Unbounded partial Hilbert algebras

G. O. S. Ekhaguere
Department of Mathematics, University of Ibadan, Ibadan, Nigeria

(Received 16 August 1988; accepted for publication 22 March 1989)

A notion of an unbounded partial Hilbert algebra is introduced and some properties and examples of such an algebra are furnished. Noncommutative versions of Arens L^ω spaces over partial Hilbert algebras are formulated and shown to be unbounded partial Hilbert algebras. Moreover, necessary and sufficient conditions for the L^ω spaces to be pure unbounded partial Hilbert algebras are established.

I. INTRODUCTION

Partial $*$-algebras have been systematically studied in recent publications by Antoine and his co-workers.[1–4] These are generalizations of $*$-algebras. In particular, a partial Op-$*$-algebra is a generalization of an Op-$*$-algebra.[5] These generalizations appear in several physical contexts.[4]

In this paper, we introduce the notion of an unbounded partial Hilbert algebra. This is an extension of the concept of an unbounded Hilbert algebra. The latter has been systematically studied by Inoue.[6,7]

The organization of this paper is as follows. In Sec. II, we review the basic notions and notation which we employ throughout the discussion. In Sec. III, we formulate our notion of unbounded partial Hilbert algebras and present some relevant results. The idea of a pure unbounded partial Hilbert algebra is also considered. Examples of unbounded partial Hilbert algebras are furnished in Sec. IV. Using the results of this paper, one readily sees that the space $L^2(\mathfrak{A},\mathfrak{B},\tau)$ introduced in Ref. 8 in the course of our study of Dirichlet forms is also an unbounded partial Hilbert algebra. To give further examples of unbounded partial Hilbert algebras, we introduce noncommutative L^ω spaces over partial Hilbert algebras in Sec. V and study a number of their properties. In Sec. VI, we show that the L^ω spaces are indeed unbounded partial Hilbert algebras and give necessary and sufficient conditions for their purity. The sort of study carried out in this paper is useful in the understanding of unbounded left partial Hilbert algebras which are themselves crucial in any formulation of a Tomita–Takesaki theory[9] in the context of partial $*$-algebras. The results of this paper generalize a number of results obtained by Inoue in Refs. 6 and 7.

II. PARTIAL $*$-ALGEBRAS

The notion of a partial $*$-algebra[2,3] is fundamental in what follows. We shall therefore briefly introduce it and, in the process, discuss certain related concepts.

Definition: A *partial* $*$*-algebra* is a triple $(\mathfrak{A},*,\Gamma)$ consisting of a complex linear space $\mathfrak{A}$, an antilinear involution $x\mapsto x*$ of $\mathfrak{A}$ into itself and a subset Γ of $\mathfrak{A}\times\mathfrak{A}$ with the following properties:

(i) (x,y_1), $(x,y_2)\in\Gamma$, and $(\lambda_1,\lambda_2)\in\mathbb{C}^2$ imply $(x,\lambda_1 y_1+\lambda_2 y_2)\in\Gamma$, where $\mathbb{C}$ = the complex numbers;

(ii) $(x,y)\in\Gamma$ implies the existence of a member $x\cdot y$ in $\mathfrak{A}$ and the map $(u,v)\mapsto u\cdot v$ of Γ into $\mathfrak{A}$ is such that if (x,y) and (x,z) lie in Γ, then $x\cdot(\alpha y+\beta z)=\alpha(x\cdot y)+\beta(x\cdot z)$, $(\alpha,\beta)\in\mathbb{C}^2$;

(iii) $(x,y)\in\Gamma$ implies $(y*,x*)\in\Gamma$ and $(x\cdot y)*=y*\cdot x*$.

Remark: (1) Let $(\mathfrak{A},*,\Gamma)$ be a partial $*$-algebra and $(x,y)\in\Gamma$. Then x (resp. y) is called a *left multiplier* (resp. a *right multiplier* of y (resp. x).

For $x\in\mathfrak{A}$, we write $M_L(x)$ [resp. $M_R(x)$] for the set of all left multipliers (resp. right multipliers) of x. Furthermore, if $\mathfrak{C}$ is a subset of $\mathfrak{A}$, then we define $M_L(\mathfrak{C})$ and $M_R(\mathfrak{C})$ by

$$M_L(\mathfrak{C})=\{x\in\mathfrak{A}:\quad x\in M_L(y)\,\forall y\in\mathfrak{C}\},$$

and

$$M_R(\mathfrak{C})=\{x\in\mathfrak{A}:\quad x\in M_R(y)\,\forall y\in\mathfrak{C}\}.$$

We remark that $M_L(\mathfrak{C})$ [resp. $M_R(\mathfrak{C})$] is the set of universal left (resp. right) multipliers of $\mathfrak{C}$.

(2) Given the notion of multipliers, we see that

$$(x,y)\in\Gamma\Leftrightarrow x\in M_L(y)\Leftrightarrow y\in M_R(x).$$

Therefore, we may replace the relation $(x,y)\in\Gamma$ by $x\in M_L(y)$ or $y\in M_R(x)$ and refer briefly to $\mathfrak{A}$ as a partial $*$-algebra. We adopt this procedure in the sequel.

(3) It is noteworthy that the partial multiplication $(x,y)\mapsto x\cdot y$ of a partial $*$-algebra $\mathfrak{A}$ is not required to be *associative*. Although it is possible to endow $\mathfrak{A}$ with a notion of associativity, such a notion may be too strong in some instances.[2] Therefore we shall employ the following weaker concept.

Definition: A partial $*$-algebra is said to be *semiassociative* if $x,y\in\mathfrak{A}$, with $y\in M_R(x)$, implies

(i) $y\cdot z\in M_R(x)$, for all $z\in M_R(\mathfrak{A})$,

and

(ii) $x\cdot(y\cdot z)=(x\cdot y)\cdot z$, for all $z\in M_R(\mathfrak{A})$.

Remark: (1) It is clear that the notion of semiassociativity may also be formulated in terms of left, rather than right, multipliers. Examples of semiassociative partial $*$-algebras may be found in Ref. 2; others will be encountered in the sequel.

(2) It is useful to note that $M_L(\mathfrak{A})$ and $M_R(\mathfrak{A})$ are algebras whenever $\mathfrak{A}$ is semiassociative.

(3) If $\mathfrak{A}$ is an arbitrary partial $*$-algebra which contains a member e with the properties that $e\in M_L(\mathfrak{A})\cap M_R(\mathfrak{A})$, $e*=e$ and $e\cdot x=x=x\cdot e$ for all $x\in\mathfrak{A}$, then $\mathfrak{A}$ is said to be *unital* and e is called the *unit* of $\mathfrak{A}$.

(4) We conclude this section by introducing some notation from the theory of unbounded linear operators on Hilbert spaces.

Let $\mathscr{H}$ be some Hilbert space and $\mathfrak{C}$ a set of closable linear operators each with a dense domain in $\mathscr{H}$ which it

 0022-2488/89/091957-08$02.50

maps into $\mathscr{H}$. For $A\in\mathbb{C}$, we write $\bar{A}$ for the minimal closed extension of A and define $\overline{\mathbb{C}}$ by

$$\overline{\mathbb{C}}=\{\bar{A}:\quad A\in\mathbb{C}\}.$$

If $A,B\in\overline{\mathbb{C}}$, then we define their strong sum $A+B$ and strong product $A\cdot B$ by $\overline{A+B}$ and $\overline{AB}$, respectively, whenever these closures exist. Furthermore, the strong scalar multiplication $\lambda\cdot C$ of $\lambda\in\mathbb{C}$ and $C\in\mathbb{C}$ is defined by

$$\lambda\cdot C=\lambda C,\ \text{if}\ \lambda\neq 0,$$

and

$$\lambda\cdot C=0\ \text{if}\ \lambda=0.$$

Finally, if $B(\mathscr{H})$ is the Banach $*$-algebra of all endomorphisms of $\mathscr{H}$ and $\mathscr{M}$ is a set of linear operators with domains and ranges in $\mathscr{H}$, we define the bounded part $\mathscr{M}_b$ of $\mathscr{M}$ by

$$\mathscr{M}_b=\mathscr{M}\cap B(\mathscr{H}).$$

III. UNBOUNDED PARTIAL HILBERT ALGEBRAS

Throughout the rest of this paper, $\mathfrak{A}$ is an arbitrary semiassociative partial $*$-algebra, unless otherwise stated.

Definition 3.1: We say that a sesquilinear form

$$\tau:\quad \mathfrak{A}\times\mathfrak{A}\to\mathbb{C}$$

is a *bitrace* on $\mathfrak{A}$ if it enjoys the following additional properties:

(i) $\tau(x,x)\geqslant 0$, $\forall\, x\in\mathfrak{A}$;

(ii) $\tau(x,x)=0$, iff $x=0$;

(iii) $\tau(y^*,x^*)=\tau(x,y)$, $\forall x,y\in\mathfrak{A}$;

(iv) $\tau(zx,y)=\tau(x,z^*y)$, $\forall x,y,z\in\mathfrak{A}$ such that $z\in M_L(x)$ and $z^*\in M_L(y)$.

Remark: Antoine[4] defines an *h form* as a non-negative sesquilinear form satisfying 3.1 (iv). It follows that a bitrace is a *faithful h* form possessing the additional property 3.1 (iii).

Notation: We denote the set of all bitraces on $\mathfrak{A}$ by $\mathrm{btr}(\mathfrak{A})$.

Remark: For $\tau\in\mathrm{btr}(\mathfrak{A})$, define $||\cdot||_\tau:\mathfrak{A}\to[0,\infty)$ by

$$||x||_\tau=(\tau(x,x))^{1/2},\quad x\in\mathfrak{A}.$$

Then, the pair $(\mathfrak{A},||\cdot||_\tau)$ is a normed space.

Definition: Let $\tau\in\mathrm{btr}(\mathfrak{A})$. Then, we say that the pair $(\mathfrak{A},\tau)$ is a *partial Hilbert algebra* if the pair $(\mathfrak{A},||\cdot||_\tau)$ is a Banach space.

Remark: Let $\tau\in\mathrm{btr}(\mathfrak{A})$. We denote the $||\cdot||_\tau$ completion of $\mathfrak{A}$ by $\mathfrak{H}_\tau$ and the extension of the involution of $\mathfrak{A}$ to $\mathfrak{H}_\tau$ by J. The map J is bijective.

We shall utilize the following result in the sequel.

Proposition 3.2: Let $\tau\in\mathrm{btr}(\mathfrak{A})$. Then, $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$ iff $M_L(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$.

Proof: Suppose that $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$. Let $x\in\mathfrak{H}_\tau$ be arbitrary and $\tau(x,y)=0$ $\forall y\in M_L(\mathfrak{A})$. Then, by 3.1 (iii),

$$\tau(y^*,Jx)=0,\quad \forall y\in M_L(\mathfrak{A}),$$

i.e.,

$$\tau(z,Jx)=0,\quad \forall z\in M_R(\mathfrak{A}),$$

since the transformation $u\to u^*$ maps $M_L(\mathfrak{A})$ onto $M_R(\mathfrak{A})$ and is bijective. Hence $Jx=0$, in view of the denseness of $M_R(\mathfrak{A})$ in $\mathfrak{H}_\tau$. Hence $x=0$, showing that $M_L(\mathfrak{A})$ is indeed dense in $\mathfrak{H}_\tau$. A similar argument holds when the roles of $M_R(\mathfrak{A})$ and $M_L(\mathfrak{A})$ are interchanged. This concludes the proof.

Remark: If $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$, then the bitrace τ is a weakly GNS *h form* in the sense of Antoine,[4] since τ is faithful.

Definition: A member τ of $\mathrm{btr}(\mathfrak{A})$ will be called *regular* if $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$.

Notation: (1) We write $\mathbf{btr}(\mathfrak{A})$ for the set of all the regular members of $\mathrm{btr}(\mathfrak{A})$.

(2) Throughout the rest of this paper, τ denotes a fixed member of $\mathbf{btr}(\mathfrak{A})$, unless otherwise stated.

Definition: For each $x\in\mathfrak{A}$, define $\pi(x)$ and $\pi'(x)$ on $M_R(\mathfrak{A})$ and $M_L(\mathfrak{A})$, respectively, as follows:

$$\pi(x)y=xy,\quad y\in M_R(\mathfrak{A}),$$

and

$$\pi'(x)z=zx,\quad z\in M_L(\mathfrak{A}).$$

We call π (resp. π') the *left* (resp. *right*) *regular representation* of $\mathfrak{A}$.

Remark: For each $x\in\mathfrak{A}$, $\pi(x)$ and $\pi'(x)$ are closable linear operators on their respective domains and $\pi(x)^*\supset\pi(x^*)$, $\pi'(x)^*\supset\pi'(x^*)$.

Notation: Define $\mathfrak{A}_0$ and $\mathfrak{A}_0'$ by

$$\mathfrak{A}_0\equiv\{x\in M_R(\mathfrak{A}):\quad \overline{\pi(x)}\in B(\mathfrak{H}_\tau)\},$$
$$\mathfrak{A}_0'\equiv\{x\in M_L(\mathfrak{A}):\quad \overline{\pi'(x)}\in B(\mathfrak{H}_\tau)\}.$$

Furthermore, let

$$\mathfrak{A}_0^2\equiv\{x\cdot y:\quad x,y\in\mathfrak{A}_0\},$$
$$\mathfrak{A}_0'^2\equiv\{x\cdot y:\quad x,y\in\mathfrak{A}_0'\}.$$

It follows from the semiassociativity of $\mathfrak{A}$ that $\mathfrak{A}_0^2$ and $\mathfrak{A}_0'^2$ are subsets of $M_R(\mathfrak{A})$ and $M_L(\mathfrak{A})$, respectively. Furthermore, $\mathfrak{A}_0^2\subset\mathfrak{A}_0$ and $\mathfrak{A}_0'^2\subset\mathfrak{A}_0'$.

Proposition 3.3: The set $\mathfrak{A}_0^2$ is dense in $\mathfrak{H}_\tau$ iff the set $\mathfrak{A}_0'^2$ is dense in $\mathfrak{H}_\tau$.

Proof: The argument is analogous to that used in the proof of Proposition 3.2.

Definition 3.4: We call the pair $(\mathfrak{A},\tau)$ an *unbounded partial Hilbert algebra* over $\mathfrak{A}_0$ (or over $\mathfrak{A}_0'$) if $\mathfrak{A}_0^2$ (resp. $\mathfrak{A}_0'^2$) is dense in $\mathfrak{H}_\tau$.

Proposition: The pair $(\mathfrak{A},\tau)$ is an unbounded partial Hilbert algebra over $\mathfrak{A}_0$ iff $(\mathfrak{A},\tau)$ is an unbounded partial Hilbert algebra over $\mathfrak{A}_0'$.

Proof: This is a restatement of Proposition 3.3. □

Definition: An unbounded partial Hilbert algebra $(\mathfrak{A},\tau)$ over $\mathfrak{A}_0$ (resp. $\mathfrak{A}_0'$) will be called *pure* if $\mathfrak{A}_0\neq\mathfrak{A}$ (resp. $\mathfrak{A}_0'\neq\mathfrak{A}$).

Remark 3.5: (1) Let π_0 (resp. π_0') denote the left (resp. right) regular representation of $\mathfrak{A}_0'$ (resp. $\mathfrak{A}_0$).

(2) For each $x\in\mathfrak{H}_\tau$, define $\pi_0(x)$ and $\pi_0'(x)$ by

$$\pi_0(x)y=\overline{\pi_0'(y)}x,\quad y\in\mathfrak{A}_0,$$
$$\pi_0'(x)y=\overline{\pi_0(y)}x,\quad y\in\mathfrak{A}_0'.$$

Then, $\pi_0(x)$ [resp. $\pi_0'(x)$], $x\in\mathfrak{H}_\tau$, is a linear operator on $\mathfrak{H}_\tau$ with domain $\mathfrak{A}_0$ (resp. $\mathfrak{A}_0'$).

(3) Notice that[10]

$$\overline{\pi_0(Jx)} = \pi_0(x)^*$$

and

$$\overline{\pi_0'(Jx)} = \pi_0'(x)^*$$

for each $x\in\mathfrak{H}_\tau$.

Proposition 3.6: For each $x\in\mathfrak{A}$,

(1) $\overline{\pi(x)} = \overline{\pi_0(x)}$, $\overline{\pi'(x)} = \overline{\pi_0'(x)}$,

(2) $\overline{\pi(x^*)} = \pi(x)^*$, $\overline{\pi'(x^*)} = \pi'(x)^*$.

Proof: (1) Let $x\in\mathfrak{A}$. To prove that $\overline{\pi(x)} = \overline{\pi_0(x)}$, first note that $\pi(x)\supset\pi_0(x)$, by the definition of these operators. Hence, $\pi_0(x)^*\supset\pi(x)^*$. Since $\pi_0(x)^* = \overline{\pi_0(x^*)}$, by Remark 3.5 (3), and $\pi(x)^*\supset\overline{\pi(x^*)}$, we have

$$\overline{\pi_0(x)} = \pi_0(x)^{**} = \pi_0(x^*)^*\supset\pi(x^*)^*\supset\overline{\pi(x)}.$$

Therefore, $\overline{\pi_0(x)} = \overline{\pi(x)}$, as claimed.

The claim $\overline{\pi'(x)} = \overline{\pi_0'(x)}$ is similarly established.

(2) Let $x\in\mathfrak{A}$. To prove $\overline{\pi(x^*)} = \pi(x)^*$, we use the result in (1) above to get

$$\overline{\pi(x^*)} = \overline{\pi_0(x^*)} = \pi_0(x)^* = \pi(x)^*.$$

We show similarly that $\overline{\pi'(x^*)} = \pi'(x)^*$. This concludes the proof.

□

Remark: The following result is readily verified. So we omit its proof.

Proposition 3.7: (1) Let λ_1, $\lambda_2\in\mathbb{C}$ and x, x_j, y, $y_j\in\mathfrak{A}$, $j = 1,2$. Then, the following relations hold:

(i) $\pi(\lambda_1x_1 + \lambda_2x_2) = \lambda_1\pi(x_1) + \lambda_2\pi(x_2)$,

(ii) $\pi(x_1\cdot x_2) = \pi(x_1)\pi(x_2)$, if $x_1\in M_L(x_2)$,

(iii) $\pi(x^*)\subset\pi(x)^*$,

(i') $\pi'(\lambda_1y_1 + \lambda_2y_2) = \lambda_1\pi'(y_1) + \lambda_2\pi'(y_2)$,

(ii') $\pi'(y_1\cdot y_2) = \pi'(y_2)\pi'(y_1)$, if $y_1\in M_L(y_2)$,

(iii') $\pi'(y^*)\subset\pi'(y)^*$.

(2) Define the involution $^\#$ on $\pi(\mathfrak{A})$ and $\pi'(\mathfrak{A})$ by

$$\pi(x)^\# = \pi(x^*), \quad x\in\mathfrak{A},$$

and

$$\pi'(y)^\# = \pi'(y^*), \quad y\in\mathfrak{A}.$$

Then, $\pi(\mathfrak{A})$ and $\pi'(\mathfrak{A})$ are partial $^\#$-algebras and

(i) $\pi(M_R(\mathfrak{A}))_b = \pi(\mathfrak{A}_0)$,

(ii) $\overline{\pi(x)^\#} = \pi(x^*)$, $\forall x\in\mathfrak{A}$,

(iii) $J\pi(x)J = \pi'(x)^\#$ on $\mathfrak{A}_0'$, $\forall x\in\mathfrak{A}$,

(i') $\pi'(M_L(\mathfrak{A}))_b = \pi'(\mathfrak{A}_0')$,

(ii') $\overline{\pi'(x)^\#} = \pi'(x)^*$, $\forall x\in\mathfrak{A}$,

(iii') $J\pi'(x)J = \pi(x)^\#$ on $\mathfrak{A}_0$, $\forall x\in\mathfrak{A}$,

(iv) $\pi(x)\pi'(y) = \pi'(y)\pi(x)$ on $\mathfrak{A}_0\cap\mathfrak{A}_0'$, $\forall x,y\in\mathfrak{A}$.

□

Remark: The sets $\overline{\pi(\mathfrak{A})}$ and $\overline{\pi'(\mathfrak{A})}$ also have the structure of partial *-algebras. More precisely, we have the following result.

Proposition 3.8: The sets $\overline{\pi(\mathfrak{A})}$ and $\overline{\pi'(\mathfrak{A})}$ are partial *-algebras of closed linear operators on $\mathfrak{H}_\tau$ under the algebraic operations of strong sum, strong multiplication, strong scalar multiplication, and the formation of adjoints.

Furthermore,

$$\overline{J\pi(x)J} = \pi'(x),$$

and

$$\overline{J\pi'(x)J} = \pi(x)^*, \quad x\in\mathfrak{A}.$$

Proof: Using Propositions 3.6 and 3.7, this is proved essentially as in the second reference of Ref. 6.

Remark: (1) In the next section, we give some examples of unbounded partial Hilbert algebras.

(2) Inoue[11] has also given a formulation of the notion of an unbounded partial Hilbert algebra. However, his Definition 5.2 of an unbounded partial Hilbert algebra is different from our Definition 3.4. Furthermore, the results contained in this section are different from those obtained by Inoue, who does not at all consider the notion of a *pure* unbounded partial Hilbert algebra. Finally, we remark that Example 4.1 and Example 4.3 of the present paper are not considered by Inoue and, moreover, the ideas and analysis in Secs. V and VI of this paper are entirely absent from Ref. 11.

IV. EXAMPLES OF UNBOUNDED PARTIAL HILBERT ALGEBRAS

In this section, we give three examples of unbounded partial Hilbert algebras.

Example 4.1: Let $\mathfrak{A}$ be the Sobolev space $H^{-1}(\mathbb{R})$, i.e., $H^{-1}(\mathbb{R})$ is the completion of the Schwartz space $\mathcal{S}(\mathbb{R})$ of C^∞, rapidly decreasing functions[12] in the norm topology given by

$$f\mapsto||f|| = \left(\int_{-\infty}^{\infty} dp\,\frac{|\tilde f(p)|^2}{1+|p|^2}\right)^{1/2},$$

where $\tilde f$ is the Fourier transform of $f\in\mathcal{S}(\mathbb{R})$. Then, $\mathfrak{A}$ is a linear space of generalized functions.[12] Define the product of two members of $\mathfrak{A}$ as their convolution, denoted by $*$, and the involution $f\mapsto f^\#$ in $\mathfrak{A}$ by

$$(f^\#)\tilde{}\,(p) = \overline{\tilde f(p)}, \quad p\in\mathbb{R}.$$

With the product and involution just introduced, $\mathfrak{A}$ is a *partial $^\#$-algebra*, since, for example, the generalized function f such that $\tilde f(p) = |p|^{1/4}$, $p\in\mathbb{R}$, lies in $\mathfrak{A}$ but $f*f$ does not lie in $\mathfrak{A}$.

Let

$$\mathfrak{B} = \{f\in\mathfrak{A}: \ ||\tilde f||_\infty < \infty\}$$

Then, it is not difficult to see that

$$M_L(\mathfrak{A}) = \mathfrak{B} = M_R(\mathfrak{A}).$$

Furthermore, it is clear that $\mathfrak{A}$ is *semiassociative*. Next, define τ: $\mathfrak{A}\times\mathfrak{A}\to\mathbb{C}$ by

$$\tau(f,g) = \int_{-\infty}^{\infty} dp\,\frac{\tilde f(p)\,\overline{\tilde g(p)}}{(1+|p|^2)^2}.$$

Then, it is straightforward to see that τ is a *bitrace* on $\mathfrak{A}$. Moreover, we have

$$\mathfrak{A}_0 = \mathfrak{B} = \mathfrak{A}_0'$$

and

$$\mathfrak{A}_0^2 = \mathfrak{B}_0 = \mathfrak{A}_0'^2,$$

where $\mathfrak{B}_0$ is some subset of $\mathfrak{B}$.

It is clear that $\mathfrak{B} \supset \mathscr{S}(\mathbb{R})$. Therefore, $\mathfrak{B}$ is dense in the Hilbert space $\mathfrak{H}_\tau$ which is the $||\cdot||_\tau$ completion of $\mathfrak{A}$, where $||f||_\tau = (\tau(f,f))^{1/2}, f \in \mathfrak{A}$. Hence, τ is a *regular bitrace.*

The set $\mathfrak{B}_0$ also contains $\mathscr{S}(\mathbb{R})$ since $f \in \mathscr{S}(\mathbb{R})$ implies $\tilde{f} \in \mathscr{S}(\mathbb{R})$ and

$$\tilde{f}(p) = \tilde{f}_1(p)\tilde{f}_2(p),$$

with

$$\tilde{f}_1(p) = (1 + |p|^2)\tilde{f}(p)$$

and

$$\tilde{f}_2(p) = (1 + |p|^2)^{-1}, \quad p \in \mathbb{R},$$

showing that $f_1, f_2 \in \mathfrak{A}_0$. Hence $\mathfrak{B}_0$ is also dense in $\mathfrak{H}_\tau$, and we conclude that the pair $(\mathfrak{A},\tau)$ is *an unbounded partial Hilbert algebra* over $\mathfrak{A}_0 = \mathfrak{A}_0'$.

Finally, the pair $(\mathfrak{A},\tau)$ is a *pure* unbounded partial Hilbert algebra over $\mathfrak{A}_0 = \mathfrak{A}_0'$ since the function f with $\tilde{f}(p) = |p|^{1/4}$, $p \in \mathbb{R}$, lies in $\mathfrak{A}$ but not in $\mathfrak{A}_0 = \mathfrak{A}_0'$.

Example 4.2: Let $\mathfrak{A}$ be a von Neumann algebra which admits a semifinite, faithful, normal trace τ satisfying $\tau(e) < \infty$, where e is the unit of $\mathfrak{A}$. We denote the involution of $\mathfrak{A}$ by $*$.

Since a trace is automatically a bitrace, it is clear that $\tau \in \mathrm{btr}(\mathfrak{A})$. Furthermore, $\mathfrak{A}$ is semiassociative since $\mathfrak{A}$ is, in fact, associative.

Next, we have

(i) $M_L(\mathfrak{A}) = \mathfrak{A} = M_R(\mathfrak{A})$,

(ii) $\mathfrak{A}_0 = \mathfrak{A} = \mathfrak{A}_0'$,

(iii) $\mathfrak{A}_0^2 = \mathfrak{A} = \mathfrak{A}_0'^2$, since e lies in $\mathfrak{A}_0 = \mathfrak{A}_0'$.

Let $\mathfrak{H}_\tau$ be the $||\cdot||_\tau$ completion of $\mathfrak{A}$, where

$$||x||_\tau = (\tau(x^*x))^{1/2}, \quad x \in \mathfrak{A}.$$

From (i), it follows that τ is a *regular* bitrace and from (iii), we conclude that the pair $(\mathfrak{A},\tau)$ is an *unbounded partial Hilbert algebra.* But since $\mathfrak{A}_0 = \mathfrak{A} = \mathfrak{A}_0'$, the pair $(\mathfrak{A},\tau)$ is not a pure unbounded partial Hilbert algebra over $\mathfrak{A}_0 = \mathfrak{A}_0'$.

Example 4.3: Let $\mathscr{H}$ be a separable Hilbert space, with inner product $\langle\cdot,\cdot\rangle$ and H a self-adjoint linear operator on a dense domain in $\mathscr{H}$ such that $\exp(-\beta H)$ is nuclear[13] for every $\beta > 0$. Let (f_n) and (λ_n) be the normalized eigenvectors and the corresponding eigenvalues of H. Then, by the nuclearity of $\exp(-\beta H)$ for each $\beta > 0$, we have that (f_n) is an orthonormal basis for $\mathscr{H}$ and also that

$$\sum_{n=1}^{\infty} e^{-\beta\lambda_n} < \infty, \text{ for } \beta > 0. \tag{4.1}$$

Let $\mathscr{D} = \cap_{\beta>0} D(e^{\beta H})$, where $D(A)$ denotes the domain of A. Observe that $\mathscr{D}$ contains (f_n). Hence $\mathscr{D}$ is dense in $\mathscr{H}$.

We introduce the weak partial Op^*-algebra $\mathscr{L}_w^+(\mathscr{D},\mathscr{H})$ on $\mathscr{D}$ as follows.[14]

Let $\mathscr{L}^+(\mathscr{D},\mathscr{H})$ denote the set of all linear operators A on $\mathscr{H}$ such that $D(A) = \mathscr{D}$ and $D(A^*) \supset \mathscr{D}$, where * is the operator adjoint. Then, $\mathscr{L}^+(\mathscr{D},\mathscr{H})$ is a linear space when equipped with the usual notions of addition and scalar multiplication. Furthermore, $\mathscr{L}^+(\mathscr{D},\mathscr{H})$ admits an involution $^+$ defined by

$$A^+ = A^* \restriction \mathscr{D}, \quad A \in \mathscr{L}^+(\mathscr{D},\mathscr{H}),$$

and a partial multiplication $\square$ defined as follows: for

$$A_1, \quad A_2 \in \mathscr{L}^+(\mathscr{D},\mathscr{H}),$$

with

$$A_2\mathscr{D} \subset D(A_1^{+*})$$

and

$$A_1^+\mathscr{D} \subset D(A_2^*),$$

put $A_1 \square A_2 \equiv A_1^{+*}A_2$. Then, endowed with the involution $^+$ and partial multiplication $\square$, the linear space $\mathscr{L}^+(\mathscr{D},\mathscr{H})$ becomes a partial Op^*-algebra, denoted by $\mathscr{L}_w^+(\mathscr{D},\mathscr{H})$.

Define $\mathscr{D}_*(\mathscr{L}_w^+(\mathscr{D},\mathscr{H}))$ by

$$\mathscr{D}_*(\mathscr{L}_w^+(\mathscr{D},\mathscr{H})) = \bigcap_{A \in \mathscr{L}_w^+(\mathscr{D},\mathscr{H})} D(A^*).$$

Notice that $\mathscr{D}_*(\mathscr{L}_w^+(\mathscr{D},\mathscr{H})) \supseteq \mathscr{D}$.

Now, define $\mathfrak{A}$ to be $\mathscr{L}_w^+(\mathscr{D},\mathscr{H})$.

In what follows, we assume that $\mathscr{L}_w^+(\mathscr{D},\mathscr{H})\cdot$ is a *self-adjoint* partial Op^*-algebra, i.e., that

$$\mathscr{D}_*(\mathscr{L}_w^+(\mathscr{D},\mathscr{H})) = \mathscr{D}$$

and also that

$$e^{-\beta H}x = xe^{-\beta H} \tag{4.2}$$

for all $x \in B(\mathscr{H}) \cap \mathscr{L}(\mathscr{D})$ and $\beta > 0$, where $\mathscr{L}(\mathscr{D})$ is the subspace of $\mathscr{L}^+(\mathscr{D},\mathscr{H})$ consisting of all operators which leave $\mathscr{D}$ invariant. From the self-adjointness of $\mathscr{L}_w^+(\mathscr{D},\mathscr{H})$, it follows, by Ref. 2, Proposition 3.4, that $\mathfrak{A}$ is *semiassociative.*

Let $\tau: \mathfrak{A} \times \mathfrak{A} \to \mathbb{C}$ be given by

$$\tau(x,y) = \sum_{n=1}^{\infty} \langle xf_n, yf_n\rangle e^{-\beta\lambda n}, \quad x,y \in \mathfrak{A}. \tag{4.3}$$

We remark that the pair $(\mathfrak{A},\tau)$ appears in Ref. 4.

From Eqs. (4.3) and (4.1), one sees that $\tau(e,e) < \infty$, where e is the unit of $\mathfrak{A}$.

Next, we have the following (Ref. 2, p. 311):

$$M_R(\mathfrak{A}) = \{a \in B(\mathscr{H}): \quad a\mathscr{D} \subset \mathscr{D}\} = B(\mathscr{H}) \cap \mathscr{L}(\mathscr{D}),$$

$$M_L(\mathfrak{A}) = \{a \in B(\mathscr{H}): \quad a^+\mathscr{D} \subset \mathscr{D}\}.$$

We remark that $M_R(\mathfrak{A})$ and $M_L(\mathfrak{A})$ are, in general, distinct.

Let π (resp. π') be the left (resp. right) regular representation of $\mathfrak{A}$.

Let $\mathfrak{H}_\tau$ be the completion of $\mathfrak{A}$ in the norm $||\cdot||_\tau$ induced by the form in (4.3). Then, we find that

(i) $\mathfrak{A}_0 = \{x \in M_R(\mathfrak{A}): \quad \overline{\pi(x)} \in B(\mathfrak{H}_\tau)\} = M_R(\mathfrak{A})$,

(ii) $\mathfrak{A}_0' = \{x \in M_L(\mathfrak{A}): \quad \overline{\pi'(x)} \in B(\mathfrak{H}_\tau)\} = M_L(\mathfrak{A})$,

(iii) $\mathfrak{A}_0^2 = M_R(\mathfrak{A}) \square M_R(\mathfrak{A}) = M_R(\mathfrak{A})$,

$$\mathfrak{A}_0'^2 = M_L(\mathfrak{A}) \square M_L(\mathfrak{A}) = M_L(\mathfrak{A}), \tag{4.4}$$

since the unit e lies in $M_L(\mathfrak{A}) \cap M_R(\mathfrak{A})$,

(iv) $\mathfrak{A}_0 \neq \mathscr{L}_w^{=}(\mathscr{D},\mathscr{H}) = \mathfrak{A}$

and

$$\mathfrak{A}_0' \neq \mathscr{L}_w^{+}(\mathscr{D},\mathscr{H}) = \mathfrak{A}.$$

Furthermore, we have the following proposition.

Proposition: τ is a regular bitrace on $\mathfrak{A}$ if $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$.

Proof: Let $M_R(\mathfrak{A})$ be dense in $\mathfrak{H}_\tau$, $x,y\in\mathfrak{A}$ and $(x_p)_{p\in\mathbb{N}}$, $(y_q)_{q\in\mathbb{N}}$ be a sequence in $M_R(\mathfrak{A})$ such that

$$\lim_{p\to\infty} ||x-x_p||_\tau = 0 = \lim_{q\to\infty} ||y-y_q||_\tau.$$

Since

$$||z||_\tau^2 = \sum_{n=1}^{\infty} ||zf_n||^2 e^{-\beta\lambda_n}$$

for $z\in\mathfrak{A}$, we may assume, as we do henceforth, that $(x_p)_{p\in\mathbb{N}}$ and $(y_q)_{q\in\mathbb{N}}$ converge to x and y, respectively, in the $|||\cdot|||_\tau$ topology on $\mathfrak{A}$, where

$$|||z|||_\tau^2 = \sum_{n=1}^{\infty} (||zf_n||^2 + ||z^+f_n||^2)e^{-\beta\lambda_n}, \quad z\in\mathfrak{A}.$$

It then follows that

$$\lim_{p,q\to\infty} \sum_{n=1}^{\infty} \langle (y-y_q)f_n,(x-x_p)f_n\rangle e^{-\beta\lambda_n} = 0$$

and

$$\lim_{p,q\to\infty} \sum_{n=1}^{\infty} \langle (y-y_q)^+f_n,(x-x_p)^+f_n\rangle e^{-\beta\lambda_n} = 0,$$

since

$$\left|\sum_{n=1}^{\infty} \langle (y-y_q)f_n,(x-x_p)f_n\rangle e^{-\beta\lambda_n}\right| \leqslant |||y-y_q|||_\tau |||x-x_p|||_\tau$$

and

$$\left|\sum_{n=1}^{\infty} \langle (y-y_q)^+f_n,(x-x_p)^+f_n\rangle e^{-\beta\lambda_n}\right| \leqslant |||y-y_q|||_\tau |||x-x_p|||_\tau. \tag{4.5}$$

From the foregoing, we get finally that

$$\begin{aligned}
\tau(x,y) &= \lim_{p,q\to\infty} \tau(x_p,y_q) \\
&= \lim_{p,q\to\infty} \sum_{n=1}^{\infty} \langle x_p e^{-\beta H/2}f_n, y_q e^{-\beta H/2}f_n\rangle \\
&= \lim_{p,q\to\infty} \operatorname{tr}((x_p e^{-\beta B/2})^+ \square (y_q e^{-\beta H/2})) \\
&= \lim_{p,q\to\infty} \operatorname{tr}(e^{-\beta H} y_q x_p^+), \text{ by using (4.2)} \\
&= \lim_{p,q\to\infty} \sum_{n=1}^{\infty} \langle f_n, e^{-\beta H} y_q x_p^+ f_n\rangle \\
&= \lim_{p,q\to\infty} \sum_{n=1}^{\infty} \langle y_q^+ f_n, x_p^+ f_n\rangle e^{-\beta\lambda_n} \\
&= \tau(y^+,x^+), \quad \text{by (4.5)}
\end{aligned}$$

where

$$\operatorname{tr}(z) = \sum_{n=1}^{\infty} \langle f_n, zf_n\rangle.$$

This concludes the proof. ⌑

Thus it is evident that if we could show that $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$, then we could draw the following conclusions.

(1) τ is a *regular* bitrace on $\mathfrak{A}$;

(2) $(\mathfrak{A},\tau)$ is an *unbounded partial Hilbert algebra*, in view of (4.4) (iii); and

(3) $(\mathfrak{A},\tau)$ is a pure unbounded partial Hilbert algebra over $\mathfrak{A}_0$ (resp. $\mathfrak{A}_0'$), in view of (4.4) (iv).

The relevant result is the following.

Proposition: $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$.

Proof: Let x be an arbitrary member of $\mathfrak{H}_\tau$. For arbitrary $f\in\mathscr{D}$, xf lies in $\mathscr{H}$ and may be expressed as follows:

$$xf = \sum_{n=1}^{\infty} \langle f_n, xf\rangle f_n,$$

whence

$$||xf||^2 = \sum_{n=1}^{\infty} |\langle f_n, xf\rangle|^2 < \infty,$$

where $||\cdot||$ is the norm induced on $\mathscr{H}$ by $\langle\cdot,\cdot\rangle$.

Next, define x_m, $1\leqslant m<\infty$, by its action on $\mathscr{H}$ as follows:

$$x_m g = \sum_{n=1}^{m} \langle x^*f_n, g\rangle f_n,$$

for arbitrary $g\in\mathscr{H}$. Then, $x_m\in M_R(\mathfrak{A})$, for each m. Furthermore,

$$\begin{aligned}
||x-x_m||_\tau^2 &= \sum_{p=1}^{\infty} \langle xf_p - x_m f_p, xf_p - x_m f_p\rangle e^{-\beta\lambda_p} \\
&= \sum_{p=1}^{\infty} \left|\left| \sum_{n=m+1}^{\infty} \langle f_n, xf_p\rangle f_n \right|\right|^2 e^{-\beta\lambda_p} \\
&= \sum_{p=1}^{\infty} \sum_{n=m+1}^{\infty} |\langle f_n, xf_p\rangle|^2 e^{-\beta\lambda_p} \\
&= \sum_{n=m+1}^{\infty} \sum_{p=1}^{\infty} |\langle f_p, x^*f_n\rangle|^2 e^{-\beta\lambda_p}.
\end{aligned}$$

Hence

$$\lim_{m\to\infty} ||x-x_m||_\tau^2 = 0.$$

Thus, since x was arbitrary in $\mathfrak{H}_\tau$, it follows that $M_R(\mathfrak{A})$ is dense in $\mathfrak{H}_\tau$. This concludes the proof.

⌑

V. L^ω SPACES OVER PARTIAL HILBERT ALGEBRAS

Let $(\mathfrak{A},\tau)$ be an unbounded partial Hilbert algebra and $\mathfrak{A}_0,\pi_0,\mathfrak{H}_\tau$ be as previously introduced. In this section, we construct L^ω-spaces over $\mathfrak{A}_0$, as done in Ref. 6 for unbounded Hilbert algebras, and study some of their properties.

Let $\mathscr{U}_0(\mathfrak{A}_0)$ be the von Neumann subalgebra of $B(\mathfrak{H}_\tau)$ generated by $\overline{\pi_0(\mathfrak{A}_0)}$. The bitrace τ induces a trace τ_0, called the *natural trace*, on $\mathscr{U}_0(\mathfrak{A}_0)$ which is defined through its action on $\overline{\pi_0(\mathfrak{A}_0)}$ as follows:

$$\tau_0(\pi_0(x)*\overline{\pi_0(y)}) \equiv \tau(x,y), \quad x,y\in\mathfrak{A}_0.$$

For $1\leqslant p\leqslant\infty$, let $L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$ be Segal's noncommutative L^p space (Refs. 15–17) over $\mathscr{U}_0(\mathfrak{A}_0)$ with respect to τ_0. Then, we make the definitions

$$L^{\omega}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\equiv\bigcap_{1\leqslant p<\infty}L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

$$L_2^{\omega}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\equiv\bigcap_{2\leqslant p<\infty}L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

$$L^{\omega}(\mathfrak{A}_0,\tau)\equiv\{x\in\mathfrak{H}_\tau:\ \overline{\pi_0(x)}\in L^{\omega}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\},$$

$$L_2^{\omega}(\mathfrak{A}_0,\tau)\equiv\{x\in\mathfrak{H}_\tau:\ \overline{\pi_0(x)}\in L_2^{\omega}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\}.$$

These are the noncommutative analogues of the L^{ω}-spaces introduced in Ref. 18 by Arens. We also set

$$L_2^p(\mathfrak{A}_0,\tau)\equiv\{x\in\mathfrak{H}_\tau:\ \overline{\pi_0(x)}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\},\quad 1\leqslant p\leqslant\infty.$$

Define $||\cdot||_{(2,p)}$ on $L_2^p(\mathfrak{A}_0,\tau)$, $1\leqslant p\leqslant\infty$, as follows:

$$||x||_{(2,p)}=\max(||x||_2,||x||_p),\quad x\in L_2^p(\mathfrak{A}_0,\tau),$$

where

$$||x||_p=||\overline{\pi_0(x)}||_p,$$

for $x\in\mathfrak{H}_\tau$ with

$$\overline{\pi_0(x)}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),\quad 2\leqslant p\leqslant\infty.$$

We have the following result.

Proposition 5.1: For $1\leqslant p\leqslant\infty$, $||\cdot||_{(2,p)}$ is a norm on $L_2^p(\mathfrak{A}_0,\tau)$ and $L_2^p(\mathfrak{A}_0,\tau)$ is $||\cdot||_{(2,p)}$ complete.

Proof: It is clear that $||\cdot||_{(2,p)}$ is a norm on $L_2^p(\mathfrak{A}_0,\tau)$ in view of the faithfulness of π_0. So let us prove the $||\cdot||_{(2,p)}$ completeness of $L_2^p(\mathfrak{A}_0,\tau)$.

Suppose that (x_n) is a Cauchy sequence in $L_2^p(\mathfrak{A}_0,\tau)$. From the completeness of $\mathfrak{H}_\tau$ and $L^p(\mathscr{U}_0(\mathfrak{A}_0,\tau_0))$, there exist $x\in\mathfrak{H}_\tau$ and $T\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$ such that

$$\lim_{n\to\infty}||x_n-x||_2=0$$

and

$$\lim_{n\to\infty}||\overline{\pi_0(x_n)}-T||_p=0.$$

We shall show that $T=\overline{\pi_0(x)}$.

Let $\langle\cdot,\cdot\rangle$ denote the inner product of $L^2(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$. Then, for $y,z\in D(T)\cap\mathfrak{A}_0$, we have

$$\lim_{n+\infty}|\tau[y,(\overline{\pi_0(x_n)}-T)z]|$$

$$=\lim_{n\to\infty}|\tau_0[\pi_0(y)^*(\overline{\pi_0(x_n)}-T)\cdot\overline{\pi_0(z)}]|$$

$$=\lim_{n\to\infty}|\tau_0[\overline{\pi_0(z)}\cdot\pi_0(y)^*(\overline{\pi_0(x_n)}-T)]|$$

$$\leqslant\lim_{n\to\infty}||\pi_0(z)\cdot\pi_0(y)^*||_{p'}||\overline{\pi_0(x_n)}-T||_p=0,$$

where

$$\frac{1}{p}+\frac{1}{p'}\cdot=1.$$

Also,

$$\lim_{n\to\infty}|\langle y,(\overline{\pi_0(x_n)}-\overline{\pi_0(x)})z\rangle|$$

$$\leqslant\lim_{n\to\infty}||y||_2||z||_2||x_n-x||_2=0.$$

Hence, $Tz=\overline{\pi_0(x)}z$, for all $z\in D(T)\cap M_R(\mathfrak{A}_0)$, from the uniqueness of limits. Since T and $\pi_0(x)$ are essentially measurable, it follows by Ref. 15, Theorem 4, that $T+\pi_0(x)$ is also essentially measurable. Hence, by Ref. 15, Lemma (1.2), $D(T)\cap\mathfrak{A}_0$ is dense in $\mathfrak{H}_\tau$ and we conclude that $T=\overline{\pi_0(x)}$. So $L_2^p(\mathfrak{A}_0,\tau)$ is $||\cdot||_{(2,p)}$ complete. ∎

Remark: (1) Since $\mathfrak{A}_0$ is $||\cdot||_\tau$ dense in $\mathfrak{H}_\tau$, one readily sees that $L_2^2(\mathfrak{A}_0,\tau)=\mathfrak{H}_\tau$ and that $L^2(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$ is isometrically isomorphic to $\mathfrak{H}_\tau$.

(2) In the sequel, we regard $L_2^{\omega}(\mathfrak{A}_0,\tau)$ [resp. $L^{\omega}(\mathfrak{A}_0,\tau)$] as endowed with the projective limit topology τ_2^{ω} (resp. τ^{ω}) generated by the norms

$$\{||\cdot||_{(2,p)}:\ 2\leqslant p<\infty\}$$

(resp. $\{||\cdot||_p:\ 1\leqslant p<\infty\}$)

of the Banach spaces

$$\{L_2^p(\mathfrak{A}_0,\tau):\ 2\leqslant p<\infty\}$$

[resp. $\{L^p(\mathfrak{A}_0,\tau):\ 1\leqslant p<\infty\}$].

(3) The next result will be used in the sequel.

Proposition 5.2: (i) If $1\leqslant p<q\leqslant 2$, then

$$L_2^2(\mathfrak{A}_0,\tau)=\mathfrak{H}_\tau\supset L_2^q(\mathfrak{A}_0,\tau)\supset L_2^p(\mathfrak{A}_0,\tau)\supset L_2^1(\mathfrak{A}_0,\tau).$$

(ii) If $2\leqslant p<q$, then

$$L_2^2(\mathfrak{A}_0,\tau)=\mathfrak{H}_\tau\supset L_2^p(\mathfrak{A}_0,\tau)\supset L_2^q(\mathfrak{A}_0,\tau)\supset L_2^{\omega}(\mathfrak{A}_0,\tau)\supset L_2^{\infty}(\mathfrak{A}_0,\tau).$$

(iii) The projective limit topology τ_2^{ω} on $L_2^{\omega}(\mathfrak{A}_0,\tau)$ is equivalent to the projective limit topology generated by the norm topologies $\{||\cdot||_{(2,n)}:\ 2\leqslant n<\infty,\ n\text{ an integer}\}$ of the Banach spaces $\{L_2^n(\mathfrak{A}_0,\tau):\ 2\leqslant n<\infty,\ n\text{ an integer}\}$.

Proof: (i) and (ii) are proved as in Ref. 7, Lemma (2.3); and (iii) follows from the fact that for each $p\in[2,\infty)$, there is a positive integer n such that $n\leqslant p<n+1$, whence we have the estimate $||x||_p\leqslant||x||_n+||x||_{n+1}$ for all $x\in L_2^{\omega}(\mathfrak{A}_0,\tau)$. ∎

Proposition 5.3: If

$$L^p(\mathscr{U}_0,(\mathfrak{A}_0),\tau_0)=L^q(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$$

and

$$L_2^p(\mathfrak{A}_0,\tau)=L_2^q(\mathfrak{A}_0,\tau),\ \text{for some } q>p\geqslant 2,$$

then

$$L_2^r(\mathfrak{A}_0,\tau)=L_2^{\omega}(\mathfrak{A}_0,\tau)\ \text{ for all } r\in[p,\infty).$$

Proof: Let

$$x\in L_2^p(\mathfrak{A}_0,\tau)=L_2^q(\mathfrak{A}_0,\tau).$$

Then

$$|\overline{\pi_0(x)}|^{q/p}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0).$$

From the inequality $2<2q/p\leqslant q$ and Proposition 5.2 (ii), we get

$$L_2^2(\mathfrak{A}_0,\tau)\supset L^{2q/p}(\mathfrak{A}_0,\tau)\supset L_2^q(\mathfrak{A}_0,\tau),$$

whence

$$x\in L_2^{2q/p}(\mathfrak{A}_0,\tau),$$

i.e.,

$$|\overline{\pi_0(x)}|^{2q/p}\in L^1(\mathscr{U}_0(\mathfrak{A}_0),\tau_0).$$

Hence

$$|\overline{\pi_0(x)}|^{q/p}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\cap L^2(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$$
$$= L_2^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0) = L_2^q(\mathscr{U}_0(\mathfrak{A}_0),\tau_0).$$

Next, we show that $|\overline{\pi_0(x)}|^{(q/p)^2}\in L_2^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$. To see this, notice first that since

$$|\overline{\pi_0(x)}|^{2q/p}\in L^1(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

it follows that

$$|\overline{\pi_0(x)}|^{q/p}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\cap L^2(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$$

whence

$$|\overline{\pi_0(x)}|^{2(q/p)^2}\in L^1(\mathscr{U}_0(\mathfrak{A}_0),\tau_0).$$

Hence

$$|\overline{\pi_0(x)}|^{(q/p)^2}\in L^2(\mathscr{U}_0,(\mathfrak{A}_0),\tau_0).$$

Furthermore, since

$$|\overline{\pi_0(x)}|^{q/p}\in L^q(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

it follows that

$$|\overline{\pi_0(x)}|^{q^2/p}\in L^1(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

whence

$$|\overline{\pi_0(x)}|^{(q/p)^2}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0).$$

Thus

$$|\overline{\pi_0(x)}|^{(q/p)^2}\in L^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\cap L^2(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$$
$$= L_2^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0).$$

By iterating the last argument, we get that

$$|\overline{\pi_0(x)}|^{(q/p)^n}\in L_2^p(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),\quad n=1,2,\ldots.$$

Since $q/p>1$, we conclude from Proposition 5.2 that $x\in L_2^\omega(\mathfrak{A}_0,\tau)$. This completes the proof. ⌑

VI. PURITY OF THE UNBOUNDED PARTIAL HILBERT ALGEBRAS $L^\omega(\mathfrak{A}_0,\tau)$ and $L_2^\omega(\mathfrak{A}_0,\tau)$

In this section, we show that $L^\omega(\mathfrak{A}_0,\tau)$ and $L_2^\omega(\mathfrak{A}_0,\tau)$ are unbounded partial Hilbert algebras and then we furnish necessary and sufficient conditions for their purity.

Notation: Define $(\mathfrak{A}_0)_b$ and $(\mathfrak{A}_0)_b^2$ by

$$(\mathfrak{A}_0)_b \equiv \{x\in\mathfrak{H}_\tau:\quad \overline{\pi_0(x)}\in B(\mathfrak{H}_\tau)\}$$

and

$$(\mathfrak{A}_0)_b^2 \equiv \{x\cdot y:\quad x,y\in(\mathfrak{A}_0)_b \text{ with } x\in M_L(y)\}.$$

Notice that

$$(\mathfrak{A}_0)_b = L_2^\infty(\mathfrak{A}_0,\tau).$$

Moreover, $(\mathfrak{A}_0)_b$ is a partial Hilbert algebra which is maximal amongst all partial Hilbert algebras contained in $\mathfrak{H}_\tau$.

Remark: The following result is proved essentially as in Ref. 7, Lemma 2.2.

Proposition 6.1: (1) For $1\leqslant p<2$, the set $(\mathfrak{A}_0)_b^2$ is dense in $L_2^p(\mathfrak{A}_0,\tau)$.

(2) For $2\leqslant p\leqslant\infty$, the set $(\mathfrak{A}_0)_b$ is dense in $L_2^p(\mathfrak{A}_0,\tau)$.

Remark: The next result discloses that $L^\omega(\mathfrak{A}_0,\tau)$ and $L_2^\omega(\mathfrak{A}_0,\tau)$ are unbounded partial Hilbert algebras.

Theorem 6.2: The space $L^\omega(\mathfrak{A}_0,\tau)$ [resp. $L_2^\omega(\mathfrak{A}_0,\tau)$] is an unbounded partial Hilbert algebra containing

$$(\mathfrak{A}_0)_b^2 \text{ [resp. } (\mathfrak{A}_0)_b\text{]}.$$

Proof: Let $x,y\in L^\omega(\mathfrak{A}_0,\tau)$. Then, there are X,Y in $L^\omega(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)$ such that $X=\overline{\pi_0(x)}$ and $Y=\overline{\pi_0(y)}$. Now, for any $p\in[1,\infty)$, we have

$$||X\cdot Y||_p\leqslant||X||_{2p}||Y||_{2p}.$$

Hence

$$X\cdot Y=\overline{\pi_0(x)}\cdot\overline{x_0(y)}\in L^\omega(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

whence

$$\overline{\pi_0(x\cdot y)}\in L^\omega(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

if $x\in M_L(y)$. This shows that $x\cdot y\in L^\omega(\mathfrak{A}_0,\tau)$ whenever $x,y\in L^\omega(\mathfrak{A}_0,\tau)$ and $x\cdot y$ is defined. Also, since

$$||X^*||_p = ||X||_p,\quad \text{for } p\in[1,\infty),$$

it follows that $x^*\in L^\omega(\mathfrak{A}_0,\tau)$ for each $x\in L^\omega(\mathfrak{A}_0,\tau)$. Hence $L^\omega(\mathfrak{A}_0,\tau)$ is a partial *-algebra. Furthermore, since every $x\in L^\omega(\mathfrak{A}_0,\tau)$ is also a member of $\mathfrak{H}_\tau$, it is easy to see that the restriction of τ to $L^\omega(\mathfrak{A}_0,\tau)\times L^\omega(\mathfrak{A}_0,\tau)$ is a bitrace.

Next, suppose that $x,y\in(\mathfrak{A}_0)_b$ with $x\in M_L(y)$. Then, $x\cdot y\in(\mathfrak{A}_0)_b^2$, and for any $p\in[1,\infty)$, we have

$$||x\cdot y||_p^p = \tau_0(|\overline{\pi_0(x\cdot y)}|^p)$$
$$= \tau_0(|\overline{\pi_0(x\cdot y)}|^{p-1}|\overline{\pi_0(x\cdot y)}|)$$
$$\leqslant||x\cdot y||_\infty^{p-1}||x\cdot y||_1$$
$$\leqslant||x||_\infty^{p-1}||y||_2^{p-1}||x||_2||y||_2$$
$$= ||x||_\infty^{p-2}||x||_2||y||_2^p$$
$$\leqslant||x||_{(2,\infty)}^p||y||_{(2,\infty)}^p.$$

Hence

$$||x\cdot y||_p\leqslant||x||_{(2,\infty)}||y||_{(2,\infty)},$$

$x,y\in(\mathfrak{A}_0)_b$, with $x\in M_L(y)$. This shows that $(\mathfrak{A}_0)^2\subset L^\omega(\mathfrak{A}_0,\tau)$. The density of $(\mathfrak{A}_0)_b^2$ in $L^\omega(\mathfrak{A}_0,\tau)$ follows from Proposition 6.1.

Similarly, we readily show that $L_2^\omega(\mathfrak{A}_0,\tau)$ is a partial *-algebra and that the restriction of τ to $L_2^\omega(\mathfrak{A}_0,\tau)\times L_2^\omega(\mathfrak{A}_0,\tau)$ is a bitrace. Moreover, $(\mathfrak{A}_0)_b$ is contained in $L_2^\omega(\mathfrak{A}_0,\tau)$ since for arbitrary $p\in(2,\infty)$ and $x\in(\mathfrak{A}_0)_b$, we have

$$||x||_p^p = \tau_0(|\overline{\pi_0(x)}|^{p-2}|\overline{\pi_0(x)}|^2)$$
$$\leqslant||x||_\infty^{p-2}||x||_2^2\leqslant||x||_{(2,\infty)}^p,$$

whence

$$||x||_p\leqslant||x||_{(2,\infty)}.$$

The density of $(\mathfrak{A}_0)_b$ in $L_2^\omega(\mathfrak{A}_0,\tau)$ follows from Proposition 6.1. This concludes the proof. ⌑

Remark: Concerning the purity of the unbounded partial Hilbert algebras $L^\omega(\mathfrak{A}_0,\tau)$ and $L_2^\omega(\mathfrak{A}_0,\tau)$, we get the following result.

Theorem 6.3: Let $(\mathfrak{A},\tau)$ be a partial Hilbert algebra over $\mathfrak{A}_0$. Then, the following conditions are equivalent:

(1) $L_2^{\omega}(\mathfrak{A}_0,\tau)$ is pure.
(2) $L^{\omega}(\mathfrak{A}_0,\tau)$ is pure.
(3) $\mathfrak{H}_\tau$ is not a partial Hilbert algebra, i.e., $(\mathfrak{A}_0)_b \neq \mathfrak{H}_\tau$.
(4) $L_2^{\omega}(\mathfrak{A}_0,\tau) \neq \mathfrak{H}_\tau$.
(5) $L_2^{p}(\mathfrak{A}_0,\tau) \neq L_2^{q}(\mathfrak{A}_0,\tau)$ for some $q > p \geqslant 2$.
(6) $L_2^{p}(\mathfrak{A}_0,\tau) \neq L_2^{2}(\mathfrak{A}_0,\tau)$ and

$$L_2^{p}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0) \neq L_2^{2}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0),$$

for each $p > 2$.

Proof: (6) $\Rightarrow$ (5); this is obvious.

(5) $\Rightarrow$ (4). Suppose (5) holds but $L_2^{\omega}(\mathfrak{A}_0,\tau) = \mathfrak{H}_\tau$.

Then, by Proposition 5.2 (ii), we have $L_2^{p}(\mathfrak{A}_0,\tau) = L_2^{q}(\mathfrak{A}_0,\tau)$, for $2 \leqslant p < q$, a contradiction.

(4) $\Rightarrow$ (3). Suppose (4) holds but $\mathfrak{H}_\tau = (\mathfrak{A}_0)_b$. Then, since

$$\mathfrak{H}_\tau \supset L_2^{\omega}(\mathfrak{A}_0,\tau) \supset (\mathfrak{A}_0)_b,$$

we have $\mathfrak{H}_\tau = L_2^{\omega}(\mathfrak{A}_0,\tau)$, a contradiction.

(3) $\Rightarrow$ (2). Suppose (3) holds but $L^{\omega}(\mathfrak{A}_0,\tau) = (\mathfrak{A}_0)_b^2$. Then,

$$\{x \in \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}: \quad \overline{\pi_0(x)} \in L^{\omega}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0)\} = \{0\},$$

i.e.,

$$x \in \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}, \quad x \neq 0, \text{ implies } x \notin L^{\omega}(\mathfrak{A}_0,\tau).$$

Thus, we must have

$$(*) \quad ||x||_p = \infty, \text{ for some } p \in [1,\infty), \text{ if } x \neq 0.$$

Now, either $p \in (2,\infty)$ or else $p \in (1,2)$.

Suppose $p \in (2,\infty)$. Then $||x||_q = \infty \ \forall\, q > p$, by Proposition 5.2 (ii). Hence

$$x \in \bigcap_{2 < t < p} (L^{t}(\mathfrak{A}_0,\tau) \cap \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}) \equiv L_2^{p-}(\mathfrak{A}_0,\tau),$$

whence

$$\mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2} \subset L_2^{p-}(\mathfrak{A}_0,\tau).$$

But clearly,

$$L_2^{p-}(\mathfrak{A}_0,\tau) \subset \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}.$$

Thus

$$L_2^{p-}(\mathfrak{A}_0,\tau) = \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}.$$

But $L_2^{p-}(\mathfrak{A}_0,\tau)$, being the projective limit of the Banach spaces $\{L^{t}(\mathfrak{A}_0,\tau) \cap \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}, ||\cdot||_{(2,t)}: 2 \leqslant t < p\}$, is not a Hilbert space unless $p = 2$. Hence, from (*), we must have $||x||_2 = \infty$, $0 \neq x \in \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}$, a contradiction.

Similarly, if $p \in [1,2)$, then it follows that $||x||_p = \infty$ for all $1 \leqslant q < p$, and $x \neq 0$, by Proposition 5.2 (i). Hence

$$x \in \bigcup_{p < t \leqslant 2} (L^{t}(\mathfrak{A}_0,\tau) \cap \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}) \equiv L_2^{p+}(\mathfrak{A}_0,\tau)$$

and we see again that

$$L_2^{p+}(\mathfrak{A}_0,\tau) = \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}.$$

But $L_2^{p+}(\mathfrak{A}_0,\tau)$, being the inductive limit of the Banach spaces $\{L_2^{t+}\mathfrak{A}_0,\tau) \cap \overline{(\mathfrak{A}_0)_b^2}, ||\cdot||_{(2,t)}: p < t \leqslant 2\}$, is not a Hilbert space unless $p = 2$. Hence from (*), we get that $||x||_2 = \infty$, $0 \neq x \in \mathfrak{H}_\tau \ominus \overline{(\mathfrak{A}_0)_b^2}$, a contradiction.

(2) $\Rightarrow$ (1). This follows from the purity of $L^{\omega}(\mathfrak{A}_0,\tau)$ and the inclusion $L_2^{\omega}(\mathfrak{A}_0,\tau) \supset L^{\omega}(\mathfrak{A}_0,\tau)$.

(1) $\Rightarrow$ (6). Suppose (1) holds but

$$L_2^{p}(\mathfrak{A}_0,\tau) = L_2^{2}(\mathfrak{A}_0,\tau)$$

and

$$L_2^{p}(\mathscr{U}_0(\mathfrak{A}_0),\tau_0) = L_2^{2}(\mathscr{U}_0(\mathfrak{A}_0)\tau_0),$$

for each $p > 2$. Then, by Proposition 5.3, we get

$$\mathfrak{H}_\tau = L_2^{2}(\mathfrak{A}_0,\tau) = L_2^{\omega}(\mathfrak{A}_0,\tau).$$

Thus $\mathfrak{H}_\tau$ is a partial Hilbert algebra since $L_2^{\omega}(\mathfrak{A}_0,\tau)$ is such. But the maximal partial Hilbert algebra contained in $\mathfrak{H}_\tau$ is $(\mathfrak{A}_0)_b$. Hence $\mathfrak{H}_\tau = (\mathfrak{A}_0)_b$, whence $L_2^{\omega}(\mathfrak{A}_0,\tau) = (\mathfrak{A}_0)_b$, a contradiction. This concludes the proof. ⌑

ACKNOWLEDGMENT

The author is grateful to the referee for providing him with helpful comments and suggestions.

[1] J.-P. Antoine and W. Karwowski, "Partial *-algebras of closed operators in Hilbert space," Publ. RIMS Kyoto Univ. **21**, 205 (1985); **22**, 507 (1986).
[2] J.-P. Antoine and F. Mathot, "Partial *-algebras of closed operators and their commutants. I. General structure," Ann. Inst. H. Poincaré **46**, 299 (1987).
[3] J.-P. Antoine, F. Mathot, and C. Trapani, "Partial *-algebras of closed operators and their commutants. II. Commutants and bicommutants," Ann. Inst. H. Poincaré **46**, 325 (1987).
[4] J.-P. Antoine, "States and representations of partial *-algebras," in *Spontaneous Symmetry Breakdown and Related Subjects*, edited by L. Michel, J. Mozrzymas, and A. Pekalski (World Scientific, Singapore, 1985), pp. 247–267.
[5] G. Lassner, "Algebras of unbounded operators and quantum dynamics," Physica A **124**, 471 (1984).
[6] A. Inoue, "On a class of unbounded operator algebras I, II, III," Pacific J. Math. **65**, 77 (1976); **66**, 411 (1976); **69**, 105 (1977).
[7] A. Inoue, "L^p-spaces and maximal unbounded Hilbert algebras," J. Math. Soc. Jpn. **30**, 667 (1978).
[8] G. O. S. Ekhaguere, "Dirichlet forms on partial *-algebras," Math. Proc. Camb. Phil. Soc. **104**, 129 (1988).
[9] M. Takesaki, "Tomita's theory of modular Hilbert algebras and its applications," *Lecture Notes in Mathematics*, Vol. 128 (Springer, Berlin, 1970).
[10] R. Pallu de La Barriere, "Algébres unitaires et espaces d'Ambrose," Ann. Ec. Norm. Sup. **70**. 381 (1953).
[11] A. Inoue, "A generalization of the Tomita–Takesaki theory to partial *Op**-algebras," preprint, Dept. of Applied Mathematics, Fakuoka Univ., Fakuoka, Japan, 1988.
[12] I. M. Gel'fand and G. E. Shilov, *Generalized functions. I. Properties and Operations* (Academic, New York, 1964).
[13] I. M. Gel'fand and N. Ya. Vilenkin, *Generalized functions IV. Applications of Harmonic Analysis* (Academic, New York, 1964).
[14] J.-P. Antoine and A. Inoue, "Unbounded generalization of von Neumann algebras by partial *Op**-algebras," preprint, Institut de Physique Theorique, Université Catholique de Louvain, Belgium, 1987.
[15] T. Ogasawara and K. Yoshinaga, "A noncommutative theory of integration for operators," J. Sci. Hiroshima Univ., **18**, 311 (1955).
[16] E. Nelson, "Non-commutative integration theory," J. Funct. Anal. **15**, 103 (1974).
[17] M. Takesaki, *Theory of operator Algebras* (Springer, Berlin, 1979), Vol. I.
[18] R. Arens, The space L^{ω} and convex topological rings, Bull. Amer. Math. Soc. **52**, 931 (1946).

SU(1,1) and the vector model for continuous bases

E. de Prunelé
Service de Physique des Atomes et des Surfaces, Centre d'Etudes Nucléaires de Saclay, 91191 Gif-sur-Yvette Cedex, France

(Received 30 November 1988; accepted for publication 19 April 1989)

The vector model is formulated in a general form that allows its application for calculating square moduli of matrix elements involving continuous bases. The model is applied to positive discrete unitary irreducible representations of SU(1,1) and the classical estimates are compared with the exact results.

I. INTRODUCTION

The vector model was originally defined for SU(2) (see, e.g., Refs. 1–4). It has recently[5] been applied to positive discrete unitary irreducible representations (UIR's) of SU(1,1) for the case of a discrete basis, i.e., a basis that diagonalizes a compact generator. It is the purpose of the present short paper to set up the formulation of the vector model in a slightly more general form that extends its domain of application. This is done in Sec. II. As a result, various square moduli of matrix elements involving continuous bases of positive discrete UIR's of SU(1,1) are evaluated and compared with the exact results in Secs. III–V. Results obtained from the vector model will be called classical estimates in the following. Section VI concludes this work.

It is emphasized that only positive discrete UIR's are considered throughout this paper.

II. BASIC FORMULATION OF THE VECTOR MODEL

The basic features of the vector model for SU(2) are now briefly recalled. The vector model for SU(2) associates to a vector in the space of the UIR considered a set of classical vectors in a three-dimensional space as follows. The UIR is specified by j, and one has, with standard notation,

$$(J_x^2+J_y^2+J_z^2)|j,m\rangle = j(j+1)|j,m\rangle, \tag{1}$$

$$J_z|j,m\rangle = m|j,m\rangle. \tag{2}$$

Let O be the origin of an orthonormal frame with axes J_x, J_y, J_z. According to the two equations above, the vector model associates to the vector $|j,m\rangle$ a set of vectors with origin O, with square modulus equal to $j(j+1)$, and with projection on the J_z axis equal to m. The remaining components relative to the other axes can be specified by an azimuthal angle φ. The probability density relative to the variable φ is supposed to be a constant by symmetry considerations.

For a proper description of the vector model, two questions remain to be answered. First, how does a group element act on the previous set of classical vectors? Second, how can one determine the classical square modulus for SU(2) matrix elements?

The first question is solved by considering the homomorphism of SU(2) onto the transformation group SO(3). Otherwise stated, the group elements are generated by the differential operators

$$J_x = -i\left(y\frac{d}{dz}-z\frac{d}{dy}\right), \tag{3}$$

and so on by circular permutation for the two other generators. Let t be the parameter for the integral curves of a one-dimensional subgroup generated by exponentiation of a real linear combination of iJ_x, iJ_y, and iJ_z. The elements of this one-dimensional group are represented by $\exp(-\tau\, d/dt)$. The action of this element on a classical vector of origin O and extremity M is specified by requiring that M moves a parameter distance τ along the integral curve of d/dt.

The second question is solved by interpreting the square modulus of exact matrix elements as a probability. The initial assumption of constant probability density for the azimuthal angle φ leads to a determined probability density for the final extremities of the set of vectors that was moved upon the action of the group element. For more details we refer to Refs. 1–4.

The vector model for UIR's of SU(1,1) within a discrete basis can be described essentially in the same way with obvious changes (see Ref. 5). In particular, the homomorphism to be considered is the one of SU(1,1) onto the transformation group SO(2,1). Thus if K_x, K_y, and K_z denote the elements of the Lie algebra of SU(1,1) (see Ref. 6),

$$[K_x,K_y] = -iK_z, \quad [K_y,K_z] = iK_x, \quad [K_z,K_x] = iK_y, \tag{4}$$

one makes the following correspondence:

$$K_x = -i\left(y\frac{d}{dz}+z\frac{d}{dy}\right), \tag{5}$$

$$K_y = -i\left(-z\frac{d}{dx}-x\frac{d}{dz}\right), \tag{6}$$

$$K_z = -i\left(x\frac{d}{dy}-y\frac{d}{dx}\right). \tag{7}$$

When one tries to apply the vector model to more general situations, as, for example, by considering positive discrete UIR's of SU(1,1) in continuous bases, one must have some assumption that generalizes the constant probability density assumption for the azimuthal angle φ. This generalized assumption will be given after some definitions and notation are introduced.

The letter γ for a vector of an UIR of SU(1,1) indicates that this vector has the eigenvalue $\gamma(\gamma+1)$ for the Casimir operator defined by

$$K^2 = K_z^2 - K_x^2 - K_y^2. \tag{8}$$

Let $|\gamma,v,(a,b,c)\rangle$ be a vector of a positive discrete UIR of SU(1,1) specified by γ, which satisfies

 0022-2488/89/091965-06$02.50

$$(aK_x + bK_y + cK_z)|\gamma,v,(a,b,c)\rangle = v|\gamma,v,(a,b,c)\rangle. \quad (9)$$

In the case $a = b = 0$, $c = 1$, one obtains the discrete basis considered in Ref. 5, and v varies from $-\gamma$ to plus infinity by steps of unity. In the case $a = c = 0$, $b = 1$, one obtains a continuous basis and v can vary continuously from minus infinity to plus infinity.[7] The set of classical vectors corresponding to $|\gamma,v,(0,1,0)\rangle$ is shown in Fig. 1. It is the intersection of a hyperboloid with the plane $K_y = v$. (For the particular cases $\gamma = -\frac{1}{2}, -1$, the geometrical pictures are different, as noted in Ref. 5, but this does not change the line of arguments.) It is now clear that the previous set of classical vectors can no longer be parametrized by a parameter that varies on a compact domain, as was the case for the angle φ when considering a discrete basis for UIR's of SU(1,1).[5] It remains to define some kind of probability density for these vectors. The name probability density is used from now on in a rather large sense since we do not require the integral of this probability density over its whole domain to converge. The situation is quite analogous to the one encountered when considering the scalar product between two improper vectors associated to a particle with well defined position and well defined momentum. This scalar product yields a plane wave and the integration of the square modulus with respect to the position or the momentum clearly diverges.

The fundamental hypothesis of the vector model that appears to be the most natural and general is the one of constant probability density with respect to the volume element $dx\,dy\,dz$ in the space generated by K_x, K_y, K_z.

It is clear that the constant probability density hypothesis with respect to the azimuthal angle φ that was used previously is a particular case of the above hypothesis.

As a direct consequence one obtains for the classical estimate of the square modulus corresponding to different bases the following expression:

$$|\langle\gamma,v,(a,b,c)|\gamma,v',(a',b',c')\rangle|^2 \simeq A \sum_n J(x_n,y_n,z_n;\gamma,v,v'), \quad (10)$$

where A is independent of v and v' and remains to be determined. Here J is the absolute value of the Jacobian,

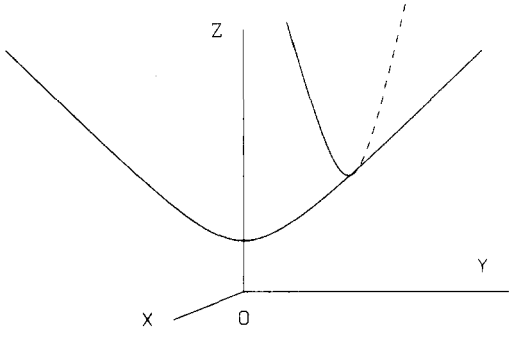


FIG. 1. Hyperboloid associated with a positive discrete UIR. The vectors whose common origin is O and whose extremities lie at the intersections of the hyperboloid with the plane $y = v$ are associated with the state $|\gamma,v,(0,1,0)\rangle$ (see the text).

$$J = \begin{vmatrix} \dfrac{dx}{d(\gamma(\gamma+1))} & \dfrac{dx}{dv} & \dfrac{dx}{dv'} \\ \dfrac{dy}{d(\gamma(\gamma+1))} & \dfrac{dy}{dv} & \dfrac{dy}{dv'} \\ \dfrac{dz}{d(\gamma(\gamma+1))} & \dfrac{dz}{dv} & \dfrac{dz}{dv'} \end{vmatrix}, \quad (11)$$

of the transformation given by

$$\gamma(\gamma+1) = z^2 - y^2 - x^2, \quad (12)$$

$$v = ax + by + cz, \quad (13)$$

$$v' = a'x + b'y + c'z, \quad (14)$$

evaluated at all the real points (x_n,y_n,z_n), which are the solutions of the three equations above, with the restriction that negative values of z_n should be rejected because we are considering positive discrete UIR's. In practice, zero, one, or two points satisfy these conditions. If there is no solution, the classical estimate is zero and one says that the matrix element lies in the classically forbidden domain.

