon(\bar{x}_\epsilon)=s(x)+\epsilon v(x). \tag{3.5}$$

Differentiation of ρ_ϵ with respect to ϵ at $\epsilon=0$ yields, for fixed $\bar{x}$ (at $\epsilon=0$),

$$\left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0}\rho_\epsilon(\bar{x})=v(x)-\rho_{\bar{x}}\left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0}\bar{x}_\epsilon$$

$$=v(x)-\rho_{\bar{x}}D^{-1}v(x). \tag{3.6}$$

Hence, the transformation operator that maps the s vector fields into the ρ vector fields is

$$\Pi=I-\rho_{\bar{x}}D_x^{-1}. \tag{3.7}$$

The time evolutions for s,

$$s_t=K(s), \tag{3.8}$$

is mapped into the corresponding time evolution for ρ:

$$\rho_t=K(s)-\rho_{\bar{x}}D^{-1}K(s), \tag{3.9}$$

where then, by using (3.1) and (3.2), the s on the right-hand side has to be expressed in terms of ρ. Another way of looking at this transformation formula for equations would be to consider the infinitesimal transformation between dependent and independent variables given by

$$d\bar{x}=\rho\,dx-dt\,D^{-1}(K(s)), \quad d\bar{t}=dt. \tag{3.10}$$

Equation (3.10) is the reciprocal transformation introduced in Ref. 3, where it is shown that the singularity hierarchy (2.20) transforms into the Kawamoto-type hierarchy having (1.1) as its base member. The reason why we did not adopt this form is clear: Namely, we wanted to study the

transformation behavior of quantities that have nothing to do with time.

Combining the transformation formulas (2.10) and (2.11) with (3.1) and (3.2) shows that the crucial transformation operator $\hat{\Pi}$ going from u to ρ (via s) has the form

$$\hat{\Pi} = (I - \rho_{\bar{x}} D_x^{-1}) D_x s \Theta(u)^{-1}, \tag{3.11}$$

where $\Theta(u)$ is the differential operator (with respect to the variable x) given in (2.3). Here we adopt the notation

$$D_x = \frac{d}{dx}, \quad D_{\bar{x}} = \frac{d}{d\bar{x}}$$

in order to distinguish between the different variables. Using (3.2) and

$$D_x = \rho D_{\bar{x}}, \quad (D_x - \rho_{\bar{x}})\rho = \rho^2 D_{\bar{x}} \tag{3.12}$$

we can write

$$\hat{\Pi} = \rho^2 D_{\bar{x}} \Theta(u)^{-1}, \tag{3.13}$$

where u and $\Theta(u)$, rewritten in terms of $\bar{x}$, are given by

$$\begin{aligned} u &= \rho\rho_{\bar{x}\bar{x}} - \tfrac{1}{2}\rho_{\bar{x}}^2, \\ \Theta(u) &= (\rho D_{\bar{x}})^3 + u\rho D_{\bar{x}} + \rho D_{\bar{x}} u. \end{aligned} \tag{3.14}$$

In order to transform the operator $\Phi(u)$ [given in (2.2)] from the u manifold to the ρ manifold we have to rewrite the $J(u)$ given in (2.4) in the same way, i.e., we have to replace all $D = D_{\bar{x}}$ by (3.12). We obtain

$$\begin{aligned} J_1(u) = {} & (\rho D_{\bar{x}})^3 + 2[(\rho D_{\bar{x}})^2 u (\rho D_{\bar{x}})^{-1} \\ & + (\rho D_{\bar{x}})^{-1} u (\rho D_{\bar{x}})^2] \\ & + 8[u^2 (\rho D_{\bar{x}})^{-1} + (\rho D_{\bar{x}})^{-1} u^2] \\ & + 3(u\rho D_{\bar{x}} + \rho D_{\bar{x}} u). \end{aligned} \tag{3.15}$$

Now, the transformation of $\Phi(u)$ is easily obtained under the application of (2.9). Renaming $\bar{x}$ by x then yields the operator given in (1.3). The base members (2.14) and (2.13) transform into (1.1) and $\rho_t = 0$. Hence half of the hierarchy disappears by this transformation. Since $\Phi(\rho)$ was obtained from a hierarchy operator it must again be hereditary.

The transformation of the symplectic–implectic factorization is a bit more involved because the independent variable x occurs explicitly in the transformation. The point is that for this transformation we need the transformation formulas for the cotangent space instead of those for the tangent space, i.e., we need Π^+ instead of Π. Usually this transformation is given canonically because we tacitly represent the cotangent space by the bilinear form

$$\langle G(u), F(u) \rangle = \int_{-\infty}^{+\infty} G(u) F(u) dx. \tag{3.16}$$

The form (3.16) is preserved whenever the transformation does not depend explicitly on the independent variable. However, in the case given by (3.1) and (3.2) the density on the right-hand side of (3.16) is transformed in the following manner:

$$G(u)F(u)dx \to G(u(\rho))F(u(\rho))\rho^{-1} d\bar{x}. \tag{3.17}$$

Equation (3.17) suggests that on the ρ manifold we now represent cotangent vectors by the bilinear form induced by this new density on the right-hand side of (3.17); certainly, this is allowed since the choice of (3.16) was only suggested for convenience because we wanted the differential operator to be antisymmetric. By having the representation for the duality between the tangent and cotangent spaces now fixed, we can compute adjoints of the operators. Because of

$$(\rho D_{\bar{x}})^+ = -\rho D_{\bar{x}}, \tag{3.18}$$

we find, for the operators given in (3.14) and (3.15),

$$\Theta(u)^+ = -\Theta(u), \tag{3.19}$$

$$J(u)^+ = -J(u) \tag{3.20}$$

and $\hat{\Pi}^+$ is

$$\hat{\Pi}^+ = \Theta(u)^{-1} \rho^2 D_{\bar{x}}. \tag{3.21}$$

The quantity (3.21) is important because $(\hat{\Pi}^+)^{-1}$ transforms covectors from the u manifold to covectors of the ρ manifold. Finally, application of (2.7) and (2.8) to the symplectic–implectic factorization on the u manifold leads to (1.30) and (1.31) (after having renamed $\bar{x}$ by x). The representation of $K_0(\rho)$ given in (1.34) in terms of the covector field $G_0(\rho)$ follows, by transformation of covectors, out of the corresponding base members of the KK equation.

IV. THE ADDITIONAL SYMMETRIES

By going from the singularity hierarchy to the Kawamoto-type equation we have seen that half of the hierarchy disappears; however, where do the additional symmetries come from?

The singularity manifold is invariant under the Möbius transform, which goes into the symmetry group given by (1.27) and (1.28). We want to use this group for the creation of additional hierarchies of symmetry generators: We do this by considering it as a Bäcklund transformation on the manifold. Again, we have to compute the crucial transformation operator Π. Thus consider the operation

$$\rho(x) \to \bar{\rho}(\bar{x}) = [(cx + d)^2/(ad - bc)]\rho(x), \tag{4.1}$$

$$x \to \bar{x} = (ax + b)/(cx + d) \tag{4.2}$$

and perform the same perturbation procedure as before:

$$\rho_\epsilon = \rho + \epsilon\nu, \tag{4.3}$$

$$\bar{x}_\epsilon = \bar{x}, \tag{4.4}$$

$$\bar{\rho}_\epsilon(\bar{x}_\epsilon) = [(cx + d)/(ad - bc)]^2 \rho_\epsilon. \tag{4.5}$$

Then

$$\left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0} \bar{\rho}_\epsilon = \left(\frac{cx + d}{ad - bc}\right)^2 \nu. \tag{4.6}$$

Hence, the transformation operator Π is given by

$$\Pi = [(cx + d)/(ad - bc)]^2, \tag{4.7}$$

which means that whenever $T(\rho(x))$ is a symmetry generator,

$$\bar{T} = [(cx + d)/(ad - bc)]^2 T(\rho(x)) \tag{4.8}$$

is again a symmetry generator (now for the variables $\bar{\rho}, \bar{x}$). Of course, this transformation scheme does not give us any new information when applied to symmetry generators given by the hierarchy since these are invariant under this transformation. However, there is an additional obvious symmetry, namely, the generator of translation invariance

$$T_0(\rho) = \rho_x,$$

which is not a member of the hierarchy [since it obviously is not invariant under the Möbius-type transformation (1.27) and (1.28)]. We apply the transformation scheme to this symmetry for the cases $a = d = 0$ and $b = c = 1$ and obtain

$$\bar{x} = 1/x, \quad \bar{\rho}(\bar{x}) = x^2\rho(x). \tag{4.9}$$

Hence,

$$\overline{T}(\bar{\rho}(\bar{x})) = x^2\rho_x \tag{4.10}$$

must be a symmetry for the Kawamoto-type equation (written for the overbarred variables). Using

$$\bar{\rho}_x = \bar{\rho}_{\bar{x}}\bar{x}_x = 2x\rho + x^2\rho_x \tag{4.11}$$

we obtain

$$\overline{T}(\bar{\rho}(\bar{x})) = \bar{x}^2\bar{\rho}_{\bar{x}} - 2\bar{x}\bar{\rho}, \tag{4.12}$$

which is exactly the symmetry given in (1.13). The symmetry in (1.12) is then obtained by commuting (4.12) with the generator of translation. Hence, the four hierarchies for equation (1.1) and their properties are established. For (1.2) the relevant quantities are obtained out of the Bäcklund transformation between (1.1) and (1.2) (see Ref. 3).

The reason that in this section we only considered special cases for the parameters a, b, c, and d is that we did not want to consider vector fields where rational functions in the independent variable occur.

The transformation behavior under the Möbius-type transformation (4.1) and (4.2) of the hierarchy starting with R_0 is rather interesting. All the fields T_n, R_n, N_n, K_n are mapped under this transformation on their negatives; hence, for equations given by these fields the transformation (4.1) and (4.2) results in a reflection of time.

V. GENERAL ASPECTS

We would like to address the question of whether or not the results found for the Kawamoto equation are accidental. In fact, they are not accidental. The results hold—with slight variations—for all equations obtained by reciprocal links from singularity hierarchies.

First we observe that there must be a Bäcklund transformation between the CDGSK equation and its singularity hierarchy since the CDGSK and KK equations are linked by a direct Bäcklund transformation. Actually, this is known for the case of the CDGSK (see Ref. 3), but in fact must be the case for every hierarchy in $1 + 1$ dimensions, leading to a successful result in the Painlevé test.

To see the above, consider the Painlevé series truncated at the constant level term. Then the constant level term $\tilde{u}$ fulfills the same equation as the original quantity u. Hence, the original quantity and the constant level term are related via an auto-Bäcklund transformation (ABT) (see Ref. 24 and 26). This ABT usually is the time derivative of the spatial part of the ABT known for the system. In fact, this ABT can be computed from the Painlevé test. Now, among the equations provided by the Painlevé test there is one where it is shown that the singularity function ϕ can be considered as a symmetry generator for the time evolution of the constant level term. Thus taking into account the ABT between the constant level term $\tilde{u}$ and u we see that the time evolution of ϕ can be expressed in terms of u.

This allows us to eliminate the time evolution of ϕ in all equations, thus obtaining—in principle—a relation between u and ϕ. This relation can be considered as a Bäcklund transformation, thus proving that out of the recursion scheme for the symmetry generators of u there arises a recursion scheme for the fields describing the different singularity flows in the hierarchy.

Now one proceeds as before: The Möbius group is a symmetry group for all members of the singularity hierarchy. This group is transformed to a Möbius group with respect to the independent variable for the equation linked reciprocally to the singularity hierarchy. Since this group is not translation invariant we obtain, out of the generator of the translation group, the generators given in (1.13) and (1.12) as starting points for further hierarchies. The only difference relative to the case reported in Sec. I is that the hereditary operator and the base member $K_0(\rho)$ will be different from the case of the Kawamoto-type equation (1.1). Actually, all computations necessary for this analysis can be done on a fairly general level. The details of the theory behind the procedure are reported in a subsequent paper.[27]

An interesting problem would be to study what results from the additional symmetries found for the Kawamoto equation if one goes all the way back to the KK equation (or the CDGSK if one desires). Most probably, these additional symmetries are then annihilated by the corresponding Bäcklund transformations because their generators may lie in the kernel of the infinitesimal form of these Bäcklund transformations. However, by using the inverses of these kernels one can possibly obtain new symmetries, which most probably can be written only in implicit form. This procedure suggests that for all completely integrable equations of fifth order in the $1 + 1$ dimension one can find additional symmetries which make the symmetry group non-Abelian. Furthermore, this probably suggests that in these cases the time-dependent symmetry groups are all isomorphic to $A[\mathrm{st}(3,\mathbb{C})]$, the affine Lie algebra associated to the traceless upper-triangular 3×3 matrices. We believe that by similar methods such a result also can be demonstrated for other integrable systems such as the Korteweg–de Vries hierarchy.

ACKNOWLEDGMENT

Both authors wish to express their gratitude to Professor Colin Rogers for the very many stimulating discussions they were allowed with him on the subject of this paper.

[1] C. Rogers, J. G. Kingston, and W. F. Shadwick, "Reciprocal-type transformations in magnetogasdynamics," J. Math. Phys. **21**, 395 (1979); C. Rogers, M. C. Nucci, and J. G. Kingston, "On reciprocal auto-Bäcklund transformations: Application to a nonlinear hierarchy," Nuovo Cimento **96**, 55 (1986); C. Rogers and M. C. Nucci, "On reciprocal Bäcklund transformations and the Korteweg–de Vries hierarchy," Phys. Scr. **33**, 289 (1986).

[2] C. Rogers, "The Harry Dym equation in $2 + 1$ dimensions: A reciprocal link with the Kodomtsev Petviashvili equation," Phys. Lett. **120**, 15 (1987).

[3] C. Rogers and S. Carillo, "On reciprocal properties of the Caudrey–Dodd–Gibbon and Kaup–Kupershmidt hierarchies," Phys. Scr. **36**, 865 (1987).

[4]S. Kawamoto, "An exact transformation from the Harry Dym equation to the modified KdV equation," J. Phys. Soc. Jpn. **54**, 2055 (1985).
[5]B. Fuchssteiner, "Application of hereditary symmetries to nonlinear evolution equations," Nonlinear Analysis TMA **3**, 849 (1979); B. Fuchssteiner, "The Lie algebra structure of nonlinear evolution equations admitting infinite dimensional Abelian symmetry groups," Prog. Theor. Phys. **65**, 861 (1981).
[6]W. Oevel, "A geometrical approach to integrable systems admitting time dependent invariants," in *Topics in Soliton Theory and Exactly Solvable Nonlinear Equations*, edited by M. J. Ablowitz, M. Kruskal, and B. Fuchssteiner (World Scientific, Singapore, 1987), pp. 108–124.
[7]P. J. Olver, "Applications of Lie groups to differential equations," *Graduate texts in Mathematics* (Springer, Berlin, 1986).
[8]B. Fuchssteiner, "The Lie algebra structure of degenerate Hamiltonian and bi-Hamiltonian systems," Prog. Theor. Phys. **68**, 1082 (1982).
[9]I. M. Gelfand and I. Y. Dorfmann, "Hamiltonian operators and algebraic structures related to them," Funkcional. Anal. i Priložen. **13**, 13 (1974); "The Schouten bracket and Hamiltonian operators," **14**, 71 (1980); "Hamiltonian operators and infinite-dimensional Lie-algebras," **15**, 23 (1981).
[10]B. Fuchssteiner, "Mastersymmetries, higher-order time-dependent symmetries and conserved densities of nonlinear evolution equations," Prog. Theor. Phys. **70**, 1508 (1983).
[11]A. S. Fokas and B. Fuchssteiner, "The hierarchy of the Benjamin–Ono equation," Phys. Lett. A **86**, 341 (1981).
[12]W. Oevel and B. Fuchssteiner, "Explicit formulas for the symmetries and conservation laws of the Kadomtsev–Petviashvili equation," Phys. Lett. A **88**, 323 (1982).
[13]B. Fuchssteiner, "On the hierarchy of the Landau–Lifshitz equation," Physica D **13**, 387 (1984).
[14]B. Fuchssteiner, "Mastersymmetries for completely integrable systems in statistical mechanics," in Proceeding of the Stiges Conference 1984, *Springer Lecture Notes in Physics*, Vol. 216, edited by L. Garrido (Springer, Berlin, 1985), pp. 305–315; E. Barouch and B. Fuchssteiner, "Master symmetries and similarity equations of the XYh-model," Stud. Appl. Math. **73**, 221 (1985).
[15]B. Fuchssteiner, "Some recent results on solitons, symmetries and conservation laws in nonlinear dynamics," in *Proceedings of the 14th* ICGTMP, edited by Y. M. Cho (World Scientific, Singapore, 1986), pp. 421–424.
[16]W. Oevel and B. Fuchssteiner, "New hierarchies of nonlinear completely integrable systems related to a change of variables for evolution parameters, Physica A **68**, 67 (1987).
[17]Tu Gui-Zhang, "A commutativity theorem for partial differential operators," Commun. Math. Phys. **77**, 289 (1980).
[18]B. Fuchssteiner, W. Oevel, and W. Wiwianka, "Computer-algebra methods for investigation of hereditary operators of higher order soliton for equations," Comput. Phys. Commun. **44**, 47.
[19]W. Oevel, private communication.
[20]P. J. Caudrey, R. K. Dodd, and J. D. Gibbon, "A new hierarchy of Korteweg–de Vries equations," Proc. R. Soc. London Ser. A **351**, 407 (1976).
[21]A. K. Sawada and A. T. Kotera, "A method for finding *N*-soliton solutions of the KdV and KdV-like equations," Prog. Theor. Phys. **51**, 1355 (1974).
[22]A. P. Fordy and J. Gibbons, "Some remarkable nonlinear transformations," Phys. Lett. A **75**, 325 (1980).
[23]W. Oevel and B. Fuchssteiner, "The bi-Hamiltonian structure of some nonlinear fifth- and seventh-order differential equations and recursion formulas for their symmetries and conserved covariants," J. Math. Phys. **23**, 358 (1982).
[24]J. Weiss, "On classes of integrable systems and the Painlevé property," J. Math. Phys. **25**, 13 (1984); "Bäcklund transformation and Painlevé property," **27**, 1293 (1986).
[25]A. S. Fokas and B. Fuchssteiner, "Bäcklund transformations for hereditary symmetries," Nonlinear Analysis TMA **5**, 423 (1981).
[26]J. Weiss, M. Tabor, and G. Carnevale, "The Painlevé property for partial differential equations," J. Math. Phys. **24**, 522 (1983); J. Weiss "Painlevé property for partial differential equations. II: Bäcklund transformations, Lax pairs, and the Schwarzian derivative," *ibid.* **24**, 1405 (1983).
[27]B. Fuchssteiner and S. Carillo, "Soliton structure versus singularity analysis, third-order completely integrable nonlinear differential equations in 1 + 1-dimension," Physica A **152**, 467 (1989).

The relations among a special type of solutions in some (D+1)-dimensional nonlinear equations

Sen-yue Lou and Guang-jiong Ni
Physics Department, Fudan University, Shanghai, People's Republic of China

(Received 26 January 1988; accepted for publication 25 January 1989)

Among a special type of solutions of some kinds of nonlinear equations in $(D+1)$ dimensions such as the cubic nonlinear Klein–Gordon, sine–Gordon, double sine–Gordon, Ginzburg–Landau equations, etc., some mapping relations exist that allow one to find a solution of one system from a solution of another. Furthermore, one can find some new solutions from a known one within the same system. Some new N-kinklike solutions are also presented.

I. INTRODUCTION

In recent years nonlinear effects in various areas of science have attracted much attention.[1,2] A striking feature of these investigations is the existence of various similarities or relations among models in quite different categories. For example, the well-known sine-Gordon (sG) system in $(1+1)$ dimensions is equivalent to the massive Thirring model,[3,4] to the two-dimensional Coulomb gas,[5] to the continuous limit of lattice x–y–z spin-$\frac{1}{2}$ model,[6] and to the massive $O(2)$ nonlinear σ model.[5] Between the nonlinear Schrödinger equation and the equation of the ferromagnetic chain, there exists gauge equivalence.[7,8] In this paper, we will treat this problem from another point of view, i.e., try to find some relations of a special type of solutions among different nonlinear equations in high dimensions: These are the sG equation[9]

$$\Phi_{tt}-\sum_{i=1}^{D}\Phi_{x_ix_i}-\frac{M}{g}\sin g\Phi=0\,, \tag{1.1}$$

the double sine–Gordon (DsG) equation[10,11]

$$\tilde{\Phi}_{tt}-\sum_{i=1}^{D}\tilde{\Phi}_{x_ix_i}-\frac{\alpha g}{2}\left[\sin\frac{g}{2}\tilde{\Phi}-2\eta\sin g\tilde{\Phi}\right]=0\,, \tag{1.2}$$

the cubic nonlinear Klein–Gordon (^{3}NKG) equation

$$\phi_{tt}-\sum_{i=1}^{D}\phi_{x_ix_i}+\lambda\phi+\mu\phi^3=0\,, \tag{1.3}$$

and the Ginzburg–Landau (GL) equation[12]

$$f_{tt}-\sum_{i=1}^{D}f_{x_ix_i}+\alpha_1 f+\alpha_2 f^3+\alpha_3 f^{-3}=0\,. \tag{1.4}$$

There are kink solutions for Eqs. (1.1)–(1.4): In Sec. II, we will find mapping relations among a special type of solutions for these equations.

Another quite interesting problem is whether we can obtain some new solutions for one particular equation from one which is known. The Bäcklund transformation is just such a powerful method; however, according to present understanding, not all nonlinear equations can be treated by this method. In Sec. III, we will also propose some algebraic formulas to obtain some new type of solutions from a known solution.

Usually, the N-kinklike soliton solutions for the $(D+1)$-dimensional nonlinear Klein–Gordon equation are constructed by the method of the base equation technique,[13] in which the solutions of a nonlinear differential equation are expressed in terms of a related linear differential equation—the base equation. In Sec. IV, we will propose some new N-soliton solutions for the ^{3}NKG equation. Actually, the models (1.1), (1.2), and (1.4) are also solved simultaneously because of the results in Sec. II. Section V will provide a summary and discussions. In Table I, various traveling wave or N-wave solutions of ^{3}NKG useful in mapping are listed.

II. MAPPING RELATIONS AMONG A SPECIAL TYPE OF SOLUTIONS OF SOME NONLINEAR EQUATIONS

To solve a nonlinear equation is usually a formidable task. However, there are many nonlinear equations in different physical fields. Thus it is quite interesting to establish some relations among different equations even though they are valid only for a special type of solutions.

General nonlinear Klein–Gordon equations in $D+1$ dimensions have the form

$$\Box\phi\equiv\left(\sum_{i=1}^{D}\partial_{x_i}^2-\partial_t^2\right)\phi=F(\phi)\,. \tag{2.1}$$

Equations (1.1)–(1.4) are special cases of (2.1). Furthermore, if we pose the conditions

$$(\tilde{\nabla}\phi)^2\equiv\sum_{i=1}^{D}(\partial_{x_i}\phi)^2-(\partial_t\phi)^2=G(\phi) \tag{2.2}$$

and

$$F(\phi)=\frac{1}{2}\frac{dG(\phi)}{d\phi}\,, \tag{2.3}$$

then we can establish mapping relations among Eqs. (1.1)–(1.4). For simplicity we shall call equation system (2.1) and (2.3) the constrained nonlinear Klein–Gordon system.

A. Mapping relation between constrained DsG and ^{3}NKG

If $\phi(X)$ is a solution of the constrained ^{3}NKG equation

$$\begin{aligned}&\Box\phi(X)=\lambda\phi(X)+\mu\phi^3(X)\,,\\&(\tilde{\nabla}\phi(X))^2=\lambda\phi^2(X)+(\mu/2)\phi^4(X)+C\,,\end{aligned} \tag{2.4}$$

then

$$\tilde{\Phi}(X)=(2n\pi/g)\pm(4/g)\tan^{-1}[\phi(bX)/a]\,, \tag{2.5}$$

with

 0022-2488/89/071614-07$02.50

TABLE I. Some solutions of [3]NKG useful in mapping. The argument V is defined in (4.6) and k is the modulus of the elliptic function $k'^2 = 1 - k^2$.

ϕ	λ	μ	C
$\sqrt{-\lambda/\mu} = 1$	± 1	∓ 1	$\mp\frac{1}{2}$
V	0	0	1
$1/V$	0	2	0
$e^{\pm V}$	1	0	0
$e^{V} + Be^{-V}$	1	0	$-4B$
$\sinh V$	1	0	1
$\cosh V$	1	0	-1
$\tanh V, \coth V$	-2	2	1
$\operatorname{sech} V$	1	-2	0
$\operatorname{csch} V$	1	2	0
$\sin V, \cos V$	-1	0	1
$B\sin V + \cos V$	-1	0	$1 + B^2$
$\tan V, \cot V$	2	2	1
$\sec V, \csc V$	-1	2	0
$\operatorname{sn} V$	$-(1 + k^2)$	$2k^2$	1
$\operatorname{cn} V$	$k^2 - k'^2$	$-2k^2$	k'^2
$\operatorname{dn} V$	$1 + k'^2$	-2	$-k'^2$
$\operatorname{tn} V$	$1 + k'^2$	$2k'^2$	1
$\operatorname{dn} V/\operatorname{sn} V$	$1 - 2k'^2$	2	$-k^2k'^2$
$\operatorname{dn} V/(1 + k\operatorname{sn} V)$	$(1 + k^2)/2$	$(k^2 - 1)/2$	$(k^2 - 1)/4$
$\operatorname{cn} V/(1 + \operatorname{sn} V)$	$(1 + k^2)/2$	$(1 - k^2)/2$	$(1 - k^2)/4$
$k\operatorname{sn} V/(1 + \operatorname{dn} V)$	$-(1 + k'^2)/2$	$k^2/2$	$k^2/4$
$\dfrac{\sqrt{1-k}\,\operatorname{dn} V}{\sqrt{1+k}\,(1 + k\operatorname{sn} V) + \sqrt{2k(1 + k\operatorname{sn} V)(1 + \operatorname{sn} V)}}$	$(1 + 6k + k^2)/8$	$-(1 - k)^2/8$	$-(1 - k)^2/16$
$\dfrac{\sqrt{1-k}\,\operatorname{dn} V}{\sqrt{2k(1 + k\operatorname{sn} V)(1 + \operatorname{sn} V)}}$	$-(1 - 6k + k^2)/4$	$2k$	$-(1 - k)^2/4$
$\dfrac{\sqrt{2k(1 + k\operatorname{sn} V)(1 + \operatorname{sn} V)}}{\sqrt{1+k}\,(1 + k\operatorname{sn} V) + \sqrt{1-k}\,\operatorname{dn} V}$	$-(1 + k^2)/4$	$k/2$	$k/4$
$\dfrac{(1 + k')\operatorname{sn} V}{\sqrt{2[\operatorname{cn}^2 V + \operatorname{dn} V + k'\operatorname{sn}^2 V]}}$	$(1 + 6k' + k'^2)/4$	$2k'$	$(1 + k')^2/4$
$\dfrac{\sqrt{2[\operatorname{cn}^2 V + \operatorname{dn} V + k'\operatorname{sn}^2 V]}}{1 + \operatorname{dn} V + (1 + k')\operatorname{sn} V}$	$(1 + k'^2)/4$	$k'/2$	$k'/4$
$\sqrt{\dfrac{1-k}{2}}\dfrac{\operatorname{cn} V}{\sqrt{(1 + k\operatorname{sn} V)(1 + \operatorname{sn} V)}}$	$-(1 - 6k + k^2)/4$	$(1 - k)^2$	$(1 - k)^2/4$
$\sqrt{\dfrac{(1+k)(1 \pm \operatorname{sn} V)}{2(1 \pm k\operatorname{sn} V)}}$	$-(1 + 6k + k^2)/4$	$2k$	$(1 + k)^2/4$
$\dfrac{2\sqrt{k'}\operatorname{sn} V \operatorname{cn} V}{\operatorname{cn}^2 V \pm k'\operatorname{sn}^2 V}$	$1 + k'^2 \mp 6k'$	$\mp 2(1 \mp k')^2$	$4k'$
$\dfrac{\operatorname{cn}^2 V - k'\operatorname{sn}^2 V}{\operatorname{cn}^2 V + k'\operatorname{sn}^2 V}$	$-2(1 + k')^2$	$2(1 - k')^2$	$(1 + k')^2$
$\dfrac{2\sqrt{k'}\operatorname{dn} V}{k' \pm \operatorname{dn}^2 V}$	$1 + k'^2 \pm 6k'$	$\mp 2(1 \pm k')^2$	$-4k'$
$\dfrac{k' - \operatorname{dn}^2 V}{k' + \operatorname{dn}^2 V}$	$-2(1 + k'^2)$	$2(1 + k')^2$	$(1 - k')^2$

$$a^2 = \frac{(-1)^{n+1}[\lambda \pm \sqrt{\lambda^2 + 2C\mu[(4\eta)^2 - 1]}]}{\mu[4\eta - (-1)^n]} \tag{2.6}$$

and

$$4b^2 = \frac{\alpha g^2[(4\eta)^2 - 1]}{4\eta\lambda \mp (-1)^{n+1}\sqrt{\lambda^2 + 2C\mu[(4\eta)^2 - 1]}} \tag{2.7}$$

must be a solution of the constrained DsG equation

$$\begin{aligned} &\Box\tilde{\Phi}(X) = (\alpha g/2)[2\eta \sin g\tilde{\Phi}(X) - \sin(g/2)\tilde{\Phi}(X)], \\ &(\tilde{\nabla}\tilde{\Phi}(X))^2 = 2\alpha[\cos(g/2)\tilde{\Phi}(X) - \eta\cos g\tilde{\Phi}(x)] + C_D, \end{aligned} \tag{2.8}$$

where C_D is a constant and

$$\begin{aligned} &X = (t, x_1, \ldots, x_D), \quad bX = (bt, bx_1, \ldots, bx_D), \\ &C_D = (2/g^2)(4b^2\lambda - 3\alpha g^2\eta). \end{aligned} \tag{2.9}$$

In (2.6) and (2.7) the choices of (i) the sign $\pm$, (ii) n even or odd, and (iii) α and η belonging to one of four regions

$$\begin{aligned} &[\alpha\eta > 0, |\eta| > \tfrac{1}{4}], \quad [\alpha\eta > 0, |\eta| < \tfrac{1}{4}], \\ &[\alpha\eta < 0, |\eta| > \tfrac{1}{4}], \quad [\alpha\eta < 0, |\eta| < \tfrac{1}{4}] \end{aligned} \tag{2.10}$$

are all determined by the real field conditions $a^2 > 0$ and $b^2 > 0$.

By using relations (2.5)–(2.7) and Table I, we can obtain many of traveling solutions and N-kinklike solutions for the constrained DsG equation.

The inverse is also true, i.e., if (2.5) with (2.6) and (2.7) is a solution of the constrained DsG equation, then $\phi(X)$ must be a solution of the constrained ^{3}NKG equation.

B. Mapping relation between constrained sG and ^{3}NKG

The mapping relation from sG to ^{3}NKG can be stated as follows.