The results obtained from Eqs. (10)–(14) will be compared with the exact results in Sec. III, and they are the starting point for the other applications considered in Secs. IV and V.

III. CONNECTION BETWEEN DIFFERENT BASES

The SU(1,1) UIR's in continuous bases have been studied by different authors (see, e.g., Ref. 7–9). The classical estimates given by Eq. (10) will be compared to the exact results obtained by Linblad and Nagel[7] for some different bases. For practical calculations, it appears more convenient to use Eq. (10) in the following form:

$$|\langle\gamma,v,(a,b,c)|\gamma,v',(a',b',c')\rangle|^2 \simeq A \sum_n [J(\gamma,v,v';x_n,y_n,z_n)]^{-1}. \quad (15)$$

The three different cases studied in Ref. 7 are now considered.

First, Eq. (15) yields for the classical estimate

$$|\langle\gamma,v,(0,0,1)|\gamma,v',(0,1,0)\rangle|^2 \simeq A\,[v^2 - \gamma(\gamma+1) - v'^2]^{-1/2}. \quad (16)$$

The normalization constant can now be determined by requiring that the integral over v' inside the classical domain be unity. One finds

$$A = 1/\pi. \quad (17)$$

This γ-independent value for A will be retained in all subsequent calculations. Equations (16) and (17) have to be compared with the result obtained from Eq. (4.17) of Ref.7, which involves gamma functions and a hypergeometric ${}_2F_1$ function of argument -1. The numerical comparison for the case $\gamma = -\frac{5}{2}$, $v = \frac{5}{2}$ is reported in Fig. 2. Only positive values of v' have been considered, since both results are even functions of v'. It is seen on this figure that the exact result decreases very rapidly outside the classical domain. It oscillates around the classical estimate inside the classical domain. At the limit of the classical domain the classical estimate diverges.

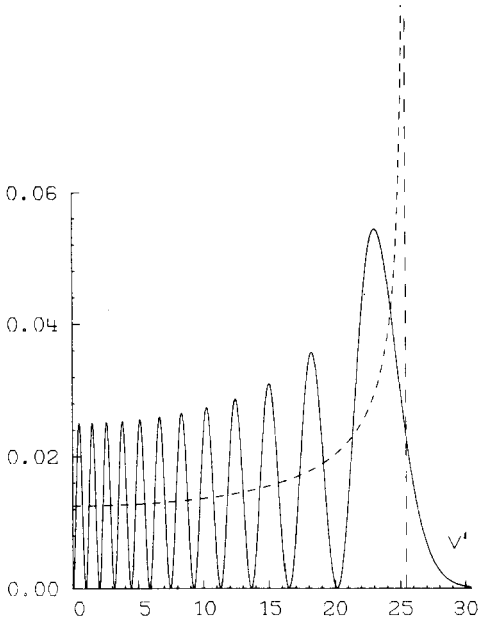


FIG. 2. Square modulus of $\langle \gamma = -\frac{5}{2}, v = \frac{51}{2}, (0,0,1) | \gamma = -\frac{5}{2}, v', (1,0,1) \rangle$ as a function of v'. The vertical dashed line corresponds to the limit of the classical domain. The dashed line curve inside the classical domain represents the classical estimate.

For the computation of the exact result we use a three term recursion relation with respect to v [see Eq. (4.4) of Ref. 7]. For $v = -\gamma$ this relation becomes a two term recursion relation and the initial value is easily computed since the hypergeometric function is then equal to unity.

Second, Eqs. (15) and (17) yield for the classical estimate

$$|\langle \gamma,v,(0,0,1)|\gamma,v',(1,0,1)\rangle|^2 \simeq [2vv' - v'^2 - \gamma(\gamma+1)]^{-1/2}/\pi. \quad (18)$$

[It is recalled that the v' corresponding to the (1,0,1) basis must be positive for positive discrete UIR's (see Ref. 7).] This has to be compared with the result obtained from Eq. (5.10) of Ref. 7, which involves gamma functions and a Whittaker function. The numerical comparison for the case $\gamma = -25, v = 35$ is reported in Fig. 3. It is seen again on this figure that the exact result decreases very rapidly outside the classical domain. It oscillates around the classical estimate inside the classical domain. At the limits of the classical domain the classical estimate diverges.

It can be noticed that for integer values of $-\gamma$ the exact result given by Eq. (5.10) of Ref. 7 can be rewritten as

$$\langle \gamma,v,(0,0,1)|\gamma,v',(1,0,1)\rangle = v^2 (v')^{1/2} R_{v,-\gamma-1}(vv'), \quad (19)$$

where $R_{n,l}(r)$ denotes the radial hydrogenic wave function corresponding to principal and orbital quantum numbers n and l. According to this relation we have used the algorithm described in Ref. 10 for the numerical computations reported in Fig. 3. A more direct and economical method, which is also valid for half-integer values of $-\gamma$, can be developed on the basis of the three term recursion relation with respect to v [see Eq. (5.3) of Ref. 7].

Third and finally, Eqs. (15) and (17) yield for the classical estimate

$$|\langle \gamma,v,(0,1,0)|\gamma,v',(1,0,1)\rangle|^2 \simeq (2\pi v')^{-1}. \quad (20)$$

This v-independent expression corresponds exactly to the exact result obtained from Eq. (5.20) of Ref. 7.

IV. MATRIX ELEMENTS OF FINITE TRANSFORMATIONS

For comparison with the exact results of Ref. 7 we shall consider successively

$$\begin{aligned} C1 &\equiv |\langle \gamma,v',(0,1,0)|\exp(-i\tau(K_x+K_z))|\gamma,v,(0,1,0)\rangle|^2, \\ C2 &\equiv |\langle \gamma,v',(0,1,0)|\exp(-i\tau K_z)|\gamma,v,(0,1,0)\rangle|^2, \\ C3 &\equiv |\langle \gamma,v',(1,0,1)|\exp(-i\tau K_z)|\gamma,v,(1,0,1)\rangle|^2, \quad (21) \\ C4 &\equiv |\langle \gamma,v',(1,0,1)|\exp(-i\tau K_y)|\gamma,v,(1,0,1)\rangle|^2. \end{aligned}$$

First we consider C1. A set of classical vectors with constant value of y, $y = v$, corresponds to the state $|\gamma,v,(0,1,0)\rangle$. According to Eqs. (5)–(7) the integral curves of the parabolic generator $K_x + K_z = -i\,d/dt$ can be parametrized as follows:

$$z + x = w, \quad (22)$$

$$x = -\tfrac{1}{2}wt^2 + (w^2 - \gamma(\gamma+1))/(2w), \quad (23)$$

$$y = wt. \quad (24)$$

Under the action of $\exp(-i\tau(K_x+K_z))$, a point on the hyperboloid, characterized by γ, w, t, is moved to the point characterized by $\gamma, w, t+\tau$. In particular, y becomes $y + w\tau$. Therefore the probability density relative to y in the final configuration is determined by the probability density relative to w in the initial configuration. This latter one is given by

$$|\langle \gamma,w,(1,0,1)|\gamma,v,(0,1,0)\rangle|^2.$$

One therefore has

$$C1 \simeq |\langle \gamma,(v'-v)/\tau,(1,0,1)|\gamma,v,(0,1,0)\rangle|^2/\tau. \quad (25)$$

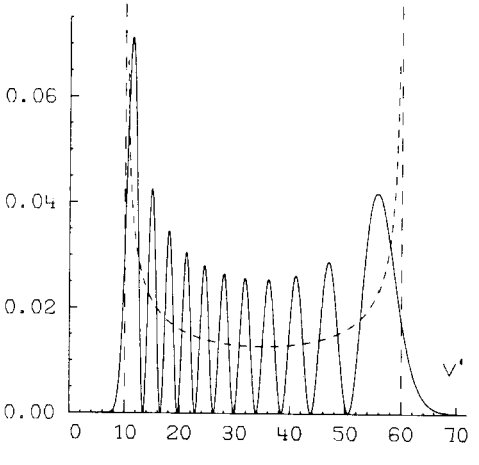


FIG. 3. Square modulus of $\langle \gamma = -25, v = 35, (0,0,1) | \gamma = -25, v', (1,0,1) \rangle$ as a function of v'. The vertical dashed lines correspond to the limits of the classical domain. The dashed line curve inside the classical domain represents the classical estimate.

Using Eq. (20) one finally obtains

$$C1 \simeq \begin{cases} \theta(v'-v)[2\pi(v'-v)]^{-1}, & \text{if } \tau \text{ is positive,} \\ \theta(v-v')[2\pi(v-v')]^{-1}, & \text{if } \tau \text{ is negative,} \end{cases} \tag{26}$$

where θ denotes the Heaviside function [$\theta(x) = 0$, if x negative, $= 1$, otherwise]. This has to be compared with the following exact expression, obtained from Eq. (6.2) of Ref. 7:

$$C1 = \begin{cases} [2\pi(v'-v)(1-\exp(-2\pi(v'-v)))]^{-1}, \\ \text{if } \tau \text{ positive,} \\ [2\pi(v'-v)(-1+\exp(2\pi(v'-v)))]^{-1}, \\ \text{if } \tau \text{ negative.} \end{cases} \tag{27}$$

The classical estimate therefore is a good approximation provided that the absolute value of $v'-v$ not be too small.

For C2 one has to consider the integral curves of K_z. After the action of $\exp(-i\tau K_z)$, the initial value v for y becomes $v\cos(\tau) + x\sin(\tau)$. Therefore the final probability density relative to y is determined by the initial probability density relative to x. This latter one is given by

$$|\langle \gamma,x,(1,0,0)|\gamma,v,(0,1,0)\rangle|^2.$$

One therefore has

$$C2 \simeq |\langle \gamma,(v'-v\cos(\tau))/\sin(\tau),(1,0,0)|\gamma,v,(0,1,0)\rangle|^2/|\sin(\tau)|. \tag{28}$$

Using Eqs. (15) and (17) one finally obtains

$$C2 \simeq (2\pi)^{-1}[\gamma(\gamma+1)\sin^2(\tau) + v^2 + v'^2 - 2vv'\cos(\tau)]^{-1/2}. \tag{29}$$

The exact result [see Eq. (6.4) of Ref. 7] involves hypergeometric functions ${}_2F_1$ but simplifies greatly in the case $v = 0$, $\gamma = -1$. For this particular case, one obtains

$$\begin{aligned} &|\langle \gamma = -1,v',(0,1,0)| \\ &\quad \times \exp(-i\tau K_z)|\gamma = -1, v = 0,(0,1,0)\rangle|^2 \\ &\quad = (2\pi v')^{-1}[\coth(\pi v') \\ &\qquad - \cos(2v'\ln|\tan(\tau/2)|)/\sinh(\pi v')]. \end{aligned} \tag{30}$$

It is seen that the classical estimate then provides a good approximation if v' is large.

For C3 one again has to consider the integral curves of K_z. After the action of $\exp(-i\tau K_z)$, the coordinate y becomes $y\cos(\tau) + x\sin(\tau)$, the coordinate x becomes $x\cos(\tau) - y\sin(\tau)$, and the coordinate z remains unchanged. The initial value v for $x+z$ thus becomes $v + (\cos(\tau) - 1)x - \sin(\tau)y$. Therefore the probability density relative to $x+z$ in the final configuration is equal to the probability density relative to $(\cos(\tau) - 1)x - \sin(\tau)y$ in the initial configuration.

One therefore has

$$C3 \simeq |\langle \gamma,v' - v,(\cos(\tau) - 1, -\sin(\tau),0)|\gamma,v,(1,0,1)\rangle|^2. \tag{31}$$

Using Eqs. (15) and (17) one finally obtains after some calculations

$$C3 \simeq (2\pi|\sin(\tau/2)|)^{-1}[vv' - \gamma(\gamma+1)\sin^2(\tau/2)]^{-1/2}. \tag{32}$$

This has to be compared with the exact result obtained from Eq. (6.6) of Ref. 7:

$$C3 = [J_{-2\gamma-1}(2|vv'|^{1/2}/|\sin(\tau/2)|)/\sin(\tau/2)]^2, \tag{33}$$

where J denotes the Bessel function regular at the origin.[11] It is seen that in the case $\gamma = -\frac{1}{2}$ and v and v' equal to zero, the two results are identical. It is also seen[11] that, in the limit where $vv'/\sin^2(\tau/2)$ goes to infinity,

$$C3 = [\pi vv']^{-1/2}\cos^2(2|vv'|^{1/2}/|\sin(\tau/2)| + (2\gamma+1)\pi/2 - \pi/4). \tag{34}$$

In this limit it is clear that the exact result oscillates around the classical estimate.

For C4 one has to consider the integral curves of the hyperbolic generator $K_y = -i\,d/dt$. According to Eqs. (5)–(7), one has the following parametrization:

$$x = -[\gamma(\gamma+1) + y^2]^{1/2}\sinh(t), \tag{35}$$

$$z = [\gamma(\gamma+1) + y^2]^{1/2}\cosh(t). \tag{36}$$

Under the action of the operator $\exp(-i\tau K_y)$, it is clear that $x+z$ becomes $(x+z)\exp(-\tau)$. Therefore C4 should be zero except for $v = v'\exp(\tau)$. The exact result given by Eq. (6.7) of Ref. 7 is

$$\begin{aligned} &\langle \gamma,v',(1,0,1)|\exp(-i\tau K_y)|\gamma,v,(1,0,1)\rangle \\ &\quad = (v/v')^{1/2}\delta(v - v'\exp(\tau)), \end{aligned} \tag{37}$$

where δ denotes the Dirac function.

V. CLEBSCH–GORDAN COEFFICIENTS

The exact expression for Clebsch–Gordan coefficients coupling two positive discrete UIR's of SU(1,1) in the y-continuous basis has been given by Mukunda and Radhakrishnan.[12] According to our previous notations, these coefficients take the form[12]

$$\begin{aligned} &\langle \gamma,v,(0,1,0);\gamma'v',(0,1,0)|\gamma'',v'',(0,1,0)\rangle \\ &\quad = \delta(v+v'-v'')D(\gamma,v,\gamma',v',\gamma''). \end{aligned} \tag{38}$$

Conservation of K_y implies that there will always be a Dirac δ function in front of the right-hand side of Eq. (38).[12] One considers γ, v, γ', and v' all fixed. The diagram of Fig. 4 corresponds to the vector relation $K'' = K + K'$. Here OO' represents a vector associated to $|\gamma,v,(0,1,0)\rangle$. The second coordinate of O' is equal to v. To the state $|\gamma',v',(0,1,0)\rangle$ are associated the vectors O'M, where M is on the inner hyperboloid. The set of points M have fixed second coordinate v' with respect to the frame O (see Fig. 4). To ensure that M belongs to an hyperboloid characterized by γ'' relative to the first frame O_{xyz}, one must have

$$\begin{aligned} c''^2 &= (z+z')^2 - (x+x')^2 \\ &= [(c^2+x^2)^{1/2} + (c'^2+x'^2)^{1/2}]^2 - (x+x')^2 \\ &\equiv f(c,c',x,x'), \end{aligned} \tag{39}$$

where

$$c''^2 \equiv \gamma''(\gamma''+1) + v''^2, \tag{40}$$

and v'' is equal to the sum of v and v'. We also define c^2 in terms of γ, v, and c'^2 in terms of γ' and v', as in Eq. (40). The classical estimate for the probability density relative to c''^2 follows by integrating over all positions of O' and M on the hyperboloids for v and v' fixed and subject to the condition given by Eq. (39). Each position of O' and M has to be

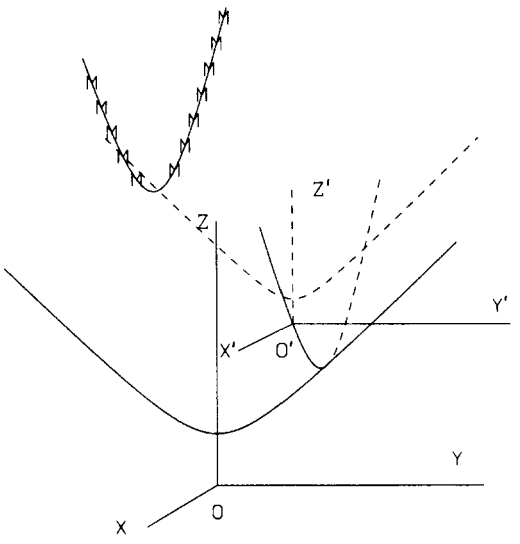


FIG. 4. Diagram corresponding to the coupling of two positive discrete UIR's in a continuous basis. The point O' moves on the intersection of the outer hyperboloid with the plane $y = v$ (see the text).

weighted by its probability density. The probability density relative to x for v fixed is, according to Eqs. (15) and (17),

$$|\langle \gamma,x,(1,0,0)|\gamma,v,(0,1,0)\rangle|^2 = (2\pi)^{-1}(c^2+x^2)^{-1/2}. \tag{41}$$

As noted in Sec. II, this is an improper probability density since its integral with respect to x diverges, as can be clearly seen from the change of variable $x = c\sinh(\varphi)$:

$$(2\pi)^{-1}\int dx(c^2+x^2)^{-1/2} = (2\pi)^{-1}\int d\varphi. \tag{42}$$

To obtain the probability density relative to c''^2 one therefore has to consider the following divergent integral, to be denoted I:

$$I = (2\pi)^{-2}\int dx\int dx'[(c^2+x^2)(c'^2+x'^2)]^{-1/2} \times\delta[f(c,c',x,x') - c''^2]. \tag{43}$$

The above integral can be factorized into two parts, one of them corresponding to the improper normalization given by Eq. (42). This is achieved by performing the above change of variables for x and x'. One finally obtains

$$I = \left[(4\pi)^{-1}\int d(\varphi-\varphi')\delta\left[cc'\cosh(\varphi-\varphi') - \frac{c''^2-c'^2-c^2}{2}\right]\right](2\pi)^{-1}\int d(\varphi+\varphi'). \tag{44}$$

The first integral in the above product should provide the probability density relative to c''^2. Using Eq. (40) one sees that the probability density relative to γ'' is obtained by multiplying by the factor $-(2\gamma''+1)$. One finally obtains for the classical estimate of the square modulus of the function D in Eq. (38) the following expression:

$$|D(\gamma,v,\gamma',v',\gamma'')|^2 \simeq[(-\gamma''-\tfrac{1}{2})/\pi][c''^4+c^4+c'^4-2c^2c'^2 - 2c''^2c'^2-2c''^2c^2]^{-1/2}. \tag{45}$$

The exact result [see Eq. (4.11) of Ref. 12] involves a hypergeometric function ${}_3F_2$ of argument unity. The general expression, however, simplifies drastically in the case $\gamma = -1$ and $v = 0$. One obtains, from Eq. (4.11) of Ref. 12,

$$|D(-1,0,\gamma',v',\gamma'')|^2 = (-\gamma''-\tfrac{1}{2})/[\pi(\gamma''+\gamma'+1)(\gamma''-\gamma')]. \tag{46}$$

It is seen that this v'-independent result corresponds exactly to the classical estimate obtained from Eq. (45).

VI. DISCUSSION

The classical estimate should not be used directly for the sake of accurate numerical computation of some given square moduli of matrix elements. The classical estimate can indeed oscillate round the exact value and is a very poor approximation at the limit of the classical domain. A reasonable numerical accuracy is, however, expected if the classical domain coincides with the exact domain. In this case the classical estimate can even coincide with the exact results as illustrated by some examples considered in this paper. Among the principal interests of the vector model we emphasize the following three points.

First, it provides the so-called classical domain. The exact results decrease very rapidly outside this domain. This is of considerable interest for noncompact groups because one thus has a criterion for truncating infinite expansions in practical calculations.[5]

Second, it provides a simple geometric interpretation of many results. Most of these results could probably be obtained by other methods, but the geometrical point of view is particularly suitable for developing a rapid and global understanding.

Third and finally, it is essential for developing algorithms for the computation of the exact results: These exact results generally satisfy three term recursion relations, which are numerically stable only when progressing toward the classical domain.[3,5,10,13]

The vector model could also be applied to the continuous UIR's of SU(1,1). (The case of negative discrete UIR's is obtained directly from the case of positive UIR's.) There should be no basic difficulties for this extension of the present work. The comparison with the exact results of continuous UIR's is, however, more difficult because they generally do not simplify for particular cases and the initialization of the three term recursion relations involved[7] becomes more complicated.

ACKNOWLEDGMENTS

The author would like to thank M. Poirier for many helpful discussions, and B. Carré and J. Pascale for a careful reading of the manuscript.

[1] E. P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra* (Academic, New York, 1959).
[2] P. J. Brussaard and H. A. Tolhoek, Physica **23**, 955 (1957).

[3]K. Schulten and R. G. Gordon, J. Math. Phys. **16**, 1971 (1975).
[4]L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics: Theory and Application*, Vol. 8, *Encyclopedia of Mathematics and its Applications* (Addison–Wesley, Reading, MA, 1981).
[5]E. de Prunelé, J. Math. Phys. **29**, 2523 (1988).
[6]We depart here from the notations of Ref. 5 where the covariant contravariant indices were very suitable for considering vector operators. In the present work they are not useful.
[7]G. Linblad and B. Nagel, Ann. Inst. H. Poincaré **13**, 27 (1970).
[8]A. O. Barut and E. C. Phillips, Commun. Math. Phys. **8**, 52 (1968).
[9]N. Mukunda, J. Math. Phys. **10**, 2086, 2092 (1969).
[10]E. de Prunelé, J. Math. Phys. **25**, 472 (1984).
[11]M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965), 5th ed.
[12]N. Mukunda and B. Radhakrishnan, J. Math. Phys. **15**, 1320 (1974).
[13]K. Schulten, and R. G. Gordon, J. Math. Phys. **16**, 1961 (1975).

Finite geometries and Clifford algebras

R. Shaw
School of Mathematics, University of Hull, Hull, HU6 7RX, United Kingdom

(Received 16 March 1989; accepted for publication 12 April 1989)

The Clifford algebra in dimension $d = 2^{m+1} - 1$, $m \geqslant 2$, is treated using the finite m-dimensional projective geometry PG(m,2) over the field of order 2. The incidence properties of the geometry help in the problem of finding a complete commuting set of operators with which to label the $2^{(d-1)/2}$ spinor states of an irreducible representation. Full details are given in the case $m = 3$, $d = 15$, thus generalizing previous work for the $m = 2$, $d = 7$ case, and various conjectures are made concerning the cases $m > 3$.

I. INTRODUCTION

We will deal throughout with an irreducible representation of the real Clifford algebra Cliff(0,d), $d = 2^{m+1} - 1$, for some $m = 2,3,...$, in which the operators $\Gamma_1, \Gamma_2,...,\Gamma_d$ (representing a set of anticommuting imaginary units generating the algebra) satisfy not only the relations

$$(\Gamma_i)^2 = -I, \quad \Gamma_i\Gamma_j = -\Gamma_j\Gamma_i, \quad i \neq j, \tag{1.1}$$

but also the relation

$$\Gamma_1\Gamma_2\cdots\Gamma_d = +I. \tag{1.2}$$

Since the cases with which we deal satisfy $d \equiv 7 \pmod 8$, then, as is well known,[1-6] the operators Γ_i can be taken to be skew-adjoint operators acting upon a Euclidean space of real spinors of dimension $2^{(d-1)/2}$.

We shall be interested in the finite group G_0 generated by the operators Γ_i,

$$G_0 = \{\pm I, \pm\Gamma_i, \pm\Gamma_i\Gamma_j, \pm\Gamma_i\Gamma_j\Gamma_k,...\}. \tag{1.3}$$

On account of the relations (1.1) and (1.2), the display (1.3) will list each element of G_0 once, and only once, provided that we impose the restrictions $i < j < k < \cdots$, and provided also that we consider products of at most $\frac{1}{2}(d-1)$ of the operators Γ_i. Consequently G_0 has order

$$|G_0| = 2^d. \tag{1.4}$$

One of our chief concerns will be to determine a suitable maximal Abelian subgroup $H^{\max}$ of G_0. A set of independent generators of $H^{\max}$ can then (after the omission of $-I$) be used as a complete commuting set of operators, whose sets of simultaneous eigenvalues will therefore distinguish between the $2^{(d-1)/2}$ different spinor states. A solution to the problem in the case $d = 7$ was spelled out in a previous paper[7] (cf. also the earlier papers of Refs. 8 and 9) where it proved to be useful to associate the seven operators Γ_i with the seven points of the two-dimensional projective geometry PG(2,2) over the field $\mathbb{F}_2 = \{0,1\}$ of order 2. The present paper will demonstrate the relevance, and elegance, of methods based upon the corresponding m-dimensional projective geometry PG(m,2) over $\mathbb{F}_2$ when applied to the case of the Clifford algebra Cliff$(0,2^{m+1}-1)$, $m = 3,4,...$.

To this end, let us quotient out G_0 by its center,

$$Z_0 = Z(G_0) = \{I, -I\}, \tag{1.5}$$

which is seen also to be the commutator subgroup of G_0. We thereby obtain an Abelian group

$$C_0 = G_0/Z_0 \tag{1.6}$$

of order 2^{d-1}. Since the square of each element of G_0 is either $+I$ or $-I$, the square of each element of C_0 is the identity 1 ($= Z_0$). Consequently,

$$C_0 \simeq (\mathbb{Z}_2)^{d-1} = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \cdots \times \mathbb{Z}_2 \quad (d-1 \text{ factors}), \tag{1.7}$$

where $\mathbb{Z}_2$ denotes the (multiplicative) group of order 2. Under the canonical projection π: $G_0 \to C_0$ let s_i denote the image of Γ_i,

$$s_i = \{\Gamma_i, -\Gamma_i\} \in C_0. \tag{1.8}$$

The elements $s_1,s_2,...,s_d$ generate C_0 and by (1.1) of course satisfy

$$(s_i)^2 = 1, \quad s_is_j = s_js_i. \tag{1.9}$$

However, they are not independent generators, since by (2) they are subject to the relation

$$s_1s_2\cdots s_d = 1. \tag{1.10}$$

Corresponding to (1.3) we have the display

$$C_0 = \{1, s_i, s_is_j, s_is_js_k, ...\}, \tag{1.11}$$

where $i < j < k < \cdots$ and where products of at most $\frac{1}{2}(d-1)$ of the s_i are to be listed.

Consider the set

$$S = \{s_1,s_2,...,s_d\}. \tag{1.12}$$

Given a subset $\alpha \subseteq S$, we will term α a small or large subset according as $|\alpha| \leqslant \frac{1}{2}(d-1)$ or $|\alpha| \geqslant \frac{1}{2}(d+1)$. With a view to an imminent geometrical interpretation let us refer to the small subsets of S as *figures*. These figures are in a one–one correspondence with the elements of C_0; for example, the figure $\{s_2,s_3,s_5\}$ corresponds to the element $s_2s_3s_5$ of C_0, and $\varnothing$ corresponds to 1. In fact *we will identify C_0 with the set of all figures of S*, the group multiplication of figures α, β being therefore defined by

$$\alpha\beta = \begin{cases} \alpha\Delta\beta, & \text{if } \alpha\Delta\beta \text{ is small,} \\ (\alpha\Delta\beta)^c & \text{if } \alpha\Delta\beta \text{ is large,} \end{cases} \tag{1.13}$$

and satisfying

$$\alpha^2 = 1 (= \varnothing) \text{ and } \alpha\beta = \beta\alpha, \quad \text{for all } \alpha,\beta \in C_0. \tag{1.14}$$

Here γ^c denotes the complement in S of a subset γ, and $\alpha\Delta\beta$ denotes the symmetric difference of the subsets α, β,

$$\alpha\Delta\beta = (\alpha \cap \beta^c) \cup (\beta \cap \alpha^c) = (\alpha \cup \beta) \setminus (\alpha \cap \beta).$$

 0022-2488/89/091971-14$02.50

As is very well known, the set S^* consisting of *all* the subsets of S is an Abelian group under the operation Δ, the identity element being the empty set $\emptyset$; moreover, each $\alpha \neq \emptyset$ is of order 2; $\alpha\Delta\alpha = \emptyset$. The extra twist in the definition (1.13) of the multiplication law for figures stems from (1.10), or in turn from (1.2): a subset α of S gives rise to the same group element in C_0 as the complementary subset α^c.

What we are doing can be phrased in slightly different languages. Loosely speaking, we can think of C_0 as the group (S^*,Δ) of all the subsets of S, *except that we identify a subset with its complement*. That is, we introduce an equivalence relation $\sim$ on subsets by

$$\alpha \sim \beta \quad \text{if and only if } \alpha = \beta \text{ or } \alpha = \beta^c. \tag{1.15}$$

Better, perhaps, we should talk in terms of *configurations*, say, rather than figures, where a configuration is defined to consist of a (nonordered) pair $\{\alpha,\alpha^c\}$ of complementary subsets of S. Multiplication of configurations is by way of

$$\{\alpha,\alpha^c\}\{\beta,\beta^c\} = \{\gamma,\gamma^c\} \tag{1.16}$$

where

$$\gamma = \alpha\Delta\beta = \alpha^c\Delta\beta^c, \quad \gamma^c = \alpha^c\Delta\beta = \alpha\Delta\beta^c \tag{1.17}$$

In other (group theory) words, we define C_0 to be the quotient of S^* by the subgroup $\{\emptyset,S\}$ (the cosets of this subgroup being the configurations $\{\alpha,\alpha^c\}$). However, we can always unambiguously (since d is odd) choose the smaller of the sets α,α^c as coset representative of the coset $\{\alpha,\alpha^c\}$, and end up with our original view of C_0 in terms of the multiplication (1.13) of figures.

At first sight, it might appear that the Abelian group $C_0 \simeq (\mathbb{Z}_2)^{d-1}$ is utterly trivial, and that our sole concern should be with the non-Abelian group G_0. Surprisingly, as we shall see in Secs. II and III, certain interesting and nontrivial questions arise, and can be settled, even at the Abelian level—once, that is, *we bring C_0 to geometric life* by viewing the d elements s_i of the set S as the points of the finite m-dimensional projective geometry $\mathrm{PG}(m,2)$, $d = 2^{m+1} - 1$.

Before going into details, we should confess at the outset that our geometrical interpretation involves a reduction in symmetry. For each identification of the elements of S with the points of $\mathrm{PG}(m,2)$ will single out certain triplet figures as privileged, to be interpreted as the lines of $\mathrm{PG}(m,2)$. Now both C_0 and G_0 come along with a rich supply of automorphisms. (In the case of the Abelian group C_0, any automorphism, other than the identity automorphism, is of course an outer automorphism.) In particular, each element of $\sigma \in \mathcal{S}_d$ (= symmetric group on d symbols) gives rise to an element, also denoted σ, of Aut C_0 by acting upon the generators $s_1,\ldots,s_d$ in the obvious way. Each $\sigma(\neq 1) \in \mathcal{S}_d$ also induces an outer automorphism $\bar{\sigma}$, say, of G_0, defined on generators $\Gamma_1,\ldots,\Gamma_d$ by

$$\bar{\sigma}(\Gamma_p) = (\mathrm{sgn}\,\sigma)\Gamma_{\sigma(p)}. \tag{1.18}$$

Now the projectivities of $\mathrm{PG}(m,2)$, which (in the present case of $\mathbb{F}_2$) can be identified with the elements $A \in \mathrm{GL}(m+1,2) \equiv \mathrm{GL}(m+1,\mathbb{F}_2)$, form a privileged subgroup Ω, say, of $\mathcal{S}_d$: elements of Ω are picked out by the property that they map lines onto lines. Thus our geometrical interpretation naturally goes along with selecting as privileged a certain subgroup, also denoted Ω, of Aut C_0. (Elements of Aut C_0 not lying in Ω will not map lines onto lines; indeed they will not even map S onto S except when they lie in $\mathcal{S}_d$.) Similarly, we shall in effect be concerned with a corresponding subgroup $\bar{\Omega}$, say, of Aut G_0, consisting of the outer automorphisms $\bar{A}$ arising from the elements $A \in \mathrm{GL}(m+1,2)$.

Nevertheless, we shall be able to turn the loss of symmetry to our advantage, since in our search for a maximal Abelian subgroup $H^{\max}$ of G_0 our attention will be usefully directed, in the first instance, to certain subgroups H of G_0 which are $\bar{\Omega}$-admissible, satisfying that is

$$\bar{A}(H) = H, \quad \text{for all } A \in \mathrm{GL}(m+1,2). \tag{1.19}$$

(Actually in the cases $m = 3, 5, 7,\ldots$, the maximal Abelian subgroups which we unearth will turn out to be preserved only under certain proper subgroups of $\bar{\Omega}$.)

II. PG $(m,2)$ AND THE GROUP C_0

A. PG$(m,2)$

In the sequel, we will need very little more than the usual elementary incidence properties shared by projective geometries over an arbitrary field $\mathbb{F}$, together with some (easily computed) numbers relevant to the case in hand when $\mathbb{F} = \mathbb{F}_2$. Nevertheless, let us commence this section by listing the relevant facts, and let us also point out the occasional special feature peculiar to the choice $\mathbb{F} = \mathbb{F}_2$. (For more information on finite projective geometries see, for example, the texts by Hirschfeld.[10,11])

Let $V(m+1)$ denote a vector space of dimension $m+1$ over $\mathbb{F}_2$. Since there are no scalars in $\mathbb{F}_2$ other than 0, 1, $V(m+1)$ is essentially nothing more than an additive Abelian group which possesses $m+1$ independent generators of order 2. Each choice of ordered basis yields an isomorphism of $V(m+1)$ with $(\mathbb{F}_2)^{m+1}$. In particular, $|V(m+1)| = 2^{m+1}$. The points of the projective geometry $\mathrm{PG}(m,2)$ can be taken to be the set

$$S_0 = S_0(m) = V(m+1)\setminus\{0\} \tag{2.1}$$

of nonzero vectors of $V(m+1)$. (For general $\mathbb{F}$, one obtains the projective points by way of an equivalence relation upon S_0, setting $y \sim x$ whenever $y = \lambda x$, $\lambda \in \mathbb{F}\setminus\{0\}$; but in our case of $\mathbb{F}_2$ the only choice of λ is 1!) A (projective) subspace of dimension r of the projective geometry $\mathrm{PG}(m,2)$ can be defined to be a subset $\alpha \subseteq S_0$ such that the extended set $\alpha \cup \{0\}$ forms a (vector) subspace of dimension $r+1$ of $V(m+1)$. Subspaces of $\mathrm{PG}(m,2)$ of (projective) dimensions $0,1,2,3,\ldots,r,\ldots,(m-1)$ are called, respectively, (projective) points, lines, planes, solids, ..., r-spaces,..., hyperplanes. The empty subset of S_0 has projective dimension -1.

We denote by $S_r = S_r(m)$ the set consisting of all the r-spaces of $\mathrm{PG}(m,2)$. Putting

$$N(r,m) = |S_r(m)|,$$

then of course, from (2.1), we have

$$N(0,m) = 2^{m+1} - 1.$$

To compute $N(r,m)$ for $r > 0$, let $M(r,m)$ denote the number of ordered r-simplices in $\mathrm{PG}(m,2)$. Here, by an ordered r-simplex we mean an ordered set of $r+1$ linearly indepen-

dent points (which therefore span a projective r-space). Clearly

$$M(r,m) = (2^{m+1}-1)(2^{m+1}-2)\cdots(2^{m+1}-2^{r-1})(2^{m+1}-2^r),$$

and

$$N(r,m) = \frac{M(r,m)}{M(r,r)} = \frac{(2^{m+1}-1)(2^m-1)\cdots(2^{m-r+1}-1)}{(2^{r+1}-1)(2^r-1)\cdots(2-1)}. \tag{2.2}$$

Incidentally, note that the order of $\mathrm{GL}(m+1,2)$ is, equally clearly, given by

$$|\mathrm{GL}(m+1,2)| = M(m,m).$$

A choice of a particular m-simplex in $\mathrm{PG}(m,2)$ is referred to as a choice of a *simplex of reference*. Usually in order to introduce (homogeneous) coordinates into an m-dimensional projective space, one has to make a choice not only of an ordered m-simplex of reference, but also to make the further choice of a "unit point," that is, a point not belonging to any face of the simplex. However, in the case of $\mathrm{PG}(m,2)$, each m-simplex defines a unique unit point, which we will call the *center* of the simplex. For if $p_1,p_2,\ldots,p_{m+1}$ are the vertices of a simplex of reference, and if $x = x_1p_1 + x_2p_2 + \cdots + x_{m+1}p_{m+1}$ lies in none of the faces of the simplex, then we require $x_i \neq 0$ for each $i = 1,2,\ldots,m+1$, whence for $x_i \in \mathbb{F}_2$ we have $x_i = 1$, and x is uniquely given to be $x = c$ where

$$c = p_1 + p_2 + \cdots + p_{m+1}. \tag{2.3}$$

For example, in the case $m = 3$, if p_1,p_2,p_3,p_4 denote the vertices of a tetrahedron of reference, and if $ijkl$ is a permutation of 1234, then p_j,p_k,p_l can be taken to be the vertices of a triangle of reference for the plane ($m = 2$) geometry of the ith face, the remaining four points of this geometry being the "midpoints" $p_k + p_l, p_j + p_l, p_j + p_k$ of the sides of the triangle along with the center $c_i = p_j + p_k + p_l$ of the triangle. The four faces together account for 14 points of $\mathrm{PG}(3,2)$, the remaining 15th point being the center $c = p_1 + p_2 + p_3 + p_4$ of the tetrahedron.

A *nonordered* set of $r+2$ points in $\mathrm{PG}(m,2)$, which consists of an r-simplex together with the center of this simplex, will be termed an r-frame. Clearly, the number of distinct r-frames is $M(r,m)/(r+2)!$ Thus in $\mathrm{PG}(2,2)$ there are seven two-frames, while in $\mathrm{PG}(3,2)$ there are 105 two-frames and 168 three-frames. The plane geometry $\mathrm{PG}(2,2)$ is unusual in that the complement of a two-frame is a line. For if the four points of a two-frame are taken to be p, q, r, and $c = p + q + r$, then the remaining three points, namely

$$p' = q + r, \quad q' = p + r, \quad r' = p + q, \tag{2.4}$$

are linearly dependent (since $p' + q' + r' = 0$) and so form a line.

The intersection $\alpha \cap \beta$ of two (projective) subspaces of $\mathrm{PG}(m,2)$ is a subspace. Their join $j(\alpha,\beta)$ is defined to be the smallest (projective) subspace containing both α and β. In terms of vector subspaces of $V(m+1)$, we have $j(\alpha,\beta)\cup\{0\} = (\alpha\cup\{0\}) + (\beta\cup\{0\})$. Grassmann's relation for the (vector) subspaces of $V(m+1)$ yields a corresponding relation for the (projective) subspaces of $\mathrm{PG}(m,2)$

$$\dim(\alpha\cap\beta) = \dim\alpha + \dim\beta - \dim j(\alpha,\beta). \tag{2.5}$$

In particular, a line not contained in a hyperplane always intersects the hyperplane in a point, and two distinct hyperplanes always intersect in an $(m-2)$-space. If subspaces α, β are skew, that is, if $\alpha\cap\beta = \varnothing$, note also from (2.5) that

$$\dim j(\alpha,\beta) = \dim\alpha + \dim\beta + 1 \quad (\text{if } \alpha,\beta \text{ are skew}). \tag{2.6}$$

In particular, the join of a pair of skew lines in solid geometry is the whole space. In the sequel, we shall use over and over again the following further elementary consequence of (2.5).

Lemma 2.1: For $r = 0,1,2,\ldots,m$, every r-space of $\mathrm{PG}(m,2)$ intersects every $(m-r)$-space.

Proof: From (2.5), $\dim(\alpha\cap\beta) \geqslant 0$, since $\dim j(\alpha,\beta) \leqslant m$.

B. Duality

Consider now the vector space $\hat{V}(m+1)$ dual to $V(m+1)$. Let us denote by $\langle\alpha,p\rangle$ the value of $\alpha\in\hat{V}(m+1)$ at the vector $p\in V(m+1)$. For given nonzero $\alpha\in\hat{V}(m+1)$, the solutions of the equation $\langle\alpha,p\rangle = 0$ form a vector hyperplane in $V(m+1)$, and indeed every hyperplane of $V(m+1)$ can be represented in this way. Consequently each (projective) hyperplane of $\mathrm{PG}(m,2)$ can be represented in terms of the nonzero solutions of such an equation $\langle\alpha,p\rangle = 0$, moreover (special nature of $\mathbb{F}_2$) for uniquely determined $\alpha\in\hat{V}(m+1)\setminus\{0\}$. Consequently, this last set can, and will, be *identified with the set of hyperplanes in* $\mathrm{PG}(m,2)$;

$$S_{m-1} = S_{m-1}(m) = \hat{V}(m+1)\setminus\{0\}. \tag{2.7}$$

Since a linear functional can take only the values 0,1, note that we have, for $\alpha\in S_{m-1}$, $p\in S_0$,

$$\langle\alpha,p\rangle = \begin{cases} 0, & \text{if } p\in\alpha, \\ 1, & \text{if } p\notin\alpha. \end{cases} \tag{2.8}$$

Incidentally, note that S_{m-1} in (2.7), just like S_0 in (2.1), can be taken as a model for an m-dimensional geometry, say the dual geometry $\mathrm{PG}(m,2)\hat{}$. The "points" of the dual geometry are the hyperplanes of the original geometry, and similarly the "r-spaces" of the dual geometry are the $(m-r-1)$-spaces of the original geometry. This leads on to the well-known *principle of duality*: a valid theorem in one geometry automatically gives rise [after reversing inclusions, interchanging joins and intersections, and replacing r-spaces by $(m-r-1)$-spaces] to a valid "dual theorem" in the dual geometry.

In the sequel, we shall need to deal with the linear characters of the Abelian group $V(m+1)$, and of its dual $\hat{V}(m+1)$, in a multiplicative fashion. To this end, let us set

$$\chi_p^\alpha = \exp(i\pi\langle\alpha,p\rangle), \quad p\in V, \quad \alpha\in\hat{V}. \tag{2.9}$$

Then the 2^{m+1} linear characters of $V(m+1)$ are given by

$$\chi^\alpha:\ p\to\chi_p^\alpha,\quad \alpha\in\hat{V}(m+1), \tag{2.10}$$

and the 2^{m+1} linear characters of $\hat{V}(m+1)$ are given by

$$\chi_p:\ \alpha\to\chi_p^\alpha,\quad p\in V(m+1). \tag{2.11}$$

Notice that the values of χ: $\hat{V}\times V\to\{\pm 1\}$ upon $S_{m-1}\times S_0$ are given, from (2.8), by

$$\chi_p^\alpha=\begin{cases}+1, & \text{if } p\in\alpha,\\ -1, & \text{if } p\notin\alpha.\end{cases} \tag{2.12}$$

C. Subgroups of C_0

As promised in Sec. I we now view the group C_0 of figures of the set S [see (1.12) and (1.13)] in geometric terms, by identifying S with the set S_0 of points of PG$(m,2)$. For any collection $\mathcal{F}$ of figures of C_0, let $\langle\mathcal{F}\rangle$ denote the subgroup of C_0 generated by $\mathcal{F}$. Of interest are the subgroups $C_1,C_2,\ldots,C_{m-1}$ generated by, respectively, the lines, planes,...,hyperplanes of PG$(m,2)$,

$$C_r=C_r(m)=\langle S_r(m)\rangle,\quad r=0,1,2,\ldots,m, \tag{2.13}$$

where of course, by (1.10), $C_m=\{1\}$.

Before proceeding further, it is well to clear up any possible confusion that could arise between the additive "group of points" $V(m+1)=S_0\cup\{0\}$ and the multiplicative group C_0 *generated by the points*, and similarly between the additive "group of hyperplanes" $\hat{V}(m+1)=S_{m-1}\cup\{0\}$ and the multiplicative group C_{m-1} *generated by the hyperplanes*. Given two distinct points, $p,q\in S_0$, we can form their sum $p+q$ within $V(m+1)$ and their product pq within C_0. Of course $p+q$ is quite different from pq, since $p+q$ is another point, namely the third point on the line $j(p,q)$, while pq is the doubleton figure $\{p,q\}$. [In the case $p=q$ then $p+q$ is the zero vector of $V(m+1)$, while $pp=p^2$ is the identity element $1=\varnothing$ of C_0.] No doubt we have been laboring the obvious—yet a surprise is in store (see the next lemma) when we consider hyperplanes.

Lemma 2.2: If α, β are distinct hyperplanes, then $\alpha+\beta=\alpha\beta$.

Before proving this lemma for general m, it may help to illustrate it in the plane, taking the seven points of PG(2,2) to be $p,q,r,c(=p+q+r)$, p', q', r', as in (2.4). Setting $\alpha=cpp'=\{c,p,p'\}$ and $\beta=cqq'=\{c,q,q'\}$, then $\alpha+\beta$ will be another line, namely the third line $\gamma=crr'=\{c,r,r'\}$ through $c=\alpha\cap\beta$. But, by (1.9), (1.10), we have

$$\alpha\beta\gamma=cpp'qq'rr'=1, \tag{2.14}$$

whence

$$\alpha\beta=\gamma=\alpha+\beta.$$

Remark: This plane result was, in effect, made use of in Ref. 7; unfortunately, in this earlier paper, the group of points and the group of lines were written multiplicatively. (They were denoted, respectively, P_8 and L_8.)

Proof of Lemma 2.2: Set $\lambda=\alpha\cap\beta$. Now through the $(m-2)$-space λ pass just three hyperplanes, namely α, β, and $\gamma=\alpha+\beta$. [This dualizes the result that on the line $j(p,q)$ there are just three points, namely p,q and $p+q$.] So we can decompose S_0 as a disjoint union of four small sets λ, α',β',γ' such that $\alpha=\lambda\alpha',\beta=\lambda\beta',\gamma=\lambda\gamma'$. Hence, using (1.10), (1.14),

$$\alpha\beta\gamma=\lambda^3\alpha'\beta'\gamma'=\lambda\alpha'\beta'\gamma'=1, \tag{2.15}$$

whence $\alpha\beta=\gamma=\alpha+\beta$.

Remark: The fact that just three hyperplanes pass through $\alpha\cap\beta$ could have been seen in terms of the result (2.8). For if a point $p\in S_0$ belongs neither to α nor to β, then $\langle\alpha,p\rangle=1=\langle\beta,p\rangle$, whence $\langle\alpha+\beta,p\rangle=1+1=0$, and so all such points p belong to the hyperplane $\alpha+\beta$. [In fact, it is quite easy to compute the number of s-spaces within general "pencils" or "stars"—see (2.18) below.]

Corollary: The subgroup $C_{m-1}(m)$ of $C_0(m)$, which is generated by the hyperplanes, is isomorphic to the group $\hat{V}(m+1)$ (of order 2^{m+1}).

This follows from the lemma since, in the case $\alpha=\beta$ not considered in the lemma, $\alpha+\alpha$ and α^2 are the identity elements, 0 and $1=\varnothing$, of the respective groups $\hat{V}(m+1),C_{m-1}(m)$.

D. Stars

The results (2.14), (2.15) are ripe for generalization. To this end, if α is a given r-space lying inside a given t-space γ, let us define $\mathrm{star}_s(\alpha,\gamma)$ to be the family of s-spaces, $(r\leqslant s\leqslant t)$, which lie inside γ but which all contain α,

$$\mathrm{star}_s(\alpha,\gamma)=\{\beta\in S_s:\ \alpha\subseteq\beta\subseteq\gamma\}. \tag{2.16}$$

(In certain cases such a family is called a pencil rather than a star.) In the special case $t=m$, i.e., $\gamma=S_0$, we adopt the notation $\mathrm{star}_s(\alpha)$,

$$\mathrm{star}_s(\alpha)=\{\beta\in S_s:\ \alpha\subseteq\beta\}$$

and in the special case $r=-1$, i.e., $\alpha=\varnothing$, we adopt the notation $S_s(\gamma)$:

$$S_s(\gamma)=\{\beta\in S_s:\ \beta\subseteq\gamma\}=\mathrm{star}_s(\varnothing,\gamma).$$

Setting now

$$N(r,s,t)=|\mathrm{star}_s(\alpha,\gamma)|,(\dim\alpha=r,\quad\dim\gamma=t), \tag{2.17}$$

we have, by duality in PG$(t,2)$,

$$N(r,s,t)=N(t-s-1,\quad t-r-1), \tag{2.18}$$

where $N(\ ,\ )$ is as given in (2.2). Consider now the product, within C_0, of all $N(r,s,t)$ members β of $\mathrm{star}_s(\alpha,\gamma)$, say

$$\mathrm{st}_s(\alpha,\gamma)=\Pi_{\beta\in\mathrm{star}_s(\alpha,\gamma)}\beta. \tag{2.19}$$

In the case when $\gamma=S_0$ we denote the corresponding product of the members of $\mathrm{star}_s(\alpha)$ by $\mathrm{st}_s(\alpha)$.

Theorem 2.3:

$$\text{(i) } \mathrm{st}_s(\alpha,\gamma)=\gamma;\ \text{(ii) } \mathrm{st}_s(\alpha)=1. \tag{2.20}$$

In each case the result is independent of the choice of s and of α.

Proof: Consider the number of times a point $p\in S_0$ occurs in the product (2.19). If $p\in\alpha$ then p occurs in each β of the star, that is, $N(r,s,t)$ times. If $p\notin\alpha$, but $p\in\gamma$, then $j(p,\alpha)$ is a $(r+1)$-space, and so the number of times such a point p occurs in (2.19) is $N(r+1,s,t)$. Since $N(r,m)$ in (2.2) is odd, we see that each point $p\in\gamma$ occurs an odd number of times in the product (2.19). But clearly a point $p\notin\gamma$ does not occur. The theorem therefore follows from (1.9) [and from (1.10) in the case part of (ii) of the theorem].

Illustrations: We have the following particular cases of Theorem 2.3.

(a) If $\lambda_1,\lambda_2,\lambda_3$ are three distinct concurrent lines of a plane α, then

$$\lambda_1\lambda_2\lambda_3 = \alpha. \tag{2.21}$$

(b) If $\lambda_1,\lambda_2,...,\lambda_7$ are the seven distinct lines of a plane α, then

$$\lambda_1\lambda_2\cdots\lambda_7 = \alpha. \tag{2.22}$$

In the case of illustrations (c), (d), (e), let us assume that we deal with the case $m = 3$ of solid projective geometry over $\mathbb{F}_2$.

(c) If $\lambda_1, \lambda_2,...,\lambda_7$ are the seven distinct lines through a point p, then

$$\lambda_1\lambda_2\cdots\lambda_7 = 1. \tag{2.23}$$

(d) If $\alpha_1, \alpha_2, \alpha_3$ are three distinct planes intersecting in a common line, then

$$\alpha_1\alpha_2\alpha_3 = 1. \tag{2.24}$$

(e) If $\alpha_1, \alpha_2,...,\alpha_7$ are the seven distinct planes through a point p, then

$$\alpha_1\alpha_2\cdots\alpha_7 = 1. \tag{2.25}$$

Remark: Illustration (b) is of the case $r = -1$, $s = 1$, $t = 2$. Taking (2.21) and (2.22) together, note that the four lines $\lambda_4,...,\lambda_7$ of a "quadrilateral" (i.e., no three are concurrent) satisfy

$$\lambda_4\lambda_5\lambda_6\lambda_7 = 1. \tag{2.26}$$

Clearly the seven lines of any plane α can be partitioned into a triplet satisfying (2.21) and a quadruplet satisfying (2.26) in precisely seven ways (since α contains seven points).

There are similar partitionings of the seven planes through a point p arising from (2.24) and (2.25). [The situation is entirely analogous, for if we had been in the general case $m > 3$, and had considered the planes $\alpha_1,...,\alpha_7$ in (2.24), (2.25) to lie in a solid σ, then (2.24), (2.25) would hold only after 1 had been replaced by σ.] Again there are precisely seven such partitionings, corresponding to the seven choices of a "common line" through p.

Corollary to Theorem 2.3: For $s < t$, $C_s \supseteq C_t$.

Proof: By Theorem 2.3 every t-space γ can be expressed as a product of s-spaces.

III. PG(3,2) AND THE SUBGROUP CHAIN $C_0 \supset C_1 \supset C_2 \supset C_3 = \{1\}$

Since we intend to study the Clifford algebra for $d = 15$ in some detail, it will prove worthwhile to persist at the Abelian level a little longer in order to obtain some detailed $m = 3$ results. By the preceding corollary, the subgroups $C_r = C_r(3)$ of $C_0 = C_0(3)$ form a chain,

$$C_0 \supseteq C_1 \supseteq C_2 \supseteq C_3 = \{1\}, \tag{3.1}$$

where $|C_0| = 2^{14}$. Now by Lemma 2.2, Corollary, we already know that the group C_2 generated by the planes is isomorphic to $\hat{V}(4)$, and so has order $2^4 = 16$. Consequently, we are curious to find out about the intervening subgroup C_1 generated by the lines of PG(3,2).

To this end, it helps to choose a distinguished point $v \in S_0$ and a distinguished plane $\delta \in S_2$ such that v does not lie on δ. We will refer to v as the *vertex* and δ as the *base*. Now each $p \in \delta$ defines a line

$$\nu_p = j(p,v), \tag{3.2}$$

and, as p varies over δ, we obtain the star consisting of the seven lines through the vertex,

$$\mathrm{star}_1(v) = \{\nu_p:\ p \in \delta\}. \tag{3.3}$$

(Incidentally, we are frequently rather sloppy in not distinguishing between v and $\{v\}$; certainly we will always prefer $p \in \delta$ to $\{p\} \subset \delta$.) Since we will be interested in the coset group C_1/C_2, it is worth noting the result for $p \neq q \in \delta$,

$$\nu_p\nu_q = \nu_{p+q} \pmod{C_2}, \tag{3.4}$$

which follows from (2.21). In a similar vein, note also that $\lambda_1\lambda_2 = (\lambda_1 + \lambda_2)\delta$ holds for $\lambda_1 \neq \lambda_2 \in S_1(\delta)$, so that we have

$$\lambda_1\lambda_2 = (\lambda_1 + \lambda_2) \pmod{C_2}. \tag{3.5}$$

Here $\lambda_1 + \lambda_2$ is the third line in δ which is concurrent with the distinct lines λ_1, λ_2 of δ, the addition being that in the dual vector space $\hat{V}(3)$ associated with the PG(2,2) geometry of the plane δ. In the next lemma, it will prove convenient to allow p,q in (3.4) to range over the extended domain $\delta \cup \{0\}$, which we identify with $V(3)$, by defining $\nu_0 = 1$. Similarly, we allow λ_1, λ_2 in (3.5) to range over $\hat{V}(3)$ by identifying the identity elements, 0 and $\varnothing$, of the groups $\hat{V}(3)$ and $C_1(2)$ under the isomorphism of Lemma 2.2, Corollary.

Lemma 3.1: (i) The disjoint cosets of C_2 in C_1 are given by

$$\{\nu_p\lambda C_2;\quad p \in V(3),\quad \lambda \in \hat{V}(3)\},$$

(ii) $C_1/C_2 \simeq V(3) \times \hat{V}(3)$,

(iii) $|C_1| = 2^{10} = 1024$.

Proof: Suppose μ is a line such that μ is neither one of the seven lines of $\mathrm{star}_1(v)$ nor one of the seven lines of $S_1(\delta)$. (There are $35 - 7 - 7 = 21$ such lines.) Set $p = \mu \cap \delta$ and $\lambda = j(v,\mu) \cap \delta$ (= projection of μ from the vertex onto the base). Then μ, λ, and ν_p are coplanar lines concurrent in p. Hence by (2.21),

$$\mu = \nu_p\lambda \pmod{C_2}. \tag{3.6}$$

Now a general element $\psi \in C_1$ is, by definition of C_1, a product of lines. Hence, modulo C_2, it follows from (3.6) that ψ is a product of lines drawn from $\mathrm{star}_1(v)$ and from $S_1(\delta)$. After using (3.4) and (3.5), we see that ψ, modulo C_2, can be expressed as a product of at most one line from $\mathrm{star}_1(v)$ and at most one line from $S_1(\delta)$. Hence, under the extended interpretation of p and λ indicated after (3.5), we have

$$\psi = \nu_p\lambda \pmod{C_2}, \quad \text{for some } p \in V(3),\quad \lambda \in \hat{V}. \tag{3.7}$$

[Incidentally, in the case $p \in \delta$, $\lambda \subset \delta$, the lines ν_p, λ in (3.7) may now be skew, while in the case of the line μ in (3.6) the lines ν_p, λ intersected at $p = \mu \cap \delta$.]

In order to conclude the proof of part (i) we now have only to check that the $8 \times 8 = 64$ cosets exhibited are distinct. Suppose then that the coset $\nu_p\lambda_1 C_2$ coincided with the coset $\nu_q\lambda_2 C_2$. Setting $p + q = r$ and $\lambda_1 + \lambda_2 = \lambda$, it would follow that $\nu_r = \lambda\alpha$ for some $\alpha \in C_2$. Clearly this entails $\nu_r = 1, \lambda = 1, \alpha = 1$, and so $\nu_p = \nu_q$ and $\lambda_1 = \lambda_2$ as desired.

Upon appealing to (3.4), (3.5) again, we have

$$(\nu_p\lambda_1 C_2)(\nu_q\lambda_2 C_2) = \nu_{p+q}(\lambda_1 + \lambda_2)C_2, \tag{3.8}$$

whence part (ii) follows. Finally, since $|V(3) \times \hat{V}(3)| = 2^3 \times 2^3 = 2^6$, and since $|C_2| = 2^4$, we have $|C_1| = 2^{10}$.

Lemma 3.2: (i) The disjoint cosets of C_1 in C_0 are given by

$$\{pC_1;\quad p\in V(4)\};$$

(ii) $C_0/C_1\simeq V(4)$.

Proof: For $p\neq q\in S_0(3)$ we have $pq(p+q)=\lambda$ where $\lambda=j(p,q)$. Thus we have

$$pq=p+q(\mathrm{mod}\ C_1). \tag{3.9}$$

It follows that C_0 is certainly the union of the listed cosets,

$$C_0=\cup_{p\in V(4)}pC_1. \tag{3.10}$$

But we know $|C_0|=2^{14}$, $|C_1|=2^{10}$, and $|V(4)|=2^4$, whence the cosets in (3.10) must be disjoint.

Part (ii) now follows from (3.9).

The coset decomposition of Lemma 3.1, namely

$$C_1=\cup_{p\in V(3),\ \lambda\in\hat{V}(3)}\nu_p\lambda C_2, \tag{3.11}$$

suggests that we introduce two subgroups which stand midway between C_1 and C_2 in the chain (3.1), namely

$$C_{2,\delta}=\langle S_2\cup\{\lambda:\ \lambda\in S_1(\delta)\}\rangle \tag{3.12}$$

and

$$C_{2,v}=\langle S_2\cup\{\nu:\ \nu\in \mathrm{star}_1(v)\}\rangle. \tag{3.13}$$

By (3.4) and (3.5), once more we have the coset decompositions

$$C_{2,\delta}=\cup_{\lambda\in\hat{V}(3)}\lambda C_2, \tag{3.14}$$

$$C_{2,v}=\cup_{p\in V(3)}\nu_p C_2, \tag{3.15}$$

and from (3.11) we have

$$C_1=\cup_{p\in V(3)}\nu_p C_{2,\delta}, \tag{3.16}$$

$$C_1=\cup_{\lambda\in\hat{V}(3)}\lambda C_{2,v}. \tag{3.17}$$

Remark: Clearly, in contrast to the subgroups C_r, $r=0$, 1, 2, 3, the subgroups $C_{2,\delta}$ and $C_{2,v}$ are not Ω-admissible subgroups, but only Ω_δ- or Ω_v-admissible, respectively, where Ω_δ arises from those projectivities of PG(3,2) which preserve the distinguished plane δ, and similarly Ω_v arises from those projectivities which preserve the distinguished point v.

When, in the next section, we climb up to the non-Abelian level, we will chiefly make use of the coset decompositions already arrived at in this section, rather than a knowledge of precisely which figures occur in the group C_1 generated by the lines of PG(3,2). Nevertheless, one is curious to discover these $|C_1|=1024$ figures which are picked from amongst the 16 384 figures of C_0. Of course C_1 contains $1\equiv\varnothing$ and the 35 lines of PG(3,2). Multiplying two distinct lines λ_1,λ_2 together, there are two possibilities: if λ_1,λ_2 intersect, then $\lambda_1\lambda_2$ is a two-frame, which accounts for 105 figures (as explained in Sec. II A), while if λ_1, λ_2 are skew, then $\lambda_1\lambda_2$ is a "skew pair," of which there are 280.

Several possibilities arise when we consider the produce $\lambda_1\lambda_2\lambda_3$ of three lines. If coplanar and concurrent, then as in (2.21), $\lambda_1\lambda_2\lambda_3$ is a plane, which accounts for a further 15 figures. If coplanar but not concurrent, then $\lambda_1\lambda_2\lambda_3$ is a single line [cf. (2.26)], which has already been counted. Another possibility occurs when $\lambda_1\lambda_2$ is a two-frame which intersects λ_3 in a single point, in which case we see that $\lambda_1\lambda_2\lambda_3$ is a three-frame, which accounts for a further 168 figures. Another possibility is that $\lambda_1,\lambda_2,\lambda_3$ are concurrent in some point p, say, but are not coplanar. Let us term the resulting figure $\lambda_1\lambda_2\lambda_3$ a "tripod." It consists of seven points, one of which is the privileged "apex" p, and the remaining six points lie in pairs along the three "legs" λ_1, λ_2, λ_3, of the tripod. Now we can choose the apex p in 15 ways, and then for given p choose the three legs in $7.6.4/3!=28$ ways. Consequently, the group C_1 contains $15\times 28=420$ tripods.

There is no need to look at further possibilities, since elementary arithmetic assures us that we have already discovered the 1024 (see Lemma 3.1) figures of C_1.

Lemma 3.3: The group $C_1=C_1(3)$ consists of the following 1024 figures: $\varnothing$; 15 planes, 35 lines; 105 two-frames; 168 three-frames; 280 skew pairs; 420 tripods.

Corollary: If lines λ_1, λ_2, λ_3 of PG(3,2) are mutually skew, then there exist lines σ, ν such that the five lines λ_1, λ_2, λ_3, σ, ν are mutually skew [and consequently exhaust all the 15 points of PG(3,2)].

Proof: The figure $\lambda_1\lambda_2\lambda_3$ clearly belongs to C_1 and consists of $15-9=6$ points. But in the list in Lemma 3.3, the only figure of size six is a skew pair.

Remark: The corollary is a well-known result of PG(3,2) geometry. We do *not* claim that our proof is superior to ones based more immediately (see, for example, Ref. 11, p. 55) upon simple incidence properties of PG(3,2)!

Remark: The lines σ, ν are uniquely determined by the three skew lines $\lambda_1,\lambda_2,\lambda_3$. The latter also uniquely determine three other skew lines μ_1,μ_2,μ_3, namely the three transversals of λ_1, λ_2, λ_3, which are thus also skew to σ and ν. One ends up with the result that given any two skew lines σ, ν, their product can be expressed as the product of three skew lines in precisely two ways,

$$\sigma\nu=\lambda_1\lambda_2\lambda_3=\mu_1\mu_2\mu_3. \tag{3.18}$$

Remark: Another corollary of Lemma 3.3, in the same vein as the above one, is that a "pierced plane" $\lambda\alpha$, consisting of a line λ not lying inside a plane α—which is thus a figure (clearly not a plane) of size $15-8=7$—*must* be a tripod. This conclusion is easily checked directly, by expressing α in the form $\lambda_1\lambda_2\lambda_3$ and using (2.23) to reduce $\lambda\lambda_1\lambda_2\lambda_3$ to the product of three lines of a tripod.

IV. PG(m,2) AND THE GROUP G_0

A. Anticommutativity and incidence properties

At long last, we return to the non-Abelian group G_0 consisting of the 2^d linear operators displayed in (1.3). These operators fall into pairs, such as $\{\Gamma_2\Gamma_3\Gamma_5,-\Gamma_2\Gamma_3\Gamma_5\}$, forming a coset of $Z_0=\{I,-I\}$ in G_0 and thus yielding a figure of C_0, such as $\{s_2,s_3,s_5\}=s_2s_3s_5$. At times, for the sake of definiteness, we will for each figure $\alpha\in C_0$ make some choice $\Gamma(\alpha)$, say, of coset representative. However, we will always insist that our choice satisfies $\Gamma(\varnothing)=+I$ and $\Gamma(s_i)\ [=\Gamma(\{s_i\})]=\Gamma_i$. When we identify, as in Sec. II C, S in (1.12) with the points S_0 of PG(m,2), we also write $\Gamma(p)=\Gamma_p$, for $p\in S_0$. Such a choice of $\Gamma(\alpha)\in G_0$ for each $\alpha\in C_0$ corresponds, at least for $|\alpha|>1$, to making a choice of "orientation" for each figure α. Modulo even permutations, there are two orderings of the points of α (for $|\alpha|>|$), and a choice of one of these will be termed a

choice of orientation. To an oriented figure $\alpha \in C_0$ will correspond a unique element $\Gamma(\alpha) \in G_0$ in the obvious manner. Thus if $\alpha = \{p,q,r\}$ is given the orientation provided by the orderings *pqr*, *qrp*, *rpq*, then we set

$$\Gamma(\alpha) = +\Gamma_p\Gamma_q\Gamma_r = +\Gamma_q\Gamma_r\Gamma_p = +\Gamma_r\Gamma_p\Gamma_q.$$

(Caution: in the case of lines, a different sign convention was adopted in Ref. 7.)

Of course for some $\alpha, \beta \in C_0$ the orientation assigned to $\alpha\beta$ will be such that $\Gamma(\alpha\beta) = -\Gamma(\alpha)\Gamma(\beta)$ rather than $+\Gamma(\alpha)\Gamma(\beta)$. Consequently, from the point of view of the Abelian group C_0, the assignment $\alpha \to \Gamma(\alpha)$ defines a multiplier representation of C_0,

$$\Gamma(\alpha)\Gamma(\beta) = c(\alpha,\beta)\Gamma(\alpha\beta), \quad \alpha,\beta \in C_0, \tag{4.1}$$

where the multiplier c takes values in $\{1, -1\}$.

Lemma 4.1: If α is an r-space, with $r \geqslant 1$, then $\Gamma(\alpha)^2 = +I$.

Proof: For any figure α, we easily check that

$$\Gamma(\alpha)^2 = (-1)^{1/2|\alpha|(|\alpha|+1)}I.$$

Hence the lemma, since, for an r-space, $|\alpha| = 2^{r+1} - 1$.

Two distinct subsets α,β of S_0 define a partition of S_0 into four subsets,

$$S_0 = (\alpha \cap \beta) \cup (\alpha \cap \beta^c) \cup (\alpha^c \cap \beta) \cup (\alpha^c \cap \beta^c). \tag{4.2}$$

Whether an individual subset has an even or an odd number of elements is of no great relevance, since in the group C_0 we identify a set with its complement. (For, since $|S_0| = d$ is odd, $|\alpha|$ is odd whenever $|\alpha^c|$ is even.) Nevertheless, given the *pair* of subsets α,β, it makes sense to decide whether one bears an even or an odd relation to the other in accordance with the following definition: if three of the four subsets in (4.2) are even, then we say that α bears an even relation to β, while if three of the subsets in (4.2) are odd, we say that α bears an odd relation to β. (Since $|S_0|$ is odd, no other possibilities arise.) Let us define

$$\epsilon(\alpha,\beta) = \begin{cases} +1, & \text{if } \alpha \text{ bears an even relation to } \beta, \\ -1, & \text{if } \alpha \text{ bears an odd relation to } \beta. \end{cases} \tag{4.3}$$

Equivalently we have

$$\epsilon(\alpha,\beta) = (-1)^{|\alpha\cap\beta| + |\alpha||\beta|}. \tag{4.3'}$$

Now note that if α is an r-space then $|\alpha|$ is odd, *except in the case* $r = -1$ *of the empty set* $\emptyset$. Consequently, if α is an r-space and β is an s-space, and if neither of α or β equals $\emptyset$, note that

$$\epsilon(\alpha,\beta) = \begin{cases} +1, & \text{if } \alpha \text{ intersects } \beta, \\ -1 & \text{if } \alpha \text{ is skew to } \beta. \end{cases} \tag{4.4}$$

Observe also that the special case of (4.4) when α is a hyperplane and β is a point p ties in with the characters of the groups $V(m+1)$, $\hat{V}(m+1)$ considered in Sec. II B,

$$\epsilon(\alpha,p) = \chi_p^\alpha. \tag{4.5}$$

Lemma 4.2: For any figures $\alpha, \beta \in C_0$ we have

$$\Gamma(\alpha)\Gamma(\beta) = \epsilon(\alpha,\beta)\Gamma(\beta)\Gamma(\alpha). \tag{4.6}$$

In particular, if α is an r-space and β is an s-space, and if neither α nor β is $\emptyset$, then $\Gamma(\alpha)$ commutes with $\Gamma(\beta)$, except in the case when α,β are skew, in which case $\Gamma(\alpha)$ anticommutes with $\Gamma(\beta)$.

Proof: The lemma is an elementary consequence of (1.1) taken in conjunction with (4.3) and (4.4).

Remark: We already begin to see that the switch from the commuting points of C_0 to the anticommuting "Γ-points" of G_0 will only strengthen the usefulness of PG$(m,2)$ methods. Incidentally note that it follows from Lemma 4.2 that $\epsilon(\alpha,\beta)$ has the multiplicative property

$$\epsilon(\alpha,\beta\gamma) = \epsilon(\alpha,\beta)\epsilon(\alpha,\gamma).$$

B. Subgroup chains

From the Corollary to Theorem 2.3 we have the subgroup chain of Abelian groups,

$$C_0 \supset C_1 \supset \cdots \supset C_{m-1} \supset C_m = \{1\}. \tag{4.7}$$

Taking inverse images under the projection π: $G_0 \to G_0/Z_0 = C_0$ we also have the subgroup chain

$$G_0 \supset G_1 \supset \cdots \supset G_{m-1} \supset G_m = Z_0 = \{\pm I\}, \tag{4.8}$$

where

$$G_r = \pi^{-1}(C_r) = \langle \pm\Gamma(\alpha); \alpha \in S_r(m) \rangle. \tag{4.9}$$

Since the chain (4.7) is a normal series for C_0, it follows (and is otherwise obvious) that the chain (4.8) is a normal series for G_0, and not merely a subnormal series, in that each subgroup G_r is a normal (invariant) subgroup of G_0.

For each $r = 0,1,...,m$, let Z_r denote the centralizer in G_0 of the subgroup G_r,

$$Z_r = Z_{G_0}(G_r).$$

Theorem 4.3: For each $r = 0,1,2,...,m$,

$$G_{m-r} \subseteq Z_r. \tag{4.10}$$

Proof: By Lemmas 2.1 and 4.2, each element of G_{m-r} commutes with every element of G_r.

Remark: Working at the Abelian level in Sec. III, it emerged only belatedly that, at least in the $m = 3$ case, the group C_1 was a proper subgroup of C_0, or equivalently by multiplying lines together one never obtains a point. By the switch to anticommuting Γ-points, this result is now (indeed, for any m) immediately obvious. For if α is a hyperplane, then $\Gamma(\alpha)$ commutes with every "Γ-line" $\Gamma(\lambda)$, $\lambda \in S_1$, and hence with every element of G_1, yet anticommutes with Γ_p for $p \notin \alpha$.

Since $C_r = G_r/Z_0$ we have the isomorphisms

$$G_r/G_s \simeq C_r/C_s, \quad r \leqslant s. \tag{4.11}$$

Consequently we can make good use of any results obtained at the Abelian level to gain knowledge of the chain (4.8). In the next section, we will do this in some detail for the case $m = 3$. Before so doing, it is worth pointing out that Theorem 4.3 strongly suggests that *we should expect a certain mod 2 periodicity involving the dimension m of the projective geometry*. For if $m = 2l$ is even, then, by Theorem 4.3, the chain

$$G_0 \supset G_1 \supset \cdots \supset G_l \supset \cdots \supset G_{2l-1} \supset G_{2l} \tag{4.12}$$

becomes Abelian at the middle term G_l. On the other hand, if $m = 2l + 1$ is odd, then the chain

$$G_0 \supset G_1 \supset \cdots \supset G_l \supset G_{l+1} \supset \cdots \supset G_{2l} \supset G_{2l+1} \quad (4.13)$$

has no middle term, but becomes Abelian at G_{l+1}.

Generalizing from the particular odd case $m = 3$ studied in Sec. III, let us in the general odd case $m = 2l + 1$ distinguish some "vertex" $v \in S_0$ and some "base" hyperplane $\delta \supset S_{2l}$. Then we can define two subgroups of C_0,

$$C_{l+1,\delta} = \langle S_{l+1} \cup \{\beta\colon\ \beta \in S_l(\delta)\}\rangle, \quad (4.14)$$

$$C_{l+1,v} = \langle S_{l+1} \cup \{\gamma\colon\ \gamma \in \mathrm{star}_l(v)\}\rangle, \quad (4.15)$$

which lie partway between C_l and C_{l+1} in the chain (4.7). Consequently we have two normal subgroups of G_0

$$\begin{aligned} G_{l+1,\delta} &= \pi^{-1}(C_{l+1,\delta}), \\ G_{l+1,v} &= \pi^{-1}(C_{l+1,v}), \end{aligned} \quad (4.16)$$

which lie partway between G_l and G_{l+1} in the chain (4.13). Now l-spaces lying inside the distinguished $2l$-space δ intersect and hence their Γ-analogs commute amongst themselves. Similarly for the l-spaces passing through the common vertex v, and of course any l-space intersects every $(l + 1)$-space. Consequently, by Lemma 4.2, we have arrived at the following Lemma.

Lemma 4.4: Let H denote either of the subgroups of G_0 in (4.16). Then H is an Abelian normal subgroup such that

$$G_l \supset H \supset G_{l+1}. \quad (4.17)$$

Remark: In the case of those subgroups in the chains (4.12), (4.13) which are Abelian, we can obviously express them in the form, for some subgroup $K_r \subset G_r$,

$$G_r = K_r \times Z_0, \quad (4.18)$$

where r is $\geqslant \frac{1}{2}m$ (m even) or r is $\geqslant \frac{1}{2}(m + 1)$ (m odd). However—cf. the corresponding discussion in Ref. 7—the choice of subgroup $K_r \subset G_r$ will be far from unique, and no choice will be a normal subgroup of G_0. Similarly we can find, in a not unique way, a subgroup $K_{l+1,\delta}$ of $G_{l+1,\delta}$ and a subgroup $K_{l+1,v}$ of $G_{l+1,v}$ such that

$$G_{l+1,\delta} = K_{l+1,\delta} \times Z_0, \quad G_{l+1,v} = K_{l+1,v} \times Z_0. \quad (4.19)$$

V. PG(3,2) AND CLIFF (0,15)

A. Dual pairs

For $m = 3$ we deal with the chain

$$G_0 \supset G_1 \supset G_2 \supset G_3 = Z_0, \quad (5.1)$$

and also with

$$G_1 \supset H \supset G_2 \quad (5.2)$$

for the two choices (4.16) of H [arising from (3.12), (3.13)],

$$\text{(i)}\ H = G_{2,\delta}, \quad \text{(ii)}\ H = G_{2,v}. \quad (5.3)$$

From our results in Secs. III and IV we already know a great deal about these chains. From Theorem 4.3 we know that

$$G_1 \subseteq Z_2, \quad G_2 \subseteq Z_1 \quad (5.4)$$

and from Lemma 4.4 that $G_{2,\delta}$ and $G_{2,v}$ are both Abelian normal subgroups of G_0. From Sec. III, taken in conjunction with (4.11), we have the detailed coset decompositions

$$G_0 = \cup_{p \in V(4)} \Gamma_p G_1, \quad (5.5)$$

$$G_1 = \cup_{p \in V(3),\ \lambda \in \hat{V}(3)} \Gamma(\nu_p)\Gamma(\lambda) G_2, \quad (5.6)$$

along with ones corresponding to (3.14)–(3.17), and we also have the isomorphisms of factor groups

$$\begin{aligned} G_0/G_1 &\simeq V(4), \\ G_1/G_2 &\simeq V(3) \times \hat{V}(3), \\ G_2/G_3 &\simeq \hat{V}(4). \end{aligned} \quad (5.7)$$

The orders of G_0, G_1, G_2 are thus 2^{15}, 2^{11}, 2^5, respectively, and the orders of both $G_{2,\delta}$ and $G_{2,v}$ are 2^8.

In this section, we sharpen these results by proving the following two theorems. The second theorem can be summarized by saying that the subgroups G_{3-r}, G_r form a *dual pair*—cf. Howe.[12]

Theorem 5.1: Both $G_{2,\delta}$ and $G_{2,v}$ are maximal Abelian normal subgroups of G_0.

Theorem 5.2: For $r = 0,1,2,3$, G_{3-r} is the full centralizer in G_0 of G_r. In particular,

$$G_1 = Z_2, \quad G_2 = Z_1. \quad (5.8)$$

Proofs: We will start with Theorem 5.2—in the course of its proof we will obtain Theorem 5.1. We clearly have $G_3 = Z_0$ and $G_0 = Z_3$, so let us look at Z_2 (= the centralizer in G_0 of G_2). Now an element g of the coset $\Gamma_p G_1$, $p \in S_0$, can not belong to Z_2, since such a g anticommutes with $\Gamma(\alpha) \in G_2$ for any choice of plane α not passing through p. It follows that $Z_2 \subseteq G_1$, and hence from (5.4) we have $Z_2 = G_1$ as desired.

Next we look at the centralizer $Z_{2,\delta}$ of $G_{2,\delta}$ in G_0. Since $G_{2,\delta} \supset G_2$ we must have $Z_{2,\delta} \subseteq Z_2 = G_1$. Now consider an element g_1 of G_1 which lies in the coset $\Gamma(\nu_p)G_{2,\delta}$, $\nu_p \in \mathrm{star}_1(v)$, of $G_{2,\delta}$ in G_1. Such an elementary g_1 can not belong to $Z_{2,\delta}$, since it anticommutes with $\Gamma(\lambda) \in G_{2,\delta}$ for any choice of the line $\lambda \in S_1(\delta)$ not passing through p. It follows that $Z_{2,\delta} \subseteq G_{2,\delta}$, and hence, since $G_{2,\delta}$ is Abelian, we have $Z_{2,\delta} = G_{2,\delta}$, i.e., we have proved Theorem 5.1 for $G_{2,\delta}$. An entirely similar argument proves that $G_{2,v}$ also is its own centralizer in G_0.

Finally we look at the centralizer Z_1 of G_1 in G_0. Since $G_1 \supset G_{2,\delta}$ we must have $Z_1 \subseteq Z_{2,\delta} = G_{2,\delta}$. Now consider an element h of $G_{2,\delta}$ which lies in the coset $\Gamma(\lambda)G_2$, $\lambda \in S_1(\delta)$, of G_2 in $G_{2,\delta}$. Such an element h can not belong to G_1, since it anticommutes with $\Gamma(\nu) \in G_1$ for any choice of line ν skew to λ. It follows that $Z_1 \subseteq G_2$ and hence from (5.4) we have $Z_1 = G_2$ as desired. So Theorems 5.1 and 5.2 have both been proved.

Remark: The proof in the last paragraph, that $Z_1 \subseteq G_2$, could be replaced by a slightly shorter one, as follows. Since G_1 contains both $G_{2,\delta}$ and $G_{2,v}$, it follows that

$$Z_1 \subseteq Z_{2,\delta} \cap Z_{2,v} = G_{2,\delta} \cap G_{2,v} = G_2, \quad (5.9)$$

where the last equality is clear from Lemma 3.1 (i).

B. Orientation conventions

In a moment, when we come to look at the action of the operators Γ_p upon a specific spinor basis, we will need to assume that the projective geometry PG(3,2) has been "totally oriented." By this we mean that every line and every plane has been given an orientation in the sense described in Sec. IV A [and that the whole space is oriented by condition

(1.2)]. As far as the planes are concerned, this can be done in such a way that the set

$$K_2 = \{I, \Gamma(\alpha);\ \alpha \in S_2\} \tag{5.10}$$

forms a group, isomorphic to $\hat{V}(4)$, with

$$\Gamma(\alpha)\Gamma(\beta) = +\Gamma(\alpha\beta) = +\Gamma(\alpha+\beta), \tag{5.11}$$

for distinct planes α, β, and indeed for general $\alpha, \beta \in K_2$ (for recall that $\Gamma(\emptyset) = +I$, and see Lemma 2.2 and its corollary). One way to achieve (5.10) is to make arbitrary choices of $\Gamma(\alpha_i)$ for the four faces α_i, $i = 1,2,3,4$, of the tetrahedron of reference (with, say, $\alpha_4 = \delta$ when we come to choose a "base" plane), and then to define $\Gamma(\alpha)$ for the remaining 11 planes by

$$\begin{aligned}
&\Gamma(\alpha_i + \alpha_j) = \Gamma(\alpha_i)\Gamma(\alpha_j), \\
&\Gamma(\alpha_i + \alpha_j + \alpha_k) = \Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_k), \\
&\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) = \Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4).
\end{aligned} \tag{5.12}$$

Thus the isomorphism $K_2 \simeq (\mathbb{Z}_2)^4$ is realized in terms of the choice of the $\Gamma(\alpha_i)$ as the generators of the four $\mathbb{Z}_2$-subgroups. We then have

$$G_2 = K_2 \times Z_0. \tag{5.13}$$

However, as discussed previously in connection with (4.18), the subgroup K_2 of G_2 is not a normal subgroup of G_0.

We would like to choose our total orientation in such a way that as many planes as possible are "well-oriented." By a WOP (well-oriented plane) we mean a plane γ such that

$$\Gamma\mathrm{st}_1(p,\gamma) = +\Gamma(\gamma) \tag{5.14+}$$

holds at each point $p \in \gamma$, rather than

$$\Gamma\mathrm{st}_1(p,\gamma) = -\Gamma(\gamma). \tag{5.14-}$$

Here we define $\Gamma\mathrm{st}_s(\alpha,\gamma)$, *assuming* $\alpha \neq \emptyset$, to be the obvious Γ-level version of $\mathrm{st}_s(\alpha,\gamma)$—that is $\Gamma\mathrm{st}_s(\alpha,\gamma)$ denotes the product of all the $\Gamma(\beta)$ as β runs through $\mathrm{star}_s(\alpha,\gamma)$. (The order is not important since, for $\alpha \neq \emptyset$, the $\Gamma(\beta)$ commute amongst themselves.) Of course, by Theorem 2.3, at each point $p \in \gamma$ either (5.14+) holds or (5.14−) holds. Note that if γ is a WOP, then for any distinct $\lambda, \mu \in S_1(\gamma)$ we will have

$$\Gamma(\lambda)\Gamma(\mu) = +\Gamma(\lambda + \mu)\Gamma(\gamma), \tag{5.15}$$

where $\lambda + \mu$ is the line of γ which is concurrent with λ, μ, the addition being that in the dual vector space $\hat{V}(3)$ associated with the PG(2,2) geometry of the plane γ.

Given a choice of vertex v and base δ we can show (see the Appendix) that, consistent with (5.11), we can totally orient PG(3,2) in such a way that δ as well as all seven planes of $\mathrm{star}_2(v)$ are WOP's. (As discussed in the Appendix, the remaining seven planes are then not well-oriented, but can be "badly-oriented in a particular way," making them into what we will call BOP's.)

Once each line has been oriented, then each ordered pair of distinct points $p, q \in S_0$ defines a sign $\eta(p,q) = -\eta(q,p) = \pm 1$ such that, for $\mu = j(p,q)$,

$$\Gamma_p\Gamma_q = \eta(p,q)\Gamma_{p+q}\Gamma(\mu), \quad p \neq q \in S_0. \tag{5.16}$$

We can extend (5.16) to cover the case $p = q \in S_0$ by setting $\eta(p,p) = -1$, $p \in S_0$, and interpreting μ as $\emptyset$. [We could further extend (5.16) to all $p, q \in V(4)$ by setting $\eta(p,0) = \eta(0,p) = \eta(0,0) = +1$, and again interpreting μ as $\emptyset$ in these cases.]

C. Clifford theory

At this stage, we could for the main part proceed directly to Sec. V D and construct explicit spinor bases. However, our results there will fall into a clearer pattern if we first of all learn what we can from the well-known theory of Clifford[13] concerning representations subduced in a normal subgroup. (Of course we refer here to A. H. Clifford, not to W. K. Clifford of Clifford algebras!) Our notation and terminology will be close to that used in Ref. 14. Let U denote the defining representations of the group G_0: $U(g) = g$. The carrier space E of U is a 128-dimensional Euclidean space of spinors.

Remark: The fact that G_0 possesses just one faithful irreducible representation U, and that the dimension is 128, can of course be easily deduced from the properties of G_0 itself, rather than, as in Sec. I, by appeal to general results on Clifford algebras. For the conjugacy classes of G_0 are in fact the cosets $\{\Gamma(\psi), -\Gamma(\psi)\}$ of Z_0, except for Z_0 itself which consists of two classes. So G_0 possesses $2 + \frac{1}{2}(|G_0| - 2) = 2^{14} + 1$ classes, and hence $2^{14} + 1$ inequivalent irreducible representations. The Abelian group $C_0 = G_0/Z_0$ of order 2^{14} accounts for 2^{14} representations, in which Z_0 is not faithfully represented, and so there exists just one irreducible representation U of G_0 which represents Z_0 faithfully. The dimension n of this representation is given by $2^{14}1^2 + n^2 = |G_0| = 2^{15}$, that is, $n = 2^7 = 128$. On various grounds (or even from our explicit results below in Sec. V D) the representation U can be taken to be real. By group averaging, it can be taken to be an orthogonal representation, whence the Γ_p, satisfying $(\Gamma_p)^2 = -I$, will be skew-adjoint, while the Γ-lines and Γ-planes will be self-adjoint operators.

We have a choice of normal subgroup of G_0. Let us first of all apply Clifford's results, using G_2 as the normal subgroup. Of the $32 = 2^5$ irreducible representations of $G_2 \simeq (\mathbb{Z}_2)^5$, the only ones involved in U are those which represent Z_0 faithfully. These are the 16 one-dimensional representations D_p, $p \in V(4) = \{0\} \cup S_0$, given by [see (2.7)–(2.12)]

$$D_p(\zeta\Gamma(\alpha)) = \zeta\chi_p^\alpha, \quad \zeta = \pm 1, \tag{5.17}$$

for $\alpha \in \hat{V}(4) = \{0\} \cup S_2$. [In terms of $G_2 = K_2 \times Z_0 \simeq \hat{V}(4) \times Z_0$ we have $D_p \simeq \chi_p \times \epsilon$, where $\epsilon(\pm I) = \pm 1$.] Under the usual action of $g \in G_0$ upon $D \in \hat{G}_2$, namely $D \to gD$ where $(gD)(n) = D(g^{-1}ng)$, $n \in G_2$, we have from (4.5), (4.6), $\Gamma_p D_q = D_{p+q}$, and so we see that the 16 representations (5.17) form a single G_0-orbit. The decomposition of $U \downarrow G_2$ into its disjoint primary constituents will be of the form

$$E = \oplus_{p \in V(4)} W_p, \tag{5.18}$$

where W_p carries the primary representations $D_p \oplus \cdots \oplus D_p$ (ν summands) of G_2. The common multiplicity ν with which the D_p occur [$= \dim W_p$ for each $p \in V(4)$] is thus

$$\nu = \dim E / |V(4)| = 128/16 = 8.$$

From $\Gamma_p D_q = D_{p+q}$ we see that the Γ_p permute the eight-dimensional subspaces W_p amongst themselves,

$$\Gamma_p(W_q) = W_{p+q}. \tag{5.19}$$

Note in particular that, if we choose, say, W_0 as "base-point," the remaining 15 W_p are given by

$$W_p = \Gamma_p(W_0), \quad p \in S_0. \tag{5.20}$$

The isotropy group of D_p ($= \{g \in G_0 : gD_p = D_p\}$) is seen from (5.17) to be the centralizer Z_2 of G_2 in G_0. That is, by Theorem 5.2, the isotropy group of each D_p is G_1. (Observe therefore that we are in a special case of Clifford's results in which the isotropy subgroup is itself a normal subgroup of G_0.) Setting T to be the eight-dimensional representation of G_1 carried by W_0,

$$T(h) = U(h)|_{W_0}, \quad h \in G_1,$$

then U is obtained, up to equivalence, as an induced representation,

$$U \simeq T(G_1) \uparrow G_0. \tag{5.21}$$

If instead we had chosen to apply Clifford's results for the choice $G_{2,\delta}$ of normal subgroup, we would have arrived at U as an induced representation from a one-dimensional representation of $G_{2,\delta}$. A similar remark applies to the choice $G_{2,v}$ of normal subgroup. We could also have chosen the larger group G_1. The various results tie in with each other via our previous coset decompositions and the well-known theorem (see e.g., Ref. 14, p. 536) concerning "inducing in stages."

If we are so minded, it is now a small further step to introduce appropriate bases for the 128-dimensional space E. Let us start with the eight-dimensional space W_0 which carries the representation $8D_0$ of G_2, and so satisfies

$$\Gamma(\alpha) w_0 = w_0, \quad \text{for all } w_0 \in W_0, \quad \alpha \in S_2. \tag{5.22}$$

The irreducible eight-dimensional representation T of G_1 carried by W_0 subduces eight one-dimensional representations of the subgroup $G_{2,\delta}$, and similarly also for the subgroup $G_{2,v}$, giving rise to two different decompositions of W_0, say

$$W_0 = \oplus_{d \in V(3)} W_{0,d}, \quad \text{and} \quad W_0 = \oplus_{\lambda \in \hat{V}(3)} W_0^\lambda, \tag{5.23}$$

into a sum of eight lines. Here $W_{0,d} = \Gamma(\nu_d) W_{0,0}$ carries χ_d, and $W_0^\lambda = \Gamma(\lambda) W_{0,0}$ carries χ^λ, where for $d \in \delta$ and $\lambda \in S_1(\delta)$,

$$\chi_d(\lambda) = \epsilon(d,\lambda) = \chi^\lambda(\nu_d). \tag{5.24}$$

(Of course we are now in a Clifford situation for the group G_1, rather than G_0, and we are considering two choices of normal subgroup of G_1.)

D. Spinor bases

Let us recapitulate the last paragraph in perhaps more familiar language: since each Γ-line commutes with all Γ-planes, the subspace W_0 which is pointwise fixed by the Γ-planes will be invariant under every Γ-line. However the Γ-lines do not commute amongst themselves, and we can either choose to simultaneously diagonalize the commuting subset $\Gamma S_1(\delta)$ or the commuting subset $\Gamma \operatorname{star}_1(v)$ (to adopt an obvious notation). In the former case, we choose a unit vector $e_{0,0}$ (in fact unique up to a $\pm$ sign) such that

$$\begin{aligned} &\Gamma(\alpha) e_{0,0} = e_{0,0}, \quad \text{for all } \alpha \in S_2, \\ &\Gamma(\lambda) e_{0,0} = e_{0,0}, \quad \text{for all } \lambda \in S_1(\delta), \end{aligned} \tag{5.25}$$

and define further vectors $e_{0,d} \in W_{0,d}$ by

$$e_{0,d} = \Gamma(\nu_d) e_{0,0}, \quad d \in \delta, \tag{5.26}$$

and thereby obtain a basis $\{e_{0,d};\ d \in V(3)\}$ for W_0 which simultaneously diagonalizes the elements of $\Gamma S_1(\delta)$,

$$\Gamma(\lambda) e_{0,d} = \epsilon(d,\lambda) e_{0,d}, \quad \lambda \in S_1(\delta). \tag{5.27}$$

In the latter case, we choose a unit vector f_0^0 such that

$$\begin{aligned} &\Gamma(\alpha) f_0^0 = f_0^0, \quad \text{for all } \alpha \in S_2, \\ &\Gamma(\nu_d) f_0^0 = f_0^0, \quad \text{for all } d \in \delta, \end{aligned} \tag{5.28}$$

and define further vectors $f_0^\lambda \in W_0^\lambda$ by

$$f_0^\lambda = \Gamma(\lambda) f_0^0, \quad \lambda \in S_1(\delta), \tag{5.29}$$

and thereby obtain a basis $\{f_0^\lambda;\ \lambda \in \hat{V}(3)\}$ for W_0, which simultaneously diagonalizes the elements of $\Gamma \operatorname{star}_1(v)$,

$$\Gamma(\nu_d) f_0^\lambda = \epsilon(p,\lambda) f_0^\lambda, \quad d \in \delta. \tag{5.30}$$

Bearing in mind (3.4), (3.5), and *adopting the orientation convention of Sec. V B, whereby δ and also each element of $st_2(v)$ is a WOP*, note that the action of the $\Gamma(\nu_c)$ upon the e-basis for W_0 is given by

$$\Gamma(\nu_c) e_{0,d} = e_{0,c+d}, \quad c,d \in V(3), \tag{5.31}$$

and that of the $\Gamma(\lambda)$, $\lambda \in S_1(\delta)$, upon the f-basis for W_0 is given by

$$\Gamma(\lambda) f_0^\mu = f_0^{\lambda+\mu}, \quad \lambda,\mu \in \hat{V}(3). \tag{5.32}$$

By use of (5.19), or no doubt from more elementary considerations, we can obtain bases for each subspace W_p, $p \in S_0$, by taking the image under Γ_p of the above bases for W_0. We thereby obtain two different bases for E, namely

$$\{e_{p,d};\ p \in V(4),\ d \in V(3)\} \tag{5.33}$$

and

$$\{f_p^\lambda;\ p \in V(4),\ \lambda \in \hat{V}(3)\}, \tag{5.34}$$

where we have defined

$$e_{p,d} = \Gamma_p e_{0,d} = \Gamma_p \Gamma(\nu_d) e_{0,0} \tag{5.35}$$

and

$$f_p^\lambda = \Gamma_p f_0^\lambda = \Gamma_p \Gamma(\lambda) f_0^0. \tag{5.36}$$

Both bases are of course orthonormal bases.

It is now straightforward to compute the effect of the Γ_p upon the spinor basis (5.33) in terms of "incidence numbers" [and involving also the orientation numbers $\eta(p,q)$ of (5.16)]. Of course the action of Γ_p upon the basis vector $e_{0,d}$ is by definition $e_{p,d}$, and so we also have $\Gamma_p e_{p,d} = -e_{0,d}$. Consequently we only have need to further compute

$$\Gamma_p e_{q,d} = \Gamma_p \Gamma_q \Gamma(\nu_d) e_{0,0}, \tag{5.37}$$

in the cases $p \neq q \in S_0$. Now, after using (5.16), the resulting Γ-line $\Gamma(j(p,q))$ can be moved through $\Gamma(\nu_d)$, thus contributing the incidence number

$$i(p,q,d) = \begin{cases} \epsilon(j(p,q),\nu_d), & \text{if } d\in\delta, \\ 1, & \text{if } d = 0. \end{cases} \tag{5.38}$$

There are now two cases to consider. In the first case the line $j(p,q)$ lies in δ, whence $\Gamma(j(p,q))e_{0,0} = e_{0,0}$, by (5.25). In the second case $j(p,q)$ meets δ in a point, say

$$c = c(p,q,\delta) = j(p,q)\cap\delta. \tag{5.39}$$

In this second case, either $\mu\equiv j(p,q)$ equals ν_c or else we have, cf. (3.6),

$$\mu = \nu_c\lambda\alpha, \tag{5.40}$$

where $\lambda\in S_1(\delta)$, $\alpha = j(v,\mu)\in st_2(v)$. In this second case we thus have, using (5.15) and (5.25) together with the agreed orientation convention,

$$\Gamma(\mu)e_{0,0} = +\Gamma(\nu_c)e_{0,0} = e_{0,c}. \tag{5.41}$$

If we extend the definition of c to be

$$c = \begin{cases} 0, & \text{if } j(p,q)\subset\delta, \\ j(p,q)\cap\delta, & \text{otherwise,} \end{cases} \tag{5.42}$$

then (5.41) applies to both of the cases. Putting the foregoing together we have the final result, valid for $p\neq q\in S_0$ and $d\in V(3)$,

$$\Gamma_p e_{q,d} = \eta(p,q)i(p,q,d)e_{p+q,c+d}, \tag{5.43}$$

where $i(p,q,d)$ and c are as defined in (5.38) and (5.42).

The effect of the Γ_p upon the spinor basis (5.34) can be derived by an entirely analogous computation. The final result, valid for $p\neq q\in S_0$ and $\mu\in\hat{V}(3)$, is

$$\Gamma_p f_q^\mu = \eta(p,q)k(p,q,\mu)f_{p+q}^{\lambda+\mu}, \tag{5.44}$$

where

$$k(p,q,\mu) = \begin{cases} \epsilon(j(p,q),\mu), & \text{if } \mu\in S_1(\delta), \\ 1, & \text{if } \mu = 0, \end{cases} \tag{5.45}$$

and

$$\lambda = \begin{cases} 0, & \text{if } j(p,q)\in st_1(v), \\ j(v,j(p,q))\cap\delta, & \text{otherwise.} \end{cases} \tag{5.46}$$

[That is, in the last line, λ is the projection from v of $j(p,q)$ onto the base δ.]

Remark: Reverting to U in (5.21) as an induced representation, let us spell out more concerning the eight-dimensional irreducible representation T of G_1 carried by the subspace W_0. (Of course, since G_1 is normal in G_0, each of the remaining subspaces W_p, $p\in S_0$, will carry a representation of G_1 which is equivalent to a conjugate of T; one sees that all 16 of those inequivalent irreducible representations of G_1 which represent Z_0 faithfully are thus involved in U.) On account of the defining property (5.22) of W_0 note that $T(\Gamma(\alpha)) = I$, for $\alpha\in K_2$, so that T is in effect a representation of the factor group G_1/K_2, which is non-Abelian and of order 128. Let us define for $d\in V(3) = \{0\}\cup\delta$ and $\lambda\in\hat{V}(3) = \{0\}\cup S_1(\delta)$,

$$Q(d) = T(\Gamma(\nu_d)), \quad R(\lambda) = T(\Gamma(\lambda)). \tag{5.47}$$

Then the representation T is determined by the properties of $Q(d)$, $R(\lambda)$,

$$Q(c)Q(d) = Q(c+d), \quad R(\lambda)R(\mu) = R(\lambda+\mu), \tag{5.48}$$

and, for $d\in\delta$, $\lambda\in S_1(\delta)$,

$$Q(d)R(\lambda) = \epsilon(d,\lambda)R(\lambda)Q(d). \tag{5.49}$$

Now the group of R-operators is generated by $R_i = R(\lambda_i)$, $i = 1,2,3$, where $\lambda_1, \lambda_1, \lambda_3$ denote the base edges of the tetrahedron of reference, and the group of Q-operators is generated by $Q_i = T(\Gamma(\nu_i))$, $i = 1,2,3$, where ν_1, ν_2, ν_3 denote the edges of the tetrahedron which are concurrent at the vertex v. Labelling these in such a way that λ_i is skew to ν_i, $i = 1,2,3$, then the six-operators $Q_1, Q_2, Q_3, R_1, R_2, R_3$ fall into three mutually commuting pairs (Q_i,R_i), $i = 1,2,3$. The operators Q_i, R_i of the ith pair anticommute and generate a group $G^{(i)}$, say, isomorphic to the dihedral group D_8 of order 8. Since Z_0 is to be represented faithfully, we require the two-dimensional irreducible representation of each of the dihedral groups $G^{(i)}$, and taking the tensor product of these two-dimensional representations, we arrive at the required eight-dimensional representation of the generators of the group of operators $T(h)$, $h\in G_1$. Notice that this last group is thus seen to be isomorphic to the central product of three copies of the dihedral group D_8.

VI. CONJECTURES AND SPECULATIONS

The general results of Sec. IV, along with the further $m = 3$ results in Sec. V A, strongly incline one to believe in the following two conjectures.

Conjecture A: (i) If $m = 2l$ then G_l is a maximal Abelian normal subgroup of G_0.

(ii) If $m = 2l+1$ then both $G_{l+1,\delta}$ and $G_{l+1,v}$ are (for any choice of base hyperplane δ and any choice of vertex v) maximal Abelian normal subgroups of G_0.

Conjecture B: For each $r = 0,1,\ldots,m$, the subgroups G_{m-r}, G_r form a dual pair of subgroups of G_0 (in the sense that either subgroup is the full centralizer in G_0 of the other subgroup).

Of course Conjecture A (i) is the special case $r = l$, $m = 2l$ of Conjecture B. Certainly, as we have just seen in Sec. V, both conjectures are true in the case $m = 3$, and they are easily seen to be true in the case $m = 2$, $d = 7$ previously studied.[7] The author believes that the methods of Sec. V could be pushed further to establish A and B in the case $m = 4$, but that different methods may have to be employed to establish their truth (or falsity!) for $m > 4$.

At any rate, the next case $m = 4$, $d = 31$ would certainly appear to merit further study, especially as the even cases seem to enjoy somewhat more symmetrical properties. In the first even case $m = 2$, all the Γ-lines commuted amongst themselves, and in the next even case $m = 4$, all the Γ-planes commute, and can thus be simultaneously diagonalized; they presumably generate, in fact, a maximal Abelian subgroup of G_0. Incidentally recall that, by a standard doubling process, $d = 31$ Clifford algebra is intimately related to $d = 32$ Clifford algebra. In particular, the 496 infinitesimal generators of the two fundamental spinor representatons of SO(32) can be taken to be the $31 + 465$ sets $(\Gamma_p;\Gamma_p\Gamma_q)$ and $(-\Gamma_p;\Gamma_p\Gamma_q)$ of $d = 31$ Clifford operators. Is it just conceivable that the special PG(4,2) properties of $d = 31$ Clifford algebra tie in, in some way, with the special nature[15] of SO(32) in the context of superstring theories?

Further conjectures can be made concerning the subgroup chains

$$C_0\supset C_1\supset\cdots\supset C_{m-1}\supset C_m = \{1\} \tag{6.1}$$

and

$$G_0 \supset G_1 \supset \cdots \supset G_{m-1} \supset G_m = Z_0, \tag{6.2}$$

in the case of general m. From the definitions of the subgroups C_r and G_r, C_r is clearly an Ω-admissible subgroup of C_0, and G_r is an $\overline{\Omega}$-admissible subgroup of G_0 [see (1.19)]. In fact G_r, being normal, is also stable under all inner automorphisms of G_0. Let the latter taken in conjunction with the outer automorphisms belonging to $\overline{\Omega}$ generate a subgroup Ω^* of Aut G_0. Then we can say that (6.1) is a Ω-series for C_0 and that (6.2) is a Ω^*-series for G_0. If the Ω-series (6.1) has the property that, for each $r = 1,2,...,m$, C_r is a maximal Ω-subgroup of C_{r-1}, then we shall term (6.1) to be a principal[16] Ω-series. A principal Ω^*-series is defined analogously.

Conjecture C: (i) The Ω-series (6.1) is a principal Ω-series for C_0.

(ii) The Ω^*-series (6.2) is a principal Ω^*-series for G_0. Sticking our neck out still further, we put forward tentatively an even stronger conjecture.

Conjecture D: (i) The group C_0 possesses the unique principal Ω-series (6.1).

(ii) The group G_0 possesses the unique principal Ω^*-series (6.2).

Remark: If C could be established then B would follow. To see this, note first of all that the centralizer Z_r of the Ω^*-admissible subgroup G_r is itself Ω^*-admissible. Secondly note that we have the inclusions

$$Z_{m-r} \supseteq Z_{m-r-1} \supseteq G_{r+1} \tag{6.3}$$

(since $G_{m-r} \subset G_{m-r-1}$, and by Theorem 4.3). Consequently, granted C, we see that

$$Z_{m-r} = G_r \quad \text{implies} \quad Z_{m-r-1} = G_{r+1} \tag{6.4}$$

(for Z_{m-r-1} can not equal $Z_{m-r} = G_r$, since an $(m-r-1)$-space can be skew to an r-space). Now $Z_{m-r} = G_r$ holds for $r = 0$, and so, granted C, the inductive step (6.4) establishes $Z_{m-r} = G_r$ for each $r = 0,1,2,...,m$.

Concerning D, at least it can be seen to be true in the cases $m = 2$ and $m = 3$.

One can speculate in another direction and enquire whether interesting algebras are brought to one's attention by thinking in PG(m,2) terms. In the case $m = 2$, the seven operators Γ_p, acting upon an eight-dimensional space, can be given an octonionic interpretation in terms of the operators $L(e_p)$ of left multiplication by the imaginary octonionic units. In the case $m = 3$, the operators $\Gamma_p \Gamma(\lambda)$, $\lambda \in \hat{V}(3)$, will be similarly associated with an algebra of dimension 128. [Another algebra of dimension 128 will be obtained by using instead the operators $\Gamma_p \Gamma(\nu_d)$, $d \in V(3)$.] There seems to be a good possibility that these algebras may have interesting properties, perhaps generalizing in some sense the composition algebra property of the octonions, and may have large automorphism groups. Probably even more worthy of investigation will be the algebra (of still higher symmetry?) of dimension 2^{15} which arises out of the case $m = 4$, $d = 31$.

ACKNOWLEDGMENTS

I would like to acknowledge an interesting and informative correspondence with the late Frank Adams, with Keith C. Hannabus, and with Eamonn A. O'Brien. This correspondence arose in connection with a (in retrospect, rather foolish!) speculation of the author's that certain well-known exceptional features of dimension 8 might show up even at the level of the finite group G_0, of order 128, associated with Cliff(0,7). In particular the author speculated—see the final remark of Ref. 7—that G_0 (in the case $m = 2$, $d = 7$) could perhaps be exceptional in its possession of a maximal Abelian normal subgroup of the kind $(\mathbb{Z}_2)^q$. But in fact, as pointed out by O'Brien and Slattery[17] (see also the earlier paper by Braden[18] which consists of a thorough account of the finite groups associated with real Clifford algebras of arbitrary signature), dimension $d = 7$ is far from exceptional in this respect. Indeed wherever $d = 8k + 7$, for $k \geqslant 0$, the group G_0, of order 2^d, is a central product (i.e., a product with amalgamation of the center) of dihedral groups of order 8, and it always contains a maximal Abelian normal subgroup $\simeq (\mathbb{Z}_2)^q$, with $q = \frac{1}{2}(d+1)$. Of course in the present paper, we have, in the case $d = 15$, arrived at two explicit examples of such a maximal subgroup, namely the subgroups $G_{2,\delta}$ and $G_{2,\upsilon}$ of Theorem 5.1.

APPENDIX: TOTAL ORIENTATIONS OF PG(3,2)

Recall the definition of a WOP (= well-oriented plane) as defined via (5.14 +), or equivalently by (5.15). Given a single plane γ, it is a simple matter to totally orient it in such a manner that it becomes a WOP. For example, let $\omega = abc$ be any line of γ and let υ be any point not on ω. Let $\nu_p = j(\upsilon,p), p \in \omega$, and let μ_p ($= \nu_p \omega \gamma$) complete the triple of lines ν_p, ω, μ_p which pass through $p \in \omega$. [Incidentally, take note that in PG(3,2), in contrast with PG(2,2), the seven lines ω, ν_a, ν_b, ν_c, μ_a, μ_b, μ_c of γ, together with 1, do *not* form a subgroup of C_0. But of course the three nonconcurrent lines μ_a, μ_b, μ_c *do* generate a subgroup of order eight, whose elements are seen to be

$$\{1, \mu_a, \mu_b, \mu_c, \nu_a\gamma, \nu_p\gamma, \nu_c\gamma, \omega\}, \tag{A1}$$

after recalling that $\nu_a \nu_b \nu_c = \gamma$, as in (2.21).] Let us now choose any orientations for the four lines $\nu_a, \nu_b, \nu_c, \omega$ by making arbitrary choices of coset representatives $\Gamma(\nu_a), \Gamma(\nu_b), \Gamma(\nu_c), \Gamma(\omega)$. We then fix the orientations of the plane γ and of the remaining three lines by defining their coset representatives to be

$$\Gamma(\gamma) = \Gamma(\nu_a)\Gamma(\nu_b)\Gamma(\nu_c) \tag{A2}$$

and

$$\Gamma(\mu_p) = \Gamma(\nu_p)\Gamma(\omega)\Gamma(\gamma), \quad p = a,b,c. \tag{A3}$$

By our definitions, (A2) and (A3), we have satisfied (5.14 +) at the four points υ, a, b, c, and an easy check shows that (5.14 +) is satisfied also at the remaining three points of γ. Hence we have made γ into a WOP.

When we now consider the entire family of 15 planes in PG(3,2) we can not of course totally orient each plane independently, since each of the 35 lines of PG(3,2) is common to three planes. On the other hand, we have the freedom of choice of $2^4 = 16$ total orientations for each plane, subject to it remaining a WOP, corresponding to the four arbitrary choices of orientations for the lines ν_a, ν_b, ν_c, ω in the preceding paragraph. In the face of such a variety of choices,

and the profusion of their mutual interactions, it is quite easy to become bewildered! The point of this Appendix is to describe a total orientation for PG(3,2) whose simplicity would be hard to better.

To this end, choose a vertex v and a base δ, and make δ into a WOP. For each line $\lambda \subset \delta$ let the three planes of $\mathrm{star}_2(\lambda)$ be δ, $\alpha(\lambda)$, $\beta(\lambda)$, where $\beta(\lambda) = \delta\alpha(\lambda)$ and $\alpha(\lambda) = j(v,\lambda)$. As λ varies over the seven lines of δ, we obtain seven "α-planes" which pass through v and seven "β-planes" which are distinct from δ and yet do not belong to $\mathrm{star}_2(v)$. For $d \in \delta$ set $\nu_d = j(v,d)$. Recall from (2.23) that $\mathrm{st}_1(v) = 1$. Let us make arbitrary choices of orientation for the seven lines ν_d of $\mathrm{st}_1(v)$ subject only to the demand that $\Gamma\mathrm{st}_1(v)$ be $+I$ rather than $-I$,

$$\prod_{d\in\delta} \Gamma(\nu_d) = +I. \tag{A4}$$

Four lines of the α-plane $\alpha(\omega)$ have now been given orientations, namely $\omega, \nu_a, \nu_b, \nu_c$, where a,b,c denote the points of the line $\omega \subset \delta$, and we can, as in the opening paragraph, make $\alpha(\omega)$, for each $\omega \in S_1(\delta)$, into a WOP by means of (A2), (A3) [with $\alpha(\omega)$ instead of γ]. In this way, all of the $3\times 7 = 21$ lines not in $S_1(\delta)$ or $\mathrm{star}_1(v)$ have now also been given orientations. So far we have eight WOP's, the seven α-planes together with the base δ. So we need to look at the remaining seven planes of PG(3,2), the seven β-planes.

In the plane $\beta(\omega)$, we need to distinguish between the points a,b,c of $\omega \subset \delta$ and the remaining four points of $\beta(\omega)$ which do not lie in δ. Consider, first of all, a point $a \in \omega$, and let the three lines of $\mathrm{star}_1(a,\beta(\omega))$ be ω, μ_a, μ'_a. Let $j(v,\mu_a)$ and $j(v,\mu'_a)$ meet δ in the lines λ_a, λ'_a, respectively. Since $\alpha(\lambda_a)$ and $\alpha(\lambda'_a)$ are WOP's, we have

$$\begin{aligned}\Gamma(\mu_a) &= \Gamma(\lambda_a)\Gamma(\nu_a)\Gamma(\alpha(\lambda_a)),\\ \Gamma(\mu'_a) &= \Gamma(\lambda'_a)\Gamma(\nu_a)\Gamma(\alpha(\lambda'_a)).\end{aligned} \tag{A5}$$

Now if $\lambda, \lambda', \lambda''$ are any three concurrent lines of δ we have, as an easy consequence of (A3) and (A2) (the latter applied three times, with $\alpha(\lambda), \alpha(\lambda'), \alpha(\lambda'')$ instead of γ, and $\lambda, \lambda', \lambda''$ instead of ω),

$$\Gamma(\alpha(\lambda))\Gamma(\alpha(\lambda'))\Gamma(\alpha(\lambda'')) = I. \tag{A6}$$

Consequently, upon using (A5) in conjunction with (A6), we can evaluate $\Gamma\mathrm{st}_1(a,\beta(\omega))$ and obtain (since $\lambda_a, \lambda'_a, \omega$ are concurrent lines of δ, and since δ is a WOP)

$$\begin{aligned}\Gamma(\omega)\Gamma(\mu_a)\Gamma(\mu'_a) &= \Gamma(\omega)\Gamma(\lambda_a)\Gamma(\lambda'_a)\Gamma(\alpha(\omega))\\ &= \Gamma(\delta)\Gamma(\alpha(\omega)).\end{aligned} \tag{A7}$$

Let us now agree to orient plane $\beta(\omega) = \delta\alpha(\omega)$ by defining

$$\Gamma(\beta(\omega)) = \Gamma(\delta)\Gamma(\alpha(\omega)). \tag{A8}$$

Taking (A7) and (A8) together, we see that the three points of a base line ω are "good" points for the plane $\beta(\omega)$ in the sense that $(5.16+)$ holds at these points rather than $(5.16-)$ [with $\beta(\omega)$ replacing γ in these equations].

Consider now a point $q \in \beta(\omega)$ which does not lie on ω, and so does not lie in δ. Let the joins of q to the points a,b,c of ω be μ_a, μ_b, μ_c, and let the projections of the latter onto δ from v be $\lambda_a, \lambda_b, \lambda_c$. Since $\alpha(\lambda_p)$, $p = a,b,c$, is a WOP we have, from (A3),

$$\Gamma(\mu_p) = \Gamma(\lambda_p)\Gamma(\nu_p)\Gamma(\alpha(\lambda_p)), \quad p = a,b,c. \tag{A9}$$

Consequently, we can compute

$$\Gamma\mathrm{st}_1(q,\beta(\omega)) \equiv \Gamma(\mu_a)\Gamma(\mu_b)\Gamma(\mu_c) \tag{A10}$$

as follows. First of all, note that $\lambda_a, \lambda_b, \lambda_c$ are concurrent lines of δ, since they are projections from v of concurrent lines of $\beta(\omega)$, and so we can make use of (A6). Secondly note that, in contrast to the corresponding computation of $\Gamma\mathrm{st}_1(a,\beta(\omega))$ in (A7), our various Γ-lines do not all commute, since λ_p is skew to ν_r for $p \neq r$. Consequently, (A9) leads to

$$\begin{aligned}&\Gamma(\mu_a)\Gamma(\mu_b)\Gamma(\mu_c)\\ &\quad= (-1)^3\Gamma(\lambda_a)\Gamma(\lambda_b)\Gamma(\lambda_c)\Gamma(\nu_a)\Gamma(\nu_b)\Gamma(\nu_c)\\ &\quad= -\Gamma(\delta)\Gamma(\alpha(\omega))\\ &\quad= -\Gamma(\beta(\omega)),\end{aligned} \tag{A11}$$

since δ and $\alpha(\omega)$ are WOP's, after using (A8). Thus the four points of $\beta(\omega)$ not lying in the base are "bad" points in the sense that $(5.16-)$ holds rather than $(5.16+)$. Let us call a totally oriented plane a BOP ($\equiv$badly oriented plane) if it possesses a line λ consisting of good points but whose complement λ^c consists of bad points. Our foregoing results can then be summarized as in the following theorem [the second part of the theorem being an immediate consequence of (A6) and (A8)].

Theorem A1: Relative to any choice of vertex v and of base δ, there exists a total orientation for PG(3,2) such that δ together with every plane of $\mathrm{star}_2(v)$ are WOP's and such that each of the remaining seven planes β is a BOP whose good line is $\beta \cap \delta$. The oriented planes, together with I, form a subgroup

$$K_2 = \{I, \Gamma(\gamma);\ \gamma \in S_2\}$$

of G_2, isomorphic to $\hat{V}(4)$, such that $G_2 = K_2 \times Z_0$.

Remark: For given v and δ there are 1024 choices of total orientation for PG(3,2) which achieve the stated results. For there are 2^4 way of making δ a WOP, and there are 2^6 ways of orienting the seven lines of $\mathrm{star}_1(v)$ subject to the one constraint (A4). The minus sign in (A11) comes about for all of these 2^{10} choices of total orientation. This strongly suggests the final conjecture.

Conjecture E: For any total orientation of PG(3,2) there exist at most eight WOP's.

[1] M. F. Atiyah, R. Bott, and A. Shapiro, Topology **3** (Suppl. 1), 3 (1964).
[2] I. R. Porteous, *Topological Geometry* (Van Nostrand, London, 1969).
[3] P. Lounesto, Found. Phys. **11**, 721 (1981).
[4] R. Coquereaux, Phys. Lett. B **115**, 389 (1982).
[5] I. M. Benn and R. W. Tucker, *An Introduction to Spinors* (Hilger, Bristol, 1987).
[6] P. Budinich and A. Trautman, *The Spinorial Chessboard* (Springer, Berlin, 1988).
[7] R. Shaw, J. Phys. A: Math. Gen **21**, 7 (1988).
[8] H. H. Goldstine and L. P. Horwitz, Math. Ann. **154**, 1 (1964).
[9] M. Rooman, Nucl. Phys. B **236**, 501 (1984).
[10] J. W. P. Hirschfeld, *Projective Geometries over Finite Fields* (Clarendon, Oxford, 1979).
[11] J. W. P. Hirschfeld, *Finite Projective Spaces of Three Dimensions* (Clarendon, Oxford, 1985).
[12] R. Howe, Lectures in Applied Maths. (Am. Math. Soc.) **21**, 179 (1985).

[13]A. H. Clifford, Ann. Math. **38**, 533 (1937).
[14]R. Shaw, *Linear Algebra and Group Representations* (Academic, New York, 1983), Vol. 2.
[15]M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge U.P., Cambridge, 1987).
[16]cf. M. Hall, *The Theory of Groups* (Chelsea, New York, 1976), 2nd ed., p. 124.
[17]E. A. O'Brien and M. C. Slattery, "Clifford algebras and finite groups," J. Phys. A: Math. Gen. (to be published).
[18]H. W. Braden, J. Math. Phys. **26**, 613 (1985).

Modular representations as a possible basis of finite physics

Felix Lev
Institute of Applied Mathematics (Khabarovsk Division), ul. Kim-Yu-Chena 65, Khabarovsk 680063, USSR

(Received 31 January 1989; accepted for publication 5 April 1989)

The approach in which physical systems are described by the elements of a linear space over a finite field, and operators of physical quantities by linear operators in this space, is discussed. The mathematical formulation of the correspondence between such a description and the conventional one is given for a case when the characteristic of the finite field is sufficiently large. The above correspondence is considered in the examples of finite analogs of representations of so(2,3), so(1,4), and osp(1,4) algebras. Finite analogs of representations of infinite-dimensional algebras are briefly discussed.

I. INTRODUCTION

The notion of the field of rational numbers Q is undoubtedly connected with an everyday experience that any physical quantity can be divided into an arbitrary number of equal parts. However, this is not obviously the case on the microscopic level. Indeed, if for example, e is the minimal electric charge in nature, then the quantity e/n for $n = 2,3,4,...$ has no sense if the division is understood conventionally.

In the Copenhagen formulation of quantum theory, the state of a system is fully defined by the vector from a separable Hilbert space, and the main physical quantities are the relative probabilities determined by the coefficients of decomposition of this vector relative to some orthogonal basis. It is natural that the notion of relative probability is introduced by means of conventional division; however, this notion is only a formal way of description of an experiment. One can notice that all information about probabilities is given by a set of integers because we conduct an experiment N times and observe that the first possibility is realized in n' cases, the second in n'' cases, and so on (recall the well-known Kronecker's expression about natural numbers). Therefore one could try to describe probabilities by integers only and normalize probabilities not on unity but on some large number N. Notice also that the notion of probability is an idealized one because in reality the number of experiments cannot be infinite.

It is also natural to deal with the field Q if we wish to introduce the notion of average value; however, it is pure mathematical notion and is not observable. Indeed, if for example, some physical quantity can assume the values 0 and 1 with equal probabilities, then its average value is 1/2 in the conventional sense but the measurement of this quantity can give only the values either 0 or 1.

Proceeding from what was mentioned above, one may believe that future fundamental physics will be based not on the field Q and its expansions but on some finite field. In this case there is also a hope that the theory will become not only discrete but even finite. In most works on this subject (see e.g., Refs. 1–7) finite fields were used for the quantization of space-time and this seems natural if one constructs the quantum theory proceeding from the transition amplitudes. In operator formalism, however, the main object of the theory is a representation of the symmetry group or algebra in the corresponding Hilbert space and the notion of space-time is secondary (in particular the coordinate operator in the nonrelativistic case and the Newton–Wigner position operator in the relativisitic one can be expressed as the functions of representation generators of the corresponding symmetry group). Therefore if one proceeds from the operator formalism, then the first thing coming to mind is to describe physical systems by the elements of some linear space over a finite field, and physical quantities by linear operators in this space.

In mathematics, the representations in the spaces over a field with a nonzero characteristic are called modular representations. An investigation of modular representations of Lie algebras and Chevalley groups has been carried out by many authors (see, e.g., Refs. 8–38 and references cited therein). It is important to note that the general theory is constructed in spaces over an algebracially closed (i.e., infinite) field. This is due in particular to the fact that in an algebraically unclosed field not every linear operator in finite-dimensional space has an eigenvector, since the characteristic equation may have no solution in this field.

Though the complete theory of modular representations is not yet constructed, there are many important results. As it will be clear from the subsequent exposition, the representations of Lie algebras and superalgebras are of major interest to us. Since the paper of Zassenhaus[9] it is known that all irreducible modular representations are finite-dimensional and many papers have dealt with the maximal dimension of such representations. For classical Lie algebras this was done in Refs. 8–10, 13, 15 and for algebras of the Cartan series in Refs. 9, 15, 20, 22–25. The complete classification of irreducible representations over an algebraically closed field has been made only for A_1 algebras.[10] The important results for A_2 and B_2 algebras are obtained in Refs. 11, 18; for A_n algebras see Refs. 18, 28, 30, and results obtained for some other algebras are in Ref. 18.

In the present paper the general theory of modular representations is not used and only representations over the Galois field containing p^2 elements (where p is a prime number) are considered. In Sec. II the notion of correspondence between modular representations and conventional representations in Hilbert spaces is introduced for the cases when

 0022-2488/89/091985-14$02.50

the number p is sufficiently large. In other words, it is a question of correct transition from "modular physics" to conventional physics in the limit $p\to\infty$. In Sec. III some auxiliary modular representations are discussed and in Secs. IV–VI we perform the explicit construction of modular analogs of representations describing particles in the de Sitter spaces. In Sec. VII we construct the modular analogs of representations describing particles in the de Sitter superspace, and Sec. VIII contains a few remarks on modular analogs of representations of infinite-dimensional Lie algebras.

In the last two years the number of papers in which the principal field in string theory is not the field of complex numbers C, but the p-adic[6,7,39–46] or adelic field,[47–49] has essentially increased. The latter choice is motivated in particular by the confidence that there is no reason to prefer any special value of p. It is clear that the physics in such versions cannot be finite. The authors of these works proceed from the transition amplitudes formulation of quantum theory. Note that though this formalism is traditionally regarded as more powerful than the operator one; the latter formalism, as has been pointed out in recent papers,[50,51] has its own merits.

The question of correspondence between modular and ordinary representations has been considered in our previous paper[52] "on the physical level of rigor" and we exposed briefly the results about modular analogs of representations of so(2,3) and so(1,4) algebras. In the present work we give the mathematical formulation of the correspondence and the results on modular analogs of all considered representations are given with proofs.

In mathematical literature on modular representations, special attention is given to the relationship between modular and ordinary representations of Chevalley groups over a field of characteristic p (see Refs. 12, 17–19, 21, 26, 27, 29, 31–34). The correspondence described in our paper differs from that considered in the above references. It is clear that if we wish to treat the conventional physics as a limit of some physics with the parameter p, then we should consider a correspondence between two cases where one of them does not contain this parameter at all.

The author hopes that in order to understand this work, the reader need not know even the specific facts about the ordinary representations of the de Sitter algebras, because the ordinary and modular cases are treated on the same basis. Note also that, as it will be clear soon, the physical meaning may have only modular analogs of representations describing particles or strings in spaces of the de Sitter but not Minkowski type.

II. CORRESPONDENCE BETWEEN MODULAR AND ORDINARY REPRESENTATIONS OF ENVELOPING ALGEBRAS

Let Z be the ring of integers and $F_p = Z/pZ$ be the residue field of characteristic p. The elements of F_p will be denoted as $0,1,\ldots, p-1$. As usual, if $a\in F_p$ then $-a$ means the elements $b\in F_p$ such that $a+b=0$ in F_p. Therefore, by definition, $-1=p-1$, $-2=p-2$, etc.

Consider F_p as the ring relative to the addition, subtraction, and multiplication. Introduce two functions from F_p into Z such that if $a=0,1,\ldots,p-1$ then $\rho(a)=\min\{a,p-a\}$ and

$$f(a)=\begin{cases} a, & \text{if } a<p-a,\\ -(p-a), & \text{if } a>p-a. \end{cases} \tag{1}$$

It is obvious that

$$\rho(a)\in[0,(p-1)/2],$$
$$\rho(-a)=\rho(a),$$
$$f(-a)=-f(a),$$

and

$$|f(a)|=\rho(a).$$

The reflection f is not a homomorphism of rings F_p and Z. However, it is in some sense an isomorphism between elements $z\in Z$ for which $|z|$ is much less than p, and elements $a\in F_p$ for which $\rho(a)$ is also much less than p. This isomorphism may be understood as follows. Denote $C(p)=p^{1/(\ln p)^{1/2}}$ and U_0 as the set of $a\in F_p$ for which $\rho(a)<C(p)$. It is easy to see that if $a_1,\ldots,a_n\in U_0$ and n_1, n_2 are natural numbers such that

$$n_1<(p-1)/2C(p),\quad n_2<\left[\ln\frac{(p-1)}{2}\right]\Big/(\ln p)^{1/2},$$

then

$$f(a_1\pm a_2\pm\cdots\pm a_n)=f(a_1)\pm f(a_2)+\cdots\pm f(a_n),$$
$$f(a_1\cdots a_n)=f(a_1)\cdots f(a_n).$$

Hence if p is sufficiently large then for a sufficiently large number of elements from U_0 the addition, subtraction, and multiplication are performed according to the same rules as for elements $z\in Z$ for which $|z|<C(p)$, and when p becomes larger, then the larger part of F_p can be put into correspondence with the elements of the ring Z.

Proceeding from the experience, modern physics states that the addition, subtraction, and multiplication of integers should be performed according to the rules for the ring Z. It cannot be excluded, however, that physics in our universe should be based on the rules for the ring F_p at some large p, and the concrete value of p will be determined from future experiments or from a more general physical theory.

It is necessary to keep in mind that even for elements from U_0 the results of division in the field F_p, generally speaking, differ from that in Q. For example, the element $1/2$ in F_p is equal to the integer $(p+1)/2$. This does not mean, however, that physics based on F_p contradicts the experience because, as it was explained in Sec. I, the observables in quantum physics are defined without division, and division can be viewed as a purely mathematical notion.

Since we wish to have the finite analog of the complex version of quantum theory, then it is natural (instead of the field C) to deal with the field F_{p^2}, consisting of p^2 elements. The actions in F_{p^2} are the natural generalizations of the actions in F_p if the elements of F_{p^2} are formally represented as $a+bi$ ($a,b\in F_p$, $i^2=-1$) and the element inverse to $a+bi$ can be written as $a/(a^2+b^2)-(b/(a^2+b^2))i$, assuming

that here and henceforth the division is understood in the sense of F_p. Therefore we must be sure that if $a \neq 0$ or $b \neq 0$ then $a^2 + b^2 \neq 0 \bmod p$. It is known that this condition is satisfied if $p-1$ is not divided by 4 (in the usual sense). Note that in the case of the quaternionic or octonionic versions of quantum theory, the analogous reasonings are not valid since, according to the Lagrange theorem, the number p can be represented as a sum of the squares of four integers.

The field F_{p^2} can be represented as the residue field $(Z+iZ)/p(Z+iZ)$. Let U be the set of elements $a+bi \in F_{p^2}$, for which $a,b \in U_0$. Then the function from F_{p^2} into $Z+iZ$, defined by the formula $f(a+bi) = f(a)+f(b)i$ (we use the same notation f), is the local isomorphism between F_{p^2} and $Z+iZ$ in the same sense that f is the local isomorphism between F_p and Z.

We will assume that the role of the state space for the system under consideration is played by the linear space V over F_{p^2}. Since we wish to have an analog of the Hilbert space, we will assume that V is supplied by the scalar product $(...,...)$ such that for $x, y \in V$, $(x,y) \in F_{p^2}$ and the equalities

$$(x,y) = \overline{(y,x)},$$

$$(ax,y) = \bar{a}(x,y),$$

and

$$(x,ay) = a(x,y) \quad (a \in F_{p^2})$$

are also understood in the sense of F_{p^2}. One should keep in mind that such a scalar product cannot define in V any positively defined metric, and thus in the modular case there is no probabilistic interpretation. We shall see, however, that in the cases considered below, the probabilistic interpretation can be restored in the limit $p \to \infty$.

In the spaces considered below, there exists a basis $\{e_1,e_2,...,e_n\}$ such that $(e_j,e_k) = 0$ for $j \neq k$ and $(e_j,e_j) \neq 0$ for all j (in general case such a basis does not exist). If

$$x = c_1e_1 + \cdots + c_ne_n \quad (c_j \in F_{p^2}),$$

the coefficients c_j in this case are uniquely defined by the formula $c_j = (e_j,x)/(e_j,e_j)$. If $\{f_1,...,f_n\}$ is another basis, then the elements $\{f_j\}$ are in $1 \leftrightarrow 1$ correspondence with $\{e_j\}$, but it may occur that for some j $(f_j,f_j) = 0$ (see Sec. IV).

As usual, if L_1 and L_2 are linear operators in V such that

$$(L_1x,y) = (x,L_2y) \quad \forall x,y \in V,$$

they are said to be conjugated: $L_2 = L_1^*$. It is easy to see that $L_1^{**} = L_1$ and thus $L_2^* = L_1$. If $L = L^*$ then the operator L is said to be Hermitian.

If $(e,e) \neq 0$, $Le = ae$, $a \in F_{p^2}$, and $L^* = L$, then it is obvious that $a \in F_p$. At the same time, if $e \neq 0$ but $(e,e) = 0$ then the element a may be imaginary, i.e., be of the form $a = bi$, $b \in F_p$ (see Sec. IV). Further, if

$$Le_1 = a_1e_1, \quad Le_2 = a_2e_2,$$

$$(e_1,e_1) \neq 0, \quad (e_2,e_2) \neq 0,$$

$$a_1 \neq a_2,$$

then as in the usual case, one has that $(e_1,e_2) = 0$. At the same time, the situation when

$$(e_1,e_1) = (e_2,e_2) = 0, \quad (e_1,e_2) \neq 0,$$

$$a_1 = b_1i, \quad a_2 = b_2i \quad (b_1,b_2 \in F_p)$$

is also possible (see Sec. IV). Thus in the modular case the eigenvalues of a Hermitian operator are not necessarily real and the eigenvectors corresponding to the different eigenvalues are not necessarily orthogonal.

We now discuss the question about the conditions under which the descriptions of state vectors on the language of Hilbert spaces and on the language of linear spaces over F_{p^2} may lead to the close physical results. In the cases considered below, all spaces V over F_{p^2} are finite-dimensional and separable Hilbert spaces H are infinite-dimensional. Let $\{\tilde{e}_1,\tilde{e}_2,...\}$ be a basis in H and $\{e_1,e_2,...,e_N\}$ be a basis in V, where $N = N(p)$ and $N(p) \to \infty$ if $p \to \infty$. If one proceeds from the approach of conventional quantum theory, then for the description of an experiment it is sufficient to deal with a set dense in H. As such a set one can take the set of vectors of the form $\tilde{c}_1\tilde{e}_1 + \cdots + \tilde{c}_{\tilde{N}}\tilde{e}_{\tilde{N}}$ where $\tilde{N}$ is individual for every vector, $\tilde{c}_j = \tilde{a}_j + \tilde{b}_ji$, and $\tilde{a}_j,\tilde{b}_j \in Q$. Take into account now that the Hilbert spaces in quantum theory are projective ones (this is because relative but not absolute probabilities have the physical meaning). Therefore one can assume that $\tilde{a}_j,\tilde{b}_j \in Z$ (compare with that noted in the Introduction). Suppose further that there exists such natural $N_1 = N_1(p)$ that

$$N_1(p) \leqslant N(p), \quad N_1(p) \to \infty$$

if $p \to \infty$ and

$$(e_j,e_k) \in U, \quad f((e_j,e_k)) = (\tilde{e}_j,\tilde{e}_k)$$

for $j,k \leqslant N_1$ (we use the same notation for the scalar products in H and V). If one deals only with the vectors from H satisfying the conditions $\tilde{c}_j = 0$ for $j > N_1$, $|a_j|, |b_j| < C(p)$ for $j \leqslant N_1$, and only with the vectors from V of the form $c_1e_1 + c_{N_1}e_{N_1}$ where

$$c_j = a_j + b_ji, \quad a_j,b_j \in U_0,$$

then it is clear that there exists $1 \leftrightarrow 1$ correspondence between such vectors, if p is so large that for the description of experiments at the existing energies it is sufficient to confine ourselves by the elements of the basis in H with the numbers $\leqslant N_1$ and by the vectors with $\tilde{a}_j,\tilde{b}_j \in f(U_0)$ for all $j \leqslant N_1$, then it is clear that there is no difference between the description of physical systems on the language of projective Hilbert spaces and the projective spaces over F_{p^2}.

In quantum theory, one usually considers the Lie algebras or superalgebras over the field of real numbers R and their representations in some Hilbert space H. Hence it is natural to consider the Lie algebras or superalgebras over F_p and their representations in spaces V over F_{p^2} as modular analog of the above representations. Let $\tilde{A}$ be a finite-dimensional Lie algebra or superalgebra over R with a basis $\{\tilde{h}_1,...,\tilde{h}_n\}$ and structure constants $\tilde{c}^{\gamma}_{\alpha\beta}$. A basis is supposed to be chosen in such a way that $\tilde{c}^{\gamma}_{\alpha\beta} \in Z$ for all $1 \leqslant \alpha,\beta,\gamma \leqslant n$ (the Chevalley basis[12]). If p is such that all $\tilde{c}^{\gamma}_{\alpha\beta}$ belong to $f(U_0)$, then the Lie algebra or superalgebra A over F_p with a basis $\{h_1,h_2,...,h_n\}$ and structure constants $c^{\gamma}_{\alpha\beta}$ such that $f(c^{\gamma}_{\alpha\beta}) = \tilde{c}^{\gamma}_{\alpha\beta}$ is the natural modular analog of $\tilde{A}$. Let $\tilde{B}$ and B be the enveloping algebras for $\tilde{A}$ and A, respectively, and let $\tilde{T}(\tilde{B})$ and $T(B)$ be the representations of $\tilde{B}$ and B in H

and V, respectively. We wish to introduce the notion of correspondence between these representations.

We will deal only with the elements $\tilde{u}$ from $\tilde{B}$ that are sums or differences of the elements $\tilde{d}\tilde{h}_{j_1}^{n_{j_1}}\cdots\tilde{h}_{j_l}^{n_{j_l}}$, where $j_1,...,j_l$ are some l numbers from the set $\{1,2,...,n\}$, n_{j_l} are natural numbers, $\tilde{d}\in Z$, and $|d|<C(p)$. Analogously, in the B algebra we will deal only with the elements u that are sums or differences of the elements $dh_{j_1}^{n_{j_1}}\cdots h_{j_l}^{n_{j_l}}$ with $d\in U_0$. It is evident that there exists the natural $1\leftrightarrow 1$ correspondence between elements $\tilde{u}$ and u. Define the notion of the order of element u as follows. The order of element $dh_{j_1}^{n_{j_1}}\cdots h_{j_l}^{n_{j_l}}$ is $n_{j_1}+\cdots+n_{j_l}$ if $d=1$ and $n_{j_1}+\cdots+n_{j_l}+1$ if $d\neq 1$ and the order of the sum or difference of such elements is equal to the sum of the orders. The order of the element $\tilde{u}$ from $\tilde{B}$ is defined analogously. It is clear that the order of the element depends on the form in which it is written. Therefore define the minimal order of the element as the minimal value of all its orders. Now the notion of the correspondence between the representations $\tilde{T}(\tilde{B})$ and $T(B)$ can be defined as follows.

Definition 1: Let $N_1=N_1(p)$ and $N_2=N_2(p)$ be such natural numbers that

$$(e_j,e_k)\in U,\quad f((e_j,e_k))=(\tilde{e}_j,\tilde{e}_k)$$

for all $j,k\leqslant N_1$, and for all elements u from B and $\tilde{u}$ from $\tilde{B}$ with the minimal orders $\leqslant N_2$ the equality

$$f((e_j,T(u)e_k))=(\tilde{e}_j,\tilde{T}(\tilde{u})\tilde{e}_k)$$

holds if u and $\tilde{u}$ are in correspondence with each other. If $N_1(p)\to\infty$, $N_2(p)\to\infty$ when $p\to\infty$, then the representations $T(B)$ and $\tilde{T}(\tilde{B})$ are said to be in correspondence with each other.

We treat Definition 1 as the formalization of correspondence between state spaces of physical systems and operators of physical quantities for conventional and modular cases. The described correspondence shows in what sense the conventional physics can be treated as the limit of finite physics if $p\to\infty$.

The notion of correspondence between ordinary and modular representations in the sense of Definition 1 can be used also in a case when H is the Hilbert space with indefinite metric. When H is finite-dimensional, the notion of correspondence can be simplified, assuming that the dimensions of H and V are equal and the equalities

$$f((e_j,e_k))=(\tilde{e}_j,\tilde{e}_k)\,,$$
$$f((e_j,T(u)e_k))=(\tilde{e}_j,\tilde{T}(\tilde{u})\tilde{e}_k)$$

hold for all j,k [i.e., there is no need to introduce the number $N_1=N_1(p)$]. Finally, if the Lie algebra or superalgebra A is infinite-dimensional, then the notion of correspondence can be generalized as follows. Let $\{\tilde{h}_1,\tilde{h}_2,...\}$ be a basis in $\tilde{A}$ and A be the Lie algebra or superalgebra over F_p with a basis $\{h_1,h_2,...,h_{N'}\}$, where $N'=N'(p)$ and $N'(p)\to\infty$ if $p\to\infty$. Let further $N_3=N_3(p)$ be a natural number such that $N_3(p)\leqslant N'(p)$, $N_3(p)\to\infty$ if $p\to\infty$ and if $1\leqslant\alpha,\beta,\gamma\leqslant N_3$ then $c_{\alpha\beta}^{\gamma}\in U_0$ and $f(c_{\alpha\beta}^{\gamma})=\tilde{c}_{\alpha\beta}^{\gamma}$. In this case A is treated as the modular analog of $\tilde{A}$. Let the elements u and $\tilde{u}$ of the enveloping algebras B and $\tilde{B}$ be constructed as above, using only the elements $h_j,\tilde{h}_j$ with $j\leqslant N_3(p)$. Then the correspondence can be defined by the repetition of Definition 1.

III. MODULAR ANALOGS OF REPRESENTATIONS OF ALGEBRAS SU(2) AND SP(2)

In quantum physics the angular momentum operator $\mathbf{L}=\{L_1,L_2,L_3\}$ can have only the values multiple to $\hbar/2$. Therefore if $\hbar/2$ (and not $\hbar$ as usual) is taken as the unit of measurement of angular momentum, then the commutation relations for the angular momentum operators can be written as

$$[L_3,L_+]=2L_+\,,\quad [L_3,L_-]=-2L_-\,,\quad [L_+,L_-]=L_3\,, \tag{2}$$

where

$$L_1=L_++L_-\,,\quad L_2=-i(L_+-L_-)\,.$$

Furthermore, if $L_3=L_3^*$, $L_-^*=L_+$, then (2) is the realization of the su(2) algebra by the Hermitian operators. Assume now that these relations are realized in a space of modular representation.

Algebra (2) possesses the Casimir operator of the second order,

$$K=L_3^2-2L_3+4L_+L_-=L_3^2+2L_3+4L_-L_+\,. \tag{3}$$

It is known that in the ordinary case all other Casimir operators are the functions of K, but in the modular case this is not so (see, e.g., Ref. 53).

Consider only the representations containing vector e_0 such that

$$L_+e_0=0\,,\quad L_3e_0=se_0\,,\quad (e_0,e_0)\neq 0\,.$$

Then (see Sec. II) $s\in F_p$. Denote

$$e_n=(L_-)^ne_0\,,\quad n=0,1,2,...\,.$$

Then it follows from (2) and (3) that

$$L_3e_n=(s-2n)e_n\,,\quad Ke_n=s(s+2)e_n\quad \text{for all } n.$$

Hence it follows from (3) that

$$L_+L_-e_n=(n+1)(s-n)e_n\,, \tag{4}$$

and since $L_+^*=L_-$ it follows from (4) that

$$(e_{n+1},e_{n+1})=(n+1)(s-n)(e_n,e_n)\,. \tag{5}$$

It is clear from (5) that $(e_n,e_n)\neq 0$ for $n=0,1,...,s$ if $s=0,1,...,p-1$ and as it is easy to see, the vectors e_n are mutually orthogonal.

As in the ordinary case, the considered representation is irreducible if $L_-e_s=0$. Therefore, as in the ordinary case, the dimension of the irreducible representation (IR) is equal to $s+1$. However, in contrast to the ordinary case, the dimension of IR cannot be more than p. There exists the general statement[13] that the maximal dimension of IR of classical Lie algebra of dimension n and rank r is equal to $p^{(n-r)/2}$.

To ensure the correspondence with ordinary representations of the su(2) algebra (see Sec. II), it is necessary to require $f(s)$, $f((e_0,e_0))>0$. If for example $(e_0,e_0)=1$, then the correspondence with the conventional case in the sense of Definition 1 can be surely guaranteed if

$$s<(\ln p)^{1/4}/2,\quad N_2(p)=[\ln p/\ln\ln p]\,.$$

It is easy to see also that if s is close to p, then the considered representations are the modular analogs of representations of the su(2) algebra with negative spin (infinite-dimensional representations in the Hilbert space with indefinite metric).

If, instead of (2), one deals with a more usual algebra in which the right-hand sides in (2) are divided by 2, then the correspondence with the ordinary case in the sense of Definition 1 cannot be achieved. In particular, in the representation with dimension 2, the vectors e_0 and e_1 are the eigenvectors of the L_3 operator with large eigenvalues $(p+1)/2$ and $(p-1)/2$. This is due to the fact that 1/2 in the usual case does not coincide with 1/2 in F_p.

We consider now the modular analogs of representations of the sp(2) algebra. Let a', a'', h be such operators that

$$[h,a'] = -2a', \quad [h,a''] = 2a'', \quad [a',a''] = h, \tag{6}$$

and $h^* = h$, $(a')^* = a''$. The Casimir operator of the second order now has the form

$$K = h^2 - 2h - 4a''a' = h^2 + 2h - 4a'a'' . \tag{7}$$

We will consider representations with the vector e_0, such that

$$a'e_0 = 0, \quad he_0 = q_0 e_0, \quad (e_0,e_0) \neq 0.$$

Then (see Sec. II) $q_0 \in F_p$ and thus $q_0 = 0,1,\ldots,p-1$. Denote $e_n = (a'')^n e_0$. Then it follows from (6) and (7) that

$$he_n = (q_0 + 2n)e_n, \quad Ke_n = q_0(q_0 - 2)e_n$$

for all n, and instead of (4), (5) we have

$$a'a''e_n = (n+1)(q_0+n)e_n , \tag{8}$$

$$(e_{n+1},e_{n+1}) = (n+1)(q_0+n)(e_n,e_n) . \tag{9}$$

The set $\{e_0,e_1,\ldots,e_N\}$ will be a basis of IR if $a''e_j \neq 0$ for $j < N$ and $a''e_N = 0$. This condition must be compatible with $a'a''e_N = 0$. Therefore, as it follows from (8), N is defined by the condition $q_0 + N = 0$ in F_p if $q_0 \neq 0$, and the case $q_0 = 0$ corresponds to zero representation.

We see that if $q_0 > 0$, then in the modular case $N = p - q_0$ and the dimension of IR is equal to $p - q_0 + 1$, in contrast to the ordinary case where IR is realized in the Hilbert space and $n = 0,1,\ldots,\infty$. It follows from (9) that $(e_n,e_n) \neq 0$ for all n and it is easy to see from (8) that the vectors e_n are mutually orthogonal. The correspondence with the representations in Hilbert space in the sense of Definition 1 can be surely guaranteed if $(e_0,e_0) = 1$, $q_0 < (\ln p)^{1/4}/2$, since, as it is easy to see, for the roles of $N_1(p)$ and $N_2(p)$ one can choose the values $[(\ln p)^{1/4}/2]$ and $[(\ln p)^{1/2}/2]$, respectively.

In the ordinary case, if q_0 is a positive integer, the finite-dimensional IR of sp(2) algebra can be obtained if one requires that $a''e_0 = 0$ instead of $a'e_0 = 0$. This IR is realized in the space with indefinite metric and has the dimension $q_0 + 1$. An analogous result exists of course in the modular case, but here IR with the maximal weight has simultaneously the minimal weight (and vice versa) for all $q_0 \in F_p$.

IV. MODULAR ANALOGS OF REPRESENTATIONS OF THE SP(4) ALGEBRA

In quantum theory, the Hilbert space of states of a physical system is a tensor product or a direct sum of tensor products of Hilbert spaces, which are the state spaces for the fundamental objects—elementary particles or strings. Therefore in Secs. IV–VI we consider the modular analogs of IR's of so(2,3) and so(1,4) algebras describing elementary particles in the de Sitter spaces. It is well known that so(2,3) and sp(4) algebras are isomorphic. The representation of the latter algebra can be realized by means of the operators $\{h_j,a_j',a_j'',b'b'',L_-,L_+\}$ $(j = 1,2)$ with the commutation relations

$$\begin{aligned}
&[a_1',b'] = [a_2',b'] = [a_1'',b''] = [a_2'',b''] \\
&\quad = [a_1',L_-] = [a_1'',L_+] \\
&\quad = [a_2',L_+] = [a_2'',L_-] = 0 , \\
&[h_j,b'] = -b', \quad [h_j,b''] = b'', \quad [h_1,L_\pm] = \pm L_\pm , \\
&[h_2,L_\pm] = \mp L_\pm , \quad [b',b''] = h_1 + h_2 , \\
&[L_+,L_-] = h_1 - h_2 , \\
&[a_1',b''] = [b',a_2''] = L_-, \quad [a_2',b''] = [b',a_1''] = L_+ , \\
&[a_1',L_+] = [a_2',L_-] = b', \\
&[a_2'',L_+] = [a_1'',L_-] = -b'' , \\
&[b',L_-] = 2a_1', \quad [b',L_+] = 2a_2', \\
&[b'',L_-] = -2a_2'', \quad [b'',L_+] = -2a_1'' ,
\end{aligned} \tag{10}$$

where it is assumed additionally that the sets (a_j',a_j'',h_j) $(j = 1,2)$ are independent algebras (6) and $L_-^* = L_+$, $(b')^* = b''$.

If we denote $L_3 = h_1 - h_2$ then the set (L_-,L_+,L_3) realizes the representation of the su(2) algebra and if we introduce the operators

$$\begin{aligned}
&L_{12} = L_3, \quad L_{23} = L_+ + L_-, \quad L_{31} = -i(L_+ - L_-) , \\
&L_{05} = h_1 + h_2, \quad L_{35} = b' + b'', \quad L_{30} = -i(b'' - b') , \\
&L_{10} = i(a_1'' - a_1' + a_2'' - a_2') , \\
&L_{15} = a_2'' + a_2' - a_1'' - a_1' , \\
&L_{20} = a_1'' + a_2'' + a_1' + a_2' , \\
&L_{25} = i(a_1'' + a_2'' - a_1' - a_2') ,
\end{aligned} \tag{11}$$

and denote $L_{\mu\nu} = -L_{\nu\mu}$ $(\mu,\nu = 0,1,2,3,5)$, then the operators $L_{\mu\nu}$, as it follows from (10), satisfy the relations

$$\begin{aligned}
&[L_{\mu\nu},L_{\rho\sigma}] \\
&\quad = -2i(g_{\mu\rho}L_{\nu\sigma} + g_{\nu\sigma}L_{\mu\rho} - g_{\mu\sigma}L_{\nu\rho} - g_{\nu\rho}L_{\mu\sigma}) ,
\end{aligned} \tag{12}$$

where $g_{\mu\nu}$ is the diagonal tensor with the components

$$g_{00} = -g_{11} = -g_{22} = -g_{33} = g_{55} = 1 .$$

Therefore the $L_{\mu\nu}$ operators realize the representation of the so(2,3) algebra by the Hermitian operators and the multiplier 2 on the right-hand part of (12) is present due to the choice of units of measurement, as it has been pointed out in Sec. III. The fact that the formulas (10)–(12) do not contain the division is crucial for the establishing of correspondence with ordinary representations of the sp(4) and so(2,3) algebras.