If

$$\Phi(X) = 2n\pi/g \pm (4/g)\tan^{-1}[\phi(bX)/a] \tag{2.11}$$

is a solution of the constrained sG equation

$$\begin{aligned} &\Box\Phi(X) + (M_1/g)\sin g\Phi(X) = 0\,, \\ &(\tilde{\nabla}\Phi(X))^2 - 2(M_1/g^2)\cos g\Phi(X) - C_s = 0\,, \end{aligned} \tag{2.12}$$

then

$$\phi_1(bX) = a\tan(g/4)\Phi(X) = \phi(bX) \tag{2.13}$$

and

$$\phi_2(bX) = \frac{a^2 - \phi^2(bX)}{2a^2\phi(bX)} = \frac{1}{a}\cot\frac{g}{2}\Phi(X) \tag{2.14}$$

are solutions of the constrained ^{3}NKG equation (2.4) with

$$\begin{aligned} &\lambda_1 = (1/8b^2)(g^2C_s - 6M_1)\,, \\ &\mu_1 = (1/8a^2b^2)(g^2C_s + 2M_1)\,, \\ &C_1 = (a^2/16b^2)(g^2C_s + 2M_1) \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} &\lambda_2 = (1/16b^2)[2C_sg^2 - 15(-1)^nM_1]\,, \\ &\mu_2 = (1/16a^2b^2)[2C_sg^2 - 31(-1)^nM_1]\,, \\ &C_2 = (a^2/32b^2)[2C_sg^2 + (-1)^nM_1]\,. \end{aligned} \tag{2.16}$$

The inverse mapping, i.e., the mapping from ^{3}NKG to sG has to be stated carefully, as follows.

If $\phi(X)$ is a solution of the constrained ^{3}NKG equation (2.4) with $2c\mu > 0$, then (2.11) is a solution of Eq. (2.12); however, if $2c\mu < 0$, (2.11) is a solution of Eq. (2.12) with $g \to g/2$, i.e., a solution of

$$\begin{aligned} &\Box\Phi(X) + (2M_2/g)\sin(g/2)\Phi(X) = 0\,, \\ &(\tilde{\nabla}\Phi(X))^2 - (8M_2/g^2)\cos(g/2)\Phi(X) - C_s' = 0\,. \end{aligned} \tag{2.17}$$

The parameters a and b in (2.11) are

$$\begin{aligned} &a_1^4 = 2C/\mu\,, \\ &b_1^2 = M_1/[(\operatorname{sgn}\mu)\sqrt{2C\mu} - \lambda]\,, \\ &g^2C_s = (6\mu + 8b_1^2\lambda) \end{aligned} \tag{2.18}$$

and

$$\begin{aligned} &a_2^2 = (1/\mu)[\lambda \pm \sqrt{\lambda^2 - 2C\mu}]\,, \\ &b_2^2 = (-1)^{n+1}M_2/\sqrt{\lambda^2 - 2C\mu}\,, \\ &g^2C_s' = 8b_2^2\lambda\,, \end{aligned} \tag{2.19}$$

respectively.

In (2.18) and (2.19), the $\pm$ sign and n even or odd are fixed by the real field conditions $a^2 > 0$, $b^2 > 0$.

C. Mapping relation between constrained ^{3}NKG and GL

Between the constrained ^{3}NKG equation (2.4) and the constrained GL equation

$$\begin{aligned} &\Box f(X) = \alpha_1 f(X) + \alpha_2 f^3(X) + \alpha_3 f^{-3}(X)\,, \\ &(\tilde{\nabla}f(X))^2 = \alpha_1 f^2(X) + (\alpha_2/2)f^4(X) \\ &\qquad - \alpha_3 f^{-2}(X) + \alpha_4 \end{aligned} \tag{2.20}$$

there also exists a correspondence

$$\begin{aligned} &f^2(X) \leftrightarrow \phi^2(X) + f_0\,, \\ &\alpha_1 \leftrightarrow \lambda - \tfrac{3}{2}f_0\mu\,, \\ &\alpha_2 \leftrightarrow \mu\,, \\ &\alpha_3 \leftrightarrow \lambda f_0^2 - (\mu/2)f_0^3 - Cf_0\,, \\ &\alpha_4 \leftrightarrow \tfrac{3}{2}f_0^2\mu + C - 2\lambda f_0\,, \end{aligned} \tag{2.21}$$

where f_0 and α_4 are also constants. With the aid of correspondence (2.21) and Table I we can obtain many solutions of GL, some of which are familiar in the literature.[12]

D. Mapping relations of traveling wave solutions between ^{3}NKG and other models

If we confine ourselves further to traveling wave solutions of some nonlinear equations (fortunately, the one-soliton or kink solutions have such a property), we can also obtain some correspondent relations. For example, between the solution u of the Korteweg–deVries (KdV) equation

$$u_t - muu_x + u_{xxx} = 0 \tag{2.22}$$

in $(1+1)$ dimensions and the solution ϕ of ^{3}NKG equation in $(D+1)$ dimensions, we have the following correspondence:

$$\begin{aligned} &u(\eta) \leftrightarrow \phi^2(\xi) + u_0\,, \\ &m \leftrightarrow 6\mu\,, \\ &v \leftrightarrow 4\lambda - 6\mu u_0\,, \\ &J \leftrightarrow 2C - 4\lambda u_0 + 3\mu u_0^2\,, \\ &K \leftrightarrow -4Cu_0 + 4\lambda u_0^2 - 2\mu u_0^3\,, \end{aligned} \tag{2.23}$$

where u_0 is an arbitrary constant,

$$\eta = x - vt\,, \tag{2.24}$$

$$\xi = \frac{\sum_{i=1}^{D} a_i x_i + vt}{\sqrt{\sum_{i=1}^{D} a_i^2 - v^2}}\,, \tag{2.25}$$

while J, K are integration constants of the KdV equation and are obtained by substituting (2.24) into (2.22) and integrating twice:

$$\left(\frac{d}{d\eta}u(\eta)\right)^2 - \frac{m}{3}u^3(\eta) - vu^2(\eta) - 2Ju(\eta) - K = 0\,. \tag{2.26}$$

Between ^{3}NKG in $(D+1)$ dimensions and the nonlinear Schrödinger equation (NLS) in $(1+1)$ dimensions,

$$i\psi_t = \beta_1\psi - \beta_2\psi_{xx} - \beta_3|\psi|^2\psi\,, \tag{2.27}$$

there also exists a mapping relation

$$\begin{aligned} &\psi(x,t) = h(\eta)e^{-i(kx-\omega t)}\,, \\ &h(\eta) \leftrightarrow \phi(\xi)\,, \end{aligned} \tag{2.28}$$

$$(\beta_1+\beta_2 k^2-\omega)/\beta_2 \leftrightarrow \lambda\,, \qquad (2.29)$$
$$-\beta_3/\beta_2 \leftrightarrow \mu\,.$$

Thus far we have found the mapping relations among constrained ^{3}NKG, sG, DsG, GL, and the correspondences among traveling wave solutions of ^{3}NKG and KdV (or NLS) equations. By using these relations and Table I, starting from a known solution of any one of the equations, one is able to find its correspondent in any one of the above-mentioned remaining equations.

III. OBTAINING SOME NEW SOLUTIONS FROM A KNOWN SOLUTION

In Sec. II we find a solution of one model from a solution of another. Now we turn to the following question: Can we obtain some new solutions of one model from a known solution within the same model? As is well known, the Bäcklund transformation is just such a method for integrable systems. In this section, we will propose some algebraic formulas for obtaining a new special type of solutions from a known solution in a constrained ^{3}NKG system. Once this is done, by means of the correspondence relations found in Sec. II, one can then turn to other systems.

(i) If $\phi(X)$ is a solution of the constrained ^{3}NKG equation (2.4) with

$$\lambda+\tfrac{1}{2}\mu+C=0 \qquad (3.1)$$

and $\sqrt{1-\phi^2}$ is real, then

$$\phi_1=\phi/(\sqrt{1-\phi^2}+\theta) \quad (\theta=\pm 1)\,, \qquad (3.2)$$
$$\phi_2=\phi/\sqrt{1-\phi^2}\,, \qquad (3.3)$$

and

$$\phi_3=\sqrt{1-\phi^2}/(\theta+\phi) \qquad (3.4)$$

are all solutions of the constrained ^{3}NKG equation with

$$\begin{aligned}&\lambda_1=\lambda+\tfrac{3}{2}C\,,\\ &\mu_1=\tfrac{1}{2}C\,,\\ &C_1=\tfrac{1}{4}C\,,\end{aligned} \qquad (3.5)$$

$$\begin{aligned}&\lambda_2=3C+\lambda\,,\\ &\mu_2=4C+2\lambda\,,\\ &C_2=C\,,\end{aligned} \qquad (3.6)$$

and

$$\begin{aligned}&\lambda_3=-\tfrac{1}{2}\lambda\,,\\ &\mu_3=\tfrac{1}{2}(C-\tfrac{1}{2}\mu)\,,\\ &C_3=\tfrac{1}{4}(C-\tfrac{1}{2}\mu)\,,\end{aligned} \qquad (3.7)$$

respectively.

Generally, for an arbitrary solution of (2.4), condition (3.1) is not satisfied. In this case one may reconstruct a new solution of ^{3}NKG, say, $B\phi(AX)$ to satisfy condition (3.1). It is easy to see that if $\phi(X)$ is a solution of ^{3}NKG, then $B\phi(AX)$ is also a solution of ^{3}NKG with

$$\begin{aligned}&\lambda'=A^2\lambda\,,\\ &\mu'=(A^2/B^2)\mu\,,\\ &C'=A^2B^2C\,.\end{aligned} \qquad (3.8)$$

Thus if λ, μ, and C do not satisfy (3.1), one may choose

$$B^2=(-\lambda\pm\sqrt{\lambda^2-2C\mu})/2C \qquad (3.9)$$

presumably such that the condition

$$C'+\lambda'+\tfrac{1}{2}\mu'=0 \qquad (3.10)$$

is satisfied. The sign $\pm$ in (3.9) is required because the condition of B^2 is positive: $B^2>0$. However, if $B^2<0$ for $+$ and $-$ (for example, λ, μ, and C are all positive or negative), then the above procedure will fail to obtain a new real solution. In this case, one may instead choose

$$B^2=(\lambda\pm\sqrt{\lambda^2-2C\mu})/2C\,, \qquad (3.11)$$

so that $C'+\frac{1}{2}\mu'-\lambda'=0$ is satisfied. Thus we can construct some new solutions because of the following property.

(ii) If $\phi(X)$ is a solution of the constrained ^{3}NKG equation (2.4) with

$$C+\tfrac{1}{2}\mu-\lambda=0\,, \qquad (3.12)$$

then

$$\phi_4=\phi/\sqrt{1+\phi^2} \qquad (3.13)$$

and

$$\phi_5=1/\sqrt{1+\phi^2} \qquad (3.14)$$

are solutions of the constrained ^{3}NKG equation (2.4) with

$$\begin{aligned}&\lambda_4=\lambda-3C\,,\\ &\mu_4=4C-2\lambda\,,\\ &C_4=C\end{aligned} \qquad (3.15)$$

and

$$\begin{aligned}&\lambda_5=\lambda-\tfrac{3}{2}\mu\,,\\ &\mu_5=2\mu-2\lambda\,,\\ &C_5=\tfrac{1}{2}\mu\,,\end{aligned} \qquad (3.16)$$

respectively.

However, if $\lambda^2-2C\mu<0$, (3.9) and (3.11) become complex. Then we must try to find some other relation to obtain new solutions; fortunately, it is possible. When $\lambda^2-2C\mu<0$, both C and μ must have the same sign (positive or negative); then we can choose

$$B^4=\mu/2C \qquad (3.17)$$

such that $\mu'=2C'$ is satisfied. Hence we have another property, as follows.

(iii) If $\phi(X)$ is a solution of the constrained ^{3}NKG equation (2.4) with

$$\mu=2C\,, \qquad (3.18)$$

then

$$\phi_6=2\phi/(1+\theta\phi^2) \quad (\theta=\pm 1) \qquad (3.19)$$

and

$$\phi_7=(1-\phi^2)/(1+\phi^2) \qquad (3.20)$$

are solutions of the ^{3}NKG equation (2.4) with

$$\begin{aligned}&\lambda_6=\lambda-6\theta C\,,\\ &\mu_6=4C-2\theta\lambda\,,\\ &C_6=4C\end{aligned} \qquad (3.21)$$

and

$$\lambda_7 = -2\lambda,$$
$$\mu_7 = 2\lambda - 4C, \tag{3.22}$$
$$C_7 = \lambda + 2C,$$

respectively.

For a special solution of ^{3}NKG, sometimes more than one of (3.1), (3.12), and (3.18) can be satisfied. For example, for the solution $\tan V$ with $\lambda = 2, \mu = 2$, and $C = 1$, V is as given in Sec. IV and conditions (3.12) and (3.18) are satisfied at the same time, while for the solution $\tanh V$ with $\lambda = -2, \mu = 2$, and $C = 1$, conditions (3.1) and (3.18) are satisfied.

To conclude this section, we wish to point out that the inverse of a known solution is also a solution of the ^{3}NKG equation. Indeed, we have the following property.

(iv) If $\phi(X)$ is a solution of the ^{3}NKG equation (2.4), then

$$\phi_8 = 1/\phi \tag{3.23}$$

is also a solution of the ^{3}NKG equation, but with

$$\lambda_8 = \lambda,$$
$$\mu_8 = 2C, \tag{3.24}$$
$$C_8 = \tfrac{1}{2}\mu.$$

IV. SOME NEW N-KINKLIKE SOLUTIONS IN $D+1$ DIMENSIONS

In Secs. II and III we presented some mapping relations among different constrained nonlinear equations and some algebraic formulas for obtaining some new solutions of a constrained equation from a known solution. In this section we shall present some concrete examples which satisfy the condition of constrained equations. Here we treat the ^{3}NKG model only in light of the results of Sec. II.

If we rewrite $\phi(X)$ as

$$\phi(X) \equiv \phi(g(X)), \tag{4.1}$$

then (2.4) becomes

$$(\Box g)\phi_g + (\tilde{\nabla} g)^2 \phi_{gg} = \lambda\phi + \mu\phi^3,$$
$$(\tilde{\nabla} g)^2 \phi_g^2 = \lambda\phi^2 + (\mu/2)\phi^4 + C. \tag{4.2}$$

Now, because of (4.1), the variable X should not appear explicitly in (4.2): then $\Box g$ and $(\tilde{\nabla} g)^2$ are only functions of g. Thus the function g satisfies the base equations

$$\Box g = A(g), \quad (\tilde{\nabla} g)^2 = B(g). \tag{4.3}$$

Generally, it is difficult to solve (4.2) and (4.3) simultaneously. Furthermore, if g is again a solution of a constrained equation, i.e.,

$$A(g) = \frac{1}{2}\frac{dB(g)}{dg}, \tag{4.4}$$

then Eq. (4.2) becomes a one-dimensionlike equation:

$$\frac{d^2}{dV^2}\phi = \lambda\phi + \mu\phi^3,$$
$$\left(\frac{d\phi}{dV}\right)^2 = \lambda\phi^2 + \frac{\mu}{2}\phi^4 + C, \tag{4.5}$$

with

$$V = \int B^{-1/2}\, dg. \tag{4.6}$$

The kinklike solution of (4.5) reads as

$$\phi = (\sqrt{-\lambda/\mu})\tanh(\sqrt{-\lambda/2})V \quad (\lambda<0, \mu>0). \tag{4.7}$$

The solitonlike solution of (4.5) is

$$\phi = \sqrt{-2\lambda/\mu}\ \mathrm{sech}(\sqrt{\lambda})V \quad (\lambda>0, \mu<0). \tag{4.8}$$

Many other solutions of (4.5) are listed in Table I.

In (4.6) $B(g)$ may be an arbitrary function of g; thus we see that the solution of any constrained equation can be related to the solutions of the constrained ^{3}NKG equation.

Therefore, we can take the base equation in the simplest form with

$$\Box g = \alpha^2 g, \tag{4.9a}$$
$$(\tilde{\nabla} g)^2 = \alpha^2 g^2 \tag{4.9b}$$

(where α is a constant) and, then,

$$V = (1/\alpha)\ln g. \tag{4.10}$$

By means of the base equation (4.9), many authors studied multiple solitonlike or kinklike solutions in high dimensions[13–15] and obtained a set of solutions of (4.9) expressed as

$$g_0 = \sum_{\gamma=1}^{N} \exp \alpha\theta_\gamma \tag{4.11}$$

with

$$\theta_\gamma = \sum_{j=1}^{D} P_\gamma^j x_j + \omega_\gamma t + \delta_\gamma, \tag{4.12}$$

$$\sum_{j=1}^{D} (P_\gamma^j)^2 - \omega_\gamma^2 = 1, \tag{4.13}$$

and

$$\sum_{j=1}^{D} (P_\gamma^j - P_{\gamma'}^j)^2 - (\omega_\gamma - \omega_{\gamma'})^2 = 0. \tag{4.14}$$

After substituting (4.11) into (4.7) [or (4.8)], one obtains the N-kink solution (or the N-soliton solution). Consequently, the solutions listed in Table I evolve into various types of N-wave solutions.

Now we are in a position to obtain some new N-kink solutions of the ^{3}NKG equation by solving new solutions of the base equation (4.9). Since (4.9) is linear, although one may combine two solutions of the type (4.11) linearly into one solution, no new information is substantially obtained. Thus we must extend (4.11) to a more general form. Fortunately, it is quite easy to verify that Eq. (4.9) has the solution

$$g_1 = \left(\sum_{\gamma=1}^{N} \exp \alpha_1\theta_\gamma\right)^{\alpha_2} \tag{4.15}$$

with (4.12)–(4.14) and

$$\alpha_1^2\alpha_2^2 = \alpha^2. \tag{4.16}$$

When $\alpha_2 = 1$, (4.15) reduces to (4.11). Now we can linearly combine various such solutions into one solution:

$$g_2 = \sum_{k=1}^{m} a_k \left(\sum_{\gamma_k=1}^{N_k} \exp \alpha_{1k}\theta_{\gamma_k}\right)^{\alpha_{2k}} \tag{4.17}$$

with

$$\alpha_{1k}^2\alpha_{2k}^2 = \alpha^2 , \tag{4.18}$$

$$\theta_{\gamma_k} = \sum_{j=1}^{D} P^j_{\gamma_k} x_j + \omega_{\gamma_k} t + \delta_{\gamma_k} , \tag{4.19}$$

while the constraint conditions are

$$\sum_{j=1}^{D} (P^j_{\gamma_k})^2 - (\omega_{\gamma_k})^2 = 1 \tag{4.20}$$

and

$$\sum_{j=1}^{D} (P^j_{\gamma_k} - P^j_{\gamma'_{k'}})^2 - (\omega_{\gamma_k} - \omega_{\gamma'_{k'}})^2 = 0 . \tag{4.21}$$

In Eq. (4.17) the a_k may all be set to 1 without the loss of generality as a result of the existence of δ_k as an arbitrary constant. The number of terms superposing in (4.17), m, can be any positive integer.

By means of the two-base equation technique first introduced in Ref. 16, it is easy to obtain the second type of solutions of (4.9):

$$g_3 = \sqrt{g_2 g_2'} , \tag{4.22}$$

where g_2 and g_2' can take the form of (4.17). Substituting g_1, g_2, and g_3 into (4.10) and then into the solutions in Table I, one can obtain many new types of N-wave (including some N kink or N soliton) solutions.

It is necessary to mention that for our N-soliton solutions, at first sight, as in some other authors' opinions,[13–15] the integer number N has an upper bound $N \leqslant 2D + 1$ because of conditions (4.13) and (4.14) (D is the space dimension); however, it is not true. Some of the equations (4.13) and (4.14) may be degenerate. Here let us give an eight-($> 2D + 1 = 5, D = 2$) kink solution of the ^{3}NKG equation explicitly,

$$\begin{aligned} g_{01} = {} & \exp(x + 2y + 2t + \delta_1) + \exp[(1/5)(-3x + 4y) + \delta_2] + \exp[(1/5)(x + 7y + 5t) + \delta_3] \\ & + \exp[(1/4)(5y + 3t) + \delta_4] + \exp[(1/10)(-2x + 11y + 5t) + \delta_5] + \exp[(1/8)(4x + 13y + 11t) + \delta_6] \\ & + \exp[(1/16)(4x + 23y + 17t) + \delta_7] + \exp[(1/5)(13x + 16y + 20t) + \delta_8] , \end{aligned} \tag{4.23}$$

$$\phi = (\sqrt{-\lambda/\mu})\,\mathrm{th}(\sqrt{-\lambda}/2)\ln g_{01} \quad (\lambda < 0, \mu > 0) , \tag{4.24}$$

where δ_γ ($\gamma = 1,\ldots,8$) are arbitrary constants.

More about N-kink solutions with $N > 2D + 1$ will be discussed elsewhere.

V. SUMMARY AND DISCUSSION

We have proposed a method of mapping so that a special type of solutions of many nonlinear systems such as ^{3}NKG, SG, DsG, GL, KdV, NLS, etc. could be found at the same time. Once one of them is known, the others are also obtained. These mapping relations have been discussed in Sec. II. On the other hand, by using the results of Sec. III, and starting from a known solution of the ^{3}NKG equation, one is able to find a group of solutions for ^{3}NKG (and then for other models). For example, beginning from e^V with $\lambda = 1$ and $\mu \to C = 0$ (see Table I), one easily finds that e^{-V}, $\sinh V$, $\tanh V$, $\cosh V$, $\mathrm{sech}\, V$, $\mathrm{csch}\, V$, and $\coth V$, etc., are all solutions of ^{3}NKG, but, of course, they are accompanied by different parameters (see Table I). Although some of the solutions of ^{3}NKG listed in Table I are singular, they are useful in mapping relations (perhaps, also, in Bäcklund transformation). Once they are mapped to sG or DsG, they will become finite and nonsingular.

It is also worthwhile to mention that each of these models contains more than one characteristic structure. For example, the parameters in ^{3}NKG are separated into four regions: $[\lambda > 0, \mu > 0]$, $[\lambda > 0, \mu < 0]$, $[\lambda < 0, \mu < 0]$, and $[\lambda < 0, \mu > 0]$, where the latter is the famous ϕ^4 model describing spontaneous symmetry breaking. Also, the DsG system contains four characteristic structures, as shown in (2.10). It is all four structures rather than one structure that should be involved in our mapping; thus it is no surprise that, say, $\tanh V$ is a solution of the ϕ^4 model, whereas $\sinh V$, $\cosh V$, etc., are not, although the latter are solutions of ^{3}NKG (as mentioned above).

In this paper our relations are valid only for a special kind of solutions between these models. Fortunately, all the traveling solutions and N-kinklike solutions belong to this category. On the other hand, the breatherlike solutions [for the (1 + 1) sG model] are beyond our control. Notice however, that the N-solitons or kinklike solutions discussed in this paper are not the traditional n-kink solutions obtained by the Bäcklund transformation method: The latter kind of n-kink solutions, which usually exists in (1 + 1) dimensions,[16] are different from ours.

In summary, by using the simple methods in this paper, many solutions of different nonlinear equations will be obtained; while some of them are already known, many are new and deserve further investigation.

ACKNOWLEDGMENT

We acknowledge the support of the Science Foundation of the National Education Committee of China under Grant No. KF 12016.

[1] R. Z. Sagdeev, *Nonlinear and Turbulent Processes in Physics* (Harwood, New York, 1984), Vols. 1–3.
[2] R. Ingermanson, Nucl. Phys. B **266**, 620 (1986).
[3] S. Coleman, Phys. Rev. D **11**, 2088 (1975).
[4] S. Mandelstam, Phys. Rev. D **11**, 3026 (1975).
[5] S. Samuel, Phys. Rev. D **18**, 1916 (1978).
[6] A. Luther, Phys. Rev. B **14**, 2153 (1976).
[7] V. E. Zakharov and A. B. Shabat, Sov. Phys. JETP **34**, 62 (1972).
[8] H. C. Fogedby, J. Phys. A: Math. Gen. **13**, 1467 (1980).
[9] G-j. Ni, J-j. Xu, and W. Chen, J. Phys. A: Math. Gen. **18**, 149 (1985).

[10]R. K. Bollough, P. J. Caudrey, and H. M. Gibbs, in *Solitons*, edited by R. K. Bollough and P. J. Caudrey, Topics in Current Physics (Springer, Berlin, 1980), Vol. 17, pp. 107–143.
[11]D. K. Cambell, M. Peyrard, and P. Sodano, Physica D **19**, 165 (1986).
[12]P. V. Christiansen, E. B. Hansen, and C. J. Sjöström, J. Low Temp. Phys. **4**, 349 (1971).
[13]P. B. Burt, Proc. R. Soc. London Ser. A **359**, 479 (1978).
[14]J. Gibbon, in *Solitons and Condensed Matter Physics*, edited by A. R. Bishop and T. Schneider (Springer, Berlin, 1978), p. 297.
[15]J. D. Gibbon, N. C. Freeman, and R. S. Johnson, Phys. Lett. A **65**, 380 (1978).
[16]J. D. Gibbon, N. Q. Freeman, and A. Davey, J. Phys. A **11**, L93 (1978).
[17]R. K. Dodd and R. K. Bullough, Proc. R. Soc. London Ser. A **351**, 499 (1976).

Foldy–Wouthuysen-type reduction of the second-order Dirac equation

Levere Hostler
Physics Department, Wilkes College, Wilkes Barre, Pennslyvania 19766

(Received 3 August 1988; accepted for publication 15 March 1989)

A nonrelativistic reduction of the second-order Dirac equation is investigated. An equation $K^2 + [\Pi_0; K - \hat{\Pi}] = m^2 + \hat{\Pi}^2$ is obtained for a "kinetic Hamiltonian" K such that Φ obeys a nonrelativistic Schrödinger-type equation $\Pi_0\Phi = K\Phi$ in addition to obeying the second-order Dirac equation. Non-self-adjoint terms in K are removed by a change of representation from Φ to $\Phi^T \equiv P\Phi$, and a closed expression $P \equiv [\frac{1}{2} + (K^\dagger - \hat{\Pi})(K - \hat{\Pi})/2(m)^2]^{1/2}$ is obtained for P. Similar results are obtained for antiparticle states, and a type of "factorization theorem" for the second-order Dirac equation is obtained in which the second-order Dirac equation is replaced by a pair of uncoupled first-order equations. Since nonperturbative techniques are used throughout, the factorization theorem holds independently of the question of convergence of the Foldy–Wouthuysen series and is valid whenever the equation $K^2 + [\Pi_0; K - \hat{\Pi}] = m^2 + \hat{\Pi}^2$ admits a solution for K. The connection with the usual linear Dirac equation is investigated, and closed expressions are described for a transformation to a representation of that equation which is "even" in the language of Foldy and Wouthuysen. The transformation described is equivalent to two consecutive "odd" unitary transformations.

I. INTRODUCTION AND CONCLUSION

The second-order Dirac equation

$$(\Pi_0 + \hat{\Pi})(\Pi_0 - \hat{\Pi})\Phi = m^2\Phi, \qquad \Pi_0 = \left(i\frac{\partial}{\partial t} - qV\right), \quad \hat{\Pi} \equiv \vec{\sigma}\cdot\left(\frac{1}{i}\nabla - q\vec{A}\right), \tag{1.1}$$

has been investigated before by a number of authors.[1–13] In Eq. (1.1) Φ is a 2×1 Pauli spinor, and the components of $\vec{\sigma}$ are the usual 2×2 Pauli matrices. The use of Eq. (1.1) to describe a spin-$\frac{1}{2}$ particle brings out a close parallel between the theory of a Dirac particle and the theory of a simple scalar particle, with the function[14]

$$\overline{\Phi} \equiv \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi}) \tag{1.2}$$

playing the role of Φ^* in the corresponding scalar theory. For example, the Dirac inner product can be shown to have the representation

$$(\Phi_2;\Phi_1) = \int d^3r \frac{1}{m^2}\overline{\Phi}_2(\overleftrightarrow{\Pi}_0)\Phi_1 \tag{1.3}$$

when expressed in terms of two solutions of Eq. (1.1). This is just the form expected from the theory of a scalar particle, except for the appearance of the dual state $\overline{\Phi}$ in place of Φ^*.

The inner product (1.3) is actually equal to the corresponding inner product of 4×1 Dirac spinors when the wave equation (1.1) is taken into account.[15] Accordingly, the inner product (1.3) inherits from the Dirac inner product the properties of being positive definite[16] and of being Lorentz invariant and a constant of the motion. The property of (1.3) of being a constant of the motion is exploited in the sequel to prove the self-adjoint nature of the final Hamiltonian obtained after transformation (Secs. II and III) and to show the equivalence of the two final Hamiltonians obtained by two different methods (Sec. IV).

A program of quantum electrodynamic calculations[17] using the second-order Dirac equation has provided the motivation for this study. Since the results of this study are expected to have a more general utility, it was decided to publish them separately. Also, the material has a certain intrinsic interest in illuminating some aspects of the Foldy–Wouthuysen transformation.

In Sec. II an equation $K^2 + [\Pi_0; K - \hat{\Pi}] = m^2 + \hat{\Pi}^2$, Eq. (2.5), is derived for a "kinetic Hamiltonian" K. The operator K is defined such that an equation $\Pi_0\Phi = K\Phi$ of a nonrelativistic Schrödinger-type is obeyed in addition to the second-order Dirac equation (1.1) itself. Non-self-adjoint terms in K are removed by a change of representation from Φ to $\Phi^T \equiv P\Phi$. The closed expression

$$P = \left[\frac{1}{2} + \frac{(K^\dagger - \hat{\Pi})(K - \hat{\Pi})}{2(m)^2}\right]^{1/2},$$

Eq. (2.7), is obtained for P.

In Sec. III it is shown that similar results hold for the dual space, the space of functions $\overline{\Phi} \equiv \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})$, and a second representation with its own Hamiltonian, $L \neq K$, is obtained.

In Sec. IV a study of the interplay between the space Φ and the dual space $\overline{\Phi}$ suggests a further transformation $\sqrt{U}$ having the effect of removing all terms in the two Hamiltonians K and L that differ. This last transformation is unitary, whereas P was self-adjoint.

Section V considers antiparticle states and obtains results entirely parallel to the results of Secs. II–IV.

In Sec. VI implications for the usual linear Dirac equation are investigated. The combined transformations P and $\sqrt{U}$ in the space Φ are found to correspond in the language of Foldy and Wouthuysen to two consecutive "odd" unitary transformations of the conventional linear Dirac equation, bringing that equation to "even" form.

The final nonrelativistic Hamiltonian, Eq. (4.9), agrees with the nonrelativistic Hamiltonians obtained by Foldy and

 0022-2488/89/071621-10$02.50

Wouthuysen[18] and by Eriksen[19] through terms of order $(1/m)^3$ except for the same $(1/2m)^3 q^2 E^2$ term for which Foldy and Wouthuysen and Eriksen differ.[20] Including the result presented here there are now three different coefficients for this term, depending upon the method. These differences are not fundamental, however: all three representations are unitarily equivalent ways of bringing the Dirac Hamiltonian to even form. One feature that the present method has in common with Eriksen's method is the ability to give formal closed expressions for the transformations involved.

The second-order Dirac equation treated here is close to the equation treated by Pauli,[21] who used an "elimination method" later amended by Achiezer and Beresteckij.[22] Achiezer and Beresteckij exploited the representation independence of the inner product in Hilbert space, a technique that characterizes the present approach as well. Aside from this, the method presented here appears quite different from the elimination method, since the present approach deals with closed expressions throughout, whereas the elimination method involves a step by step series expansion closer to the original Foldy–Wouthuysen approach.