The modular representations of algebra (10) have been investigated for the first time by Braden,[11] however his results do not cover the case of modular analogs of representa-

tions in Hilbert space, describing elementary particles (see the end of Sec. V). Nevertheless, in analogy with Braden's work[11] and with papers[54,55] in which the ordinary representations of the so(2,3) algebra in Hilbert space have been constructed, we will use the basis in which h_j, K_j $(j=1,2)$ operators are diagonal and K_j is the operator (7) for algebra (a_j', a_j'', h_j). In analogy with Braden's work[11] we introduce further the operators,

$$\begin{aligned}
\tilde{A}^{++} &= b''(h_1-1)(h_2-1) - a_1'' L_-(h_2-1) \\
&\quad - a_2'' L_+(h_1-1) + a_1'' a_2'' b', \\
\tilde{A}^{+-} &= L_+(h_1-1) - a_1'' b', \\
\tilde{A}^{-+} &= L_-(h_2-1) - a_2'' b', \\
\tilde{A}^{--} &= b',
\end{aligned} \tag{13}$$

and consider their action only on the space of "minimal" $\mathrm{sp}(2)\times\mathrm{sp}(2)$ vectors, i.e., such vectors x, so that $a_j' x = 0$ for $j=1,2$ (Braden[11] works in the space of "maximal" vectors such that $a_j'' x = 0$ for $j=1,2$, while the approach of Refs. 54, 55 is not quite convenient for us because it uses irrational numbers explicitly.)

It is easy to see that if x is a minimal vector such that $h_j x = \tilde{q}_j x$, then $\tilde{A}^{++}x$ is the minimal eigenvector of the operators h_j with the values $\tilde{q}_j + 1$ $(j=1,2)$, $\tilde{A}^{+-}x$ with the values $\tilde{q}_1 + 1, \tilde{q}_2 - 1$, $\tilde{A}^{-+}x$ with the values $\tilde{q}_1 - 1, \tilde{q}_2 + 1$, and $\tilde{A}^{--}x$ with the values $\tilde{q}_j - 1$.

The algebra sp(4) possesses the Casimir operators of the second and the fourth orders I_2 and I_4 and in the modular case there exist additional independent invariants. We need only the action of I_2 on the space of minimal vectors,

$$I_2 = (h_1^2 + h_2^2) - 2(h_1 - 2h_2) + 2(L_-L_+ - b''b'). \tag{14}$$

It follows from (10) and (14) that

$$\begin{aligned}
&[\tilde{A}^{--}, \tilde{A}^{+-}] \\
&\quad = [\tilde{A}^{--}, \tilde{A}^{-+}] = [\tilde{A}^{++}, \tilde{A}^{+-}] \\
&\quad = [\tilde{A}^{++}, \tilde{A}^{-+}] = 0, \\
&[\tilde{A}^{--}, \tilde{A}^{++}] \\
&\quad = (h_1 + h_2 - 2)[\tfrac{1}{2}(h_1+h_2)^2 - 2(h_1+h_2) - \tfrac{1}{2} I_2], \\
&[\tilde{A}^{+-}, \tilde{A}^{-+}] \\
&\quad = (h_1 - h_2)[2 + \tfrac{1}{2} I_2 - \tfrac{1}{2}(h_1 - h_2)^2].
\end{aligned} \tag{15}$$

In analogy with the construction of ordinary representations with positive energy[54,55] we require the existence of the vector e_0 satisfying the conditions

$$\begin{aligned}
&(e_0, e_0) \neq 0, \\
&a_j' e_0 = b' e_0 = L_+ e_0 = 0, \\
&h_j e_0 = q_j e_0 \quad (j=1,2).
\end{aligned}$$

Then (see Sec. II) $q_j \in F_p$. As it follows from (14) the invariant I_2 assumes in the representation space the numerical value

$$I_2 = q_1^2 + q_2^2 - 2(q_1 + 2q_2), \tag{16}$$

and as it follows from (15), the basis in the space of minimal vectors can be chosen in the form $\tilde{e}_{nk} = (\tilde{A}^{++})^n (\tilde{A}^{-+})^k e_0$ $(n,k = 0,1,\ldots)$. Then it follows from (15) and (16) that

$$\begin{aligned}
&\tilde{A}^{--}\tilde{A}^{++}\tilde{e}_{nk} \\
&\quad = (n+1)(q_1+q_2+n-2)(q_1+n) \\
&\qquad \times (q_2+n-1)\tilde{e}_{nk}, \qquad (17) \\
&(\tilde{e}_{n+1,k}, \tilde{e}_{n+1,k}) \\
&\quad = (q_1+n-k-1)(q_2+n+k-1)(q_1+n) \\
&\qquad \times (q_2+n-1)(n+1)(q_1+q_2+n-2)(\tilde{e}_{nk}, \tilde{e}_{nk}), \qquad (18) \\
&\tilde{A}^{+-}\tilde{A}^{-+}\tilde{e}_{nk} \\
&\quad = (k+1)(q_1-q_2-k)(q_1-k-2) \\
&\qquad \times (q_2+k-1)\tilde{e}_{nk}, \qquad (19) \\
&(\tilde{e}_{n,k+1}, \tilde{e}_{n,k+1}) \\
&\quad = \frac{q_2+n+k-1}{q_1+n-k-2}(q_1-k-2)(q_2+k-1) \\
&\qquad \times (k+1)(q_1-q_2-k)(\tilde{e}_{nk}, \tilde{e}_{nk}). \qquad (20)
\end{aligned}$$

The full basis of the representation space can be chosen in the form $(a_1'')^{n_1}(a_2'')^{n_2}\tilde{e}_{nk}$ where, as it follows from the results of Sec. III,

$$\begin{aligned}
&n_1 = 0,1,\ldots,N_1(n,k), \\
&n_2 = 0,1,2,\ldots,N_2(n,k), \\
&N_1(n,k) = p - q_1 - n + k, \\
&N_2(n,k) = p - q_2 - n - k.
\end{aligned}$$

As it follows from (11), in the ordinary case, IR with given q_1, q_2 describes the particle with the spin s and the de Sitter mass $m = q_1 + q_2$. It is known that for IR's with positive energy, the following classification exists: IR's with $m - s > 2$ (massive particles[54,55]), IR's with $m - s = 2$ (massless particles[54–56]), and two Dirac singletons[57] with $m = 1, s = 0$ and $m = 2, s = 1$.

The modular analog of singletons can be investigated very simply. Indeed, it follows from (17)–(20) that in the modular, as well as in the ordinary case, n and k assume the values

$$n = 0,1, \quad k = 0 \quad \text{for} \quad m = 1, \quad s = 0$$

and

$$n = 0, \quad k = 0,1 \quad \text{for} \quad m = 2, \quad s = 1.$$

Thence it follows from the results of Sec. III that in the modular case the space of IR has the dimension

$$D = (p^2+1)/2 \quad \text{if} \quad m = 1, \quad s = 0$$

and

$$D = (p^2-1)/2 \quad \text{if} \quad m = 2, \quad s = 1.$$

We shall not attempt to classify all modular IR's of algebra (10) and confine ourselves to representations which may correspond to real particles. Therefore, considering modular analogs of massive and massless cases we suppose that if m and s are represented as $0,1,\ldots,p-1$, then $m + s < p$ what surely takes place for real particles.

We consider first the modular analog of the massive case, when $m - s = 3,4,\ldots$. Then it follows from (19) and (20) that k can assume only the values $0,1,\ldots,s$, as well as in

the ordinary case. At the same time it follows from (17), (18) that in contrast to the ordinary case (where $n=0,1,...,\infty$), in the modular one $n=0,1,...,N$ where

$$N=\begin{cases}p+2-m, & \text{for even } m-s,\\ (p-m-s)/2, & \text{for odd } m-s.\end{cases} \tag{21}$$

Hence the space of minimal vectors has the dimension $(s+1)(N+1)$ and IR turns out to be finite-dimensional (and even finite since the field F_{p^2} is finite). If $D(m,s)$ is the dimension of IR with the given m and s, then it can be easily shown by the direct calculation that

$$\begin{aligned}&D(m,s)\\ &\quad=\tfrac{1}{3}(s+1)(p+3-m)[p^2-\tfrac{1}{2}p(m-s)\\ &\qquad+\tfrac{1}{4}(m-3)^2-\tfrac{1}{4}(s+1)^2],\quad \text{for even } m-s,\\ &D(m,s)\\ &\quad=((p+2-m-s)/24)(s+1)[(p-m)^2\\ &\qquad+(p-m)(s+1)-s],\quad \text{for odd } m-s.\end{aligned} \tag{22}$$

The actions here must be understood of course in the usual sense but not in the sense of F_p.

The matrix elements of every representation operator can be determined in massive and singleton cases by means of (10), (15), (17), (19), and the definition of e_0, but in the massless case this is not so, since there exist minimal vectors with the eigenvalues of operators h_1 and h_2 equal to unity: if $s\neq 0$ then

$$(h_2-1)e_0=0,\quad (h_1-1)e_0\neq 0,$$
$$(h_2-1)\tilde{e}_{0s}\neq 0,\quad (h_1-1)\tilde{e}_{0s}=0$$

and if $s=0$ then

$$(h_1-1)e_0=(h_2-1)e_0=0.$$

Therefore instead of (13) we introduce the operators

$$\begin{aligned}A^{++}&=\tilde{A}^{++}\frac{1}{(h_1-1)(h_2-1)}\\ &=b''-a_1''L_-\frac{1}{h_1-1}-a_2''L_+\frac{1}{h_2-1}\\ &\quad+a_1''a_2''b'\frac{1}{(h_1-1)(h_2-1)},\\ A^{+-}&=\tilde{A}^{+-}\frac{1}{h_1-1}=L_+-a_1''b'\frac{1}{h_1-1},\\ A^{-+}&=\tilde{A}^{-+}\frac{1}{h_2-1}=L_--a_2''b'\frac{1}{h_2-1},\\ A^{--}&=\tilde{A}^{--}=b',\end{aligned} \tag{23}$$

and assume additionally that if $m-s=2$, $s\neq 0$, then

$$\begin{aligned}A^{++}e_0&=(b''-a_1''L_-(h_1-1)^{-1})e_0,\\ A^{-+}e_0&=L_-e_0,\\ A^{++}e_{0s}&=(b''-a_2''L_+(h_2-1)^{-1})e_{0s},\\ A^{+-}e_{0s}&=L_+e_{0s},\end{aligned} \tag{24}$$

and if $m-s=2$, $s=0$ then

$$A^{+-}e_0=A^{-+}e_0=0,\quad A^{++}e_0=b''e_0,$$

where

$$e_{nk}=(A^{++})^n(A^{-+})^ke_0.$$

So defined A operators as well as $\tilde{A}$ operators transform the minimal vectors into minimal ones and instead of (17)–(20) we have

$$\begin{aligned}&A^{--}A^{++}e_{nk}=a_{nk}e_{nk},A^{+-}A^{-+}e_{nk}\\ &\qquad=b_{nk}e_{nk},\\ &(e_{n+1,k}e_{n+1,k})=a_{nk}(e_{nk},e_{nk}),\\ &(e_{n,k+1},e_{n,k+1})=b_{nk}(e_{nk},e_{nk}),\\ &a_{nk}=\frac{(n+1)(q_1+q_2+n-2)(q_1+n)(q_2+n-1)}{(q_1+n-k-1)(q_2+n+k-1)},\\ &b_{nk}=\frac{(k+1)(q_1-q_2-k)(q_1-k-2)(q_2+k-1)}{(q_1+n-k-2)(q_2+n+k-1)},\end{aligned} \tag{25}$$

and as it follows from (24), in the massless case,

$$\begin{aligned}&b_{0k}=(k+1)(s-k),\quad a_{0k}=0(k\neq 0),\\ &a_{00}=q_1 \text{ if } s\neq 0,\quad a_{00}=2 \text{ if } s=0.\end{aligned} \tag{26}$$

As it follows from (25) and (26), the numbers (n,k) in the massless case assume only the values $(0,k)$ for $k=0,1,...,s$ and $(n,0)$ for $n=1,2,...,p-1-s$. Therefore it is easy to calculate that in the massless case,

$$\begin{aligned}&D(m,s)\\ &\quad=\frac{p^3}{3}+\frac{p^2(s+1)}{2}-p\left(s^2+\frac{s}{2}-\frac{1}{6}\right)+\frac{s(s^2-1)}{6}.\end{aligned} \tag{27}$$

Thus we have constructed modular analogs of all IR's of the sp(4) algebra with positive energy. The proof of irreducibility in the modular case is rather simple. Indeed, vector e_0 is a cyclic one by construction. Let x be an arbitrary vector from V and

$$\begin{aligned}&x=\sum_{n_1,n_2,n,k}c(n_1,n_2,n,k)(a_1'')^{n_1}(a_2'')^{n_2}e_{nk},\\ &c(n_1,n_2,n,k)\neq 0.\end{aligned}$$

Let $\tilde{N}_1$ and $\tilde{N}_2$ be the maximal values of n_1 and n_2 in this decomposition. Then $y=(a_1')^{\tilde{N}_1}(a_2')^{\tilde{N}_2}x$ is a linear combination of nonzero minimal vectors. Acting on y by the powers of operators A^{--} and A^{+-} we can obviously obtain a nonzero vector proportional to e_0. Therefore every vector from V is a cyclic one and the representation is irreducible.

It is easy to see that different elements of the constructed basis are mutually orthogonal; they are such that $(e,e)\neq 0$ and all eigenvalues of the operators h_j, K_j $(j=1,2)$ belong to F_p. Therefore as it follows from (11), the eigenvalues of the z component of the orbital angular momentum L_{12} and the de Sitter analog of the energy L_{05} also belong to F_p. However, generally speaking, this is not the case for other Hermitian operators. Consider for simplicity a massive case when $s=0$, m is even, and we take the operator L_{35} which is the de Sitter analog of the z component of the momentum operator. Consider the subspace generated by the vectors

$$\begin{aligned}&X_1=(a_1'')^{p-m/2-1}e_0,\quad X_2=(a_1'')^{p-m/2-1}b''e_0,\\ &X_3=(2/m)(a_1'')^{p-m/2}a_2''e_0.\end{aligned} \tag{28}$$

It is easy to establish the following relations:

$$b''X_1 = X_2, \quad b''X_2 = 2X_3, \quad b''X_3 = 0,$$
$$b'X_1 = 0, \quad b'X_2 = -2X_1, \quad b'X_3 = -X_2. \tag{29}$$

Therefore the considered subspace is invariant relative to the action of $L_{35} = b' + b''$. Solving the characteristic equation for the L_{35} operator in this subspace, we can see that there exist three eigenvectors

$$\psi_1 = X_1 + X_3, \quad \psi_2 = X_1 - iX_2 - X_3, \quad \psi_3 = X_1 + iX_2 - X_3$$

such that

$$L_{35}\psi_1 = 0, \quad L_{35}\psi_2 = 2i\psi_2, \quad L_{35}\psi_3 = -2i\psi_3.$$

It is easy to calculate further that

$$(X_1,X_1) = (X_3,X_3) = c, \quad (X_2,X_2) = -2c$$

where

$$c = \frac{(p - m/2 - 1)!}{(m/2 - 1)!}(e_0,e_0) \quad \text{in} \quad F_p.$$

Thence from the orthogonality of vectors X_1,X_2, and X_3, it follows that

$$(\psi_1,\psi_2) = (\psi_1,\psi_3) = (\psi_2,\psi_2) = (\psi_3,\psi_3) = 0,$$
$$(\psi_1,\psi_1) = 2c, \quad (\psi_2,\psi_3) = 4c.$$

The above example corroborates what has been said in Sec. II about eigenvectors and eigenvalues of Hermitian operators in the modular case. The "anomalies" of such a type may occur obviously in the region far from the existing energies where the residue modulo p plays an essential role.

V. CORRESPONDENCE BETWEEN REPRESENTATIONS OF THE sp(4) ALGEBRA OVER THE FIELDS F_{p^2} and C

It is clear that the above correspondence in the sense of Definition 1 may be achieved only if m and s are much less than p (the estimate will be given below) and we assume that this is the case. Consider first the case of massive particles and even $m - s$. Then it is clear that q_1 and q_2 [more exactly $f(q_1)$ and $f(q_2)$] are much less than p. Therefore there exists the correspondence between operators h_1 and h_2 in the sense of Sec. II. Furthermore, since formulas (10), (13), (18) do not contain the division and in (20) division is performed as usual [since if $n = 0$ then the denominator is cancelled by the multiplier $(q_1 - k - 2)$ in the numerator and if $n \neq 0$ then

$$(\tilde{A}^{-+}\tilde{e}_{nk}, \tilde{A}^{-+}\tilde{e}_{nk})$$
$$= (\tilde{A}^{-+}\tilde{A}^{++}\tilde{e}_{n-1,k}, \tilde{A}^{-+}\tilde{A}^{++}\tilde{e}_{n-1,k})$$
$$= (\tilde{A}^{++}\tilde{A}^{-+}\tilde{e}_{n-1,k}, \tilde{A}^{++}\tilde{A}^{-+}\tilde{e}_{n-1,k})$$
$$= (\tilde{A}^{++}\tilde{e}_{n-1,k+1}, \tilde{A}^{++}\tilde{e}_{n-1,k+1})],$$

then there is correspondence between the representation spaces (in the sense of Sec. II) if a basis is constructed by means of $\tilde{A}$ operators. It is clear from the results of Sec. III that the correspondence between operators a_j', a_j'', h_j $(j = 1,2)$ also exists in the considered case. However, it is easy to see that the matrix elements of operators b', b'', L_-, L_+ in the basis $\{(a_1'')^{n_1}(a_2'')^{n_2}\tilde{e}_{nk}\}$ contain the division nontrivially and thus there is no correspondence in the sense of Sec. II. Therefore, as it follows from (11), the correspondence does exist between operators $L_{\mu\nu}$ if $\mu,\nu \neq 3$ and does not exist between operators $L_{3\mu}$.

Since the physical quantities corresponding to the operators which are nondiagonal in the chosen orthogonal basis are not observable, then the absence of the correspondence between nondiagonal operators is possibly not essential from the physical point of view. Nevertheless the following problem arises. Is it possible to choose a basis in such a way that the correspondence takes place in the sense of Definition 1?

Consider the set of vectors

$$|n_1,n_2,n,k\rangle = (a_1'')^{n_1}(a_2'')^{n_2}(b'')^n(L_-)^k e_0,$$

where n_1,n_2,n,k run over the same values as above. Our nearest aim is to show that these elements form the basis.

It follows from (15) and (23) that

$$A^{+-}(A^{++})^l = (A^{++})^l A^{+-} - \frac{h_2 - 2}{h_2 + 1 - 2},$$
$$A^{-+}(A^{++})^l = (A^{++})^l A^{-+} - \frac{h_1 - 2}{h_1 + l - 2}. \tag{30}$$

Thence and from (23), (25), it follows that

$$e_{n+1,k} = b''e_{nk} - \frac{q_1 - k - 2}{(q_1 + n - k - 2)(q_1 + n - k - 1)} a_1'' e_{n,k+1}$$
$$- \frac{(q_2 + k - 2)k(s + 1 - k)}{(q_2 + n + k - 2)(q_2 + n + k - 1)} a_2'' e_{n,k-1} - \frac{a_{n-1,k}}{(q_1 + n - k - 1)(q_2 + n + k - 1)} a_1'' a_2'' e_{n-1,k}. \tag{31}$$

It can be shown now that

$$e_{nk} = (b'')^n(L_-)^k e_0 + \sum_n' c_{i_1 i_2 il}|i_1 i_2 il\rangle, \tag{32}$$

where $c_{i_1 i_2 il} \in F_p$ are some coefficients and Σ_n' means the sum over such i_1, i_2, i, l that $i_1 + i_2 + i = n$, $i < n$. Indeed if $n = 0$ the formula (32) is valid since $e_{0k} = (L_-)^k e_0$ and if $n \neq 0$ the validity of (32) can be established by induction with the help of formula (31).

It follows from (32) that

$$|n_1,n_2,n,k\rangle = (a_1'')^{n_1}(a_2'')^{n_2}e_{nk}$$
$$- (a_1'')^{n_1}(a_2'')^{n_2}\sum_n' c_{i_1 i_2 il}|i_1 i_2 il\rangle. \tag{33}$$

Thence if $n = 0$ then the restrictions on n_1,n_2 are the same as above and if

$$n_1 = N_1(n,k) + 1$$

or

$$n_2 = N_2(n,k) + 1,$$

then the vector $|n_1,n_2,n,k\rangle$ can be expressed through vectors $|n_1+i_1,n_2+i_2,i,l\rangle$ with $i<n$, since for such n_1 or n_2,

$$(a_1'')^{n_1}(a_2'')^{n_2}e_{nk}=0.$$

Thence the restrictions on n_1 and n_2 remain valid if $0\leqslant n\leqslant N$, where N is the maximal value of n (in the massless case N depends on k if $s\neq 0$). Finally, it follows from the same formula (33) that vector $|n_1,n_2,N+1,k\rangle$ can be expressed through $|n_1+i_1,n_2+i_2,i,l\rangle$ with $i\leqslant N$ because $e_{N+1,k}=0$.

Since $D(m,s)$ elements $(a_1'')^{n_1}(a_2'')^{n_2}e_{nk}$ form a basis and can be expressed through $D(m,s)$ elements $|n_1,n_2,n,k\rangle$ according to (33), then, according to the Steinitz theorem (see, e.g., § 20 in Ref. 58), the elements $|n_1,n_2,n,k\rangle$ also form a basis.

The direct calculation by means of formulas (10) shows that if

$$n<N_1,\quad n_1<N_1(n,k),\quad n_2<N_2(n,k),$$

then

$$\begin{aligned}
L_+|n_1,n_2,n,k\rangle &= k(s+1-k)|n_1,n_2,n,k-1\rangle + n_2|n_1,n_2-1,n+1,k\rangle + 2n|n_1+1,n_2,n-1,k\rangle,\\
L_-|n_1,n_2,n,k\rangle &= |n_1,n_2,n,k+1\rangle + n_1|n_1-1,n_2,n+1,k\rangle + 2n|n_1,n_2+1,n-1,k\rangle,\\
a_1'|n_1,n_2,n,k\rangle &= n_1(q_1+n-k+n_1-1)|n_1-1,n_2,n,k\rangle + n(n-1)|n_1,n_2+1,n-2,k\rangle + n|n_1,n_2,n-1,k+1\rangle,\\
a_2'|n_1,n_2,n,k\rangle &= n_2(q_2+n+k+n_2-1)|n_1,n_2-1,n,k\rangle + nk(s+1-k)|n_1,n_2,n-1,k-1\rangle\\
&\quad + n(n-1)|n_1+1,n_2,n-2,k\rangle,\\
b'|n_1,n_2,n,k\rangle &= n_1n_2|n_1-1,n_2-1,n,k\rangle + n(q_1+q_2+n+2n_1+2n_2-1)|n_1,n_2,n-1,k\rangle\\
&\quad + n_1k(s+1-k)|n_1-1,n_2,n,k-1\rangle + n_2|n_1,n_2-1,n,k+1\rangle,\\
h_1|n_1,n_2,n,k\rangle &= (q_1+n-k+2n_1)|n_1,n_2,n,k\rangle,\\
h_2|n_1,n_2,n,k\rangle &= (q_2+n+k+2n_2)|n_1,n_2,n,k\rangle,\\
a_1''|n_1,n_2,n,k\rangle &= |n_1+1,n_2,n,k\rangle,\quad a_2''|n_1,n_2,n,k\rangle = |n_1,n_2+1,n,k\rangle,\quad b''|n_1,n_2,n,k\rangle = |n_1,n_2,n+1,k\rangle.
\end{aligned}\tag{34}$$

Let us agree that if one of the numbers n_1,n_2,n exceeds the maximal allowed one by 1 then $|n_1,n_2,n,k\rangle$ means the vector which is expressed through the basis elements by means of formula (33) with $(a_1'')^{n_1}(a_2'')^{n_2}e_{nk}=0$ and the coefficients $c_{i_1i_2il}$ which in principle can be found from (32). Under such an agreement, the formulas (34) define formally the matrix elements of representation operators in the basis $|n_1,n_2,n,k\rangle$ for all $n\leqslant N$,

$$n_1\leqslant N_1(n,k),\quad n_2\leqslant N_2(n,k).$$

These formulas make it possible to calculate matrix elements of representation operators of the enveloping algebra and scalar products of the basis elements $|n_1,n_2,n,k\rangle$.

In a case of representation in the Hilbert space, $|n_1,n_2,n,k\rangle$ can be also chosen as the basis elements but now, in the massive case, for example, $n_1,n_2,n=0,1,2,\ldots,\infty$. The matrix elements of the representation operators in this basis have been calculated in Ref. 59 and it is clear that the result can be formally represented in the form (34). Therefore it is easy to see that the correspondence in the sense of Definition 1 can be only in the massive case if $m-s$ is even and in the massless case if $m=2$, $s=0$ (the latter case sometimes is also classified as the massive one). If for example

$$(e_0,e_0)=1,\quad q_1,q_2,n_1,n_2,n<\tfrac{1}{2}(\ln p)^{1/10},$$
$$N_2(p)=[(\ln p)^{1/2}/4],$$

then the above mentioned correspondence is surely satisfied.

At the end of this section we briefly discuss the question about modular analogs of ordinary finite-dimensional representations of the sp(4) algebra. They can be obtained if one requires the existence of the vector $\tilde{e}_0$ (instead of e_0) such that

$$a_j''\tilde{e}_0=b''\tilde{e}_0=L_+\tilde{e}_0=0,\quad h_j\tilde{e}_0=\tilde{q}_j\tilde{e}_0,$$

$\tilde{q}_j$ are positive integers and $\tilde{q}_1\geqslant\tilde{q}_2$. Then it can be shown that the representation contains only the maximal $\mathrm{sp}(2)\times\mathrm{sp}(2)$ vectors with the eigenvalues of h_1 and h_2 operators equal to $\tilde{q}_1-n-k$ and $\tilde{q}_2-n+k$, respectively, where

$$k=0,1,\ldots,\quad \tilde{q}_1-\tilde{q}_2,$$
$$n=0,1,\ldots,\tilde{q}_2.$$

Thence proceeding from that mentioned at the end of Sec. III it is easy to calculate directly that, instead of (22), the dimension of IR is given by

$$D(\tilde{q}_1,\tilde{q}_2)=\tfrac{1}{6}(\tilde{q}_1-\tilde{q}_2+1)(\tilde{q}_2+1)(\tilde{q}_1+2)(\tilde{q}_1+\tilde{q}_2+3).\tag{22'}$$

This formula was obtained for the first time by Antoine and Speiser,[60] proceeding from the well-known Weyl formula for the dimensions of ordinary IR's. The analogous result has a place in the modular case as well. Namely, Braden[11] has investigated the case when

$$0\leqslant\tilde{q}_2,\tilde{q}_1-\tilde{q}_2\leqslant p-1,\quad \tilde{q}_1+\tilde{q}_2<p$$

and has shown that the formula (22') defines the dimension of the modular IR when $\tilde{q}_1+\tilde{q}_2<p-2$.

It is easy to see that in the modular case, the representation with

$$a_j''\tilde{e}_0 = b''\tilde{e}_0 = L_+\tilde{e}_0 = 0, \quad h_j\tilde{e}_0 = \tilde{q}_j\tilde{e}_0$$

is equivalent to the representation with

$$a_j'e_0 = b'e_0 = L_+e_0 = 0,$$

$$h_je_0 = q_je_0 \quad \text{if} \quad q_1 = p - \tilde{q}_2,\ q_2 = p - \tilde{q}_1.$$

Since the modular analog of representations describing elementary particles corresponds to the case when q_1 and q_2 are much less than p (see above), then Braden's analysis does not cover this case. In principle, the information about modular characters and dimensions of modular IR's for the wide class of Lie algebras and Chevalley groups (including the algebra of B_2 type) can be obtained proceeding from the general approach developed by Jantzen.[18] However, it is clear that in order to establish the correspondence in the sense of Definition 1, one needs to have explicit formulas for matrix elements in each specific case.

VI. MODULAR ANALOGS OF REPRESENTATIONS OF SO(1,4) ALGEBRA

A representation of the so(1,4) algebra can be defined by means of (12), but now $\mu,\nu,\rho,\sigma = 0,1,2,3,4$ and $g_{44} = -1$. Let $\mathbf{J}'$ and $\mathbf{J}''$ be vector operators forming two independent su(2) algebras (i.e., $[\mathbf{J}',\mathbf{J}''] = 0$) and let R_{ij} $(i,j = 1,2)$ be the operators which satisfy the commutation relations

$$\begin{aligned}
&[J_3',R_{1j}] = R_{1j}, \quad [J_3',R_{2j}] = -R_{2j},\\
&[J_3'',R_{i1}] = R_{i1}, \quad [J_3'',R_{i2}] = -R_{i2},\\
&[J_+',R_{1j}] = [J_+'',R_{i1}] = [J_-',R_{2j}] = [J_-'',R_{i2}] = 0,\\
&[J_+',R_{2j}] = R_{1j}, \quad [J_+'',R_{i2}] = R_{i1},\\
&[J_-',R_{1j}] = R_{2j}, \quad [J_-'',R_{i1}] = R_{i2},\\
&[R_{11},R_{12}] = 2J_+', \quad [R_{11},R_{21}] = 2J_+'',\\
&[R_{11},R_{22}] = -(J_3' + J_3''),\\
&[R_{12},R_{21}] = J_3' - J_3'', \quad [R_{12},R_{22}] = -2J_-'',\\
&[R_{21},R_{22}] = -2J_-',
\end{aligned} \tag{35}$$

and hermiticity relations $R_{11}^* = R_{22}$, $R_{12}^* = -R_{21}$. Then $L_{\mu\nu}$ operators are expressed through $\mathbf{J}',\mathbf{J}'',R_{ij}$ as follows:

$$\mathbf{L} = \mathbf{J}' + \mathbf{J}'', \quad \mathbf{B} = \mathbf{J}'' - \mathbf{J}',$$

$$L_{01} = i(R_{11} - R_{22}), \quad L_{02} = R_{11} + R_{22},$$

$$L_{03} = -i(R_{12} + R_{21}), \quad L_{04} = R_{12} - R_{21}, \tag{36}$$

where

$$\mathbf{L} = (L_{23},L_{31},L_{12}), \quad \mathbf{B} = (L_{14},L_{24},L_{34}).$$

For a construction of IR's we will use the method of $su(2)\times su(2)$ shift operators developed by Hughes[61] and first applied by him for the investigation of ordinary IR's of the group SO(5). We use the basis in which the operators J_3',J_3'',K',K'' are diagonal where K' and K'' are the Casimir operators (3) for the algebras $\mathbf{J}'$ and $\mathbf{J}''$, respectively.

Let x be a maximal $su(2)\times su(2)$ vector, i.e., $J_+'x = J_+''x = 0$. The following operators act on the set of maximal vectors invariantly:

$$\begin{aligned}
&\tilde{A}^{++} = R_{11}, \quad \tilde{A}^{+-} = R_{12}(J_3'' + 1) - J_-''R_{11},\\
&\tilde{A}^{-+} = R_{21}(J_3' + 1) - J_-'R_{11},\\
&\tilde{A}^{--} = -R_{22}(J_3' + 1)(J_3'' + 1) + J_-''R_{21}(J_3' + 1)\\
&\qquad + J_-'R_{12}(J_3'' + 1) - J_-'J_-''R_{11}.
\end{aligned} \tag{37}$$

The direct calculation by means of (35) shows that

$$\begin{aligned}
[\tilde{A}^{++},\tilde{A}^{+-}] &= [\tilde{A}^{++},\tilde{A}^{+-}] = [\tilde{A}^{--},\tilde{A}^{+-}]\\
&= [\tilde{A}^{--},\tilde{A}^{-+}] = 0.
\end{aligned} \tag{38}$$

We consider only the modular analog of massive representations, i.e., representations with the vector e_0 such that $(e_0,e_0) \neq 0$, $\mathbf{J}'e_0 = 0$. Since $\mathbf{J}''e_0 = \mathbf{L}e_0$, we require also that $L_+e_0 = 0$, $L_3e_0 = se_0$. These relations show that $\mathbf{J}'$ may be treated as the de Sitter analog of the conventional momentum and the subspace generated by the vectors $(L_-)^k e_0 (k = 0,1...,s)$, as the set of rest states (for a more detailed discussion see Ref. 62). It follows from (35) that $\tilde{A}^{--}e_0 = \tilde{A}^{-+}e_0 = 0$.

Let $I_2 = -\frac{1}{2}L_{\mu\nu}L^{\mu\nu}$ (the summation over μ,ν) be the Casimir operator of the second order. In a case of induced representations in the Hilbert space, $I_2 = m^2 + 9 - s(s+2)$, where m is the de Sitter mass. However, in the so(1,4) case, the mass cannot be defined as the lowest value of the energy, and the parameter m has no clear algebraic sense. Therefore in the modular case, we simply introduce the element $w \in F_p$ such that $I_2 = w + 9 - s(s+2)$. Then the calculation analogous to that carried out in Sec. IV gives

$$\begin{aligned}
&\tilde{A}^{--}\tilde{A}^{++}\tilde{e}_{nk} = -\frac{(n+1)(n+s+2)}{4}[w + (2n+s+3)^2]\tilde{e}_{nk},\\
&\tilde{A}^{-+}\tilde{A}^{+-}\tilde{e}_{nk} = -\frac{(k+1)(s-k)}{4}[w + 1 + (2k-s)(2k+2-s)]e_{nk},\\
&(\tilde{e}_{n+1,k},\tilde{e}_{n+1,k}) = \frac{(n+1)(n+s+2)[w + (2n+s+3)^2]}{4(n+k+2)(s-k+n+2)}(\tilde{e}_{nk},\tilde{e}_{nk}),\\
&(\tilde{e}_{n,k+1},\tilde{e}_{n,k+1}) = \frac{(k+1)(s-k)[w + 1 + (2k-s)(2k+2-s)](s-k+n+1)}{4(s-k+n+2)}(\tilde{e}_{nk},\tilde{e}_{nk}),
\end{aligned} \tag{39}$$

where

$$\tilde{e}_{nk} = (\tilde{A}^{++})^n(\tilde{A}^{+-})^k e_0.$$

Thence it follows that $k = 0,1,...,s$ as well in the ordinary case, but in the modular case $n = 0,1,...,N$, where N depends on whether the condition $w + (2n + s + 3)^2 = 0, \mod p$ can be satisfied for some n. If this condition cannot be satisfied (for example, if $w = m^2$, $m \in F_p$) then $N = p - s - 2$. However, it may be that there exists such N that

$$f(w) + (2n + s + 3)^2 < p \quad \text{for} \quad n = 0,1...,N-1,$$

and

$$f(w) + (2N + s + 3)^2 = p.$$

The orthogonal basis in the representation space can be chosen in the form $(J'_-)^{n'}(J''_-)^{n''}\tilde{e}_{nk}$, where

$$k = 0,1,...,s, \quad n = 0,1,...,N,$$

$$n' = 0,1,...,n + k, \quad n'' = 0,1,...,s - k + n.$$

The irreducibility can be proved in the analogy with Sec. IV and the dimension of IR is obviously equal to

$$D(w,s) = \sum_{K=0}^{S}\sum_{n=0}^{N}(k + n + 1)(s - k + n + 1)$$
$$= (N + 1)(s + 1)\left[\tfrac{1}{3}(N^2 + \tfrac{7}{2}N + 3)\right.$$
$$\left. + \tfrac{1}{2}s(N + \tfrac{1}{3}s + \tfrac{5}{6})\right]. \qquad (40)$$

It is clear from (39) that if $f(w)$ and $f(s)$ are positive and much less than p, and if one works with the operators

$$A^{++} = 4\tilde{A}^{++}(J'_3 + 2)(J''_3 + 2),$$
$$A^{+-} = 2\tilde{A}^{+-}(J''_3 + 2),$$

instead of $\tilde{A}^{++}, \tilde{A}^{+-}$, then the correspondence in the sense of Sec. II can be achieved between the representation spaces and the operators $\mathbf{J}', \mathbf{J}''$. However, in contrast to the so(2,3) case, we do not succeed in finding a basis in which the correspondence between representations takes place in the sense of Definition 1. In our opinion this is due to the following circumstance. In the so(2,3) case, particles are described by the IR of the discrete series. We require that e_0 are the eigenvectors of the operators of the Cartan subalgebra, and are killed by the operators which are negative relative to this subalgebra (such a representation is said to be a Verma module). In this case, the representation is fully defined and, in particular, the action of I_2 is also defined. In the so(1,4) case, however, particles are described by the IR of the principal series. From three conditions $\mathbf{J}'e_0 = 0$, only two of them are obviously independent and the extra independent condition is that e_0 is the eigenvector of I_2 (such a representation is not a Verma module). Since I_2 is quadratic in the representation operators of the so(1,4) algebra, then we cannot construct a basis analogous to that from Sec. V.

VII. MODULAR ANALOGS OF REPRESENTATIONS OF THE OSP(1,4) SUPERALGEBRA

The superalgebra osp(1,4) is a generalization of the sp(4) algebra. A representation of osp(1,4) can be realized by the operators d_j, d_j^* ($j = 1,2$) such that if α,β,γ are some of these operators, then

$$[\alpha,\{\beta,\gamma\}] = \langle\alpha,\beta\rangle\gamma + \langle\alpha,\gamma\rangle\beta, \qquad (41)$$

where the form $\langle\alpha,\beta\rangle$ is skew symmetric, $\langle d_1,d_1^*\rangle = \langle d_2,d_2^*\rangle = 1$, and other independent values are equal to zero. For the representation operators of the sp(4) algebra we have

$$h_j = \{d_j,d_j^*\}, \quad a'_j = d_j^2, \quad a''_j = d_j^{*2},$$
$$b' = \{d_1,d_2\}, \quad b'' = \{d_1^*,d_2^*\}, \qquad (42)$$
$$L_+ = \{d_1^*,d_2\}, \quad L_- = \{d_1,d_2^*\}.$$

We introduce the operators

$$A_j^- = d_j, \quad A_j^+ = d_j^* - a''_j d_j (h_j - 1)^{-1}. \qquad (43)$$

Consider the action of these operators only on the set of minimal sp(2)×sp(2) vectors (see Sec. IV). Then A_j^- decreases the eigenvalue of h_j by unity, A_j^+ increases this value by unity and the operators (43) transform the minimal vectors into minimal ones.

Introduce now the notations

$$A_{12}^{++} = A^{++}, \quad A_{12}^{+-} = A^{+-},$$
$$A_{12}^{-+} = A^{-+}, \quad A_{12}^{--} = A^{--}$$

and assume that $A_{ij}^{\epsilon\epsilon'} = A_{ji}^{\epsilon'\epsilon}$ where ϵ and ϵ' assume the values $+$ or $-$. Then the direct calculation by means of (41), (42), and (23) gives

$$(A_j^\epsilon)^2 = 0, \quad A_j^- A_j^+ = h_j - h_j(h_j - 1)^{-1}A_j^+ A_j^-,$$
$$A_{ij}^{\epsilon\epsilon'} = \{A_i^\epsilon, A_j^{\epsilon'}\},$$
$$A_i^- A_{ij}^{+\epsilon} = A_{ij}^{+\epsilon} A_i^- + A_j^\epsilon - (h_i - 1)^{-1}A_i^+ A_{ij}^{-\epsilon},$$
$$A_i^+ A_{ij}^{-\epsilon} = A_{ij}^{-\epsilon} A_i^+ - A_j^\epsilon + (h_i - 1)^{-1}A_{ij}^{+\epsilon} A_i^-,$$
$$[A_i^\epsilon, A_{ij}^{\epsilon\epsilon'}] = 0 \quad (i,j = 1,2, \ i \neq j) \qquad (44)$$

[these relations are valid also for osp(1,2n) if $i,j = 1,2,...,n$].

We consider only modular analogs of representations of the osp(1,4) superalgebra with positive energy. All such representations have been found in Refs. 63 and 64. In analogy with these papers we require the existence of the vector e_0 such that

$$d_j e_0 = L_+ e_0 = 0, \quad h_j e_0 = q_j e_0.$$

Then in particular $A_j^- e_0 = A_{21}^{-+} e_0 = 0$. Denote $m = q_1 + q_2$, $s = q_1 - q_2$.

According to Refs. 63–65 all representations with positive energy are classified as follows: the representation combining both Dirac singletons, representations combining two massless particles (m,s) and $(m + 1, s + 1)$ if $m - s = 2$, $s \neq 0$, and representations combining only massive particles. Note at once that the modular analog of the Dirac supermultiplet can be investigated without any problems. As in the ordinary case, IR under consideration has the property $[\alpha,\beta] = \frac{1}{2}\langle\alpha,\beta\rangle$ and the basis in the representation space can be chosen in the form $(d_1^*)^{n_1}(d_2^*)^{n_2}e_0$ but in the modular case, $n_1,n_2 = 0,1,...,p-1$ and the dimension of IR is equal to p^2. Note that (see Sec. IV) the number of bosonic states (with the even spin) is not equal to the number of fermionic ones (with the odd spin).

To investigate the general case, consider the vectors

$$e_0, e_1 = A_1^+ e_0, \quad e_2 = A_2^+ - (s+1)^{-1} A_{12}^{-+} A_1^+ e_0,$$
$$e_3 = (A_1^+ A_2^+ - [(q_2-1)/(q_1-1)] A_2^+ A_1^+) e_0. \tag{45}$$

It follows from the definition of e_0 and (44) that these vectors are killed by the operators A_{21}^{--}, A_{12}^{--}, and furthermore the following relations take place:

$$A_j^- e_0 = 0 \quad (j=1,2), \quad A_1^+ e_0 = e_1,$$
$$A_2^+ e_0 = e_2 + (s+1)^{-1} A_{12}^{-+} e_1,$$
$$A_1^- e_1 = q_1 e_0, \quad A_2^- e_1 = A_1^+ e_1 = 0,$$
$$A_2^+ e_1 = [(q_1-1)/(m-2)](A_{12}^{++} e_0 - e_3),$$
$$A_1^- e_2 = -[(q_2-1)/(s+1)] A_{12}^{-+} e_0,$$
$$A_2^- e_2 = [s(q_2-1)/(s+1)] e_0,$$
$$A_1^+ e_2 = [s/(m-2)(s+1)][(q_1-1)e_3 + (q_2-1)A^{++}e_0],$$
$$A_2^+ e_2 = -[q_1 - 1/(s+1)(m-2)] A^{-+}(A^{++}e_0 - e_3),$$
$$A_1^- e_3 = [(m-1)/(q_1-1)](q_1 e_2 + [(q_2-1)/(s+1)] A^{-+} e_1),$$
$$A_2^- e_3 = -[(m-1)(q_2-1)/(q_1-1)] e_1,$$
$$A_1^+ e_3 = -[(q_2-1)/(q_1-1)] A^{++} e_1,$$
$$A_2^+ e_3 = A^{++}(e_2 + (s+1)^{-1} A^{-+} e_1),$$
$$(e_1,e_1) = q_1 (e_0,e_0),$$
$$(e_2,e_2) = [s(q_2-1)/(s+1)](e_0,e_0),$$
$$(e_3,e_3) = [q_1(q_2-1)(m-1)(m-2)/(q_1-1)^2](e_0,e_0). \tag{46}$$

Thence it follows that if (m,s) is a massive representation of the sp(4) algebra, then the representation space of the osp(1,4) superalgebra can be decomposed into the direct sum of the representation spaces corresponding to (m,s), $(m+1,s-1)$, $(m+1,s+1)$, $(m+2,s)$ if $s \neq 0$ and to $(m,0)$, $(m+1,1)$, $(m+2,0)$ if $s=0$. Now let (m,s) be a massless representation with $s \neq 0$. Then e_1 generates the representation $(m+1,s+1)$. Take into account that according to the results of Sec. IV,

$$A^{-+} A^{++} e_j = A^{++} A^{-+} e_j = 0 \quad (j=0,1).$$

Therefore as it follows from (46), the representation space of the osp(1,4) superalgebra can be decomposed into the direct sum of representation spaces corresponding to (m,s) and $(m+1,s+1)$.

Thus we have shown that in the modular case there take place the same decompositions into IR's of the sp(4) algebra as in the ordinary case.[63–65] We now discuss the question about the correspondence of representations under consideration in the sense of Definition 1. In the case of the Dirac supermultiplet, the correspondence obviously does not take place (however it does take place if the unit of measurements of angular momentum is $\hbar/4$ instead of $\hbar/2$). As it follows from the results of Sec. V, the hope of correspondence may be only in the massive case for even $m-s$. However, decomposition into the IR's of the sp(4) algebra is not convenient, since the formulas (46) contain the division nontrivially, and in analogy with the ordinary case[63,64] we can choose instead the basis of the form

$$(a_1'')^{n_1} (a_2'')^{n_2} (b'')^n \beta^* (L_-)^k e_0, \tag{47}$$

where $k=0,1,\ldots,s$, and β^* is one of the operators $1, d_1^*, d_2^*, [d_1^*, d_2^*]$. In the modular case the existence of such a basis is not obvious, since n_1, n_2, n assume only a finite number of values. Proceeding from the results of Sec. IV, it can be shown that for even $m-s$ the maximal value of n is equal to $p+2-m-\tilde{n}(\beta^*)$ where

$$\tilde{n}(1) = 0, \quad \tilde{n}(d_1^*) = \tilde{n}(d_2^*) = 1, \quad \tilde{n}([d_1^*, d_2^*]) = 2,$$

the maximal value of n_1 is equal to $p+1-q_1-n+k-\tilde{n}_1(\beta^*)$ where

$$\tilde{n}_1(1) = \tilde{n}_1(d_2^*) = 0,$$
$$\tilde{n}_1(d_1^*) = \tilde{n}_1([d_1^*, d_2^*]) = 1,$$

and the maximal value of n_2 is equal to $p+1-q_2-n-k-\tilde{n}_2(\beta^*)$ where

$$\tilde{n}_2(1) = \tilde{n}_2(d_1^*) = 0,$$
$$\tilde{n}_2(d_2^*) = \tilde{n}_2([d_1^*, d_2^*]) = 1.$$

The basis (47) is analogous to the basis from Sec. V; however, such a choice of basis does not ensure the correspondence in the sense of Definition 1. This is obvious from the relation

$$d_1^* d_2^* e_0 = \tfrac{1}{2}(b'' + [d_1^*, d_2^*]) e_0.$$

Thus there is no correspondence in the sense of Sec. II between supersymmetry operators. Otherwise speaking, dealing with matrix elements of supersymmetry operators in the modular case, one cannot be confined to the elements F_p from U_0 even at low energies. Is it not the reason due to which the supersymmetry is not yet discovered experimentally?

VIII. A FEW REMARKS ON MODULAR ANALOGS OF REPRESENTATIONS OF INFINITE-DIMENSIONAL ALGEBRAS

In the ordinary case, the representation of the Kac–Moody algebra associated with some Lie algebra $\tilde{A}$ is defined by means of operators T_n^j ($j=1,2,\ldots,\dim A$, $n \in Z$) such that $T_n^{j*} = T_{-n}^j$ and

$$[T_n^j, T_{n'}^k] = i f^{jkl} T_{n+n'}^l + \kappa n \delta^{jk} \delta_{n+n',0}, \tag{48}$$

where f^{jkl} are the structure constants of $\tilde{A}$, the summation over the repeated indices (here and henceforth) is meant, and κ is a central element having a numerical value in any IR. The modular analog of the above representation can be evidently defined by the same formulas, but now $n, n', f^{jkl}, \kappa \in F_p$.

In the ordinary case it is well known that if the operators L_n are defined by the formula

$$L_n = \text{s-lim} \frac{1}{2\beta} \sum_{n'=-N}^{N} \mathcal{N}\{T_{n+n'}^j T_{-n'}^j\}, \tag{49}$$

where s-lim means the strong limit and $\mathcal{N}$ means the normal ordering (such that

$$\mathcal{N}\{T^j_n T^k_{-n}\} = T^k_{-n} T^j_n \text{ if } n>0$$

and

$$\mathcal{N}\{T^j_n T^k_{n'}\} = T^j_n T^k_{n'},$$

in all other cases), and then if

$$f^{ijk}f^{ijl} = c\delta^{kl}, \quad \beta = \kappa + \tfrac{1}{2}c,$$

these operators form the representation of the Virasoro algebra,

$$[L_n, L_{n'}] = (n-n')L_{n+n'} + \frac{\kappa}{12\beta} n(n^2-1)\delta_{n+n',0} \dim \tilde{A}. \tag{50}$$

As the modular analog of such a construction, it is natural to define L_n in the form

$$L_n = \frac{1}{2\beta} \sum_{n' \in F_p} \mathcal{N}\{T^j_{n+n'} T^j_{-n'}\}, \tag{51}$$

where it is meant that in the definition of normal ordering, the elements 1,2,..., $(p-1)/2$ from F_p are considered as positive and the elements $(p+1)/2$,..., $p-1$ as negative. Then it is easy to calculate that if $\beta = \kappa$, then

$$[L_n, L_{n'}] = (n-n')L_{n+n'} - \tfrac{1}{4} n\delta_{n+n',0} \dim A. \tag{52}$$

In the ordinary case, the correspondence between the representations of Kac–Moody and Virasoro algebras is used for the description of strings not only in a flat space, but also in the de Sitter space,[66] and the central element of the Virasoro algebra is connected with the critical dimension of the space-time. Therefore one can come to a conclusion that the critical dimension strongly depends on whether the physics is based on the field C or F_{p^2}, or, otherwise speaking, on whether the number p is finite or infinite.

It is known that many infinite-dimensional Lie algebras have (finite-dimensional) modular analogs. In the case of simple finite-dimensional restricted Lie algebras over an algebraically closed field, the problem of classification has been investigated in Refs. 67–69, and Ref. 70 contains the complete solution of this problem if $p>7$. The representation theory of such algebras contains yet only partial results and most of the papers apparently deal with the determination of maximal dimensions of IR's (see Sec. I).

IX. DISCUSSION AND CONCLUSIONS

In the present work we have given the mathematical formulation of the correspondence between the description of physical systems on the language of spaces and operators over the field F_{p^2} and on the conventional language of projective Hilbert spaces. This correspondence explains in what sense the conventional physics can be treated as the limit of "modular" physics when $p \to \infty$. Among considered concrete cases, the clearest correspondence seems to be between modular representations of the so(2,3) algebra and its representations in a Hilbert space. Indeed if in the case of the basis considered in Sec. V one formally goes to the limit $p \to \infty$ and introduces the rational and irrational numbers, then one obtains an ordinary representation of the so(2,3) algebra in a Hilbert space. Then by means of the standard contraction procedure, one can (if it is desirable) obtain the conventional representation of the Poincaré algebra, integrate it to a global representation of the Poincaré group, etc. Otherwise speaking, if the radius of the de Sitter space is denoted by R, then the transition from the modular representation of the de Sitter algebra to a conventional representation of the Poincaré algebra can be performed in the succession $p \to \infty$, $R \to \infty$, but not vice versa. In particular, it is stated *a priori* in the considered approach that the cosmological constance differs from zero.

The modular analog of the representations of the de Sitter algebras looks like natural way of quantization of such quantities as energy, mass, and momentum, since even for the elementary particles these quantities in the de Sitter units are very large and the requirement that they assume only integer values does not obviously contradict an experiment (in conventional units the quantum of the above quantities is the value of the order $\hbar/R$). At the same time, if one proceeds from the Planck system of units, then the modular analogs of representations describing particles and strings is unphysical, since masses and energies of elementary particles in these units are much less than one.

Proceeding from the modular analogs of representations describing elementary particles, it is possible to construct, in principle, modular analogs of operators in the Fock representation for any quantum field theory in the de Sitter space. The case when the theory contains only fermions is of particular interest, since due to the fermionic commutation relations, the number of particles in the modular version of a theory will be finite (of the order p^3) and all of the theory will be also discrete and finite.

We have pointed out that in the modular case, the probabilistic interpretation can be restored only in the limit $p \to \infty$. However, if one proceeds from the point of view that modular representations are the basis of a more fundamental physics, then they must have a physical interpretation for the finite values of p, and this interpretation is more fundamental than the conventional interpretation of quantum theory. Without such an interpretation it is unclear, in particular, in which experiment the number p can be defined (if it exists).

The very possibility of constructing the physics without actual infinity seems to be very attractive, and therefore the investigation of representations in spaces over F_{p^2} and other finite fields is of indubitable interest.

ACKNOWLEDGMENTS

The author is grateful to M. V. Borovoj, E. V. Gaidukov, M. A. Olshanetsky, V. N. Tolstoj, and I. V. Volovich for numerous useful discussions, and to S. V. Harlamova for remarks concerning the translation of the manuscript.

[1] H. R. Coish, "Elementary particles in a finite world geometry," Phys. Rev. **114**, 383 (1959).

[2] I. S. Shapiro, "Weak interactions in the theory of elementary particles with finite space," Nucl. Phys. **21**, 474 (1960).

[3] S. Wolfram, "Statistical mechanics of cellular automata," Rev. Mod. Phys. **55**, 601 (1983).

[4]O. Martin, A. M. Odlyzko, and S. Wolfram, "Algebraic properties of cellular automata," Comm. Math. Phys. **93**, 219 (1984).
[5]Yu. N. Babaev, "Gravitational field in discrete compact space, Dirac monopole and the nature of gravitational forces," preprint, Inst. Nucl. Res. Acad. Sci. USSR 0481 (1986).
[6]I. V. Volovich, "*P*-adic space-time and string theory," Teor. Mat. Fiz. **71**, 337 (1987); "*P*-adic string," Class. Quant. Grav. **4**, L 83 (1987); "Harmonic analysis and *p*-adic strings," Lett. Math. Phys. **16**, 61 (1988).
[7]B. Grossman, "*P*-adic strings, the Weil conjectures and anomalies," Phys. Lett. B **197**, 101 (1987).
[8]R. Brauer and C. Nesbitt, "On the modular characters of groups," Ann. Math. **42**, 556 (1941).
[9]H. Zassenhaus, "The representations of Lie algebras of prime characteristic," Proc. Glasgow Math. Assoc. **2**, 1 (1954).
[10]A. N. Rudakov and I. R. Shafarevich, "Irreducible representations of simple three-dimensional Lie algebra over a field of finite characteristic," Mat. Zametki **2**, 439 (1967).
[11]B. Braden, "Restricted representations of classical Lie algebras of type A_2 and B_2," Bull. Am. Math. Soc. **73**, 482 (1967); thesis, Univ. of Oregon, Eugene, OR (1966).
[12]R. Steinberg, *Lectures on Chevalley Groups*, (Yale U.P., New Haven, CT, 1967).
[13]A. N. Rudakov, "On the representations of classical semisimple Lie algebras in characteristic *p*," Izv. Akad. Nauk SSSR, Ser. Mat. **34**, 735 (1970).
[14]J. E. Humphreys, "Modular representations of classical Lie algebras and semisimple groups," J. Algebra **19**, 51 (1971).
[15]B. Yu. Veisfeiler and V. G. Kac, "On the irreducible representations of *p*-Lie algebras," Funk. Anal. Priloz. **5**, 28 (1971).
[16]A. N. Rudakov, "Weights of modular representations," Mat. Zametki **11**, 397 (1972).
[17]L. Dornhoff, *Group Representation Theory. Part B: Modular Representation Theory* (Dekker, New York, 1972).
[18]J. C. Jantzen, "Zur Characterformel gewisser Darstellungen helbeinfacher Gruppen und Lie-Algebren," Math. Z. **140**, 127 (1974); "Darstellungen helbeinfacher Gruppen und kontravariante Formen," J. Reine Angew. Math. **290**, 117 (1977).
[19]J. E. Humphreys, "Ordinary and modular representations of Chevalley groups," Lect. Notes Math. **528**, 1 (1976).
[20]A. A. Milner, "Irreducible representations of modular Lie algebras," Izv. Akad. Nauk SSSR, Ser. Mat. **39**, 1240 (1975).
[21]B. M. Puttaswamaiach and J. D. Dixon, *Modular Representations of Finite Groups* (Academics, New York, 1977).
[22]A. N. Grishkov, "Irreducible representations of modular Lie algebras," Mat. Zametki **30**, 21 (1981).
[23]J. Schue, "Representations of Lie *p*-algebras," Lect. Notes Math. **933**, 191 (1982).
[24]A. A. Premet, "Irreducible restricted representations of Hamilton and contact *p*-algebras," preprint, Inst. Mat. Akad. Nauk BSSR 14/171 (1983).
[25]Ya. S. Krylyuk, "On the maximal dimension of irreducible representations of simple *p*-Lie algebras of Cartan series *S* and *H*," Mat. Sbornik **123**, 109 (1984).
[26]D. Benson, "Modular representations theory: new trends and methods," Lect. Notes Math. **1081**, 1 (1984).
[27]R. W. Carter, *Finite Groups of Lie Type. Conjugacy Classes and Complex Characters* (Wiley, New York, 1985).
[28]A. N. Panov, "Irreducible representations of Lie algebra $\mathrm{sl}(n)$ over a field of positive characteistic," Mat. Sbornik **128**, 21 (1985).
[29]A. I. Kostrikin and I. A. Chubarov, "Representations of finite groups," Itogi nauki VINITI: Algebra, Topology, Geometry **23**, 119 (1985).
[30]R. Dipper and J. Gordon, "Identification of the irreducible modular representations of $GL_n(g)$," J. Algebra **104**, 266 (1986).
[31]J. E. Humphreys, "Projective modules for Sp $(4,p)$ in characteristic *p*," J. Algebra **104**, 80 (1986).
[32]J. L. Alperin, *Local Representation Theory. Modular Representation Theory as an Introduction to the Local Representation Theory of Finite Groups* (Cambridge U.P., Cambridge, 1986).
[33]C. W. Curtis, "Topics in the theory of representations of finite groups," Lect. Notes Math. **1185**, 58 (1986).
[34]J. C. Jantzen, "Modular representations of reductive groups," Lect. Notes Math. **1185**, 118 (1986).
[35]J. E. Humphreys, "The Steinberg representations," Bull. Am. Math. Soc. **16**, 247 (1987).
[36]E. M. Friedlander and B. J. Parshall, "Representations of mod *p* Lie algebras," Bull. Am. Math. Soc. **17**, 129 (1987).
[37]A. Colembiowski, "Zur Berechnung modular irreduzibler Matrixdarstellungen symmetrisher Gruppen mit Hilfe eines Verfahrens von M. Clausen," Bayreuth Math. Schr. **25**, 135 (1987).
[38]S. D. Smith, "Geometric techniques in representation theory," Geom. dedic. **25**, 355 (1988).
[39]P. G. O. Freund and M. Olson, "Non-Archimedean strings," Phys. Lett. B **199**, 186 (1987); *P*-adic dynamical systems, Nucl. Phys. B **297**, 86 (1987).
[40]I. Ya. Aref'eva, B. G. Dragovic, and I. V. Volovich, "Open and closed p-adic strings and quadratic extensions of number fields," preprint, Inst. Phys. Yugoslavia 14 (1988).
[41]P. H. Frampton and Y. Okada, "*P*-adic string *N*-point function," Phys. Rev. Lett. **60**, 484 (1988).
[42]Z. Ryzak, "Scattering amplitudes from higher dimensional *p*-adic world sheet," Phys. Lett. B **207**, 411 (1988).
[43]B. L. Spokoiny, "Quantum geometry of non-archimedean particles and strings," Phys. Lett. B **207**, 401 (1988).
[44]H. Yamakoshi, "Arithmetic of strings," Phys. Lett. B **207**, 426 (1988).
[45]R. B. Zhang, "Lagrangian formulation of open and closed *p*-adic strings," Phys. Lett. B **209**, 229 (1988).
[46]L. Brekke, P. G. O. Freund, M. Olson, and E. Witten, "Non-Archimedean string dynamics," Nucl. Phys. B **302**, 365 (1988).
[47]P. G. O. Freund and E. Witten, "Adelic string amplitudes,"' preprint, LASSNS-HEP 42 (1987); Phys. Lett. B **199**, 191 (1987).
[48]Yu. I. Manin, "Reflection on arithmetical physics," talk at the Poiana–Brasov school on strings and conformal field theory (1987).
[49]I. Ya. Aref' eva, B. G. Dragovic, and I. V. Volovich, "On the adelic string amplitudes," preprint, Inst. Phys. Yugosalvia 13 (1988).
[50]A. Le Clair, "An operator formulation of the superstring," Nucl. Phys. B **303**, 189 (1988).
[51]L. Alvarez Gaume, C. Gomez, G. Moore, and C. Vafa, "Strings in the operator formalism," Nucl. Phys. B **303**, 455 (1988).
[52]F. M. Lev, "Some representations of de Sitter algebras over the finite field and their possible physical interpretation," Yad. Fiz. **48**, 903 (1988).
[53]K. V. Kozerenko, "Full list of invariants of sl(2) algebra in a case of field with finite characteristic," Funk. Anal. Priloz. **22**, 73 (1988).
[54]C. Fronsdal, "Elementary particles in a curved space," Rev. Mod. Phys. **37**, 221 (1965).
[55]N. T. Evans, "Discrete series for the universal covering group of the $3+2$ de Sitter group," J. Math. Phys. **8**, 170 (1967).
[56]C. Fronsdal, "Elementary particles in a curved space. IV. Massless particles," Phys. Rev. D **12**, 3819 (1975).
[57]P. A. M. Dirac, "A remarkable representation of the $3+2$ de Sitter group," J. Math. Phys. **4**, 901 (1963); L. Castell and W. Heidenreich, SO(3,2)-invariant scattering and Dirac singletons, Phys. Rev. D **24**, 371 (1981).
[58]B. L. Van Der Varden, *Algebra II* (Springer-Verlag, New York, 1967).
[59]O. Castanos and M. Moshinsky, "Matrix representation of the generators of symplectic algebras: I. The case of sp(4,*R*)," J. Phys. A: Math. Gen. **20**, 513 (1987).
[60]J. P. Antoine and D. Speiser, "Characters of irreducible representations of the simple groups II," J. Math. Phys. **5**, 1560 (1964).
[61]J. W. B. Hughes, "$\mathrm{SU}(2)\times\mathrm{SU}(2)$ shift operators and representations of SO(5)," J. Math. Phys. **24**, 1015 (1983).
[62]F. M. Lev, "Some group-theoretical aspects of the SO(1,4)-invariant theory," J. Phys. A: Math. Gen. **21**, 599 (1988).
[63]W. Heidenreich, "All linear unitary irreducible representations of de Sitter supersymmetry with positive energy," Phys. Lett. B **110**, 461 (1982).
[64]I. Inaba, T. Maekawa, and Y. Yamamoto, "Infinite dimensional representations of the graded Lie algebra (Sp(4):4). Representation of the para-Bose operators with real order of quantization," J. Math. Phys. **23**, 954 (1982).
[65]C. Fronsdal, "Dirac supermultiplet," Phys. Rev. D **26**, 1988 (1982).
[66]H. J. De Vega and N. Sanchez, "A new approach to string quantization in curved space-times," Phys. Lett. B **197**, 320 (1987).
[67]A. I. Kostrikin and I. R. Shafarevich, Graded Lie algebras of finite characteristic," Izv. Akad. Nauk SSSR. Ser. Mat. **33**, 251 (1969).
[68]V. G. Kac, "On the classification of simple Lie algebras over a field with nonzero characteristic," Izv. Akad. Nauk SSSR, Ser. Mat. **34**, 385 (1970).
[69]R. E. Block, "The classification problem for simple Lie algebras of characteristic *p*," Lect. Not. Math. **933**, 38 (1982).
[70]R. E. Block and R. L. Wilson, "Classification of the restricted simple Lie algebras," J. Algebra **114**, 115 (1988).

A simple proof of the result that the Wigner transformation is of finite order

M. A. Rashid
Mathematics Department, Ahmadu Bello University, Zaria, Nigeria

(Received 23 February 1988; accepted for publication 29 March 1989)

Using a set of functions with linear span that is dense in $L^2(R^{2N})$, a simple proof is constructed of the result (discovered and proved by Várilly and Gracia-Bondia recently [J. Math. Phys. **28**, 2390 (1987)]) that the Wigner transformation is of order 24 and that its sixth power is the inverse Fourier transform (or Fourier cotransform).

I. INTRODUCTION

In a recent paper, Várilly and Gracia-Bondia[1] have established the unexpected result that the Wigner transformation W from $L^2(R^{2N})$ onto itself, given by

$$(Wf)(x,y) = (2\pi)^{-1/2N}\int_{R^N} f(2^{-1/2}(x+y),t) \times \exp(2^{-1/2}i(x-y)^T t)dt, \tag{1}$$

is of order 24 and that its sixth power is the inverse Fourier transform (or Fourier contransform) where the Fourier transform F on $L^2(R^{2N})$ is defined by

$$(Ff)(u) = (2\pi)^{-N}\int_{R^{2N}} f(v)\exp(-iu^T v)dv. \tag{2}$$

In the above two equations and in the sequel, we shall generally follow the notation and the conventions introduced in Ref. 1. In particular u, v, z, and η are $2N\times 1$ column vectors whereas x, y, t, α, and β are $N\times 1$ column vectors. We shall also, whenever necessary, partition z and η as $z=(x/y)$, $\eta=(\alpha/\beta)$. Also $u,v,z\in R^{2N}$, $\eta\in\mathbb{C}^{2N}$, and t, appearing as an integration variable, $\in R^N$.

In this paper, we use a set of functions that resemble the functions used in the coherent state presentation and whose linear span is dense in $L^2(R^{2N})$ and show that this set is not only invariant under the Wigner transformation but that this transformation is represented on this set by a very simple $2N\times 2N$ complex matrix of order 24. Again the inverse Fourier transform is represented on this set by iI_{2N}, which is also the sixth power of the matrix representing the Wigner transformation. Two other operators R and Φ associated with the Wigner transformation also have a very simple $2N\times 2N$ matrix representation on the set of functions we use. These operators were introduced in Ref. 1 and were defined on $L^2(R^{2N})$ by

$$(Rf)(x,y) = f(2^{-1/2}(x+y),2^{-1/2}(x-y)) \tag{3}$$

and

$$(\Phi f)(x,y) = (2\pi)^{-1/2N}\int_{R^N} f(x,t)\exp(iy^T t)dt. \tag{4}$$

Obviously

$$R^2 = \mathrm{Id}, \tag{5}$$

and

$$W = R\Phi. \tag{6}$$

II. PROOF OF THE MAIN RESULTS

For $\eta\in\mathbb{C}^{2N}$ and $z\in R^{2N}$, where η and z are $2N\times 1$ column vectors, we define[2]

$$f_\eta(z) = \exp(-\tfrac{1}{2}z^T z + 2\eta^T z - \eta^T\eta) \tag{7}$$

$$= \exp(-\tfrac{1}{2}(x^T x + y^T y) + 2(\alpha^T x + \beta^T y) - (\alpha^T\alpha + \beta^T\beta)). \tag{8}$$

The set S, defined by

$$S = \{f_\eta(z):\ \eta\in\mathbb{C}^{2N}\}, \tag{9}$$

is such that its linear span is dense in $L^2(R^{2N})$ because this set contains all the *translates* of a single Gaussian, a set whose span is well known to be dense in $L^2(R^{2N})$. We now compute the Wigner transform (Wf_η) of f_η. Indeed, using the definition in Eqs. (1) and (8) and noting that the phase-space variable z and the variable η are partitioned as (x/y) and (α/β), we have

$$(Wf_\eta)(x,y) = (2\pi)^{-1/2N}\int_{R^N}\exp\left[-\frac{1}{2}\left(\left(\frac{x+y}{\sqrt{2}}\right)^T\left(\frac{x+y}{\sqrt{2}}\right)+t^T t\right) + 2\left(\alpha^T\frac{x+y}{\sqrt{2}}+\beta^T t\right) - (\alpha^T\alpha+\beta^T\beta) + i\left(\frac{x-y}{\sqrt{2}}\right)^T t\right]dt. \tag{10}$$

Integrating and then simplifying, we arrive at

$$(Wf_\eta)(x,y) = \exp\left[-\frac{1}{2}(x^T x + y^T y) + 2\left(\frac{\alpha+i\beta}{\sqrt{2}}\right)^T x + 2\left(\frac{\alpha-i\beta}{\sqrt{2}}\right)^T y - \left(\frac{\alpha+i\beta}{\sqrt{2}}\right)^T\left(\frac{\alpha+i\beta}{\sqrt{2}}\right) + \left(\frac{\alpha-i\beta}{\sqrt{2}}\right)^T\left(\frac{\alpha-i\beta}{\sqrt{2}}\right)\right]. \tag{11}$$

 0022-2488/89/091999-02$02.50

On comparing the above result with Eq. (8), we immediately find

$$(Wf_\eta)(x,y) = f_{A\eta}(x,y), \tag{12}$$

where A is the $2N \times 2N$ complex matrix

$$A = \frac{1}{\sqrt{2}} \begin{pmatrix} I_N & iI_N \\ I_N & -iI_n \end{pmatrix}. \tag{13}$$

In the above, I_N is the $N \times N$ identity matrix.

Again starting from Eqs. (2) and (7) we can similarly show that the Fourier transform $(Ff_\eta)(z)$ of $f_\eta(z)$ is given by

$$Ff_\eta(z) = \exp(-\tfrac{1}{2}z^T z + 2(-i\eta)^T z - (-i\eta)^T(-i\eta)), \tag{14}$$

which, on comparing with Eq. (7), results in

$$(Ff_\eta)(z) = f_{-i\eta}(z). \tag{15}$$

Similarly

$$(F^{-1}f_\eta)(z) = f_{i\eta}(z), \tag{16}$$

where the inverse Fourier transform F^{-1} is defined on $L^2(R^{2N})$ by

$$(F^{-1}f)(u) = (2\pi)^{-N} \int_{R^{2N}} f(v)\exp(iu^T v)dv. \tag{17}$$

From Eq. (16), it is obvious that in the space of η's, *which index the elements of the set S*, the inverse Fourier transform F^{-1} is represented by the complex $2N \times 2N$ matrix

$$i\begin{pmatrix} I_N & 0_N \\ 0_N & I_N \end{pmatrix}, \tag{18}$$

where I_N and 0_N are the $N \times N$ identity and zero matrices, respectively.

To prove the main result of our paper, we now just have to find the various powers of the matrix A, which can be done trivially. In particular we find that

$$A^6 = i\begin{pmatrix} I_N & 0_N \\ 0_N & I_N \end{pmatrix} \tag{19}$$

and

$$A^{24} = \begin{pmatrix} I_N & 0_N \\ 0_N & I_N \end{pmatrix}, \tag{20}$$

which proves that the Wigner transformation is of order 24 and that its sixth power is the inverse Fourier transform. Note that for this purpose we have to extend W and F from the set S first by linearity to the linear span of S and then by continuity to the whole of $L^2(R^{2N})$. This last step is possible because the linear span of S is dense in $L^2(R^{2N})$.

III. THE OPERATORS R AND Φ

For the operators R and Φ defined in Eqs. (3) and (4) above, using the technique in Sec. II, we have

$$(Rf_\eta)(z) = f_{B\eta}(z) \tag{21}$$

and

$$(\Phi f_\eta)(z) = f_{C\eta}(z), \tag{22}$$

where the $2N \times 2N$ matrices B and C are given by

$$B = \frac{1}{\sqrt{2}} \begin{pmatrix} I_N & I_N \\ I_N & -I_N \end{pmatrix} \tag{23}$$

and

$$C = \begin{pmatrix} I_N & 0_N \\ 0_N & iI_N \end{pmatrix}. \tag{24}$$

The property $W = R\Phi$ given in Eq. (6) above is represented by

$$A = BC, \tag{25}$$

in terms of the matrices A, B, and C.

Note that the matrices A, B, and C are all *unitary* with determinant equal to $(-i)^N$, $(-1)^N$, and $(i)^N$, respectively. (Thus, in general, these are not unimodular.)

ACKNOWLEDGMENT

The author expresses his sincerest thanks to the referee for making some suggestions that have resulted in a still shorter proof than the one suggested in the first submission. The referee's suggestions have also added clarity and completeness.

[1] J. C. Várilly and J. M. Gracia-Bondia, J. Math. Phys. **28**, 2390 (1987).

[2] The function $f_\eta(z)$ can be expressed as a product of generating functions for $H_m(x)e^{-(1/2)m^2}$ as follows from

$$\sum_{m=0}^{\infty} \frac{\alpha^m}{m!} H_m(x)e^{-(1/2)x^2} = e^{-(1/2)x^2 + \alpha x - \alpha^2}.$$

In the above, $H_m(x)$ is the Hermite polynomial of degree m.

Multivariable Wilson polynomials

M. V. Tratnik
Center for Nonlinear Studies and Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received September 1988; accepted for publication 19 April 1989)

A multivariable biorthogonal generalization of the Wilson polynomials is presented. These are four distinct families, which in a special case occur in two complex conjugate pairs, that satisfy four biorthogonality relations among them. An interesting limit case is the multivariable continuous dual Hahn polynomials.

I. INTRODUCTION

Wilson[1] and Askey and Wilson[2] introduce a family of hypergeometric orthogonal polynomials in a single variable that include as special or limiting cases all the classical orthogonal polynomials and many related families. Their discrete analog, also known as Racah polynomials, provide an explicit representation for the $6j$ symbols of angular momentum theory. Further interesting properties and applications are discussed by various authors. Dunkl[3] finds that the Wilson polynomials are connection coefficients between different bases for the solution space of a linear difference equation. Letessier and Valent[4] show that they can be interpreted as transition probabilities for a birth and death process. Miller[5] uses local symmetry techniques to obtain an elegant orthogonality proof and an elementary evaluation of the norm. Rao *et al.*[6] deduce a three term recurrence relation while Montaldi and Zucchelli[7] present still another orthogonality proof.

These important polynomials can be expressed as the following hypergeometric series:

$$P_n(x) = (a+b)_n (a+c)_n (a+d)_n \, {}_4F_3\left(\begin{matrix} -n, n+a+b+c+d-1, a-ix, a+ix \\ a+b, a+c, a+d \end{matrix} ; 1 \right), \tag{1.1}$$

where a, b, c, and d are complex parameters, n is a non-negative integer, and $(a+b)_n \equiv \Gamma(n+a+b)/\Gamma(a+b)$ denotes the usual Pochhammer symbol. These are polynomials of degree n in x^2 or of degree $2n$ in x (the latter interpretation extends to the multivariable case). One can show, by iterating a transformation satisfied by the ${}_4F_3$ hypergeometric function,[1] that $P_n(x)$ is symmetric under the interchange of all four parameters a, b, c, and d, which is a continuous analog of the symmetries of the $6j$ symbols. When the real parts of a, b, c, and d are positive the Wilson polynomials satisfy a continuous orthogonality relation on the real line

$$\int_{-\infty}^{\infty} dx \, P_n(x) P_m(x) w(x) = \delta_{nm} h_n, \tag{1.2}$$

where the weight function $w(x)$ is given by

$$w(x) = \frac{\Gamma(a+ix)\Gamma(a-ix)\Gamma(b+ix)\Gamma(b-ix)\Gamma(c+ix)\Gamma(c-ix)\Gamma(d+ix)\Gamma(d-ix)}{\Gamma(2ix)\Gamma(-2ix)}, \tag{1.3}$$

and the normalization constant h_n is

$$h_n = 4\pi n!(n+a+b+c+d-1)_n \frac{\Gamma(n+a+b)\Gamma(n+a+c)\Gamma(n+a+d)\Gamma(n+b+c)\Gamma(n+b+d)\Gamma(n+c+d)}{\Gamma(2n+a+b+c+d)}. \tag{1.4}$$

If the parameters a, b, c, and d are real or if they occur in complex conjugate pairs, or a combination of both (with positive real parts), then the polynomials $P_n(x)$ are real and the weight function can be expressed as a modulus squared:

$$w(x) = \left| \frac{\Gamma(a+ix)\Gamma(b+ix)\Gamma(c+ix)\Gamma(d+ix)}{\Gamma(2ix)} \right|^2, \tag{1.5}$$

which is real and positive.

Two interesting limit cases are the continuous Hahn and continuous dual Hahn polynomials. The continuous Hahn family is obtained by setting

$$a = a' + \tfrac{1}{2}i\omega, \quad b = b' - \tfrac{1}{2}iw, \quad c = c' + \tfrac{1}{2}iw, \quad d = d' - \tfrac{1}{2}iw, \quad x = x' - \tfrac{1}{2}w, \tag{1.6}$$

dividing (1.1) by $n!w^n$, and then taking the limit $w \to \infty$. The resulting polynomials are (dropping the primes)

$$H_n(x) = \frac{i^n}{n!}(a+b)_n (a+d)_n \, {}_3F_2\left(\begin{matrix} -n, n+a+b+c+d-1, a+ix \\ a+b, a+d \end{matrix} ; 1 \right), \tag{1.7}$$

 0022-2488/89/092001-11$02.50

and from (1.2)–(1.4) one deduces the orthogonality relation

$$\int_{-\infty}^{\infty} dx\, H_n(x)H_m(x)\Gamma(a+ix)\Gamma(b-ix)\Gamma(c+ix)\Gamma(d-ix)$$
$$=\delta_{nm}2\pi(n+a+b+c+d-1)_n\,\frac{\Gamma(n+a+b)\Gamma(n+a+d)\Gamma(n+b+c)\Gamma(n+c+d)}{n!\Gamma(2n+a+b+c+d)}. \tag{1.8}$$

The continuous dual Hahn polynomials $D_n(x)$ result upon dividing (1.1) by d^n and taking the limit $d\to\infty$,

$$D_n(x)=(a+b)_n(a+c)_n\,{}_3F_2\left(\begin{matrix}-n,a+ix,a-ix\\ a+b,a+c\end{matrix};1\right), \tag{1.9}$$

and these satisfy the orthogonality relation

$$\int_{-\infty}^{\infty} dx\, D_n(x)D_m(x)\,\frac{\Gamma(a+ix)\Gamma(a-ix)\Gamma(b+ix)\Gamma(b-ix)\Gamma(c+ix)\Gamma(c-ix)}{\Gamma(2ix)\Gamma(-2ix)}$$
$$=\delta_{nm}4\pi n!\Gamma(n+a+b)\Gamma(n+a+c)\Gamma(n+b+c), \tag{1.10}$$

as follows from (1.2)–(1.4).

In Sec. II we introduce the multivariable Wilson polynomials and their associated weight function and calculate a multiple Mellin–Barnes-type integral that is the norm of the weight function. In Sec. III we deduce the biorthogonality relations satisfied by these multivariable polynomials and in Sec. IV we consider two limits, the multivariable continuous Hahn and continuous dual Hahn polynomials, as well as a few special cases.

II. MULTIVARIABLE WILSON POLYNOMIALS

The extension to p variables $x_1,x_2,\ldots,x_p$ is given by the following four families:

$$P\left(\begin{matrix}x_1,x_2,\ldots,x_p\\ n_1,n_2,\ldots,n_p\end{matrix}\middle|\begin{matrix}a_1,a_2,\ldots,a_p\\ b_1,b_2,\ldots,b_p\end{matrix}\middle|\, c,d\right)=\left[\prod_{k=1}^{p}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(a_k+b_k)}\right](A+c)_N(A+d)_N$$
$$\times F^{2:2;\ldots;2}_{2:1;\ldots;1}\left(\begin{matrix}N+A+B+c+d-1,A-iX:-n_1,a_1+ix_1;\ldots;-n_p,a_p+ix_p\\ A+c,A+d:a_1+b_1;\ldots;a_p+b_p\end{matrix}\right), \tag{2.1}$$

$$\overline{P}\left(\begin{matrix}x_1,x_2,\ldots,x_p\\ n_1,n_2,\ldots,n_p\end{matrix}\middle|\begin{matrix}a_1,a_2,\ldots,a_p\\ b_1,b_2,\ldots,b_p\end{matrix}\middle|\, c,d\right)=\left[\prod_{k=1}^{p}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(a_k+b_k)}\right](B+c)_N(B+d)_N$$
$$\times F^{2:2;\ldots;2}_{2:1;\ldots;1}\left(\begin{matrix}N+A+B+c+d-1,B+iX:-n_1,b_1-ix_1;\ldots;-n_p,b_p-ix_p\\ B+c,B+d:a_1+b_1;\ldots;a_p+b_p\end{matrix}\right), \tag{2.2}$$

$$Q\left(\begin{matrix}x_1,x_2,\ldots,x_p\\ n_1,n_2,\ldots,n_p\end{matrix}\middle|\begin{matrix}a_1,a_2,\ldots,a_p\\ b_1,b_2,\ldots,b_p\end{matrix}\middle|\, c,d\right)=\left[\prod_{k=1}^{p}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(a_k+b_k)}\right](c-iX)_N(d-iX)_N$$
$$\times F^{2:2;\ldots;2}_{2:1;\ldots;1}\left(\begin{matrix}-N-c-d+1,B+iX:-n_1,a_1+ix_1;\ldots;-n_p,a_p+ix_p\\ -N-c+iX+1,-N-d+iX+1:a_1+b_1;\ldots;a_p+b_p\end{matrix}\right), \tag{2.3}$$

$$\overline{Q}\left(\begin{matrix}x_1,x_2,\ldots,x_p\\ n_1,n_2,\ldots,n_p\end{matrix}\middle|\begin{matrix}a_1,a_2,\ldots,a_p\\ b_1,b_2,\ldots,b_p\end{matrix}\middle|\, c,d\right)=\left[\prod_{k=1}^{p}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(a_k+b_k)}\right](c+iX)_N(d+iX)_N$$
$$\times F^{2:2;\ldots;2}_{2:1;\ldots;1}\left(\begin{matrix}-N-c-d+1,A-iX:-n_1,b_1-ix_1;\ldots;-n_p,b_p-ix_p\\ -N-c-iX+1,-N-d-iX+1:a_1+b_1;\ldots;a_p+b_p\end{matrix}\right), \tag{2.4}$$

where $F^{q:l_1;\ldots;l_p}_{r:v_1;\ldots;v_p}$ is the generalized Kampé de Fériet hypergeometric series[8] defined as

$$F^{q:l_1;\ldots;l_p}_{r:v_1;\ldots;v_p}\left(\begin{matrix}\alpha_1,\ldots,\alpha_q:\beta_1^{(1)},\ldots,\beta_{l_1}^{(1)};\ldots;\beta_1^{(p)},\ldots,\beta_{l_p}^{(p)};\\ \gamma_1,\ldots,\gamma_r:\xi_1^{(1)},\ldots,\xi_{v_1}^{(1)};\ldots;\xi_1^{(p)},\ldots,\xi_{v_p}^{(p)};\end{matrix}z_1,z_2,\ldots,z_p\right)$$
$$=\sum_{\{j_k\}}\frac{\Pi_{i=1}^{q}(\alpha_i)_J\Pi_{i=1}^{l_1}(\beta_i^{(1)})_{j_1}\cdots\Pi_{i=1}^{l_p}(\beta_i^{(p)})_{j_p}}{\Pi_{i=1}^{r}(\gamma_i)_J\Pi_{i=1}^{v_1}(\xi_i^{(1)})_{j_1}\cdots\Pi_{i=1}^{v_p}(\xi_i^{(p)})_{j_p}}\frac{z_1^{j_1}}{j_1!}\frac{z_2^{j_2}}{j_2!}\cdots\frac{z_p^{j_p}}{j_p!}, \tag{2.5}$$

and $\{j_k\}$ denotes summation indices $j_1,j_2,\ldots,j_p$, which run over all non-negative integers. We use the convention that $1/\Gamma(-n)=0$, $n=0,1,2,\ldots$, and we have introduced the following shorthand notation:

$$X\equiv\sum_{k=1}^{p}x_k,\quad N\equiv\sum_{k=1}^{p}n_k,\quad J\equiv\sum_{k=1}^{p}j_k,\quad A\equiv\sum_{k=1}^{p}a_k,\quad B\equiv\sum_{k=1}^{p}b_k. \tag{2.6}$$

The overbars in (2.2) and (2.4) denote distinct families of polynomials and should not be confused with complex conjugation. Members of a given family are labeled by the set of p non-negative integers $n_1,n_2,\ldots n_p$ and the degree of a polynomial is given by $2N$, where N is the sum of these integers as defined in (2.6). These polynomials are associated with the following multivariable weight function:

$$w\left(x_1,x_2,\ldots,x_p \left| \begin{matrix} a_1,a_2,\ldots,a_p \\ b_1,b_2,\ldots,b_p \end{matrix} \right| c,d\right)$$
$$= \left[\prod_{k=1}^{p} \Gamma(a_k+ix_k)\Gamma(b_k-ix_k)\right] \frac{\Gamma(A-iX)\Gamma(B+iX)\Gamma(c+iX)\Gamma(c-iX)\Gamma(d+iX)\Gamma(d-iX)}{\Gamma(2iX)\Gamma(-2iX)}, \tag{2.7}$$

where the $2p+2$ complex parameters $a_1,a_2,\ldots,a_p$, $b_1,b_2,\ldots,b_p$, c, and d are assumed to have positive definite real parts but are otherwise arbitrary. When no ambiguity arises we simply write $P_n(x)$, $\overline{P}_n(x)$, $Q_n(x)$, $\overline{Q}_n(x)$, and $w(x)$ for the polynomials and weight function, respectively.

In the special case $p=1$ all four families (2.1)–(2.4) reduce to the familiar single-variable Wilson polynomials (1.1). For $P_n(x)$ this is obvious upon comparing (2.1) and (1.1). The remaining three families (for a single variable) are equivalent to (1.1) through a transformation formula satisfied by the ${}_4F_3$ hypergeometric series of unit argument.[1] Also the weight function (2.7) for $p=1$ obviously reduces to (1.3).

In general the four families $P_n(x)$, $\overline{P}_n(x)$, $Q_n(x)$, and $\overline{Q}_n(x)$ are distinct. However, the barred and unbarred pairs are related through the following transformations:

$$\overline{P}\left(\begin{matrix} x_1,x_2,\ldots,x_p \\ n_1,n_2,\ldots,n_p \end{matrix} \left| \begin{matrix} a_1,a_2,\ldots,a_p \\ b_1,b_2,\ldots,b_p \end{matrix} \right| c,d\right) = P\left(\begin{matrix} -x_1,-x_2,\ldots,-x_p \\ n_1,n_2,\ldots,n_p \end{matrix} \left| \begin{matrix} b_1,b_2,\ldots,b_p \\ a_1,a_2,\ldots,a_p \end{matrix} \right| c,d\right)$$
$$= P^*\left(\begin{matrix} x_1,x_2,\ldots,x_p \\ n_1,n_2,\ldots,n_p \end{matrix} \left| \begin{matrix} b_1^*,b_2^*,\ldots,b_p^* \\ a_1^*,a_2^*,\ldots,a_p^* \end{matrix} \right| c^*,d^*\right), \tag{2.8}$$

$$\overline{Q}\left(\begin{matrix} x_1,x_2,\ldots,x_p \\ n_1,n_2,\ldots,n_p \end{matrix} \left| \begin{matrix} a_1,a_2,\ldots,a_p \\ b_1,b_2,\ldots,b_p \end{matrix} \right| c,d\right) = Q\left(\begin{matrix} -x_1,-x_2,\ldots,-x_p \\ n_1,n_2,\ldots,n_p \end{matrix} \left| \begin{matrix} b_1,b_2,\ldots,b_p \\ a_1,a_2,\ldots,a_p \end{matrix} \right| c,d\right)$$
$$= Q^*\left(\begin{matrix} x_1,x_2,\ldots,x_p \\ n_1,n_2,\ldots,n_p \end{matrix} \left| \begin{matrix} b_1^*,b_2^*,\ldots,b_p^* \\ a_1^*,a_2^*,\ldots,a_p^* \end{matrix} \right| c^*,d^*\right), \tag{2.9}$$

where the asterisk denotes complex conjugation. Notice also that the weight function is invariant under these transformations. Furthermore, all four families and the weight function are symmetric under the interchange of c and d. These results are apparent from (2.1)–(2.4) and (2.7), and we are assuming the variables $x_1,x_2,\ldots,x_p$ are real. From the second relation in each of (2.8) and (2.9) we see that, in the special case $a_k=b_k^*$, $k=1,2,\ldots,p$ and $c=d^*$ or c and d real, the barred families are simply the complex conjugates of the unbarred families:

$$\overline{P}_n(x)=P_n^*(x), \quad \overline{Q}_n(x)=Q_n^*(x). \tag{2.10}$$

Also, in this special case, the weight function is real and positive:

$$w(x)=\left|\left[\prod_{k=1}^{p}\Gamma(a_k+ix_k)\right]\frac{\Gamma(A-iX)\Gamma(c+iX)\Gamma(d+iX)}{\Gamma(2iX)}\right|^2 \tag{2.11}$$

(recall we are assuming $x_1,x_2,\ldots,x_p$ are real).

Returning to general parameter values (with real parts greater than zero) we introduce the following alternate representations of $Q_n(x)$ and $\overline{Q}_n(x)$ ($L\equiv\Sigma_{k=1}^{p}l_k$):

$$Q_n(x)=\left[\prod_{k=1}^{p}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(a_k+b_k)}\right](c+d)_N(B+c)_N\sum_{\{l_k\}}\left[\prod_{k=1}^{p}\binom{n_k}{l_k}\right]\frac{(c+iX)_L}{(c+d)_L}$$
$$\times\frac{(c-iX)_L}{(B+c)_L}(-1)^L F_{1:1;\ldots;1}^{1:2;\ldots;2}\left(\begin{matrix} B+iX:-n_1+l_1,b_1-ix_1;\ldots;-n_p+l_p,b_p-ix_p \\ L+B+c:a_1+b_1;\ldots;a_p+b_p \end{matrix}\right), \tag{2.12}$$

$$\overline{Q}_n(x)=\left[\prod_{k=1}^{p}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(a_k+b_k)}\right](c+d)_N(A+d)_N\sum_{\{l_k\}}\left[\prod_{k=1}^{p}\binom{n_k}{l_k}\right]\frac{(d-iX)_L}{(c+d)_L}$$
$$\times\frac{(d+iX)_L}{(A+d)_L}(-1)^L F_{1:1;\ldots;1}^{1:2;\ldots;2}\left(\begin{matrix} A-iX:-n_1+l_1,a_1+ix_1;\ldots;-n_p+l_p,a_p+ix_p \\ L+A+d:a_1+b_1;\ldots;a_p+b_p \end{matrix}\right), \tag{2.13}$$

which are used in Sec. III to determine biorthogonality relations. To demonstrate the equivalence of (2.12) and (2.3) we begin with the following multiple summation theorem[9]:

$$F_D^{(p)}(\alpha,\beta_1,\beta_2,\ldots,\beta_p;\gamma;1,1,\ldots,1)=\frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta_1-\beta_2-\cdots-\beta_p)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta_1-\beta_2-\cdots-\beta_p)}, \tag{2.14}$$

where $F_D^{(p)}$ is the p-variable Lauricella hypergeometric series, a special case of (2.5),

$$F_D^{(p)}(\alpha,\beta_1,\beta_2,\ldots,\beta_p;\gamma;z_1,z_2,\ldots,z_p)\equiv F\begin{matrix}1:1;\ldots;1\\1:0;\ldots;0\end{matrix}\left(\begin{matrix}\alpha:\beta_1;\ \beta_2;\ \ldots;\beta_p;\\ \gamma:\ -\ ;\ -\ ;\ldots;\ -\ ;\end{matrix}\ z_1,z_2,\ldots,z_p\right). \tag{2.15}$$

Setting $\alpha = J+B+iX, \beta_k = -n_k+j_k+l_k$, and $\gamma = -N+J-c+iX+1$ in (2.14) gives the identity ($R\equiv\Sigma_{k=1}^p r_k$)

$$\frac{\Gamma(N+B+c)}{\Gamma(J+L+B+c)} = \sum_{\{r_k\}}\left[\prod_{k=1}^p \binom{n_k-j_k-l_k}{r_k}\right]\frac{\Gamma(J+R+B+iX)}{\Gamma(J+B+iX)}\frac{\Gamma(N-J-R+c-iX)}{\Gamma(L+c-iX)}, \tag{2.16}$$

where we have also used the reflection formula[10]

$$\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z). \tag{2.17}$$

Substituting (2.16) into (2.12) and bringing the $\{l_k\}$ sum on the inside yields

$$\begin{aligned} Q_n(x) = &\sum_{\{j_k\}}\left[\prod_{k=1}^p \binom{n_k}{j_k}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(j_k+a_k+b_k)}\frac{\Gamma(j_k+b_k-ix_k)}{\Gamma(b_k-ix_k)}\right](-1)^J \\ &\times\sum_{\{r_k\}}\left[\prod_{k=1}^p \binom{n_k-j_k}{r_k}\right]\frac{\Gamma(J+R+B+iX)}{\Gamma(B+iX)}\frac{\Gamma(N-J-R+c-iX)}{\Gamma(c-iX)} \\ &\times\sum_{\{l_k\}}\left[\prod_{k=1}^p \binom{n_k-j_k-r_k}{l_k}\right]\frac{\Gamma(N+c+d)}{\Gamma(L+c+d)}\frac{\Gamma(L+c+iX)}{\Gamma(c+iX)}(-1)^L. \end{aligned} \tag{2.18}$$

Then the $\{l_k\}$ sum is performed by setting $\alpha = c+iX, \beta_k = -n_k+j_k+r_k$, and $\gamma = c+d$ in (2.14) and using the reflection formula (2.17) again. In this manner one finds

$$\sum_{\{l_k\}}\left[\prod_{k=1}^p \binom{n_k-j_k-r_k}{l_k}\right]\frac{\Gamma(N+c+d)}{\Gamma(L+c+d)}\frac{\Gamma(L+c+iX)}{\Gamma(c+iX)}(-1)^L = \frac{\Gamma(N+c+d)}{\Gamma(N-J-R+c+d)}\frac{\Gamma(N-J-R+d-iX)}{\Gamma(d-iX)}. \tag{2.19}$$

Substituting this into (2.18) and reversing the $\{r_k\}$ sum, $r_k\to n_k-j_k-r_k$, gives the following expression:

$$\begin{aligned} Q_n(x) = &\sum_{\{j_k\}}\left[\prod_{k=1}^p \binom{n_k}{j_k}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(j_k+a_k+b_k)}\frac{\Gamma(j_k+b_k-ix_k)}{\Gamma(b_k-ix_k)}\right](-1)^J \\ &\times\sum_{\{r_k\}}\left[\prod_{k=1}^p \binom{n_k-j_k}{r_k}\right]\frac{\Gamma(N-R+B+iX)}{\Gamma(B+iX)}\frac{\Gamma(R+c-iX)}{\Gamma(c-iX)}\frac{\Gamma(R+d-iX)}{\Gamma(d-iX)}\frac{\Gamma(N+c+d)}{\Gamma(R+c+d)}. \end{aligned} \tag{2.20}$$

Then interchanging the order of the $\{j_k\}$ and $\{r_k\}$ sums yields

$$\begin{aligned} Q_n(x) = &\sum_{\{r_k\}}\left[\prod_{k=1}^p \binom{n_k}{r_k}\right]\frac{\Gamma(N-R+B+iX)}{\Gamma(B+iX)}\frac{\Gamma(R+c-iX)}{\Gamma(c-iX)}\frac{\Gamma(R+d-iX)}{\Gamma(d-iX)}\frac{\Gamma(N+c+d)}{\Gamma(R+c+d)} \\ &\times\left[\prod_{k=1}^p\sum_{j_k}\binom{n_k-r_k}{j_k}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(j_k+a_k+b_k)}\frac{\Gamma(j_k+b_k-ix_k)}{\Gamma(b_k-ix_k)}(-1)^{j_k}\right]. \end{aligned} \tag{2.21}$$

Now each j_k sum is performed by using the Chu–Vandermonde theorem[11]:

$${}_2F_1(-n,\alpha;\beta;1) = \frac{\Gamma(n+\beta-\alpha)\Gamma(\beta)}{\Gamma(n+\beta)\Gamma(\beta-\alpha)}. \tag{2.22}$$

This leads to

$$\begin{aligned} Q_n(x) = &\sum_{\{r_k\}}\left[\prod_{k=1}^p \binom{n_k}{r_k}\frac{\Gamma(n_k+a_k+b_k)}{\Gamma(n_k-r_k+a_k+b_k)}\frac{\Gamma(n_k-r_k+a_k+ix_k)}{\Gamma(a_k+ix_k)}\right] \\ &\times\frac{\Gamma(N+c+d)}{\Gamma(R+c+d)}\frac{\Gamma(N-R+B+iX)}{\Gamma(B+iX)}\frac{\Gamma(R+c-iX)}{\Gamma(c-iX)}\frac{\Gamma(R+d-iX)}{\Gamma(d-iX)}; \end{aligned} \tag{2.23}$$

and upon reversing the summations, $r_k\to n_k-r_k$, we obtain representation (2.3). The equivalence of (2.13) and (2.4) then follows by either of transformations (2.9) and interchanging c and d.

Next we derive the following multiple Mellin–Barnes-type integral:

$$\begin{aligned} &\int_{-\infty}^{\infty}dx_p\cdots\int_{-\infty}^{\infty}dx_1\left[\prod_{k=1}^p\Gamma(a_k+ix_k)\Gamma(b_k-ix_k)\right]\frac{\Gamma(A-iX)\Gamma(B+iX)\Gamma(c+iX)\Gamma(c-iX)\Gamma(d+iX)\Gamma(d-iX)}{\Gamma(2iX)\Gamma(-2iX)} \\ &\quad = 2(2\pi)^p\left[\prod_{k=1}^p\Gamma(a_k+b_k)\right]\frac{\Gamma(A+c)\Gamma(A+d)\Gamma(B+c)\Gamma(B+d)\Gamma(c+d)}{\Gamma(A+B+c+d)}, \end{aligned} \tag{2.24}$$

which is the norm of the weight function. The integration contours are the real axes and recall that the complex parameters $a_1,a_2,\ldots,a_p$, $b_1,b_2,\ldots,b_p$, c and d are assumed to have positive definite real parts.

We begin with a change of variables from $x_1,x_2,\ldots,x_p$ to $X,x_2,\ldots,x_p$ in which case the multiple integral becomes

$$\int_{-\infty}^{\infty} dX \int_{-\infty}^{\infty} dx_p \cdots \int_{-\infty}^{\infty} dx_2\, \Gamma(a_1+iX-iX_2^p)\Gamma(b_1-iX+iX_2^p)\left[\prod_{k=2}^{p}\Gamma(a_k+ix_k)\Gamma(b_k-ix_k)\right]$$
$$\times\frac{\Gamma(A-iX)\Gamma(B+iX)\Gamma(c+iX)\Gamma(c-iX)\Gamma(d+iX)\Gamma(d-iX)}{\Gamma(2iX)\Gamma(-2iX)}, \tag{2.25}$$

where we have introduced the following shorthand notation:

$$X_j^l \equiv \sum_{k=j}^{l} x_k, \quad A_j^l \equiv \sum_{k=j}^{l} a_k, \quad B_j^l \equiv \sum_{k=j}^{l} b_k. \tag{2.26}$$

The x_2 integration can now be performed by using the integral formula[12]

$$\int_{-\infty}^{\infty} dx\, \Gamma(\alpha+ix)\Gamma(\beta+ix)\Gamma(\gamma-ix)\Gamma(\delta-ix) = (2\pi)\frac{\Gamma(\alpha+\gamma)\Gamma(\alpha+\delta)\Gamma(\beta+\gamma)\Gamma(\beta+\delta)}{\Gamma(\alpha+\beta+\gamma+\delta)},$$
$$\mathrm{Re}(\alpha),\mathrm{Re}(\beta),\mathrm{Re}(\gamma),\mathrm{Re}(\delta)>0, \tag{2.27}$$

which gives

$$\int_{-\infty}^{\infty} dx_2\, \Gamma(a_1+iX-iX_2^p)\Gamma(b_1-iX+iX_2^p)\Gamma(a_2+ix_2)\Gamma(b_2-ix_2)$$
$$= (2\pi)\frac{\Gamma(a_1+b_1)\Gamma(a_2+b_2)}{\Gamma(a_1+a_2+b_1+b_2)}\Gamma(a_1+a_2+iX-iX_3^p)\Gamma(b_1+b_2-iX+iX_3^p). \tag{2.28}$$

From (2.27), (2.28), and induction one can show that

$$\int_{-\infty}^{\infty} dx_l \cdots \int_{-\infty}^{\infty} dx_2\, \Gamma(a_1+iX-iX_2^p)\Gamma(b_1-iX+iX_2^p)\left[\prod_{k=2}^{l}\Gamma(a_k+ix_k)\Gamma(b_k-ix_k)\right]$$
$$= (2\pi)^{l-1}\left[\prod_{k=1}^{l}\Gamma(a_k+b_k)\right]\frac{\Gamma(A_1^l+iX-iX_{l+1}^p)\Gamma(B_1^l-iX+iX_{l+1}^p)}{\Gamma(A_1^l+B_1^l)}, \tag{2.29}$$

and if we set $l=p$ in (2.29) and substitute in (2.25) we find the multiple integral becomes

$$(2\pi)^{p-1}\left[\prod_{k=1}^{p}\Gamma(a_k+b_k)\right][\Gamma(A+B)]^{-1}\int_{-\infty}^{\infty} dX\,\Gamma(A+iX)\Gamma(A-iX)$$
$$\times\frac{\Gamma(B+iX)\Gamma(B-iX)\Gamma(c+iX)\Gamma(c-iX)\Gamma(d+iX)\Gamma(d-iX)}{\Gamma(2iX)\Gamma(-2iX)}. \tag{2.30}$$

This is now simply proportional to the norm of the single-variable Wilson weight function given by (1.2)–(1.4) with $n=m=0$. Using this result in (2.30) then gives the right-hand side of (2.24).

The multivariable norm is thus evaluated from a knowledge of the single-variable result, which in turn relies on the following nontrivial ${}_5F_4(1)$ summation theorem[11]:

$${}_5F_4\left(\begin{matrix}2\alpha,\alpha+1,\alpha+\beta,\alpha+\gamma,\alpha+\delta\\ \alpha,\alpha-\beta+1,\alpha-\gamma+1,\alpha-\delta+1\end{matrix};1\right)=\frac{\Gamma(\alpha-\beta+1)\Gamma(\alpha-\gamma+1)\Gamma(\alpha-\delta+1)\Gamma(-\alpha-\beta-\gamma-\delta+1)}{\Gamma(2\alpha+1)\Gamma(-\beta-\gamma+1)\Gamma(-\beta-\delta+1)\Gamma(-\gamma-\delta+1)},$$
$$\mathrm{Re}(\alpha+\beta+\gamma+\delta)<1. \tag{2.31}$$

Thus the multiple Mellin–Barnes integral (2.24) is also associated with this same *single*-variable summation theorem. This is in complete analogy with the multivariable Appell weight function

$$w(x_1,x_2,\ldots,x_p) = x_1^{\nu_1}x_2^{\nu_2}\cdots x_p^{\nu_p}(1-x_1-x_2-\cdots-x_p)^{\nu_{p+1}}, \tag{2.32}$$

whose norm can be related, in an analogous manner, to that of the single-variable Jacobi weight $w(x)=x^{\alpha_1}(1-x)^{\alpha_2}$. The multivariable extensions of orthogonal polynomials discussed in this paper are, roughly speaking, along the same lines as the Appell, Lauricella, and Kampé de Fériet multivariable generalizations of hypergeometric functions.[8,9]

There are other quite different and important multivariable extensions of hypergeometric series. Holman, Biedenharn, and Louck[13] define a generalization which they call well-poised in $SU(n)$, $n\geqslant 2$, which are closely related to Racah and Wigner coefficients for $SU(n)$. Holman[14] defines a general hypergeometric series in $U(n)$ and proves extensions of several classical summation theorems including a terminating form of the ${}_5F_4(1)$ sum (2.31). Milne defines q-analogs of hypergeometric series well-poised in $SU(n)$ (see Ref. 15) as well as in $U(n)$ (see Ref. 16) and proves a number of important properties for them.[17–20] These include a $U(n)$ generalization[17] of the q-binomial theorem, a $SU(n)$ (see Ref. 18) extension of the q-analog of the ${}_5F_4(1)$ summa-

tion theorem (2.31), and also two U(n) (see Ref. 16) generalizations of the q-analog of (2.31). Further summation theorems and transformations of this kind are discussed by Gustafson,[21] who also introduces[22] a multivariable orthogonal (as opposed to biorthogonal) generalization of the Racah polynomials.

The discrete counterparts to the polynomials presented in this paper will be discussed in a future publication. These are associated with a different weight function than Gustafson's polynomials and are biorthogonal as opposed to orthogonal. The difference in these two families is a reflection of the different hypergeometric series to which they are related: the Kampé de Fériet series (2.5) in the first case and the U(n) series in Gustafson's case.

III. BIORTHOGONALITY

In this section we deduce the four biorthogonality relations satisfied by the multivariable Wilson polynomials. We begin by demonstrating that the inner product of $P_n(x)$ with $\overline{P}_m(x)$ vanishes if $N \neq M$.

Using representations (2.1) and (2.2) and the integral formula (2.24), one can easily deduce

$$\int_{-\infty}^{\infty} dx_p \cdots \int_{-\infty}^{\infty} dx_1 \, P_n(x)\overline{P}_m(x)w(x)$$
$$\equiv P_n \cdot \overline{P}_m = \beta_{nm} \sum_{\{j_k\}}\sum_{\{l_k\}} \left[\prod_{k=1}^{p} \binom{n_k}{j_k}\binom{m_k}{l_k} \frac{\Gamma(j_k+l_k+a_k+b_k)}{\Gamma(j_k+a_k+b_k)\Gamma(l_k+a_k+b_k)}\right]$$
$$\times \frac{\Gamma(N+J+A+B+c+d-1)\Gamma(M+L+A+B+c+d-1)}{\Gamma(J+L+A+B+c+d)}(-1)^{J+L}, \tag{3.1}$$

where β_{nm} is some constant. If we assume $N > M$ then we can write

$$\frac{\Gamma(N+J+A+B+c+d-1)}{\Gamma(J+L+A+B+c+d)} = \sum_{r=0}^{N-L-1} \xi_r J^r, \tag{3.2}$$

where ξ_r are some constants independent of J. Substituting this into (3.1) gives

$$P_n \cdot \overline{P}_m = \beta_{nm} \sum_{\{l_k\}} \Gamma(M+L+A+B+c+d-1)(-1)^L$$
$$\times \left[\prod_{k=1}^{p} \binom{m_k}{l_k} \frac{1}{\Gamma(l_k+a_k+b_k)}\right] \sum_{\{j_k\}} \sum_{r=0}^{N-L-1} \xi_r J^r \left[\prod_{k=1}^{p} \binom{n_k}{j_k} \frac{\Gamma(j_k+l_k+a_k+b_k)}{\Gamma(j_k+a_k+b_k)}\right](-1)^J, \tag{3.3}$$

and if we then introduce a set of real variables $z_1, z_2, \ldots, z_p$ this can be written as

$$P_n \cdot \overline{P}_m = \beta_{nm} \sum_{\{l_k\}} \Gamma(M+L+A+B+c+d-1)(-1)^L$$
$$\times \left[\prod_{k=1}^{p} \binom{m_k}{l_k} \frac{1}{\Gamma(l_k+a_k+b_k)} \left(\frac{\partial}{\partial z_k}\right)^{l_k}_{z_k=1} z_k^{l_k+a_k+b_k-1}\right] \sum_{\{j_k\}} \sum_{r=0}^{N-L-1} \xi_r J^r \left[\prod_{k=1}^{p} \binom{n_k}{j_k} (-z_k)^{j_k}\right]. \tag{3.4}$$

The $\{j_k\}$ sums can be performed as follows:

$$\sum_{\{j_k\}} \sum_{r=0}^{N-L-1} \xi_r J^r \left[\prod_{k=1}^{p} \binom{n_k}{j_k} (-z_k)^{j_k}\right] = \sum_{\{j_k\}} \sum_{r=0}^{N-L-1} \xi_r \left(\sum_{k=1}^{p} z_k \frac{\partial}{\partial z_k}\right)^r \left[\prod_{k=1}^{p} \binom{n_k}{j_k} (-z_k)^{j_k}\right]$$
$$= \sum_{r=0}^{N-L-1} \xi_r \left(\sum_{k=1}^{p} z_k \frac{\partial}{\partial z_k}\right)^r \sum_{\{j_k\}} \left[\prod_{k=1}^{p} \binom{n_k}{j_k} (-z_k)^{j_k}\right]$$
$$= \sum_{r=0}^{N-L-1} \xi_r \left(\sum_{k=1}^{p} z_k \frac{\partial}{\partial z_k}\right)^r \left[\prod_{k=1}^{p} (1-z_k)^{n_k}\right]. \tag{3.5}$$

Then the inner product becomes

$$P_n \cdot \overline{P}_m = \beta_{nm} \sum_{\{l_k\}} \Gamma(M+L+A+B+c+d-1)(-1)^L \left[\prod_{k=1}^{p} \binom{m_k}{l_k} \frac{1}{\Gamma(l_k+a_k+b_k)} \left(\frac{\partial}{\partial z_k}\right)^{l_k}_{z_k=1} z_k^{l_k+a_k+b_k-1}\right]$$
$$\times \sum_{r=0}^{N-L-1} \xi_r \left(\sum_{k=1}^{p} z_k \frac{\partial}{\partial z_k}\right)^r \left[\prod_{k=1}^{p} (1-z_k)^{n_k}\right]. \tag{3.6}$$

Clearly if any factor of $(1-z_k)$ survives the differentiations it will vanish upon setting $z_k = 1$. The total degree of these factors is N while the highest-order derivative acting on them is of order $N-1$. Thus at least one factor of $(1-z_k)$, for some k, will survive in every term after the differentiations and then will vanish upon setting $z_k = 1$. This demonstrates that $P_n \cdot \overline{P}_m$ vanishes for $N > M$, but then expression (3.1) is symmetric in n_k and m_k so the same argument follows for $M > N$, and thus

$$\int_{-\infty}^{\infty} dx_p \cdots \int_{-\infty}^{\infty} dx_1\, P_n(x)\overline{P}_m(x)w(x) = 0, \quad \text{if } N \neq M, \tag{3.7}$$

which, however, says nothing of polynomials of the same degree ($N = M$).

Next we demonstrate the analogous result for the other two families $Q_n(x)$ and $\overline{Q}_m(x)$. Using representations (2.12) and (2.13) and the integral formula (2.24), we obtain

$$\begin{aligned} Q_n \cdot \overline{Q}_m = \xi_{nm} \sum_{\{i_k\}}\sum_{\{r_k\}}\sum_{\{j_k\}}\sum_{\{l_k\}} \Bigg[\prod_{k=1}^{p} \binom{m_k}{i_k}\binom{m_k - i_k}{r_k}\binom{n_k}{j_k}\binom{n_k - j_k}{l_k} \\ \times \frac{\Gamma(i_k + j_k + a_k + b_k)}{\Gamma(i_k + a_k + b_k)\Gamma(j_k + a_k + b_k)} \Bigg] \frac{\Gamma(M + c + d)}{\Gamma(R + c + d)} \frac{\Gamma(N + c + d)}{\Gamma(L + c + d)} \\ \times \frac{\Gamma(I + L + A + c)\Gamma(J + R + B + d)\Gamma(L + R + c + d)}{\Gamma(I + R + J + L + A + B + c + d)} (-1)^{I+R+J+L}, \end{aligned} \tag{3.8}$$

where ξ_{nm} is some constant ($I \equiv \Sigma_{k=1}^{p} i_k$). If we substitute the identity

$$\begin{aligned} \frac{1}{\Gamma(I + R + J + L + A + B + c + d)} \\ = \frac{1}{\Gamma(N + I + R + A + B + c + d)} \sum_{\{t_k\}} \left[\prod_{k=1}^{p} \binom{n_k - j_k - l_k}{t_k} \right] \frac{\Gamma(I + N - J - T + A + c)}{\Gamma(I + L + A + c)} \frac{\Gamma(J + R + T + B + d)}{\Gamma(J + R + B + d)} \end{aligned} \tag{3.9}$$

($T \equiv \Sigma_{k=1}^{p} t_k$), which is deduced from (2.14), and then interchange the order of the $\{l_k\}$ and $\{t_k\}$ sums, (3.8) becomes

$$\begin{aligned} Q_n \cdot \overline{Q}_m = \xi_{nm} \sum_{\{i_k\}}\sum_{\{r_k\}}\sum_{\{j_k\}}\sum_{\{t_k\}}\sum_{\{l_k\}} \Bigg[\prod_{k=1}^{p} \binom{m_k}{i_k}\binom{m_k - i_k}{r_k}\binom{n_k}{j_k}\binom{n_k - j_k}{t_k}\binom{n_k - j_k - t_k}{l_k} \\ \times \frac{\Gamma(i_k + j_k + a_k + b_k)}{\Gamma(i_k + a_k + b_k)\Gamma(j_k + a_k + b_k)} \Bigg] \frac{\Gamma(M + c + d)}{\Gamma(R + c + d)} \frac{\Gamma(N + c + d)}{\Gamma(L + c + d)} \\ \times \frac{\Gamma(I + N - J - T + A + c)\Gamma(J + R + T + B + d)\Gamma(L + R + c + d)}{\Gamma(N + I + R + A + B + c + d)} (-1)^{I+R+J+L}. \end{aligned} \tag{3.10}$$

The $\{l_k\}$ sum is now performed by using (2.14) giving

$$\begin{aligned} Q_n \cdot \overline{Q}_m = \xi_{nm} \sum_{\{i_k\}}\sum_{\{r_k\}}\sum_{\{j_k\}}\sum_{\{t_k\}} \Bigg[\prod_{k=1}^{p} \binom{m_k}{i_k}\binom{m_k - i_k}{r_k}\binom{n_k}{j_k}\binom{n_k - j_k}{t_k} \\ \times \frac{\Gamma(i_k + j_k + a_k + b_k)}{\Gamma(i_k + a_k + b_k)\Gamma(j_k + a_k + b_k)} \Bigg] \frac{R!}{(R - N + J + T)!} \frac{\Gamma(M + c + d)}{\Gamma(N - J - T + c + d)} \\ \times \frac{\Gamma(I + N - J - T + A + c)\Gamma(J + R + T + B + d)}{\Gamma(N + I + R + A + B + c + d)} (-1)^{I+R+T}, \end{aligned} \tag{3.11}$$

where ξ_{nm} has been redefined. We reverse the $\{t_k\}$ sums, $t_k \to n_k - j_k - t_k$, and then interchange the order of the $\{j_k\}$ and $\{t_k\}$ summations to obtain

$$\begin{aligned} Q_n \cdot \overline{Q}_m = \xi_{nm} \sum_{\{i_k\}}\sum_{\{r_k\}}\sum_{\{t_k\}}\sum_{\{j_k\}} \Bigg[\prod_{k=1}^{p} \binom{m_k}{i_k}\binom{m_k - i_k}{r_k}\binom{n_k}{t_k}\binom{n_k - t_k}{j_k} \\ \times \frac{\Gamma(i_k + j_k + a_k + b_k)}{\Gamma(i_k + a_k + b_k)\Gamma(j_k + a_k + b_k)} \Bigg] \frac{R!}{(R - T)!} \frac{\Gamma(M + c + d)}{\Gamma(T + c + d)} \\ \times \frac{\Gamma(I + T + A + c)\Gamma(N + R - T + B + d)}{\Gamma(N + I + R + A + B + c + d)} (-1)^{N+I+R+J+T}. \end{aligned} \tag{3.12}$$

Then the $\{j_k\}$ sums are performed by the Chu–Vandermonde theorem (2.22), yielding

$$\begin{aligned} Q_n \cdot \overline{Q}_m = \xi_{nm} \sum_{\{i_k\}}\sum_{\{r_k\}}\sum_{\{t_k\}} \Bigg[\prod_{k=1}^{p} \binom{m_k}{i_k}\binom{m_k - i_k}{r_k}\binom{n_k}{t_k} \frac{i_k!}{\Gamma(n_k - t_k + a_k + b_k)(i_k - n_k + t_k)!} \Bigg] \\ \times \frac{R!}{(R - T)!} \frac{\Gamma(M + c + d)}{\Gamma(T + c + d)} \frac{\Gamma(I + T + A + c)\Gamma(N + R - T + B + d)}{\Gamma(N + I + R + A + B + c + d)} (-1)^{I+R}. \end{aligned} \tag{3.13}$$

Now the $\{i_k\}$ sum is brought on the inside, the indices are redefined $i_k \to i_k + n_k - t_k$, and the sum is performed by using (2.14) again, resulting in

$$Q_n\cdot\overline{Q}_m=\xi_{nm}\sum_{\{r_k\}}\sum_{\{t_k\}}\left[\prod_{k=1}^{p}\binom{m_k}{r_k}\binom{n_k}{t_k}\frac{\Gamma(a_k+b_k)}{\Gamma(n_k-t_k+a_k+b_k)}\frac{(m_k-r_k)!}{(m_k-n_k-r_k+t_k)!}\right]$$
$$\times\frac{R!}{(R-T)!}\frac{\Gamma(M+c+d)}{\Gamma(T+c+d)}(-1)^{R+T},\tag{3.14}$$

where ξ_{nm} has been redefined again. Finally we interchange the remaining two sums and use (2.14) once more to obtain

$$Q_n\cdot\overline{Q}_m=\xi_{nm}\sum_{\{t_k\}}\left[\prod_{k=1}^{p}\binom{n_k}{t_k}\frac{\Gamma(a_k+b_k)}{\Gamma(n_k-t_k+a_k+b_k)}\frac{m_k!}{(m_k-n_k+t_k)!}\right]\frac{1}{(M-N)}\frac{\Gamma(M+c+d)}{\Gamma(T+c+d)}\frac{1}{\Gamma(-T)}(-1)^T,\tag{3.15}$$

where we are assuming $N\neq M$. Clearly every term in this sum vanishes and thus

$$\int_{-\infty}^{\infty}dx_p\cdots\int_{-\infty}^{\infty}dx_1\,Q_n(x)\overline{Q}_m(x)w(x)=0,\quad\text{if } N\neq M,\tag{3.16}$$

in analogy with (3.7).

Next we show that the two families $P_n(x)$ and $Q_n(x)$ are biorthogonal in all the indices $n_1,n_2,\ldots,n_p$. From (2.1), (2.12), and the multiple integral formula (2.24), we deduce

$$P_n\cdot Q_m=2(2\pi)^p\left[\prod_{k=1}^{p}\Gamma(n_k+a_k+b_k)\Gamma(m_k+a_k+b_k)\right]\Gamma(N+A+d)\Gamma(M+B+c)\Gamma(M+c+d)$$
$$\times\sum_{\{j_k\}}\sum_{\{i_k\}}\sum_{\{l_k\}}\left[\prod_{k=1}^{p}\binom{n_k}{j_k}\binom{m_k}{i_k}\binom{m_k-i_k}{l_k}\frac{\Gamma(i_k+j_k+a_k+b_k)}{\Gamma(i_k+a_k+b_k)\Gamma(j_k+a_k+b_k)}\right]$$
$$\times\frac{\Gamma(N+A+c)}{\Gamma(J+A+c)}\frac{\Gamma(N+J+A+B+c+d-1)}{\Gamma(N+A+B+c+d-1)}\frac{\Gamma(J+L+A+c)\Gamma(I+B+d)}{\Gamma(I+J+L+A+B+c+d)}(-1)^{I+J+L}.\tag{3.17}$$

Using (2.14) to sum the $\{l_k\}$ summations one finds

$$P_n\cdot Q_m=2(2\pi)^p\left[\prod_{k=1}^{p}\Gamma(n_k+a_k+b_k)\Gamma(m_k+a_k+b_k)\right]\Gamma(N+A+c)\Gamma(N+A+d)\Gamma(M+B+c)\Gamma(M+B+d)$$
$$\times\sum_{\{j_k\}}\sum_{\{i_k\}}\left[\prod_{k=1}^{p}\binom{n_k}{j_k}\binom{m_k}{i_k}\frac{\Gamma(i_k+j_k+a_k+b_k)}{\Gamma(i_k+a_k+b_k)\Gamma(j_k+a_k+b_k)}\right]$$
$$\times\frac{\Gamma(N+J+A+B+c+d-1)}{\Gamma(N+A+B+c+d-1)}\frac{\Gamma(M+c+d)}{\Gamma(M+J+A+B+c+d)}(-1)^{I+J}.\tag{3.18}$$

Then the Chu–Vandermonde theorem (2.22) is applied to the $\{i_k\}$ sums, resulting in

$$P_n\cdot Q_m=2(2\pi)^p\left[\prod_{k=1}^{p}\Gamma(n_k+a_k+b_k)\right]\Gamma(N+A+c)\Gamma(N+A+d)\Gamma(M+B+c)\Gamma(M+B+d)$$
$$\times\sum_{\{j_k\}}\left[\prod_{k=1}^{p}\binom{n_k}{j_k}\frac{j_k!}{(j_k-m_k)!}\right]\frac{\Gamma(N+J+A+B+c+d-1)}{\Gamma(N+A+B+c+d-1)}\frac{\Gamma(M+c+d)}{\Gamma(M+J+A+B+c+d)}(-1)^{J-M}.\tag{3.19}$$

Finally, we redefine the summation indices, $j_k\to j_k+m_k$, and apply (2.14) once more to obtain

$$P_n\cdot Q_m=2(2\pi)^p\left[\prod_{k=1}^{p}\Gamma(n_k+a_k+b_k)\frac{n_k!}{(n_k-m_k)!}\right]\frac{1}{(M-N)!}$$
$$\times\frac{\Gamma(N+A+c)\Gamma(N+A+d)\Gamma(M+B+c)\Gamma(M+B+d)\Gamma(M+c+d)}{(N+M+A+B+c+d-1)\Gamma(N+A+B+c+d-1)},\tag{3.20}$$

which is clearly zero unless $n_k=m_k$ for every k; that is,

$$\int_{-\infty}^{\infty}dx_p\cdots\int_{-\infty}^{\infty}dx_1\,P_n(x)Q_m(x)w(x)=h_n\prod_{k=1}^{p}\delta_{n_km_k},\tag{3.21}$$

where the normalization constant is given by

$$h_n=2(2\pi)^p\left[\prod_{k=1}^{p}\Gamma(n_k+a_k+b_k)n_k!\right]\frac{\Gamma(N+A+c)\Gamma(N+A+d)\Gamma(N+B+c)\Gamma(N+B+d)\Gamma(N+c+d)}{(2N+A+B+c+d-1)\Gamma(N+A+B+c+d-1)}.\tag{3.22}$$

Applying either of the transformations in (2.8) and (2.9) to (3.21) then gives

$$\int_{-\infty}^{\infty} dx_p \cdots \int_{-\infty}^{\infty} dx_1\, \overline{P}_n(x)\overline{Q}_m(x)w(x) = h_n \prod_{k=1}^{p} \delta_{n_k m_k}, \tag{3.23}$$

where h_n is again given by (3.22).

The four biorthogonality relations are schematically depicted below

$$\begin{array}{ccc} P_n(x) & \Leftrightarrow & Q_n(x) \\ \updownarrow & & \updownarrow \\ \overline{P}_n(x) & \Leftrightarrow & \overline{Q}_n(x) \end{array}, \tag{3.24}$$

where the horizontal arrows denote biorthogonality in all the indices $n_1,n_2,\ldots,n_p$ while the vertical arrows denote biorthogonality only for polynomials of different degrees $N \neq M$.

IV. SPECIAL AND LIMIT CASES

Analogous to the single-variable case there exists a limit to the multivariable continuous Hahn polynomials. These are obtained by setting

$$a_k = a'_k + \tfrac{1}{2}iw_k, \quad b_k = b'_k - \tfrac{1}{2}iw_k, \quad c = c' + \tfrac{1}{2}iW, \quad d = d' - \tfrac{1}{2}iW, \quad x_k = x'_k - \tfrac{1}{2}w_k, \quad W \equiv \sum_{k=1}^{p} w_k, \tag{4.1}$$

dividing each polynomial family by W^N, dividing the weight function by $\Gamma(A+c)\Gamma(B+d)$, and then taking the limit $W \to \infty$, in which manner one finds

$$\begin{aligned} &\lim_{W\to\infty} W^{-N} P_n(x) = \lim_{W\to\infty} W^{-N} \overline{P}_n(x) = \left[\prod_{k=1}^{p} n_k!\right] H_n(x'), \\ &\lim_{W\to\infty} W^{-N} Q_n(x) = \lim_{W\to\infty} W^{-N} \overline{Q}_n(x) = \left[\prod_{k=1}^{p} n_k!\right] \overline{H}_n(x'), \\ &\lim_{W\to\infty} \frac{w(x)}{\Gamma(A+c)\Gamma(B+d)} = \left[\prod_{k=1}^{p} \Gamma(a'_k + ix'_k)\Gamma(b'_k - ix'_k)\right] \Gamma(c' + iX')\Gamma(d' - iX'). \end{aligned} \tag{4.2}$$

In particular, both $P_n(x)$ and $\overline{P}_n(x)$ reduce to the same family of multivariable Hahn polynomials $H_n(x')$. Similarly both $Q_n(x)$ and $\overline{Q}_n(x)$ reduce to the same family of biorthogonal counterparts $\overline{H}_n(x')$. Then from (3.24) we deduce that each of the families $H_n(x')$ and $\overline{H}_m(x')$ are orthogonal with themselves only for different degrees $N \neq M$, and biorthogonal to each other in all the indices $n_1,n_2,\ldots,n_p$. Also in the special case of (2.10) we find that both families are real. These polynomials and their discrete analogs have already been discussed in detail.[23,24]

Another interesting limit case, not yet known, is the multivariable continuous dual Hahn polynomials. These result upon dividing the polynomial families by d^N, dividing the weight function by $\Gamma^2(d)$, and then taking the limit $d \to \infty$. In this manner one finds

$$\begin{aligned} &\lim_{d\to\infty} d^{-N} P_n(x) = \lim_{d\to\infty} d^{-N} \overline{Q}_n(x) = D_n(x), \\ &\lim_{d\to\infty} d^{-N} \overline{P}_n(x) = \lim_{d\to\infty} d^{-N} Q_n(x) = \overline{D}_n(x), \\ &\lim_{d\to\infty} \frac{w(x)}{\Gamma^2(d)} = \left[\prod_{k=1}^{p} \Gamma(a_k + ix_k)\Gamma(b_k - ix_k)\right] \frac{\Gamma(A-iX)\Gamma(B+iX)\Gamma(c+iX)\Gamma(c-iX)}{\Gamma(2iX)\Gamma(-2iX)}, \end{aligned} \tag{4.3}$$

where now the two families $P_n(x)$ and $\overline{Q}_n(x)$ both reduce to the same continuous dual Hahn polynomials $D_n(x)$, whereas $\overline{P}_n(x)$ and $Q_n(x)$ both reduce to the same biorthogonal counterparts $\overline{D}_n(x)$. These are given by

$$D_n(x) = \left[\prod_{k=1}^{p} \frac{\Gamma(n_k + a_k + b_k)}{\Gamma(a_k + b_k)}\right] (A+c)_N F^{1:2;\ldots;2}_{1:1;\ldots;1}\left(\begin{matrix} A - iX: -n_1, a_1 + ix_1; \ldots; -n_p, a_p + ix_p \\ A + c: a_1 + b_1; \ldots; a_p + b_p \end{matrix}\right), \tag{4.4}$$

$$\overline{D}_n(x) = \left[\prod_{k=1}^{p} \frac{\Gamma(n_k + a_k + b_k)}{\Gamma(a_k + b_k)}\right] (B+c)_N F^{1:2;\ldots;2}_{1:1;\ldots;1}\left(\begin{matrix} B + iX: -n_1, b_1 - ix_1; \ldots; -n_p, b_p - ix_p \\ B + c: a_1 + b_1; \ldots; a_p + b_p \end{matrix}\right), \tag{4.5}$$

and the four biorthogonalities satisfied by the Wilson polynomials, in this limit, imply the single relation

$$\int_{-\infty}^{\infty} dx_p \cdots \int_{-\infty}^{\infty} dx_1\, D_n(x)\overline{D}_m(x)w(x) = h_n \prod_{k=1}^{p} \delta_{n_k m_k}, \tag{4.6}$$

where h_n is now given by

$$h_n = 2(2\pi)^p \left[\prod_{k=1}^{p} \Gamma(n_k + a_k + b_k) n_k!\right] \Gamma(N+A+c)\Gamma(N+B+c). \tag{4.7}$$

In the special case, $a_k = b_k^*$, $k = 1,2,\ldots,p$, $c = c^*$, the weight function is real and positive and the barred polynomials are

simply complex conjugates of the unbarred polynomials, $\bar{D}_n(x) = D_n^*(x)$; thus in this special case (4.6) describes a conventional orthogonality relation.

Returning to the multivariable Wilson polynomials we consider a few specific cases of the parameters for which the weight function takes on interesting forms. If $a_k = b_k = l_k + 1$, $c = l_{p+1} + 1$, $d = l_{p+2} + 1$, where $l_1, l_2, \ldots, l_p, l_{p+1}, l_{p+2}$ ($L \equiv \Sigma_{k=1}^{p} l_k$), are non-negative integers, then, using known properties of the gamma function,[10] we can write, apart from a multiplicative constant,

$$w(x) = \left[\prod_{k=1}^{p} \frac{x_k}{\sinh(\pi x_k)} \prod_{j_k=1}^{l_k} \left\{1 + \frac{x_k^2}{j_k^2}\right\}\right] X^4 \frac{\cosh(\pi X)}{\sinh^2(\pi X)} \left[\prod_{j=1}^{L+p-1} \left\{1 + \frac{X^2}{j^2}\right\}\right] \left[\prod_{j_{p+1}=1}^{l_{p+1}} \left\{1 + \frac{X^2}{j_{p+1}^2}\right\}\right] \left[\prod_{j_{p+2}=1}^{l_{p+2}} \left\{1 + \frac{X^2}{j_{p+2}^2}\right\}\right], \tag{4.8}$$

which in the simplest case, $l_1 = l_2 = \cdots = l_p = l_{p+1} = l_{p+2} = 0$, just becomes

$$w(x) = \left[\prod_{k=1}^{p} \frac{x_k}{\sinh(\pi x_k)}\right] X^4 \frac{\cosh(\pi X)}{\sinh^2(\pi X)} \left[\prod_{j=1}^{p-1} \left\{1 + \frac{X^2}{j^2}\right\}\right]. \tag{4.9}$$

If, on the other hand, $a_k = b_k = l_k + \frac{1}{2}$, $c = l_{p+1} + \frac{1}{2}$, $d = l_{p+2} + \frac{1}{2}$, then, apart from a multiplicative constant,

$$\begin{aligned} w(x) = {} & \left[\prod_{k=1}^{p} \operatorname{sech}(\pi x_k) \prod_{j_k=1}^{l_k} \left\{1 + \frac{x_k^2}{(j_k - \frac{1}{2})^2}\right\}\right] X^2 \operatorname{sech}(\pi X) \\ & \times \left[\prod_{j=1}^{L+(1/2)p-1} \left\{1 + \frac{X^2}{j^2}\right\}\right] \left[\prod_{j_{p+1}=1}^{l_{p+1}} \left\{1 + \frac{X^2}{(j_{p+1} - \frac{1}{2})^2}\right\}\right] \left[\prod_{j_{p+2}=1}^{l_{p+2}} \left\{1 + \frac{X^2}{(j_{p+2} - \frac{1}{2})^2}\right\}\right], \end{aligned} \tag{4.10}$$

for p even, and

$$\begin{aligned} w(x) = {} & \left[\prod_{k=1}^{p} \operatorname{sech}(\pi x_k) \prod_{j_k=1}^{l_k} \left\{1 + \frac{x_k^2}{(j_k - \frac{1}{2})^2}\right\}\right] X \sinh(\pi X) \operatorname{sech}^2(\pi X) \\ & \times \left[\prod_{j=1}^{L+(1/2)p-1/2} \left\{1 + \frac{X^2}{(j - \frac{1}{2})^2}\right\}\right] \left[\prod_{j_{p+1}=1}^{l_{p+1}} \left\{1 + \frac{X^2}{(j_{p+1} - \frac{1}{2})^2}\right\}\right] \left[\prod_{j_{p+2}=1}^{l_{p+2}} \left\{1 + \frac{X^2}{(j_{p+2} - \frac{1}{2})^2}\right\}\right], \end{aligned} \tag{4.11}$$

for p odd, which in the simplest case, respectively, reduce to

$$\begin{aligned} w(x) &= \left[\prod_{k=1}^{p} \operatorname{sech}(\pi x_k)\right] X^2 \operatorname{sech}(\pi X) \left[\prod_{j=1}^{(1/2)p-1} \left\{1 + \frac{X^2}{j^2}\right\}\right], \\ w(x) &= \left[\prod_{k=1}^{p} \operatorname{sech}(\pi x_k)\right] X \sinh(\pi X) \operatorname{sech}^2(\pi X) \left[\prod_{j=1}^{(1/2)p-1/2} \left\{1 + \frac{X^2}{(j - \frac{1}{2})^2}\right\}\right]. \end{aligned} \tag{4.12}$$

V. DISCUSSION

That these polynomials do not form a complete set is obvious since they are of even degree, but a given family is not complete even for even degree polynomials. One easily finds by considering a few simple cases that $P_n(x)$ and $\overline{P}_n(x)$ cannot be expanded in terms of one another, and similarly for $Q_n(x)$ and $\overline{Q}_n(x)$, which if possible would have implied further orthogonalities. It appears the only pairs that can be expanded in terms of each other are $P_n(x)$ and $\overline{Q}_n(x)$ for one, and $\overline{P}_n(x)$ and $Q_n(x)$ for the other, which is consistent with the known biorthogonalities. The expansion coefficients are themselves related to a family of discrete biorthogonal polynomials, which we will consider in a future publication in the more general context discussed below.

There are several extensions of these multivariable Wilson polynomials currently being investigated. Relaxing the restriction that the real parts of the parameters are greater than zero leads to biorthogonalities that are partly continuous and partly discrete or completely discrete; the latter case being the multivariable biorthogonal Racah polynomials. It also appears that these polynomials have a natural q-extension as in the single variable case.[25,26]

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the United States Department of Energy.

[1] J. A. Wilson, SIAM J. Math. Anal. **11**, 690 (1980).
[2] R. Askey and J. Wilson, SIAM J. Math. Anal. **13**, 651 (1982).
[3] C. F. Dunkl, Pac. J. Math. **92**, 57 (1981).
[4] J. Letessier and G. Valent, SIAM J. Appl. Math. **46**, 393 (1986).
[5] W. Miller, SIAM J. Math. Anal. **18**, 1221 (1987).
[6] K. S. Rao, T. S. Santhanam, and R. A. Gustafson, J. Phys. A **20**, 3041 (1987).
[7] E. Montaldi and G. Zucchelli, J. Math. Phys. **28**, 2040 (1987).
[8] H. M. Srivastava and Per W. Karlsson, *Multiple Gaussian Hypergeometric Series* (Ellis Horwood, West Sussex, 1985), p. 38.
[9] P. Appell and J. Kampé de Fériet, *Fonctions Hypergéométriques et Hypersphériques* (Gauthier-Villars, Paris, 1926).
[10] A. Erdelyi, A. Magnus, F. Oberhettinger, and F. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1955), Vols I–III.
[11] W. N. Bailey, *Generalized Hypergeometric Series* (Cambridge U.P., London, 1935).
[12] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980), p. 655.
[13] W. J. Holman, III, L. C. Biedenharn, and J. D. Louck, SIAM J. Math. Anal. **7**, 529 (1976).
[14] W. J. Holman, III, SIAM J. Math. Anal. **11**, 523 (1980).

[15]S. C. Milne, Adv. Math. **58**, 1 (1985).
[16]S. C. Milne, J. Math. Anal. Appl. **122**, 223 (1987).
[17]S. C. Milne, Adv. Math. **57**, 34 (1985).
[18]S. C. Milne, Adv. Math. **57**, 14 (1985).
[19]S. C. Milne, Adv. Math. **57**, 71 (1985).
[20]S. C. Milne, J. Math. Anal. Appl. **118**, 263 (1986).
[21]R. A. Gustafson, SIAM J. Math. Anal. **18**, 1576 (1987).
[22]R. A. Gustafson, SIAM J. Math. Anal. **18**, 495 (1987).
[23]M. V. Tratnik, J. Math. Phys. **29**, 1529 (1988).
[24]M. V. Tratnik, J. Math. Phys. **30**, 627 (1989).
[25]R. Askey and J. Wilson, SIAM J. Math. Anal. **10**, 1008 (1979).
[26]R. Askey and J. Wilson, Mem. Am. Math. **54**, Num. 319 (1985).

Additive decomposition for the product of two θ_3 functions and modular equations

Emilio Montaldi
Dipartimento di Fisica, Universita' di Milano, Istituto Nazionale di Fisica Nucleare, Sezione di Milano via Celoria 16, 20133 Milano, Italy

Giuseppe Zucchelli
Centro CNR, Dipartimento di Biologia, Universita' di Milano via Celoria 26, 20133 Milano, Italy

(Received 3 January 1989; accepted for publication 12 April 1989)

By using a general additive decomposition for the product of two θ_3 functions, a simple and unified derivation of the modular equations of degree 3, 5, and 7 is given.

I. INTRODUCTION

Some years ago, Boon *et al.*[1] showed that the description of harmonic oscillator states on rational von Neumann lattices in the kq representation leads to the following *additive* decomposition of $\theta_3(nx|n\tau)$:

$$n\theta_3(nx|n\tau)=\sum_{r=0}^{n-1}\theta_3\left(x+\frac{r\pi}{n}\middle|\frac{\tau}{n}\right). \tag{1.1}$$

Corresponding results for the other θ functions were given later,[2] together with a simple direct proof based on the definition of such functions as infinite series; furthermore, it was pointed out that similar additive decompositions may be found for powers of θ functions. For instance, one has

$$2n\theta_3(0|\tau)\theta_3(nx|n^2\tau)=\sum_{r=0}^{2n-1}\theta_3^2\left(\frac{x}{2}+\frac{r\pi}{2n}\middle|\frac{\tau}{2}\right). \tag{1.2}$$

In particular, for $n=1$ and $x=0$, this identity becomes

$$2\theta_3^2(0,q^2)=\theta_3^2(0,q)+\theta_4^2(0,q) \tag{1.3}$$

(as usual, $q=e^{i\pi\tau}$). It is interesting to observe that (1.3) is nothing but a relation between moduli of Jacobian elliptic functions, namely,

$$\tfrac{1}{2}k_2k_2'=k^2k'^{1/2}(1+k')^{-3}, \tag{1.4}$$

where

$$k^{1/2}=\frac{\theta_2(0,q)}{\theta_3(0,q)},\quad k'^{1/2}=\frac{\theta_4(0,q)}{\theta_3(0,q)}, \tag{1.5}$$

$$k_2^{1/2}=\frac{\theta_2(0,q^2)}{\theta_3(0,q^2)},\quad k_2'^{1/2}=\frac{\theta_4(0,q^2)}{\theta_3(0,q^2)}. \tag{1.6}$$

To see this, we recall that[3]

$$\prod_{n=1}^{\infty}(1-q^{2n-1})^6=2q^{1/4}k^{-1/2}k', \tag{1.7}$$

$$\prod_{n=1}^{\infty}(1+q^{2n-1})^6=2q^{1/4}(kk')^{-1/2}, \tag{1.8}$$

$$\prod_{n=1}^{\infty}(1+q^{2n})^6=\frac{1}{4}q^{-1/2}kk'^{-1/2}, \tag{1.9}$$

$$\theta_3(0,q)=G(q)\prod_{n=1}^{\infty}(1+q^{2n-1})^2, \tag{1.10}$$

$$\theta_4(0,q)=G(q)\prod_{n=1}^{\infty}(1-q^{2n-1})^2, \tag{1.11}$$

$$G(q)=\prod_{n=1}^{\infty}(1-q^{2n}). \tag{1.12}$$

From Eqs. (1.8) and (1.10), we have

$$\theta_3(0,q)=G(q)(2q^{1/4}k^{-1/2}k'^{-1/2})^{1/3}; \tag{1.13}$$

similarly, from (1.7) and (1.11),

$$\theta_4(0,q)=G(q)(2q^{1/4}k^{-1/2}k')^{1/3}. \tag{1.14}$$

Furthermore, from (1.12) and (1.9),

$$G(q^2)=G(q)(\tfrac{1}{4}q^{-1/2}kk'^{-1/2})^{1/6}, \tag{1.15}$$

so that, if we replace q by q^2, Eq. (1.13) becomes

$$\theta_3(o,q^2)=G(q)(\tfrac{1}{4}q^{-1/2}kk'^{-1/2})^{1/6}\times(2q^{1/2}k_2^{-1/2}k_2'^{-1/2})^{1/3} \tag{1.16}$$

and, by inserting (1.13), (1.14), and (1.16) into (1.3), Eq. (1.4) follows at once. Conversely since[3]

$$\prod_{n=1}^{\infty}(1+q^n)^6=\frac{1}{2}q^{-1/4}k^{1/2}k'^{-1}, \tag{1.17}$$

or, with q replaced by q^2,

$$\prod_{n=1}^{\infty}(1+q^{2n})^6=\frac{1}{2}q^{-1/2}k_2^{1/2}k_2'^{-1}, \tag{1.18}$$

we have, by comparing (1.18) with (1.9),

$$k_2^{1/2}k_2'^{-1}=\frac{1}{2}kk'^{-1/2}, \tag{1.19}$$

whence (recall that $k^2+k'^2=k_2^2+k_2'^2=1$)

$$k_2=[(1-k')/k]^2,\quad k_2'=2k^{-2}k'^{1/2}(1-k'). \tag{1.20}$$

From this, we recover Eq. (1.4) and hence, by performing the above calculations in the reverse order, Eq. (1.3).

The aim of this paper is to provide a simple proof of the formulas[4]

$$(kk_3)^{1/2}+(k'k_3')^{1/2}=1, \tag{1.21}$$

$$k^{1/2}-k_5^{1/2}=2^{2/3}(kk_5)^{1/12}(k'k_5')^{1/3}, \tag{1.22}$$

$$(kk_7)^{1/4}+(k'k_7')^{1/4}=1, \tag{1.23}$$

where

$$k_\nu^{1/2}=\frac{\theta_2(0,q^\nu)}{\theta_3(0,q^\nu)},\quad k_\nu'^{1/2}=\frac{\theta_4(0,q^\nu)}{\theta_3(0,q^\nu)}. \tag{1.24}$$

 0022-2488/89/092012-04$02.50

The only tool needed in our derivation is the following additive decomposition for the product of two θ_3 functions:

$$\theta_3(x|r\tau)\theta_3(y|s\tau) = \sum_{l=0}^{r+s-1} e^{il^2s\pi\tau + 2ily}\theta_3(x+y+ls\pi\tau|(r+s)\tau) \times\theta_3(ry-sx+lrs\pi\tau|rs(r+s)\tau) \tag{1.25}$$

r and s being positive integers.

This equation is a simple consequence of the (fairly obvious) series rearrangement property[5]

$$\sum_{n=-\infty}^{\infty} c_n = \sum_{l=0}^{k-1}\sum_{n=-\infty}^{\infty} c_{m+kn+l}, \tag{1.26}$$

where k is a positive integer and m an arbitrary integer; indeed, one has

$$\theta_3(x|r\tau)\theta_3(y|s\tau) = \sum_{m,n=-\infty}^{\infty} \exp(im^2r\pi\tau + in^2s\pi\tau + 2imx + 2iny)$$

$$= \sum_{l=0}^{r+s-1}\sum_{m,n=-\infty}^{\infty} \times\exp\{im^2r\pi\tau + i[m+(r+s)n+l]^2s\pi\tau + 2imx + 2i[m+(r+s)n+l]y\},$$

i.e., with $m\to m-sn$:

$$\theta_3(x|r\tau)\theta_3(y|s\tau) = \sum_{l=0}^{r+s-1} \exp(il^2s\pi\tau + 2ily) \times\sum_{m,n=-\infty}^{\infty} \exp\{i\pi\tau[(r+s)m^2 + 2lsm + rs(r+s)n^2 + 2rsln] + 2i[m(x+y)+n(ry-sx)]\},$$

which is just (1.25).

We are now ready to prove Eqs. (1.21)–(1.23).

II. PROOF OF (1.21)

In the general formula (1.25), we put $r=3$, $s=1$, $x=9\pi\tau/2$, and $y=3\pi\tau/2$. By using[3]

$$\theta_3(x+\pi\tau/2|\tau) = q^{-1/4}e^{-ix}\theta_2(x|\tau), \tag{2.1}$$

$$\theta_3(x+\pi\tau|\tau) = q^{-1}e^{-2ix}\theta_3(x|\tau), \tag{2.2}$$

it is easily seen that

$$\theta_3(3\pi\tau/2|\tau) = q^{-9/4}\theta_2(0|\tau),$$

$$\theta_3(9\pi\tau/2|3\tau) = q^{-27/4}\theta_2(0|3\tau),$$

$$\theta_3(6\pi\tau + l\pi\tau|4\tau) = q^{-(3l+9)}\theta_2(l\pi\tau|4\tau),$$

so that, in the present case, Eq. (1.25) takes the form

$$\theta_2(0|\tau)\theta_2(0|3\tau) = \sum_{l=0}^{3} e^{il^2\pi\tau}\theta_2(l\pi\tau|4\tau)\theta_3(3l\pi\tau|12\tau) \tag{2.3}$$

or also[6]

$$\theta_2(0|\tau)\theta_2(0|3\tau) = \frac{1}{4}\sum_{l=0}^{3} e^{il^2\pi\tau}\left[\theta_3\left(\frac{l\pi\tau}{2}|\tau\right) - \theta_4\left(\frac{l\pi\tau}{2}|\tau\right)\right] \cdot\left[\theta_3\left(\frac{3l\pi\tau}{2}|3\tau\right) + \theta_4\left(\frac{3l\pi\tau}{2}|3\tau\right)\right]. \tag{2.4}$$

If we write explicitly on the rhs, considerable simplifications occur as a result of the quasidoubly periodic character of the θ functions. A straightforward calculation gives

$$\theta_2(0|\tau)\theta_2(0|3\tau) = \frac{1}{2}[\theta_3(0|\tau)\theta_3(0|3\tau) - \theta_4(0|\tau)\theta_4(0|3\tau) + \theta_2(0|\tau)\theta_2(0|3\tau)],$$

whence

$$\theta_2(0|\tau)\theta_2(0|3\tau) + \theta_4(0|\tau)\theta_4(0|3\tau) = \theta_3(0|\tau)\theta_3(0|3\tau), \tag{2.5}$$

i.e., by recalling (1.5) and (1.24) with $\nu=3$,

$$(kk_3)^{1/2} + (k'k_3')^{1/2} = 1. \tag{2.6}$$

Q.E.D.

As an interesting application, let us consider the double series[7]

$$S(\alpha) = \sum_{m,n=-\infty}^{\infty} (-1)^{(m+1)(n+1)} e^{-\alpha(m^2+n^2+mn)} = \sum_{m,n=-\infty}^{\infty} (-1)^{(n+1)} e^{-\alpha[3m^2+(n+m)^2]} + e^{-\alpha} \times\sum_{m,n=-\infty}^{\infty} \exp(-\alpha[3m(m+1)+(n+m)(n+m+1)]) = -\sum_{m,n=-\infty}^{\infty} (-1)^{(n-m)} e^{-\alpha(3m^2+n^2)} + e^{-\alpha} \times\sum_{m,n=-\infty}^{\infty} \exp(-\alpha[3m(m+1)+n(n+1)]) = -\theta_4(0,q)\theta_4(0,q^3) + \theta_2(0,q)\theta_2(0,q^3), \quad q=e^{-\alpha}. \tag{2.7}$$

For $\alpha=\pi/\sqrt{3}$, we have[8]

$$k = (\sqrt{3}+1)/2\sqrt{2} = k_3', \quad k' = (\sqrt{3}-1)/2\sqrt{2} = k_3 \tag{2.8}$$

and thus $(kk_3)^{1/2} = (k'k_3')^{1/2} = \frac{1}{2}$, i.e.,

$$S(\pi/\sqrt{3}) = 0. \tag{2.9}$$

The same conclusion can be reached by using Jacobi's imaginary transformation.