In Appendix A an alternative derivation of the Hamiltonian (4.9) using a step by step series approach is described. This derivation offers a closer parallel between the present method and the Pauli elimination method. In the elimination method a nonrelativistic structure is achieved by iterating an energy eigenvalue equation in order to push unwanted dependence on the eigenvalue to higher and higher order. In the method of the appendix a nonrelativistic structure is achieved by iterating the time-dependent second-order Dirac equation in order to push unwanted time derivatives of the wave function to higher and higher order. The method of Pauli *et al.* has been shown to lead to results in agreement with the method of Eriksen.[23] Accordingly, the final Hamiltonian (4.9) differs from that of the elimination method in the third order, although, again, the difference is zero modulo unitary equivalence.

It is worth noting that our results for particles and antiparticles taken together constitute a type of "factorization theorem" for the second-order Dirac equation according to which the second-order Dirac equation may be replaced by a pair of uncoupled first-order equations. Since nonperturbative techniques are used throughout, this factorization theorem holds independently of the question of convergence of the Foldy–Wouthuysen series,[24] and is valid whenever the equation $K^2 + [\Pi_0;\, K - \hat{\Pi}] = m^2 + \hat{\Pi}^2$ admits a solution for K. On the other hand, aside from certain special cases discussed below, little is known at the present time about this equation for K.

Explicit closed formal expressions for K are obtainable in certain classic special cases: a free particle and a particle in a static magnetic field. In Appendix B a solution for K in the case of a static spherically symmetric electric potential is exhibited. As discussed in the Appendix, this K illustrates the above factorization theorem but lacks positiveness properties that one would be inclined to impose on a physically acceptable nonrelativistic limit.

Finally, it seems clear that the methods contained herein apply equally well to a scalar particle.

II. NONRELATIVISTIC REPRESENTATION FOR Φ

The basic idea explored here is to investigate the possibility of the existence of special solutions of Eq. (1.1) that satisfy in addition to Eq. (1.1) an equation having a "nonrelativistic Schrödinger equation" type of structure:

$$\Pi_0 \Phi = K\Phi, \tag{2.1}$$

in which K is a suitable operator not involving time derivatives of Φ, and representing the Hamiltonian less the potential energy term. For definiteness in the following K will be called the kinetic Hamiltonian, although sometimes in order to avoid awkward language the term is shortened to just "Hamiltonian." There is a definite advantage in working entirely in terms of gauge invariant quantities and writing the equation of motion in the gauge invariant form (2.1) involving the kinetic Hamiltonian.

In order to investigate the possibility of obeying Eq. (2.1) in addition to the original Eq. (1.1), Eq. (2.1) is used to evaluate the time derivatives in Eq. (1.1):

$$\begin{aligned}
&(\Pi_0 + \hat{\Pi})(\Pi_0 - \hat{\Pi})\Phi \\
&\quad = (\Pi_0\Pi_0 - [\Pi_0;\hat{\Pi}] - (\hat{\Pi})^2)\Phi \\
&\quad = (\Pi_0 K - [\Pi_0;\hat{\Pi}] - (\hat{\Pi})^2)\Phi \\
&\quad = ([\Pi_0;K] + K\Pi_0 - [\Pi_0;\hat{\Pi}] - (\hat{\Pi})^2)\Phi \\
&\quad = (K^2 + [\Pi_0;K - \hat{\Pi}] - (\hat{\Pi})^2)\Phi.
\end{aligned}$$

Note that in the $\Pi_0\Pi_0\Phi$ term the K that is introduced in place of the right-hand factor of Π_0 will intervene between the Φ and the left-hand factor of Π_0, preventing us from making a further direct replacement of the left-hand factor by K. The remedy is to interchange the order of the factors $\Pi_0 K$ before obtaining K^2 for this term, but this entails a commutator correction (line 4). Taking into account the second-order Dirac equation (1.1), the above calculation tells us that

$$(K^2 + [\Pi_0;K - \hat{\Pi}] - (\hat{\Pi})^2)\Phi = m^2\Phi \tag{2.2}$$

for any Φ simultaneously obeying Eqs. (1.1) and (2.1).

The two functions Φ and $\dot{\Phi}$ may be arbitrarily prescribed at one moment of time for the second-order Dirac equation. Equivalently, we may prescribe the pair of functions $(\Phi, \Pi_0\Phi)$ as initial data, and are at liberty to focus on the linear manifold M of solutions of the second-order Dirac equation for which the initial data has the form $(\Phi, \Pi_0\Phi \equiv K\Phi)_i$ in accordance with Eq. (2.1). Now for Eq. (2.1) Φ_i alone is a complete set of initial data, so that in (2.2) we can replace Φ by $\Phi = V\Phi_i$, where V is the time evolution operator for Eq. (2.1),

$$(K^2 + [\Pi_0;K - \hat{\Pi}] - (\hat{\Pi})^2)V\Phi_i = m^2 V\Phi_i. \tag{2.3}$$

On the other hand Φ_i varies freely as Φ ranges over the linear manifold M, so the identity (2.3) can be true for all solutions in M if and only if the operator statement

$$(K^2 + [\Pi_0;K - \hat{\Pi}] - (\hat{\Pi})^2)V = m^2 V, \tag{2.4}$$

holds. Dropping the factor V, assumed to be nonsingular, we obtain the operator identity,

$$K^2 + [\Pi_0;K - \hat{\Pi}] = m^2 + \hat{\Pi}^2. \tag{2.5}$$

This operator identity is a necessary and sufficient condition

that both Eqs. (1.1) and (2.1) shall be obeyed for all time, for any Φ in M.

To solve Eq. (2.5), K is expressed in the form $K = m + T$, and then the equation is iterated to obtain an expansion for T in ascending powers of $1/m$. The result is

$$\begin{aligned} K = {} & (1/2\lambda) + \lambda\hat{\Pi}^2 - \lambda^2[\Pi_0;[\Pi_0;\hat{\Pi}]] - \lambda^3(\hat{\Pi}^2)^2 \\ & - \lambda^3([\Pi_0;\hat{\Pi}])^2 + \lambda^3[\Pi_0;[\Pi_0;\hat{\Pi}^2]] + \lambda\,[\Pi_0;\hat{\Pi}] \\ & - \lambda^2[\Pi_0;\hat{\Pi}^2] - \lambda^3(\hat{\Pi}^2[\Pi_0;\hat{\Pi}] + [\Pi_0;\hat{\Pi}]\hat{\Pi}^2) \\ & + \lambda^3[\Pi_0;[\Pi_0;[\Pi_0;\hat{\Pi}]]] + O(\lambda^4), \\ & \lambda \equiv 1/2m. \end{aligned} \tag{2.6}$$

Another solution of the equation for K is obtained by substituting the form $K = -m + T$, and iterating. This solution, which describes negative energy states, can be obtained from Eq. (2.6) by the substitution $\lambda \to -\lambda$.

The operator K defined through Eq. (2.5) is not self-adjoint. To investigate this the Dirac inner product (1.3) is employed. Equation (2.1) is used to reduce the inner product as follows:

$$\begin{aligned} (\Phi_2;\Phi_1) &= \int d^3r \frac{1}{m^2}(\Phi_2)^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})(\vec{\Pi}_0 + \overleftarrow{\Pi}_0)\Phi_1 \\ &= \int d^3r \frac{1}{m^2}(\Phi_2)^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})((\overleftarrow{\Pi}_0 + \hat{\Pi}) \\ &\quad + (\vec{\Pi}_0 - \hat{\Pi}))\Phi_1 \\ &= \int d^3r \frac{1}{m^2}(\Phi_2)^\dagger(m^2 + (\overleftarrow{\Pi}_0 - \hat{\Pi})(\vec{\Pi}_0 - \hat{\Pi}))\Phi_1 \\ &= \int d^3r \frac{1}{m^2}(\Phi_2)^\dagger(m^2 + (K^\dagger - \hat{\Pi})(K - \hat{\Pi}))\Phi_1 \\ &= \int d^3r\, 2(\Phi_2)^\dagger(P)^2\Phi_1 \end{aligned}$$

in which[25]

$$P \equiv \left[\frac{1}{2} + \frac{(K^\dagger - \hat{\Pi})(K - \hat{\Pi})}{2(m)^2}\right]^{1/2}. \tag{2.7}$$

It is seen that the inner product (1.3) can be brought to the form

$$(\Phi_2;\Phi_1) = \int d^3r\, 2(\Phi_2^T)^\dagger\Phi_1^T, \tag{2.8}$$

appropriate for a nonrelativistic Schrödinger equation. In Eq. (2.8) Φ^T represents a transformed state

$$\Phi^T \equiv P\Phi, \tag{2.9}$$

which can be thought of as a "nonrelativistic representation" of Φ.

The kinetic Hamiltonian in the Φ^T representation is computed as follows:

$$\begin{aligned} & \Pi_0\Phi = K\Phi, \\ & P\Pi_0P^{-1}\Phi^T = PKP^{-1}\Phi^T, \\ & [P;\Pi_0]P^{-1}\Phi^T + \Pi_0\Phi^T = PKP^{-1}\Phi^T, \\ & \Pi_0\Phi^T = (PKP^{-1} + [\Pi_0;P]P^{-1})\Phi^T. \end{aligned}$$

With

$$K^T \equiv (PKP^{-1} + [\Pi_0;P]P^{-1}), \tag{2.10}$$

the wave equation

$$\Pi_0\Phi^T = K^T\Phi^T$$

in the transformed variables is obtained. By means of the transformation P the nonrelativistic form (2.1) of the equation of motion has been preserved, and the nonrelativistic structure (2.8) of the dot product in Hilbert space has been achieved. It follows that the new kinetic Hamiltonian is self-adjoint. To prove this recall that in nonrelativistic quantum mechanics the fact that the inner product (2.8) is a constant of the motion follows from the Schrödinger equation assuming a self-adjoint Hamiltonian. Here the argument is just reversed. The inner product (2.8) is known to be a constant of the motion and the self-adjoint nature of the Hamiltonian must be deduced. Thus

$$\begin{aligned} 0 &= i\frac{\partial}{\partial t}(\Phi_2;\Phi_1) = i\frac{\partial}{\partial t}\int d^3r\, 2(\Phi_2^T)^\dagger\Phi_1^T \\ &= \int d^3r\, 2(\Phi_2^T)^\dagger(\vec{\Pi}_0 - \overleftarrow{\Pi}_0)\Phi_1^T, \\ 0 &= \int d^3r\, 2(\Phi_2^T)^\dagger(K^T - (K^T)^\dagger)\Phi_1^T. \end{aligned}$$

Since the states Φ_2^T and Φ_1^T may be arbitrarily prescribed at one moment of time, it follows that the integrand of this last integral must vanish:

$$(K^T)^\dagger = K^T. \tag{2.11}$$

The series expansion of K^T can be obtained by substituting the series (2.6) for K in Eq. (2.7) for P and expanding, and then applying Eq. (2.10),

$$\begin{aligned} K^T = {} & (1/2\lambda) + \lambda\hat{\Pi}^2 - \lambda^3(\hat{\Pi}^2)^2 - \tfrac{1}{2}\lambda^2[\hat{\Pi};[\Pi_0;\hat{\Pi}]] \\ & + \lambda^3[\hat{\Pi};[\Pi_0;(\hat{\Pi})^2]] - \lambda^3([\Pi_0;\hat{\Pi}])^2 \\ & - \lambda^2[\Pi_0;[\Pi_0;\hat{\Pi}]] \\ & + \lambda^3[\Pi_0;[\Pi_0;(\hat{\Pi})^2]] + O(\lambda^4). \end{aligned} \tag{2.12}$$

Note that all terms here are self-adjoint, in accord with the theorem just proven. The transformation P that accomplishes this change of the kinetic Hamiltonian is

$$\begin{aligned} P = \exp\{ & -\lambda\hat{\Pi} + \lambda^2\hat{\Pi}^2 + \tfrac{2}{3}\lambda^3\hat{\Pi}^3 - \lambda^3[\Pi_0;[\Pi_0;\hat{\Pi}]] \\ & - \lambda^3[\hat{\Pi};[\Pi_0;\hat{\Pi}]] + O(\lambda^4)\}. \end{aligned} \tag{2.13}$$

To within unitary equivalence, the above results complete the goal of finding an acceptable nonrelativistic representation of the second-order Dirac equation. However, as mentioned above, in Sec. III a certain asymmetry between the space Φ and the dual space $\overline{\Phi}$ will be exhibited. This asymmetry will be investigated in Sec. IV, where a unitary transformation to remove the asymmetry is obtained. A further simplification of the Hamiltonian results as a bonus for making this final transformation.

III. NONRELATIVISTIC REPRESENTATION FOR $\overline{\Phi}$

Consider now the effect of shifting the point of view to the dual space, the space of states $\overline{\Phi}$. The space $\overline{\Phi}$ is treated as primary and the space Φ now becomes the dual space. Then an alternate procedure for the nonrelativistic reduction of the second-order Dirac equation is suggested as fol-

lows. The nonrelativistic limit will now refer to states $\bar{\Phi}$ obeying an equation

$$\bar{\Phi}\overleftarrow{\Pi}_0 = \bar{\Phi}L, \tag{3.1}$$

of the nonrelativistic Schrödinger-type in addition to obeying the dual form

$$\bar{\Phi}(\overleftarrow{\Pi}_0 + \hat{\Pi})(\overleftarrow{\Pi}_0 - \hat{\Pi}) = \bar{\Phi}m^2 \tag{3.2}$$

of the second-order Dirac equation. Proceeding along lines entirely parallel to Sec. II, an equation[26]

$$L^2 - [\Pi_0;L + \hat{\Pi}] = m^2 + \hat{\Pi}^2, \tag{3.3}$$

for the calculation of L can be obtained. Setting $L = m + T$ and developing T in a series of ascending powers of $1/m$ by iteration of Eq. (3.3), leads to the following expression for L:

$$\begin{aligned} L = {} & (1/2\lambda) + \lambda\hat{\Pi}^2 + \lambda^2[\Pi_0;[\Pi_0;\hat{\Pi}]] - \lambda^3(\hat{\Pi}^2)^2 \\ & - \lambda^3([\Pi_0;\hat{\Pi}])^2 + \lambda^3[\Pi_0;[\Pi_0;\hat{\Pi}^2]] + \lambda[\Pi_0;\hat{\Pi}] \\ & + \lambda^2[\Pi_0;\hat{\Pi}^2] - \lambda^3(\hat{\Pi}^2[\Pi_0;\hat{\Pi}] + [\Pi_0;\hat{\Pi}]\hat{\Pi}^2) \\ & + \lambda^3[\Pi_0;[\Pi_0;[\Pi_0;\hat{\Pi}]]] + O(\lambda^4). \end{aligned} \tag{3.4}$$

Equation (3.3) admits a second series solution, obtained by making the replacement $\lambda \to -\lambda$. Again, the second solution describes the negative frequency states.

Before proceeding to round out the parallel with the earlier results for Φ, there is a question of consistency that must be addressed. Suppose that Φ is given by a simultaneous solution of Eqs. (2.1) and (1.1), and that $\bar{\Phi}$ is calculated by the prescription (1.2). A definite equation for $\bar{\Phi}$ may already be implied by these conditions. It must be checked whether that equation indeed has the requisite form (3.1). To investigate this the form $\bar{\Phi} \equiv \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})$ is substituted into Eq. (3.1) in order to determine the constraints on L,

$$\begin{aligned} & \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})\overleftarrow{\Pi}_0 = \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})L, \\ & \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})(\overleftarrow{\Pi}_0 + \hat{\Pi}) = \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})(L + \hat{\Pi}), \\ & \Phi^\dagger m^2 = \Phi^\dagger(K^\dagger - \hat{\Pi})(L + \hat{\Pi}). \end{aligned}$$

Since $\Phi^\dagger$ at one instant of time can be arbitrary, this last time implies the identity

$$m^2 = (K^\dagger - \hat{\Pi})(L + \hat{\Pi}). \tag{3.5}$$

Since the steps leading to Eq. (3.5) are reversible, this calculation answers the above question of consistency: the dual $\bar{\Phi}$ of a simultaneous solution Φ of Eqs. (2.1) and (1.1) obeys the nonrelativistic Schrödinger-type equation (3.1) with L given by Eq. (3.5). That Eq. (3.5) indeed gives an L that obeys the former equation (3.3) for L is shown by the following direct calculation:

$$\begin{aligned} & L + \hat{\Pi} = m^2/(K^\dagger - \hat{\Pi}), \\ & L^2 = [m^2/(K^\dagger - \hat{\Pi})]\{m^2 - K^\dagger\hat{\Pi} - \hat{\Pi}K^\dagger + 2\hat{\Pi}^2\} \\ & \qquad \times [1/(K^\dagger - \hat{\Pi})] + \hat{\Pi}^2, \\ & [\Pi_0;L + \hat{\Pi}] = -[m^2/(K^\dagger - \hat{\Pi})] \\ & \qquad \times [\Pi_0;K^\dagger - \hat{\Pi}][1/(K^\dagger - \hat{\Pi})] \\ & \qquad = -[m^2/(K^\dagger - \hat{\Pi})]\{(K^\dagger)^2 - m^2 - \hat{\Pi}^2\} \\ & \qquad \times [1/(K^\dagger - \hat{\Pi})]. \end{aligned}$$

This last step involves Eq. (2.5). Continuing, the expression $L^2 - [\Pi_0;L + \hat{\Pi}]$ occurring in Eq. (3.3) is formed:

$$\begin{aligned} & L^2 - [\Pi_0;L + \hat{\Pi}] \\ & \quad = [m^2/(K^\dagger - \hat{\Pi})](K^\dagger - \hat{\Pi})^2[1/(K^\dagger - \hat{\Pi})] + \hat{\Pi}^2, \\ & L^2 - [\Pi_0;L + \hat{\Pi}] = m^2 + \hat{\Pi}^2. \end{aligned}$$

Because of the multiple valuedness of the solutions of Eq. (3.3) it must be checked that the positive frequency expression (3.4) for L actually results from the substitution of the expression (2.6) obtained previously for K into the formula (3.5) connecting the two kinetic Hamiltonians. At the same time the L obtained from Eq. (3.5) and the K used as input data are found to both describe the same positive or negative frequency type of solutions.

Having established the consistency of the approach, the parallel between the space Φ and the dual space $\bar{\Phi}$ will be pursued further. The Dirac inner product can be expressed in terms of the dual states as follows[27]:

$$\begin{aligned} (\Phi_2;\Phi_1) &= \int d^3r \frac{1}{m^2}\bar{\Phi}_2(\vec{\Pi}_0 + \overleftarrow{\Pi}_0)\Phi_1 \\ &= \int d^3r \frac{1}{m^2}\bar{\Phi}_2(\vec{\Pi}_0 + \overleftarrow{\Pi}_0)(\vec{\Pi}_0 + \hat{\Pi})\frac{(\bar{\Phi}_1)^\dagger}{m^2} \\ &= \int d^3r \frac{1}{m^2}\bar{\Phi}_2(\overleftarrow{\Pi}_0 + \hat{\Pi} + (\vec{\Pi}_0 - \hat{\Pi})) \\ &\quad \times (\vec{\Pi}_0 + \hat{\Pi})\frac{(\bar{\Phi}_1)^\dagger}{m^2} \\ &= \int d^3r \frac{1}{m^2}\bar{\Phi}_2((\overleftarrow{\Pi}_0 + \hat{\Pi})(\vec{\Pi}_0 + \hat{\Pi}) + m^2) \\ &\quad \times \frac{(\bar{\Phi}_1)^\dagger}{m^2} \\ &= \int d^3r \frac{1}{m^2} 2\bar{\Phi}_2\left(\frac{(L + \hat{\Pi})(L^\dagger + \hat{\Pi})}{2m^2} + \frac{1}{2}\right) \\ &\quad \times (\bar{\Phi}_1)^\dagger. \end{aligned}$$

This result can be put in the form [analogous to Eq. (2.8)]

$$(\Phi_2;\Phi_1) = \int d^3r \frac{1}{m^2} 2\bar{\Phi}_2^T(\bar{\Phi}_1^T)^\dagger \tag{3.6}$$

appropriate for a nonrelativistic Schrödinger equation. The change of representation needed here is from $\bar{\Phi}$ to $\bar{\Phi}^T$, where

$$\bar{\Phi}^T \equiv \bar{\Phi}Q, \tag{3.7}$$

and

$$Q \equiv \left[\frac{1}{2} + \frac{(L + \hat{\Pi})(L^\dagger + \hat{\Pi})}{2m^2}\right]^{1/2}. \tag{3.8}$$

Again, the transformed Hamiltonian is self-adjoint. The appropriate equations this time [corresponding to Eqs. (2.10) and (2.12)] are

$$\begin{aligned} & L^T \equiv (Q^{-1}LQ - Q^{-1}[\Pi_0;Q]), \\ & \bar{\Phi}^T\Pi_0 = \bar{\Phi}^T L^T, \end{aligned} \tag{3.9}$$

and

$$L^T = (L^T)^\dagger = (1/2\lambda) + \lambda\hat{\Pi}^2 - \lambda^3(\hat{\Pi}^2)^2 - \tfrac{1}{2}\lambda^2[\hat{\Pi};[\Pi_0;\hat{\Pi}]] - \lambda^3[\hat{\Pi};[\Pi_0;(\hat{\Pi})^2]] - \lambda^3([\Pi_0;\hat{\Pi}])^2 + \lambda^2[\Pi_0;[\Pi_0;\hat{\Pi}]] + \lambda^3[\Pi_0;[\Pi_0;(\hat{\Pi})^2]] + O(\lambda^4). \tag{3.10}$$

The transformation Q that brings about this change in the kinetic Hamiltonian is

$$Q = \exp\{\lambda\hat{\Pi} + \lambda^2\hat{\Pi}^2 - \tfrac{2}{3}\lambda^3\hat{\Pi}^3 - \lambda^3[\hat{\Pi};[\Pi_0;\hat{\Pi}]] + \lambda^3[\hat{\Pi}_0;[\Pi_0;\hat{\Pi}]] + O(\lambda^4)\}. \tag{3.11}$$

IV. UNITARY EQUIVALENCE OF THE TWO NONRELATIVISTIC REPRESENTATIONS

The transformed kinetic Hamiltonian (3.10) obtained by working in the dual space differs somewhat from the transformed kinetic Hamiltonian (2.12) obtained originally for the space Φ. To investigate this the relationship between the states Φ^T and $\bar{\Phi}^T$ is studied. Thus

$$\bar{\Phi}^T = \bar{\Phi}Q = \Phi^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})Q = \Phi^\dagger(K^\dagger - \hat{\Pi})Q = (\Phi^T)^\dagger P^{-1}(K^\dagger - \hat{\Pi})Q,$$

or

$$\bar{\Phi}^T = m(\Phi^T)^\dagger U \tag{4.1}$$

where

$$U \equiv P^{-1}((K^\dagger - \hat{\Pi})/m)Q. \tag{4.2}$$

As the notation suggests, the operator U is unitary. This may be proven by investigating $UU^\dagger$ as follows:

$$UU^\dagger = P^{-1}\left(\frac{K^\dagger - \hat{\Pi}}{m}\right)Q^2\left(\frac{K - \hat{\Pi}}{m}\right)P^{-1} = P^{-1}\left(\frac{K^\dagger - \hat{\Pi}}{m}\right)\left(\frac{1}{2} + \frac{(L + \hat{\Pi})(L^\dagger + \hat{\Pi})}{2m^2}\right) \times\left(\frac{K - \hat{\Pi}}{m}\right)P^{-1} = P^{-1}\frac{1}{m^2}\left(\frac{(K^\dagger - \hat{\Pi})(K - \hat{\Pi})}{2} + \frac{(K^\dagger - \hat{\Pi})(L + \hat{\Pi})(L^\dagger + \hat{\Pi})(K - \hat{\Pi})}{2m^2}\right)P^{-1}$$

or, using Eq. (3.5),

$$UU^\dagger = P^{-1}\frac{1}{m^2}\left(\frac{(K^\dagger - \hat{\Pi})(K - \hat{\Pi})}{2} + \frac{(m^2)^2}{2m^2}\right)P^{-1} = P^{-1}P^2P^{-1} = 1.$$

The unitary transformation U provides the key to the relationship between the two nonrelativistic representations that have been found above in Secs. II and III. Equation (4.1) can be written in the form

$$\bar{\Phi}^T = m(\Phi^T)^\dagger\sqrt{U}\sqrt{U},$$

or

$$\bar{\Phi}^T\sqrt{U^\dagger} = m(\sqrt{U^\dagger}\Phi^T)^\dagger, \tag{4.3}$$

a form which treats the Hilbert space Φ and the dual space $\bar{\Phi}$ on equal footing. This form suggests further changes of representation according to

$$\Phi^{NR} \equiv \sqrt{U^\dagger}\Phi^T \tag{4.4}$$

and

$$\bar{\Phi}^{NR} \equiv \bar{\Phi}^T\sqrt{U^\dagger}. \tag{4.5}$$

In the new representation there is a simple relationship between the Hilbert space and the dual space. By Eq. (4.3) this relationship is

$$\bar{\Phi}^{NR} = m(\Phi^{NR})^\dagger. \tag{4.6}$$

Of course, since this final transformation is unitary, the kinetic Hamiltonians in the two representations remain self-adjoint and the nonrelativistic structure of the inner products (2.8) and (3.6) is preserved. It can be shown that in the final representation the two kinetic Hamiltonians are actually equal. The proof of this is similar to the above proof that K^T is self-adjoint:

$$\bar{\Phi}_2^{NR}\overleftarrow{\Pi}_0 = \bar{\Phi}_2^{NR}L^{NR},$$
$$m(\Phi_2^{NR})^\dagger\overleftarrow{\Pi}_0 = m(\Phi_2^{NR})^\dagger L^{NR},$$
$$(\Phi_2^{NR})^\dagger\overleftarrow{\Pi}_0\Phi_1^{NR} = (\Phi_2^{NR})^\dagger L^{NR}\Phi_1^{NR}.$$

A similar calculation starting with the equation $\vec{\Pi}_0\Phi_1^{NR} = K^{NR}\Phi_1^{NR}$ leads to

$$(\Phi_2^{NR})^\dagger\vec{\Pi}_0\Phi_1^{NR} = (\Phi_2^{NR})^\dagger K^{NR}\Phi_1^{NR}.$$

The two corresponding formulas are manipulated in a usual way to obtain the integral identity:

$$i\frac{\partial}{\partial t}\int d^3r(\Phi_2^{NR})^\dagger\Phi_1^{NR} = \int d^3r(\Phi_2^{NR})^\dagger(K^{NR} - L^{NR})\Phi_1^{NR}.$$

But since the last transformation was unitary, the time derivative on the left has the value

$$i\frac{\partial}{\partial t}\int d^3r(\Phi_2^{NR})^\dagger\Phi_1^{NR} = i\frac{\partial}{\partial t}\int d^3r(\Phi_2^T)^\dagger\Phi_1^T = i\frac{\partial}{\partial t}(\Phi_2;\Phi_1) = 0.$$

Accordingly, the identity

$$\int d^3r(\Phi_2^{NR})^\dagger(K^{NR} - L^{NR})\Phi_1^{NR} = 0$$

is obtained, which then implies the equality of the two kinetic Hamiltonians, since the states Φ_2^{NR} and Φ_1^{NR} can be arbitrarily prescribed at one moment of time:

$$K^{NR} = L^{NR}. \tag{4.7}$$

When K^{NR} is expanded in ascending powers of $1/m$ the following result is obtained:

$$K^{NR} = L^{NR} = (1/2\lambda) + \lambda\hat{\Pi}^2 - \lambda^3(\hat{\Pi}^2)^2 - \tfrac{1}{2}\lambda^2[\hat{\Pi};[\Pi_0;\hat{\Pi}]] + \lambda^3[\Pi_0;[\Pi_0;(\hat{\Pi})^2]] - \lambda^3([\Pi_0;\hat{\Pi}])^2 + O(\lambda^4). \tag{4.8}$$

The commutators in Eq. (4.8) can be worked out explicitly. For static field there results

$$K^{NR} = L^{NR} = (1/2\lambda) + \lambda\Pi^2(1-\lambda^2\Pi^2) - q\lambda\tfrac{1}{2}\vec{\sigma}\cdot\vec{B}(1-2\lambda^2\Pi^2) - q\lambda(1-2\lambda^2\Pi^2)\tfrac{1}{2}\vec{\sigma}\cdot\vec{B} - q^2\lambda^3(E^2+B^2) - \tfrac{1}{2}q\lambda^2\nabla\cdot\vec{E} - \tfrac{1}{2}q\lambda^2\vec{\sigma}\cdot(\vec{E}\times\vec{\Pi} - \vec{\Pi}\times\vec{E}). \quad (4.9)$$

The result (4.9) is the final nonrelativistic kinetic Hamiltonian obtained for a charged Fermion in a static electromagnetic field. Note that the potential energy term qV must be added to the expression (4.9) to give the full nonrelativistic Hamiltonian. As indicated in the Introduction, the representation (4.9) agrees through third order with the representations obtained by Pauli and Achiezer and Beresteckij, Foldy and Wouthuysen, and Eriksen, except for the one third-order term, $q^2\lambda^3E^2$. This difference goes away when the Hamiltonians are compared modulo unitary equivalence. The U that brings the kinetic Hamiltonian to its final form (4.9) is[28]:

$$U = \exp(-2\lambda^2[\Pi_0;\hat{\Pi}] + 4\lambda^4(\hat{\Pi}^2[\Pi_0;\hat{\Pi}] + [\Pi_0;\hat{\Pi}]\hat{\Pi}^2) - 2\lambda^4[\Pi_0;[\Pi_0;[\Pi_0;\hat{\Pi}]]] + O(\lambda^5)). \quad (4.10)$$

V. ANTIPARTICLE STATES

Entirely analogous results may be obtained for states Φ describing antiparticles. We shall seek such states as simultaneous solutions of the second-order Dirac equation and the nonrelativistic Schrödinger-type equation

$$\Pi_0\Phi = -L'\Phi \quad (5.1)$$

that are orthogonal to all states considered before. Using the subscripts $+/-$ to signal particle/antiparticle, the orthogonality condition can be explored as follows:

$$0 = (\Phi_+;\Phi_-) = \int d^3r\frac{1}{m^2}\bar{\Phi}_+(\overleftarrow{\Pi}_0 + \vec{\Pi}_0)\Phi_-$$

$$= \int d^3r\frac{1}{m^2}(\Phi_+)^\dagger(\overleftarrow{\Pi}_0 - \hat{\Pi})[(\overleftarrow{\Pi}_0 + \hat{\Pi}) + (\vec{\Pi}_0 - \hat{\Pi})]\Phi_-$$

$$= \int d^3r\frac{1}{m^2}(\Phi_+)^\dagger[m^2 + (\overleftarrow{\Pi}_0 - \hat{\Pi})(\vec{\Pi}_0 - \hat{\Pi})]\Phi_-$$

$$= \int d^3r\frac{1}{m^2}(\Phi_+)^\dagger[m^2 + (K^\dagger - \hat{\Pi})(-L' - \hat{\Pi})]\Phi_-.$$

Since Φ_+ and Φ_- at one moment of time can be artibrary functions, the square bracket in this last line must vanish:

$$m^2 = (K^\dagger - \hat{\Pi})(L' + \hat{\Pi}).$$

In view of Eq. (3.5) this shows that

$$L' = L. \quad (5.2)$$

An approach that just uses Eq. (5.1) to evaluate time derivatives in the second-order Dirac equation would lead to an equation for L' identical to Eq. (3.3), but with L' appearing instead of L. Although of course $L' = L$ is one solution of that equation, such an approach cannot determine L' unambiguously, because of the multiple valuedness of the solutions of Eq. (3.3). The above orthogonality considerations bypass this uniqueness question entirely.

A transformation law of the antiparticle states that will reduce the dot product of two antiparticle states to the nonrelativistic Schrödinger form and thereby render the transformed kinetic Hamiltonian self-adjoint can be discovered by investigating $(\Phi_2;\Phi_1)$. Using by now familiar methods we find the identity

$$(\Phi_2;\Phi_1) = \int d^3r\,2(\Phi_2)^\dagger\left\{\frac{1}{2} + \frac{(L^\dagger + \hat{\Pi})(L + \hat{\Pi})}{2m^2}\right\}\Phi_1. \quad (5.3)$$

The operator sandwiched between states here would be Q^2, but for the order of factors. It turns out that there is a trick for interchanging the order of these factors. We write $(K - \hat{\Pi})/m$ in polar form as follows:

$$\frac{K - \hat{\Pi}}{m} \equiv e^{iA}\left[\frac{(K^\dagger - \hat{\Pi})(K - \hat{\Pi})}{m^2}\right]^{1/2} \quad (5.4)$$

thereby defining a self-adjoint operator A. If Eq. (5.4) is multiplied on the right by the adjoint of itself the identity

$$(K - \hat{\Pi})(K^\dagger - \hat{\Pi}) = e^{iA}(K^\dagger - \hat{\Pi})(K - \hat{\Pi})e^{-iA} \quad (5.5)$$

emerges, an identity that can be used to interchange the order of factors in a product $(K - \hat{\Pi})(K^\dagger - \hat{\Pi})$. By virtue of the connection $(L + \hat{\Pi}) = m^2/(K^\dagger - \hat{\Pi})$ between L and K, it becomes possible to develop a relation analogous to Eq. (5.5) enabling the order of factors in a product $(L^\dagger + \hat{\Pi})(L + \hat{\Pi})$ to be reversed:

$$(L^\dagger + \hat{\Pi})(L + \hat{\Pi}) = \frac{m^2}{(K - \hat{\Pi})}\frac{m^2}{(K^\dagger - \hat{\Pi})} = \frac{(m^2)^2}{(K^\dagger - \hat{\Pi})(K - \hat{\Pi})} = \frac{(m^2)^2}{e^{-iA}(K - \hat{\Pi})(K^\dagger - \hat{\Pi})e^{iA}} = e^{-iA}\frac{(m^2)^2}{(K - \hat{\Pi})(K^\dagger - \hat{\Pi})}e^{iA} = e^{-iA}\frac{m^2}{(K^\dagger - \hat{\Pi})}\frac{m^2}{(K - \hat{\Pi})}e^{iA} = e^{-iA}(L + \hat{\Pi})(L^\dagger + \hat{\Pi})e^{iA}.$$

we have proved

$$(L^\dagger + \hat{\Pi})(L + \hat{\Pi}) = e^{-iA}(L + \hat{\Pi})(L^\dagger + \hat{\Pi})e^{iA}. \quad (5.6)$$

Now we can process Eq. (5.3)

$$(\Phi_2;\Phi_1) = \int d^3r\,2(\Phi_2)^\dagger\left\{\frac{1}{2} + \frac{e^{-iA}(L + \hat{\Pi})(L^\dagger + \hat{\Pi})e^{iA}}{2m^2}\right\}\Phi_1 = \int d^3r\,2(\Phi_2)^\dagger e^{-iA}\left\{\frac{1}{2}\right.$$

$$+\frac{(L+\hat{\Pi})(L^{\dagger}+\hat{\Pi})}{2m^2}\Bigg\} e^{iA}\Phi_1$$

$$=\int d^3r\, 2(e^{iA}\Phi_2)^{\dagger}Q^2e^{iA}\Phi_1.$$

The desired effect is obtained by a change of representation from Φ to Φ^T, where

$$\Phi^T\equiv Qe^{iA}\Phi. \tag{5.7}$$

Next we investigate the dual $\overline{\Phi}$ of an antiparticle state. Results for $\overline{\Phi}$ whose derivation involves by now familiar methods will be summarized briefly. The state $\overline{\Phi}$ is required to satisfy simultaneously the second-order Dirac equation and a nonrelativistic Schrödinger-type equation as follows:

$$\overline{\Phi}\Pi_0=-\hat{\Pi}K'. \tag{5.8}$$

The relation

$$K'=K \tag{5.9}$$

can be derived by again exploiting the orthogonality between particle and antiparticle states. Also, it can be shown that the change of representation from $\overline{\Phi}$ to $\overline{\Phi}^T$, where

$$\overline{\Phi}^T\equiv\overline{\Phi}e^{iA}P, \tag{5.10}$$

will render the inner product to two antiparticle states in the form

$$(\Phi_2;\Phi_1)=\int d^3r\frac{2}{m^2}\overline{\Phi}_2(\overline{\Phi}_1)^{\dagger}$$

appropriate for the nonrelativistic Schrödinger-type wave equation (5.8) for the antiparticles, at the same time rendering the transformed kinetic Hamiltonian of the antiparticles self-adjoint.

The transformed Hamiltonians for antiparticles corresponding to Eqs. (5.1) and (5.8) will in general not be equal. This asymmetry will now be removed using the approach of Sec. IV. We start by looking at the relationship between $\overline{\Phi}^T$ and Φ^T,

$$\overline{\Phi}^T\equiv\overline{\Phi}e^{iA}P=\Phi^{\dagger}(\overleftarrow{\Pi}_0-\hat{\Pi})e^{iA}P$$

$$=\Phi^{\dagger}(-L^{\dagger}-\hat{\Pi})e^{iA}P=(\Phi^T)^{\dagger}Q^{-1}e^{iA}$$

$$\times(-L^{\dagger}-\hat{\Pi})e^{iA}P$$

$$=-(\Phi^T)^{\dagger}Q^{-1}e^{iA}\frac{m^2}{K-\hat{\Pi}}e^{iA}P$$

$$=-(\Phi^T)^{\dagger}Q^{-1}e^{iA}\frac{m^2}{e^{iA}\sqrt{(K^{\dagger}-\hat{\Pi})(K-\hat{\Pi})}}e^{iA}P.$$

So far we have the identity:

$$\overline{\Phi}^T=-(\Phi^T)^{\dagger}Q^{-1}e^{iA}\frac{m^2}{\sqrt{(K^{\dagger}-\hat{\Pi})(K-\hat{\Pi})}}P. \tag{5.11}$$

The calculation can be completeted by noting the relation[29]:

$$e^{-iA}Qe^{iA}=\frac{mP}{\sqrt{(K^{\dagger}-\hat{\Pi})(K-\hat{\Pi})}}, \tag{5.12}$$

so that Eq. (5.11) becomes $\overline{\Phi}^T=-m(\Phi^T)^{\dagger}e^{iA}$, or

$$\overline{\Phi}^Te^{-iA/2}=-m(e^{-iA/2}\Phi^T)^{\dagger}, \tag{5.13}$$

a form that treats the space Φ and the dual space $\overline{\Phi}$ on equal footing. Equation (5.13) suggest the final change of representations

$$\Phi^{\rm NR}\equiv e^{-iA/2}\Phi^T=e^{-iA/2}Qe^{iA}\Phi, \tag{5.14}$$

and

$$\overline{\Phi}^{\rm NR}\equiv\overline{\Phi}^Te^{-iA/2}=\overline{\Phi}e^{iA}Pe^{-iA/2}, \tag{5.15}$$

with respect to which Eq. (5.13) takes the form

$$\overline{\Phi}^{\rm NR}\equiv-m(\Phi^{\rm NR})^{\dagger} \tag{5.16}$$

analogous to Eq. (4.6). By use of Eq. (5.16) it is possible to show along the lines of Sec. IV that the final transformed antiparticle Hamiltonians corresponding to Eqs. (5.1) and (5.8) agree and remain self-adjoint.

VI. CONNECTION WITH THE LINEAR DIRAC EQUATION

The above investigations yield particle Φ_+ or antiparticle Φ_- solutions of the second-order Dirac equation according as the wave equation (2.1) or (5.1) is taken to be obeyed simultaneously with the second-order Dirac equation. From a physical standpoint it is clear that the most general solution Φ of the second-order Dirac equation (1.1) can be expressed as a linear combination of the particle and antiparticle parts: $\Phi=\Phi_++\Phi_-$. The corresponding Dirac wave function is then[30]

$$\Psi\equiv\begin{bmatrix}\Phi_++\Phi_-\\(\overline{\Phi}_+)^{\dagger}/m+(\overline{\Phi}_-)^{\dagger}/m\end{bmatrix}, \tag{6.1}$$

in which so far all states refer to the original representation. Now let all be expressed in terms of the nonrelativistic representation. Then the above expression for Ψ can be reduced as follows:

$$\Psi\equiv\begin{bmatrix}P^{-1}\sqrt{U}(\Phi_+)^{\rm NR}+e^{-iA}Q^{-1}e^{iA/2}(\Phi_-)^{\rm NR}\\((\overline{\Phi}_+)^{\rm NR}\sqrt{U}Q^{-1}/m)^{\dagger}+((\overline{\Phi}_-)^{\rm NR}e^{iA/2}P^{-1}e^{-iA}/m)^{\dagger}\end{bmatrix}$$

$$\Psi\equiv\begin{bmatrix}P^{-1}\sqrt{U}(\Phi_+)^{\rm NR}+e^{-iA}Q^{-1}e^{iA/2}(\Phi_-)^{\rm NR}\\Q^{-1}\sqrt{U^{\dagger}}\Phi_+^{\rm NR}-e^{iA}P^{-1}e^{-iA/2}\Phi_-^{\rm NR}\end{bmatrix}.$$

These expressions incorporate Eqs. (2.9), (4.4), (4.5), (4.6), (3.7), (5.14), (5.15), and (5.16). The result for Ψ can be written in the form

$$\Psi=V\Psi^{\rm NR}, \tag{6.2}$$

where

$$V=\begin{bmatrix}(\sqrt{2}P)^{-1}e^{-iA/2} & e^{-iA}(\sqrt{2}Q)^{-1}e^{iA/2}\\(\sqrt{2}Q)^{-1}e^{iA/2} & -e^{iA}(\sqrt{2}P)^{-1}e^{-iA/2}\end{bmatrix}, \tag{6.3}$$

an equation that incorporates the relation[31]

$$U=e^{-iA}. \tag{6.4}$$

The spinor $\Psi^{\rm NR}$ is defined by

$$\Psi^{\rm NR}\equiv\begin{bmatrix}\sqrt{2}(\Phi_+)^{\rm NR}\\\sqrt{2}(\Phi_-)^{\rm NR}\end{bmatrix}. \tag{6.5}$$

It can be shown that the operator V is unitary. One way of doing this is to first write P and Q in Eq. (6.3) in terms of an operator θ defined as follows:

$$\tan(\theta)\equiv\sqrt{(K^{\dagger}-\hat{\Pi})(K-\hat{\Pi})/m^2}. \tag{6.6}$$

Equation (2.7) for P gives

$$P = 1/\sqrt{2}\cos(\theta), \tag{6.7}$$

and from Eq. (5.12) we get

$$Q = e^{iA}\left[1/\sqrt{2}\sin(\theta)\right]e^{-iA}. \tag{6.8}$$

Now the Dirac matrix V can be written,

$$V = \begin{bmatrix} \cos(\theta)e^{-iA/2} & \sin(\theta)e^{-iA/2} \\ e^{iA}\sin(\theta)e^{-iA/2} & -e^{iA}\cos(\theta)e^{-iA/2} \end{bmatrix}. \tag{6.9}$$

Having achieved the form (6.9), it is quite easy to verify the unitarity of V. The factors in Eq. (6.9) are noncommutative, but occur in just the right order to permit a trouble-free calculation when the relation $VV^\dagger = 1$ is tested. Accordingly, the transformation to Φ_+^{NR} and Φ_-^{NR} correspond in the space of the linear Dirac equation to a unitary transformation to a nonrelativistic representation of that equation.

From the defining equation (6.6) it follows that an expansion of θ in ascending powers of $1/m$ begins with the term $\theta = \pi/4$. If θ is written

$$\theta \equiv \pi/4 - \xi/2, \tag{6.10}$$

then ξ will be an infinitesimal in the parameter $1/m$. In terms of ξ Eq. (6.9) reads

$$V = \begin{bmatrix} \cos(\xi/2)e^{-iA/2} & \sin(\xi/2)e^{-iA/2} \\ -e^{iA}\sin(\xi/2)e^{-iA/2} & e^{iA}\cos(\xi/2)e^{-iA/2} \end{bmatrix} T,$$

$$T \equiv \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}. \tag{6.11}$$

The matrix T is the matrix that takes Ψ from the present representation of the linear Dirac equation to the standard one[32]:

$$\Psi_{STD} = T\Psi. \tag{6.12}$$

The factor T in Eq. (6.11) can be eliminated by converting all to the standard representation: $\Psi^{STD} = TV\Psi^{NR} = V^{STD}\Psi^{NR}$, where $V^{STD} = TV$. In terms of

$$\hat{\xi} \equiv e^{iA/2}\xi e^{-iA/2}, \tag{6.13}$$

there results

$$V^{STD} = T\begin{bmatrix} e^{-iA/2} & 0 \\ 0 & e^{iA/2} \end{bmatrix}\begin{bmatrix} \cos(\hat{\xi}/2) & \sin(\hat{\xi}/2) \\ -\sin(\hat{\xi}/2) & \cos(\hat{\xi}/2) \end{bmatrix}T,$$

or, in canonical form,

$$V^{STD} = \exp\left(\begin{bmatrix} 0 & -iA/2 \\ -iA/2 & 0 \end{bmatrix}\right)\exp\left(\begin{bmatrix} 0 & -\hat{\xi}/2 \\ \hat{\xi}/2 & 0 \end{bmatrix}\right). \tag{6.14}$$

The transformation V^{STD} corresponds in the space of the ordinary linear Dirac equation to exactly two consecutive canonical transformations that are both "odd" in the terminology of Foldy and Wouthuysen. It has been verified explicitly through terms $O(\lambda^4)$ that the transformation V^{STD} transforms the Dirac equation into an "even" form with a kinetic Hamiltonian given exactly by the expression (4.8) but with λ replaced by $\beta_{STD}\lambda$. For this calculation A is provided by Eqs. (4.10) and (6.4) and $\hat{\xi}/2$ was found to have the value

$$\frac{\hat{\xi}}{2} = \lambda\hat{\Pi} - \tfrac{4}{3}\lambda^3\hat{\Pi}^3 + \lambda^3[\Pi_0;[\Pi_0;\hat{\Pi}]] + O(\lambda^5). \tag{6.15}$$

The following formula provided a convenient means of calculating ξ:

$$\xi = \sin^{-1}\left(\frac{1-(K^\dagger-\hat{\Pi})(K-\hat{\Pi})/m^2}{1+(K^\dagger-\hat{\Pi})(K-\hat{\Pi})/m^2}\right). \tag{6.16}$$

ACKNOWLEDGMENT

This research was supported by the National Science Foundation under Grant No. PHY-8603769.

APPENDIX A: SOLUTION BY SERIES

In the following a direct series approach to the above results will be described. The approach is entirely equivalent to the earlier method and, as noted before, provides a more direct parallel to the original Foldy–Wouthuysen method and to the elimination method of Pauli *et al.* The procedure begins by writing in Eq. (1.1) $\Phi \equiv e^{-imt}\chi$, leading to the equation

$$\Pi_0\chi = \lambda\hat{\Pi}^2\chi + \lambda[\Pi_0;\hat{\Pi}]\chi - \lambda(\Pi_0)^2\chi. \tag{A1}$$

Note that in Eq. (A1) the last term on the right-hand side is the only term that spoils the desired eventual form $\Pi_0\chi = K\chi$ of the nonrelativistic equation of motion: this is because the term involves time derivatives of χ. Iteration is used to push the terms on the right-hand side involving time derivatives of χ out to higher and higher orders in the parameter $1/m$. As mentioned in the Introduction, the technique for doing this has a parallel in the elimination method of Pauli *et al.*, who work with a corresponding stationary state equation. Their method uses iteration to push unwanted terms involving the eigenvalue out to higher and higher orders.

The right-hand factor Π_0 of the $-\lambda(\Pi_0)^2\chi$ term on the right-hand side of Eq. (A1) is "evaluated" by applying Eq. (A1) itself. It is true that in this iteration still higher powers of Π_0 are introduced, but it turns out that such additional terms are of higher order in the parameter $1/m$, and eventually after sufficiently many iterations, can be neglected to any desired degree of approximation.

To illustrate the technique the first iteration is carried out explicitly:

$$\Pi_0\chi = \lambda\hat{\Pi}^2\chi + \lambda[\Pi_0;\hat{\Pi}]\chi - \lambda^2\Pi_0\hat{\Pi}^2\chi - \lambda^2\Pi_0[\Pi_0;\hat{\Pi}]\chi + \lambda^2(\Pi_0)^3\chi.$$

The Π_0's on the right-hand side that occur inside commutators, as in the second λ term and the second λ^2 term, contain no time derivatives that can act on χ and are left alone. Factors of Π_0 not so "absorbed" in commutators are the only factors evaluated by iteration. However, if such a factor of Π_0 does not stand exactly to the left of χ, it must be commuted to that position before it can be evaluated by iteration. The additional commutators thereby produced are free of time derivatives that can act on χ, and just contribute additional terms to the eventual Hamiltonian,

$$\Pi_0\chi = \lambda\hat{\Pi}^2\chi + \lambda[\Pi_0;\hat{\Pi}]\chi - \lambda^2[\Pi_0;\hat{\Pi}^2]\chi - \lambda^2\hat{\Pi}^2\Pi_0\chi - \lambda^2[\Pi_0[\Pi_0;\hat{\Pi}]]\chi - \lambda^2[\Pi_0;\hat{\Pi}]\Pi_0\chi + \lambda^2(\Pi_0)^3\chi.$$

At this point the expression is poised for another iteration in

which, again, all factors of Π_0 standing immediately to the left of χ are evaluated using the equation itself. It is clear that this procedure must lead to the same kinetic Hamiltonian as Eq. (2.5), since in both cases the only input information is the same second-order Dirac Equation and the equation $\Pi_0\chi = K\chi$ which will eventually be satisfied here to arbitrary accuracy.

Once K has been obtained to the desired degree of accuracy, the next step is to remove the non-self-adjoint terms from K. For this purpose a transformed wave function $\chi^T \equiv e^A\chi$ is introduced, leading to a new wave equation in the form $e^A\Pi_0 e^{-A}\chi^T = e^A K e^{-A}\chi^T$. The wave equation in the new representation can be obtained with the help of the identity

$$e^A Q e^{-A} = Q + [A;Q] + (1/2!)[A;[A;Q]] + (1/3!)[A;[A;[A;Q]]] + \cdots . \quad \text{(A2)}$$

Successive transformations are carried out in which the non-self-adjoint terms of the kinetic Hamiltonian are removed order by order, starting with those of lowest order in $1/m$. It has been found by experience that at each step the non-self-adjoint terms of lowest order have a $[\Pi_0; \cdots]$ structure. Such terms can be removed by choosing A so that the term $[A;\Pi_0]$ on the left-hand side of the new wave equation

$$e^A\Pi_0 e^{-A}\chi^T = e^A K e^{-A}\chi^T$$

balances out the unwanted $[\Pi_0; \cdots]$ term contained in K itself on the right-hand side. For reference, the three A's needed to remove in succession non-self-adjoint terms of orders λ, λ^2, and λ^3 are

$$\begin{aligned} A_1 &= -\lambda\hat{\Pi}, \\ A_2 &= \lambda^2\hat{\Pi}^2, \\ A_3 &= \lambda^3(\tfrac{2}{3}\hat{\Pi}^3 - [\Pi_0;[\Pi_0;\hat{\Pi}]] - [\hat{\Pi};[\Pi_0;\hat{\Pi}]]). \end{aligned} \quad \text{(A3)}$$

The resultant of the three transformations is equivalent to a single transformation by e^A, where

$$e^A = e^{A_3}e^{A_2}e^{A_1}. \quad \text{(A4)}$$

When the three exponentials are combined,[33] exactly the expression (2.13) obtained before for P results. Analogous results are obtained for L, the kinetic Hamiltonian in the dual space.

Next, terms in L^T and K^T that differ are removed order by order. Again, experience has shown that at each step the lowest order terms in L^T and K^T that differ have the $[\Pi_0; \cdots]$ type of structure, and accordingly are removable. Since self-adjoint terms are being removed, the transformations involved at this stage are unitary. Accurate to and including third order in $1/m$, only one such unitary transformation is needed to remove the terms that differ in the two Hamiltonians. That transformation is found to be in agreement with Eq. (4.10).

APPENDIX B: SPHERICALLY SYMMETRIC POTENTIALS

It has been noticed that in the case of a spherically symmetric potential, two exact solutions of Eq. (2.5) for the kinetic Hamiltonian can be found. It is puzzling that these "extra" solutions do not provide examples illustrating the nonrelativistic reduction in the sense of the above formalism. However, they do provide examples illustrating the factorization theorem for the second-order Dirac equation that was mentioned in the Introduction. Accordingly, these extra solutions may be used to illustrate most of our equations.

The additional solutions in question were suggested by the work of Biedenharn and Horwitz[34,35] and are

$$K = \pm m\eta_3 + \vec{\sigma}\cdot\vec{p}, \qquad \eta_3 \equiv \Lambda/|\Lambda|, \quad \Lambda \equiv (\vec{\sigma}\cdot\vec{L} + 1). \quad \text{(B1)}$$

The operator η_3 is essentially the η_3 of Biedenharn and Horwitz. It anticommutes with $\vec{\sigma}\cdot\vec{p}$ and has eigenvalues ± 1, depending on the orbital and total angular momentum quantum numbers.

With the help of this K it is possible to illustrate most of the above equations. For example, the relations $L = \pm m\eta_3 - \vec{\sigma}\cdot\vec{p}$, $P = Q = 1$, and $U = \eta_3 = e^{i\pi(\eta_3 - 1)/2}$ can be derived, corresponding to Eqs. (3.5), (2.7), (3.8), and (4.2), respectively. When

$$K^{NR} = \sqrt{U^\dagger}K\sqrt{U} + [\Pi_0;\sqrt{U^\dagger}]\sqrt{U}$$

is computed the expression

$$K^{NR} = \eta_3(\pm m - i\hat{p}) \quad \text{(B2)}$$

results. In an effort to compare this with the kinetic Hamiltonian of Eq. (4.8), Eq. (B2) is rewritten[36]

$$K^{NR} = \epsilon(\eta_3(\pm m - i\hat{p}))\sqrt{m^2 + p^2}, \quad \text{(B3)}$$

in which $\epsilon(A)$ is the sign function: $\epsilon(A) \equiv \pm 1$ according as the eigenvalue of A is positive or negative. The kinetic Hamiltonian (B3) is thus a type of square root Klein–Gordon operator. As such, Eq. (B3) might appear at first to qualify as a nonrelativistic representation of the Dirac particle. However, the square root is taken sometimes with the plus sign and sometimes with the minus sign, according to the subspace of Hilbert space in which it acts. This behavior is at variance with physical expectations for a proper nonrelativistic reduction.

[1] O. Laporte and G. E. Uhlenbeck, Phys. Rev. **37**, 1380 (1931).
[2] R. P. Feynman and M. Gell-Mann, Phys. Rev. **109**, 193 (1958).
[3] L. M. Brown, Phys. Rev. **111**, 957 (1958).
[4] M. Tonin, Nuovo Cimento **14**, 1108 (1959).
[5] W. R. Theis, Fortschr. Phys. **7**, 559 (1959).
[6] H. Pietschmann, Acta Phys. Austriaca **14**, 63 (1961).
[7] L. M. Brown, "Two-component fermion theory," in *Lectures in Theoretical Physics* (Interscience, New York, 1962), Vol. IV.
[8] L. M. Brown, "Quantum electrodynamics at high energy," in *Topics in Theoretical Physics*, Proceedings of the Liperi Summer School in Theoretical Physics, 1967, edited by C. Cronstrom (Gordon and Breach, New York, 1969), p. 113.
[9] L. C. Hostler, J. Math. Phys. **23**, 1179 (1982).
[10] L. C. Hostler, J. Math. Phys. **24**, 2366 (1983).
[11] L. C. Biedenharn and L. P. Horwitz, Found. Phys. **14**, 953 (1984).
[12] L. C. Hostler, J. Math. Phys. **26**, 1348 (1985); **27**, 2208 (E) (1986).
[13] Levere C. Hostler, J. Math. Phys. **27**, 2423 (1986).

[14] The notation $i(\overleftarrow{\partial}/\partial t)$ shall mean minus differentiation of all standing to the left. An example of the use of this notation is $\Phi^\dagger\overleftarrow{\Pi}_0 = \Phi^\dagger(i(\overleftarrow{\partial}/\partial t) - qV) = -i(\partial\Phi^\dagger/\partial t) - \Phi^\dagger qV$. A related notation is $\overleftrightarrow{\Pi}_0 \equiv \overleftarrow{\Pi}_0 + \overrightarrow{\Pi}_0$. For simplicity the arrows are not written explicitly over space derivatives, where their direction is clear from the context. Also, right arrows over Π_0 are not normally indicated explicitly: all Π_0's are

understood to act to the right, except where explicitly indicated to the contrary.

[15]A representation in which the Dirac matrices are $\vec{\alpha} \equiv \left[\begin{smallmatrix} 0 & \vec{\sigma} \\ \vec{\sigma} & 0 \end{smallmatrix}\right], \beta \equiv \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ is assumed. In this representation the connection between the wave functions of the second-order Dirac equation and the linear Dirac equation is particularly simple, $\Psi = \left[\begin{smallmatrix} \Phi \\ \dot{\Phi}/m \end{smallmatrix}\right]$. The equivalence of the two inner products, (1.3) and the familiar Dirac inner product, follows from this equation for Ψ upon taking into account the second-order Dirac equation.

[16]In order to evaluate the inner product (1.3) of Φ into itself, the values of $\Phi, \dot{\Phi}$, and $\ddot{\Phi}$ at one moment of time are needed! Only when these quantities are chosen in a way that respects the second-order Dirac equation will positive definiteness be obtained. In the course of the paper, the wave functions are transformed a number of times, and sometimes by a nonunitary transformation. It is a part of the strategy of our approach that each inner product in whatever representation is kept actually equal in value to the inner product of the corresponding 4×1 Dirac spinors. Accordingly, all inner products retain the original physical interpretation of the Dirac inner product. It will be noted that a number of familiar operators, for example the linear momentum operator, lack self-adjointness with respect to the inner product (1.3). Because of this lack of self-adjointness, the second-order Dirac equation may not be well suited for all types of problems. It has been the authors's experience that Feynman graph calculations in quantum electrodynamics provide a class of problems for which the second-order Dirac equation is especially well suited.

[17]L. C. Hostler, J. Math. Phys. **28**, 720 (1987); **29**, 1930 (1988).

[18]L. L. Foldy and S. A. Wouthuysen, Phys. Rev. **78**, 29 (1950).

[19]E. Eriksen, Phys. Rev. **111**, 1011 (1958).

[20]Various treatments of the Foldy–Wouthuysen transformation are compared in the review article by E. de Vries, Fortschr. Phys. **18**, 149 (1970).

[21]W. Pauli, Z. Phys. **43**, 601 (1927).

[22]A. I. Achiezer and W. B. Beresteckij, *Quantum Electrodynamics* (Interscience, New York, 1965).

[23]E. de Vries, See Ref. 20.

[24]The question of convergence of the Foldy–Wouthuysen series is a very complicated matter. The Foldy–Wouthuysen series for the Coulomb potential has been shown to diverge [L. C. Biedenharn and L. P. Horwitz (Ref. 11)]. In general it may well be that the Foldy–Wouthuysen series is at best an asymptotic expansion. Some convergence questions are considered by E. Eriksen (Ref. 19). A question of unitarity of the Foldy–Wouthuysen transformation is considered by J. Kupersztych [Phys. Rev. Lett. **42**, 483 (1979)] and the problem of stability of numerical calculations involving the Foldy–Wouthuysen transformation is considered by H. Wallmeier and W. Kutzelnigg [Phys. Rev. A **28**, 3092 (1983)].

[25]Note that the operator under the radical is self-adjoint and positive definite. Accordingly, a unique square root operator can be defined through the use of the spectral theorem.

[26]The argument here parallels exactly that leading up to Eq. (2.5). We specify $\bar{\Phi}$ and $\bar{\Phi}\overleftarrow{\Pi}_0$ as initial data for Eq. (3.2) and we restrict consideration to the linear manifold N of solutions $\bar{\Phi}$ that evolve out of initial conditions of the special form $(\bar{\Phi}, \bar{\Phi}\overleftarrow{\Pi}_0 \equiv \bar{\Phi}L)_i$, where $\bar{\Phi}_i$ is arbitrary. For Eq. (3.1) the function $\bar{\Phi}_i$ is already a complete set of initial data and we have a relation $\bar{\Phi} = \bar{\Phi}_i W$, where W is the time evolution operator of Eq. (3.1). The equation corresponding to Eq. (2.3) is here $\bar{\Phi}_i W(L^2 - [\Pi_0; L + \hat{\Pi}]) = \bar{\Phi}_i W(m^2 + \hat{\Pi}^2)$ from which the desired conclusion follows, since $\bar{\Phi}_i$ can be arbitrary, and W is assumed to be nonsingular.

[27]The two identities $\Phi = (\Pi_0 + \hat{\Pi})\bar{\Phi}^\dagger/m^2$ and $(\Pi_0 - \hat{\Pi})\ (\Pi_0 + \hat{\Pi})\bar{\Phi}^\dagger = m^2\bar{\Phi}^\dagger$ are required in this calculation. The second identity is just the adjoint of Eq. (3.2), the equation of the dual state. The first equation follows from the second-order Dirac equation (1.1), on noting that, by Eq. (1.2), $(\Pi_0 - \hat{\Pi})\Phi = \bar{\Phi}^\dagger$.

[28]Only the first $O(\lambda^2)$ term is actually needed here. For reference the expression for U is provided accurate through $O(\lambda^4)$.

[29]Proof:

$$
\begin{aligned}
e^{-iA}Qe^{iA} &= e^{-iA}\left[\frac{1}{2} + \frac{(L+\hat{\Pi})(L^\dagger+\hat{\Pi})}{2m^2}\right]^{1/2} e^{iA} \\
&= e^{-iA}\left[\frac{1}{2} + \frac{1}{2m^2}\frac{m^2}{(K^\dagger-\hat{\Pi})}\frac{m^2}{(K-\hat{\Pi})}\right]^{1/2} e^{iA} \\
&= e^{-iA}\left[\frac{1}{2} + \frac{1}{2m^2}\frac{(m^2)^2}{(K-\hat{\Pi})(K^\dagger-\hat{\Pi})}\right]^{1/2} e^{iA} \\
&= \left[\frac{1}{2} + \frac{1}{2m^2}\frac{(m^2)^2}{(K^\dagger-\hat{\Pi})(K-\hat{\Pi})}\right]^{1/2} \\
&= \left[\frac{1}{2}\frac{(K^\dagger-\hat{\Pi})(K-\hat{\Pi}) + m^2}{(K^\dagger-\hat{\Pi})(K-\hat{\Pi})}\right]^{1/2} \\
&= \frac{mP}{\sqrt{(K^\dagger-\hat{\Pi})(K-\hat{\Pi})}}.
\end{aligned}
$$

[30]Recall the connection $\Psi = \left[\begin{smallmatrix} \Phi \\ \dot{\Phi}/m \end{smallmatrix}\right]$ between the Dirac wave function Ψ and the wave function Φ of the second-order Dirac equation (see Ref. 15 above).