III. PROOF OF (1.22)

In the general formula (1.25), we put $r=5$, $s=1$, $x=0$, and $y=\pi\tau/2$. By using (2.1) and (2.2), we first have

$$q^{-1/4}\theta_2(0|\tau)\theta_3(0|5\tau)$$
$$=2\theta_2\left(\frac{\pi\tau}{2}|6\tau\right)\theta_2\left(\frac{5\pi\tau}{2}|30\tau\right)$$
$$+2\theta_3\left(\frac{\pi\tau}{2}|6\tau\right)\theta_3\left(\frac{5\pi\tau}{2}|30\tau\right)$$
$$+2q^2\theta_3\left(\frac{3\pi\tau}{2}|6\tau\right)\theta_3\left(\frac{15\pi\tau}{2}|30\tau\right). \tag{3.1}$$

On the other hand, if we put $r=5$, $s=1$, $x=5\pi\tau/2$, and $y=0$, we get

$$q^{-1/4}\theta_2(0|5\tau)\theta_3(0|\tau)$$
$$=2\theta_2\left(\frac{\pi\tau}{2}|6\tau\right)\theta_3\left(\frac{5\pi\tau}{2}|30\tau\right)$$
$$+2\theta_3\left(\frac{\pi\tau}{2}|6\tau\right)\theta_2\left(\frac{5\pi\tau}{2}|30\tau\right)$$
$$+2q^2\theta_3\left(\frac{3\pi\tau}{2}|6\tau\right)\theta_3\left(\frac{15\pi\tau}{2}|30\tau\right). \tag{3.2}$$

From (3.1) and (3.2), it follows that

$$q^{-1/4}[\theta_2(0|\tau)\theta_3(0|5\tau)-\theta_3(0|\tau)\theta_2(0|5\tau)]$$
$$=2\left[\theta_3\left(\frac{\pi\tau}{2}|6\tau\right)-\theta_2\left(\frac{\pi\tau}{2}|6\tau\right)\right]$$
$$\cdot\left[\theta_3\left(\frac{5\pi\tau}{2}|30\tau\right)-\theta_2\left(\frac{5\pi\tau}{2}|30\tau\right)\right] \tag{3.3}$$

or also[6]

$$q^{-1/4}[\theta_2(0|\tau)\theta_3(0|5\tau)-\theta_3(0|\tau)\theta_2(0|5\tau)]$$
$$=2\left[\theta_4\left(\frac{\pi\tau}{4}\,\middle|\,\frac{3\tau}{2}\right)\theta_4\left(\frac{5\pi\tau}{4}\,\middle|\,\frac{15\tau}{2}\right)\right]. \tag{3.4}$$

Now, by recalling that[3]

$$\theta_4(x,q)=G(q)\prod_{n=1}^{\infty}(1-2q^{2n-1}\cos 2x+q^{4n-2}), \tag{3.5}$$

a straightforward calculation gives

$$\theta_4\left(\frac{\pi\tau}{4}\,\middle|\,\frac{3\tau}{2}\right)=\prod_{n=1}^{\infty}(1-q^n),$$
$$\theta_4\left(\frac{5\pi\tau}{4}\,\middle|\,\frac{15\pi}{2}\right)=\prod_{n=1}^{\infty}(1-q^{5n})$$

and (3.4) becomes

$$q^{-1/4}[\theta_2(0|\tau)\theta_3(0|5\tau)-\theta_3(0|\tau)\theta_2(0|5\tau)]$$
$$=2\prod_{n=1}^{\infty}(1-q^n)(1-q^{5n}). \tag{3.6}$$

This is indeed equivalent to (1.22), because[3]

$$\prod_{n=1}^{\infty}(1-q^n)=(4\pi^{-3}q^{-1/4}k^{1/2}k'^2K^3)^{1/6}, \tag{3.7}$$

$$\prod_{n=1}^{\infty}(1-q^{5n})=(4\pi^{-3}q^{-5/4}k_5^{1/2}k_5'^2K_5^3)^{1/6}, \tag{3.8}$$

where $K=(\pi/2)\theta_3^2(0|\tau)$ and $K_5=(\pi/2)\theta_3^2(0|5\tau)$ are the complete elliptic integrals of the first kind, having modulus k and k_5, respectively.

As an example, let us take $q=e^{-\pi}$; then, as it is well known, $k=k'=2^{-1/2}$ and (1.22) gives

$$k_5=2^{3/2}(5^{1/4}+1)^{-4}[(5^{1/2}-1)/2]^8,$$
$$k_5'=2^{-5/2}(5^{1/4}+1)^4[(5^{1/2}-1)/2]^4. \tag{3.9}$$

It may be observed that other choices for r, s, x, and y in Eq. (1.25) are of course possible; for instance, by repeating the above calculations with $r=5$, $s=1$, $x=\pi/2$, and $y=0$ and $r=5$, $s=1$, $x=0$, and $y=\pi/2$, respectively, we get

$$k_5'^{1/2}-k'^{1/2}=2^{2/3}(k'k_5')^{1/12}(kk_5)^{1/3}. \tag{3.10}$$

A further relation is obtained from (3.6) with the replacement $q\to -q$. The result is

$$q^{-1/4}[\theta_2(0|\tau)\theta_4(0|5\tau)+\theta_4(0|\tau)\theta_2(0|5\tau)]$$
$$=2\prod_{n=1}^{\infty}(1-q^{2n})(1+q^{2n-1})(1-q^{10n})$$
$$\times(1+q^{5(2n-1)}), \tag{3.11}$$

whence, by using once more the familiar formulas for the infinite products,

$$(kk_5')^{1/2}+(k'k_5)^{1/2}=2^{2/3}(kk'k_5k_5')^{1/12}. \tag{3.12}$$

A final remark is in order. Let us put

$$2^{1/4}q^{1/24}u=\prod_{n=1}^{\infty}(1+q^{2n-1})=[2q^{1/4}(kk')^{-1/2}]^{1/6}, \tag{3.13}$$

$$2^{1/4}q^{5/24}v=\prod_{n=1}^{\infty}(1+q^{5(2n-1)})=[2q^{5/4}(k_5k_5')^{-1/2}]^{1/6}, \tag{3.14}$$

i.e.,

$$u=(2kk')^{-1/12},\quad v=(2k_5k_5')^{-1/12}. \tag{3.15}$$

Now, from (1.22) and (3.10), we have

$$(kk_5')^{1/2}-(kk')^{1/2}-(k_5k_5')^{1/2}+(k'k_5)^{1/2}$$
$$=2^{4/3}(kk'k_5k_5')^{5/12}$$

or also, by recalling (3.12) and (3.15),

$$(u/v)^3+(v/u)^3=2(u^2v^2-1/u^2v^2). \tag{3.16}$$

This is the modular equation of the fifth degree quoted by Ramanujan[9] in his celebrated paper "Modular equations and approximations to π," which has received renewed attention in recent evaluations of π, with a surprisingly high number of decimal figures.[10]

IV. PROOF OF (1.23)

In the general formula (1.25) we put $r=7$, $s=1$, and $x=y=0$; then

$$\theta_3(0|\tau)\theta_3(0|7\tau)$$
$$=\sum_{l=0}^{7}e^{il^2\pi\tau}\theta_3(l\pi\tau|8\tau)\theta_3(7l\pi\tau|56\tau). \tag{4.1}$$

Similarly, with $r=7$, $s=1$, and $x=y=\pi/2$,

$$\theta_3(0|\tau)\theta_3(0|7\tau)+\theta_4(0|\tau)\theta_4(0|7\tau)$$
$$=2\sum_{l=0}^{3} e^{4il^2\pi\tau}\theta_3(2l\pi\tau|8\tau)\theta_3(14l\pi\tau|56\tau), \tag{4.3}$$

From (4.1) and (4.2), we get

$$\theta_3(0|\tau)\theta_3(0|7\tau)+\theta_4(0|\tau)\theta_4(0|7\tau)$$
$$=2\sum_{l=0}^{3} e^{4il^2\pi\tau}\theta_3(2l\pi\tau|8\tau)\theta_3(14l\pi\tau|56\tau), \tag{4.3}$$

which can be simplified in the usual manner, leading to

$$\theta_3(0|\tau)\theta_3(0|7\tau)+\theta_4(0|\tau)\theta_4(0|7\tau)$$
$$=\sum_{l=2}^{4}\theta_l(0|2\tau)\theta_l(0|14\tau). \tag{4.4}$$

This can be rewritten as

$$\theta_3(0|\tau)\theta_3(0|7\tau)\left[1+(k'k_7')^{1/2}\right]$$
$$=\theta_3(0|2\tau)\theta_3(0|14\tau)$$
$$\cdot\left[1+(k_2k_{14})^{1/2}+(k_2'k_{14}')^{1/2}\right]. \tag{4.5}$$

Now, from

$$\prod_{n=1}^{\infty}(1-q^n)^6=4\pi^{-3}q^{-1/4}k^{1/2}k'^2K^3, \tag{4.6}$$

we have, with $q\to q^2$,

$$\prod_{n=1}^{\infty}(1-q^{2n})^6=4\pi^{-3}q^{-1/2}k_2^{1/2}k_2'^2K_2^3. \tag{4.7}$$

On the other hand,

$$\prod_{n=1}^{\infty}(1-q^{2n})^6=2\pi^{-3}q^{-1/2}kk'K^3 \tag{4.8}$$

and so, by comparison with (4.7) [also recall (1.20)]

$$K_2=[(1+k')/2]K, \quad \text{i.e.,}$$
$$\theta_3(0|2\tau)=[(1+k')/2]^{1/2}\theta_3(0|\tau). \tag{4.9}$$

Therefore Eq. (4.5) becomes

$$1+(k'k_7')^{1/2}$$
$$=\tfrac{1}{2}\left[(1+k')(1+k_7')\right]^{1/2}$$
$$+\tfrac{1}{2}\left[(1-k')(1-k_7')\right]^{1/2}+(k'k_7')^{1/4},$$

which, by a straightforward manipulation, leads to (1.23).

As an example, let us take $q=e^{-\pi}$; then, $k=k'=2^{-1/2}$, and

$$k_7^{1/4}+k_7'^{1/4}=2^{1/8}, \quad k_7^2+k_7'^2=1 \quad (k_7'>k_7). \tag{4.10}$$

It is convenient to put $t=(2k_7k_7')^{1/4}$ and $\xi=t+1/t$, i.e., $t=\frac{1}{2}(\xi\pm\sqrt{\xi^2-4})$. As one easily shows, ξ satisfies the equation

$$\xi^2-8\sqrt{2}\xi+18=0,$$

whose acceptable root (in order to have a real t) is $\xi=(\sqrt{7}+1)^2/\sqrt{2}$.

From this, it follows that

$$t=\tfrac{1}{8}\left[(1+\sqrt{7})/\sqrt{2}\pm 7^{1/4}\right]^3. \tag{4.11}$$

Now, if we write

$$k_7^{1/4}=\tfrac{1}{2}\cdot 2^{1/8}(1-\sqrt{1-\alpha}), \quad k_7'^{1/4}=\tfrac{1}{2}\cdot 2^{1/8}(1+\sqrt{1-\alpha}), \tag{4.12}$$

so that $t=(2k_7k_7')^{1/4}=2^{-3/2}\alpha$, we must choose the lower sign in (4.11), in order to have $\alpha<1$. Thus the modulus associated to $q=e^{-7\pi}$ is

$$k_7=2^{-7/2}(1-\sqrt{1-\alpha})^4 \tag{4.13}$$

and the complementary modulus is

$$k_7'=2^{-7/2}(1+\sqrt{1-\alpha})^4, \tag{4.14}$$

α being given by

$$\alpha=\left[\frac{(\sqrt{4+\sqrt{7}}-7^{1/4})}{\sqrt{2}}\right]^3. \tag{4.15}$$

In conclusion, although the results (1.21)–(1.23) are well known in the literature, we believe that their derivation by means of the (very general) additive decomposition for the product $\theta_3(x|r\tau)\theta_3(y|s\tau)$, Eq. (1.25), is of some intrinsic interest. It should also be possible to obtain, along similar lines, formulas for k_{2r+1}, $r\geqslant 4$.

[1]M. Boon, J. Zak, and I. J. Zucker, J. Math. Phys. **24**, 316 (1983).
[2]M. Boon, M. L. Glasser, J. Zak, and I. J. Zucker, J. Phys. A: Math. Gen. **15**, 3439 (1982).
[3]E. T. Whittaker and G. N. Watson, *Modern Analysis* (Cambridge U.P., Cambridge, 1958), Chaps. XXI and XXII.
[4]See, for instance, A. Cayley, *Elliptic Functions* (Dover, New York, 1961), Chap. VIII.
[5]See, in this context, H. M. Srivastava and H. L. Manocha, *A Treatise on Generating Functions* (Horwood, Chichester, 1984), Chap. 2.
[6]Recall that $\theta_{3/4}(x|\tau)=\theta_3(2x|4\tau)\pm\theta_2(2x|4\tau)$.
[7]This series was brought to our attention by R. Ferrari, who met it in the study of two-dimensional electrons in a strong magnetic field. He also guessed, by using the computer, that $S(\pi/\sqrt{3})=0$.
[8]Recall the well-known Legendre's result concerning the modulus $k=\sin(\pi/12)$.
[9]S. Ramanujan, *Collected Papers* (Cambridge U.P., London, 1927).
[10]J. M. Borwein and P. B. Borwein, Sci. Am. **258**(2), 66 (1988).

A Mellin transform technique for the heat kernel expansion

A. P. C. Malbouisson, M. A. R. Monteiro, and F. R. A. Simão
Centro Brasileiro de Pesquisas Físicas, CNPq/CBPF, Rua Dr. Xavier Sigaud, 150, 22290, Rio de Janeiro, RJ, Brazil

(Received 1 December 1987; accepted for publication 19 April 1989)

It is shown, for a wide class of operators, that the solution to the associated heat equation may be obtained as a series. This is accomplished using the inverse Mellin transform of the kernel of the sth power of the operator, together with the analytic properties of the kernel in the complex s plane.

I. INTRODUCTION

A renewed interest in the study of the solutions of the heat equation associated to elliptic operators, in the form of an asymptotic expansion, was observed over these last few years in several situations. Examples of such situations are found in the regularization of operator determinants associated with Grassmann variables in the path integral approach to quantum field theory, in the calculation of non-Abelian anomalies,[1,2] or in studies of field theories in curved spaces.[3] For instance, in two-dimensional QCD[1,4] we may write the generating functional, after integration over the fermionic fields, as

$$Z=\int \mathscr{D}A(x)\det \not{D}\exp\left[-\frac{1}{4}\int d^2x\, G_{\mu\nu a}G^{\mu\nu a}\right] \tag{1}$$

(apart from gauge fixing terms), where $A(x)$ is the gauge field, $G_{\mu\nu a}$ is the gauge field strength tensor, and $D=i(\partial+A(x))$. The determinant of $\not{D}$ appearing in (1) comes from the integration over the fermionic fields; it is a divergent quantity and must be regularized. One of the most popular ways is the proper-time regularization method.[5] The regularized determinant is given in terms of the proper-time regularization parameter ϵ, through the diagonal part of a function[1] $F(\epsilon,x,y)$, which obeys the "heat equation," associated to the operator $\not{D}^2$ [see Eq. (2)]. The regularized determinant $\det \not{D}(\epsilon)$ $(\epsilon\to 0)$ must be known and then the behavior of $F(\epsilon,x,y)$, as $\epsilon\to 0$, must be found.

Beyond the particular example we have just given, we are led, in general, to the study of the solutions of the heat equation,

$$\frac{d}{dt}F(t,x,y)=HF(t,x,y), \tag{2}$$

where t is a ("time") parameter and x and y are points of a D-dimensional compact manifold without boundary; the operator H acts on the x variable. In the case of QCD_2, t is the proper-time regularization parameter ϵ and H is the operator $\not{D}^2$. For more generality, H may be taken as an order $\mathbf{m}$ pseudodifferential operator.[6] Particularly important is the asymptotic behavior of the diagonal part of $F(t,x,y)$ as $t\to 0$. This is usually done by means of the de Witt *anzatz*,[3]

$$F(t,x,y)=F_0(t,x,y)\sum_{l=0}^{\infty}t^l a_l(x,y),$$

where F_0 is the solution of the "free heat equation."

In this paper we propose a new simple Mellin transform method to obtain an asymptotic expansion to the solution of the heat equation $F(t;x,y)$, the so-called *heat kernel*. This is done using the rigorous results of Seeley[7] on the analytic structure of the kernel $K(s;x,y)$ associated to the sth power of the operator H of [Eq. (2)], H^s, in the complex s plane. Our expansion may be seen as an alternative to the de Witt *anzatz* in the case where the residues of the diagonal part of $K(s;x,y)$ can be calculated. We remark that Mellin transform methods for obtaining asymptotic behaviors have been used in other contexts. Indeed, one of us used such methods to demonstrate theorems on the asymptotic behavior of Feynman amplitudes.[8]

To proceed, we note that it may be shown that the Green's function of H^s, $Z(s,x,y)$, is related to the Seeley's kernel of H^s by $Z(s,x,y)=K(-s,x,y)$, and to the solution of the heat equation (2) by a Mellin transform. So

$$K(s,x,y)=\frac{1}{\Gamma(-s)}\int_0^{\infty}dt\, t^{-s-1}F(t,x,y). \tag{3}$$

Conversely, the inverse Mellin transform gives

$$F(t,x,y)=\int_{-\infty}^{+\infty}\frac{d\,\mathrm{Im}\,s}{2i\pi}\,t^s\,\Gamma(-s)K(s,x,y), \tag{4a}$$

provided $K(s,x,y)$ can be extended to the whole complex s plane. The s integration in (4) goes parallel to the imaginary axis and $\mathrm{Re}(s)$ must belong to the analyticity domain of $K(s,x,y,)$. One of Seeley's results[7] is that within the approximation made to construct the power operator H^s, it has a continuous kernel for $\mathrm{Re}(s)<-D/\mathbf{m}$. The diagonal elements $K(s,x,y)$ extend to meromorphic functions of s, having as only singularities simple poles located at $s=(j-D)/\mathbf{m}, j=0,1,2,\ldots$; their residues can, in principle, be calculated[7] from the symbol (generalization of the characteristic polynomial) of H^s, using the formula

$$\mathrm{Res}\,K(s=(j-D)|\overline{\mathbf{m}})$$

$$=\frac{1m}{i\mathbf{m}(2\pi)^{D+1}}\iint_{|\xi|=1,\Gamma}\lambda^{(j-D)/\mathbf{m}}\times b_{-\mathbf{m}-j}(\lambda,\xi)d\lambda\,d\xi, \tag{4b}$$

where Γ is a curve coming from ∞, going along the ray of minimal growth to a small circle around the origin, then going back to infinity. The quantities $b(\lambda,\xi)_{-\mathbf{m}-j}$ are obtained from the coefficients of the symbol. Here $|\xi|=1$ means that the set of variables $\{\xi\}$ is constrained to be at the surface of the D-dimensional unit sphere. The off-diagonal elements $K(s,x,y)$ $x\neq y$ extend to entire functions of s.

 0022-2488/89/092016-03$02.50

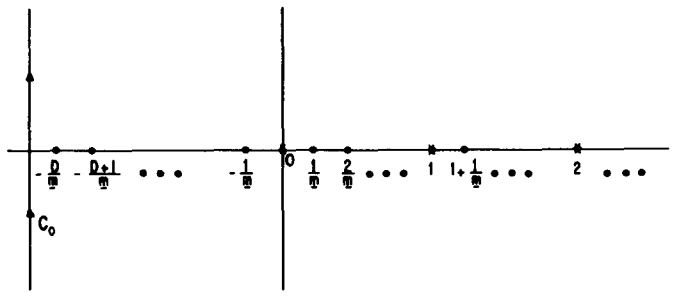


FIG. 1. Poles of $\Gamma(-s)K(s,x,x)$ for a general pseudodifferential operator H; K is the approximate kernel of H^s as given by Ref. 6.

For the diagonal elements, the general analytic structure of the integrand in (4a) is displayed in Fig. 1. The inverse Mellin transform (4a) is unambiguously defined if we take the integration along a line $\mathbb{C}_0$ in the "initial" analyticity domain of $K(s,x,y)$, $\mathrm{Re}(s) < -D/\bar{m}$. Then we may obtain an expansion in t by displacing the integration contour to the right, picking up successively the contributions from the poles. Then if the kernel K has a good behavior at infinite s, together with the vanishing[7] of the residues of K at positive integer s, we are left with the remaining contributions from the residues at the poles. Thus the diagonal elements of the solution of the heat equation (2) are expressible as the following series:

$$F(t,x,x) = -\sum_{l=0}^{\infty} t^l \frac{(-1)^l}{l!} K(l;x,x) - \sum_j t^{(j-D)/m}\,\Gamma[(D-j)/\mathbf{m}]R_j(x), \qquad (5)$$

where the sum over $j = 0,1,2,\ldots$, excludes the terms such that $(j-D)/\bar{m} = 0,1,2,\ldots$, and R_j are the residues of $K(s,x,x)$ at $s = (j-D)/\bar{m}$.

This is the result we would like to present here. In the following section we illustrate our method with two simple examples.

II. THE LAPLACIAN IN A RIEMANNIAN METRIC

We consider the operator $H = -\nabla^2 + P$, where ∇^2 is the Laplacian in a Riemannian manifold and P the projection on the constants. In this case, $K(s;x,x)$ has poles at the values $s = j - D/2$, $j = 0,1,2,\ldots$, and if the dimension D is even, these poles are in finite number,[7] located at $s = j - D/2$, $j = 0,1,\ldots,D/2-1$. The residue at $s = -P/2$ can be calculated in geodesic coordinates using formula (4b), and we can obtain the leading term to the expansion (5) in differential form [We remember the definition of Seeley's kernel, as such, that if H acts on a manifold M, then $H^s f(x) = \int_M dy\, K(s,x,y)f(y)$.]:

$$t^{-D/2}\,\Gamma(D/2)[(2\pi)^{-D}/2]|S^{D-1}|dv,$$

where $|S^{D-1}|$ is the surface area of the unit sphere in $\mathbb{R}^D$, and dv the volume element in the manifold.

III. THE EUCLIDEAN LAPLACIAN

Of course we do not know the exact kernel in the general case, but we indeed know it in at least a particular one, and it may be instructive to see what happens in this case. Let us take H to be the Euclidean Laplacian operator, $H = \partial^2$, in D dimensions. The Green's function of H^k, for real integer positive k, is given in Ref. 9. Starting from this we perform the extension from positive integers k to complex s values, obtaining the exact kernel of H^s as a meromorphic function of s, for any dimensionality D,

$$K_L(s,x,y) = (-1)^{-s}\frac{e^{i\pi(D/2)}\Gamma(D/2+s)(P+i0)^{-D/2-s}}{4^{-s}\Gamma(-s)\pi^{D/2}}, \quad \mathrm{Re}(s) < D/2, \qquad (6)$$

where P is the quadratic form $-\Sigma_{i=1}^{D}(x_i-y_i)^2 \equiv -(x-y)^2$. We note that the original restriction[9] $(-D/2) < k < 0$ for integer k implies, after performing the analytic continuation, the "initial" domain of analyticity $(-D/2) < \mathrm{Re}(s) < 0$ for $K(s,x,y)$. In this domain the inverse Mellin transform (6) is unambiguously defined. Then starting from an integral along a line $\mathbb{C}_0$ parallel to the imaginary axis in the region $(-D/2) < \mathrm{Re}(s) < 0$ (Fig. 2) we sum up the contributions from the poles, obtaining the result

$$F(x,y,t) = (-1)^D(4\pi t)^{D/2}\exp[-(x-y)^2/4t], \qquad (7)$$

which is the well-known solution to the free heat equation.

In the case of the Laplacian L, the analytic structure in s of the exact kernel $K_L(s,x,y)$ of L^s, multiplied by $\Gamma(-s)$, is shown in Fig. 2. It is rather different from that corresponding to the kernel of the power H^s of *any* pseudodifferential operator H, as obtained from Ref. 7 (Fig. 1). This is simply due to the fact that $\mathbb{R}^D$ does not belong to the class of spaces treated by Seeley in his work.

IV. CONCLUSION

The expansion (5) for the heat kernel, which is the result we would like to present here, looks rather different from the de Witt's *anzatz* currently used. It could give new results when applied to more realistic examples. However, it is not our purpose in this short note to make physical applications. These will be the subject of a forthcoming paper. Our work must be understood as a new method to obtain an

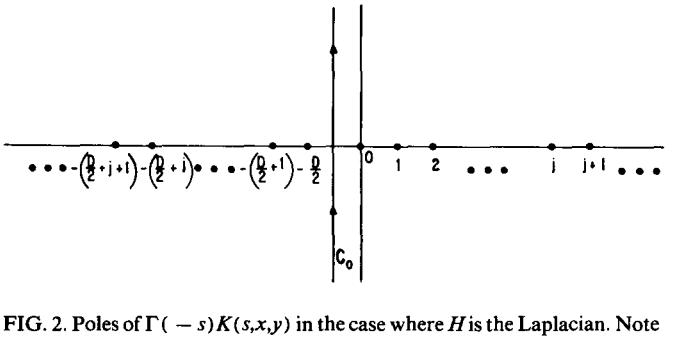


FIG. 2. Poles of $\Gamma(-s)K(s,x,y)$ in the case where H is the Laplacian. Note that $K(s,x,y)$ is the exact kernel of H^s.

asymptotic expansion to the heat kernel, which could lead to new results in physical situations.

ACKNOWLEDGMENT

We are indebted to Dr. J. A. Mignaco and Dr. A. Ferraz de Camargo for fruitful discussions.

[1]A. M. Polyakov and P. B. Wilyman, Phys. Lett. B **131**, 121 (1983); O. Alvarez, Nucl. Phys. B **238**, 61 (1984); R. E. Gamboa Saravi, F. A. Schaposnik, and J. E. Solomin, Phys. Rev. D **30**, 1353 (1984); L. C. L. Botelho and M. A. do Rego Monteiro, *ibid.* **30**, 2242 (1984); A. P. Balachandran, G. Marmo, Y. P. Nair, and C. G. Traherm, *ibid.* **25**, 2713 (1982).

[2]K. Fujikawa, Phys. Rev. Lett. **42**, 1195 (1979); Phys. Rev. D **21**, 2848 (1980).

[3]See B. de Witt, *The Dynamical Theory of Groups and Fields* (Gordon and Breach, New York, 1965); N. D. Birrel and P. C. W. Daries, *Quantum Fields in Curved Space* (Cambridge U. P., Cambridge, 1982); G. Cognola and S. Zerbini, "Some Physical Applications of the Heat Kernel Expansion," Preprint, Dipartimento di Fisica, Università di Trento, Ist. Naz. di Fisica Nucleare, Sezione di Padova (Italia), 1987.

[4]R. E. Gamboa Saravi, F. A. Schaposnik, and J. E. Solomin, Nucl. Phys. B **185**, 239 (1981).

[5]S. W. Hawking, Commun. Math. Phys. **55**, 133 (1977); K. D. Rothe and B. Schroer, in *Field Theoretical Methods in Particle Physics*, NATO Advanced Study Institutes Series B, edited by Werner Rühl (Plenum, New York, 1980), Vol. 55, p. 249.

[6]L. Hörmander, Commun. Pure Appl. Math. **18**, 501 (1965); J. J. Kohn and L. Miremberg, *ibid.* **18**, 269 (1965).

[7]R. T. Seeley, Am. Math. Soc. Proc. Symp. Pure Math. **10**, 288 (1967).

[8]M. Bergère, C. de Calan, and A. P. C. Malbouisson, Commun. Math. Phys. **62**, 137 (1978); C. de Calan and A. P. C. Malbouisson, Ann. Inst. Henri Poincaré **32**, 91 (1980).

[9]I. M. Gellfand and G. E. Chirlov, *Les Distributions* (Dunod, Paris, 1962), Vol. 1, pp. 273–277.

Sensitive terms in the path integral: Ordering and stochastic options

B. Gaveau
Department of Mathematics, University of Paris VI, 4 pl. Jussieu, 75252, Paris Cedex 05, France
L. S. Schulman
Department of Physics, Clarkson University, Potsdam, New York 13676

(Received 14 February 1988; accepted for publication 8 March 1989)

In a previous publication [Phys. Rev. D **36**, 1135 (1987)] it was shown how the stochastic process associated with the solution of the Schrödinger or diffusion equation can be derived by an infinity of "Gaussian tricks." In this article the method is extended to differential operators of the form $(p - eA/c)^2$, $\partial c^2(x)\partial$, and $\Delta = (1/\sqrt{g})\partial_\mu(\sqrt{g}g^{\mu\nu}\partial_\nu)$. In this formulation the relation between operator ordering ambiguities and time labeling in the functional integral is immediate. In particular, it is clear where the choice between Ito and Stratonovich integrals enters.

I. INTRODUCTION

In a previous publication[1] we gave a new derivation of the path integral using the Trotter formula and the explicit introduction of a stochastic variable at each time step by means of the "Gaussian trick" or the "uncompleting of the square." In this paper we extend the method to situations where the Lagrangian contains sensitive terms and where problems of operator ordering may arise. In particular, we will show where the choice between the Ito and Stratonovich stochastic integrals enters.

The operators that we consider are

$$L_1 = \frac{1}{2}\left(\frac{\partial}{\partial \mathbf{x}} - \frac{ie}{c}\mathbf{A}\right)^2 - V, \tag{1}$$

with V and $\mathbf{A}$ potentials

$$L_2 = \frac{1}{2}\frac{\partial}{\partial x}(c(x))^2\frac{\partial}{\partial x}, \tag{2}$$

with $c(x)$ a function, and

$$L_3 = \Delta = \frac{1}{\sqrt{g}}\frac{\partial}{\partial x^\mu}\left(\sqrt{g}g^{\mu\nu}\frac{\partial}{\partial x^\nu}\right), \tag{3}$$

with $g_{\mu\nu}(x)$ a Riemannian metric, using the usual notation, e.g., $g = \det g_{\mu\nu}$. For simplicity we work with the heat equation $\partial f/\partial t = L_i f$ and evaluate $\exp(tL_i)$, $i = 1,2,3$.

We recall the method of Ref. 1 for the case L_1 with $\mathbf{A} = 0$ (which we call L_0). By the Trotter formula,

$$e^{tL_0} = \lim_{n\to\infty}\left[\exp\left(\frac{1}{2}\left(\frac{\partial^2}{\partial x^2}\right)\epsilon\right)\exp(-\epsilon V)\right]^n, \tag{4}$$

with $\epsilon = t/n$. We also define $\epsilon' \equiv \sqrt{\epsilon}$ and use the notation $\partial = \partial/\partial\mathbf{x}$. The essential step is to write

$$e^{(\epsilon/2)\partial^2} = (2\pi)^{-3/2}\int_{-\infty}^{\infty} d^3y\, e^{-\mathbf{y}^2/2 + \mathbf{y}\cdot\epsilon'\partial} = \langle e^{\epsilon'\mathbf{G}\cdot\partial}\rangle, \tag{5}$$

so that $\exp(\frac{1}{2}\epsilon\,\partial^2)$ is the expectation of a vector valued Gaussian random variable of mean zero, variance 1. A variable $\mathbf{G}_k$, $k = 1,\ldots,n$, is defined for each term in the product (4) and the operators $\exp(\epsilon'\mathbf{G}_k\cdot\partial)$ commuted past the factors $\exp(-\epsilon V)$ by means of the general formula

$$e^{u\partial}f(v) = f(u+v)e^{u\partial}. \tag{6}$$

The result is

$$e^{tL_0} = \lim_{n\to\infty}\left\langle \exp\left[-\epsilon\sum_{j=1}^{n} V\left(\cdot + \epsilon'\sum_{i=j}^{n}\mathbf{G}_i\right)\right]\times\exp\left(\sum_{j=1}^{n}\epsilon'\mathbf{G}_j\cdot\partial\right)\right\rangle, \tag{7}$$

with the dot in the argument of V referring to the argument of the function on which e^{tL_0} acts. In the continuum limit, (7) becomes

$$e^{tL_0} = E\left[\exp\left(\int_0^t ds V(\cdot + \mathbf{b}(t) - \mathbf{b}(s))\right)\times\exp(\mathbf{b}(t)\cdot\partial)\right], \tag{8}$$

the expectation being over Wiener measure. Here $\mathbf{b}(t)$ is Brownian motion and is the limit of $\epsilon'\Sigma_{j=1}^n\mathbf{G}_j$, so that $\langle\mathbf{b}\rangle = 0$ and $\langle b_\alpha(s)b_\beta(s')\rangle = \delta_{\alpha\beta}\delta(s-s')$. As we explained in Ref. 1 (and in more detail in Ref. 2), Eq. (8) is the usual path integral representation for the propagator (the Feynman–Kac formula), although the form in which it is written is not that most commonly used in the physics literature.

II. VECTOR POTENTIAL

We now apply the same method to L_1, that is to say, deal with the presence of the vector potential $\mathbf{A}(x)$. We will drop the $V(x)$ from L_1 since it is handled exactly as for L_0. The first step is *not* to apply the Trotter formula,[3] but merely to write

$$e^{tL_1} = \{\exp[\epsilon(\partial - \mathbf{a}(x))^2]\}^n, \tag{9}$$

with $\mathbf{a} = ie\mathbf{A}/c$. The Gaussian trick is applied to each term in the product to yield

$$e^{tL_1} = \left\langle\prod_k \exp[\epsilon'\mathbf{G}_k\cdot(\partial - \mathbf{a})]\right\rangle. \tag{10}$$

The expectation in (10) is over all random variables $\mathbf{G}_k$, $k = 1,\ldots,n$. At this point the usual Trotter formula approach would be to separate the terms in the argument of the exponent (giving terms of the form $e^{\epsilon' G\partial}e^{\epsilon' Ga}$), but this does not work here because the error is of order ϵ'^2 and

 0022-2488/89/092019-04$02.50

$\epsilon' = \sqrt{\epsilon} = O(1/\sqrt{n})$. Terms of order $\epsilon'^2 = O(1/n)$ must be kept. Recall

$$e^{A+B} = e^A e^B e^{-(1/2)[A,B]} e^{\text{cubic terms}}$$
$$= e^{+(1/2)[A,B]} e^B e^A e^{\text{cubic terms}}. \tag{11}$$

We apply formula (11) to (10), keeping $O(\epsilon'^2)$ but dropping $O(\epsilon'^3)$. The k th term is

$$\exp(\epsilon' \mathbf{G}_k \cdot \partial) \exp(-\epsilon' \mathbf{G}_k \cdot \mathbf{a})$$
$$\times \exp\left(\frac{\epsilon}{2} \sum_{\alpha,\beta=1}^{3} \mathbf{G}_{k\alpha} \mathbf{G}_{k\beta} \cdot \partial_\alpha a_\beta\right). \tag{12}$$

Because only $O(\epsilon)$ terms need be retained we can consider the expectation of the last exponential in the product (12) independently of the expectation of the product of the first two exponentials in (12). By independence, $\langle \mathbf{G}_{k\alpha} \mathbf{G}_{k\beta} \rangle = \delta_{\alpha\beta}$ (the Greek indices refer to the vector components) and the argument of the last exponent can simply be replaced by $\frac{1}{2}\epsilon \partial \cdot \mathbf{a}$. Following the steps used for L_0 we obtain

$$e^{tL_1} = \lim_{n\to\infty} \left\langle \exp\left[-\sum_{k=1}^{n} \epsilon' \mathbf{G}_k \cdot \mathbf{a}\left(\cdot + \epsilon' \sum_{i=k}^{n} \mathbf{G}_i\right)\right] \right.$$
$$\times \exp\left[\frac{1}{2}\epsilon \sum_{k=1}^{n} (\partial \cdot \mathbf{a})\left(\cdot + \epsilon' \sum_{i=k}^{n} \mathbf{G}_i\right)\right]$$
$$\left. \times \exp\left(\sum_{k=1}^{n} \epsilon' \mathbf{G}_k \cdot \partial\right)\right\rangle. \tag{13}$$

Details of the passage to the continuum limit are now important and we define

$$\mathbf{b}_k = \epsilon' \sum_{j=0}^{k} \mathbf{G}_k, \quad \mathbf{G}_0 \equiv 0,$$
$$(\Delta \mathbf{b})_k = \mathbf{b}_k - \mathbf{b}_{k-1} = \epsilon' \mathbf{G}_k, \tag{14}$$

leading to

$$e^{tL_1} = \lim_{n\to\infty} \left\langle \exp\left[-\sum_{k=1}^{n} (\Delta \mathbf{b})_k \cdot \mathbf{a}(\cdot + \mathbf{b}_n - \mathbf{b}_{k-1})\right] \right.$$
$$\times \exp\left[\frac{1}{2}\epsilon \sum_{k=1}^{n} (\partial \cdot \mathbf{a})(\cdot + \mathbf{b}_n - \mathbf{b}_{k-1})\right]$$
$$\left. \times \exp(\mathbf{b}_n \cdot \partial)\right\rangle. \tag{15}$$

A continuum limit can now be written, but it is clear from (15) that a particular form of the stochastic integral has been chosen, to wit the argument of $\mathbf{a}$ involves $\mathbf{b}_{k-1}$, while the $\Delta \mathbf{b}_k$ in the stochastic integral involves both $\mathbf{b}_{k-1}$ and the later variable $\mathbf{b}_k$. By using the other of the two expressions in Eq. (11) we could get the other choice of order with an opposite sign for the divergence term $\partial \cdot \mathbf{a}$. These alternatives provide the representation of the propagator using an Ito, retarded, prescription or its opposite, a fully advanced prescription. (In a gauge in which $\partial \cdot \mathbf{a} = 0$ the choice would be irrelevant.)

We next see in what way the Stratonovich stochastic integral is naturally generated. In this form there will be no $\partial \cdot \mathbf{a}$ term, so that for quantum mechanical applications it is generally preferred.[4]

Let n be even and consider two successive terms in the product (10). For each term we use (11) to split off the gradient; in the first term we split off the gradient to the left, in the second it is split off to the right. Thus

$$\exp[\epsilon' \mathbf{G}_{2k} \cdot (\partial - \mathbf{a})] \exp[\epsilon' \mathbf{G}_{2k-1} \cdot (\partial - \mathbf{a})]$$
$$= \exp(\epsilon' \mathbf{G}_{2k} \cdot \partial) \exp(-\epsilon' \mathbf{G}_{2k} \cdot \mathbf{a})$$
$$\times \exp(+(\epsilon/2)[\mathbf{G}_{2k} \cdot \partial, \mathbf{G}_{2k} \cdot \mathbf{a}])$$
$$\times \exp(-(\epsilon/2)[\mathbf{G}_{2k-1} \cdot \partial, \mathbf{G}_{2k-1} \cdot \mathbf{a}])$$
$$\times \exp(-\epsilon' \mathbf{G}_{2k-1} \cdot \mathbf{a}) \exp(\epsilon' \mathbf{G}_{2k-1} \cdot \partial). \tag{16}$$

Again, because we need retain only $O(\epsilon)$, we can consider expectations of the intermediate terms (involving the commutators) separately. But these terms are now seen to cancel each other because $\mathbf{G}_{2k-1}$ and $\mathbf{G}_{2k}$ are identically distributed. Therefore by splitting successive terms in opposite order in the product (10), the commutators disappear — at the expense of extra bookkeeping for the variables $\mathbf{G}_k$. The propagator becomes

$$e^{tL_1} = \lim_{n\to\infty} \langle \exp(\epsilon' \mathbf{G}_{2\nu} \cdot \partial) \exp(-\epsilon'(\mathbf{G}_{2\nu} + \mathbf{G}_{2\nu-1}) \cdot \mathbf{a}) \cdots$$
$$\times \exp(\epsilon'(\mathbf{G}_3 + \mathbf{G}_2) \cdot \partial) \exp(-\epsilon'(\mathbf{G}_2 + \mathbf{G}_1) \cdot \mathbf{a})$$
$$\times \exp(\epsilon'(\mathbf{G}_1 + \mathbf{G}_0) \cdot \partial) \rangle, \tag{17}$$

where we have used $\nu = n/2$. We now perform the usual permuting to bring all translation operators to the right. The result is

$$e^{tL_1} = \lim_{\nu\to\infty} \left\langle \exp\left[-\epsilon' \sum_{k=1}^{\nu} (\mathbf{G}_{2k} + \mathbf{G}_{2k-1}) \right.\right.$$
$$\left. \cdot \mathbf{a}\left(\cdot + \epsilon' \sum_{l=k}^{\nu} (\mathbf{G}_{2l+1} + \mathbf{G}_{2l})\right)\right]$$
$$\left. \times \exp\left(\epsilon' \sum_{k=0}^{\nu} (\mathbf{G}_{2k+1} + \mathbf{G}_{2k}) \cdot \partial\right)\right\rangle, \tag{18}$$

with $\mathbf{G}_{2\nu+1} \equiv 0$. At this point, comparison with (13) shows the essential difference between the present method (which will lead to the Stratonovitch integral) and the former (Ito) one. In (13) the increment multiplying $\mathbf{a}$ is $\mathbf{G}_k$ and it is also in the argument of $\mathbf{a}$.[5] In (18) the increment is $\mathbf{G}_{2k} + \mathbf{G}_{2k-1}$, while the argument of $\mathbf{a}$ contains $\mathbf{G}_{2k}$ but not $\mathbf{G}_{2k-1}$. This last observation is the essential feature of our demonstration and is the way the "midpoint" rule enters the path integral.

The calculation proceeds by defining $\mathbf{b}_k = \epsilon' \Sigma_{j=0}^{k} \mathbf{G}_j$ with $k = 1,\ldots,n$ and $\mathbf{b}_0 \equiv 0$. Equation (18) becomes

$$e^{tL_1} = \lim_{n\to\infty} \left\langle \exp\left[-\sum_{k=1}^{\nu} (\mathbf{b}_{2k} - \mathbf{b}_{2k-2}) \right.\right.$$
$$\left. \left. \cdot \mathbf{a}(\cdot + \mathbf{b}_n - \mathbf{b}_{2k-1})\right] \exp(\mathbf{b}_n \cdot \partial)\right\rangle. \tag{18a}$$

Again taking $s = k\epsilon$, in continuous time notation the expression (18a) becomes

$$e^{tL_1} = \lim_{n\to\infty} \left\langle \exp\left[-\sum (\mathbf{b}(s) - \mathbf{b}(s - 2\epsilon)) \right.\right.$$
$$\left. \left. \cdot \mathbf{a}(\cdot + \mathbf{b}(t) - \mathbf{b}(s - \epsilon))\right] \exp(\mathbf{b}(t) \cdot \partial)\right\rangle. \tag{18b}$$

The sum in (18a) or (18b) is a particular prescription for the integral $\int d\mathbf{b} \cdot \mathbf{a}$, and as such is a variant of the Ito integral

[in Eq. (13)] encountered previously. It is in fact the Stratonovich integral and from (18b) it is clear that for a given time interval $dt = 2\epsilon$ between $s - 2\epsilon$ and s, the time argument for $\mathbf{b}(\cdot)$ in $\mathbf{a}$ is $s - \epsilon$, that is, at the midpoint. In probabilistic notation (18b) can be written as

$$e^{tL_1} = E\left[\exp\left(-\int_0^t d\mathbf{b}\cdot\mathbf{a}(\cdot + \mathbf{b}(t) - \mathbf{b}(s))\right) \times \exp(\mathbf{b}(t)\cdot\partial)\right], \tag{19}$$

with the understanding that $\mathbf{b}(s)$, within the argument of $\mathbf{a}$, is evaluated at the midpoint. That is, the integral is the limit of a sum of terms,

$$[\mathbf{b}(s+\epsilon) - \mathbf{b}(s)]\cdot\mathbf{a}(\cdot + \mathbf{b}(t) - \mathbf{b}(s + \tfrac{1}{2}\epsilon)).$$

III. SPATIAL DEPENDENCE IN THE COEFFICIENT OF THE SECOND DERIVATIVE

The operator $L_2 = \frac{1}{2}\partial c^2\,\partial$ calls for a different strategy. It is clearly one of the quantizations of the classical Lagrangian $\frac{1}{2}c(x)^2\dot{x}^2$ and it is self-adjoint with respect to the measure dx. Other operator choices are possible, for example, $\frac{1}{2}c\,\partial^2 c$ or $\frac{1}{4}(c\,\partial c\,\partial + \partial c\,\partial c)$ and will arise from other orderings. In fact, our scheme will not produce L_2 exactly but an expression differing from it by a potential term—something easily corrected.

As usual we write

$$e^{tL_2} = (e^{\epsilon\,\partial c^2\,\partial/2})^n, \tag{20}$$

with $\epsilon = t/n$. Now consider the operator

$$F(\lambda) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dy\, e^{\lambda y\partial}\frac{1}{c}e^{-y^2/2c^2}e^{\lambda y\partial} \equiv \int dy\,\phi(y). \tag{21}$$

Clearly F is self-adjoint and $F(0) = 1$. Here F is an even function of λ: For $\lambda \to -\lambda$, let $y \to -y$. We have in mind that λ^2 will essentially be ϵ so we want to expand F in powers of λ up to and including λ^2. Consider

$$\frac{\partial F}{\partial\lambda} = \partial\int dy\, y\phi + \text{adjoint}. \tag{22}$$

At $\lambda = 0$ this is zero, as expected. Next, consider

$$\frac{\partial^2 F}{\partial\lambda^2} = \partial^2\int dy\, y^2\phi + \partial\int dy\, y^2\phi\partial + \text{adjoint}. \tag{23}$$

At $\lambda = 0$ this gives

$$\left.\frac{\partial^2 F}{\partial\lambda^2}\right|_{\lambda=0} = \partial^2 c^2 + \partial c^2\,\partial + \text{adjoint} = 4\,\partial c^2\,\partial + (c^2)'', \tag{24}$$

where $(c^2)''$ is second derivative of c^2. Thus

$$F(\lambda) = 1 + (\lambda^2/2)(4\,\partial c^2\,\partial + (c^2)'') + O(\lambda^4). \tag{25}$$

Comparing to $\exp(\epsilon\,\partial c^2\,\partial/2) = 1 + (\epsilon/2)\partial c^2\,\partial + O(\epsilon^2)$ we let $\lambda = \sqrt{\epsilon}/2 = \epsilon'/2$. It follows that to order ϵ

$$e^{(1/2)\epsilon\,\partial c^2\,\partial} = F(\sqrt{\epsilon}/2)e^{-\epsilon(c^2)''/8} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{(\epsilon'/2)y\partial}\frac{1}{c} \times e^{-y^2/2c^2}e^{(\epsilon'/2)y\partial}e^{-\epsilon(c^2)''/8}. \tag{26}$$

The operator $F(\epsilon')$ is in a sense $\langle\exp(\epsilon' G\partial)\rangle$ but the random variable G, implicitly defined, has a position dependent variance $c(x)^2$. Thus when taking the expectation of a function of $\partial = \partial/\partial x$—which does not commute with $c(x)$—more explicit specification of the meaning of the expectation must be given. Therefore we will continue to write the explicit integral form of F rather than employ bracket notation.

Remark: We can interpret the expression (26) in a rather different way. The x-dependent Gaussian factor $e^{-y^2/2c^2}$ plays two roles: first, it is a weight on the infinitesimal part of the path between time t and $t + \epsilon'$. Secondly, it is also an operator acting on functions of x (by ordinary multiplication). This points to a more general kind of path integral where the weights over the paths would themselves be operators. In fact this is the situation that was met in the path integral representation for the Dirac equation.[1] In the formula (26) and, more generally, in path integral representations of elliptic or parabolic scalar equations, the situation is easier for the following reasons: the various path weights at each time are commutative, because they are Gaussian weights with value in the operator algebra of multiplication by scalar functions. Finally, this leads, after reordering, to an ordinary measure on the path space and to stochastic differential equations. On the other hand, this is no longer true for the Dirac equation or even for matrix valued elliptic systems. This is also what happens in quantum field theory.

Continuing our development, the propagator is then written

$$e^{tL_2} = \prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}}\int dy_k\,\exp\left(\frac{1}{2}\epsilon' y_k\partial\right)\frac{1}{c(\cdot)}\exp\left[\frac{-y_k^2}{2c(\cdot)^2}\right] \times\exp\left(\frac{1}{2}\epsilon' y_k\partial\right)\exp\left[\frac{-\epsilon(c^2)''}{8}\right], \tag{27}$$

with time ordering and limit understood. By the usual steps (27) becomes

$$e^{tL_2} = \prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}}\int dy_k\,\frac{1}{c(\cdot+\eta_k)}\exp\left(-\sum_{k=1}^{n}\frac{(\Delta b)_k^2}{2c^2(\cdot+\eta_k)}\right) \times\exp\left(-\frac{\epsilon}{8}(c^2)''(\cdot+\eta_k)\right)\exp(b_n\partial), \tag{28}$$

where we have defined

$$\eta_k = b_n - b_k + \tfrac{1}{2}(\Delta b)_k, \qquad b_k = \epsilon'\sum_{j=1}^{k} y_j, \text{ and } (\Delta b)_k \equiv \epsilon' y_k, \tag{29}$$

and the $\frac{1}{2}\Delta b_k$ can be dropped from the argument of $(c^2)''$ since this is an ordinary potential-like term. The continuum limit of (28) is

$$e^{tL_2} = E\left[\exp\left(-\frac{1}{8}\int_0^t ds\,(c^2)''(\cdot + b(t) - b(s))\right)\exp(b(t)\partial)\right].$$

In this expectation the weight assigned to paths is position dependent. In calculating that weight, the function c is evaluated at the midpoint position as in the Stratonovich stochastic integral.

IV. RIEMANNIAN METRIC

The operator L_3 is expressed in a similar way. Define

$$F(\lambda) = (2\pi)^{-d/2}\int d^d y\, g^{-1/4} e^{\lambda y^\mu \partial_\mu} g \times e^{-g_{\mu\nu}y^\mu y^\nu/2} e^{\lambda y^\mu \partial_\mu} g^{-1/4}, \tag{30}$$

with d the dimension of the space. By our usual steps this leads to the exponential of the operator,

$$\tilde{L}_3 = g^{-1/4}\partial_\mu (g^{1/2} g^{\mu\nu})\partial_\nu g^{-1/4}, \tag{31}$$

based on the $O(\epsilon)$ relation

$$\exp(\epsilon \tilde{L}_3) = F(\sqrt{\epsilon/2})\exp(-(\epsilon/4) g^{-1/2}\partial_\mu \partial_\nu (g^{1/2} g^{\mu\nu})). \tag{32}$$

We have not quite recovered the Laplacian because our integration volume is implicitly $d^d x$. By going to the Riemannian volume, $g^{1/2} d^d x$, we effectively replace the functions f on which $\tilde{L}_3$ acts by $f \to \psi = g^{-1/4} f$. Under this transformation the operator $\tilde{L}_3$ of Eq. (31) becomes $g^{-1/2}\partial_\mu (g^{1/2} g^{\mu\nu}\partial_\mu)$, the usual Laplacian.

As in our earlier cases, the relation (32) leads to midpoint evaluation of stochastic integrals, namely, the Stratonovich form.

By changes in the ordering scheme of objects of the form of our $F(\lambda)$, potentials can be added to the effective generator, and these potentials are of the same sort as those derived in earlier work.[6] The present calculations also make it clear why ordering errors can so easily escape notice. For us, y is basically $\Delta x/\sqrt{\Delta t}$ and the c^2 of L_2 would be $O(\hbar)$ if we were working with quantum mechanics. Matters of operator ordering gave us an effective potential $(c^2)''$. Since the leading term $y^2/2c^2$ is $O(1/c^2)$, the relative size of the correction is $O(\hbar^2)$ and semiclassical results will not generally be affected by ordering problems. Of course this is an old observation and we are only noting the natural way in which it emerges from the present considerations.

The form of F given in Eq. (30), while manifestly Hermitian, is not covariant and requires a noninvariant potential [cf. the second factor in (32)] to produce the Laplacian. Clearly one could experiment with slightly different ordering to obtain a covariant form at all stages, but we expect that a coordinate-free expression right from the start would be the most efficient way to proceed.

V. DISCUSSION

Our goal in this article has been to develop further a new method for the introduction of stochastic processes into the solution of differential equations. This method was previously used to derive the Feynman–Kac formula for the solution of the Schrödinger equation with Hamiltonian $H = (1/2m)p^2 + V$. We have now shown how to handle situations where the coefficients of the derivative terms are themselves functions, the simplest case being Hamiltonians containing vector potentials. Under these circumstances we obtain the midpoint rule of the functional integral, although it is also clear how alternative forms (e.g., Ito rather than Stratonovitch) can arise. The natural appearance of the midpoint rule arises in our method because we introduce our stochastic variables in a manifestly Hermitian way—that is, sprinkling our $\partial/\partial x$'s symmetrically with respect to x-dependent terms.

Similar ideas have been developed for the case of a fourth-order elliptic operator of the type $-\Delta^2 - V$ in Ref. 7. But the path lives in the complex space. On the other hand, in Ref. 8, Hochberg has constructed signed measures of infinite total variation to treat the case of $-\Delta^2$, and in Ref. 9 the fundamental solutions of evolution equations have been used to construct formal measures on path spaces. However, these authors do not consider any disentangling of products of noncommuting operators, which is one of the natural goals for developing path integral formulas.

ACKNOWLEDGMENTS

One of us (LS) wishes to thank the University of Paris VI for its kind hospitality.

This work was supported in part by the United States Department of Energy and by NSF Grant Nos. 85-18806 and 88-11106.

[1] B. Gaveau and L. S. Schulman, Phys. Rev. D **36**, 1135 (1987).

[2] L. S. Schulman, in *Path Summation: Achievements and Goals*, edited by S. O. Lundqvist, A. Ranfagni, V. Sa-yakanit, and L. S. Schulman (World Scientific, Singapore, 1988).

[3] If V were present we would first apply the Trotter formula to separate it from the $(\partial - a)^2$ terms and subsequently take the square root of the $(\partial - a)^2$ terms, as described in the text.

[4] Further recent justification of this noncontroversial assertion (within the context of phase space path integrals) can be found in H. Fukutaka and T. Kashiwa, Prog. Theor. Phys. **80**, 151 (1988). Of course preferences for mathematically equivalent formulations are necessarily matters of taste. One such subjective argument for the midpoint rule is the fact that gauge invariance is manifest. For details see L. S. Schulman, in *Functional integration and its applications*, edited by A. M. Arthurs (Clarendon, Oxford, 1975). See also M. M. Mizrahi, J. Math. Phys. **16**, 2201 (1975). Note, however, that with the Ito integral one does not lose gauge invariance. It is retained when one uses the additional second derivative term in the Ito formula. Historically the midpoint preference goes back to the first paper on the path integral [R. P. Feynman, Rev. Mod. Phys. **20**, 367 (1948)], where Feynman states (p. 376) that any other prescription would change the wave equation. Use of the Ito formula (not yet discovered at the time of Feynman's writing) would allow the restoration of the wave equation but, again invoking subjectivity, such use seems perverse in the quantum context. See also L. S. Schulman, "Feynman and Beyond," in the *Proceedings of the Third meV to MeV Path Integral Conference, 1989* (World Scientific, Singapore, to be published).

[5] With the opposite factoring scheme [alluded to following (15)] the increment multiplying **a** is entirely absent from **a**'s argument.

[6] D. W. McLaughlin and L. S. Schulman, J. Math. Phys. **12**, 2520 (1971).

[7] B. Gaveau and Ph. Sainty, Lett. Math. Phys. **15**, 345 (1988).

[8] K. Hochberg, Ann. Probab. **6**, 433 (1978).

[9] L. Ehrenpreis, Proceedings of the International Congress of Mathematicians (Stockholm, 1962); Y. Daletskii, Russian Math. Surveys **22**, 1 (1967).

A classical particle in heat bath under the influence of external noise

J. Mencia Bravo
Departamento de Matemática Aplicada III, Universidad Politécnica de Cataluña, Avd. Alcalde Rovira Roure, 177, 250086 Lerida, Spain

R. M. Velasco
Departamento de Fisica, UAM Iztapalapa, 09340 México D.F., Mexico

J. M. Sancho
Departamento de Estructura y Constituyentes de la Materia, Universidad de Barcelona, Diagonal 647, 08028 Barcelona, Spain

(Received 4 January 1989; accepted for publication 12 April 1989)

A very simple model of a classical particle in a heat bath under the influence of external noise is studied. By means of a suitable hypothesis, the heat bath is reduced to an internal colored noise (Ornstein–Uhlenbeck noise). In a second step, an external noise is coupled to the bath. The steady state probability distributions are obtained.

I. INTRODUCTION

Our interest and the main purpose of this paper is the study of a system under the influence of internal and external fluctuations from the microscopic point of view of Hamiltonian dynamics.

The study of a system interacting with its surroundings is an interesting problem and has deserved great attention.[1] The methods to derive the equations of motion of such a system have been very different; they go from the projection operator technique[2] to the direct elimination of the heat bath surrounding the system.[3–9] The main results arising in such methods are expressed through generalized Langevin equations or the corresponding kinetic equations for the probability density. Those methods start with a selection of the relevant variables for the system, and they eliminate the irrelevant ones to obtain the behavior of the quantities we are really interested in. The relevant variable equations of motion contain some characteristics of the irrelevant variables which manifest themselves as a noise. This is usually called internal noise.

On the other hand, the behavior of a system driven by an external noise has also been studied through the introduction of stochastic terms in the phenomenological equation of motion of the relevant variables. To be specific, let us think of the equation of motion for a relevant variable x,

$$\dot{x} = v(a,x), \tag{1.1}$$

where a is the parameter which will be stochastically driven. Then Eq. (1.1) becomes

$$\dot{x} = v(a + \mu(t),x), \tag{1.2}$$

where $\mu(t)$ is a stochastic process or noise with well-defined statistics. Equation (1.2) is called a stochastic differential equation. The problems associated with this kind of equation have been extensively studied in the literature.[10]

The joint study of internal and external noises was considered by means of a master equation.[11] The internal noise was scaled with the size of the system, and the external noise was introduced through the parameters of the probability transitions. This approach does not incorporate the opportunity to study the effect of thermal noise.

In this work, we are interested in the following explicit problem: A classical particle is immersed in a heat bath of normal modes under the influence of an external stochastic field. In Sec. II we study this simple system without external noise in order to obtain a new representation for the internal noise. Section III is devoted to the study of this system when the heat bath is coupled with an external delta-correlated noise, and in Sec. IV we consider the coupling with an Ornstein–Uhlenbeck (OU) external noise. In all those cases, we found the stationary solution for the probability density. Finally, we make some comments about the results we have obtained.

II. INTERNAL NOISE

We start with the classical problem of a particle of mass M coupled to a heat bath of N normal modes. The Hamiltonian is given by

$$H = \frac{p^2}{2M} + V(x) + \frac{1}{2}\sum_{\alpha}^{N}\left\{\frac{p_\alpha^2}{m_\alpha} + m_\alpha w_\alpha^2 (q_\alpha - a_\alpha(x))^2\right\}, \tag{2.1}$$

where x, p are the coordinate and the momentum of the particle, respectively, (q_α,p_α) are the variables associated to the α normal mode, w_α is the corresponding frequency, and the quantity $a_\alpha(x)$ measures the interaction between the particle and the bath. $V(x)$ is the potential energy of the particle.

The method we will follow here was developed by Zwanzig,[8] and applied to several systems by Lindenberg *et al.*[9] The Hamilton equations for the normal mode variables (q_α,p_α) are immediately solved and their direct substitution in the Hamilton equations for the particle gives us [8,9]

$$\dot{x} = p/M, \tag{2.2}$$

$$\dot{p} = -V'(x) - \frac{1}{M}\int_0^t dt'\beta(t')P(t-t') + f_0(t), \tag{2.3}$$

where we have assumed that the interaction between the particle and the heat bath is linear,

 0022-2488/89/092023-05$02.50

$$a_\alpha(x) = \theta_\alpha x. \tag{2.4}$$

The last term on the rhs of Eq. (2.3) is the fluctuating force,

$$f_0(t) = \sum_1^N \theta_\alpha \{[q_\alpha(0) - a_\alpha(x(0))] m_\alpha w_\alpha{}^2 \cos(w_\alpha t) + p_\alpha(0) w_\alpha \sin(w_\alpha t)\}. \tag{2.5}$$

It depends on the coupling function θ_α and the normal mode initial conditions, which we will assume to be canonically distributed. The statistical properties of this fluctuating force are determined according to that assumption. At first we notice that $f_0(t)$ has a Gaussian distribution function with zero mean, and the correlation satisfies a fluctuation–dissipation relation,

$$\langle f_0(t) f_0(t') \rangle = K_B T \beta(t - t'), \tag{2.6}$$

$$\beta(t) = \sum_1^N m_\alpha w_\alpha{}^2 \theta_\alpha{}^2 \cos(w_\alpha t). \tag{2.7}$$

It is well known that a nonlinear interaction between the normal modes and the particle leads us to an equation with multiplicative noise.[9] But here we are much more interested in the analysis of internal noise properties. In order to accomplish this fact, we realize that the number of the normal modes of the heat bath should be very large ($N \gg 1$), and we will also assume that the frequencies are distributed according to a Lorentzian function. This assumption resembles broadly the behavior of the hydrodynamical modes in a macroscopic system.[12]

The frequency distribution we propose is

$$g(w) = 2N/\pi\tau(w^2 + \tau^{-2}), \tag{2.8}$$

where τ^{-1} is the cutoff frequency and the distribution is normalized to the number of normal modes.

The coupling function $\theta_\alpha = \theta(w)$ is assumed to scale with the number of oscillators N in such a way that the final results were independent of N,

$$\theta(w) = \theta_0/\sqrt{N\tau}\, w \ (\theta_0 \text{ is a constant}). \tag{2.9}$$

The masses m_α are put all equal to m. The preceding assumptions transform our internal noise $f_0(t)$ in an OU process. In particular, the correlation function (2.6) is obtained transforming the sum in (2.7) in a w integral using (2.8) as a weight. The explicit expression we get is

$$\langle f_0(t) f_0(t') \rangle = (m K_B T \theta_0{}^2/\tau) e^{-|t-t'|/\tau}, \tag{2.10}$$

where $m K_B T \theta_0^2$ is the intensity of this noise and τ (the inverse of the cutoff frequency) is its correlation time.

It is well known that the OU noise becomes a delta-correlated noise when the correlation time τ goes to zero, i.e., when the relaxation time of the normal modes is very small compared with the macroscopic time of the particle.

The set of Langevin equations (2.2), (2.3) and the statistical properties of the internal noise (2.10) define a non-Markovian problem. To circumvent this difficulty, we use the well-known procedure of expanding the variables' space in order to have a set of equations without the memory function and a delta-correlated noise.[13,14]

An equivalent set of equations to (2.2), (2.3), and (2.10) can be written as

$$\begin{aligned} \dot{x} &= p/M, \\ \dot{p} &= -V'(x) + R(t), \\ \dot{R} &= -\frac{R}{\tau} - \frac{m\theta_0{}^2}{M\tau} p + \frac{\Gamma(t)}{\tau}, \end{aligned} \tag{2.11}$$

where

$$R(t) = -\frac{1}{M} \int_0^t dt' \beta(t - t') P(t') + f_0(t). \tag{2.12}$$

$R(t)$ is an additional variable and $\Gamma(t)$ is a Gaussian, zero mean, and delta-correlated noise,

$$\langle \Gamma(t) \Gamma(t') \rangle = 2m K_B T \theta_0{}^2 \delta(t - t'). \tag{2.13}$$

The evolution equation for the probability density $P(x,p,R;t)$ is the following Fokker–Planck equation:

$$\begin{aligned} \frac{\partial P}{\partial t} = &-\frac{p}{M}\frac{\partial P}{\partial x} + (V'(x) - R)\frac{\partial P}{\partial p} \\ &+ \frac{\partial}{\partial R}\left(\frac{R}{\tau} + \frac{m\theta_0{}^2}{M\tau} p\right) P \\ &+ K_B T \frac{m\theta_0{}^2}{\tau^2} \frac{\partial^2}{\partial R^2} P. \end{aligned} \tag{2.14}$$

The stationary solution $P_{st}(x,P,R)$ is given by

$$P_{st}(x,p,R) \approx \exp\left\{ -\frac{1}{K_B T}\left(\frac{p^2}{2M} + V(x) + \frac{\tau R^2}{2m\theta_0{}^2}\right)\right\}. \tag{2.15}$$

It is a canonical distribution corresponding to the heat bath temperature, modified by the characteristic of the internal noise and the coupling of the heat bath with the particle through the variable $R(t)$. In spite of those couplings, it is also easy to prove that they are not relevant to studying the statics of the particle, because the variable $R(t)$ can be eliminated by a simple integration

$$\begin{aligned} P_{st}(x,p) &= \int dR\, P_{st}(x,p,R) \\ &\approx \exp\left\{ -\frac{1}{K_B T}\left(\frac{p^2}{2M} + V(x)\right)\right\}. \end{aligned} \tag{2.16}$$

The result is the usual Maxwell–Boltzmann distribution function as we could expect. From the point of view of the stationary solution, it does not matter whether $\beta(t)$ is delta-correlated or not. The difference between those two cases will be in the dynamics of the system.

III. INTERNAL VERSUS EXTERNAL DELTA-CORRELATED NOISE

We are interested now in the effects caused by the presence of an external noise. If we start considering a direct coupling between the external noise and the particle coordinates through a parameter in the potential, then we will get the same results obtained when we introduce an external noise in a system described by a generalized Langevin equation. We do not study that problem here, because its approach is standard.[10]

The interesting case occurs when the coupling with the external noise is through the heat bath. The interaction we will study here is also the simplest one. It is linear in the bath

coordinates. The Hamiltonian (2.1) is modified by adding a new term,

$$H_{\text{int}} = \frac{1}{2}\sum_{l}^{N} \phi_\alpha q_\alpha \epsilon(t). \tag{3.1}$$

The function $\phi_\alpha(t)$ measures the interaction intensity and $\epsilon(t)$ is the external noise which we will assume to be Gaussian. The correlation function of $\epsilon(t)$ will be specified later.

The elimination of the heat bath variables follows the same way as in the previous section, leading to the generalized Langevin equation

$$\dot{x} = p/M, \tag{3.2}$$

$$\dot{p} = -V'(x) - \frac{1}{M}\int_0^t dt'\,\beta(t')p(t-t') + f_0(t) + \pi(t). \tag{3.3}$$

The quantities $\beta(t)$ and $f_0(t)$ are the same as in (2.5) and (2.7). Note that $\pi(t)$ is a fluctuating force related with the external noise through

$$\pi(t) = -\int_0^t \Phi(t-t')\epsilon(t_1)dt', \tag{3.4}$$

where

$$\Phi(t) = \sum_{l}^{N} \theta_\alpha \phi_\alpha w_\alpha \sin(w_\alpha t). \tag{3.5}$$

The statistical properties of $\pi(t)$ are determined by the normal mode distribution (2.8), the couplings of the system with the bath (2.4) and the bath with the external noise (3.1), and the external noise itself.

Now the coupling function $\phi_\alpha = \phi(w)$ in (3.1) is also chosen to scale with the system size in order to obtain a finite result in the continuous limit,

$$\phi(w) = \sqrt{\tau N^{-1}}\, w\Phi_0 \quad (\Phi_0 = \text{constant}). \tag{3.6}$$

A direct substitution of Eqs. (2.8), (2.9), and (3.6) in Eq. (3.5) leads to

$$\Phi(t) = (\theta_0\Phi_0/\tau)e^{-t/\tau}, \tag{3.7}$$

which allows the calculation of the correlation function for the $\pi(t)$ noise. So far, the calculation we have done in this section is independent of the correlation function for the external noise $\epsilon(t)$. To continue, it is necessary to specify that function. First we will assume that $\epsilon(t)$ is a delta-correlated noise,

$$\langle \epsilon(t)\epsilon(t')\rangle = 2D\delta(t-t'). \tag{3.8}$$

By definition, it is independent of the internal noise properties. Then

$$\langle \epsilon(t) f_0(t')\rangle = 0. \tag{3.9}$$

The correlation function of $\pi(t)$ noise is immediately calculated and it is given by

$$\langle \pi(t)\pi(t')\rangle = D(\theta_0\Phi_0)^2\tau^{-1}e^{-|t-t'|/\tau}, \tag{3.10}$$

where we have neglected the transient terms ($t,t' > \tau$). Equation (3.10) shows how the heat bath dresses the external noise. Although the external noise is a delta-correlated one, the particle sees it as an external OU noise with the same correlation time as the internal noise but with an intensity depending on the couplings and the external noise intensity.

To construct the Fokker–Planck equation, we will follow the same procedure as in Sec. II with few changes. Equation (3.3) has two noises and according to Eqs. (2.10) and (3.10), both have the same correlation time. We define an effective Gaussian noise $\sigma(t) = f_0(t) + \pi(t)$, which has zero mean and a correlation function given by

$$\langle \sigma(t)\sigma(t')\rangle = (mK_{\text{B}}T + D\Phi_0^2)\theta_0^2\tau^{-1}e^{-|t-t'|/\tau}. \tag{3.11}$$

Note that $\sigma(t)$ is also an OU noise, but there is not a fluctuation dissipation relation because the external noise is present.

The set of equivalent Markovian equations is the same as (2.11) but the intensity of the delta-correlated noise is given now by $(mK_{\text{B}}T + D\Phi_0^2)\theta_0^2$, which changes the diffusion coefficient in the corresponding Fokker–Planck equation,

$$\frac{\partial P}{\partial t} = -\frac{p}{M}\frac{\partial P}{\partial x} + (V'(x) - R)\frac{\partial p}{\partial \rho} + \frac{\partial}{\partial R}\left(\frac{R}{\tau} + \frac{m\theta_0^2}{M\tau}p\right)P + (mK_{\text{B}}T + D\Phi_0^2)\frac{\theta_0^2}{\tau^2}\frac{\partial^2 P}{\partial R^2}. \tag{3.12}$$

The stationary solution has the same qualitative features as Eq. (2.14), but the "temperature" has changed according to the new diffusion coefficient

$$P_{\text{st}}(x,p,R) \approx \exp\left\{-\frac{1}{K_{\text{B}}T + (D/m)\Phi_0^2}\left(\frac{p^2}{2M} + V(x) + \frac{\tau}{2m\theta_0^2}R^2\right)\right\}. \tag{3.13}$$

This is a canonical distribution with an effective diffusion which depends on the external noise intensity and the coupling between the noise and the heat bath.

Once again the correlation time τ coming from the internal noise is irrelevant to the statics because the integration in the additional R variable is not coupled to the variables (x,p), giving us the usual exponential distribution function (2.15) but with a new diffusion coefficient.

IV. INTERNAL VERSUS EXTERNAL ORNSTEIN–UHLENBECK NOISE

In this section, we consider that the external noise is an OU process with a correlation function given by

$$\langle \epsilon(t)\epsilon(t')\rangle = (D/\tau')e^{-|t-t'|/\tau'}, \tag{4.1}$$

where D and τ' are the intensity and the correlation time,

respectively. A direct calculation gives us the correlation function of $\pi(t)$,

$$\langle \pi(t)\pi(t')\rangle = \frac{D(\theta_0\Phi_0)^2}{(\tau'/\tau)^2-1}\frac{\tau'}{\tau}\left\{\frac{1}{\tau}\exp\left(-\frac{|t-t'|}{\tau'}\right) - \frac{1}{\tau'}\exp\left(-\frac{|t-t'|}{\tau}\right)\right\}, \tag{4.2}$$

where we have neglected the transient terms. The dressed external noise $\pi(t)$ now has a more complicated correlation function with two correlation times, τ and τ'. In order to see what is the meaning of Eq. (4.2), let us assume that the external correlation time is bigger than the internal one. Then the dressed noise is dominated by the external noise,

$$\langle \pi(t)\pi(t')\rangle = [D(\theta_0\Phi_0)^2/\tau']e^{-|t-t'|/\tau'}. \tag{4.3}$$

On the other hand, when the external correlation time is smaller than the internal one, we then recover (3.10).

Now let us see what is the physical situation corresponding to (4.3). The heat bath degrees of freedom act on the system as an internal noise, and relax with a characteristic time τ. A physical intuition leads us to think that the internal degrees of freedom should relax faster than any real external noise. So we will take a delta-correlated internal noise ($\tau = 0$), and then the dressed noise correlation is dominated by the external noise (4.3). The set of Langevin equations (3.2) and (3.3) is simplified,

$$\dot{x} = p/M,$$
$$\dot{p} = -V'(x) - (m\theta_0^2/M)p + f_0(t) + \pi(t), \tag{4.4}$$

where we have taken the white noise limit for $f_0(t)$ [$\tau = 0$ in (2.10)].

Note that

$$\langle f_0(t)f_0(t')\rangle = 2K_B T m\theta_0^2\delta(t-t') \tag{4.5}$$

and the term $-m\theta_0^2 p/M$ in Eq. (4.4) comes from the memory function which is now $\beta(t-t') = 2m\theta_0^2\delta(t-t')$.

To study the problem described by the set of Eqs. (4.4), we define an effective Gaussian OU noise,

$$\Omega(t) = f_0(t) + \pi(t), \tag{4.6}$$

with an intensity D_R and a correlation time τ_R given by

$$D_R = \int_0^\infty <\Omega(t)\Omega(0)dt, \tag{4.7}$$

$$\tau_R = \frac{1}{D_R}\int_0^\infty t<\Omega(t)\Omega(0)dt. \tag{4.8}$$

These definitions give

$$D_R = m\theta_0^2(K_B T + D\Phi_0^2/m), \tag{4.9}$$
$$\tau_R = D(\theta_0\Phi_0)^2\tau'/D_R,$$

which are some combinations of the parameters characterizing the noises and their interactions. The problem has been reduced to a set of equations with an effective OU noise. The stationary solution has been discussed in the literature.[15]

The stationary distribution is now

$$P_{st}(x,p) \approx \exp\left\{-\frac{p^2}{D_p 2M} + \frac{V(x)}{D_x}\right\}, \tag{4.10}$$

where

$$D_p = \frac{D_R}{m\theta_0^2(1+m\theta_0^2\tau_0/M)}, \quad D_x = \frac{D_R}{m\theta_0^2}. \tag{4.11}$$

Now we have different values for the p-diffusion and the x-diffusion coefficients. We notice that the D_x coefficient is the same as in Eq. (3.15), but the D_p coefficient has changed again and it shows a correction due to the external noise correlation time. Finally, we mention that τ' can not be eliminated as it was done in earlier cases. The interaction between internal and external noises will now effect the static properties of our system.