[31]The following proof starts with the defining equation (4.2) of U in line 1, then employs Eq. (5.4) to yield line 2, and uses Eq. (5.12) to yield line 3:

$$
\begin{aligned}
U &= P^{-1}\left(\frac{K^\dagger - \hat{\Pi}}{m}\right)Q \\
&= P^{-1}\left[\frac{(K^\dagger-\hat{\Pi})(K-\hat{\Pi})}{m^2}\right]^{1/2} e^{-iA}Qe^{iA}e^{-iA} \\
&= P^{-1}\left[\frac{(K^\dagger-\hat{\Pi})(K-\hat{\Pi})}{m^2}\right]^{1/2} \\
&\quad \times \frac{mP}{\sqrt{(K^\dagger-\hat{\Pi})(K-\hat{\Pi})}} e^{-iA} \\
&= e^{-iA}.
\end{aligned}
$$

[32]The "standard representation" is defined by the values $(\vec{\alpha})_{\mathrm{STD}} = \left[\begin{smallmatrix} 0 & \vec{\sigma} \\ \vec{\sigma} & 0 \end{smallmatrix}\right]$, and $\beta_{\mathrm{STD}} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ of the Dirac matrices.

[33]The Baker–Campbell–Hausdorff theorem is needed in general to combine the exponentials, since noncommuting quantities are involved. This theorem states that:

$$e^A e^B = \exp(A + B + \tfrac{1}{2}[A;B] + \tfrac{1}{12}[A-B;[A;B]]\cdots)$$

see for example E. Eriksen, J. Math. Phys. **9**, 790 (1968), and references cited therein.

[34]See the reference cited above.

[35]The Hamiltonian of Biedenharn and Horwitz was devised in order to provide a chiral factorization of Kramers equation. Here their technique provides a factorization of the second-order Dirac equation. Thus

$$
\begin{aligned}
&(\eta_3\Pi_0 - \eta_3\hat{p} + m)(\eta_3\Pi_0 - \eta_3\hat{p} - m) \\
&\quad = (\eta_3\Pi_0 - \eta_3\hat{p})^2 - m^2 \\
&\quad = (\Pi_0)^2 - \Pi_0\hat{p} - \eta_3\hat{p}\eta_3\Pi_0 + \eta_3\hat{p}\eta_3\hat{p} - m^2 \\
&\quad = (\Pi_0)^2 - \Pi_0\hat{p} + \hat{p}\Pi_0 - \hat{p}^2 - m^2
\end{aligned}
$$

leading to the factorization

$$
\begin{aligned}
&(\eta_3\Pi_0 - \eta_3\hat{p} + m)(\eta_3\Pi_0 - \eta_3\hat{p} - m) \\
&\quad = (\Pi_0 + \hat{p})(\Pi_0 - \hat{p}) - m^2.
\end{aligned}
$$

Note that this factorization obtains for any spherically symmetric potential, and not just for the Coulomb case considered by Biedenharn and Horwitz.

[36]This equation is to be interpreted in the light of the spectral theorem, which allows one to write for any self-adjoint operator $f(A) = \Sigma_{A'} f(A')|A'\rangle\langle A'|$.

Generalized coherent states for a three-dimensional relativistic model of the oscillator

N. M. Atakishiyev[a)]
Instituto de Investigaciones en Matemáticas Aplicadas y en Systemas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

R. M. Mir-Kasimov
Joint Institute for Nuclear Research, Dubna 141980, USSR

(Received 13 April 1988; accepted for publication 25 January 1989)

The SU(1,1) coherent states for a relativistic model of the oscillator in the configurational **r** representation are considered. Classical equations of motion in the generalized phase space are obtained with the help of the path integral for the transition amplitude between SU(1,1) coherent states. It is shown that the use of the semiclassical Bohr–Sommerfeld quantization rule yields the exact expression for the energy levels.

I. INTRODUCTION

In the present work the generalized coherent states (CS) for an isotropic oscillator in the relativistic configurational representation are considered. The concept of relativistic configurational representation or **r** space is introduced in the following way.[1] The mass shell $p_0^2 - \mathbf{p}^2 = m^2c^2$ of a particle of mass m from the geometrical point of view realizes the Lobachevsky space, whose group of motions is the Lorentz group SO(3,1). The eigenfunctions of the corresponding Casimir operator (or the Laplace–Beltrami operator) in this space,

$$\langle \mathbf{r}|\mathbf{p}\rangle = [[p_0 - (\mathbf{pn})]/mc]^{-i(mc/\hbar)r-1},$$

$$\mathbf{r} = r\mathbf{n}, \quad 0 \leqslant r < \infty, \quad \mathbf{n}^2 = 1,$$

are the generating functions for matrix elements of the principal series for unitary irreducible representations of the SO(3,1) group and form a complete orthogonal system of functions in the momentum Lobachevsky space. The formal apparatus of the Fourier transformation over these functions[2–4] is used to pass to the relativistic **r** space introduced as a three-dimensional set of variables $\mathbf{r} = r\mathbf{n}$, where r is an eigenvalue of the Casimir operator. In the nonrelativistic limit (i.e., when $r \gg \hbar/mc$ and $|\mathbf{p}| \ll mc$) we come to the usual three-dimensional configurational space and the relativistic "plane wave" $\langle \mathbf{r}|\mathbf{p}\rangle$ goes into the Euclidean plane wave $\exp((i/\hbar)\mathbf{pr})$. In the relativistic **r** space the Euclidean geometry is realized and, in particular, there exist mutually commuting "generators" of the translations (see Ref. 5).

The quasipotential approach[6–8] became the dynamical basis for applying this construction, which is transparent from the physical and group-theoretic viewpoint. The equations of the quasipotential type for the relativistic two-particle amplitude $A(\mathbf{p},\mathbf{q})$ and wave function $\psi_q(\mathbf{p})$, obtained in the framework of the diagram technique of the covariant approach,[7] have an "absolute" meaning in the sense of the momentum space geometry, i.e., they look like the nonrelativistic Lippmann–Schwinger equations. The formalism emerging in the relativistic **r** space exhibits many important features of nonrelativistic quantum mechanics. Its essential difference from quantum mechanics is that the Hamiltonian in this scheme is a differential-difference operator with step equal to the Compton wavelength of particle $\lambda = \hbar/mc$. The technique of difference differentiation was developed and analogs of the important functions of the continuous analysis were obtained to fit this formalism.[9,10] Relativistic generalizations of the exactly solvable problems of quantum mechanics were considered.[10–15] In particular, solutions of the equation with Coulomb potential, which corresponds to the exchange by a massless particle, were found.[11] A three-dimensional model of the harmonic oscillator, having U(3) symmetry, was also studied in detail.[12,14,15] The explicit form of the wave functions in the spherical system of coordinates $\mathbf{r} = (r,\theta,\varphi)$ was found. The raising and lowering operators for the radial and orbital quantum numbers were defined and the dynamical symmetry group was constructed by the Infeld–Hull factorization method.[14,15] The generating function, orthogonality, and various recurrence relations for the radial wave function were obtained.[16]

As is known the nonrelativistic harmonic oscillator has been extensively used in the various fields of theoretical physics—statistical mechanics, theory of superconductivity, nuclear physics, and so on (see, for example, Ref. 17). Interest in the harmonic oscillator was revived after the appearance of the quark models, which have made it possible to describe the basic features of hadronic structure (mass spectra, decay widths, and so forth). The further development of the quark models has led to the necessity of constructing the relativistic wave functions of compound particles and, in particular, the relativistic harmonic-oscillator models.[18–23]

The characteristic feature of the harmonic oscillator is the existence of a class of solutions in the form of coherent states, closely related with the unitary representations of the Heisenberg–Weyl group.[24–26] The use of the CS makes it possible to apply more transparent classical language to describe the quantum phenomena. Later on, the generalized CS, associated with the unitary representations of an arbitrary Lie group, have been defined.[27–30] This has led to the possibility of applying this approach to a wider range of physical problems. In this paper the generalized CS for a three-dimensional relativistic model of the harmonic oscillator are considered. In the spherical system of coordinates

[a)] Permanent address: Physics Institute, Baku 370143, USSR.

 0022-2488/89/071631-04$02.50

$\mathbf{r} = (r,\theta,\varphi)$ a path integral[31] for the transition amplitude (propagator) between SU(1,1) CS is constructed[32–34] (for a more extensive list of references see Ref. 35). The partition function corresponding to the radial part of the Hamiltonian is calculated and classical equations of motion in a generalized phase space are obtained. It is shown that the use of the semiclassical Bohr–Sommerfeld quantization rule yields the exact expression for the energy levels of the oscillator considered.

II. A RELATIVISTIC MODEL OF THE HARMONIC OSCILLATOR

Because of the O(3) symmetry of a model[12,15,16,36] the dependence of the wave function

$$\Psi_{nlm}(\mathbf{r}) = r^{-1}\chi_n^l(r)\,Y_{lm}(\theta,\varphi) \tag{1}$$

on the angles θ and φ is described by the spherical harmonics $Y_{lm}(\theta,\varphi)$. Therefore a three-dimensional problem is reduced to finding the eigenvalues and eigenfunctions of the radial part of a Hamiltonian $\tilde{H}_l(r)\chi_n^l(r) = E_{nl}\chi_n^l(r)$. In the case considered the model is specified by the finite-difference operator

$$\tilde{H}_l(r) = \frac{mc^2}{2}\left\{\exp\left(-i\frac{d}{d\tilde{r}}\right) + \left[1+\left(\frac{\hbar\omega}{mc^2}\right)^2\tilde{r}^{(2)}\right]\left[1+\frac{l(l+1)}{\tilde{r}^{(2)}}\right]\exp\left(i\frac{d}{d\tilde{r}}\right)\right\}, \tag{2}$$

where $\tilde{r} = r/\lambda$ is a dimensionless variable, $\lambda = \hbar/mc$ is the Compton wavelength, $\exp(\alpha\, d/dx)f(x) = f(x+\alpha)$, and

$$x^{(\beta)} = i^{\beta}\Gamma(\beta - ix)\Gamma^{-1}(-ix) \equiv i^{\beta}(-ix)_{\beta}$$

by definition. The radial part of the wave function (1) has the form

$$\chi_n^l(r) = N_n^l(-\tilde{r})^{(l+1)}M_{\nu}(\tilde{r})\mathscr{P}_n^{\nu,l}(\tilde{r}^2), \tag{3}$$

whereas the functions $(-\tilde{r})^{(l+1)}$ and

$$M_{\nu}(\tilde{r}) = (\hbar\omega/mc^2)^{i\tilde{r}}\Gamma(\nu + i\tilde{r}),$$

$2\nu = 1 + [1 + 4(mc^2/\hbar\omega)^2]^{1/2}$, define the asymptotic behavior of $\chi_n^l(r)$ as $r\to 0$ and $r\to\infty$, respectively. The polynomial part of the wave function $\chi_n^l(r)$ is expressed through the dual Hahn polynomials (see Refs. 15, 16, and 36, where the normalization constant N_n^l is also given)

$$\mathscr{P}_n^{\nu,l}(\tilde{r}^2) = (\nu+\tfrac{1}{2})_n^{-1}\mathscr{W}_n^{(0)}(-\tilde{r}^2 - \tfrac{1}{4}; l+\tfrac{1}{2}, \tfrac{1}{2}-\nu).$$

A space of square-integrable on the $[0,\infty)$ eigenfunctions $\chi_n^l(r)$ of the radial part of the Hamiltonian $\tilde{H}_l(r) \equiv 2\hbar\omega K_0$ is a direct sum of infinite-dimensional SU(1,1) irreducible subspaces $D^+(\varkappa_l)$ (with a fixed value of the orbital quantum number l and the radial quantum number n being equal to 0, 1, 2,...) characterized by the eigenvalues $\varkappa_l = \frac{1}{2}(\nu + l + 1)$ of the invariant Casimir operator

$$K^2 = K_0^2 - \tfrac{1}{2}(K_+K_- + K_-K_+) = \varkappa_l(\varkappa_l - 1)\hat{I}.$$

The generators K_0 and $K_{\pm}$ satisfy the commutation relations

$$[K_-,K_+] = 2K_0, \quad [K_0,K_{\pm}] = \pm K_{\pm}$$

of the spectrum generating Lie algebra of SU(1,1) [or homomorphic groups $SO(2,1)\sim Sp(2,R)\sim SL(2,R)$]. They are realized by the finite-difference operators, and their action on $\chi_n^l(r)$ is defined by the formulas[15]

$$\begin{aligned} &K_0\chi_n^l(r) = (n+\varkappa_l)\chi_n^l(r),\\ &K_+\chi_n^l(r) = a_{n+1}^l\chi_{n+1}^l(r),\\ &K_-\chi_n^l(r) = a_n^l\chi_{n-1}^l(r),\\ &a_n^l = [n(n+2\varkappa_l-1)]^{1/2}. \end{aligned} \tag{4}$$

From (4) it follows that

$$\chi_n^l(r) = [n!(2\varkappa_l)_n]^{-1/2}K_+^n\chi_0^l(r). \tag{5}$$

In the nonrelativistic limit $\chi_n^l(r)$ coincides with the radial part of the Schrödinger wave function for the three-dimensional oscillator.

III. SU(1,1) COHERENT STATES

In the Hilbert space of a unitary irreducible representation $D^+(\varkappa_l)$ of the dynamical group $\widetilde{SU}(1,1)$ the coherent state wave function $\langle r|\zeta,\varkappa_l\rangle$ is defined by acting with the operator $D(\alpha) = \exp(\alpha K_+ - \alpha^* K_-)$ on $\chi_0^l(r)$, i.e.,

$$\begin{aligned} \langle r|\zeta,\varkappa_l\rangle &= D(\alpha)\chi_0^l(r)\\ &= (1-|\zeta|^2)^{\varkappa_l}\exp(\zeta K_+)\chi_0^l(r), \end{aligned} \tag{6}$$

where $\alpha = -(\tau/2)e^{-i\phi}$, $\zeta = -\tanh(\tau/2)e^{-i\phi}$, and τ and ϕ are group parameters (see Refs. 30 and 32–34). From (5) and (6) it follows that the decomposition of $\langle r|\zeta,\varkappa_l\rangle$ over the basis functions $\chi_n^l(r)$ has the form

$$\langle r|\zeta,\varkappa_l\rangle = (1-|\zeta|^2)^{\varkappa_l}\sum_{n=0}^{\infty}\left[\frac{(2\varkappa_l)_n}{n!}\right]^{1/2}\zeta^n\chi_n^l(r). \tag{7}$$

The transition amplitude (propagator) between the SU(1,1) CS is defined as a sum of the "partial" amplitudes, i.e.,

$$\begin{aligned} K(\zeta',\zeta;T) &= \sum_{l=0}^{\infty}K_l(\zeta',\zeta;T),\\ K_l(\zeta',\zeta;T) &= \langle\zeta',\varkappa_l|e^{-(i/\hbar)T\tilde{H}_l(r)}|\zeta,\varkappa_l\rangle\\ &= \langle\zeta',\varkappa_l|e^{-2i\omega TK_0}|\zeta,\varkappa_l\rangle. \end{aligned} \tag{8}$$

Using (7) it is easy to show that

$$\begin{aligned} K_l(\zeta',\zeta;T) &= e^{-2i\omega T\varkappa_l}\langle\zeta',\varkappa_l|\zeta e^{-2i\omega T},\varkappa_l\rangle\\ &= e^{-2i\omega T\varkappa_l}[(1-|\zeta|^2)(1-|\zeta'|^2)]^{\varkappa_l}\\ &\quad\times(1-\zeta\zeta'^*e^{-2i\omega T})^{-2\varkappa_l}. \end{aligned}$$

The partition function for the relativistic model of the oscillator considered is given as

$$\begin{aligned} Z &= \mathrm{Tr}\,K(\zeta',\zeta;-i\hbar\beta)\\ &= [2e^{\nu\hbar\omega\beta}(1-e^{-\hbar\omega\beta})\sinh\beta\hbar\omega]^{-1}\\ &= e^{-(\nu-1/2)\beta\hbar\omega}Z_{\mathrm{NR}}, \end{aligned}$$

where Z_{NR} is the partition function for the nonrelativistic three-dimensional oscillator.

IV. CLASSICAL EQUATIONS OF MOTION IN THE GENERALIZED PHASE SPACE

By analogy with the one-dimensional case[35,37] to each "partial" amplitude (8) it is possible to associate a path integral

$$K_l(\zeta',\zeta;T) = \int \mathcal{D}\rho_{\varkappa_l}(\zeta)\exp\left[\frac{i}{\hbar}\int_0^T \mathscr{L}_l(\zeta,\dot\zeta;\zeta^*,\dot\zeta^*)dt\right] \tag{9}$$

with the classical Lagrange function

$$\mathscr{L}_l(\zeta,\dot\zeta;\zeta^*,\dot\zeta^*) = \frac{i\hbar\varkappa_l}{1-|\zeta|^2}[\dot\zeta(t)\zeta^*(t)-\zeta(t)\dot\zeta^*(t)] - \mathscr{H}_l(\zeta,\zeta) \tag{10}$$

in a curved phase space in the form of a Lobachevsky plane (see Refs. 30 and 32–34, and 38). The classical Euler–Lagrange equations corresponding to (9),

$$\frac{d}{dt}\left(\frac{\partial\mathscr{L}_l}{\partial\dot\zeta}\right)=\frac{\partial\mathscr{L}_l}{\partial\zeta},\quad \frac{d}{dt}\left(\frac{\partial\mathscr{L}_l}{\partial\dot\zeta^*}\right)=\frac{\partial\mathscr{L}_l}{\partial\zeta^*}, \tag{11}$$

are obtained by variation of the action

$$S=\int_0^T \mathscr{L}_l\,dt.$$

If we take into account that in the case considered

$$\mathscr{H}_l(\zeta,\zeta)\equiv\mathscr{H}_l(\tau)=2\varkappa_l\hbar\omega\cosh\tau,$$

then Eqs. (11) in the terms of the group parameters τ and ϕ will have the simple form $\dot\tau=0$ and $\dot\phi=2\omega$. Evidently, the solutions of these equations are $\tau=\text{const}$ and $\phi=2\omega t+\phi_0$, i.e., the classical motion in the phase space will be oscillator-like.

V. ENERGY EIGENVALUES

To find possible values for the energy $E_l=\mathscr{H}_l(\tau)$ of a classical system described by the Lagrangian (10), let us express it through the parameters τ and ϕ:

$$\mathscr{L}_l(\tau,\phi)=\hbar\varkappa_l(\cosh\tau-1)\dot\phi-2\hbar\omega\varkappa_l\cosh\tau \equiv\hbar\varkappa_l\tilde{\mathscr{L}}(\tau,\phi). \tag{12}$$

The introduction of the momentum $\tilde p=\partial\tilde{\mathscr{L}}/\partial\dot\phi=\cosh\tau-1$, canonically conjugate to the "coordinate" ϕ, makes it possible to write (12) in a more compact form

$$\tilde{\mathscr{L}}(\tau,\phi)=\tilde p\dot\phi-2\omega(\tilde p+1).$$

Now substituting (12) in (9) we arrive at the representation

$$K_l(\zeta',\zeta;T)=\int\mathcal{D}\rho_{\varkappa_l}(\zeta)\times\exp\left[i\varkappa_l\int_0^T\tilde{\mathscr{L}}(\tau,\phi)dt\right]. \tag{13}$$

Since when $\hbar\to0$ the parameter $\varkappa_l$ characterizing an irreducible representation $D^+(\varkappa_l)$ of the dynamical group SU(1,1) behaves like $\varkappa_l\simeq mc^2/\hbar\omega$, from (13) it follows that for $\varkappa_l$ sufficiently large the motion becomes quasiclassical (cf. Ref. 34). Therefore, as $\varkappa_l\to\infty$ we can make use of the Bohr–Sommerfeld quantization rule

$$\oint\tilde p\,d\phi=\frac{2\pi}{\varkappa_l}n,\quad n=0,1,2,...,$$

from which it follows that the momentum $\tilde p=n/\varkappa_l$. Consequently, the energy of the system considered is equal to

$$E_l=\mathscr{H}_l(\tau)=2\hbar\omega\varkappa_l\cosh\tau=2\hbar\omega\varkappa_l(\tilde p+1) =2\hbar\omega(n+\varkappa_l). \tag{14}$$

Thus the semiclassical Bohr–Sommerfeld quantization rule yields, for the energy levels of the relativistic three-dimensional oscillator (2), expression (14), which coincides with the exact one. We recall in this connection that in the nonrelativistic case application of the quasiclassical JWKB method to the Schrödinger equation with the harmonic oscillator potential also gives the exact values for the energy levels.

ACKNOWLEDGMENTS

We are grateful to V. G. Kadyshevsky, V. I. Man'ko, and M. V. Savel'yev for valuable discussions. One of us (NMA) would like to thank K. B. Wolf for his hospitality at the Instituto de Investigaciones en Mathemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Cuernavaca.

[1]V. G. Kadyshevsky, R. M. Mir-Kasimov, and N. B. Skachkov, Nuovo Cimento A **55**, 233 (1968).
[2]I. S. Shapiro, Sov. Phys. Dokl. **1**, 91 (1956).
[3]N. Ya. Vilenkin and Ya. A. Smorodinsky, Sov. Phys. JETP **19**, 1209 (1964).
[4]I. M. Gel'fand, M. I. Graev, and N. Ya. Vilenkin, *Generalized Functions*, Vol. 5 (Academic, New York, 1966).
[5]V. G. Kadyshevsky, R. M. Mir-Kasimov, and N. B. Skachkov, Sov. J. Part. Nucl. **2**, 69 (1972).
[6]A. A. Logunov and A. N. Tavkhelidze, Nuovo Cimento **29**, 380 (1963).
[7]V. G. Kadyshevsky, Nucl. Phys. B **6**, 125 (1968).
[8]A. Klein and T-S. H. Lee, Phys. Rev. **10**, 4308 (1974).
[9]V. G. Kadyshevsky, R. M. Mir-Kasimov, and N. B. Skachkov, Yad. Fiz. **9**, 212 (1969).
[10]V. G. Kadyshevsky, R. M. Mir-Kasimov, and M. Freeman, Yad. Fiz. **9**, 646 (1969).
[11]M. Freeman, M. D. Mateev, and R. M. Mir-Kasimov, Nucl. Phys. B **12**, 197 (1969).
[12]A. D. Donkov, V. G. Kadyshevsky, M. D. Mateev, and R. M. Mir-Kasimov, Teor. Mat. Fiz. **8**, 61 (1971).
[13]N. M. Atakishiyev, R. M. Mir-Kasimov, and Sh. M. Nagiyev, Teor. Mat. Fiz. **44**, 47 (1980).
[14]N. M. Atakishiyev, R. M. Mir-Kasimov, and Sh. M. Nagiyev, *Proceedings of the 4th International Seminar on High Energy Physics and Quantum Field Theory*, Protvino, 1982, Vol. 2 (Institute for High Energy Physics, Protvino, 1982), p. 180.
[15]N. M. Atakishiyev, R. M. Mir-Kasimov, and Sh. M. Nagiyev, Ann. Phys. (Leipzig) **42**, 25 (1985).
[16]N. M. Atakishiyev, Teor. Mat. Fiz. **58**, 254 (1984).
[17]M. Moshinsky, *The Harmonic Oscillator in Modern Physics: From Atoms to Quarks* (Gordon and Breach, New York, 1969).
[18]H. Yukawa, Phys. Rev. **91**, 415 (1953).
[19]M. A. Markov, Suppl. Nuovo Cimento **3** (Ser. X), 760 (1956).
[20]P. N. Bogolubov, V. A. Matveev, and B. V. Struminsky, Preprint R-2442, Joint Institute for Nuclear Research, Dubna, 1965.
[21]R. P. Feynman, M. Kislinger, and F. Ravndal, Phys. Rev. D **3**, 2706 (1971).
[22]Y. S. Kim and M. E. Noz, Phys. Rev. D **8**, 3521 (1973).
[23]Y. S. Kim and M. E. Noz, Phys. Rev. D **15**, 335 (1977).
[24]R. J. Glauber, Phys. Rev. Lett. **10**, 84 (1963).
[25]R. J. Glauber, Phys. Rev. **130**, 2529 (1963).
[26]R. J. Glauber, Phys. Rev. **131**, 2766 (1963).
[27]J. R. Klauder, J. Math. Phys. **4**, 1055 (1963).

[28]J. R. Klauder, J. Math. Phys. **4**, 1058 (1963).
[29]I. A. Malkin and V. I. Man'ko, Preprint No. 15, P. N. Lebedev Institute of Physics, USSR Academy of Sciences, Moscow, 1971.
[30]A. M. Perelomov, Commun. Math. Phys. **26**, 222 (1972).
[31]D. Peak and A. Inomata, J. Math. Phys. **10**, 1422 (1969).
[32]A. M. Perelomov, Sov. Phys. Usp. **20**, 703 (1977).
[33]C. C. Gerry and S. Silverman, J. Math. Phys. **23**, 1995 (1982).
[34]C. C. Gerry, J. B. Togeas, and S. Silverman, Phys. Rev. D **28**, 1939 (1983).
[35]N. M. Atakishiyev and R. M. Mir-Kasimov, Teor. Mat. Fiz. **67**, 68 (1986).
[36]N. M. Atakishiyev, Ann. Phys. (Leipzig) **42**, 31 (1985).
[37]N. M. Atakishiyev, *Proceedings of the XVIII International Symposium*, Ahrenshoop, GDR (Institut für Hochenergiephysik, Berlin-Zeuthen, 1984), pp. 56–67.
[38]F. A. Berezin, Commun. Math. Phys. **40**, 153 (1975).

Gauge dependences of the covariant effective action in QED

Herbert Nachbagauer, Ulrike Kraemmer, and Anton Rebhan
Institut für Theoretische Physik, Technische Universität Wien, Karlsplatz 13, A-1040 Vienna, Austria

(Received 11 November 1988; accepted for publication 8 March 1989)

A recent computation of the (nonlocal) covariant effective action of QED in weak constant external electromagnetic fields is generalized to general Lorentz gauges and its gauge dependence is displayed. The latter is then removed by the prescription for a gauge and parametrization independent effective action according to Vilkovisky. The geometry, which in this framework is ascribed to the field configuration space of QED, is made explicit, and it is shown that Vilkovisky's effective action coincides with the conventional Landau gauge case.

I. INTRODUCTION

In quantum field theory the effective action, unlike the classical action functional, is not a scalar field on the infinite-dimensional space of field configurations. This gives rise to ambiguities in the effective action that disappear only on the physical subspace where the effective field equations are fulfilled, but they leave their imprints on these effective field equations themselves. In particular, in gauge field theories the effective field equations (as well as all other derivatives of the effective action) depend on the gauge fixing parameters introduced in perturbation theory even *on* the physical subspace.

A unique effective action, which is free of all such ambiguities, also off the physical mass shell, has been proposed not long ago by Vilkovisky.[1] The key ingredient in this construction is an invariant affine connection on the configuration space, which singles out a certain, however unusual, parametrization of fields through geodesic normal coordinates. In this paper we shall study this modification of the conventional framework in the example of quantum electrodynamics (QED) whose comparative simplicity allows a rather explicit demonstration of how this new concept works.

First, in Sec. II, we deal with QED in the usual field parametrization, following a recent work by Ostrovsky and Vilkovisky,[2] where a manifestly covariant calculation of the parts of the one-loop effective action bilinear in the spinor fields has been performed using a generalized Schwinger–DeWitt technique,[3] which transcends the standard methods of perturbation theory. Using the Feynman gauge, where the free photon propagator is as simple as possible, the following structure was found:

$$W_{\psi\bar{\psi}} = \int dx\, \bar{\psi}(x) \Big[(\not\partial + m)(1 + e^2\Sigma_2(\not\partial)) + e^2 m \Sigma_1(\not\partial) + \frac{e^3}{2m}(\sigma F)\Sigma_3(\not\partial) + \Big[\not\partial + m, \frac{e^3}{2m^2}(\sigma F)\Big]\Sigma_4(\not\partial)\Big]\psi(x) + O(F^2) + O(\partial F) + O(e^4), \tag{1.1}$$

where $\Sigma_i(\not\partial)$ are nonlocal operators built from the Dirac operator $\not\partial$.

Here we report the results of the corresponding calculation in general Lorentz gauges in order to display the gauge dependence of the conventional effective action. We find that a further contribution to (1.1) appears in this general case, viz.,

$$(e^3/m^3)(\not\partial + m)(\sigma F)(\not\partial + m)\Sigma_5(\not\partial), \tag{1.2}$$

and all but one of the operators $\Sigma_i(\not\partial)$ are gauge dependent. Apart from this gauge dependence the very presence of the terms proportional to $(\not\partial + m)$ in (1.1) and (1.2) might appear to be an off-shell artifact, since they vanish in the on-shell effective action, but not in the effective field equations. The same can be said of the nonlocality of the operators $\Sigma_i(\not\partial)$, which equally disappears when the iterative solution of the effective field equations is inserted back into the effective action. In Ref. 2 the question has been raised whether the situation might be qualitatively different for the novel effective action introduced by Vilkovisky.

This issue is investigated in Sec. III where we consider the manifestly gauge independent and reparametrization invariant effective action of Ref. 1. We make the geometry explicit, which in this framework is ascribed to the field configuration space of QED, and we demonstrate how gauge independence emerges in this case. Finally we find that Vilkovisky's effective action coincides with the conventional Landau gauge result.

This paper is to be understood as a sequel to Ref. 2. Hence we shall be very brief on the details of the calculation of the effective action, referring to Ref. 2 repeatedly.

II. CALCULATION OF THE NONLOCAL EFFECTIVE ACTION IN GENERAL LORENTZ GAUGES

We start from the QED Lagrangian in the general Lorentz gauge

$$\mathcal{L} = -\tfrac{1}{4}\mathbf{F}_{\mu\nu}\mathbf{F}^{\mu\nu} - \bar{\psi}(\not\partial + m)\psi - (1/2\lambda)(\partial_\mu \mathbf{A}^\mu - \partial_\mu A^\mu)^2, \tag{2.1}$$

where $\nabla_\mu = \partial_\mu - ieA_\mu$, $\not\partial = \gamma^\mu\nabla_\mu$, $F_{\mu\nu} = (i/e)[\nabla_\mu,\nabla_\nu]$ and $A^\mu = \langle \mathbf{A}^\mu \rangle$, $\psi = \langle \psi \rangle$ are the mean fields, which appear as arguments of the effective action. As in Ref. 2, we shall restrict ourselves to the case of a flat, Euclidean, 2ω-dimensional space with metric $g_{\mu\nu}$. Generally, the notations and conventions employed are the same as in Ref. 2.