V. CONCLUDING REMARKS

Here we summarize the results obtained in this paper. At first we considered the simplest Hamiltonian system to study the behavior of an open subsystem immersed in a heat bath. The heat bath coordinates were eliminated and we found the way to model an Ornstein–Uhlenbeck noise. In this case, we have obtained that the internal noise correlation time is not a relevant quantity in the stationary solution of the Fokker–Planck equation. The delta-correlated external noise, coupled with the heat bath and with an OU internal noise, also has that property. But the diffusion coefficient was modified by the coupling parameters, and as a consequence there is not a fluctuation–dissipation relationship.

From the point of view of a more realistic situation, we have found that the most interesting case corresponds to a delta-correlated internal noise and an OU external noise. The stationary solution has two renormalized diffusion coefficients, one of them depending explicitly of the external noise correlation time. Obviously the fluctuation–dissipation relationship does not hold, because the external noise is present.

The approach presented here opens the possibility of studying more complicated systems with nonlinear couplings.

ACKNOWLEDGMENTS

J.M.B. thanks M. San Miguel for helping in his introduction to this field.

Partial financial support from the program CONACYT-CSIC (Mexico–Spain) is acknowledged. Two of us (J.M.B. and J.M.S.) also acknowledge the financial support of Direccíon General de Investigacíon Científica y Técnica, Project no. PR87-0014 (Spain).

[1] R. Zwanzig, in *Lectures in Theoretical Physics*, edited by W. E. Brittin, B. W. Downs, and J. Downs (Interscience, New York, 1961), Vol.3; H. Mori, Prog. Theor. Phys. **34**, 423 (1965).
[2] R. Zwanzig, Phys. Rev. **124**, 983 (1961).
[3] H. Grabert, *Projection Operator Techniques in Nonequilibrium Statistical Mechanics* (Springer, Berlin, 1982).
[4] P. Grigolini, in *Advances in Chemical Physics*, edited by M. W. Evans, P. Grigolini, and G. Pastori Parravicini (Wiley, New York, 1985).
[5] N. G. Van Kampen and I. Oppenheim, Physica A **138**, 231 (1986).
[6] J. J. Brey, J. M. Casado, and M. Morillo, Physica A **121**, 122 (1984).
[7] G. W. Ford, M. Kac, and P. Mazur, J. Math. Phys. **6**, 504 (1965).
[8] R. Zwanzig, J. Stat. Phys. **9**, 215 (1973).
[9] K. Lindenberg and V. Seshadri, Physica A **109**, 483 (1981); V. Seshadri and K. Lindenberg, *ibid.* A **115**, 501 (1982).

[10]W. Horsthemke and R. Lefever, *Noise Induced Transitions* (Springer, Berlin, 1984).
[11]M. San Miguel and J. M. Sancho, Phys. Lett. A **90**, 455 (1982); J. M. Sancho and M. San Miguel, J. Stat. Phys. **37**, 151 (1984).
[12]P. Resibois and M. de Leener, *Classical Kinetic Theory of Fluids* (Wiley–Interscience, New York, 1977).
[13]F. Marchesoni and P. Grigolini, J. Chem. Phys. **78**, 6287 (1983).
[14]H. Risken, *The Fokker–Planck Equation, Methods of Solution and Applications* (Springer, Berlin, 1984).
[15]L. Ramirez-Piscina and J. M. Sancho, Phys. Rev. A **37**, 4469 (1988).

Bound on the mass gap for a stochastic contour model at low temperature

Lawrence E. Thomas
Department of Mathematics, University of Virginia, Charlottesville, Virginia 22903

(Received 28 December 1988; accepted for publication 19 April 1989)

A two-dimensional stochastic contour model is considered in which the state space consists of simple closed lattice contours about a fixed origin and the evolution in the state space is a continuous time reversible jump process. An upper bound is obtained on the mass gap for the model, which goes rapidly to zero, in a low temperature limit. A connection between this model and statistical mechanical and droplet models is discussed.

I. INTRODUCTION

This paper concerns a simple model for *stochastic geometry*, i.e., a stochastic process in which the state space is a space of curves or a space of (hyper-) surfaces. The motivation for considering such processes comes from various examples in statistical physics, field theory, and nonequilibrium phenomena: We have in mind the Peierls' contours or surfaces of Ising and lattice gauge models, interfaces in dynamical chemical reaction models, and droplet models.[1-7] At the level of computational physics, certain Monte Carlo algorithms for treating problems of statistical physics and, for example, problems of self-avoiding random walk can be regarded as processes of the above sort (cf. Refs. 8–15). Mathematically, we can ask familiar questions about these processes, questions concerning ergodicity, typical geometries, etc.

Here, we consider what may be about the simplest process, for which the state space consists of simple closed lattice contours in two dimensions and the process is a continuous time evolution (jump process) on this state space. Physically, the dynamics are those of a model for a two-dimensional droplet in which the contour is the boundary of the droplet. The dynamics bear some relation to those of Glauber for the two-dimensional stochastic Ising model at low temperature.[16] The latter can be thought of as a "gas" of Peierls' contours evolving in time; if we disregard interactions between contours and instead concentrate on the evolution of a single contour, we obtain a model which is essentially that considered here. (Insisting that the contours encircle a fixed origin is just to break translational invariance.) The evolution of the model will be given by a self-adjoint semigroup, so that the process is reversible. We will be primarily concerned with the rate at which the process approaches equilibrium.

Let X be the space of simple closed lattice contours $\{\gamma\}$ of arbitrary length on $\mathbb{Z}^2$ (the two-dimensional integer lattice in $\mathbb{R}^2$) which bound an area containing a fixed origin, say at $(\frac{1}{2},\frac{1}{2})$, in $\mathbb{R}^2$. For convenience, we will assume X contains the null contour $\gamma=\varnothing$ as well. A non-null contour of X consists of unit length line segments between integer lattice points; since γ is simple, an integer lattice point p in γ has exactly two unit line segments hitting p. In statistical mechanics language, X is the space of Peierls' contours about a fixed site.

Let π_β be the probability measure on X (defined for β sufficiently large),

$$\pi_\beta\{\gamma\}=\mathbb{Z}_{(\beta)}^{-1}e^{-\beta|\gamma|}, \tag{1.1}$$

where $|\gamma|$ is the length of γ and

$$\mathbb{Z}(\beta)=\sum_{\gamma\in X}e^{-\beta|\gamma|} \tag{1.2}$$

is the normalizing partition function.

We define a non-negative self-adjoint operator G acting in the Hilbert space $\mathscr{H}=l^2(X,\pi_\beta)$ with the inner product $\langle\,,\rangle$ (cf. Ref. 8). Let $j(\gamma,\gamma')$ be non-negative, uniformly bounded *speed functions* which, moreover, satisfy the following conditions.

(i) *Local motion condition:* $j(\gamma,\gamma')=0$ unless $\gamma\Delta\gamma'$ is the perimeter of a unit lattice square.

(ii) *Detailed balance condition:* For all $\gamma,\gamma'\in X$,

$$\pi_\beta(\gamma)j(\gamma,\gamma')=\pi_\beta(\gamma')j(\gamma',\gamma). \tag{1.3}$$

Under the hypotheses (i) and (ii) the operator G having the Dirichlet form

$$\langle f,Gf\rangle=\frac{1}{2}\sum_{\gamma,\gamma'}j(\gamma,\gamma')|f(\gamma')-f(\gamma)|^2\pi_\beta(\gamma) \tag{1.4}$$

generates a (self-adjoint) Markov semigroup $\exp(-tG)$ which corresponds to a reversible Markovian jump process $\gamma(t)$ and has π_β as an invariant measure (see Ref. 8 for details). On its domain $\mathscr{D}(G)$,

$$Gf(\gamma)=-\sum_{\gamma'}j(\gamma,\gamma')(f(\gamma')-f(\gamma)),\quad f\in\mathscr{D}(G). \tag{1.5}$$

Clearly, G has $f\equiv 1$ as an eigenvector, with the eigenvalue zero. A natural question to ask is whether G has a nonzero mass gap m, i.e., whether $m=\inf \operatorname{spec} G|1^\perp>0$. If $m>0$, then the process approaches equilibrium exponentially fast. A related question concerns the random time τ for the process $\gamma(t)$ to shrink to $\gamma=\varnothing$, $\tau=\inf\{t\,|\gamma(t)=0\}$. If $\sup_\gamma|\gamma|^{-1}E_\gamma(\tau)=\infty$ [where $|\gamma|$ is the length of γ and $E_\gamma(\ )$ is the path space expectation starting at γ], then $m=0$.[8] In fact, in Ref. 8, it was shown that if the speed functions decay in the sense that

$$\lim_{n\to\infty}\frac{1}{n}\sup_{|\gamma|=n}\sum_{\gamma'}j(\gamma,\gamma')=0, \tag{1.6}$$

then the sup above is infinite and thus $m=0$.

 0022-2488/89/092028-07$02.50

Sokal and Thomas have conjectured that if the nonzero $j(\ ,\ )$'s do not decay, e.g., if

$$j(\gamma,\gamma') = c(\beta)\exp - (\beta/2)(|\gamma| - |\gamma'|)$$

for $\gamma\Delta\gamma'$ the boundary of a unit lattice square, then $E_\gamma(\tau) \sim A(\gamma)$, the area inside γ (Ref. 8); see, also, the arguments of Huse and Fisher[17] and Droz and Gunton,[18] who consider related problems. Assuming this area law to hold, it follows that there would be no mass gap. In fact, we are unable to prove this area law or that the mass gap is zero [the difficulties stemming from $\gamma(t)$ developing into a highly ramified configuration], although there is compelling numerical evidence for the former and hence the latter, at least for β large. We do show here, and this is the principal result of the paper, that if the speed functions remain bounded as $\beta\to\infty$, then the mass gap $m = m(\beta)$ goes to zero rapidly $m(\beta) \lesssim \beta e^{-4\beta}$; see Theorem 2.10, as well as the concluding remarks in Sec. II.

We also show that a certain operator G_∞, which can be regarded as the $\beta\to\infty$ limit of the operators $G(\beta)$ defined as above, has a corresponding process exhibiting at least an area law $E_\gamma(\tau) \gtrsim cA(\gamma)$. We remark that Marchand and Martin[3] and Rost[6] (see, also, an account of the latter work in Liggett[7]) have considered a related ($\beta = \infty$) process with, however, the initial configuration γ taken to be the boundary of an infinite quadrant and with motions that do not locally change the length of γ. Marchand and Martin and Rost do allow the speed functions $j(\gamma,\gamma')$ to be asymmetric, corresponding to an external magnetic field (whereas here G_∞ has only symmetric speed functions if $|\gamma| = |\gamma'|$): By relating the process to a one-dimensional exclusion process, these authors obtain the asymptotic shape of $\gamma(t)$. For negative magnetic field, $\gamma(t)$ has a limiting distribution[3]; for zero magnetic field (corresponding to symmetric speed functions), $t^{-1/2}\gamma(t)$ approaches a known convex curve in probability and the area of the quadrant is eaten away linearly in time; and finally, for positive magnetic field, $t^{-1}\gamma(t)$ approaches a known convex curve, almost surely.[6,7]

We conclude the introduction with some remarks concerning the means used to estimate the mass gap. Conceivably, the estimate could be shown by exhibiting trial functions orthogonal to $f_0 \equiv 1$ and having Dirichlet form expectation suitably small. However, we never succeeded with this strategy and so rather we employ a probabilistic approach, some ideas of which were employed in Refs. 2, 8, and 9. The main ideas of this approach are these: We consider a particular family of operators $G(\beta)$ in which the speed functions remain bounded, $\beta\to\infty$, and which for each β satisfy detailed balance. We then construct a sequence of perturbations of $G(\beta)$, $G_n(\beta)$ with $G_n(\beta) - G(\beta)$ compact (in fact, finite rank), non-negative [acting in $l^2(\pi_\beta)$], and such that each $G_n(\beta)$ is also a Markov generator. Finally, we obtain a *lower bound* on the expected time for the process associated with $G_n(\beta)$ to leave a certain finite domain $D_n \subset X$, from which we infer an upper bound on the mass gap of $G_n(\beta)$ and hence of $G(\beta)$. We note that our bound is actually a bound on what might be called the essential mass gap, i.e., the infimum of the essential spectrum of $g(\beta)$. In particular, the estimate is unchanged if a finite number of speed functions are changed.

II. EXPECTED ESCAPE TIME ESTIMATES AND STOCHASTIC CONTOUR MODELS

In this section we review at a somewhat abstract level a connection between an expected escape time and the mass gap and then go on to apply these ideas to the stochastic contour models. Some of the ideas appear in Ref. 8, but we include them here where they have been cast into a convenient form for the problems at hand.

A. Expected escape times and mass gaps

Let X be a countable set, π a probability measure on X, and G a self-adjoint Markov generator acting in $l^2(X,\pi)$ and having the Dirichlet form

$$\langle Gf,g\rangle = \frac{1}{2}\sum_x\sum_{x'} j(x,x')(f(x') - f(x))(g(x') - g(x))\pi\{x\} \tag{2.1}$$

and such that for f in the domain of G,

$$Gf(x) = -\sum_{x'} j(x,x')(f(x') - f(x)). \tag{2.2}$$

Here, the $j(x,x')$ are non-negative and satisfy the detailed balance

$$j(x,x')\pi\{x\} = j(x',x)\pi\{x'\} \tag{2.3}$$

and such that for each x,

$$\sum_{x'} j(x,x') < \infty. \tag{2.4}$$

Clearly, $f \equiv 1$ is an eigenvector of G, with the eigenvalue zero.

For D a finite set of X, let G_D be the operator obtained from G by imposing Dirichlet boundary conditions on the complement of D. In other words, G_D acts in $l^2(D,\pi)$ and has the Dirichlet form

$$\langle G_D f,g\rangle_{l^2(D,\pi)} = \langle G\tilde{f},\tilde{g}\rangle_{l^2(X,\pi)}, \tag{2.5}$$

where $\tilde{f},\tilde{g}$ are extensions of f and g, respectively, to all of X by setting $\tilde{f}(x) = \tilde{g}(x) = 0$, $x \notin D$.

Let m be the mass gap of G, i.e.,

$$m \equiv \inf \text{ spec } G|1^\perp \tag{2.6}$$

and set

$$m_D = \inf \text{ spec } G_D. \tag{2.7}$$

Lemma 2.1 (Proposition 3.3 of Ref. 8): The mass gap m for G satisfies

$$m\pi(D^c) \leqslant m_D. \tag{2.8}$$

Thus an upper bound on m_D provides an upper bound on m.

Let $x(t)$ be (the right continuous) jump process associated with the semigroup $\exp(-tG)$; henceforth, to avoid trivialities we will assume $x(t)$ is irreducible. Let τ_D be the time for $x(t)$ to escape from D, i.e., $\inf\{t\,|x(t) \notin D\}$. In the following, E will denote the path space expectation of the process. A bound relating the expected escape time to m_D is given by the following lemma.

Lemma 2.2: Assume $0 < \theta < 1$. Then m_D has an implicit upper bound in terms of the expected escape time starting at $x \in D$ given by

$$m_D < \theta E_x(\tau_D)^{-1}(\ln\|g\|_{l^2(D,\pi)} - \tfrac{1}{2}\ln(\pi\{x\}) - \ln((1-\theta)m_D)), \tag{2.9}$$

where

$$g(x) \equiv \sum_{x' \in D^c} j(x,x'). \tag{2.10}$$

Proof: If $m_D = 0$, there is nothing to prove. For $\epsilon < m_D$, the function $f_\epsilon(x) = E_x(e^{\epsilon\tau_D})$ satisfies the differential equation

$$(G - \epsilon)f_\epsilon(x) = 0, \quad x \in D, \tag{2.11}$$

with $f_\epsilon(x) = 1$, $x \in D^c$. [To see this, note that

$$M_t = f_\epsilon(x(t))e^{\epsilon t}$$

is a martingale and hence has an expectation constant in time. By optional stopping, we have that

$$f_\epsilon(x) = E_x(M_{\tau_D}) = E_x(f_\epsilon(X(\tau_D))e^{\epsilon\tau_D}) = E_x(e^{\epsilon\tau_D}).]$$

Now, Eq. (2.10) can be written as

$$(G_D - \epsilon)f_\epsilon(x) = g(x), \quad x \in D, \tag{2.11'}$$

with g defined by Eq. (2.10), so that

$$f_\epsilon(x) = (\pi\{x\})^{-1}\langle \delta_x, (G_D - \epsilon)^{-1} g\rangle_{l^2(D,\pi)} \leqslant (\pi\{x\})^{-1/2}(m_D - \epsilon)^{-1}\|g\|_{l^2(D,\pi)} \tag{2.12}$$

by the Schwarz inequality.

Also, by Hölder's inequality we have that

$$E_x(\tau_D) = \lim_{\rho \downarrow 0} \frac{1}{\rho}(E_x(e^{\rho\tau_D}) - 1) < \lim_{\rho \downarrow 0} \frac{1}{\rho}(E_x(e^{\theta m_D \tau_D})^{\rho/\theta m_D} - 1) = (\theta m_D)^{-1} \ln f_{\theta m_D}(x), \tag{2.13}$$

which, combined with inequality (2.12), gives the conclusion of Lemma 2.2. ∎

It remains to give to a means for obtaining a lower bound on the expected time to exit D in order to apply Lemmas 2.1 and 2.2. The reader should not be alarmed that the following Lyapunov function estimate involves m_D. We assume the complement of D to be decomposed into two disjoint pieces $D^c = D_{\mathrm{I}} \cup D_{\mathrm{II}}$ (one of which could be empty).

Lemma 2.3: Let $m_D > 0$. Suppose that there is a function $h(x)$ defined on X and a constant $c > 0$ independent of $x \in D$ such that

$$Gh(x) \leqslant c, \quad x \in D. \tag{2.14}$$

Then

$$E_x(\tau_D) \geqslant c^{-1}\left(h(x) - \sup_{y \in D_{\mathrm{I}}} h(y) - \pi^{-1/2}(x) m_D^{-1}\|k\|_{l^2(D,\pi)}\right), \tag{2.15}$$

where

$$k(x) \equiv \sum_{x' \in D_{\mathrm{II}}} j(x,x')h(x'). \tag{2.16}$$

Proof: Again we employ a martingale argument. Let M_t be the martingale

$$M_t = h(x(t)) + \int_0^t Gh(x(s))ds. \tag{2.17}$$

By optional stopping and by inequality (2.14),

$$h(x) = E_x(M_{\tau_D}) \leqslant E_x(h(x_{\tau_D})) + cE(\tau_D), \tag{2.18}$$

so that

$$cE(\tau_D) \geqslant h(x) - E_x(h(x(\tau_D))\chi_{\mathrm{I}}) - E_x(h(x(\tau_D))\chi_{\mathrm{II}}) \geqslant h(x) - \sup_{y \in D_{\mathrm{I}}} h(y) - E_x(h(x(\tau_D))\chi_{\mathrm{II}}), \tag{2.19}$$

where χ_{I} and χ_{II} are indicator functions for the events that $x(t)$ exits into D_{I} or D_{II}, respectively. It remains to estimate the last term on the rhs of inequality (2.19). Call this term $f(x)$. Then, by a martingale argument similar to those used above, it is easy to see that $f(x)$ satisfies

$$Gf(x) \equiv G_D f(x) - k(x) = 0, \quad x \in D, \tag{2.20}$$

with

$$f(x) = h(x), \quad x \in D_{\mathrm{II}}, \tag{2.21}$$

$$= 0, \quad x \in D_{\mathrm{I}}. \tag{2.22}$$

Thus,

$$f(x) = (\pi\{x\})^{-1}\langle \delta_x, G_D^{-1} k\rangle \leqslant (\pi\{x\})^{-1/2} m_D^{-1}\|k\|_{l^2(D,\pi)}, \tag{2.23}$$

which, combined with (2.19), gives the assertion of Lemma 2.3. ∎

In outline, we will bound the mass gap m of G (or actually an operator that majorizes G) as follows. If $m = 0$ there is nothing to prove; if $m > 0$, then m_D is *uniformly* bounded away from zero as a function of D provided that $\pi(D^c)$ is uniformly bounded away from zero. We will consider a sequence of domains D_n with this property and such that $|D_n| \to \infty$. It will turn out that the terms involving m_D on the rhs of inequalities (2.9) and (2.15) will be negligible relative to the other terms, so that the real issue is finding a suitable Lyapunov function $h(x)$ or, given $h(x)$, finding a suitable $G' \geqslant G$, for which $h(x)$ satisfies the hypotheses of Lemma 2.3.

B. Stochastic contour models

Here, we obtain an estimate on the expected escape time from various sets $D_n \subset X$ for some stochastic contour models and an upper bound on the mass gap for these models. Throughout this subsection $G(\beta)$ will be the particular generator, with the speed function $j(\gamma,\gamma') = j(\gamma,\gamma',\beta)$ equal to zero if $\gamma \Delta \gamma'$ is not the perimeter of a unit lattice square, and otherwise given by

$$j(\gamma,\gamma') = e^{-\beta \pm \beta} \quad \text{if} \quad |\gamma'| = |\gamma| \mp 2 = \tfrac{1}{2}(1 + e^{-2\beta}) \quad \text{if} \quad |\gamma'| = |\gamma| \tag{2.24}$$

and, although it plays no essential role in the analysis,

$$j(\gamma,\varnothing) = 1, \quad |\gamma| = 4, \qquad j(\varnothing,\gamma) = e^{-4\beta}, \quad |\gamma| = 4. \tag{2.25}$$

In fact, we will estimate some expected escape times for a sequence of operators $\{G_n(\beta)\}$, $G_n(\beta) \geqslant G(\beta)$, thus obtain-

ing an estimate on the mass gaps for $G_n(\beta)$ and hence the mass gap estimate for $G(\beta)$.

For $n = 1,2,...$, let γ_n be a square contour of sides n, so that $|\gamma_n| = 4n$ and $A(\gamma_n) = n^2$, where $A(\gamma)$ is the area closed by γ. Define the subsets D_n, S_n, and T_n [these are, for each n, the subsets D, D_{I}, and D_{II} of Lemma (2.3)] by

$$S_n = \{\gamma \subset X \,|\, A(\gamma) \leqslant \tfrac{1}{4}n^2\}, \tag{2.26}$$

$$T_n = \{\gamma \subset X \setminus S_n \,||\gamma| \geqslant 5n\}, \tag{2.27}$$

$$D_n = X \setminus (S_n \cup T_n). \tag{2.28}$$

Note that $\gamma_n \in D_n$. In what follows, $G_n = G_n(\beta)$ will denote the operator acting in $l^2(D_n,\pi)$ obtained from $G(\beta)$ by imposing Dirichlet boundary conditions on the complement of D_n, τ_n will denote the escape time for the process associated with $\exp(-tG(\beta))$ from D_n, and $m_n = m_n(\beta) = \inf \operatorname{spec} G_n(\beta)$. Throughout this subsection, the role of the Lyapunov function h of Lemma 2.3 will be played by the area $h(\gamma) = A(\gamma)$.

Our immediate goal is to obtain an upper bound on $G_n A(\gamma)$ in terms of the geometry of γ. Given γ, we will say that a unit lattice square q (which we assume includes its edges and vertices) is *inside* or *outside* γ according to whether q is contained in the (closed) region bounded by γ or not. We set, for $m = 1,2,3$,

$J^0_m(\gamma)$ [resp., $J^i_m(\gamma)$]

= number of unit lattice squares q *outside* (resp., *inside*) γ such that $q \cap \gamma$ consists of m unit length edges of γ. In the case $m = 2$, the two edges should be connected at a common vertex of γ.

The J's are in effect counting the number of transitions immediately possible from γ. We also need to enumerate some of the *forbidden* transitions, which are forbidden in the sense that they lead to a γ' not in X. We set

$F^{1,0}_1(\gamma)$ [resp., $F^{1,i}_1(\gamma)$]

= number of unit lattice squares *outside* (resp., *inside*) γ such that $q \cap \gamma$ is the union of one unit length edge of γ along with one or two lattice vertices in γ disjoint from the edge,

$F^{2,0}_1(\gamma)$ [resp., $F^{2,i}_1(\gamma)$]

= number of unit lattice squares *outside* (resp., *inside*) γ such that $q \cap \gamma$ is the union of two (parallel) unit length edges of γ which are disjoint,

$F^0_1(\gamma)$ [resp., $F^i_1(\gamma)$]

$= F^{1,0}_1(\gamma) + 2F^{2,0}_1$ [resp., $F^{1,i}_1(\gamma) + 2F^{2,i}_1(\gamma)$],

$F^0_2(\gamma)$ [resp., $F^i_2(\gamma)$]

= number of unit lattice squares *outside* (resp., *inside*) γ such that $q \cap \gamma$ is the union of two unit length edges of γ connected at a common lattice point, together with a lattice point of γ, disjoint from the edges.

Lemma 2.4: The following relations hold:

$$J^0_1(\gamma) + 2J^0_2(\gamma) + 3J^0_3(\gamma) + F^0_1(\gamma) + 2F^0_2(\gamma) = |\gamma|, \tag{2.29}$$

$$J^i_1(\gamma) + 2J^i_2(\gamma) + 3J^i_3(\gamma) + F^i_1(\gamma) + 2F^i_2(\gamma) = |\gamma|, \tag{2.30}$$

$$J^0_2(\gamma) - J^i_2(\gamma) + 2J^0_3(\gamma) - 2J^i_3(\gamma) + F^0_2(\gamma) - F^i_2(\gamma) = -4. \tag{2.31}$$

Proof: Relations (2.29) and (2.30) are simply an enumeration of the unit lattice squares outside (resp., inside) and tangent to γ, taking into account the length of the tangency.

Relation (2.31) is essentially the Gauss–Bonnet theorem for lattice contours: If γ is traversed in a counterclockwise direction, then the total number of right turns minus the number of left turns is -4. Associated with J^0_2 squares and F^0_2 squares are $J^0_2(\gamma) + F^0_2(\gamma)$ right turns; with J^0_3 squares are associated $2J^0_3(\gamma)$ right turns. Similar remarks hold for inside squares. ■

Lemma 2.5: For $\gamma \in X$, we have that

$$GA(\gamma) \leqslant 2(1 - 3e^{-2\beta}) + \tfrac{1}{2}(1 + e^{-2\beta})(F^0_2(\gamma) - F^i_2(\gamma)) + e^{-2\beta}(F^0_1(\gamma) - F^i_1(\gamma)). \tag{2.32}$$

Proof: Since $A(\gamma) - A(\gamma') = \pm 1$ according to whether $\gamma \Delta \gamma'$ is the perimeter of a lattice unit square outside or inside γ, we have that

$$-GA(\gamma) \geqslant e^{-2\beta}(J^0_1(\gamma) - J^i_1(\gamma)) + \tfrac{1}{2}(1 + e^{-2\beta}) \times (J^0_2(\gamma) - J^i_2(\gamma)) + J^0_3(\gamma) - J^i_3(\gamma). \tag{2.33}$$

(We have inequality rather than equality since conceivably $\gamma \Delta \gamma'$ could encircle the origin, which is a forbidden transition.) Subtracting relation (2.30) from (2.29) we obtain

$$J^0_1(\gamma) - J^i_1(\gamma) = -2(J^0_2(\gamma) - J^i_2(\gamma)) - 3(J^0_3(\gamma) - J^i_3(\gamma)) - (F^0_1(\gamma) - F^i_1(\gamma)) - 2(F^0_2(\gamma) - F^i_2(\gamma)). \tag{2.34}$$

Thus the rhs of (2.33) equals

$$\begin{aligned} &\tfrac{1}{2}(1 - 3e^{-2\beta})(J^0_2(\gamma) - J^i_2(\gamma) + 2(J^0_3(\gamma) - J^i_3(\gamma))) \\ &\quad - e^{-2\beta}(F^0_1(\gamma) - F^i_1(\gamma) + 2F^0_2(\gamma) - 2F^i_2(\gamma)) \\ &= -2(1 - 3e^{-2\beta}) - \tfrac{1}{2}(1 + e^{-2\beta})(F^0_2(\gamma) - F^i_2(\gamma)) \\ &\quad - e^{-2\beta}(F^0_1(\gamma) - F^i_1(\gamma)) \end{aligned}$$

by relation (2.31). This completes the proof of Lemma (2.5). ■

Thus $Ga((\gamma))$ fails to be bounded above because of "intrusions" into γ, resulting in a deficiency of possible immediate outward transitions (the F^0_1 and F^0_2 terms count these deficiencies) unless the F^i_1 and F^i_2 terms compensate. Hence, the basic idea is to perturb G with some additional (nonlocal) transitions which are positive and which, in effect, reduce the rhs of Eq. (2.32) for those γ's where this quantity is large and positive. The mass gap for the new operator, which we then estimate by Lemmas 2.1, 2.2, and 2.3, provides a bound on the mass gap for G.

To further amplify these remarks, suppose that for each γ with

$$U(\gamma) \equiv \tfrac{1}{2}(1 + e^{-2\beta})(F^0_2(\gamma) - F^i_2(\gamma)) + e^{-2\beta}(F^0_1(\gamma) - F^i_1(\gamma)) \tag{2.35}$$

positive there exists a $\gamma' = \gamma'(\gamma)$ with $|\gamma'| = |\gamma|$,

$A(\gamma') > A(\gamma)$, and $U(\gamma') = -U(\gamma)$. Then the operator $\tilde{G}$ defined by

$$\begin{aligned}
\tilde{G}f(\gamma) &= Gf(\gamma) - (A(\gamma'(\gamma) - A(\gamma))^{-1} \\
&\quad \times U(\gamma)(f(\gamma'(\gamma)) - f(\gamma)), \\
&\quad \text{if } U(\gamma) \geqslant 0, \\
\tilde{G}f(\gamma') &= Gf(\gamma') - (A(\gamma'(\gamma) - A(\gamma))^{-1} \\
&\quad \times U(\gamma)(f(\gamma) - f(\gamma')) \\
&\quad \text{if } U(\gamma') < 0 \text{ and } \gamma' = \gamma'(\gamma) \\
&\quad \text{for some } \gamma, \\
\tilde{G}f(\gamma) &= Gf(\gamma), \quad \text{otherwise}
\end{aligned} \tag{2.36}$$

has a quadratic form majorizing that of G, so that its mass gap exceeds that of G. However, Lemmas 2.1 and 2.3 imply an area law for the expected time to exit from D_n for $\tilde{G}$; hence, $\tilde{G}$ and therefore, G has zero mass gap. The reader might imagine other schemes where γ is mapped to a *family* of contours. Unfortunately, we cannot carry out such a program, although in a certain sense explained below, we do carry it out to lowest order in powers of $e^{-2\beta}$.

We begin with the following definition. Let γ be a contour. Then a connected segment g of γ is called an *intrusion* of γ if (i) there exists a unit length lattice bond $b(g)$, not in γ and which we refer to as a neck, with endpoints the same as those of g such that $\gamma' = b(g) \cup (\gamma - g) \in X$ and $A(\gamma') > A(\gamma)$. (ii) $b(g)$ is one side of a unit lattice square q outside γ such that $\gamma'' = \gamma \Delta q \notin X$. (iii) g is maximal in the sense that g is not contained properly in another segment satisfying properties (i) and (ii). If g is of length $2r+1$ we shall refer to g as an r-intrusion: See Fig. 1, which illustrates a contour γ with four intrusions. Note that associated with each intrusion is a forbidden transition at or near the neck of the intrusion.

The next step is to define a family of transformations among the contours $\mathscr{I}_p^n(\cdot)_n$, $n,r = 1,2,\ldots$. Set

$$k_p^n \equiv \sup(1, 15(4e^{-\beta})^{2r} n). \tag{2.37}$$

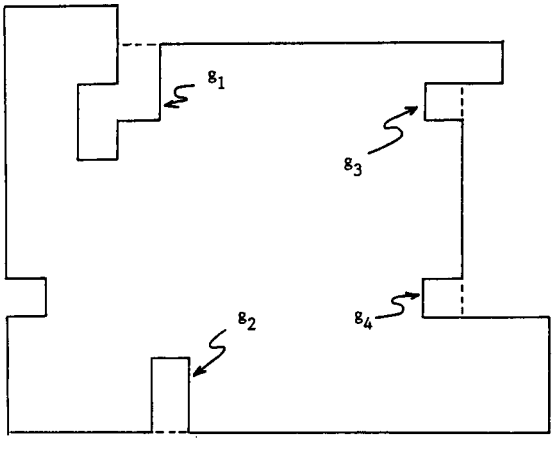


FIG. 1. A contour with four intrusions; g_1 is a four-intrusion, g_2 is a two-intrusion, and g_3 and g_4 are one-intrusions. The dashed lines are the unit length necks.

Given a contour γ, let $g_1,\ldots,g_m$ be the set of all of its r-intrusions (the set could be empty). Then $\mathscr{I}_p^n(\cdot)$ is defined by

$$\begin{aligned}
\mathscr{I}_r^n(\gamma) &= \gamma, \quad \text{if } m < k_r^n, \\
\mathscr{I}_r^n(\gamma) &= \Big(\gamma - \overset{m}{\cup} g_i\Big) \cup \overset{m}{\cup} g(g_i) \\
&\quad \text{if } m \geqslant k_r^n,
\end{aligned} \tag{2.38}$$

where $b(g_i)$ is the neck across g. Thus provided that m is sufficiently large, $\mathscr{I}_r^n(\gamma)$ is the new contour obtained from γ by removing its r-intrusions and then filling the gaps with the unit length necks. Note that if $A(\mathscr{I}_r^n(\gamma)) \geqslant A(\gamma')$ and the inequality is strict if the number of r-intrusions is sufficiently large; moreover, if $\mathscr{I}_r^n(\gamma) \neq \gamma$, then $|\mathscr{I}_r^n(\gamma)| = |\gamma| - 2mr$. Define

$$\mathscr{I}^n(\gamma) = \mathscr{I}_1^n \circ \mathscr{I}_2^n \circ \cdots \circ \mathscr{I}_l^n(\gamma), \quad \text{for} \quad l = |\gamma|. \tag{2.39}$$

We next define the operator perturbations P_n acting on functions defined on D_n:

$$\begin{aligned}
P_n f(\gamma) &= -p_n(\gamma)(f(\mathscr{I}^n(\gamma)) - f(\gamma)) \\
&\quad - \sum_{\gamma' \in D_n} q_n(\gamma,\gamma')(f(\gamma') - f(\gamma)),
\end{aligned} \tag{2.40}$$

where

$$p_n(\gamma) = (|\gamma| - \mathscr{I}^n(\gamma))/(A(\mathscr{I}^n(\gamma)) - A(\gamma)) \tag{2.41}$$

provided that $\mathscr{I}^n(\gamma) \neq \gamma$, zero otherwise, and q_n is obtained from p_n using the detailed balance

$$q_n(\gamma,\gamma') = p_n(\gamma') \exp -\beta(|\gamma| - |\gamma'|). \tag{2.42}$$

[Note that if $\gamma \in D_n$, $\mathscr{I}^n(\gamma) \in D_n$ also since $\mathscr{I}^n$ does not decrease the area of γ nor increase its length.]

We proceed to estimate the second term on the rhs of Eq. (2.39), applied to the area.

Lemma 2.6: For β sufficiently large, there is a $c(\beta)$ independent of n, γ such that for $\gamma \in D_n$,

$$-\sum_{\substack{\gamma' \in D_n \\ \mathscr{I}^n(\gamma') = \gamma}} q_n(\gamma,\gamma')(A(\gamma') - A(\gamma)) < c(\beta). \tag{2.43}$$

Proof: Clearly the lhs of inequality (2.43) is equal to

$$-\sum_{\substack{\gamma' \in D_n \\ \mathscr{I}^n(\gamma') = \gamma}} (|\gamma'| - |\gamma|) e^{-\beta(|\gamma'| - |\gamma|)}. \tag{2.44}$$

If γ' contributes a term to the sum (2.44), then γ' can be obtained from γ by inserting, say, k_1 one-intrusions, k_2 two-intrusions, ..., k_m m-intrusions, where for each i, either $k_i = 0$ or $k_i \geqslant k_i^n$. Such a term contributes

$$\begin{aligned}
&2(k_1 + 2k_2 + \cdots + mk_m) \\
&\quad \times \exp[-2\beta(k_1 + 2k_2 + \cdots + mk_m)]
\end{aligned} \tag{2.45}$$

to the sum. An obvious bound on the number of ways k_j j-intrusions may be placed along γ is $\binom{|\gamma|}{k_j}$. There are fewer than 3^{2j} j-intrusions possible emanating from a given bond of γ. Thus the sum is bounded by

$$2\sum_{\substack{k_j>k_j^n \\ \text{or } k_j=0}}\prod\binom{|\gamma|}{k_j}(k_1+2k_2+\cdots+mk_m)$$
$$\times(3e^{-\beta})^{2(k_1+2k_2+\cdots+mk_m)}$$
$$=z\frac{\partial}{\partial z}\prod_j\left(1+\sum_{k_j>k_j^n}\binom{|\gamma|}{k_j}z^{2jk_j}\right), \tag{2.46}$$

where $z=3e^{-\beta}$. Following a Chebyshev inequality strategy, we have that if r is defined by $(zr)^{2j}=(k_j^n/|\gamma|)(1-k_j^n/|\gamma|)^{-1}$, then

$$\sum_{k_j>k_j^n}\binom{|\gamma|}{k}z^{2jk}<r^{-2jk_j^n}\sum_{k>0}\binom{|\gamma|}{k}(zr)^{2jk}$$
$$=z^{2jk_j^n}(zr)^{-2jk_j^n}(1+(zr)^{2j})^{|\gamma|}$$
$$<\exp\left(k_j^n\ln\left(\frac{z^{2j}e|\gamma|}{k_j^n}\right)\right)$$
$$\leqslant\exp k_j^n\ln\left(\frac{(3e^{-\beta})^{2j}e5n}{15(4e^{-\beta})^{2j}n}\right)<\left(\frac{3}{4}\right)^{2j}; \tag{2.47}$$

the last steps follow from the fact that for $\gamma\in D_n$, $|\gamma|\leqslant 5n$ and the definition of k_j^n, Eq. (2.37).

Thus the product in expression (2.46) is bounded by $\Pi_j(1+(\frac{3}{4})^{2j})$, which evidently converges. In fact, this estimate would be uniform for z in a complex disk about the origin, so that by Cauchy's integral formula for the derivative, it is evident that the derivatives will be bounded in any smaller disk. This concludes the proof of Lemma 2.6. ■

The following lemma is a measure of how good $A(\gamma)$ is as a Lyapunov function for $G+P_n$.

Lemma 2.7: There exist non-negative functions $c_1(\beta)$, $c_2(\beta)$ which are independent of n and which are bounded $\beta\to\infty$ such that for each $\gamma\in D_n$,

$$(G+P_n)A(\gamma)\leqslant c_1(\beta)+c_2(\beta)e^{-4\beta}n. \tag{2.48}$$

Proof: Combining the results of Lemmas 2.6 and 2.7; equality (2.32) and inequality (2.43), and the definition of P_n, equality (2.40), we obtain that

$$(G+P_n)A(\gamma)<2+F_2^0(\gamma)+e^{-2\beta}F_1^0(\gamma)$$
$$+c(\beta)-(|\gamma|-\mathscr{J}^n(\gamma)), \tag{2.49}$$

where, again,

$$|\gamma|-\mathscr{J}^n(\gamma)=2\sum_j{}'jk_j(\gamma), \tag{2.50}$$

with k_j the number of j intrusions; the sum extends over j such that $k_j\geqslant k_j^n$.

Now every F_1^0 or F_2^0 unit lattice square is either tangent to an intrusion on tangent to the neck of an intrusion. Among the F_1^0 squares we distinguish two sets: (i) those tangent to the neck of a one-intrusion and (ii) those tangent to the neck of a j-intrusion or tangent to a j-intrusion $j\geqslant 2$. The cardinality of set (i) is clearly $k_1(\gamma)$, the number of one-intrusions; the cardinality of set (ii) is designated by $\hat{F}_1^0(\gamma)$ [so that $F_1^0(\gamma)=k_1(\gamma)+\hat{F}_1^0(\gamma)$]. Note that all F_2^0 squares are tangent to j-intrusions with $j\geqslant 2$. Every j-intrusion has at least two unit length segments *not* tangent to an F_1^0 or F_2^0 square, so that the number of unit length segments of a j-intrusion tanget to an F_1^0 or F_2^0 square is not longer than $2j-1$. It follows that

$$\hat{F}_1^0(\gamma)+F_3^0(\gamma)\leqslant 2\sum_{j\geqslant 2}jk_j(\gamma). \tag{2.51}$$

Inequalities (2.15) and (2.49) imply that

$$(G+P_n)A(\gamma)\leqslant 2+c(\beta)+e^{-2\beta}k_1(\gamma)\chi^n(\gamma)$$
$$+2\sum_{j\geqslant 2}jk_j(\gamma), \tag{2.52}$$

where $\chi^n(\gamma)=1$ if $k_1(\gamma)<k_1^n$ and zero otherwise, and the sum extends over j with $k_j(\gamma)<k_j^n$. However, the rhs of inequality (2.52) is bounded by

$$2+c(\beta)+\left(15e^{-2\beta}(4e^{-\beta})^2+30\sum_{j\geqslant 2}j(4e^{-\beta})^{2j}\right)n$$
$$\equiv c_1(\beta)+c_2(\beta)e^{-4\beta}n, \tag{2.53}$$

which concludes the proof of Lemma 2.7. ■

This is the key estimate. We employ this estimate along with Lemmas 2.1, 2.2, and 2.3 to bound the mass gap of $G(\beta)$.

Lemma 2.8: Suppose $m(\beta)=\inf\text{spec } G(\beta)|\mathbf{1}^{\perp}>0$. Then the expected time for the process associated with $G+P_n$ to escape from D_n satisfies, for β and n sufficiently large,

$$E_{\gamma_n}(\tau_{D_n})\geqslant c_3(\beta)ne^{4\beta}, \tag{2.54}$$

with $c_3(\beta)$ bounded away from zero, $\beta\to\infty$.

Proof: Since the mass gap of $G+P_n$ exceeds that of G and $\pi_\beta(D_n^c)\to 1$, $n\to\infty$, we have that $m_{D_n}(\beta)\equiv\inf\text{spec}(G+P_n)_{D_n}>m(\beta)(1-\epsilon)$ for any $\epsilon>0$ for n sufficiently large by Lemma 2.1, where $(G+P_n)_{D_n}$ is obtained from $G+P_n$ by imposing Dirichlet boundary conditions on D_n^c. Let

$$t_n(\gamma)=\sum_{\gamma'\in T_n}j(\gamma,\gamma')A(\gamma') \tag{2.55}$$

(noting that P_n contributes no terms to t_n). Then for n sufficiently large,

$$t_n(\gamma)\leqslant 2|\gamma|\sup_{j(\gamma,\gamma')\neq 0}A(\gamma')$$
$$\leqslant 2\cdot 5n\cdot\left(\frac{5}{4}n+1\right)^2<16n^3, \tag{2.56}$$

so that

$$\|t_n\|_{l^2(D_n)}\leqslant 16n^3(\pi\{\gamma|5n-2\leqslant|\gamma|\leqslant 5n\})^{1/2}$$
$$<cn^4\mathbb{Z}_{(\beta)}^{-1/2}(3e^{-\beta})^{(5n-2)/2} \tag{2.57}$$

for some constant independent of β by the usual Peierls' estimate on the number of contours of a given length about a given point. Thus by Lemmas 2.3, 2.7, and the fact that $\Sigma_{\gamma\in S_n}A(\gamma)\leqslant\frac{1}{4}n^2$,

$$E_{\gamma_n}(\tau_{D_n})\geqslant(c_1(\beta)+c_2(\beta)e^{-4\beta}n)^{-1}(A(\gamma_n)-\sup_{\gamma\in S_n}A(\gamma)$$
$$-m_{D_n}^{-1}cn^4(3e^{-\beta})^{(5n-2)/2}e^{2\beta n})>c_4(\beta)e^{4\beta}n \tag{2.58}$$

for some $c_4(\beta)$ bounded away from 0 for β and n sufficiently large. ■

Finally, this estimate implies a bound on the mass gap for $G + P_n$ and hence for G.

Lemma 2.9: The mass gap $m(\beta)$ for $G(\beta)$ satisfies the estimate

$$m(\beta) = \inf \operatorname{spec} G(\beta)|1^{\perp} < c_5(\beta)\beta e^{-4\beta} \tag{2.59}$$

for some $c_5(\beta)$ bounded for $\beta \to \infty$.

Proof: Here we use Lemmas 2.1, 2.2, and 2.8. Again, $m(\beta)$ is less than the mass gap of $G + P_n$, which is less than $m_{D_n}(\beta)(\pi_\beta^{-1}(D_n^c))^{-1} < m_{D_n}(\beta)(1+\epsilon)$ for any $\epsilon > 0$ for n sufficiently large. We will assume that $m_{D_n}(\beta)$ is uniformly bounded away from zero in $n \to \infty$ since otherwise there is nothing to prove. Let

$$g_n(\gamma) = \sum_{\gamma' \in D_n^c} j(\gamma,\gamma').$$

Then clearly $g_n(\gamma) \leqslant 2|\gamma| < 10n$ for $\gamma \in D_n$ and $\|g_n\|_{l^2(D_n)} < 10n(\pi_\beta(D_n))^{1/2} < 1$ for n sufficiently large. Then for $\theta < 1$, n sufficiently large, and the initial x of Lemma 2.2 identified with γ_n, we have that

$$\begin{aligned} m_{D_n}(\beta) &< (\theta c_3(\beta) n e^{4\beta})^{-1}(\ln\|g_n\|_{l^2(D_n)} \\ &\quad - \tfrac{1}{2}\ln \pi_\beta\{\gamma_n\} - \ln(1-\theta)m_{D_n}) \\ &< (\theta c_3(\beta) n e^{4\beta})^{-1}(2n\beta + \tfrac{1}{2}\ln \mathbf{Z}(\beta) \\ &\quad - \ln(1-\theta)m_{D_n}) < c_5(\beta)\beta e^{-4\beta} \end{aligned}$$

for some $c_5(\beta)$ bounded for $\beta \to \infty$. ■

Theorem 2.10: Let $G(\beta)$ be any Markov generator acting in $l^2(X,\pi_\beta)$, with the speed functions $j(\gamma,\gamma') = j(\gamma,\gamma',\beta)$ satisfying the local motion and detailed balance conditions (i) and (ii) of Sec. I and which, moreover, satisfy

$$j(\gamma,\gamma',\beta) \leqslant c_0$$

for some constant c_0 independent of γ, β. Then there exists a c depending only on c_0 such that the mass gap of $G(\beta)$ is less than $c\beta e^{-4\beta}$ for β sufficiently large.

Proof: The Dirichlet form for $G(\beta)$ is bounded above by a suitable constant times the Dirichlet form for the "standard" generator considered above, from which Theorem 2.10 follows.

Remark: The operator G_∞ having the speed functions defined by Eqs. (2.24) and (2.25) with $\beta = \infty$ generates a semigroup corresponding to a stochastic process $\gamma(t)$ in which transitions in which $\gamma(t)$ increases its length are forbidden. If the process starts in a configuration in which there are no intrusions, then it is not hard to see that $\gamma(t)$ can never have any intrusions (the process is certainly not irreducible) and consequently, $G_\infty A\{\gamma(t)\} \leqslant 2$ by inequality (2.32). Lemma 2.3 (with D_{II} empty) then implies that the expected time for $\gamma(t)$ to shrink to the zero contour satisfies $E_\gamma(\tau) \geqslant \frac{1}{2}A(\gamma)$, where γ is any contour without intrusions.

To conclude this section, we can ask whether the mass gap bound for $G(\beta)$ can be improved. We have seen that each contour γ has associated with it a certain set of intrusions. Contours with a large density of intrusions, e.g., $k_j(\gamma) \gtrsim e^{-2\beta_j}|\gamma|$ [and having only a low density of compensating extrusions which are defined analogously to intrusions, but with $A(\gamma') < A(\gamma)$ and the inside squares replacing the outside squares] have a large $GA(\gamma)$ value, leading to a poor lower bound on the expected escape time. Effectively, we dealt with contours of this sort by constructing non-negative perturbations P_n such that $(G + P_n)A(\gamma)$ is small for them; we did not attempt to compensate for contours having small intrusion density, in particular, where $k_j(\gamma) \lesssim e^{-2\beta_j}|\gamma|$.

Is there a way to compensate for contours having intrusions with these low densities? In fact, the author found a more elaborate, but *ad hoc* perturbation P_n' of G defined in terms of a mapping $\mathscr{F}_n'$ analogous to $\mathscr{F}_n$ of Sec. I, but which, moreover, maps contours with one- and two-intrusions to one- and two-extrusions. This perturbation scheme, which exploits the negativity of the $F^i(\gamma)$ terms of $GA(\gamma)$, Eq. (2.32), leads to an estimate $(G + P_n')A(\gamma) \leqslant e^{-6\beta_o}(|\gamma|)$ and then ultimately to the mass gap estimate $\beta e^{-6\beta}$ for G. Thus this scheme follows the program outlined in Eq. (2.36) to order $e^{-6\beta}|\gamma|$, whereas the basic scheme outlined in this section went to order $e^{-4\beta}|\gamma|$. The general problem of compensating for j-intrusions with $j \geqslant 3$ along these lines remain open.

ACKNOWLEDGMENT

I am indebted to J.-P. Eckmann and A. Malaspinas for providing a computer program for the models in Sec. I.

[1] D. Ruelle, *Statistical Mechanics, Rigorous Results* (Benjamin, New York, 1969).
[2] L. E. Thomas, to appear in Commun. Math. Phys.
[3] J.-P. Marchand and P. A. Martin, J. Stat. Phys. **44**, 491 (1986).
[4] J. B. Keller, J. Rubinstein, and P. Sternberg, Stanford University preprint, 1988.
[5] L. E. Thomas and Z. Yin, J. Math. Phys. **27**, 2475 (1986).
[6] H. Rost, Z. Wahrsch. Verw. Gebiete **58**, 41 (1981).
[7] T. M. Liggett, *Interacting Particle Systems* (Springer, New York, 1985), Chap. VIII, Sec. 5.
[8] A. D. Sokal and L. E. Thomas, J. Stat. Phys. **51**, 907 (1988).
[9] A. D. Sokal and L. E. Thomas, J. Stat. Phys. **54**, 797 (1989).
[10] A. D. Sokal, to appear in *Proceedings of the 8th International Congress of Mathematical Physics*, Marseille, France (1986).
[11] S. Caracciolo and A. D. Sokal, J. Phys. A **19**, 797 (1986).
[12] M. Karowski, H. J. Thun, W. Helfrich, and F. S. Rys, J. Phys. A **16**, 4073 (1983).
[13] M. Karowski and H. J. Thun, Phys. Rev. Lett. **54**, 2556 (1985).
[14] M. Karowski and F. S. Rys, J. Phys. A **19**, 2599 (1986).
[15] M. Karowski, J. Phys. A **19**, 3375 (1986).
[16] R. J. Glauber, J. Math. Phys. **4**, 294 (1963).
[17] D. A. Huse and D. S. Fisher, Phys. Rev. B **35**, 6841 (1987).
[18] M. Droz and J. D. Gunton, *Introduction to the Theory of Metastable and Unstable States, Lecture Notes in Physics*, Vol. 183 (Springer, New York, 1983).

Dynkin's method of computing the terms of the Baker–Campbell–Hausdorff series

Asok Bose
Département de Physique, Université de Montréal, C.P. 6128, Succ "A", Montréal, Québec H3C 3J7, Canada

(Received 8 November 1988; accepted for publication 26 April 1989)

The infinite series for $\log(\exp X \exp Y)$ for noncommuting X and Y is expressible in terms of iterated commutators of X and Y except for the linear term $X + Y$. Dynkin derived an explicit expression for the terms as a sum of iterated commutators over a certain set of sequences. This paper presents a practical algorithm for applying Dynkin's formula and gives several illustrative examples.

I. INTRODUCTION

Let X and Y be two noncommuting indeterminates. The Baker–Campbell–Hausdorff theorem[1–4] asserts that $\log(\exp X \exp Y)$ is expressible in terms of iterated commutators of X and Y except the linear term $X + Y$. This theorem is useful in the theory of linear differential equations,[5] group theory,[6] and physics.[7] Hausdorff[1] derived an iterative method of computing the terms of this series—usually called the BCH series—and obtained terms to fifth degree. However, the procedure based on the intricate Baker–Hausdorff differentiation[1] is not effective. Varadarajan derived a recursion formula for calculating the terms: "unfortunately, the calculations become complicated very rapidly." [3] In two papers written in 1947, Dynkin[8,9] radically simplified the problem. The author gave a new proof of the Baker–Campbell–Hausdorff theorem and derived an effective procedure for determining the terms of the BCH series.

Though complicated in appearance, Dynkin's formula[9] is quite straightforward. We sketch the author's derivation of this classical result. Using the formal power series which define the exponential function and $\log(1+z)$ we have

$$\log(\exp X \exp Y) = \log\left(1 + \sum_{\substack{n,m=0 \\ n+m>0}}^{\infty} \frac{X^n}{n!}\frac{Y^m}{m!}\right)$$
$$= \sum \frac{(-1)^{k-1}}{k} \frac{1}{p_1!q_1!\cdots p_k!q_k!} \times X^{p_1}Y^{q_1}\cdots X^{p_k}Y^{q_k}. \quad (1)$$

The summation goes over all possible systems of non-negative integers $(p_1,q_1,\ldots,p_k,q_k)$ satisfying $p_i + q_i > 0$ $(i = 1,2,\ldots,k)$. From (1) one can easily show that the linear term is $X + Y$ and the second degree term is $(XY - YX)/2$. However, (1) is not at all practical for obtaining higher degree terms. Collecting, in (1), all the terms for which $p_1 + p_2 + \cdots + p_k = p$, $q_1 + q_2 + \cdots + q_k = q$, we get $P_{p,q}(X,Y)$, the homogeneous component of degree p in X and degree q in Y. Dynkin[8,9] defined "a linear mapping ψ of the free associative algebra $\mathscr{R}$ into the free Lie algebra L by means of the formula

$$\psi(x_1x_2\cdots x_n) = (1/n)[x_1[x_2[\cdots[x_{n-1},x_n]\cdots]]]" \ [10]$$

and derived the decisive result[9]

$$P_{p,q}(X,Y) = \frac{1}{p+q}\sum \frac{(-1)^{k-1}}{k} \frac{1}{p_1!q_1!\cdots p_k!q_k!} \times [X^{p_1}Y^{q_1}\cdots X^{p_k}Y^{q_k}], \quad (2)$$

where

$$[X^{p_1}Y^{q_1}\cdots X^{p_k}Y^{q_k}] = [\overbrace{X\cdots[X}^{p_1}[\overbrace{Y\cdots[Y}^{q_1}[X\cdots[Y[\overbrace{X\cdots[X}^{p_k}[\overbrace{Y\cdots[Y,Y]}^{q_k}$$

and the summation goes over all possible systems of non-negative integers $(p_1,q_1,\ldots,p_k,q_k)$ satisfying

$$\sum_{i=1}^{k} p_i = p, \quad \sum_{i=1}^{k} q_i = q, \quad p_i + q_i > 0. \quad (3)$$

We call such a system of non-negative integers simply a sequence. (See also Bourbaki[2] and Varadarajan.[3])

Richtmyer and Greenspan[11] determined the terms of the BCH series by computer. These authors used a different computational method and gave no detail of their "fairly intricate bit of programming." Since their results are not linearly independent, it is not possible to compare them with other results.

Goldberg[12] expanded $\log(e^X e^Y)$ in terms of words in X and Y, and derived useful formulas for the expansion coefficients. The Goldberg coefficients and the coefficients derived from (2) are related. Recently, Newman and Thompson[13] implemented Goldberg's algorithm on a microcomputer and computed these coefficients of words of length up to 20.

This paper presents a simple and efficient algorithm for applying Dynkin's formula and gives illustrative examples.

II. ALGORITHM

Equation (2) shows that the iterated commutators are of two types: (a) The rightmost element is Y and its nearest neighbor is X^r, where $r = 1,2,\ldots$, i.e., $(p_1q_1) = (r1)$. (b) The rightmost element is X and its nearest neighbor is Y^r, i.e., $(q_2p_1q_1) = (r10)$.

Our tool for generating the sequences is, naturally, partitions of numbers. First, we determine the terminal sequences. We denote the distinct arrangements of $p-1$

 0022-2488/89/092035-03$02.50

(10)'s and $q-1$ (01)'s by $A(i)$, $i=1$ to N, where N is the binomial coefficient $\binom{p+q-2}{p-1}$. The terminal sequences $T(j)$ are of the form $A(i)11$ and $A(i)0110$. We reduce these sequences to their shortest possible lengths and set $T(2i-1)=A(i)11$ and $T(2i)=A(i)0110$. One notes that $T(2i-1)$ and $T(2i)$ represent the same commutator except for the sign. Two sequences are equivalent if one can be transformed into the other without 10 and 01—X and Y—crossing each other, i.e., if they are related to the same arrangement $A(i)$. It follows that each terminal sequence defines an equivalence class $C[i]$ and all its equivalent elements represent the same commutator. Thus the task of generating the sequences is divided into smaller and easier ones.

We now rewrite (2) as follows

$$(p+q)P_{p,q}(X,Y)=\sum_{C[i]} W(p_1,q_1,\ldots,p_k,q_k)\times[X^{p_k}Y^{q_k}\cdots X^{p_2}Y^{q_2}X^{p_1}X^{q_1}], \quad (4)$$

where

$$W(p_1,\ldots,q_k)=\frac{(-1)^{k-1}}{k}\left(\prod_{i=1}^{k} p_i!q_i!\right)^{-1}.$$

To calculate the coefficient $W(C[i])$ of a class $C[i]$, we do not generate all the elements of $C[i]$ and thereby make a substantial reduction in computation. We introduce the DZ transformation

$$p_iq_i \Rightarrow p_i00q_i \quad (5)$$

and observe that a DZ transformation acting on W changes only its first factor. Let σ be a NCZ (i.e., no contiguous zeros) sequence and $F(\sigma)$ represent the family σ, i.e., σ together with all its offspring generated by applying all possible DZ transformations to σ. It is not difficult to see that

$$W(F(\sigma))=FW(\sigma). \quad (6)$$

That is, the coefficient of $F(\sigma)$ can be obtained economically by applying all possible DZ transformations to $W(\sigma)$ without at all generating the offspring of σ.

This leads us to the following scheme for determining the coefficient of an equivalence class $C[i]$, i.e., a commutator.

Step 1: Start with the terminal sequence reduced to its shortest possible length.

Step 2: Use the seed to generate all the other NCZ elements of $C[i]$ by repeated partitions of p_i's and q_i's as follows:

$$\text{(i) } p_iq_i \Rightarrow 10p_i-1q_i \quad (p_i>1),$$
$$\text{(ii) } p_iq_i \Rightarrow p_iq_i-101 \quad (q_i>1) \quad (7)$$

(subject to $p_i+q_i>0$).

Step 3: Let σ be an NCZ sequence of $C[i]$. Compute $W(\sigma)$ and apply all possible DZ transformations to $W(\sigma)$. Repeat the operation for all the NCZ elements of $C[i]$ and add up all their contributions.

III. EXAMPLES

We now apply our results. The sequences are given in Table I.

Example 1: First, we compute $P_{2,2}(X,Y)$. The arrangements are 11 and 0110. Hence the terminal sequences are 1111, 1210, 0121, and 011110. The construction of the classes of sequences is immediate. To compute $FW(1111)$, we note that we can choose n 11's in $\binom{2}{n}$ ways. We have

$$FW(C[1])=-1/2+\binom{2}{1}/3-1/4=-1/12,$$
$$FW(C[2])=(-1/2+1/3)/2!+(1/3-1/4)=0,$$
$$FW(C[3])=FW(C[2]) \quad \text{and} \quad FW(C[4])=1/12.$$

Hence

$$4P_{2,2}(X,Y)=-1/12[XYXY]+0[XY^2X]+0[YX^2Y]+1/12[YXYX].$$

Using the identity

$$[YXXY]=[XYXY] \quad (8)$$

we get

$$P_{2,2}(X,Y)=-1/24[XYXY]=-1/24[X[Y[X,Y]]].$$

Example 2: We now compute $P_{2,3}(X,Y)$. The arrangements are 12, 0111, and 0210. Using (7) we construct the classes and see that $FW(C[6])=FW(C[4])$. We compute

$$FW(C[1])=(-1/2+\binom{2}{1}/3-1/4)/2!+(1/3-\binom{2}{1}/4+1/5)=-1/120.$$

Similarly, we calculate the coefficients of the other classes and obtain

TABLE I. Classes of NCZ sequences (i.e., no contiguous zeros).

$P_{2,2}$

$C[1]$	$C[2]$	$C[3]$	$C[4]$
1111	1210 110110	0121 011011	011110

$P_{2,3}$

$C[1]$	$C[2]$	$C[3]$	$C[4]$
1211 110111	1310 120110 110210 $11(01)^210$	011111	011210 01110110

$C[5]$	$C[6]$
0221 $(01)^221$ 021011 $(01)^21011$	021110 $(01)^21110$

$P_{4,1}$

$C[1]$	$C[2]$
41 1031 2021 $(10)^221$ 3011 102011 201011 $(10)^311$	3110 102110 201110 $(10)^21110$

TABLE II. Partitions of 7 and 8. (The entries in the parentheses give the numbers of permutations of the parts.)

$7(1), 61(2), 52(2), 43(2), 51^2(3),$ $421(6), 3^21(3), 32^2(3), 41^3(4),$ $321^2(12), 2^31(4), 31^4(5), 2^21^3(10),$ $21^5(6), 1^7(1)$
$8(1), 71(2), 62(2), 53(2), 4^2(1),$ $61^2(3), 521(6), 431(6), 42^2(3),$ $3^22(3), 51^3(4), 421^2(12), 3^21^2(6),$ $32^21(12), 2^4(1), 41^4(5), 321^3(20),$ $2^31^2(10), 31^5(6), 2^21^4(15), 21^6(7),$ $1^8(1)$

$$P_{2,3}(X,Y) = \{-1/120[XY^2XY] + 1/180[XY^3X] + 1/30[YXYXY] - 1/120[YXY^2X] - 1/120[Y^2X^2Y] - 1/120[Y^2XYX]\}/5$$
$$= -1/360[XY^2XY] + 1/120[YXYXY].$$

By (8), we also have

$$P_{2,3}(X,Y) = -1/360[XY^2XY] + 1/120[Y^2X^2Y].$$

Example 3: We compute $P_{4,1}(X,Y)$. Proceeding as before we get the results given in Table I. It is easy to see that the contributions of $C[2]$ and the last half of $C[1]$ cancel each other. Hence

$$P_{4,1}(X,Y) = \{FW(41 + 1031 + 2021 + (01)^221)\}/5$$
$$= -1/720[X[X[X[X,Y]]]].$$

Example 4: Here we compute $P_{1,8}(X,Y)$. The starting sequences are 0711 and 0810. In this case we use partitions of 7 and 8 to generate the sequences. Note that a partition with m parts yields a sequence with $k = m + 1$ and of weight equal to the number of distinct permutations of m parts. Table II lists the partitions of 7 and 8 together with the numbers of permutations of parts. Using this table, (4) and the SUM function of the computer algebra system MACSYMA to do rational arithmetic, we obtain

$$P_{1,8}(X,Y) = \{1/151200[Y^7XY] - 1/1209600[Y^8X]\}/9$$
$$= -1/1209600[Y^8X].$$

Example 5: We compute the coefficients of the following commutators:

(a) $[XYXYXYXYXYX]$,

(b) $[X^2YXYXYXY]$,

(c) $[X^2Y^2XYX]$.

(a) The sequence is 111111111110. The DZ transformation is $11 \Rightarrow 1001$. Now one can choose n 11's in $\binom{5}{n}$ ways. Hence

$$FW(\sigma) = -1/6 + \binom{5}{1}(1/7 - 1/10) - \binom{5}{2}(1/8 - 1/9) + 1/11$$
$$= -1/2772.$$

(b) Here the sequence is 21111111. Partition of 2 generates 1011111111. Hence

$$FW(\sigma) = \{-1/4 + \binom{4}{1}(1/5 + 1/7) - \binom{4}{2}/6 - 1/8\}/2! + \{1/5 - \binom{4}{1}(1/6 + 1/8) + \binom{4}{2}/7 + 1/9\}$$
$$= -1/(560) + 1/(630) = -1/7!.$$

(c) Here the sequence is 221110. Repeated partitions of 2 generate 10121110, 21011110, and 1011011110. Therefore

$$FW(\sigma) = (1/3 - 1/2 + 1/5)/4 + (-1/4 + 2/5 - 1/6) + (1/5 - 1/3 + 1/7)$$
$$= 1/840.$$

The numerical values of the corresponding Goldberg coefficients[13] are identical to these values of the coefficients of the commutators.

IV. CONCLUSION

The present algorithm is simple and rapid. The procedure is not recursive and computes directly the coefficient of a commutator. We are constructing software to implement the algorithm on a computer.

ACKNOWLEDGMENTS

I would like to thank Antoine Bose for stimulating discussions and the referee for bringing Refs. 11 and 13 to my attention.

[1] F. Hausdorff, Ber. Verh. Saechs. Akad. Wiss. Leipzig, Math.-Phys. Kl. **58**, 19 (1906).

[2] N. Bourbaki, *Groupes et Algèbres de Lie*, Chaps. 2 and 3 (Hermann, Paris, 1972), pp. 51–56.

[3] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representations* (Prentice–Hall, Englewood Cliffs, NJ, 1974), pp. 114–146.

[4] D. Z. Djokovic, Math. Z. **143**, 209 (1975).

[5] W. Magnus, Commun. Pure Appl. Math. **7**, 649 (1954).

[6] W. Magnus, A. Karrass, and D. Solitar, *Combinatorial Group Theory* (Interscience, New York, 1966), pp. 379–388.

[7] K. O. Friedrichs, Commun. Pure Appl. Math. **6**, 1 (1953); G. H. Weiss and A. A. Maradudin, J. Math. Phys. **3**, 771 (1962); R. M. Wilcox, *ibid.* **8**, 962 (1967); B. Milenik and J. Plebanski, Ann. Inst. H. Poincaré **12**, 215 (1970).

[8] E. B. Dynkin, Dokl. Akad. Nauk. SSSR **57**, 323 (1947); Math. Rev. **9**, 132 (1947).

[9] E. B. Dynkin, Mat. Sb. **25**, 155 (1949); Math. Rev. **11**, 80 (1949).

[10] E. B. Dynkin, Usp. Mat. Nauk **5**, 135 (1950); Am. Math. Soc. Transl. **9**, 470 (1950).

[11] R. D. Richtmyer and S. Greenspan, Commun. Pure Appl. Math. **18**, 107 (1965).

[12] K. Goldberg, Duke Math. J. **23**, 13 (1956).

[13] M. Newman and R. C. Thompson, Math. Comp. **48**, 265 (1987).

Characteristic functional structure of infinitesimal symmetry mappings of classical dynamical systems. IV. Classical (velocity-independent) mappings of second-order differential equations

Gerald H. Katzin
Department of Physics, North Carolina State University, Raleigh, North Carolina 27695-8202
Jack Levine
Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205

(Received 15 December 1988; accepted for publication 29 March 1989)

In a recent paper [J. Math. Phys. **26**, 3080 (1985)], the first of this series, it was shown that velocity-dependent symmetry mappings of second-order dynamical systems have a characteristic functional structure which is the same for all dynamical systems. The present paper continues the investigation of this characteristic functional structure. A shortened, improved, conceptionally simpler proof of the existence of the characteristic structure is given. It is shown how this characteristic functional structure may be formally incorporated into a new procedure for the determination of classical (velocity-independent) symmetries of dynamical systems. In addition to providing insight into symmetry analysis this new procedure is also a practical alternative to existing symmetry methods for those dynamical systems for which there exist sufficient knowledge of the form of the dynamical solution, so that it is feasible to determine constants of motion by inversion (or the converse). Application of the procedure to time-dependent linear systems readily gives the complete classical symmetry group for the one-dimensional case (with generators in a somewhat simpler form than those obtained by Leach, Research Report AM-79:05, La Trobe Univ., Bundoora, Australia, 1979) and the n-dimensional isotropic case [with generators the same as those obtained by Lopez, J. Math. Phys. **29**, 1097 (1988)]. By inspection the formalism shows all n-dimensional time-dependent linear systems admit at least a $2n$-parameter classical symmetry group and all decoupled n-dimensional time-dependent linear systems admit at least a $3n$-parameter classical symmetry group. Other applications to linear and nonlinear systems are given.

I. INTRODUCTION

In a recent paper Katzin and Levine[1] showed that all (Noether and non-Noether) infinitesimal velocity-dependent symmetry mappings $[\delta x^i = \xi^i(\dot{x},x,t)\delta a, \delta t = \xi^0(\dot{x},x,t)\delta a, i = 1,...,n]$ of second-order dynamical systems were expressible in a form with a *characteristic functional structure* which was the same for all dynamical systems and was manifestly dependent upon constants of motion of the system.[2,3] This characteristic structure was formulated by means of an auxiliary symmetry mapping function $Z^i(\dot{x},x,t)$ [introduced by the relation $\xi^i(\dot{x},x,t) = Z^i(\dot{x},x,t) + \dot{x}^i\xi^0(\dot{x},x,t)$]. The formalism developed in Ref. 1, which determined the characteristic functional structure of the function Z^i, was essentially independent of the velocity dependence or independence of the symmetry mappings δx^i or δt. We now continue our investigations of this characteristic functional structure by examining in detail how this fundamental property of symmetry mappings may be utilized in the determination of *classical* [velocity-independent, in that $\delta x^i = \xi^i(x,t)\delta a$, $\delta t = \xi^0(x,t)\delta a$] symmetry mappings.

As pointed out in Remark 2.1 of Ref. 1 (and earlier by Sarlet and Cantrijn[4] with regard to Noether symmetries), classical mappings are determined by auxiliary symmetry mapping functions $Z^i(\dot{x},x,t)$ which are linear in $\dot{x}^i$. Our method for incorporating the above-mentioned characteristic functional structure into a procedure which determines classical symmetries is based upon the requirement that the Z^i satisfy this linearity property in addition to being expressed in a form which exhibits the characteristic functional structure. The resulting novel procedure gives a new perspective to the process of formulating classical symmetry conditions in that it formally makes use of both the dynamical solution and the set of $2n$ functionally independent constants of motion obtained by inversion of the solution. The symmetry conditions obtained by this new procedure determine (otherwise arbitrary) constants of motion which occur as part of the characteristic functional structure of Z^i and thereby select from the totality of general (classical and nonclassical) symmetry mappings admitted by the system those that are classical. One advantage we have observed in this new formalism is that for certain problems the symmetry conditions may be satisfied by inspection (as will be shown), thus allowing immediate determination of symmetry mappings which would not otherwise be so readily apparent. Since considerable information pertaining to the dynamical system is actually incorporated into the procedure it is not surprising that such is the case.

Thus far, however, for practical purposes the applicability of the procedure appears to be limited in general to those systems for which there exist sufficient knowledge of the form of the dynamical solution, so that determination of the

 0022-2488/89/092038-16$02.50

constants of motion by inversion is feasible (or the converse). Such systems include the general class of time-dependent, n-dimensional, $n \geqslant 1$ linear systems which we shall treat in detail. In addition to showing the versatility of the procedure we obtain the classical symmetries of two additional dynamical systems described by nonlinear dynamical equations.

In Sec. II elements of symmetry theory pertaining to the characteristic functional structure of the auxiliary symmetry mapping function Z^i are summarized. In addition, a new proof of this functional structure of Z^i is given. This proof is shorter and conceptionally simpler than the original version contained in Ref. 1.

In Sec. III conditions for determining classical symmetry mappings of a general n-dimensional dynamical system are formulated in a manner that makes use of the characteristic functional structure of the auxiliary symmetry mapping function Z^i. These conditions are divided into two cases—those for one-dimensional systems and those for systems of dimension $n > 1$.

Section IV begins an analysis of linear systems. The dynamical equations and constants of motion for a general time-dependent, n-dimensional, $n \geqslant 1$ linear system are synthesized in terms of the dynamical solution functions. The symmetry conditions developed in Secs. II and III are specialized to treat such linear systems. The synthesized system, its concomitant constants of motion, and its associated symmetry conditions are further specialized for the analysis of decoupled systems. Inspection of the symmetry conditions for general linear systems shows that all n-dimensional linear systems admit at least a $2n$-parameter group of classical symmetry mappings; moreover, it is found by inspection that all decoupled n-dimensional linear systems admit at least a $3n$-parameter group of classical symmetries. In both cases the symmetry mappings are given.

In Sec. V the symmetry conditions derived in Sec. IV are solved to obtain the complete eight-parameter group of classical symmetries for a general time-dependent one-dimensional linear system. The symmetry mappings are expressed as functions of the solution functions of the dynamical equations. The five-parameter subgroup of Noether symmetries are obtained. (The procedure for extracting th Noether symmetries is indicated in an associated Appendix). The results obtained are consistent with the work of Lutzky,[5] Leach,[6,7] and Lopez[8] (based upon other methods).

In Sec. VI the conditions obtained in Sec. IV for determining the classical symmetries of n-dimensional, $n > 1$, decoupled time-dependent linear systems are applied to obtain the complete group of symmetries of a specific two-dimensional time-dependent system.

In Sec. VII the symmetry conditions (obtained in Sec. IV) applicable to n-dimensional ($n > 1$) decoupled linear systems are further specialized to treat the case in which the system is isotropic. The complete $(n^2 + 4n + 3)$-parameter (maximal) group of symmetries is obtained and without basis change is found to be identical to that obtained by another method by Lopez.[8]

In Sec. VIII the complete group of classical symmetries is obtained for a one-dimensional nonlinear equation (known to be transformable into linear form) to illustrate the general $n = 1$ theory given in Sec. III.