 0022-2488/89/071635-05$02.50

Here $\lambda = 1$ identifies the Feynman gauge, where the classical photon propagator is given by

$$-\frac{g_{\mu\nu}}{\partial^2}\,\delta(x,y) = \int_0^\infty \frac{d\tau}{(4\pi\tau)^\omega}\, e^{-\sigma(x,y)/2\tau} g_{\mu\nu} \tag{2.2}$$

in the proper-time representation, $\sigma(x,y)$ being the geodesic interval (or "world function") between x and y. In the general gauge case (2.2) becomes

$$\begin{aligned} &-\frac{1}{\partial^2}\left[g_{\mu\nu} - (1-\lambda)\frac{\partial_\mu \partial_\nu}{\partial^2}\right]\delta(x,y) \\ &= \int_0^\infty \frac{d\tau}{(4\pi\tau)^\omega}\, e^{-\sigma(x,y)/2\tau}\left[\frac{1}{2}(1+\lambda)g_{\mu\nu}\right. \\ &\quad \left. + \frac{1}{4}(1-\lambda)\frac{\sigma_\mu \sigma_\nu}{\tau}\right], \end{aligned} \tag{2.3}$$

where $\sigma_\mu(x,y) = (\partial/\partial x^\mu)\sigma(x,y)$.

At one-loop order the parts of the effective action bilinear in the mean spinor fields are given by

$$\begin{aligned} W^{(1)}_{\bar\psi\psi} = e^2 \int dx \int dy \int_0^\infty \frac{d\tau}{(4\pi\tau)^\omega}\, e^{-\sigma(x,y)/2\tau} \\ \times \bar\psi(x)\gamma^\mu G(x,y)\gamma^\nu \psi(y) \\ \times\left(\frac{1+\lambda}{2} g_{\mu\nu} + \frac{1-\lambda}{4\tau}\sigma_\mu \sigma_\nu\right). \end{aligned} \tag{2.4}$$

Here $G(x,y)$ denotes the Green's function of the Dirac operator $\not\nabla + m$. In the approximation of a weak constant electromagnetic field it can be represented by[2]

$$\begin{aligned} G(x,y) = \int_0^\infty \frac{ds}{(4\pi s)^\omega}\, e^{-\sigma(x,y)/2s - sm^2} \\ \times\left[-m + \frac{iems}{2} F_{\mu\nu}\gamma^\mu\gamma^\nu + \left\{-\frac{1}{2s}\gamma^\rho\right.\right. \\ \left.\left. + \frac{ie}{2} F_{\mu\nu}\left(\gamma^\mu g^{\rho\nu} + \frac{1}{2}\gamma^\rho\gamma^\mu\gamma^\nu\right)\right\}\right. \\ \left.\times\sigma_\rho(x,y)\right]\hat a_0(x,y) + O(F^2) + O(\partial F). \end{aligned} \tag{2.5}$$

Here the zeroth-order Seeley–DeWitt[3] coefficient $\hat a_0$, which is the parallel displacement along the geodesic, is factored out. Its action on the spinor field is given by

$$\begin{aligned} \hat a_0(x,y)\psi(y) &= \sum_{n=0}^\infty \frac{(-1)^n}{n!}\,\sigma^{\mu_1}\cdots\sigma^{\mu_n}\nabla_{\mu_1}\cdots\nabla_{\mu_n}\psi(x) \\ &=: e^{-\sigma^\mu\nabla_\mu}\psi(x). \end{aligned} \tag{2.6}$$

Upon changing one integration variable, say y^μ, to σ^μ, which in the absence of gravity does not introduce a Jacobian, the evaluation of (2.5) reduces to a computation of Gaussian integrals with noncommuting sources[2,8] and their moments,

$$\begin{aligned} &\langle\sigma^{\mu_1}\cdots\sigma^{\mu_n}\rangle \\ &:= \frac{1}{(4\pi u)}\int\left(\prod_{\beta=1}^{2\omega} d\sigma^\beta\right)\sigma^{\mu_1}\cdots\sigma^{\mu_n} e^{-g_{\mu\nu}\sigma^\mu\sigma^\nu/4u} e^{-\sigma^\alpha\nabla_\alpha}, \end{aligned} \tag{2.7}$$

where

$$1/u := 1/s + 1/\tau.$$

Compared to the Feynman gauge calculation we now need the first four instead of two moments. In $O(\partial F)$, $O(F^2)$ one obtains

$$\langle 1\rangle = e^{u\nabla_\alpha\nabla^\alpha} = (1 + (ue/2)(\sigma F))e^{uH}, \quad H := \not\nabla^2, \tag{2.8}$$

$$\langle\sigma^\mu\rangle = -2u(g^{\mu\nu} + ieuF^{\mu\nu})\nabla_\nu e^{u\nabla^2}, \tag{2.9}$$

$$\langle\sigma^\mu\sigma^\nu\rangle = 2u(g^{\mu\nu} + 2u\nabla^{(\mu}\nabla^{\nu)} - 4ieu^2\nabla_\rho F^{\rho(\mu}\nabla^{\nu)})e^{u\nabla^2}, \tag{2.10}$$

$$\begin{aligned} \langle\sigma^\mu\sigma^\nu\sigma^\rho\rangle = 4u^2(&-3ieug^{(\mu\nu}F^{\rho)\alpha}\nabla_\alpha - 3g^{(\mu\nu}\nabla^{\rho)} \\ &- 2u\nabla^{(\mu}\nabla^\nu\nabla^{\rho)} + 6ieu^2F^{\alpha(\mu}\nabla^\nu\nabla^{\rho)}\nabla_\alpha)e^{u\nabla^2}, \end{aligned} \tag{2.11}$$

where $(\sigma F) := (i/2)\,[\gamma^\mu,\gamma^\nu]\,F_{\mu\nu}$ and parentheses around the indices denote symmetrization with unit weight, so that for commuting quantities $\alpha_{(\mu}\alpha_\nu\alpha_{\rho)} = \alpha_\mu\alpha_\nu\alpha_\rho$.

With (2.8)–(2.11), the functional (2.4) reduces to integrals in the proper-time parameters s and τ, whose ultraviolet divergences can be regularized by analytic continuation in the space-time dimension $2\omega\in\mathbb{N}$ to $2\omega\in\mathbb{C}$.

The evaluation of (2.4) with general $\omega\in\mathbb{C}$ can be found in the Appendix. For the four-dimensional case, one has to consider the Laurent series expansion around $\omega = 2$, where neglecting the terms of order $(\omega - 2)$ the integrals become elementary and yield the following result. [For $\lambda = 1$ we reproduce the results of Ref. 2, except for the polynomials in N in Σ_4, where we get $(N+\frac{5}{2})(1+N)$ instead of $2(1+N)$.]

$$\begin{aligned} W^{(1)}_{\bar\psi\psi} = \frac{e^2}{(4\pi)^2}\int dx\,\bar\psi(x)\left\{m\left[3\left(-\frac{1}{2-\omega} + \ln\frac{m^2}{\tilde\mu^2}\right) - 4 + \lambda N + (3-\lambda-\lambda N)N\ln\Lambda^{-1}\right]\right. \\ + (\not\nabla + m)\lambda\left[-\frac{1}{2-\omega} + \ln\frac{m^2}{\tilde\mu^2} - 2 - N + (2+N)N\ln\Lambda^{-1}\right] \\ + \frac{e}{2m}(\sigma F)[-(1+N)(1+2N) + 2(1+N)^2N\ln\Lambda^{-1}] \\ + \left[\not\nabla + m, \frac{e}{2m^2}(\sigma F)\right](1+N)[(2N+3) - 2(1+N)^2\ln\Lambda^{-1} + \lambda(-\tfrac{1}{2}(1+2N) + (1+N)N\ln\Lambda^{-1})] \\ \left. + (1-\lambda)\frac{e}{m^3}(\not\nabla + m)\sigma F(\not\nabla + m)(1+N)^2[1 - \tfrac{1}{2}(1+2N)\ln\Lambda^{-1}]\right\}\psi(x) \\ + O(\omega - 2) + O(F^2) + O(\partial F), \end{aligned} \tag{2.12}$$

where

$$N:=(m^2-H)/H$$

is a nonlocal operator annihilating $\psi(x)$ on mass shell and

$$\Lambda:=N/(N+1)=(m^2-H)/m^2.$$

Here $\tilde{\mu}$ is the mass scale introduced by dimensional regularization.

The result (2.12) is seen to be highly gauge dependent. Notable exceptions are the mass counterterm

$$-\Sigma_1(-m)=-\frac{e^2}{(4\pi)^2}\left[3\left(-\frac{1}{2-\omega}+\ln\frac{m^2}{\tilde{\mu}^2}\right)-4\right] \tag{2.13}$$

and the operator $\Sigma_3(\not{\partial})$ even off the physical mass shell. On-shell $\Sigma_3(-m)$ determines the anomalous magnetic moment in a weak constant electromagnetic field.

All other terms depend on the gauge fixing parameter λ. The gauge dependence of the first two structure functions Σ_1 and Σ_2 is already well known from calculations of the corresponding Green's functions.[9] With $\lambda=0$ (Landau gauge) the entire contribution Σ_2 vanishes, whereas for Σ_1 the gauge choice $\lambda=3$ (the Abrikosov–Yennie gauge[5]) is special: If a renormalization condition is formulated so that the renormalized electron propagator has unit residue at the physical mass pole, the wave function counterterm is determined by

$$\Sigma_2(-m)-\Sigma_1'(-m).$$

Now the term $(3-\lambda-\lambda N)N\ln\Lambda^{-1}$ in Σ_1 gives rise to an on-shell divergence when differentiated—unless $\lambda=3$. (In Ref. 2 the on-shell fermion self energy, rather than the on-shell propagator, has been normalized, which does not lead to infrared divergences.)

As for the remaining contributions proportional to $F_{\mu\nu}$, only the last one containing $(\not{\partial}+m)\sigma F(\not{\partial}+m)$ can be removed by a suitable gauge choice: this time $\lambda=1$ (the Feynman gauge) is singled out. The term with the anticommutator $\{\not{\partial}+m,\sigma F\}$, however, is present for any value of λ.

All these gauge dependences are very unsatisfactory from the point of view of the effective field equations, where in contrast to the on-shell effective action the gauge parameter λ will not drop out in general.

In the next section we shall consider the gauge and reparametrization invariant definition of the off-shell effective action[1] to tackle these equations.

III. THE GAUGE AND PARAMETRIZATION INDEPENDENT EFFECTIVE ACTION IN QED

In Ref. 1 it has been pointed out that gauge dependence of the off-shell effective action is a manifestation of the more general field parametrization dependence of the conventional definition of the effective action.

However, if the (infinite-dimensional) field configuration space is naturally endowed with an affine connection, a preferred parametrization is given by geodesic normal coordinates.

For a large class of gauge field theories, a connection conforming to the gauge structure of the field configuration space has been constructed in Ref. 1. The necessary condition to achieve gauge fixing independence turns out to be that geodesics thus defined on the whole configuration space project onto geodesics on the quotient space of gauge orbits. This implies that covariant functional derivatives of the gauge generators (i.e., the tangent vectors to the gauge orbits) again have to be proportional to gauge generators.[1,4] In QED this requirement is violated by regarding the usual parametrization through $A_\mu(x)$ and $\psi(x)$ as "Cartesian," as is implicit in the conventional formalism.

Following the construction of Ref. 1 and the prescription of Ref. 6 for the inclusion of Fermi fields, an affine, torsionless connection with the above property is given by (in the conventional parametrization)

$$\begin{aligned}
\Gamma^{\psi}_{\psi A''_\mu}&=\Gamma^{\psi}_{A''_\mu\psi}=-ie\left(\frac{\partial_\mu}{\partial^2}\right)''\delta(x-x')\delta(x-x''),\\
\Gamma^{\bar\psi}_{\bar\psi A''_\mu}&=\Gamma^{\bar\psi}_{A''_\mu\bar\psi}=ie\left(\frac{\partial_\mu}{\partial^2}\right)''\delta(x-x')\delta(x-x''),\\
\Gamma^{\psi}_{A'_\mu A''_\nu}&=-e^2\psi(x)\left(\frac{\partial_\mu}{\partial^2}\right)'\left(\frac{\partial_\nu}{\partial^2}\right)''\delta(x-x')\delta(x-x''),\\
\Gamma^{\bar\psi}_{A'_\mu A''_\nu}&=-e^2\bar\psi(x)\left(\frac{\partial_\mu}{\partial^2}\right)'\left(\frac{\partial_\nu}{\partial^2}\right)''\delta(x-x')\delta(x-x''),
\end{aligned} \tag{3.1}$$

as the only nonvanishing connection components. [Here we use the notation $A'_\mu=A_\mu(x')$, etc.]

With (3.1) the gauge generators of QED, which are given by

$$D^{A_\mu}_{x''}=\frac{\partial}{\partial x^\mu}\delta(x-x'), \tag{3.2}$$

$$D^{\psi}_{x'}=-ie\psi(x)\delta(x-x') \tag{3.3}$$

are even found to be covariantly constant vector fields on the configuration space. For (3.2) this is seen immediately; only the covariant functional differentiation of (3.3) needs closer inspection, and, indeed, one finds

$$D^{\psi}_{x'';A'_\mu}=\int dx'''(\Gamma^{\psi}_{A'_\mu\psi'''}D^{\psi'''}_{x''}+\Gamma^{\psi}_{A'_\mu A'''_\nu}D^{A'''_\nu}_{x''})\equiv 0, \tag{3.4}$$

$$D^{\psi}_{x'';\psi}=\frac{\delta}{\delta\psi'}D^{\psi}_{x''}+\int dx'''\,\Gamma^{\psi}_{\psi A'''_\mu}D^{A'''_\nu}_{x''}\equiv 0. \tag{3.5}$$

QED is also special in that the invariant connection (3.1) is flat. [The vanishing of the curvature functional entails that the original version of the reparametrization invariant effective action proposed by Vilkovisky[1] and the modified one due to DeWitt[4] coincide. (For their difference in the general case cf. Refs. 6 and 7.)] To verify this assertion, the only nonobvious relations to be checked are

$$R^{\psi}_{A'_\mu\psi''A'''_\nu}=\frac{\delta}{\delta\psi''}\Gamma^{\psi}_{A'_\mu A'''_\nu}-\int dx''''\,\Gamma^{\psi}_{\psi'''A'''_\nu}\Gamma^{\psi'''}_{\psi''A'_\mu}=0, \tag{3.6}$$

$$\begin{aligned}
R^{\psi}_{\psi A''_\mu A'''_\nu}=\int dx''''(&\Gamma^{\psi}_{\psi'''A''_\mu}\Gamma^{\psi'''}_{A'''_\nu\psi}\\
&-\Gamma^{\psi}_{\psi'''A'''_\nu}\Gamma^{\psi'''}_{A''_\mu\psi})=0,
\end{aligned} \tag{3.7}$$

$$\begin{aligned}
R^{\psi}_{A'_\mu A''_\nu A'''_\lambda}=\int dx''''(&\Gamma^{\psi}_{\psi'''A''_\nu}\Gamma^{\psi'''}_{A'''_\lambda A'_\mu}\\
&-\Gamma^{\psi}_{\psi'''A'''_\lambda}\Gamma^{\psi'''}_{A''_\nu A'_\mu})=0.
\end{aligned} \tag{3.8}$$

The effect of going over to geodesic normal coordinates corresponding to (3.1) in place of the conventional parametrization of the configuration space is that all ordinary functional derivatives are to be replaced by covariant ones.

We will now show that this removes the gauge fixing dependences from the off-shell effective action. In the case of $W^{(1)}_{\bar\psi\psi}$, as computed above, the only modification takes place in the vertex

$$S_{,\bar\psi A'_\mu\psi''}|_{\psi=0} = -ie\gamma_\mu\delta(x-x')\delta(x-x''), \tag{3.9}$$

which is replaced by

$$\begin{aligned} S_{;\bar\psi A'_\mu\psi''}|_{\psi=0} &= S_{,\bar\psi A'_\mu\psi''} \\ &\quad - \int dx'''(\Gamma^{\bar\psi'''}_{\bar\psi A'_\mu}S_{,\bar\psi'''\psi''} \\ &\quad - S_{,\bar\psi\psi'''}\Gamma^{\psi'''}_{\psi''A'_\mu})\Big|_{\psi=0} \\ &= -ie\gamma_\nu\left(\delta^\nu_\mu - \frac{\partial^\nu\partial_\mu}{\partial^2}\right)'\delta(x-x')\delta(x-x''). \end{aligned} \tag{3.10}$$

(Here S denotes the classical action functional.)

By this the photon propagator becomes sandwiched between two transverse projection operators. The photon propagator in an arbitrary linear gauge imposed by the gauge breaking term

$$\frac{1}{2\xi}\int dx(f^\nu A_\nu)^2 = \frac{1}{2\xi}\int dx\, A_\mu f^{\dagger\mu}f^\nu A_\nu \tag{3.11}$$

reads

$$\begin{aligned} \Delta_{A_\mu A_\nu} = -\frac{1}{\partial^2}\Bigg[g_{\mu\nu} - \frac{\partial_\mu f_\nu}{(f\cdot\partial)} - \frac{f^\dagger_\mu\partial_\nu}{(f^\dagger\cdot\partial)} \\ + \frac{\partial_\mu\partial_\nu(f\cdot f^\dagger + \xi\partial^2)}{(f\cdot\partial)(f^\dagger\cdot\partial)}\Bigg]\delta(x,y), \end{aligned} \tag{3.12}$$

if $[\partial_\mu, f_\nu] = 0$. By (3.10) this general propagator effectively reduces to the one of the Landau gauge [$\lambda = 0$ in (2.3)]:

$$\left(\delta^\rho_\mu - \frac{\partial_\mu\partial^\rho}{\partial^2}\right)\Delta_{A_\rho A_\sigma}\left(\delta^\sigma_\nu - \frac{\partial_\nu\partial^\sigma}{\partial^2}\right) = \Delta^{\text{Landau}}_{A_\mu A_\nu}, \tag{3.13}$$

losing any memory of the original gauge breaking terms.

Thus the conclusion is that the gauge independent effective action of Ref. 1 coincides with the ordinary Landau gauge result. The question raised in Ref. 2, whether the unique effective action might lead to a qualitatively different situation (e.g., as to nonlocality), is answered in the negative. The only distinguished feature of $\lambda = 0$ in the result (2.12) is the vanishing of Σ_2 in the one-loop approximation. However, the other terms that appear to be off-shell artifacts and which have been deemed unsatisfactory in Ref. 2 are not removed. If the gauge independent effective action of Ref. 1 is regarded as the final answer, then these terms have to be taken seriously as they contribute to the now unambiguous effective field equations.

ACKNOWLEDGMENTS

We would like to thank Professor W. Kummer for a critical reading of the manuscript. One of the authors (A. R.) is grateful to G. A. Vilkovisky for informing him about Ref. 2 prior to publication as well as for interesting discussions and his kind hospitality.

APPENDIX: THE EFFECTIVE ACTION $W^{(1)}_{\bar\psi\psi}$ IN ARBITRARY DIMENSIONS

In this Appendix we give the results for the effective action $W^{(1)}_{\bar\psi\psi}$ in arbitrary dimensions 2ω.

Insertion of the formulas (2.8)–(2.11) into (2.4) yields the following integral representations of the operators $\Sigma_i(\not\partial)$ introduced in (1.1) and (1.2):

$$\begin{aligned} \Sigma_1(\not\partial) &= (1+\lambda)(-\omega J_{000} + (\omega-1)J_{100}) \\ &\quad - (1-\lambda)(\omega J_{010} + HJ_{111} + (\omega+1)J_{110} + HJ_{211}), \end{aligned} \tag{A1}$$

$$\Sigma_2(\not\partial) = (1+\lambda)(-(\omega-1)J_{100}) + (1-\lambda)((\omega+1)J_{100} + HJ_{211}), \tag{A2}$$

$$\begin{aligned} \Sigma_3(\not\partial) &= ((1+\lambda)/2)(-(2\omega-3)J_{101} + (\omega-1)J_{201} + (\omega-2)J_{001}) \\ &\quad - ((1-\lambda)/2)((2-\omega)J_{011} + (\omega+2)J_{211} + HJ_{312} - m^2J_{112}), \end{aligned} \tag{A3}$$

$$\begin{aligned} \Sigma_4(\not\partial) &= ((1+\lambda)/2)m^2(-(\omega-1)J_{201} - (3-\omega)J_{101}) \\ &\quad + ((1-\lambda)/2)m^2((\omega+2)J_{211} + HJ_{312} - (\omega+1)J_{111} - HJ_{212} - 2m^2J_{112}), \end{aligned} \tag{A4}$$

$$\Sigma_5(\not\partial) = -((1-\lambda)/2)m^4J_{112}, \tag{A5}$$

where

$$\begin{aligned} J_{a,b,c} &= \frac{1}{(4\pi)^\omega}\int_0^\infty\int_0^\infty d\tau\, ds \\ &\quad \times\frac{\tau^a s^{b+c}}{(s+\tau)^{a+b+\omega}}\exp\left(\frac{s\tau}{s+\tau}H - sm^2\right) \\ &= \frac{\Gamma(b+\omega-1)\Gamma(c+2-\omega)\Gamma(a+1)}{(4\pi)^\omega\Gamma(a+b+\omega)} \\ &\quad \times{}_2F_1\left(c+2-\omega, a+1; a+b+\omega; \frac{H}{m^2}\right). \end{aligned} \tag{A6}$$

According to dimensional regularization ω has been generalized to $\omega\in\mathbb{C}$.

The above expressions can be considerably simplified by the following integral relations:

$$J_{a,b,c} = J_{a+1,b,c} + J_{a,b+1,c}, \tag{A7}$$

$$m^2J_{a,b,c} = HJ_{a+1,b,c} - (c-1+2-\omega)J_{a,b,c-1}, \tag{A8}$$

$$HJ_{a,b,c} = (b+\omega-2)J_{a,b-1,c-1} - aJ_{a-1,b,c-1}, \tag{A9}$$

which yields

$$\Sigma_1 = (1+\lambda)(-\omega J_{000} + (\omega-1)J_{100}) - (1-\lambda)(\omega-1)(J_{010}+J_{100}), \quad \text{(A10)}$$

$$\Sigma_2 = -2\lambda(\omega-1)J_{100}, \quad \text{(A11)}$$

$$\Sigma_3 = -J_{111} + (\omega-2)J_{021}, \quad \text{(A12)}$$

$$\Sigma_4 = ((1+\lambda)/2)m^2(-J_{101} - J_{201} + (\omega-2)J_{111}) + ((1-\lambda)/2)m^2(\omega-1)(J_{101} - 3J_{201}), \quad \text{(A13)}$$

$$\Sigma_5 = ((1-\lambda)/2)m^2(\omega-1)(J_{101} - 2J_{201}). \quad \text{(A14)}$$

The pattern of gauge parameter (λ) dependences, as discussed in the text, remains the same for all dimensions (with the notable exception of $\omega = 1$): Σ_2 and Σ_5 vanish in the Landau gauge and Feynman gauge, respectively, and Σ_3 is gauge independent, even off shell.

The *on-shell* value of Σ_3 is given by

$$\Sigma_3|_{\not p = -m} = (\omega-3)/2, \quad \text{(A15)}$$

corresponding to an anomalous magnetic moment

$$g - 2 = (e^2/4\pi^2)(3-\omega) + O(e^4), \quad \text{(A16)}$$

which in the present one-loop order is a linear function of the space-time dimension $D = 2\omega$.

[1] G. A. Vilkovisky, in *Quantum Theory of Gravity*, edited by S. M. Christensen (Hilger, London, 1984); Nucl. Phys. B **234**, 125 (1984).
[2] A. A. Ostrovksy and G. A. Vilkovisky, J. Math. Phys. **29**, 702 (1988).
[3] J. S. Schwinger, Phys. Rev. **82**, 664 (1951); B. S. DeWitt, *Dynamical Theory of Groups and Fields* (Gordon and Breach, New York, 1965); A. O. Barvinsky and G. A. Vilkovisky, Phys. Rep. **119**, 1 (1985).
[4] B. S. DeWitt in *Quantum Field Theory and Quantum Statistics*, edited by I. A. Batalin, C. J. Isham, and G. A. Vilkovisky (Hilger, London, 1987), Vol. 1.
[5] A. A. Abrikosov, Sov. Phys. JETP **5**, 1174 (1957); H. M. Fried and D. R. Yennie, Phys. Rev. **112**, 1391 (1958); Phys. Rev. Lett. **4**, 580 (1960).
[6] A. Rebhan, in *Tev Physics Proceedings of the 12th Johns Hopkins Workshop on Current Problems in Particle Theory* (World Scientific, Singapore, 1988); Nucl. Phys. B **288**, 832 (1987).
[7] A. Rebhan, Nucl. Phys. B **298**, 726 (1988).
[8] A. O. Barvinsky and G. A. Vilkovisky, Nucl. Phys. B **282**, 163 (1987); *ibid.* in press.
[9] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

Some remarks on Einstein–Maxwell solitons

Humberto Salazar I.
Esc. Cs. Fis-mat. Universidad Autónoma de Puebla, Av. Sn. Claudio y 14 Sur, Puebla, Mexico and Sektion Physik, FSU Jena, Max-Wien-Platz 1, DDR-6900 Jena, German Democratic Republic

(Received 11 October 1988; accepted for publication 22 March 1989)

A short expression is given for the gravitational and electromagnetic potential of N Kerr–Newman particles. The free parameters of the solution and some limiting cases are discussed.

I. INTRODUCTION

The N-soliton solution of the Einstein–Maxwell stationary axisymmetric equations (Ernst equations) has been found by several authors.[1–3] The arising solution has the interpretation of N Kerr–Newman particles placed on the z axis on arbitrary background. However, from the solution given by these authors, the discussion of the free parameters and other important matters, such as the linear dependence of the gravitational potential ξ and the electromagnetic potential ϕ, is not at all direct. In this work, starting from the result and formulation of Neugebauer and Kramer,[1] we give a short expression for the resulting ξ and ϕ potentials of N aligned Kerr–Newman particles. The obtained expression enables us to see clearly that this general solution involves $3N$ complex constants associated with the complex mass, complex charge, rotation, and the place on the z axis for each Kerr–Newman particle. Furthermore this simple expression could be used to discuss the equilibrium achieved by avoiding line singularities on the z axis for two Kerr–Newman particles.

In Sec. II, we describe the formulation of Neugebauer and Kramer and find a short expression, by means of $(N+1)\times(N+1)$ determinants for the potentials ξ and ϕ. In Sec. III, we discuss the interpretation of the free constants generated by every soliton step and relate them with the physical quantities associated to the Kerr–Newman particles. We discuss also the conditions for the linear dependence of the potentials ξ and ϕ, and show that the conform stationary solution, characterized by $\xi = 0$, is a particular case of this general solution. We finish by discussing the possibilities of generating Kerr–Newman black holes.

II. THE N-SOLITON SOLUTION

The Einstein–Maxwell equations for stationary axisymmetric fields, in terms of the complex potentials (ξ,ϕ), read[4]

$$f\Delta\xi = (\nabla\xi + 2\bar{\phi}\nabla\phi)\nabla\xi,$$
$$f\Delta\phi = (\nabla\xi + 2\bar{\phi}\nabla\phi)\nabla\phi, \quad f = \mathrm{Re}\,\xi + \phi\bar{\phi}. \tag{1}$$

These nonlinear field equations are the integrability conditions of the linear problem[1]

$$\Omega_{,\zeta} = \left[\begin{pmatrix} B_1 & 0 & E_1 \\ 0 & A_1 & 0 \\ -F_1 & 0 & \frac{1}{2}(A_1+B_1) \end{pmatrix} + \lambda \begin{pmatrix} 0 & B_1 & 0 \\ A_1 & 0 & -E_1 \\ 0 & -F_1 & 0 \end{pmatrix}\right]\Omega,$$
$$\Omega_{,\bar{\zeta}} = \left[\begin{pmatrix} B_2 & 0 & E_2 \\ 0 & A_2 & 0 \\ -F_2 & 0 & \frac{1}{2}(A_2+B_2) \end{pmatrix} + \frac{1}{\lambda} \begin{pmatrix} 0 & B_2 & 0 \\ A_2 & 0 & -E_2 \\ 0 & -F_2 & 0 \end{pmatrix}\right]\Omega, \tag{2}$$

where the 3×3 pseudopotential matrix $\Omega = \Omega(\lambda,\zeta,\bar{\zeta})$ is normalized to

$$\Omega(1,\zeta,\bar{\zeta}) = \begin{pmatrix} \bar{\xi} + 2\phi\bar{\phi} & 1 & \sqrt{2}i\phi \\ \xi & -1 & -\sqrt{2}i\phi \\ -2i\bar{\phi}f^{1/2} & 0 & \sqrt{2}f^{1/2} \end{pmatrix}. \tag{3}$$

The constant spectral parameter κ is hidden in λ,

$$\lambda = \lambda(\kappa) = [(\kappa - i\bar{\zeta})/(\kappa + i\zeta)]^{1/2}. \tag{4}$$

Equations (2)–(4) imply expressions for the λ-independent matrix components $A_1,\ldots,F_2$ in terms of the potentials ξ and ϕ, and their partial derivatives.

For any given initial solution (ξ_0, ϕ_0) there is an associated matrix Ω_0 satisfying the linear problem (2) and the normalization (3). By means of the ansatz (see also Ref. 5)

$$\Omega = T\Omega_0, \quad T = T(\lambda,\zeta,\bar{\zeta}) \equiv T(\lambda),$$

$$T(\lambda) = \alpha(\kappa)(\kappa + i\zeta)^{n/2}\sum_{s=0}^{n} X_s\lambda^s, \quad n = 2N,$$
$$T(1) = \sum_{s=0}^{n} X_s, \tag{5}$$
$$X_{2s} = \begin{pmatrix} a_{2s} & 0 & b_{2s} \\ 0 & c_{2s} & 0 \\ ch_{2s} & 0 & d_{2s} \end{pmatrix},$$
$$X_{2s+1} = \begin{pmatrix} 0 & f_{2s+1} & 0 \\ g_{2s+1} & 0 & h_{2s+1} \\ 0 & j_{2s+1} & 0 \end{pmatrix}.$$

With λ-independent 3×3 matrices X_s and a suitably chosen constant $\alpha(\kappa)$, one constructs from Ω_0 a new matrix Ω which again obeys (2) and (3). The matrices X_s can be completely determined from the following system of algebraic equations:

 0022-2488/89/071640-03$02.50

$$\sum_{s=0}^{n} X_s \lambda_\kappa^s P_\kappa = 0, \quad P_\kappa = \Omega_0(\lambda_\kappa) C_\kappa = \begin{pmatrix} p_\kappa \\ q_\kappa \\ r_\kappa \end{pmatrix},$$

$$\begin{aligned} &T_{11}(1) - T_{12}(1) = 1, \quad T_{13}(1) + T_{23}(1) = 0, \\ &T_{22}(1) - T_{21}(1) = 1, \quad T_{13}(1) - T_{32}(1) = 0, \\ &T_{33}(1) = \bar{T}_{33}(1) = (1 + T_{12}(1) + T_{21}(1))^{1/2}. \end{aligned} \tag{6}$$

The zeros of the det T and the constant vectors C_κ are restricted by

$$\lambda_{3m} = \lambda_{3m-2}, \quad \lambda_{3m-1} = 1/\bar{\lambda}_{3m}, \quad m = 1,...,N, \tag{7}$$

$$C_{3m+1} \sigma C_{3m}^{\dagger} = 0 = C_{3m-1} \sigma C_{3m-2}^{\dagger},$$

$$\sigma := \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \tag{8}$$

for a given Ω_0 and with the prescribed values λ_κ and C_κ satisfying (7) and (8), one obtains from (5) the transformation matrix T and, consequently, Ω. From Ω one can read immediately

$$\xi = \Omega_{21}(1), \quad \phi = (i/\sqrt{2})\Omega_{23}(1). \tag{9}$$

From Eq. (9) we can see that the new complex potentials ξ, ϕ involve only the second row of the transformation matrix T, $T_{2\mu}$, i.e., involves only the $N+1$ matrix elements c_{2s}, $s = 0,...,N$ and the $2N$ matrix elements g_{2s+1}, h_{2s+1}; $s = 0,...,N-1$. From Eq. (6), the system of $3N+1$ algebraic equations for the $3N+1$ matrix elements c_{2s}, g_{2s+1}, and h_{2s+1} are

$$\begin{aligned} &\sum_{s=0}^{N-1} g_{2s+1} \lambda_\kappa^{2s+1} p_\kappa + \sum_{s=0}^{N} c_{2s} \lambda_\kappa^{2s} q_\kappa \\ &\quad + \sum_{s=0}^{N-1} h_{2s+1} \lambda^{2s+1} r_\kappa = 0, \quad \kappa = 1,...,3N, \\ &\sum_{s=0}^{N} c_{2s} - \sum_{s=0}^{N-1} g_{2s+1} = 1. \end{aligned} \tag{10}$$

Solving this system we arrive at an expression for the complex potentials ξ and ϕ as a quotient of the determinants of $(3N+1)\times(3N+1)$ (Ref. 1). We will reduce the system of Eqs. (10) to a system of $N+1$ algebraic equations.

First we redefine the zeros λ_κ of $\det(T)$ as[6]

$$\begin{aligned} &\lambda_M := \lambda_{3m} = \lambda_{3m-2} = [(\kappa_M - i\bar{\zeta})/(\kappa_M + i\zeta)]^{1/2}, \\ &\lambda_{\dot{M}} := \lambda_{3m-1} = [(\kappa_{\dot{M}} - \bar{\zeta})/(\kappa_{\dot{M}} + i\zeta)]^{1/2} \end{aligned} \tag{11}$$

and the polynomials $c(\lambda)$, $g(\lambda)$, and $h(\lambda)$ as

$$\begin{aligned} &c(\lambda) = \sum_{s=0}^{N} c_{2s} \lambda^{2s}, \quad g(\lambda) = \sum_{s=0}^{N-1} g_{2s+1} \lambda^{2s}, \\ &h(\lambda) = \sum_{s=0}^{N-1} h_{2s+1} \lambda^{2s}. \end{aligned} \tag{12}$$

Then from Eqs. (10) and the condition $\lambda_{3m} = \lambda_{3m-2}$, we get

$$\begin{aligned} &g(\lambda_A) = (1/\beta_A) h(\lambda_A), \\ &c(\lambda_A) = -(\alpha A/\beta_A) \lambda_A h(\lambda_A), \\ &c(1) - g(1) = 1, \quad A = 1,...,N, \end{aligned} \tag{12'}$$

and

$$\begin{aligned} &g(\lambda_{\dot{A}}) \gamma_{\dot{A}} \lambda_{\dot{A}} + C(\lambda_{\dot{A}}) + h(\lambda_{\dot{A}}) \delta_{\dot{A}} \lambda_{\dot{A}} = 0, \\ &\dot{A} = \dot{1},..., \dot{N}, \end{aligned} \tag{13}$$

where we have defined $\alpha_M, \beta_M, \gamma_{\dot{M}}$, and $\delta_{\dot{M}}$ by

$$\begin{aligned} &\alpha_M = \frac{p_{3m-2} r_{3m} - p_{3m} r_{3m-2}}{q_{3m-2} r_{3m} - q_{3m} r_{3m-2}}, \\ &\beta_M = \frac{p_{3m-2} q_{3m} - p_{3m} q_{3m-2}}{q_{3m-2} r_{3m} - q_{3m} r_{3m-2}}, \\ &\lambda_{\dot{M}} = \frac{p_{2m-1}}{q_{3m-1}}, \quad \delta_{\dot{M}} = \frac{\lambda_{3m-1}}{q_{3m-1}}. \end{aligned} \tag{14}$$

The set of N of Eqs. (12) permits us to know the matrix elements c_{2s} and g_{2s+1} in terms of the N matrix elements h_{2s+1}. From this result we now only need to compute the h_{2s+1} elements, which can be achieved by solving the set of Eqs. (13). Actually the final expression for the complex potential ξ, ϕ depends only on $g(1)$ and $h(1)$ [we must remember that $c(1) = 1 + g(1)$]. From Eqs. (5) and (9) we have

$$\begin{aligned} &\xi = \xi_0 + 2 f_0 g(1) - 2i\bar{\phi}_0 f_0^{1/2} h(1), \\ &\phi = \phi_0 + i f_0^{1/2} h(1), \end{aligned} \tag{15}$$

Then, solving Eqs. (12) and (13), the expressions for $g(1)$ and $h(1)$ are

$$g(1) = -\frac{\Delta_1}{\Delta}, \quad h(1) = -\frac{\Delta_2}{\Delta}, \quad \Delta = \det S_{\dot{A}B},$$

$$\Delta_1 = \begin{vmatrix} 0 & 1 & \cdot & \cdot & 1 \\ 1 & S_{\dot{1}1} & \cdot & \cdot & S_{\dot{1}N} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ 1 & S_{\dot{N}1} & \cdot & \cdot & S_{\dot{N}N} \end{vmatrix},$$

$$\Delta_2 = \begin{vmatrix} 0 & \beta_1 & \cdot & \cdot & \beta_N \\ 1 & S_{\dot{1}1} & \cdot & \cdot & S_{\dot{1}N} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ 1 & S_{\dot{N}1} & \cdot & \cdot & S_{\dot{N}N} \end{vmatrix}, \tag{16}$$

where

$$\begin{aligned} &S_{\dot{A}B} = 1 - [1/(\kappa_{\dot{A}} - \kappa_B)](\alpha_B R_B - \eta_{\dot{A}B} R_{\dot{A}}), \\ &\eta_{\dot{A}B} = \gamma_{\dot{A}} + \beta_B \delta_{\dot{A}}, \\ &R_A = [\rho^2 + (z - \kappa_A)^2]^{1/2}, \\ &R_{\dot{A}} = [\rho^2 + (z - \kappa_{\dot{A}})^2]^{1/2}. \end{aligned} \tag{17}$$

The constraints (7) and (8) in terms of the new functions $\alpha_{\dot{M}}, \beta_M, \gamma_{\dot{M}}, \delta_{\dot{M}}$ and $\kappa_M, \kappa_{\dot{M}}$ read

$$\bar{\gamma}_{\dot{M}} = 1/\alpha_M, \quad \bar{\delta}_{\dot{M}} = -\beta_M/\alpha_M, \tag{18}$$

and

$$\kappa_M = \bar{\kappa}_{\dot{M}}.$$

Equations (15)–(17) together with the restriction (18) describe the more general solution generated by soliton methods on arbitrary background; it involves the seed solution and $3N$ new complex constants.

III. SOME LIMITS OF THE *N*-SOLITON SOLUTION

To give a first interpretation of the general solution and the constants that it involves, we take the following particular case. Let the seed solution be a flat space (Minkowskian), i.e., $\xi_0 = 1, \phi_0 = 0$. Then the matrix Ω_0 associated with this space-time is

$$\Omega_0 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}.$$

Because the C_κ's are constant vectors, from Eqs. (6) and (14) α_M and β_M are also constants. We can define six new real constants z_A, a_A, m_A, n_A, e_A, and g_A instead of $\kappa_A, \alpha_A, \beta_A$ by

$$\kappa_A = -i\sigma_A + z_A, \quad \alpha_A = \frac{i(a_A + \sigma_A)}{m_A + in_A}, \qquad \beta_A = -\frac{i(e_A + ig_A)}{m_A + in_A}, \tag{19}$$

where

$$-\sigma_A{}^2 = m_A{}^2 + n_A{}^2 - a_A{}^2 - e_A{}^2 - g_A{}^2. \tag{20}$$

It is very easy to read the meaning of the new constants by taking the simpler case $N = 1 (\xi_0 = 1, \phi_0 = 0)$. For this case we get the hyperextreme Kerr–Newmann solution with $M_1 = m_1 + in_1$ complex mass (mass and NUT parameter), $e_1 + ig_1$ complex charge (electric and magnetic charge), and a_1 being the rotation parameter. Also, Z_1 denotes the place of the source on the z axis (for this solution this parameter can be taken equal to zero).

For the case $N = 2$ we obtain immediately from Eqs. (15)–(20) the twelve-parametric, two-soliton solution.[3]

Another interesting particular case, which follows in a very natural way from our formulation, is the linearly dependent case (for the ξ and ϕ potential). A subcase of this last one is the superposition of N collinear Kerr–Newmann sources with gravitational attraction and electrostatic repulsion balanced (in the sense that $|e_A + ig_A| = |m_A + in_A|$). Let us consider this case in a detailed form. From Eqs. (15)–(18) the complex potentials ξ, ϕ can be linearly dependent only for flat background $\xi_0 = 1, \quad \phi_0 = 0$ and $\beta_1 = \beta_2 = \cdots = \beta_N = -i\beta$, i.e., when the quotient of the complex charge and mass is the same for every source. From Eqs. (15)–(18) and this condition we arrive at

$$\xi = 1 + 2g(1), \quad \phi = \beta g(1).$$

Moreover, if $|\beta| = 1$, i.e., if the electrostatic repulsion and gravitational attraction are balanced for each source, then by means of the gauge transformation $\gamma' = \phi + \beta$ we get $\xi' = 0$. By rescaling the electromagnetic potential ϕ, we can always make $\beta = 1$. Using that $\sigma_A{}^2 = a_A{}^2$ we arrive finally at the well-known conform stationary solution

$$V := \frac{1}{1+\phi} = 1 + \sum_{A=1}^{N} \frac{e_A}{R_A}, \quad \xi' = 0.$$

The explicit metric for the last conform stationary solution in the case $N = 2$ was obtained by Perjes,[7] Parker *et al.*,[8] and Kobishe and Parker.[9] We get a static solution by taking the rotation parameter a_A equal to zero for each source.

To end this section we want to discuss the limitations of this method to generate black hole solutions. As Neugebauer and Kramer[1] have pointed out (see also Ref. 10), due to the restriction given by Eq. (7) we can not generate Kerr–Newmann black holes. However, if the background is flat space and several of the N solitions have no electromagnetic charges ($\beta_M = 0$ for some or all M) then we can show that the two possibilities $\lambda_M = (\bar{\lambda}_M)^{-1}$ or $|\lambda_M| = |\lambda_{\dot{M}}| = 1$ are present. The second one enables us to generate Kerr black holes. So we can generate from flat space—via Eqs. (15)–(17)—the superposition of N particles placed on the z axis: M of them being Kerr black holes and N–M being hyperextreme Kerr–Newmann particles. In particular we can obtain the potential associated with a Schwarzschild solution placed between two hyperextreme Kerr–Newmann particles. This configuration could be an approximate model of a mass endowed with multipolar electromagnetic moments.

ACKNOWLEDGMENTS

The author would like to acknowledge Dr. D. Kramer and Dr. G. Neugebauer for their comments on this work, and also the FSU, Jena, where part of this work was done, for its hospitality.

The author acknowledges the financial support received from the Consejo Nacional de Ciencia y Tecnología and from the Sistema Nacional de Investigadores.

[1] G. Neugebauer and D. Kramer, J. Phys. A: Math. Gen. **16**, 1927 (1983).
[2] G. A. Aleksejev, Proceedings of General Relativity and Gravitation 10, Padova, Italy (1983).
[3] G. A. Aleksejev, Proceedings of General Relativity and Gravitation 11, Stockholm, Sweden (1986).
[4] F. Ernst, Phys. Rev. **168**, 1415 (1968).
[5] D. Kramer, Class. Quant. Gravit. **1**, L45 (1984).
[6] In this notation, $\dot{A}$ does not mean the use of spinors.
[7] Z. Perjes, Phys. Rev. Lett. **27**, 1668 (1971).
[8] L. Parker, R. Ruffini, and D. Wilkins, Phys. Rev. D **7**, 2874 (1973).
[9] R. Kobiske and L. Parker, Phys. Rev. D **10**, 2321 (1974).
[10] G. Neugebauer and D. J. Kramer, in *Galaxies Axisymmetric System and Relativity*, edited by M. MacCallum (Cambridge U.P., Cambridge, 1985).

A note on the existence of a solution for a model of Mc Laughlin, Moloney, and Newell for optical ring cavities

Frank Merle
Ecole Normale Superieure, 45 Rue d'Ulm, 75230 Paris, France

(Received 12 April 1988; accepted for publication 2 February 1989)

A problem that arises in a model introduced by Mc Laughlin, Moloney, and Newell [Phys. Rev. Lett. **51**, 75 (1983)] concerning the theory of optical ring cavities is solved. Namely, the equation $i\,\partial v/\partial t = \Delta v - |v|^2 v + \epsilon_0 |v|^4 v$, $\epsilon_0 > 0$, $v(0,\cdot) = \varphi(\cdot)$, where $v(t,x)$: $\mathbb{R}\times\mathbb{R}^2 \to \mathbb{C}$, is considered. The solution of this equation is denoted by $v_\varphi(t,x)$. In addition, let $T\in\mathbb{R}$, $\Theta\in\mathbb{R}$, $0 < R < 1$, and $a(\cdot)$: $\mathbb{R}^2 \to \mathbb{C}$. It is proved that there exists a φ such that $a(\cdot) + Rv_\varphi(T,\cdot) = \varphi(\cdot)$. Finally, the properties of such a φ are examined and various extensions are given.

I. INTRODUCTION

In this paper, we show the existence of a fixed point for a map involving the solution of the equation

$$\begin{aligned} i\frac{\partial v}{\partial t} &= -\Delta v - |v|^2 v + \epsilon_0 |v|^4 v, \\ v(0,\cdot) &= \varphi(\cdot), \end{aligned} \tag{1.1}$$

where Δ is the Laplace operator on $\mathbb{R}^2$, $\epsilon_0 > 0$ is given, v is a function from $\mathbb{R}\times\mathbb{R}^2$ to $\mathbb{C}$, and φ is a suitable function from $\mathbb{R}^2$ to $\mathbb{C}$.

More precisely, under appropriate conditions on φ, (1.1) has a unique solution in $\mathscr{C}(\mathbb{R},H^1)$. We denote this solution by $v_\varphi(t,x)$ and we define $v_\varphi(t)$ to be the function $x \to v_\varphi(t,x)$. Let $T\in\mathbb{R}$, $\Theta\in\mathbb{R}$, $0 < R < 1$ be given and $a(\cdot)$: $\mathbb{R}^2 \to \mathbb{C}$ satisfying suitable conditions. The problem is to find φ such that $a + Re^{i\Theta} v_\varphi(T) = \varphi$ [(P.1)].

Such a problem has physical motivations. It arises in the theory of optical ring cavities (see Mc Laughlin, Moloney, and Newell[1] or Ref. 2 for a detailed discussion).

A laser beam in such a cavity is represented by the infinite-dimensional map (G_n), where G_n satisfies

$$\begin{aligned} &\text{for } n\geqslant 0, \quad i\frac{\partial G_n}{\partial t} = -\Delta G_n + N(|G_n|^2)G_n, \quad \text{for } t\in[0,T], \\ &G_0(0,\cdot) = \varphi(\cdot) \quad \text{and} \quad \text{for } n\geqslant 1,\ G_n(0,\cdot) = a(\cdot) + Re^{i\Theta} G_{n-1}(T,\cdot), \end{aligned} \tag{1.2}$$

where $a(\cdot)$ looks like a Gaussian and R is a factor of loss $(0 < R < 1)$. For N, we take a saturated nonlinearity, namely, $N(x) = -x + \epsilon_0 x^2$, where ϵ_0 is non-negative and small. We then look for a fixed point of (1.2). This means that we seek (G_n) such that $\forall n$, $G_n(t,\cdot) = G_0(t,\cdot)$, for $t\in[0,T]$. That is, we seek φ such that $F(\varphi) = \varphi$, where $F(\varphi) = a + Re^{i\Theta} v_\varphi(T)$.

We will prove that (P.1) has a solution by applying the Schauder theorem.

The paper is organized as follows: In Sec. II, we find a suitable set K such that $F(K) \subset K$. In Sec. III, we state and prove our main result. We further derive some results on φ satisfying $F(\varphi) = \varphi$ and on the sequence G_n. In Sec. IV, we examine slightly more general situations and make some further comments.

Let us first establish some notational conventions and briefly recall some results on $v_\varphi(\cdot)$.

All function spaces are defined on $\mathbb{R}^2$. We denote the function space $H^1 \cap \{\varphi; \int |X|^2 |\varphi|^2 < +\infty\}$ by X, and we consider it to have its natural topology: $|\varphi|_X^2 = |\varphi|_{H^1}^2 + \int |x|^2 |\varphi|^2$.

Furthermore, we suppose that $a(\cdot)\in X$.

In Ref. 3, Ginibre and Velo show that for $\varphi\in X$, (1.1) has a unique solution in $\mathscr{C}(\mathbb{R},X)$. In addition, for $t\in\mathbb{R}$,

$$|v_\varphi(t)|_{L^2} = |\varphi|_{L^2}; \tag{1.3}$$

$$E(v_\varphi(t)) = |\nabla v_\varphi(t)|_{L^2}^2 - \frac{1}{2}\int |v_\varphi(t)|^4 + \frac{\epsilon_0}{3}\int |v_\varphi(t)|^6 = E(\varphi); \tag{1.4}$$

$$\begin{aligned} &\frac{d}{dt}\int |x|^2 |v_\varphi(t)|^2 = -4\,\mathrm{Im}\int r\bar{v}_\varphi(t)\frac{d}{dr}(v_\varphi(t)), \quad \text{where } r = |x|, \\ &\frac{d^2}{dt^2}\int |x|^2 |v_\varphi(t)|^2 = 2|\nabla v_\varphi(t)|_{L^2}^2 - \frac{1}{2}\int |v_\varphi(t)|^4 + 2\frac{\epsilon_0}{3}\int |v_\varphi(t)|^6. \end{aligned} \tag{1.5}$$

Let us set

$$E_+(\varphi) = |\nabla\varphi|_{L^2}^2 + \frac{\epsilon_0}{3}|\varphi|_{L^6}^6, \quad \text{for } \varphi\in X.$$

 0022-2488/89/071643-05$02.50

II. CONSTRUCTION OF AN APPROPRIATE STABLE SET FOR F

Since F is defined on X and the immersion of X in L^2 is compact (see Sec. III), we look for a bounded convex set $K \subset X$ such that $F(K) \subset K$. The first idea is to consider sets of the form

$$\left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E(\varphi) \leqslant k_1,\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\},$$

where $k_0, k_1, k_2 \in \mathbb{R}^+$, since we have information on the $|v_\varphi(T)|_{L^2}$, $E(v_\varphi(T))$, $\int |x|^2 |v_\varphi(T)|^2$. However, such a set is not convex because $E(\cdot)$ is not convex. We are therefore led to look for sets conserved by F of the form

$$\left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1,\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\}.$$

In contrast with the former set, such a set is convex. Let us observe that we could also seek such a set K of the form

$$\left(\mathrm{Con}\left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E(\varphi) \leqslant k_1\right\}\right) \cap \left\{\varphi \in X;\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\},$$

where Con designates the closed convex hull (this will be discussed in Remark 3.1 below).

To find K, we use (1.3)–(1.5) and the fact that $0 < R < 1$. We proceed in several steps: First, we find a set $\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0\}$ stable by F. Then we look for a set of the form

$$\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1\}$$

stable by F. Finally, we find a set conserved by F of the form we want.

Proposition 2.1: There exists k_0 such that the set $\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0\}$ is stable under F.

Proof: We have for $\varphi \in X$

$$|F(\varphi)|_{L^2} = |a + Re^{i\Theta} v_\varphi(T)|_{L^2} \leqslant |a|_{L^2} + R\,|v_\varphi(T)|_{L^2}.$$

We thus derive from (1.3) that

$$|F(\varphi)|_{L^2} \leqslant |a|_{L^2} + R\,|\varphi|_{L^2}. \tag{2.1}$$

If we assume that $|a|_{L^2} + Rk_0 \leqslant k_0$, then $\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0\}$ is stable under F. Therefore, the choice of $k_0 = |a|_{L^2}/(1-R)$ is suitable. Hence Proposition 2.1 is proved.

Now, we want to compare $E_+(\varphi)$ and $E(\varphi)$ for φ such that $|\varphi|_{L^2} \leqslant k_0$.

Lemma 2.1: (i) $E(\varphi) \leqslant E_+(\varphi)$.

(ii) There exists $c = c(k_0)$ such that, for $|\varphi|_{L^2} \leqslant k_0$, we have $E_+(a + Re^{i\Theta}\varphi) \leqslant c + RE(\varphi)$.

Proof: Part (i) follows from the definition of $E(\cdot)$ and $E_+(\cdot)$.

(ii) $E_+(a + Re^{i\Theta}\varphi) = |\nabla a + Re^{i\Theta}\nabla\varphi|_{L^2}^2 + \epsilon_0/3\,|a + Re^{i\Theta}\varphi|_{L^6}^6$.

On one hand, we note that

$$|\nabla a + Re^{i\Theta}\nabla\varphi|_{L^2}^2 \leqslant |\nabla a|_{L^2}^2 + 2R\,|\nabla a|_{L^2}|\nabla\varphi|_{L^2} + R^2|\nabla\varphi|_{L^2}^2 \leqslant c + R\,|\nabla\varphi|_{L^2}^2,$$

where c is appropriately chosen. On the other hand it follows after some calculation that

$$|a + Re^{i\Theta}\varphi|_{L^6}^6 \leqslant (R\,|\varphi|_{L^6} + |a|_{L^6})^6.$$

Thus by the same argument as before, $|a + Re^{i\Theta}\varphi|_{L^6}^6 \leqslant R^5|\varphi|_{L^6}^6 + c$ $(R^5 > R^6)$. Therefore we have

$$E_+(a + Re^{i\Theta}\varphi) \leqslant c + R\,|\nabla\varphi|_{L^2}^2 + \epsilon_0 R^5/3|\varphi|_{L^6}^6$$
$$\leqslant c + R(|\nabla\varphi|_{L^2} - 1/2|\varphi|_{L^4}^4 + \epsilon_0/3|\varphi|_{L^6}^6) + \epsilon_0(R^5 - R)/3|\varphi|_{L^6}^6 + R/2|\varphi|_{L^4}^4.$$

In addition, the Hölder inequality implies that $|\varphi|_{L^4}^4 \leqslant |\varphi|_{L^2}|\varphi|_{L^6}^3 \leqslant k_0|\varphi|_{L^6}^3$. Finally, we have

$$E_+(a + Re^{i\Theta}\varphi) \leqslant c + RE(\varphi) - \epsilon_0(R - R^5)/3|\varphi|_{L^6}^6 + c|\varphi|_{L^6}^3 \leqslant c + RE(\varphi),$$

for c which depends on k_0 (since $R - R^5 > 0$). Thus Lemma 2.1 is established.

We then easily derive the following proposition from Lemma 2.1.

Proposition 2.2: There is a k_1 such that the set

$$\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1\}$$

is stable by F.

Proof: We claim it is a consequence of Lemma 2.1 and (1.4). From Lemma 2.1, there is a c such that, for $|\varphi|_{L^2} \leqslant k_0$, we have

$$E(\varphi) \leqslant E_+(\varphi),$$
$$E_+(a + Re^{i\Theta}\varphi) \leqslant c + RE(\varphi).$$

Since $|v_\varphi(T)|_{L^2} = |\varphi|_{L^2} \leqslant k_0$,

$$E_+(F(\varphi)) = E_+(a + Re^{i\Theta}v_\varphi(T)) \leqslant c + RE(v_\varphi(T)).$$

From (1.4) and Lemma 2.1, it follows that

$$E_+(F(\varphi)) \leqslant c + RE(\varphi) \leqslant c + RE_+(\varphi), \tag{2.2}$$

where c depends on k_0. From (2.2) we derive the existence of k_1 such that $|\varphi|_{L^2} \leqslant k_0$ and $E_+(\varphi) \leqslant k_1$ implies that $E_+(F(\varphi)) \leqslant k_1$. Proposition 2.1 ends the proof.

Now, let us turn to a bound for $|\nabla v_\varphi(t)|_{L^2}$, for $t \in \mathbb{R}$ involving $|\varphi|_{L^2}$, $E_+(\varphi)$.

Lemma 2.2: Assume that $|\varphi|_{L^2} \leqslant k_0$. Then, for $t \in \mathbb{R}$,

$$|\nabla v_\varphi(t)|_{L^2}^2 \leqslant c + E_+(\varphi), \quad \text{where } c = c(k_0).$$

Proof: From (4), we have $E(v_\varphi(t)) = E(\varphi) \leqslant E_+(\varphi)$. Thus

$$|\nabla v_\varphi(t)|_{L^2}^2 - \frac{1}{2}\int |v_\varphi(t)|^4 + \frac{\epsilon_0}{3}\int |v_\varphi(t)|^6 \leqslant E_+(\varphi).$$

Using the same argument as before (in Lemma 2.2), we derive that

$$|\nabla v_\varphi(t)|_{L^2}^2 + \frac{\epsilon_0}{6}\int |v_\varphi(t)|^6 \leqslant c + E_+(\varphi),$$

where $c = c(k_0)$. This yields the result.

Now, we can construct the stable set we wanted.

Proposition 2.3: There are k_0, k_1, k_2 such that

$$\left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1,\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\}$$

is stable under F.

Proof: From Proposition 2.2, we know that there are k_0, k_1 such that the set

$$\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1\}$$

is conserved by F.

Let φ be such that $|\varphi|_{L^2} \leqslant k_0$ and $E_+(\varphi) \leqslant k_1$. From (1.5), we derive that

$$\frac{d}{dt}\int |x|^2 |v_\varphi(t)|^2 = -4 \operatorname{Im} \int r\bar{\varphi}\varphi_r + \int_0^t \frac{d^2}{dt^2}\int |x|^2 |v_\varphi(t)|^2.$$

Thus, as a consequence of Lemma 2.3, (1.5), and the Hölder inequality for the term $\int r\bar{\varphi}\varphi_r$, we have, for $t \in [0,T]$,

$$\frac{d}{dt}\int |x|^2 |v_\varphi(t)|^2 \leqslant c + c\left(\int |x|^2 |\varphi|^2\right)^{1/2}$$

where $c = c(T)$. Therefore, by integrating we obtain

$$\int |x|^2 |v_\varphi(T)|^2 \leqslant c + c\left(\int |x|^2 |\varphi|^2\right)^{1/2} + \int |x|^2 |\varphi|^2.$$

On the other hand, by the same argument as in the proof of Proposition 2.2, $\int |x|^2 |F(\varphi)|^2 \leqslant c + R^{3/2} \int |x|^2 |v_\varphi(T)|^2$. Hence it follows that

$$\int |x|^2 |F(\varphi)|^2 \leqslant c + R\int |x|^2 |\varphi|^2, \tag{2.3}$$

where c depends on k_0 and k_1. This ends the proof of Proposition 2.3.

Remark 2.1: As mentioned above, we can seek a stable set of the form

$$\operatorname{Con}\left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E(\varphi) \leqslant k_1\right\} \cap \left\{\varphi \in X;\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\}$$

(where Con designates the closed convex hull in L^2 as before). The proof of conservation under F of sets of this type is the same except for Proposition 2.2. Indeed, it is based on two lemmas.

Lemma 2.2': Let $k_0 > 0$. Then there exists $c = c(k_0)$ such that $\forall k_1$,

$$\operatorname{Con}\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E(\varphi) \leqslant k_1\} \subset \{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E(\varphi) \leqslant k_1 + c\}.$$

Lemma 2.2": Let $k_0 > 0$. Then there exists $c = c(k_0)$ such that,

$$\text{for } |\varphi|_{L^2} \leqslant k_0, \quad E(a + Re^{i\Theta}\varphi) \leqslant c + RE(\varphi).$$

III. THE MAIN RESULT

In this section, we conclude the existence of a fixed point for F and give some applications of the estimates obtained in Sec. II. More precisely, for φ such that $F(\varphi) = \varphi$, we estimate $|\varphi|_X$ and state continuity result on φ. In addition, we derived a uniform bound for the sequence G_n, where (G_n) is defined by (1.2).

Theorem: Let $0 < R < 1$, $T \in \mathbb{R}$, $\Theta \in \mathbb{R}$, and $a(\cdot) \in X$. Then there exists $\varphi \in X$ such that

$$F(\varphi) = a + Re^{i\Theta} v_\varphi(T) = \varphi.$$

Proof: As mentioned in the Introduction, we apply Schauder's fixed point theorem: Let K be a convex compact subset of L^2 and let F be defined on K. Suppose that $F_{|K}$ is continuous in L^2 and $F(K) \subset K$. Then there exists $\varphi \in K$ such that $F(\varphi) = \varphi$.

Let us set

$$K = \left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1,\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\},$$

where k_0, k_1, and k_2 are defined in Sec. III. Since $K \subset X$, F is defined on K and, from Proposition 2.3, $F(K) \subset K$. In addition, we can easily check that K is convex. Thus to prove the theorem it suffices to show that K is a compact subset of L^2 and that $F_{|K}$ is continuous in the L^2 norm.

(i) First, we prove that K is compact in L^2. Let us consider a sequence (φ_n) in K. We can easily check that

$$\forall \rho > 0,\ \forall n,\ \int_{|x|>\rho} \varphi_n^2 \leqslant \frac{k_2}{\rho^2}.$$

Since $\forall n$, $|\varphi_n|_{L^2} \leqslant k_0$ and $|\nabla\varphi_n|_{L^2} \leqslant k_1$, by a classical argument there exists φ_∞ and a subsequence (which we also denote by φ_n) such that $\varphi_n \to \varphi_\infty$ in L^2. In particular, $\varphi_n \rightharpoonup \varphi_\infty$ weakly in X. Therefore $\varphi_\infty \in X$ and $|\varphi_\infty|_{L^2} \leqslant k_0$. Since $E_+(\cdot)$ is convex and strongly continuous in H^1,

$$E_+(\varphi_\infty) \leqslant \limsup E_+(\varphi_n) \leqslant k_1.$$

By the same argument,

$$\int |x|^2 |\varphi_\infty|^2 \leqslant \limsup \int |x|^2 |\varphi_n|^2 \leqslant k_2.$$

Thus $\varphi_\infty \in K$ and it follows that K is a compact set in L^2.

(ii) Let us prove that the restriction of F to $K(F_{|K})$ is continuous in the L^2 norm. We claim that it is a consequence of the inequalities proved by Ginibre and Velo.[3] First, let us remark that there is a c_K such that $\forall t \in \mathbb{R}$, $\forall \varphi \in K$, $|v_\varphi(t)|_{H^1} \leqslant c_K$ [(1.4) and Lemma 2.2]. We recall that if $v_\varphi(\cdot)$ is the solution of (1.1), we have[3]

$$\forall t,\ v_\varphi(t) = U(t)\varphi - i\int_0^t U(t-s)(-|v_\varphi(s)|^2 v_\varphi(s) + \epsilon_0 |v_\varphi(s)|^4 v_\varphi(s))ds, \tag{3.1}$$

where $U(\cdot)$ is the group generated by $i\Delta$ (the Schrödinger group). Moreover, there is a c such that

$$\forall t,\ \forall v \in L^4,\ |U(t)v|_{L^4} \leqslant c|v|_{L^{4/3}}/t^{1/2},$$

$$\forall t,\ \forall v \in H^1,\ |U(t)v|_{H^1} = |v|_{H^1}.$$

Now, consider $\varphi^1,\varphi^2\in K$. First, we estimate $|v_{\varphi^1}(t)-v_{\varphi^2}(t)|_{L^4}$, which then yields an estimate of $|v_{\varphi^1}(t)-v_{\varphi^2}(t)|_{L^2}$. Using the definitions of $v_{\varphi^1}(\cdot)$ and $v_{\varphi^2}(\cdot)$ and the inequalities above, after some computation (see Ref. 3 for similar computations), we obtain

$$|v_{\varphi^1}(t)-v_{\varphi^2}(t)|_{L^4}$$
$$\leqslant |U(t)(\varphi^1-\varphi^2)|_{L^4}+c\int_0^t \frac{1}{(t-s)^{1/2}}|v_{\varphi^1}(s)-v_{\varphi^2}(s)|_{L^4}(|v_{\varphi^1}(s)|_{L^4}^2+|v_{\varphi^2}(s)|_{L^4}^2+|v_{\varphi^1}(s)|_{L^\infty}^4+|v_{\varphi^2}(s)|_{L^\infty}^4\,ds).$$

It follows from the Sobolev inequalities and the properties of $U(\cdot)$ that

$$\forall t,\ |v_{\varphi^1}(t)-v_{\varphi^2}(t)|_{L^4}$$
$$\leqslant c_K|\varphi^1-\varphi^2|_{L^2}^{1/2}+c_K\int_0^t \frac{1}{(t-s)^{1/2}}|v_{\varphi^1}(s)-v_{\varphi^2}(s)|_{L^4}\,ds.$$

We denote $\sup_{t\in[0,t]}|v_{\varphi^1}(t)-v_{\varphi^2}(t)|_{L^4}$ by $M(t)$. $\forall t$, $M(t)\leqslant c_K|\varphi^1-\varphi^2|_{L^2}^{1/2}+c_K\int_0^t M(s)ds$. As a consequence of the Gronwall lemma, we have $|v_{\varphi^1}(T)-v_{\varphi^2}(T)|_{L^4}\leqslant c_K|\varphi^1-\varphi^2|_{L^2}^{1/2}e^{c_K}\leqslant c_K|\varphi^1-\varphi^2|_{L^2}^{1/2}$. Furthermore, using (1.5) and Lemma 2.3 we can check that

$$\int(|x|^2+1)|v_{\varphi^1}(T)-v_{\varphi^2}(T)|^2\leqslant c_K.$$

Then,

$$|v_{\varphi^1}(T)-v_{\varphi^2}(T)|_{L^{3/2}}\leqslant c_K(1/\sqrt{|x|^2+1}\in L^6).$$

Thus, using the Hölder inequality, we obtain

$$|v_{\varphi^1}(T)-v_{\varphi^2}(T)|_{L^2}\leqslant |v_{\varphi^1}(T)-v_{\varphi^2}(T)|_{L^{3/2}}^{3/5}|v_{\varphi^1}(T)-v_{\varphi^2}(T)|_{L^4}^{2/5}\leqslant c_K|\varphi^1-\varphi^2|_{L^2}^{1/5}.$$

Therefore, we easily derive that

$$|F(\varphi^1)-F(\varphi^2)|_{L^2}\leqslant c_K|\varphi^1-\varphi^2|_{L^2}^{1/5}. \tag{3.2}$$

It follows that $F_{|K}$ is continuous in L^2. This completes the proof of the theorem.

Remark 3.1: In the proof above, L^2 can be replaced by L^p with p such that $p\in[2,+\infty)$.

Remark 3.2: We may assume that $\epsilon_0=1$ to prove that F has a fixed point. Indeed, if we denote by φ the solution of

$$i\frac{\partial v}{\partial t}=-\Delta v-|v|^2v+|v|^4v,$$
$$v(0,\cdot)=\varphi(\cdot),$$
$$a(\epsilon_0^{1/2},\cdot)\epsilon_0^{1/2}+Re^{i\Theta}v_\varphi(T/\epsilon_0^{3/2},\cdot)=\varphi(\cdot).$$

Then, $\epsilon_0^{-1/2}\varphi(\epsilon_0^{-1/2},\cdot)=\Psi(\cdot)$ is a solution of

$$i\frac{\partial v}{\partial t}=-\Delta v-|v|^2v+\epsilon_0|v|^4v,$$
$$v(0,\cdot)=\Psi(\cdot),$$
$$a(\cdot)+Re^{i\Theta}v_\Psi(T,\cdot)=\Psi(\cdot).$$

We now give estimates of the norm $|\varphi|_X$ for φ satisfying $F(\varphi)=\varphi$. Indeed, we show that, for a fixed R, $|\varphi|_X$ is estimated by $|a|_X$.

Proposition 3.1: For fixed $0<R<1$, there exist h_R^1,h_R^2 such that h_R^1,h_R^2 are continuous functions and $\forall x>0$, $0<h_R^1(x)\leqslant h_R^2(x)$;

$$\lim_{x\to 0}h_R^i=0,\ \lim_{x\to\infty}h_R^i=\infty,\ \text{for } i=1,2;$$
$$h_R^1(|a|_X)\leqslant|\varphi|_X\leqslant h_R^2(|a|_X),$$

for φ a solution of the problem (P.1) in X.

Proof: Let us consider φ such that $a+Re^{i\Theta}v_\varphi(T)=\varphi$. On one hand, we have

$$|a|_{L^2}+R|v_\varphi(T)|_{L^2}\leqslant|\varphi|_{L^2}.$$

Thus (3) yields

$$|\varphi|_{L^2}\leqslant|a|_{L^2}/(1-R).$$

On the other hand,

$$|a|_{L^2}\leqslant R|v_\varphi(T)|_{L^2}+|\varphi|_{L^2}.$$

Therefore $|a|_{L^2}/(1+R)\leqslant|\varphi|_{L^2}$. Moreover, for $|a|_X\leqslant c_K$, the calculations in Sec. II [namely, (2.2)] yield the fact that $E_+(\varphi)$ is bounded by a constant that goes to zero as c_K goes to zero. Using (2.3), we obtain the same result for $\int|x|^2|\varphi|^2$.

Proposition 3.1 then follows.

Another application of the calculations in Sec. II is a uniform bound for G_n, where G_n is a sequence satisfying (1.2). Namely, $\forall G_0(\cdot)\in X$ there is a $c>0$ such that $\forall n$, $\forall t\in[0,T]$, $|G_n(t)|_X\leqslant c$.

Proposition 3.2: Consider a sequence G_n that satisfies (1.2). Then, we have

$$\text{(i)}\ \lim_{n\to+\infty}|G_n(t)|_{L^2}\leqslant k_0,$$

$$\text{(ii)}\ \lim_{n\to+\infty}E_+(G_n(t))\leqslant k_1,$$

$$\text{(iii)}\ \lim_{n\to+\infty}\int|x|^2|G_n(t)|^2\leqslant k_2,$$

where k_0, k_1, and k_2 are the constants defined in Sec. II.

Proof: (i) is easily derived from (2.1).

Then using computations similar to those in Sec. II, $\forall\epsilon>0$, *for* n large, we have

$$E_+(G_n(0))\leqslant c(k_0)+\epsilon+RE_+(G_n(0)).$$

Thus (ii) follows [since $k_1=c(k_0)/(1-R)$]. The same argument allows us to prove (iii).

Remark 3.3: It is easy to see that if we assume that

$(\epsilon_n, R_n, \Theta_n, T_n) \to (\epsilon_0, R, \Theta, T)$ and $a_n \to a$ in X as $n \to +\infty$, and if we donate by φ_n a solution of the problem in X,

$$i\frac{\partial v}{\partial t} = -\Delta v - |v|^2 v + \epsilon_n |v|^4 v,$$

$$v(0,\cdot) = \varphi(\cdot),$$

$$F_n(\varphi) = a_n(\cdot) + R_n e^{i\Theta_n} v_\varphi(T_n,\cdot) = \varphi(\cdot),$$

then the set $\{\varphi_n, n \in \mathbb{N}\}$ is precompact in L^2 and the limit points of the sequence φ_n are solutions of (P.1) [namely, $a + Re^{i\Theta} v_\varphi(T) = \varphi$].

To show this, using the same calculations as in Sec. II [(2.1)–(2.3)], we derive that there are k_0, k_1, and k_2 such that

$$K = \left\{\varphi \in X;\ |\varphi|_{L^2} \leqslant k_0,\ E_+(\varphi) \leqslant k_1,\ \int |x|^2 |\varphi|^2 \leqslant k_2\right\}$$

is conserved by F_n, $\forall n$. Therefore $\forall n$, $\varphi_n \in K$ and it is easy to conclude.

IV. FURTHER RESULTS AND COMMENTS

First, we generalize the theorem slightly.

(i) We remark that we can suppose $R > 1$ in the theorem. Indeed, using the time-reversible character of the Schrödinger equation, we may find φ such that $a + Re^{i\Theta} v_\varphi(T) = \varphi$. Indeed, this is equivalent to finding Ψ [$\Psi = v_\varphi(T)$] such that

$$-e^{-i\Theta} a/R + v_\Psi(-T)\, e^{-i\Theta}/R = \Psi.$$

(ii) Furthermore, the theorem is still true if we consider the solution of the equation

$$i\frac{\partial v}{\partial t} = -\Delta v - |v|^2 v + f(v),$$

$$v(0,\cdot) = \varphi(\cdot),$$

where $v(t,x)$: $\mathbb{R} \times \mathbb{R}^N \to \mathbb{C}$ and f satisfies appropriate conditions (see Ref. 3). Namely, we suppose that (a) $f(0) = 0$, f is a continuously differentiable function from $\mathbb{C}$ to $\mathbb{C}$, and there exists g: $\mathbf{R} \to \mathbf{R}$ such that $\forall z, f(z) = g(|z|^2)z$; (b) there exist real numbers p_1, p_2, r_1, r_2, and c such that

$$1 \leqslant p_1 \leqslant p_2 < 2^* - 1, \quad 2 \leqslant r_1 \leqslant r_2 < 2^*,$$

$$r_2/\bar{r}_1 \leqslant p_1 \leqslant p_2 \leqslant r_1/\bar{r}_2,$$

$$\forall z, \left|\frac{\partial f}{\partial z(z)}\right| + \left|\frac{\partial f}{\partial \bar{z}(z)}\right| \leqslant c(|z|^{p_1 - 1} + |z|^{p_2 - 1}),$$

where $2^* = 2N/(N-2)$ if $N > 2$, $2^* = +\infty$ (otherwise $2^* = +\infty$) and $\bar{r}$ is the conjugate exponent of r ($1/\bar{r} + 1/r = 1$); and (c) there exists c and $p_3 < 1 + 4/N$ such that

$$\forall x \in \mathbb{R}, \int_0^x f(s)ds \geqslant -c - cx^{p_3}.$$

Then there exists $\varphi \in X$ such that $a + Re^{i\Theta} v_\varphi(T) = \varphi$ (the proof is the same as before).

Finally, let us state some open questions concerning problem (P.1).

First, an important problem is the question of uniqueness of solutions of (P.1).

Additionally, since for $R \neq 1$ the problem (P.1) has a solution, we may ask if it also has a solution for $R = 1$. In other words, $\forall \Theta$, $\forall T$, $\forall a(\cdot) \in X$ is there φ such that $a + e^{i\Theta} v_\varphi(T) = \varphi$?

Lastly, is there a solution of (P.1) when we take a non-saturated linearity ($\epsilon_0 = 0$)? For $|a|_X$ small, we can check that the problem (P.1) still has one solution.

[1] D. W. Mc Laughlin, J. V. Moloney, and A. C. Newell, Phys. Rev. Lett. **51**, 75 (1983).

[2] A. J. Lichtenberg and M. A. Lieberman, *Regular and Stochastic Motion* (Springer, Berlin, 1983).

[3] J. Ginibre and G. Velo, J. Funct. Anal. **32**, 1 (1979).

Exact solvability of the Mullins nonlinear diffusion model of groove development

P. Broadbridge
Mathematics Department, La Trobe University, Bundoora, Victoria 3083, Australia

(Received 5 October 1988; accepted for publication 2 February 1989)

The Mullins equation for the development of a surface groove by evaporation–condensation is $y_t = y_{xx}/1 + y_x^2$. It is pointed out that this is the equation of the potential for the field variable Θ satisfying the nonlinear diffusion equation $\Theta_t = \partial_x [\Theta_x/1 + \Theta^2]$. The latter has already been solved exactly, with boundary conditions corresponding exactly to those specified by Mullins. The depth of a groove at a grain boundary is predicted exactly without first making the linear (small-slope) approximation. For some types of initial data, the Cauchy problem may be solved for some related equations.

I. INTRODUCTION

In the development of surface grooves by the mechanism of evaporation–condensation,[1] the theoretical profile satisfies the nonlinear diffusion equation

$$y_t = [D(0)/(1 + y_x^2)]y_{xx}, \tag{1}$$

with $D(0)$ constant and $(x,t) \in \mathbb{R} \times [0, \infty)$. Mullins[1] considered the evolution of a single groove originating at $x = 0$ from a persistent grain boundary intersecting the surface of a hot polycrystal. The groove is assumed to be symmetric about $x = 0$, so that the domain may be reduced to the half-line $x \geqslant 0$. In this case, the appropriate boundary conditions are

$$y_x(0,t) = C_0 \text{ (constant)} \tag{2a}$$

and

$$y_x(x,t) \to 0 \text{ as } x \to \infty. \tag{2b}$$

The initial condition is taken to be

$$y(x,0) = 0. \tag{3}$$

In order to obtain an analytic solution, Mullins[1] replaced (1) by the linear diffusion equation

$$y_t = a y_{xx}, \tag{4}$$

which is valid in the small-slope approximation $y_x^2 \ll 1$. The main point of this paper is to show that the full nonlinear boundary value problem (1)–(3) is exactly solvable, even without the small-slope approximation. In addition, it is pointed out that one may exactly solve a wider class of equations

$$y_t = f(y_x)y_{xx}, \tag{5}$$

subject to the same initial-boundary conditions (2) and (3). Finally, the exact solvability of the Cauchy problem for the examples of Eq. (5) and some generalizations recently investigated by Kitada[2] and by Kitada and Umehara[3] will be addressed.

II. TRANSFORMATION TO THE STANDARD NONLINEAR DIFFUSION PROBLEM

Following the simple transformation

$$\Theta = C_0^{-1} y_x, \; y = C_0 \int_{\infty}^{x} \Theta(x_1,t)dx_1, \tag{6}$$

and taking note of (2b), Eq. (5) becomes

$$\int_{\infty}^{x} \Theta_t \, dx = D(\Theta)\Theta_x,$$

where $D(\Theta) = f(C_0\Theta)$. By differentiating, we obtain

$$\Theta_t = \partial_x [D(\Theta)\Theta_x]. \tag{7}$$

This is the standard general nonlinear diffusion equation, which has been extensively studied because of its many applications to heat and mass transfer.[4] The boundary condition (2) implies

$$\Theta(0,t) = 1, \tag{8a}$$

$$\Theta(x,t) \to 0 \text{ as } x \to \infty, \tag{8b}$$

and the initial condition is

$$\Theta(x,0) = 0. \tag{9}$$

Fujita[5] showed how to obtain an exact parametric solution to the boundary value problem (7)–(9) when $D(\Theta)$ is of the form

$$D = D(0)/(1 + 2\alpha\Theta + \beta\Theta^2). \tag{10}$$

This includes the form of D arising from the Mullins equation (1), namely, that with $\alpha = 0$ and $\beta = C_0^2 > 0$,

$$D = D(0)/(1 + C_0^2\Theta^2). \tag{11}$$

However, Fujita carried out the solution only for the three special subcases

$\beta = 0, \quad \alpha \neq 0$ (Fujita[6]),

$\beta = \alpha^2 \neq 0$ (Fujita[7]),

$\alpha < 0, \quad \beta > 0, \quad 0 < -\alpha/\beta < 1$ (Fujita[5]).

Although, for practical purposes, Fujita's third subcase requires the existence of a local minimum in the diffusivity function, the same method of solution applies to the current monotonic case (11). The solution, although remaining complicated, contains fewer parameters as a result of α being zero.

Although it at first appears that an extra step [the integration (6)] is required to obtain Mullins' dependent variable y from Fujita's variable Θ, in practice this requires no extra labor. A new variable ψ, which is defined in an intermediate step of Fujita's method,[5] is shown, in the next section, to be closely related to $yt^{-1/2}$ which is a natural scaling in-

 0022-2488/89/071648-04$02.50

variant for the Mullins problem. The solution for ψ can be quite easily obtained from the solution for Θ.

III. THE EXACT SOLUTION TO THE MULLINS PROBLEM

The boundary value problem (1)–(3) is invariant under the one-parameter group of scaling transformations

$$\bar{x} = e^s x, \quad \bar{y} = e^s y, \quad \bar{t} = e^{2s} t. \tag{12}$$

A set of independent invariants for this group is $\{xt^{-1/2}, yt^{-1/2}\}$. Therefore, we seek a self-similar solution[4] of the form

$$\rho = g(\eta),$$

where $\rho = \frac{1}{2} y[D(0)t]^{-1/2}$ and

$$\eta = \tfrac{1}{2} x[D(0)t]^{-1/2}. \tag{13}$$

Equation (1) then reduces to the ordinary differential equation

$$2\left[\rho - \eta \frac{d\rho}{d\eta}\right] = \frac{1}{1 + (d\rho/d\eta)^2} \frac{d^2\rho}{d\eta^2}. \tag{14}$$

The initial-boundary conditions (2) and (3) reduce to

$$\frac{d\rho}{d\eta} = C_0 \text{ and } \lim_{\eta \to \infty} \rho = 0. \tag{15}$$

Now

$$\frac{d\rho}{d\eta} = \frac{\partial y}{\partial x} = C_0 \Theta. \tag{16}$$

In his solution to the problem (6)–(10), Fujita[5] introduced an intermediate variable ψ, that, in the current case of $\alpha = 0$ and $\beta = \theta_0$, satisfies

$$\frac{d\psi}{d\Theta} = 2\eta \quad \text{[Eq. (19) and (20) of Fujita[5]]},$$

$$\frac{d\psi}{d\Theta} = 0 \text{ at } \Theta = 1 \quad \text{[Eq. (29) of Fujita[5]]},$$

and

$$\psi \to 0 \text{ as } \Theta \to 0 \quad \text{[Eq. (29) of Fujita[5]]}.$$

Hence

$$\tfrac{1}{2}\psi = \int_0^{\Theta} \eta(\Theta_1) d\Theta_1$$

$$= \eta\Theta - \int_{\infty}^{\eta} \Theta(\eta_1) d\eta_1$$

$$= \eta\Theta - C_0^{-1}\rho, \text{ by (6) and (16)}.$$

Rearranging this equation, one obtains

$$\rho = C_0[\eta\Theta - \tfrac{1}{2}\psi]. \tag{17}$$

From Fujita's exact parametric solution

$$[0,1] \ni \theta \to (\eta, \Theta),$$

we have

$$\Theta = C_0^{-1} \tan[F(\theta,\epsilon)] \quad (0 < \Theta \leqslant \Theta_*), \tag{18a}$$

$$\Theta = C_0^{-1} \tan[\tan^{-1} C_0 - F(\theta;\epsilon) + F(\theta_m;\epsilon)] \quad (\Theta_* \leqslant \Theta \leqslant 1), \tag{18b}$$

$$\eta = \epsilon^{-1/2}\{\theta \sin[F(\theta;\epsilon)] + (1 - \theta^2 - \epsilon \ln \theta)^{1/2} \cos[F(\theta;\epsilon)]\} \quad (0 < \Theta \leqslant \Theta_*), \tag{19a}$$

$$\eta = \epsilon^{-1/2}\{\theta \sin[\tan^{-1} C_0 - F(\theta;\epsilon) + F(\theta_m;\epsilon)] - (1 - \theta^2 - \epsilon \ln \theta)^{1/2} \cos[\tan^{-1} C_0 - F(\theta;\epsilon) + F(\theta_m;\epsilon)]\} \quad (\Theta_* \leqslant \Theta \leqslant 1), \tag{19b}$$

$$\psi = 2\epsilon^{-1/2} C_0^{-1} \theta \sec[F(\theta;\epsilon)] \quad (0 < \Theta \leqslant \Theta_*), \tag{20a}$$

$$\psi = 2\epsilon^{-1/2} C_0^{-1} \theta \sec[\tan^{-1} C_0 - F(\theta;\epsilon) + F(\theta_m;\epsilon)] \quad (\Theta_* \leqslant \Theta \leqslant 1), \tag{20b}$$

where the function $F(\theta;\epsilon)$ is given by

$$F(\theta;\epsilon) = \int_0^{\theta} (1 - q^2 - \epsilon \ln q)^{-1/2} dq \tag{21a}$$

$$= \int_0^{\sin^{-1}(\theta)} \left(1 - \frac{\epsilon \ln \sin \phi}{\cos^2 \phi}\right)^{-1/2} d\phi. \tag{21b}$$

The parameters Θ_*, ϵ, and θ_m are defined by

$$\epsilon = [1 - (1 + C_0^2)\theta_m^2]/\ln \theta_m, \tag{22}$$

$$\tan^{-1} C_0 = 2F(1;\epsilon) - F(\theta_m;\epsilon), \tag{23}$$

and

$$\Theta_* = C_0^{-1} \tan[F(1;\epsilon)]. \tag{24}$$

The parameter θ_m may be found by numerically solving the transcendental equation (23). Then, ϵ and Θ_* may by computed from θ_m. Unlike the model considered by Fujita,[5] here $\theta = \theta_m$ corresponds to a spatial boundary since, from (19b) and (22), it is clear that $\eta = 0$ corresponds to $\theta = \theta_m$. Therefore, from (17) and (20b), the value of ρ at the boundary $\eta = 0$ is

$$\rho(0) = -\epsilon^{-1/2} \theta_m \sqrt{1 + C_0^2}$$

or

$$\rho(0) = \frac{-\theta_m[-\ln \theta_m]^{1/2}}{[\theta_m^2 - 1/(1 + C_0^2)]^{1/2}}, \text{ by (22)}. \tag{25}$$

The value of $\rho(0)$ may be determined from the slope at $x = 0$, according to (22), (23), and (25). This then provides a correction to the linear approximation, in which

$$y(0,t) = -2C_0[D(0)t]^{1/2} i\text{erfc}(0),$$

and

$$\rho(0) = C_0\, i\text{erfc}(0) = -C_0 \pi^{-1/2}. \tag{26}$$

For a grain boundary with slope $C_0 = 1$, Eq. (25) predicts that the depth of the developing groove is 16% lower than that suggested by Eq. (26). However, for small to moderate values of C_0, the linear model is quite accurate. For example, with $C_0 = 0.4$, the linear model overestimates $\rho(0)$ by only 3.6%.

In order to evaluate (21b), the function

$$(\ln \sin \phi)/\cos^2 \phi$$

appearing in the integrand, has been represented in the neighborhood of $\phi = \pi/2$, as a series

$$\frac{\ln \sin \phi}{1-\sin^2 \phi} = \frac{\ln(1-\delta)}{\delta(2-\delta)} \quad (\text{where } \delta = 1-\sin\phi)$$
$$= \frac{-1}{2-\delta} - \sum_{j=2}^{\infty} \frac{1}{j} \frac{\delta^{j-1}}{2-\delta}.$$

IV. OTHER EXACTLY SOLVABLE PROBLEMS OF THE MULLINS TYPE

It is clear that given any exactly solvable boundary value problem of the type (7) and (8), there is a related exactly solvable Mullins-type problem [Eqs. (5), (2), and (3)] obtained by the transformation

$$\Theta = y_x, \tag{27a}$$

$$y = \int_{\infty}^{x} \Theta(x,t)\,dx_1, \tag{27b}$$

$$f(y_x) = D(\Theta). \tag{27c}$$

A nonlinear diffusivity $D(\Theta)$ of the Fujita class (10) will yield an exact complicated parametric solution. In another original approach due to Philip,[8] one may propose an admissible explicit form for the exact solution and then deduce the exact form for the diffusivity. In direct contrast to the Fujita models, this approach tends to produce simpler explicit solutions but with more complicated forms for the diffusivity. A collection of exactly solvable models has been produced in this way by Philip.[8] Each of these now leads to an exactly solvable model of the Mullins type, through the simple transformation (27).

Kitada[2] and Kitada and Umehara[3] have rigorously established smoothing properties for the Cauchy initial data problem associated with Eq. (1). In this problem, a prescribed initial surface profile is allowed to evolve freely in time, in the absence of imposed grain boundaries. Equation (1) is considered on the domain $\mathbb{R}\times[0,\infty)$ and with a prescribed initial condition

$$y(x,0) = \alpha(x). \tag{28}$$

The author has not been able to solve the Cauchy problem for the original Mullins equation (1). However, if it is assumed that $y \to 0$ as $x \to \infty$, then under the simple transformation (27), the more general equation (5) transforms to the standard general nonlinear diffusion equation (7). Provided that α is differentiable, the initial condition (28) transforms to

$$\Theta(x,0) = \alpha'(x). \tag{29}$$

In the special case that

$$D(\Theta) = D(0)/(1+\alpha\Theta)^2, \tag{30}$$

the method of Knight and Philip[9] produces an exact parametric solution when $\alpha'(x)$ is an even function monotonically decreasing to zero on $[0,\infty)$. In the context of the smoothing problem considered here, this provides an exact solution when the initial profile is an odd function $\alpha(x)$ which becomes horizontal for large values of $|x|$. For example, the solution given explicitly by Knight and Philip[9] with the initial data

$$\Theta(x,0) = \begin{cases} \Theta_0, & |x|<1, \\ 0, & |x|>1. \end{cases} \tag{31}$$

can be used to describe the smoothing of an initial surface dislocation

$$y(x,0) = \begin{cases} -\Theta_0, & x<-1, \\ x\Theta_0, & |x|\leqslant 1, \\ \Theta_0, & x>1. \end{cases} \tag{32}$$

The governing equation in this case is

$$y_t = [D(0)/(1+\alpha y_x)^2]y_{xx}. \tag{33}$$

Although this is not the same as Mullins' equation (1), it shares many of the same features. Its exact solutions may be used to examine the effect of nonlinearity, to test numerical solution schemes, and to test rigorous estimates for the decay of surface irregularities.

Assuming that the integral in (27b) exists, the transformation (27) relates a conservation field equation

$$\Theta_t = D_x v(t,x,\Theta,\Theta_1,\Theta_2,\ldots,\Theta_n) \tag{34}$$

to an equation for the corresponding potential

$$y_t = v(t,x,y_1,y_2,\ldots,y_{n+1}). \tag{35}$$

Here, a subscript integer j represents the jth spatial derivative, for example,

$$\Theta_j = \frac{\partial^j \Theta}{\partial x^j},$$

and D_x represents the total spatial derivative,

$$D_x v = \frac{\partial v}{\partial x} + \sum_{j=0}^{n} \frac{\partial v}{\partial \Theta_j}\Theta_{j+1}.$$

Among the second-order nonlinear examples of Eq. (34), some have known exact nontrivial solutions and these lead to special solvable second-order nonlinear examples of Eq. (35), which generalizes the Mullins-type equation (5). Some of the exactly solvable nonlinear diffusion equations may incorporate nonlinear convection or spatial heterogeneity.

In a nonlinear diffusion–convection equation, the function v in Eq. (34) takes the form $v = D(\Theta)\Theta_1 + K(\Theta)$ for some functions D and K. The best known exactly solvable example is the Burgers equation,[10] with D constant and K quadratic. In this case, Eq. (35) is known as the potential Burgers equation. Another nonlinear diffusion–convection equation with known exact solutions is the Fokas–Yortsos–Rosen equation,[11] with $D(\Theta)$ as in Eq. (3) and

$$K(\Theta) = \nu/(1+\alpha\Theta) + \beta\Theta + \gamma \quad (\alpha,\nu,\beta,\gamma \text{ constant}).$$

Even if closed-form solutions are not available, a large class of equations of the type (34), subject to boundary and initial conditions (8) and (9), may be solved by the quasianalytic series method of Philip.[12]

Finally, for nonlinear diffusion in a scale-heterogeneous medium,[13] the governing field equation is a conservation equation (34), with

$$v = \lambda(x)D(\Theta)\theta_x + \lambda'(x)G(\Theta),$$

for some positive functions λ, D, and G. The class of such equations that possess Lie–Bäcklund symmetries was found by Broadbridge[14] and for each member of this class, exact solutions may be found for the potentials that satisfy the corresponding class of equations (35).

[1]W. W. Mullins, J. Appl. Phys. **28**, 333 (1957).
[2]A. Kitada, J. Math. Phys. **27**, 1391 (1986); **28**, 2982 (1987).
[3]A. Kitada and H. Umehara, J. Math. Phys. **28**, 536 (1987).
[4]J. Crank, *The Mathematics of Diffusion* (Clarendon, Oxford, 1975); P. L. Sachdev, *Nonlinear Diffusive Waves* (Cambridge U.P., Cambrige, 1987).
[5]H. Fujita, Text. Res. J. **24**, 234 (1954).
[6]H. Fujita, Text. Res. J. **22**, 757 (1952).
[7]H. Fujita, Text. Res. J. **22**, 823 (1952).
[8]J. R. Philip, Aust. J. Phys. **13**, 1 (1960).
[9]J. H. Knight and J. R. Philip, J. Eng. Math. **8**, 219 (1974).
[10]E. R. Benton and G. W. Platzman, Quart. Appl. Math. **30**, 195 (1972).
[11]A. S. Fokas and Y. C. Yortsos, SIAM J. Appl. Math. **42**, 318 (1982); G. Rosen, Phys. Rev. Lett. **49**, 1844 (1982); C. Rogers, M. P. Stallybrass, and D. L. Clements, Nonlinear Anal. Theory Meth. Appl. **7**, 785 (1982); P. Broadbridge and I. White, Water Resour. Res. **24**, 145 (1988). G. C. Sander, J.-Y. Parlange, V. Kühnel, W. L. Hogarth, D. Lockington, and J. P. J. O'Kane, J. Hydrol. **97**, 341 (1988).
[12]J. R. Philip, Soil Sci. **83**, 345 (1957).
[13]J. R. Philip, Aust. J. Soil Res. **5**, 1 (1967).
[14]P. Broadbridge, J. Math Phys. **29**, 622 (1988).

ERRATUM

Erratum: Liouville theorem for the Yang–Mills self-duality equations [J. Math. Phys. 29, 2303 (1988)]

Shahn Majid[a)]
Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138

(Received 13 March 1989; accepted for publication 13 March 1989)

Due to a typographical error, the left-hand side of Eq. (1.5) is missing. Equation (1.5) should read:

$$[F_{\mu\nu},F_{\alpha\beta}] \equiv W(F,F)_{\mu\nu\alpha\beta} + [1/(n-2)](g_{\nu\alpha}F^2_{\mu\beta} - g_{\mu\alpha}F^2_{\nu\beta} - g_{\nu\beta}F^2_{\mu\alpha} + g_{\mu\beta}F^2_{\nu\alpha}), \tag{1.5}$$

[a)] Present address: Department of Mathematics and Computer Science, University of Wales, University College of Swansea, Singleton Park, Swansea SA2 8PP, United Kingdom.

 0022-2488/89/071652-01$02.50