In Sec. IX, as a second illustration of the general $n = 1$ theory of Sec. III, the classical symmetries of a nonlinear dynamical system are obtained.

II. ELEMENTS OF SYMMETRY THEORY

Consider a dynamical system described by the second-order differential equation

$$E^i(\ddot{x},\dot{x},x,t) \equiv \ddot{x}^i - F^i(\dot{x},x,t) = 0, \quad i = 1,...,n. \tag{2.1}$$

Assume (2.1) has as its complete solution

$$x^i = \phi^i(c_1,...,c_{2n},t) \equiv \phi^i(c,t), c_A = \text{const}, \quad A = 1,...,2n. \tag{2.2}$$

Equations (2.2) and those obtained by differentiation of (2.2) with respect to t may in principle be solved for the c_A to obtain $2n$ functionally independent constants of motion:

$$C_A(\dot{x},x,t) \overset{t}{=} c_A. \tag{2.3}$$

Remark 2.1: The notation $\overset{t}{=}$ denotes equality on the dynamical paths (2.2). □

A dynamical system (2.1) is said to admit an infinitesimal symmetry mapping defined by

$$\bar{x}^i = x^i + \delta x^i, \quad \delta x^i \equiv \xi^i(\dot{x},x,t)\delta a, \tag{2.4}$$

$$\bar{t} = t + \delta t, \quad \delta t \equiv \xi^0(\dot{x},x,t)\delta a \tag{2.5}$$

iff[9]

$$\delta E^i \equiv \frac{\partial E^i}{\partial \ddot{x}^j}\delta\ddot{x}^j + \frac{\partial E^i}{\partial \dot{x}^j}\delta\dot{x}^j + \frac{\partial E^i}{\partial x^j}\delta x^j + \frac{\partial E^i}{\partial t}\delta t \overset{\circ}{=} 0, \tag{2.6}$$

where[1]

$$\delta\ddot{x}^I \equiv (\ddot{\xi}^i - 2\ddot{x}^i\dot{\xi}^0 - \dot{x}^i\ddot{\xi}^0)\delta a, \tag{2.7}$$

$$\delta\dot{x}^i \equiv (\dot{\xi}^i - \dot{x}^i\dot{\xi}^0)\delta a. \tag{2.8}$$

Remark 2.2: The notation $\overset{\circ}{=}$ denotes "whenever $E^i = 0$"; Eq. (2.1) is to be used to eliminate all time derivatives of the x^i which are higher than $\dot{x}^i$ in the functions $\phi(t,x,dx/dt,...,dx^{(\alpha)}/dt^{(\alpha)})$ and which may appear as expressions or equations subject to $\overset{\circ}{=}$. □

Remark 2.3: In the background material summarized in this section we shall assume general velocity-dependent mappings. Subsequent restrictions on the theory that are dictated by the requirement that the mappings be *classical*, that is, based upon velocity-independent mapping functions $\xi^i(x,t)$, $\xi^0(x,t)$, will be discussed in Sec. III. □

By introduction of an auxiliary mapping function $Z^i(\dot{x},x,t)$ by means of the relation

$$\xi^i(\dot{x},x,t) = Z^i(\dot{x},x,t) + \dot{x}^i\xi^0(\dot{x},x,t) \tag{2.9}$$

it follows from Eq. (2.34) of Ref. 1 that

$$\delta E^i \equiv \Delta E^i + \left(\frac{dE^i}{dt}\right)\delta t, \tag{2.10}$$

where

$$\Delta E^i = (\ddot{Z}^i + J^i_j\dot{Z}^j + K^i_jZ^j)\delta a, \tag{2.11}$$

$$J^i_j(\dot{x},x,t)\equiv\frac{\partial E^i}{\partial \dot{x}^j}, \tag{2.12}$$

$$K^i_j(\dot{x},x,t)\equiv\frac{\partial E^i}{\partial x^j}. \tag{2.13}$$

The identity (2.10) and the fact that $dE^i/dt \overset{\circ}{=} 0$ [which follows from (2.1) and Remark 2.2] show that

$$\delta E^i \overset{\circ}{=} \Delta E^i. \tag{2.14}$$

It follows from (2.6) and (2.14) that the functions ξ^i, ξ^0, which are solutions to the symmetry condition (2.6), are related by (2.9) to the functions Z^i, which are solutions of the partial differential equations obtained by the formal expansion of the auxiliary symmetry condition

$$\ddot{Z}^i + J^i_j\dot{Z}^j + K^i_j Z^i \overset{\circ}{=} 0. \tag{2.15}$$

In Ref. 1 it was proved that the solutions Z^i of (2.15) are expressible in a form with a characteristic functional structure which is the same for all dynamical systems. We next give an alternative proof of this property of Z^i solutions which is shorter and which we believe to be conceptionally simpler than that given in Ref. 1.

Alternative Proof: The $2n$ functionally independent constants of motion (2.3) are used to define the transformation

$$C_A = C_A(\dot{x},x,t) \tag{2.16}$$

between the $2n$ variables $\dot{x}^1,...,\dot{x}^n$, $x^1,...,x^n$ and the $2n$ variables $C_1,...,C_{2n}$, with t acting as a parameter. The inverse transformation that follows from (2.16) has the form

$$x^i = \phi^i(C,t), \tag{2.17}$$

$$\dot{x}^i = \psi^i(C,t), \quad \psi^i\equiv\frac{\partial\phi^i}{\partial t}. \tag{2.18}$$

It is to be noted that the functions $\phi^i(C,t)$ in (2.17) are of the same functional form as the respective functions $\phi^i(c,t)$ which appear in (2.2).

Equations (2.17) and (2.18) may be used to express any function $f(\dot{x},x,t)$ as a function $F(C,t)$ in that

$$f(\dot{x},x,t) = f[\psi(C,t),\phi(C,t),t] = F(C,t). \tag{2.19}$$

By this transformation procedure we find for the following functions, which appear in (2.15), that

$$Z^i(\dot{x},x,t) = Z^i[\psi(C,t),\phi(C,t),t]\equiv z^i(C,t), \tag{2.20}$$

$$J^i_j(\dot{x},x,t) = J^i_j[\psi(C,t),\phi(C,t),t]\equiv j^i_j(C,t), \tag{2.21}$$

$$K^i_j(\dot{x},x,t) = K^i_j[\psi(C,t),\phi(C,t),t]\equiv k^i_j(C,t). \tag{2.22}$$

From the transformation (2.16) and the fact that the functions $C_A(\dot{x},x,t)$ are constants of motion, it follows that the variables C_A satisfy $\dot{C}_A \overset{\circ}{=} 0$ (see Remark 2.2). Using this property of the C's we find from (2.20) that

$$\dot{Z}^i \overset{\circ}{=} \frac{\partial z^i(C,t)}{\partial t}, \tag{2.23}$$

$$\ddot{Z}^i \overset{\circ}{=} \frac{\partial^2 z^i(C,t)}{\partial t\,\partial t}. \tag{2.24}$$

Making use of (2.20)–(2.24) we find that the variable transformation (2.17) and (2.18) takes the auxiliary symmetry equation (2.15) into the form

$$\frac{\partial^2 z^i(C,t)}{\partial t\,\partial t} + j^i_j(C,t)\frac{\partial z^i(C,t)}{\partial t} + k^i_j(C,t)z^j(C,t) \overset{\circ}{=} 0. \tag{2.25}$$

Equation (2.25) is a system of linear differential equations in which t is the independent variable and the C's are parameters. The solution to (2.25) is therefore expressible in the form[9]

$$z^i(C,t) = B^A(C_1,...,C_{2n})g^i_A(C_1,...,C_\sigma,t), \quad 0\leqslant\sigma\leqslant 2n, \tag{2.26}$$

where the B^A are arbitrary functions of the C's.

To complete the proof we use (2.16) to express (2.26) in terms of the original variables $\dot{x}$, x, and t to obtain

$$Z^i(\dot{x},x,t) = B^A[C_1(\dot{x},x,t),...,C_{2n}(\dot{x},x,t)] \times g^i_A[C_1(\dot{x},x,t),...,C_\sigma(\dot{x},x,t),t], \quad 0\leqslant\sigma\leqslant 2n. \tag{2.27}$$

The functions Z^i in (2.27) exhibit the above-mentioned functional structure, which is characteristic of solutions of the auxiliary symmetry condition (2.15). □

Remark 2.4: If $Z^i(\dot{x},x,t)$ given by (2.27), wherein the functions (constants of motion) $B^A[C(\dot{x},x,t)]$ are arbitrary, are used in (2.9), then for each arbitrarily chosen function $\xi^0(\dot{x},x,t)$ the $\xi^i(\dot{x},x,t)$ so determined will in general define a *velocity-dependent* symmetry mapping (in that either ξ^i or ξ^0 or both are explicitly velocity-dependent). Such mappings were discussed in detail in Ref. 1. With appropriate choices for the functions $B^A(C)$ it may be possible to obtain Z^i in (2.27), so that both ξ^i and ξ^0 are each *velocity-independent* functions (in which case ξ^0 could not be arbitrarily chosen). An analysis of this situation will be developed in the sections to follow. □

III. CONDITIONS FOR CLASSICAL SYMMETRY MAPPINGS

The symmetry theory outlined in Sec. II includes both *velocity-independent* (classical) mappings and *velocity-dependent* mappings. We now assume the mapping functions ξ_i,ξ^0 that appear in (2.4) and (2.5) to be *velocity-independent* in that they are of the form $\xi^i(x,t),\xi^0(x,t)$. For such classical mappings we find, by use of (2.9) and (2.11)–(2.15), that the following theorem is readily proved.

Theorem 3.1: A necessary and sufficient condition that a dynamical system

$$E^i(\ddot{x},\dot{x},x,t) = \ddot{x}^i - F^i(\dot{x},x,t) = 0, \quad i = 1,...,n \tag{2.1′}$$

admits a classical (velocity-independent) symmetry mapping

$$\bar{x}^i = x^i + \delta x^i, \quad \delta x^i\equiv\xi^i(x,t)\delta a, \tag{3.1}$$

$$\bar{t} = t + \delta t, \quad \delta t\equiv\xi^0(x,t)\delta a, \tag{3.2}$$

in that

$$\delta E^i \overset{\circ}{=} 0, \tag{2.6′}$$

is that the auxiliary symmetry conditions

$$\ddot{Z}^i + J^i_j \dot{Z}^j + K^i_j Z^j \overset{\circ}{=} 0 \tag{2.15'}$$

be satisfied by the linear (in $\dot{x}^i$) auxiliary mapping functions

$$Z^i(\dot{x}^i x,t) = \xi^i(x,t) - \dot{x}^i \xi^0(x,t). \tag{3.3}$$

□

Remark 3.1: The notation $Z^i(\dot{x}^i,x,t)$ indicates that the only $\dot{x}$ variable present in the ith component Z^i is $\dot{x}^i$. (There is no similar restriction on the x variables.) □

It follows from (3.3) that we may therefore state the following corollary to Theorem 3.1.

Corollary 3.1.1: For a dynamical system (2.1') all admitted classical symmetry mappings (3.1) and (3.2) are defined by the symmetry mapping functions $\xi^i(x,t), \xi^0(x,t)$, which are expressible in the form

$$\xi^i(x,t) = Z^i(\dot{x}^i,x,t) - \frac{1}{n}\dot{x}^i \frac{\partial Z^j(\dot{x}^j,x,t)}{\partial \dot{x}^j}, \tag{3.4}$$

$$\xi^0(x,t) = -\frac{1}{n}\frac{\partial Z^j(\dot{x}^j,x,t)}{\partial \dot{x}^j}, \tag{3.5}$$

where $Z^i(\dot{x}^i,x,t)$ is described in Theorem 3.1. □

Remark 3.2: For Lagrangian systems formulations similar to those contained in Theorem 3.1 and Corollary 3.1.1 hold for those $Z^i(\dot{x}^i,x,t)$ that define *classical Noether* symmetries (Sarlet and Cantrijn[4]). □

It follows from Theorem 3.1 and Corollary 3.1.1 that the problem of obtaining classical symmetry mappings (if they exist) may be formulated as conditions for obtaining linear Z^i of the form (3.3). These linearity conditions may be expressed as conditions on the functions $B^A(C)$ which appear in (2.27). To derive such conditions we first obtain alternative necessary and sufficient conditions for the functions Z^i $(n>1)$ or $Z^1 \equiv Z$ $(n=1)$ to satisfy the linearity requirement (3.3).

From (3.3) for $n>1$ it is seen that

$$\frac{\partial Z^i}{\partial \dot{x}^j} = 0, \quad i \neq j, \quad n>1, \tag{3.6}$$

$$\frac{\partial Z^i}{\partial \dot{x}^i} - \frac{\partial Z^j}{\partial \dot{x}^j} = 0, \quad i,j \text{ not summed}, \quad n>1 \tag{3.7}$$

are necessary conditions on the functions $Z^i(\dot{x},x,t)$ in order for (3.3) to hold. To show that conditions (3.6) and (3.7) are also sufficient note that (3.6) implies that the only $\dot{x}$ variable present in Z^i is $\dot{x}^i$ and hence we may write $Z^i = Z^i(\dot{x}^i,x,t)$. For such Z^i it follows from (3.7) that $\partial Z^i/\partial \dot{x}^i = \partial Z^j/\partial \dot{x}^j = \mu(x,t)$, i,j not summed; hence by integration we obtain $Z^i = \mu(x,t)\dot{x}^i + \lambda^i(x,t)$, which is of the form (3.3).

When $n=1$ conditions (3.6) and (3.7) do not apply. For this case it follows from (3.3) that $(x^1 \equiv x)$.

$$\frac{\partial^2 Z}{\partial \dot{x}\,\partial \dot{x}} = 0, \quad n=1, \tag{3.8}$$

which is clearly necessary and sufficient for $Z(\dot{x},x,t)$ to be linear in $\dot{x}$. [Note that for $n>1$ Eqs. (3.6) and (3.7) imply $\partial^2 Z^i/\partial \dot{x}^i\,\partial \dot{x}^i = 0$, i not summed.]

For the case $n>1$ by use of (2.27) we may express (3.6) and (3.7) in terms of the functions $B^A(C)$, $g^i_A(C,t)$, and their derivatives to obtain, respectively,

$$\frac{\partial Z^i}{\partial \dot{x}^j} = \left(\frac{\partial B^A}{\partial C_K} g^i_A + B^A \frac{\partial g^i_A}{\partial C_K}\right)\frac{\partial C_k}{\partial \dot{x}^j} = 0, \quad i \neq j,\ n>1, \tag{3.9}$$

$$\frac{\partial Z^i}{\partial \dot{x}^i} - \frac{\partial Z^j}{\partial \dot{x}^j} = \left(\frac{\partial B^A}{\partial C_K} g^i_A + B^A \frac{\partial g^i_A}{\partial C_K}\right)\frac{\partial C_K}{\partial \dot{x}^i} - \left(\frac{\partial B^A}{\partial C_K} g^j_A + B^A \frac{\partial g^j_A}{\partial C_K}\right)\frac{\partial C_K}{\partial \dot{x}^j} = 0,$$
$$i,j \text{ not summed}, \quad n>1. \tag{3.10}$$

Similarly, for the case $n=1$ by use of (2.27) we express (3.8) in the form $(g^1_A \equiv g_A)$

$$\frac{\partial^2 Z}{\partial \dot{x}\,\partial \dot{x}} = \left(\frac{\partial^2 B^A}{\partial C_J\,\partial C_K} g_A + 2\frac{\partial B^A}{\partial C_J}\frac{\partial g_A}{\partial C_K} + B^A \frac{\partial^2 g_A}{\partial C_J\,\partial C_K}\right)\frac{\partial C_J}{\partial \dot{x}}\frac{\partial C_K}{\partial \dot{x}} + \left(\frac{\partial B^A}{\partial C_K} g_A + B^A \frac{\partial g_A}{\partial C_K}\right)\frac{\partial^2 C_K}{\partial \dot{x}\,\partial \dot{x}} = 0, \quad n=1. \tag{3.11}$$

By transformations of the form (2.17) and (2.18) the functions $\partial C_I(\dot{x},x,t)/\partial \dot{x}^j$ that appear in (3.9) and (3.10) and functions $\partial C_I(\dot{x},x,t)/\partial \dot{x}$ and $\partial^2 C_I(\dot{x},x,t)/\partial \dot{x}\,\partial \dot{x}$ that appear in (3.11) are expressed as functions of the C_I and t. The resulting equations [(3.16)–(3.18)] are given in the theorem that follows.

Theorem 3.2: Necessary and sufficient conditions for the auxiliary symmetry equation

$$\ddot{Z}^i + J^i_j \dot{Z}^j + K^i_j Z^j \overset{\circ}{=} 0 \tag{2.15'}$$

to admit solutions Z^i (linear in $\dot{x}^i$) of the form

$$Z^i(\dot{x},x,t) = \xi^i(x,t) - \dot{x}^i \xi^0(x,t), \tag{3.3'}$$

are that Z^i satisfy the conditions

$$n=1: \quad \frac{\partial^2 Z}{\partial \dot{x}\,\partial \dot{x}} = 0 \quad (Z^1 \equiv Z,\ x^1 \equiv x), \tag{3.8'}$$

$$n>1: \begin{cases} \dfrac{\partial Z^i}{\partial \dot{x}^j} = 0, \quad i \neq j, & (3.6') \\ \dfrac{\partial Z^i}{\partial \dot{x}^i} - \dfrac{\partial Z^j}{\partial \dot{x}^j} = 0, \ i,j \text{ not summed}. & (3.7') \end{cases}$$

All solutions to conditions (3.8'), for $n=1$, or (3.6') and (3.7'), for $n>1$, are expressible in the form

$$(\text{all } n) \quad Z^i(\dot{x},x,t) = B^A(C) g^i_A(C,t). \tag{2.27'}$$

In (2.27') the constants of motion $C_A(\dot{x},x,t)$ and the functions $g^i_A(C,t)$ are defined in Sec. II and the constants of motion $B_A(C)$ satisfy the conditions

$$n=1: \quad \eta_{JKA}(C,t)\frac{\partial^2 B^A(C)}{\partial C_J\,\partial C_K} + \eta_{JA}(C,t)\frac{\partial^2 B^A(C)}{\partial C_J} + \eta_A(C,t)B^A(C) = 0, \quad A,J,K = 1,2, \tag{3.12}$$

$$n>1:\begin{cases}\rho^i_{jKA}(C,t)\dfrac{\partial B^A(C)}{\partial C_K}+\rho^i_{jA}(C,t)B^A(C)=0,\quad i\neq j,\quad i,j=1,\ldots,n,\quad A,K=1,\ldots,2n, & (3.13)\\ [\rho^i_{iKA}(C,t)-\rho^j_{jKA}(C,t)]\dfrac{\partial B^A(C)}{\partial C_K}+[\rho^i_{iA}(C,t)-\rho^j_{jA}(C,t)]B^A(C)=0, & \\ \quad i,j=1,\ldots,n,\ i,j \text{ not summed},\quad A,K=1,\ldots,2n, & (3.14)\end{cases}$$

where $[g^1_A(C,t)\equiv g_A(C,t), n=1]$

$$n=1:\begin{cases}\eta_{JKA}(C,t)\equiv P_J(C,t)P_K(C,t)g_A(C,t), & (3.15)\\ \eta_{JA}(C,t)\equiv 2P_J(C,t)P_K(C,t)\dfrac{\partial g_A(C,t)}{\partial C_k}+Q_J(C,t)g_A(C,t), & (3.16)\\ \eta_A(C,t)\equiv P_J(C,t)P_K(C,t)\dfrac{\partial^2 g_A(C,t)}{\partial C_J\,\partial C_K}+Q_J(C,t)\dfrac{\partial g_A(C,t)}{\partial C_J}, & (3.17)\\ P_K(C,t)\equiv\dfrac{\partial C_K(\dot{x},x,t)}{\partial \dot{x}}, & (3.18)\\ Q_K(C,t)\equiv\dfrac{\partial^2 C_K(\dot{x},x,t)}{\partial \dot{x}\,\partial \dot{x}},\quad A,J,K=1,2 \text{ in (3.15)–(3.19)}, & (3.19)\end{cases}$$

$$n>1:\begin{cases}\rho^i_{jKA}(C,t)\equiv P_{Kj}(C,t)g^i_A(C,t), & (3.20)\\ \rho^i_{jA}(C,t)\equiv P_{Kj}(C,t)\dfrac{\partial g^i_A(C,t)}{\partial C_K}, & (3.21)\\ P_{Kj}(C,t)\equiv\dfrac{\partial C_K(\dot{x},x,t)}{\partial \dot{x}^j},\quad i,j=1,\ldots,n,\ A,K=1,\ldots,2n \text{ in (3.20)–(3.22)}. & (3.22)\end{cases}$$

□

Corollary 3.2.1: The functions

$$B^A=k^A,\quad k^A\equiv\text{const},\ A=1,\ldots,2n \tag{3.23}$$

will satisfy (i) Eq. (3.12) for the case $n=1$ if [refer to (3.17)]

$$\eta_A(C,t)=0, \tag{3.24}$$

or (ii) Eqs. (3.13) and (3.14) for the case $n>1$ if [refer to (3.21)]

$$\rho^i_{jA}(C,t)=0. \tag{3.25}$$

Conditions (3.24) for the case $n=1$ or (3.25) for the case $n>1$ will be satisfied if

$$\frac{\partial g^i_A(C,t)}{\partial C_K}=0,\quad A,K=1,\ldots,2n. \tag{3.26}$$

□

Remark 3.3: Equation (3.26) is satisfied for all linear dynamical equations. □

Remark 3.4: With the exception of the $B^A(C)$ and their derivatives all functions appearing in (3.12) for the case $n=1$ and in (3.13) and (3.14) for the case $n>1$ are assumed to be known functions of the C's and t. With the C's regarded as the independent variables and t as a parameter, Eq. (3.12) for $n=1$ or Eqs. (3.13) and (3.14) for $n>1$ must hold for all applicable values of t. In principle this leads to conditions which in general will be partial differential equations in $B^A(C)$.

Alternatively, by means of (2.3) Eq. (3.12) for $n=1$ or Eqs. (3.13) and (3.14) for $n>1$ may be evaluated on dynamical paths. This path evaluation procedure in effect replaces each constant of motion C_A which appears in (3.12)–(3.14) by its constant path value c_A. Therefore, each equation obtained by this procedure will be identical in form to its precursor. Clearly, on each path these path evaluated equations for $B^A(c)$ must hold for all applicable values of t and one again is led to conditions which in principle determine the functional form of B^A. □

When used in (2.27′) of Theorem 3.2 the $B^A(C)$ obtained as solutions (not all zero) to the conditions referred to in Remark 3.4 will determine the most general functions Z^i having the linear form (3.3) of Theorem 3.1 and therefore, by (3.4) and (3.5) of Corollary 3.1.1, will determine the most general classical symmetry mapping functions $\xi^i(x,t)$ and $\xi^0(x,t)$ admitted by a dynamical system (2.1). If the only solutions $B^A(C)$ of the above-mentioned conditions are $B^A=0$, then the associated dynamical system does not admit classical symmetries.

Remark 3.5: For certain dynamical systems it may happen that (3.12) for the case $n=1$ or (3.13) and (3.14) for the case $n>1$ lead to equations expressible in the form

$$\sum_{\alpha=1}^{r} f_\alpha(B)\tau_\alpha(t)=0, \tag{3.27}$$

where $f_\alpha(B)$ denotes functions of the $B(C)$'s and their derivatives with respect to the C's. For dynamical systems that lead to equations of this type it appears this new method for determining classical symmetries could be of particular value. Such is the case for all linear dynamical systems, as will be shown in Sec. IV.

Equations (3.12) or (3.13) and (3.14) may still be tractable even though they do not lead to equations of the form

(3.27). We illustrate this situation by applying Theorem 3.1 to determine the classical symmetries of a nonlinear dynamical equation (see Sec. IX). □

Remark 3.6: For certain dynamical systems the functions $g_A^i(C,t)$ and the constants of motion $C_A(\dot{x},x,t)$ may be such that it is possible to determine *by inspection* the functions $B^A(C)$ such that (2.27′) of Theorem 3.2 gives $Z^i(\dot{x},x,t)$ of the desired linear form, (3.3′) of Theorem 3.2, and thereby obtain in a simple fashion classical mappings admitted by the system. Such is the case with the example given in Sec. VIII, as the reader may verify, and as illustrated in Remark 9.1 for the example given in Sec. IX. □

IV. SPECIALIZATION OF THEOREM 3.2 FOR THE CASE OF LINEAR DYNAMICAL SYSTEMS

The work of Lewis[10] in formulating a constant of motion (invariant) for the one-dimensional time-dependent oscillator appears to have generated the current interest in the analysis of time-dependent linear systems. Several different methods for determination of the symmetries and/or constants of motion for various classes of such linear systems have since appeared. Many authors have contributed to this research. We shall not attempt to give an exhaustive literature review but mention only those papers that are most pertinent to our work. The interested reader may consult the bibliographies of the references cited for additional literature on this subject.

Leach,[6,7] in related papers, has obtained by phase space transformations and other techniques the complete symmetry group for a one-dimensional time-dependent oscillator and a general one-dimensional time-dependent linear system. Also, for the general one-dimensional time-dependent linear system, but in Newtonian form, Aguire and Krause,[11] by a different transformation procedure, obtained the complete symmetry group in finite form. Lopez[8] has contributed to the analysis of an n-dimensional time-dpendent linear system by establishing conditions on the coefficients in the dynamical equations for the system to admit a symmery group of maximal dimension n^2+4n+3. Lopez obtained this maximal parameter symmetry group by transformation techniques (applied to the dynamical equations in Newtonian form) for the case of n-dimensional time-dependent, *isotropic*, linear systems. The maximal (n^2+4n+3)-parameter group of infinitesimal symmetry mappings was known to exist for the n-dimensional time-independent attractive and repulsive oscillators. (See, for example, Katzin *et al.*[12]) Various techniques that lead to constants of motion of time-dependent linear systems are available. See, for example, Katzin and Levine,[13] Eliezer and Gray,[14] Lewis and Leach,[15] Lutzky,[16] Prince and Eliezer,[17] and Colegrave and Mannan.[18]

As a prerequisite to specializing Theorem 3.2 to the class of *linear* dynamical systems we first summarize some properties of a general system of linear second-order differential equations:

$$E^i \equiv \ddot{x}^i - R_j^i(t)\dot{x}^j - S_j^i(t)x^j - T^i(t) = 0, \quad i,j = 1,...,n. \tag{4.1}$$

The solution to (4.1) is of the form

$$x^i = c_A \phi_A^i(t) + \phi_0^i(t), \quad c_A = \text{const}, \quad A = 1,...,2n, \tag{4.2}$$

where the functions $\phi_A^i(t)$ and $\phi_0^i(t)$ are twice differentiable and where

$$W \equiv \left| \frac{\phi_A^i(t)}{\psi_A^i(t)} \right| \neq 0, \tag{4.3}$$

with

$$\psi_A^i(t) \equiv \dot{\phi}_A^i(t). \tag{4.4}$$

In the partitioned $2n \times 2n$ determinant (4.3) the indices $i = 1,...,n$ denote rows and the indices $A = 1,...,2n$ denote columns (so, e.g., ψ_A^1 denotes row $n+1$, etc.).

It is a straightforward matter (as described below) to determine the coefficients $R_j^i(t)$, S_j^i, and $T^i(t)$ of (4.1) and a set of $2n$ functionally independent constants of motion of this system from any given solution set $\phi_A^i(t)$ and $\phi_0^i(t)$.

As a preliminary step we obtain, from (4.2),

$$\dot{x}^i = c_A \psi_A^i(t) + \psi_0^i(t), \tag{4.5}$$

$$\ddot{x}^i = c_A \dot{\psi}_A^i(t) + \dot{\psi}_0^i(t), \tag{4.6}$$

where

$$\psi_0^i(t) \equiv \dot{\phi}_0^i(t). \tag{4.7}$$

With use of (4.3) the $2n$ equations (4.2) and (4.5) may be solved for the $2n$ constants c_A to obtain $2n$ functionally independent constants of motion [refer to (2.3)] $C_A(\dot{x},x,t)$, which are linear polynomials in x^i and $\dot{x}^i$ with coefficients determined by the solution functions $\phi_A^i(t)$, $\phi_0^i(t)$ of (4.2) and their derivatives. These linear constants of motion are given in detail in Theorem 4.1.

By eliminating the constants c_A from (4.6) by means of (2.3) and the above-described linear constants of motion (4.1) and then comparing the resulting equation with (4.1) we may readily express the coefficients $R_j^i(t)$, $S_j^i(t)$, and $T^i(t)$ which appear in (4.1) in terms of the solution functions $\phi_A^i(t)$, $\phi_0^i(t)$ of (4.2) and their derivatives. These coefficients are also given in detail in Theorem 4.1.

Theorem 4.1: A linear dynamical system

$$E^i \equiv \ddot{x}^i - R_j^i(t)\dot{x}^j - S_j^i(t)x^j - T^i(t) = 0, \quad i,j = 1,...,n \tag{4.1′}$$

has a solution of the form

$$x^i = c_A \phi_A^i(t) + \phi_0^i(t), \quad c_A = \text{const.} \quad A = 1,...,2n. \tag{4.2′}$$

The solution functions $\phi_A^i(t)$ and $\phi_0^i(t)$ are twice differentiable and satisfy the requirement (i denotes rows, A denotes columns)

$$W = \left| \frac{\phi_A^i(t)}{\psi_A^i(t)} \right| \neq 0, \tag{4.3′}$$

where

$$\psi_A^i(t) \equiv \dot{\phi}_A^i(t). \tag{4.4′}$$

The coefficients in (4.1′) are expressible in the form

$$R_j^i(t) = W^{-1}\dot{\psi}_A^i \Psi_A^j, \tag{4.8}$$

$$S_j^i(t) = W^{-1}\dot{\psi}_A^i \Phi_A^j, \tag{4.9}$$

$$T^i(t) = \dot{\psi}_0^i - R_j^i \psi_0^i - S_j^i \phi_0^i, \tag{4.10}$$

where [with reference to (4.3′)]

$$\Phi_A^i(t) \equiv \text{cof}(\phi_A^i \text{ of } W), \quad \Psi_A^i(t) \equiv \text{cof}(\psi_A^i \text{ of } W) \tag{4.11}$$

and

$$\psi_0^i(t) \equiv \dot{\psi}_0^i(t). \tag{4.7′}$$

The dynamical system (4.1′) admits $2n$ functionally independent constants of motion [obtained by inverting (4.2) and its t derivative]

$$C_A(\dot{x},x,t) = \alpha_{Ai}(t)\dot{x}^i + \beta_{Ai}(t)x^i + \gamma_A(t), \tag{4.12}$$

where

$$\alpha_{Ai}(t) \equiv W^{-1}\Psi_A^i, \tag{4.13}$$

$$\beta_{Ai}(t) \equiv W^{-1}\Phi_A^i, \tag{4.14}$$

$$\gamma_A(t) \equiv -W^{-1}(\phi_0^i\Phi_A^i + \psi_0^i\Psi_A^i). \tag{4.15}$$

□

Remark 4.1: It is to be noted that any given twice-differentiable functions $\phi_A^i(t)$ and $\phi_0^i(t)$, $i = 1,...,n$, $A = 1,...,2n$ [which satisfy (4.3)] will determine by (4.8)–(4.10) the coefficients $R_j^i(t)$, $S_j^i(t)$, and $T^i(t)$ and hence a synthesized system of equations (4.1) which have as solution functions the given $\phi_A^i(t)$ and $\phi_0^i(t)$. It also follows that the given functions $\phi_A^i(t)$ and $\phi_0^i(t)$ will determine the constants of motion $C_A(\dot{x},x,t)$ in (4.12) of such a synthesized system (4.1). □

We now consider the specialization of the characteristic functional structure of the auxiliary symmetry mapping function Z^i in (2.27) for the case of the linear dynamical system (4.1). For the dynamical equation (4.1) the auxiliary symmetry condition (2.15) takes the form [refer to (2.12) and (2.13)]

$$\ddot{Z}^i - R_j^i(t)\dot{Z}^i - S_j^i(t)Z^i \overset{\circ}{=} 0. \tag{4.16}$$

Following the procedure outlined in Sec. II, the system of partial differential equations (4.16) has an associated system of ordinary differential equations [refer to (2.20)–(2.25)]

$$\ddot{z}^i - R_j^i(t)\dot{z}^j - S_j^i(t)z^j \overset{\circ}{=} 0. \tag{4.17}$$

Since (4.17) is of the same form as (4.1), it follows that its solution will be of the form (4.2), that is

$$z^i(C,t) = B^A(C)\phi_A^i(t). \tag{4.18}$$

Hence for the linear dynamical system (4.1) the functions g_A^i [appearing in (2.26)] are defined by

$$g_A^i = \phi_A^i(t). \tag{4.19}$$

Therefore [refer to (2.27) and the comments that follow], the *general* solution to (4.16) is

$$Z^i(\dot{x},x,t) = B^A[C(\dot{x},x,t)]\phi_A^i(t), \tag{4.20}$$

where the B^A appearing in (4.20) are arbitrary functions of the constants of motion $C_A(\dot{x},x,t)$ in (4.12).

Remark 4.2: The auxiliary mapping function Z^i in (4.20) exhibits the characteristic functional structure associated with all velocity-dependent symmetry mappings of the general time-dependent linear system (4.1); (refer to Remark 2.1). □

Conditions on the $B(C)$'s of (4.20) for $Z^i(\dot{x},x,t)$ to be linear in $\dot{x}^i$ (and thereby determine classical symmetry mappings) will now be obtained by specializing Theorem 3.2 to the case of the linear dynamical system (4.1). By use of (2.12), (2.13), (3.12)–(3.22), (4.19), and the formulas given in Theorem 4.1, we find for the linear system (4.1) that Theorem 3.2 may be restated as follows.

Theorem 4.2: For a linear dynamical system

$$E^i \equiv \ddot{x}^i - R_j^i(t)\dot{x}^j - S_j^i(t)x^j - T^i(t) = 0, \quad i,j = 1,...,n \tag{4.1′}$$

(described in Theorem 4.1) to admit solutions Z^i of the auxiliary symmetry condition

$$\ddot{Z}^i - R_j^i(t)\dot{Z}^j - S_j^i(t)Z^j \overset{\circ}{=} 0 \tag{4.16′}$$

which are of the linear form

$$Z^i(\dot{x},x,t) = \xi^i(x,t) - \dot{x}^i\xi^0(x,t), \tag{3.3′}$$

it is necessary and sufficient that Z^i be of the form

$$Z^i = B^A(C)\phi_A^i(t), \quad A = 1,...,2n, \tag{4.20′}$$

where the functions $B^A(C)$ satisfy the conditions ($\phi_A^1 \equiv \phi_A$ when $n = 1$)

$$n = 1: \quad (\phi_1)^3\frac{\partial^2 B^1}{\partial C_2\,\partial C_2} + (\phi_1)^2\phi_2\left(\frac{\partial^2 B^2}{\partial C_2\,\partial C_2} - 2\frac{\partial^2 B^1}{\partial C_1\,\partial C_2}\right) + \phi_1(\phi_2)^2\left(\frac{\partial^2 B^1}{\partial C_1\,\partial C_1} - 2\frac{\partial^2 B^2}{\partial C_2\,\partial C_2}\right) + (\phi_2)^3\frac{\partial^2 B^2}{\partial C_1\,\partial C_1} = 0, \tag{4.21}$$

$$n > 1 \begin{cases} \Psi_K^j(t)\phi_A^i(t)\dfrac{\partial B^A}{\partial C_K} = 0, \quad i \neq j, \quad i,j = 1,...,n, \quad A,K = 1,...,2n, & (4.22) \\ [\Psi_K^i(t)\phi_A^i(t) - \Psi_K^j(t)\phi_A^j(t)]\dfrac{\partial B^A}{\partial C_K} = 0, \quad i,j \text{ not summed.} & (4.23) \end{cases}$$

□

Remark 4.3: That all linear systems (4.1′) lead to conditions of the form (3.27) is evident by inspection of Eqs. (4.21)–(4.23). □

For the case $n = 1$ Eq. (4.21) of Theorem 4.2 will be solved (in Sec. V) and the results used to obtain the complete group of classical symmetries for the one-dimensional linear system [(4.1) with $n = 1$].

For the case $n > 1$ Eqs. (4.22) and (4.23) of Theorem

4.2 obviously are satisfied when $B^A = k^A$, $k^A = \text{const}$. This solution also follows from (4.19) and Corollary 3.2.1, as indicated by (3.23). From (4.20′) and (3.3′) of Theorem 4.2, Theorem 3.1, and Corollary 3.1.1 the above-mentioned B solutions allow us to state the following corollary to Theorem 4.2.

Corollary 4.2.1: Every n-dimensional, $n>1$, linear dynamical system (refer to Theorem 4.1)

$$\ddot{x}^i - R^i_j(t)\dot{x}^j - S^i_j(t)x^j - T^i(t) = 0, \quad i,j = 1,...,n \tag{4.1′}$$

admits at least a $2n$-parameter classical symmetry mapping determined by the mapping function

$$\xi^i(x,t) = k^A\phi^i_A(t), \quad \xi^0(x,t) = 0. \tag{4.24}$$

These mapping functions define a $2n$-parameter Abelian (sub)group of transformations with the generators

$$X_A = \phi^i_A(t)\partial_i, \quad A = 1,...,2n. \tag{4.25}$$

□

A specialized class of the linear system (4.1) is defined by n equations of the form

$$E^i \equiv \ddot{x}^i - R^i(t)\dot{x}^i - S^i(t)x^i - T^i(t) = 0. \tag{4.26}$$

It is to be noted that in (4.26) the only dependent variable in the ith equation is x^i. We shall therefore refer to (4.26) as a *decoupled* linear system. The (general) linear system (4.1) reduces to the decoupled form (4.26) iff

$$R^i_j(t) = \delta^i_j R^i(t), \quad S^i_j(t) = \delta^i_j S^i(t). \tag{4.27}$$

When $n=1$ every linear system (4.1) is obviously always of the decoupled form (4.26).

For a nondecoupled linear system (4.1) each solution $x^i(t)$, $i=1,...,n$, (4.2) will in general contain $2n$ distinct solution functions $\phi^i_A(t)$, $A=1,...,2n$ determined by the homogeneous portion of all n equations of the system. In contradistinction for each i, the solution $x^i(t)$ of a decoupled linear system (4.26) will contain only two solution functions which are determined by the homogeneous portion of only the ith equation of the system (4.26). Hence when the linear system (4.1) is of the decoupled form (4.26) we may simplify the solution (4.2) by taking

$$\phi^i_A(t) = 0, \quad A \neq 2i-1 \text{ or } 2i \tag{4.28}$$

to obtain

$$x^i = c_{2i-1}\phi^i_{2i-1}(t) + c_{2i}\phi^i_{2i}(t) + \phi^i_0(t), \tag{4.29}$$

where the solution functions ϕ^i_{2i-1} and ϕ^i_{2i} must satisfy [refer to (4.3)]

$$W_i \equiv \begin{vmatrix} \phi^i_{2i-1} & \phi^i_{2i} \\ \psi^i_{2i-1} & \psi^i_{2i} \end{vmatrix} \neq 0. \tag{4.30}$$

It is noted by use of (4.28), (4.3), and the repeated use of the Laplace expansion that for the decoupled system (4.26) the function W in (4.3) reduces to

$$W(n) = (-1)^{n(n+3)/2} W_1 W_2 \cdots W_n \neq 0. \tag{4.31}$$

From the form of $W(n)$ in (4.31) it can be shown for the decoupled system (4.26) that the cofactors (4.11) may be expressed in the form

$$\Phi^j_{2i-1} = \delta^j_i W(n) W_i^{-1} \psi^i_{2i}, \tag{4.32}$$

$$\Phi^j_{2i} = -\delta^j_i W(n) W_i^{-1} \psi^i_{2i-1}, \tag{4.33}$$

$$\Psi^j_{2i-1} = -\delta^j_i W(n) W_i^{-1} \phi^i_{2i}, \tag{4.34}$$

$$\Psi^j_{2i} = \delta^j_i W(n) W_i^{-1} \phi^i_{2i-1}. \tag{4.35}$$

If use is made of (4.12)–(4.15), (4.28), (4.32)–(4.35), and Theorem 4.1, we may state the following corollary to Theorem 4.2.

Corollary 4.2.2: If an n-dimensional, $n>1$, linear dynamical system (4.1) is decoupled in that it is of the form

$$E^i \equiv \ddot{x}^i - R^i(t)\dot{x}^i - S^i(t)x^i - T^i(t) = 0 \tag{4.26′}$$

and its solution is expressed in the form

$$x^i(t) = c_{2i-1}\phi^i_{2i-1}(t) + c_{2i}\phi^i_{2i}(t) + \phi^i_0(t), \tag{4.29′}$$

where ϕ^i_{2i-1} and ϕ^i_{2i} satisfy ($\psi^i_A \equiv \dot{\phi}^i_A$, $A = 2i, 2i-1$)

$$W_i \equiv \phi^i_{2i-1}\psi^i_{2i} - \phi^i_{2i}\psi^i_{2i-1} \neq 0, \tag{4.36}$$

then in Theorem 4.2 Eqs. (4.22) and (4.23) reduce, respectively, to

$$\phi^i_{2i-1}\phi^j_{2j}\frac{\partial B^{2i-1}}{\partial C_{2j-1}} + \phi^i_{2i}\phi^j_{2j}\frac{\partial B^{2i}}{\partial C_{2j-1}} - \phi^i_{2i-1}\phi^j_{2j-1} \times \frac{\partial B^{2i-1}}{\partial C_{2j}} - \phi^i_{2i}\phi^j_{2j-1}\frac{\partial B^{2i}}{\partial C_{2j}} = 0, \quad i \neq j, \quad i,j \text{ not summed} \tag{4.37}$$

and

$$W_i^{-1}\left[(\phi^i_{2i-1})^2\frac{\partial B^{2i-1}}{\partial C_{2i}} + \phi^i_{2i-1}\phi^i_{2i}\left(\frac{\partial B^{2i}}{\partial C_{2i}} - \frac{\partial B^{2i-1}}{\partial C_{2i-1}}\right) - (\phi^i_{2i})^2\frac{\partial B^{2i}}{\partial C_{2i-1}}\right] = W_j^{-1}\left[(\phi^j_{2j-1})^2\frac{\partial B^{2j-1}}{\partial C_{2j}} + \phi^j_{2j-1}\phi^j_{2j}\left(\frac{\partial B^{2j}}{\partial C_{2j}} - \frac{\partial B^{2j-1}}{\partial C_{2j-1}}\right) - (\phi^j_{2j})^2\frac{\partial B^{2j}}{\partial C_{2j-1}}\right], \quad i,j \text{ not summed.} \tag{4.38}$$

The constants of motion appearing in (4.37) and (4.38) have the form

$$C_{2i-1} = W_i^{-1}(-\phi^i_{2i}\dot{x}^i + \psi^i_{2i}x^i + \psi^i_0\phi^i_{2i} - \phi^i_0\psi^i_{2i}), \tag{4.39}$$

$$C_{2i} = W_i^{-1}(\phi^i_{2i-1}\dot{x}^i - \psi^i_{2i-1}x^i - \psi^i_0\phi^i_{2i-1} + \phi^i_0\psi^i_{2i-1}). \tag{4.40}$$

□

For *any* system of n, $n>1$, decoupled dynamical equations (4.26) it is easily shown that Eqs. (4.37) and (4.38) of Corollary 4.2.2 are satisfied by the functions

$$B^{2i-1} = \mu_i C_{2i-1} + \gamma_i, \quad B^{2i} = \mu_i C_{2i} + \nu_i, \quad \mu_i,\gamma_i,\nu_i \equiv \text{const.} \tag{4.41}$$

The B's in (4.41) may be expressed as functions of $\dot{x}$,x, and t by use of the constant of motion formulas (4.39) and (4.40). The resulting B functions, along with (4.28), determine by (4.20′) of Theorem 4.2 linear Z^i of the form (3.3′); such Z^i by Theorem 3.1 and Corollary 3.1.1 allow us to state the following result.

Corollary 4.2.3: Every n-dimensional, $n>1$, decoupled linear dynamical system (refer to Corollary 4.2.2)

$$\ddot{x}^i - R^i(t)\dot{x}^i - S^i(t)x^i - T^i(t) = 0 \tag{4.42}$$

admits at least a $3n$-parameter classical symmetry mapping determined by the mapping functions

$$\xi^i(x,t) = \mu_i[x^i - \phi_0^i(t)] + \gamma_i \phi_{2i-1}^i(t) + \nu_i \phi_{2i}^i(t), \quad \xi^0(x,t) = 0. \tag{4.43}$$

These mapping functions define a $3n$-parameter (sub)group of transformations with the generators

$$M_i \equiv [x^i - \phi_0^i(t)]\partial_i, \quad G_i \equiv \phi_{2i-1}^i(t)\partial_i, \quad N_i \equiv \phi_{2i}^i(t)\partial_i, \tag{4.44}$$

which have the following group structure:

$$[M_i,M_j] = 0, \quad [M_i,G_j] = -G_j\delta_j^i, \quad [M_i,N_j] = -N_i\delta_j^i,$$
$$[G_i,G_j] = 0, \quad [G_i,N_j] = 0, \quad [N_i,N_j] \equiv 0. \tag{4.45}$$

□

In Sec. VI a specific two-dimensional decoupled dynamical system which illustrates Corollaries 4.2.2 and 4.2.3 is analyzed.

An additional illustration of Corollaries 4.2.2 and 4.2.3 is found in Sec. VII, where the complete group of classical symmetries of the class of n-dimensional *isotropic* (decoupled) linear systems is determined by the above-described techniques.

V. EXAMPLE: CLASSICAL SYMMETRIES OF ALL ONE-DIMENSIONAL SECOND-ORDER LINEAR DYNAMICAL SYSTEMS

We shall now use Theorem 3.1, Corollary 3.1.1, and Theorem 4.2 to obtain the complete group of classical symmetries for *all* one-dimensional second-order linear dynamical systems (4.1), which with $n = 1$ may be written in the form

$$E = \ddot{x} - R(t)\dot{x} - S(t)x - T(t) = 0. \tag{5.1}$$

Remark 5.1: For the case $n = 1$ ($A = 1,2$) unneeded coordinate indices will be suppressed, for example, $x^1 = x$, $\phi_A^1 \equiv \phi_A$, $R_1^1 \equiv R, \alpha_{A1} \equiv \alpha_A$, etc. □

When $n = 1$ the formulas in Theorem 4.1 take relatively simple forms. They are listed below for convenience. For the one-dimensional system (5.1) the solution (4.2′) reduces to

$$x = c_1\phi_1(t) + c_2\phi_2(t) + \phi_0(t), \quad c_1,c_2 \text{ const} \tag{5.2}$$

and the Wronskian (4.3′) takes the form

$$W = \phi_1\dot{\phi}_2 - \phi_2\dot{\phi}_1 \neq 0. \tag{5.3}$$

The coefficients appearing in (5.1) are obtained from (4.8)–(4.10):

$$R(t) = W^{-1}(\phi_1\ddot{\phi}_2 - \phi_2\ddot{\phi}_1), \tag{5.4}$$
$$S(t) = W^{-1}(\dot{\phi}_2\ddot{\phi}_1 - \dot{\phi}_1\ddot{\phi}_2), \tag{5.5}$$
$$T(t) = \ddot{\phi}_0 - R(t)\dot{\phi}_0 - S(t)\phi_0. \tag{5.6}$$

Two constants of motion are obtained from (4.12)–(4.15) for the case $n = 1$:

$$C_1(\dot{x},x,t) = \alpha_1(t)\dot{x} + \beta_1(t)x + + \gamma_1(t) \overset{t}{=} c_1, \tag{5.7}$$

$$C_2(\dot{x},x,t) = \alpha_2(t)\dot{x} + \beta_2(t)x + \gamma_2(t) \overset{t}{=} c_2, \tag{5.8}$$

where

$$\alpha_1(t) = -W^{-1}\phi_2, \quad \alpha_2(t) = W^{-1}\phi_1, \tag{5.9}$$
$$\beta_1(t) = W^{-1}\dot{\phi}_2, \quad \beta_2(t) = -W^{-1}\dot{\phi}_1, \tag{5.10}$$
$$\gamma_1(t) = W^{-1}(\dot{\phi}_0\phi_2 - \phi_0\dot{\phi}_2),$$
$$\gamma_2(t) = -W^{-1}(\dot{\phi}_0\phi_1 - \phi_0\dot{\phi}_1). \tag{5.11}$$

It is of interest to note from (5.3) and (5.4) that in this $n = 1$ case

$$\dot{W} = RW. \tag{5.12}$$

We now turn to the solution of (4.21) for the functions $B^1(C_1,C_2)$, $B^2(C_1,C_2)$, where C_1 and C_2 are defined by (5.7)–(5.11). Equation (4.21) may be expressed in the form

$$\frac{\partial^2 B^1}{\partial C_2\,\partial C_2}\sigma^3 + \left(\frac{\partial^2 B^2}{\partial C_2\,\partial C_2} - 2\frac{\partial^2 B^1}{\partial C_1\,\partial C_2}\right)\sigma^2 + \left(\frac{\partial^2 B^1}{\partial C_1\,\partial C_1} - 2\frac{\partial^2 B^2}{\partial C_1\,\partial C_2}\right)\sigma + \frac{\partial^2 B^2}{\partial C_1\,\partial C_1} = 0, \tag{5.13}$$

where

$$\sigma(t) \equiv \phi_1(t)/\phi_2(t). \tag{5.14}$$

We note that Eq. (5.13) is of the form (3.27) discussed in Remark 3.5. Since in (5.13) the functions C_A and t are treated as independent variables (refer to Remark 3.4), it follows by successive differentiations of (5.13) with respect to t [note by (5.3) that $\dot{\sigma} = W/\phi_1^2 \neq 0$] that the coefficients of the various σ terms in (5.13) are zero. Hence we are led to the following conditions on $B^1(C_1,C_2)$, $B^2(C_1,C_2)$:

$$\frac{\partial^2 B^1(C_1,C_2)}{\partial C_2\,\partial C_2} = 0, \tag{5.15}$$
$$\frac{\partial^2 B^2(C_1,C_2)}{\partial C_1\,\partial C_1} = 0, \tag{5.16}$$
$$\frac{\partial^2 B^2(C_1,C_2)}{\partial C_2\,\partial C_2} - 2\frac{\partial^2 B^1(C_1,C_2)}{\partial C_1\,\partial C_2} = 0, \tag{5.17}$$
$$\frac{\partial^2 B^1(C_1,C_2)}{\partial C_1\,\partial C_1} - 2\frac{\partial^2 B^2(C_1,C_2)}{\partial C_1\,\partial C_2} = 0. \tag{5.18}$$

Equations (5.15)–(5.18) are easily solved to obtain

$$B^I(C_1,C_2) = \lambda^I_{JK}C_JC_K + \mu^I_JC_J + \nu^I, \quad I,J,K = 1,2, \tag{5.19}$$

where the constants λ^I_{JK} ($\equiv \lambda^I_{KJ}$) satisfy the restrictions

$$\lambda^1_{22} = 0, \ \lambda^2_{11} = 0, \ \lambda^2_{22} - 2\lambda^1_{12} = 0, \quad \lambda^1_{11} - 2\lambda^2_{22} = 0, \tag{5.20}$$

but are otherwise arbitrary, and the constants μ^I_J and ν^I are arbitrary.

It is a straightforward calculation to express B^I in (5.19) as a polynomial in $\dot{x}$ by use of the constants of motion C_1 and C_2 given by (5.7)–(5.11). The resulting polynomial, when used in (4.20′) with $n = 1$, determines the desired function $Z(\dot{x},x,t)$, which is linear in $\dot{x}$ (refer to Theorem 4.2). When this so-obtained linear Z is used in (3.4) and (3.5) of Corollary 3.1.1 we obtain [note that (5.20) has been used]

$$\xi(x,t) = \lambda^1_{11}W^{-1}(\dot{\phi}_2x^2 + \Gamma_2x + \Theta_2\phi_0) - \lambda^2_{22}W^{-1}(\dot{\phi}_1x^2 + \Gamma_1x + \Theta_1\phi_0) + \mu^1_1W^{-1}(\phi_1\dot{\phi}_2x - \Theta_2\phi_1) - \mu^1_2W^{-1}(\phi_1\dot{\phi}_1x - \Theta_1\phi_1)$$

$$+\mu_1^2 W^{-1}(\phi_2\dot{\phi}_2 x - \Theta_2\phi_2)$$
$$-\mu_2^2 W^{-1}(\phi_2\dot{\phi}_1 x - \Theta_1\phi_2) + \nu^1\phi_1 + \nu^2\phi_2, \quad (5.21)$$

$$\xi^0(x,t) = \lambda^1_{11} W^{-1}(\phi_2 x - \phi_0\phi_2)$$
$$-\lambda^2_{22} W^{-1}(\phi_1 x - \phi_0\phi_1)$$
$$+\mu_1^1 W^{-1}\phi_1\phi_2 - \mu_2^1 W^{-1}\phi_1\phi_1$$
$$+\mu_1^2 W^{-1}\phi_2\phi_2 - \mu_2^2 W^{-1}\phi_2\phi_1, \quad (5.22)$$

where

$$\Theta_A \equiv \phi_0\dot{\phi}_A - \phi_A\dot{\phi}_0, \quad \Gamma_A \equiv \phi_A\dot{\phi}_0 - 2\phi_0\dot{\phi}_A, \quad A = 1,2. \quad (5.23)$$

The mapping functions $\xi(x,t)$ in (5.21) and ξ^0 in (5.22) define the most general classical symmetries of the one-dimensional linear system (5.1). The eight arbitrary constants $\lambda^1_{11}, \lambda^2_{22}, \mu_1^1, \mu_2^1, \mu_1^2, \mu_2^2, \nu_1$, and ν^2 that appear in the mapping functions (5.21) and (5.22) determine, respectively, the following eight generators [each of the form $X \equiv \xi(x,t)\partial_x + \xi^0(x,t)\partial_t$]

$$\Lambda_1 \equiv W^{-1}(\dot{\phi}_2 x^2 + \Gamma_2 x + \Theta_2\phi_0)\partial_x + W^{-1}(\phi_2 x - \phi_0\phi_2)\partial_t, \quad (5.24)$$

$$\Lambda_2 \equiv W^{-1}(\dot{\phi}_2 x^2 + \Gamma_1 x + \Theta_1\phi_0)\partial_x + W^{-1}(\phi_1 x - \phi_0\phi_1)\partial_t, \quad (5.25)$$

$$M_1^1 \equiv W^{-1}(\phi_1\dot{\phi}_2 x - \Theta_2\phi_1)\partial_x + W^{-1}\phi_1\phi_2\,\partial_t, \quad (5.26)$$

$$M_2^1 \equiv W^{-1}(\phi_1\dot{\phi}_1 x - \Theta_1\phi_1)\partial_x + W^{-1}\phi_1\phi_1\,\partial_t, \quad (5.27)$$

$$M_1^2 \equiv W^{-1}(\phi_2\dot{\phi}_2 x - \Theta_2\phi_2)\partial_x + W^{-1}\phi_2\phi_2\,\partial_t, \quad (5.28)$$

$$M_2^2 \equiv W^{-1}(\phi_2\dot{\phi}_1 x - \Theta_1\phi_2)\partial_x + W^{-1}\phi_2\phi_1\,\partial_t, \quad (5.29)$$

$$N_1 \equiv \phi_1\,\partial_x, \quad (5.30)$$

$$N_2 \equiv \phi_2\,\partial_x. \quad (5.31)$$

The generators (5.24)–(5.31) determine the complete eight-parameter group of classical symmetries of the linear dynamical equation (5.1).

Remark 5.2: In the basis with the generators G_α, $\alpha = 1,\ldots,8$ defined by the basis change $G_1 = -(M_1^1 + M_2^2)$, $G_2 = M_2^1 - M_1^2$, $G_3 = N_2$, $G_4 = N_1$, $G_5 = -(M_2^1 + M_1^2)$, $G_6 = M_1^1 - M_2^2$, $G_7 = -\Lambda_2$, and $G_8 = -\Lambda_1$, the eight-parameter symmetry group (5.24)–(5.31) is isomorphic to both the eight-parameter symmetry group determined by Lutzky[5] in his analysis of the simple harmonic oscillator and the eight-parameter symmetry group determined by Leach[6] in his analysis of the time-dependent oscillator. See, also, Leach[7] and Lopez.[8] □

The linear dynamical system (5.1) is a Lagrangian system with the Lagrangian

$$L(\dot{x},x,t) = W^{-1}[\tfrac{1}{2}\dot{x}^2 + \tfrac{1}{2}S(t)x^2 + T(t)x]. \quad (5.32)$$

If for some function $\tau(x,t)$ there exist mappings of the form (3.1) and (3.2) such that

$$\delta L + L\frac{d}{dt}\delta t = -\frac{d\tau}{dt}\delta a, \quad (5.33)$$

a Lagrangian dynamical system is said to admit classical Noether symmetries. If a Lagrangian system admits such Noether mappings they will be a subgroup of the classical symmetries determined by variation of the dynamical equation (Lagrange's equation) described in Sec. II.[1] Hence to obtain the mapping functions $\xi(x,t)$ and $\xi^0(x,t)$, which determine the complete group of classical Noether symmetries of the linear system (5.1) characterized by the Lagrangian (5.32), we need only to find what restrictions the Noether symmetry condition (5.33) places upon the general classical symmetry mapping functions (5.21) and (5.22) and obtain the associated function $\tau(x,t)$.

The details of the above-described procedure for obtaining classical Noether symmetries of the linear dynamical system (5.1) are given in the Appendix. It is found that the dynamical system (5.1) admits a five-parameter group of classical Noether mappings determined by the mapping functions

$$\xi(x,t) = W^{-1}(\tfrac{1}{2}\dot{M}x - \tfrac{1}{2}\phi_0\dot{M} + \dot{\phi}_0 M) + N, \quad (5.34)$$

$$\xi^0(x,t) = W^{-1}M, \quad (5.35)$$

where

$$M(t) = -\mu_2^1(\phi_1)^2 + 2\mu_1^1\phi_1\phi_2 + \mu_1^2(\phi_2)^2, \quad (5.36)$$

$$N(t) = \nu^1\phi_1 + \nu^2\phi_2, \quad (5.37)$$

and the associated function $\tau(x,t)$ takes the form

$$\tau(x,t) = -\tfrac{1}{4}W^{-2}\{(\ddot{M} - R\dot{M})x^2$$
$$+ 2[2\ddot{\phi}_0 M + \dot{\phi}_0(M - 2RM)$$
$$- \phi_0(\ddot{M} - R\dot{M})]x + \phi_0^2(\ddot{M} - R\dot{M} - 2SM)$$
$$- 2(\phi_0\dot{\phi}_0\dot{M} - \dot{\phi}_0^2 M)\}$$
$$- W^{-1}(\dot{N}x - \phi_0\dot{N} + \dot{\phi}_0 N). \quad (5.38)$$

It is to be noted that the five parameters $\mu_1^1, \mu_2^1, \mu_1^2, \nu^1$, and ν^2 that appear in the classical Noether mapping functions (5.34), (5.35), and (5.38) also appear in the mapping functions (5.21) and (5.22), which determine the more general group of classical symmetry mappings for the dynamical system (5.1). Corresponding, respectively, to the above-mentioned five parameters we obtain from (5.34) and (5.35) the five generators of the complete classical Noether group [refer to (5.26)–(5.31)]: $\tilde{M}_1^1 \equiv M_1^1 + M_2^2$, M_2^1, M_1^2, N_1, and N_2.

VI. EXAMPLE: CLASSICAL SYMMETRIES OF A TWO-DIMENSIONAL DECOUPLED LINEAR SYSTEM

We now use Corollary 4.2.2 and Theorem 4.2 to obtain the classical symmetires of the two-dimensional decoupled system

$$\ddot{x}^1 - 6t^{-2}x^1 = 0, \quad (6.1)$$

$$\ddot{x}^2 - 12t^{-2}x^2 = 0. \quad (6.2)$$

Solutions to (6.1) and (6.2) may be expressed in the forms [refer to (4.28)–(4.30)]

$$x^1 = c_1 t^3 + c_2 t^{-2}, \quad (6.3)$$

$$x^2 = c_3 t^4 + c_4 t^{-3}, \quad (6.4)$$

so that

$$\phi_1^1 = t^3, \quad \phi_2^1 = t^{-2}, \quad \phi_3^1 = 0, \quad \phi_4^1 = 0, \quad \phi_0^1 = 0, \quad (6.5)$$

$$\phi_1^2 = 0, \quad \phi_2^2 = 0, \quad \phi_3^2 = t^4, \quad \phi_4^2 = t^{-3}, \quad \phi_0^2 = 0, \quad (6.6)$$

and

$$W_1 = -5, \quad W_2 = -7. \tag{6.7}$$

Inverting (6.3) and (6.4) we find the constants of motion to be [refer to (4.39) and (4.40)]

$$C_1 = \tfrac{1}{5}t^{-2}\dot{x}^1 + \tfrac{2}{5}t^{-3}x^1, \tag{6.8}$$

$$C_2 = -\tfrac{1}{5}t^{3}\dot{x}^1 + \tfrac{3}{5}t^{2}x^1, \tag{6.9}$$

$$C_3 = \tfrac{1}{7}t^{-3}\dot{x}^2 + \tfrac{3}{7}t^{-4}x^2, \tag{6.10}$$

$$C_4 = -\tfrac{1}{7}t^{4}\dot{x}^2 + \tfrac{4}{7}t^{3}x^2. \tag{6.11}$$

For the case $n = 2$ two equations are determined by (4.37): With use of (6.5) and (6.6) they may be expressed in the respective forms

$$\frac{\partial B^1}{\partial C_3} + t^{-5}\frac{\partial B^2}{\partial C_3} - t^7\frac{\partial B^1}{\partial C_4} - t^2\frac{\partial B^2}{\partial C_4} = 0, \tag{6.12}$$

$$t^2\frac{\partial B^3}{\partial C_1} + t^{-5}\frac{\partial B^4}{\partial C_1} - t^7\frac{\partial B^3}{\partial C_2} - \frac{\partial B^4}{\partial C_2} = 0. \tag{6.13}$$

Equation (4.38) leads to one equation when $n = 2$. With use of (6.5) and (6.6) this equation may be expressed in the form

$$\begin{aligned}&\frac{1}{7}t^8\frac{\partial B^3}{\partial C_4} - \frac{1}{5}t^6\frac{\partial B^1}{\partial C_2} \\ &\quad + t\left(\frac{1}{7}\frac{\partial B^4}{\partial C_4} - \frac{1}{7}\frac{\partial B^3}{\partial C_3} - \frac{1}{5}\frac{\partial B^2}{\partial C_2} + \frac{1}{5}\frac{\partial B^1}{\partial C_1}\right) \\ &\quad + \frac{1}{5}t^{-4}\frac{\partial B^2}{\partial C_1} - \frac{1}{7}t^{-6}\frac{\partial B^4}{\partial C_3} = 0.\end{aligned} \tag{6.14}$$

The linear independence of the coefficients in each of Eqs. (6.12)–(6.14) leads to the conditions

$$\frac{\partial B^1(C)}{\partial C_\alpha} = 0, \quad \alpha = 2,3,4, \tag{6.15}$$

$$\frac{\partial B^2(C)}{\partial C_\beta} = 0, \quad \beta = 1,3,4, \tag{6.16}$$

$$\frac{\partial B^3(C)}{\partial C_\gamma} = 0, \quad \gamma = 1,2,4, \tag{6.17}$$

$$\frac{\partial B^4(C)}{\partial C_\delta} = 0, \quad \delta = 1,2,3, \tag{6.18}$$

$$\frac{1}{7}\left[\frac{\partial B^4(C)}{\partial C_4} - \frac{\partial B^3(C)}{\partial C_3}\right] + \frac{1}{5}\left[\frac{\partial B^1(C)}{\partial C_1} - \frac{\partial B^2(C)}{\partial C_2}\right] = 0. \tag{6.19}$$

The solution to the system (6.15)–(6.19) is readily found to be

$$B^1 = a_1C_1 + b_1, \tag{6.20}$$

$$B^2 = a_2C_2 + b_2, \tag{6.21}$$

$$B^3 = a_3C_3 + b_3, \tag{6.22}$$

$$B^4 = (-\tfrac{7}{5}a_1 + \tfrac{7}{5}a_2 + a_3)C_4 + b_4. \tag{6.23}$$

The $B^A, A = 1,...,4$ given by (6.20)–(6.23) are next expressed as functions of $\dot{x}, x$, and t by use of the constants of motion (6.8)–(6.11). The resulting formulas for B^A, along with the functions ϕ^i_A in (6.5) and (6.6), are used in (4.20′) of Theorem 4.2 to obtain the functions

$$Z^1 = \tfrac{1}{5}(2a_1 + 3a_2)x^1 + b^1t^3 + b^2t^{-2} + \tfrac{1}{5}(a_1 - a_2)t\dot{x}^1, \tag{6.24}$$

$$Z^2 = (-\tfrac{4}{5}a_1 + \tfrac{4}{5}a_2 + a_3)x^2 + b^3t^4 + b^4t^{-3} + \tfrac{1}{5}(a_1 - a_2)t\dot{x}^2. \tag{6.25}$$

It now follows from (6.24), (6.25), and (3.3′) of Theorem 4.2 (or alternatively by Corollary 3.1.1) that the classical symmetry mapping (3.1) and (3.2) for the dynamical system (6.1) and (6.2) is determined by the mapping functions

$$\xi^1(x,t) = a_1(\tfrac{2}{5}x^1) + a_2(\tfrac{3}{5}x^1) + b_1t^3 + b_2t^{-2}, \tag{6.26}$$

$$\xi^2(x,t) = a_1(-\tfrac{4}{5}x^2) + a_2(\tfrac{4}{5}x^2) + a_3x^2 + b_3t^4 + b_4t^{-3}, \tag{6.27}$$

$$\xi^0(x,t) = a_1(-\tfrac{1}{5}t) + a_2(\tfrac{1}{5}t). \tag{6.28}$$

The mapping functions (6.26)–(6.28) determine a seven-parameter group with the parameters $a_1, a_2, a_3, b_1, b_2, b_3$, and b_4. The corresponding generators may be chosen to be

$$\begin{aligned}&A_1 = 2x^1\partial_1 - 4x^2\partial_2 - t\partial_t, \quad A_2 = 3x^1\partial_1 + 4x^2\partial_2 + t\partial_t, \\ &A_3 = x^2\partial_2, \quad B_1 = t^3\partial_1, \quad B_2 = t^{-2}\partial_1, \quad B_3 = t^4\partial_2, \\ &B_4 = t^3\partial_2.\end{aligned} \tag{6.29}$$

Corollary 4.2.3 is applicable to the dynamical system (6.1) and (6.2). By use of (6.5) and (6.6) we formulate the six $(3n, n = 2)$ generators M_1, M_2, G_1, G_2, N_1, and N_2 defined by (4.44) of Corollary 4.2.3. A comparison of these generators with the seven generators in (6.29) shows $M_1 = (A_1 + A_2)/5$, $M_2 = A_3$, $G_1 = B_1$, $G_2 = B_3$, $N_1 = B_2$, and $N_2 = B_4$, which verifies Corollary 4.2.3.

VII. EXAMPLE: CLASSICAL SYMMETIRES OF AN n-DIMENSIONAL, n > 1, ISOTROPIC LINEAR SYSTEM

If the coefficients appearing in the decoupled linear system (4.26) are chosen to be

$$R^i(t) = R(t), \quad S^i(t) = S(t), \quad i = 1,...,n, \tag{7.1}$$

then each of the n dynamical equations will be of the same form except for their nonhomogeneous terms $T^i(t)$. The dynamical system is then said to be *isotropic* and has the dynamical equations

$$E^i \equiv \ddot{x}^i - R(t)\dot{x}^i - S(t)x^i - T^i(t) = 0, \quad i = 1,...,n. \tag{7.2}$$

We apply our new method to obtain the classical symmetries of the isotropic system (7.2). The reader may wish to compare this approach with that of Lopez.[8]

With reference to (4.28)–(4.30) the solution to (7.2) may be expressed in the form

$$x^i(t) = c_{2i-1}\phi_1(t) + c_{2i}\phi_2(t) + \phi^i_0(t), \tag{7.3}$$

where [refer to (4.4)]

$$w \equiv W_i = \phi_1\psi_2 - \phi_2\psi_1 \neq 0, \quad i = 1,...,n, \tag{7.4}$$

$$\phi_1(t) \equiv \phi^i_{2i-1}(t), \quad \phi_2(t) \equiv \phi^i_{2i}(t), \quad i = 1,...,n, \tag{7.5}$$

$$\psi_1(t) \equiv \psi^i_{2i-1}(t), \quad \psi_2(t) \equiv \psi^i_{2i}(t), \quad i = 1,...,n. \tag{7.6}$$

The constants of motion (4.39) and (4.40) simplify by means of (7.4)–(7.6), so that for the isotropic system (7.2) they have the form

$$C_{2i-1} = w^{-1}(-\phi_2\dot{x}^i + \psi_2x^i + \psi^i_0\phi_2 - \phi^i_0\psi_2), \tag{7.7}$$

$$C_{2i} = w^{-1}(\phi_1 \dot{x}^i - \psi_1 x^i - \psi_1 x^i - \psi_o^i \phi_1 + \phi_0^i \psi_1). \quad (7.8)$$

It follows from (7.4) and (7.5) that for the isotropic system (7.2) Eqs. (4.37) and (4.38) of Corollary 4.2.2 take the form, respectively,

$$\tau^2 \frac{\partial B^{2i}}{\partial C_{2j-1}} + \tau\left(\frac{\partial B^{2i-1}}{\partial C_{2j-1}} - \frac{\partial B^{2i}}{\partial C_{2j}}\right) - \frac{\partial B^{2i-1}}{\partial C_{2j}} = 0, \quad i \neq j, \quad (7.9)$$

$$\tau^2\left(\frac{\partial B^{2i}}{\partial C_{2i-1}} - \frac{\partial B^{2j}}{\partial C_{2j-1}}\right) - \tau\left(\frac{\partial B^{2i}}{\partial C_{2i}} - \frac{\partial B^{2j}}{\partial C_{2j}} - \frac{\partial B^{2i-1}}{\partial C_{2i-1}} + \frac{\partial B^{2j-1}}{\partial C_{2j-1}}\right) - \left(\frac{\partial B^{2i-1}}{\partial C_{2i}} - \frac{\partial B^{2j-1}}{\partial C_{2j}}\right) = 0, \quad i,j \text{ not summed}, \quad (7.10)$$

where

$$\tau(t) \equiv \phi_{2i}^i / \phi_{2i-1}^i = \phi_2/\phi_1, \ i = 1,...,n. \quad (7.11)$$

By successive differentiation of both (7.9) and (7.10) with respect to t (recall C's and t are to be treated as independent variables) and noting from (7.11) and (7.4) that $\dot{\tau} \neq 0$, we find that the $B(C)$'s must satisfy the conditions

$$\frac{\partial B^{2i}}{\partial C_{2j-1}} = 0, \quad i \neq j, \quad (7.12)$$

$$\frac{\partial B^{2i-1}}{\partial C_{2j-1}} = \frac{\partial B^{2i}}{\partial C_{2j}}, \quad i \neq j, \quad (7.13)$$

$$\frac{\partial B^{2i-1}}{\partial C_{2j}} = 0, \quad i \neq j, \quad (7.14)$$

$$\frac{\partial B^{2i-1}}{\partial C_{2i}} = \frac{\partial B^{2j-1}}{\partial C_{2j}}, \quad \text{all } i,j, \quad (7.15)$$

$$\frac{\partial B^{2i}}{\partial C_{2i}} - \frac{\partial B^{2i-1}}{\partial C_{2i-1}} = \frac{\partial B^{2j}}{\partial C_{2j}} - \frac{\partial B^{2j-1}}{\partial C_{2j-1}}, \quad \text{all } i,j, \quad (7.16)$$

$$\frac{\partial B^{2i}}{\partial C_{2i}} = \frac{\partial B^{2j}}{\partial C_{2j-1}}, \quad \text{all } i,j. \quad (7.17)$$

Equations (7.12)–(7.17) may be solved in a straightforward fashion to obtain ($k = 1,...,n$)

$$B^{2i} = \sum_k \beta_0^k C_{2k} C_{2i} + \sum_k \alpha_0^k C_{2k} C_{2i-1} + \sum_{k \neq i} \gamma_0^{ik} C_{2k} + \sigma_0^i C_{2i} + \alpha_0 C_{2i-1} + \mu_0^i, \quad (7.18)$$

$$B^{2i-1} = \sum_k \alpha_0^k C_{2k-1} C_{2i-1} + \sum_k \beta_0^k C_{2k-1} C_{2i} + \sum_{k \neq i} \gamma_0^{ik} C_{2k-1} + \tau_0^i C_{2i-1} + \beta_0 C_{2i} + \nu_0^i, \quad (7.19)$$

where

$$\sigma_0^i - \tau_0^i = u_0, \quad \text{all } i. \quad (7.20)$$

The $B^A(C)$'s, $A = 1,...,2n$ defined by (7.18)–(7.20) are next expressed as functions of $\dot{x}, x$, and t by means of (7.7) and (7.8). The resulting functions, together with the ϕ's in (7.5), are employed in (4.20′) of Theorem 4.2 to obtain the functions Z^i [wherein the parameters σ_0^i, $i = 1,...,n$ have been expressed in terms of τ_0^i and u_0 by (7.20), the notational change $\tau_0^i \equiv \gamma_0^{ii}$ has been adopted, and use has been made of the definitions (4.4) and (4.7)]:

$$\begin{aligned} Z^i = {} & \sum_k \alpha_0^k w^{-1}[-(x^k - \phi_0^k)\phi_2 \dot{x}^i + \dot{\phi}_2 x^i x^k - \Theta_2^i x^k - \phi_0^k \dot{\phi}_2 x^i + \Theta_2^i \phi_0^k] \\ & + \sum_k \beta_0^k w^{-1}[(x^k - \phi_0^k)\phi_1 \dot{x}^i - \dot{\phi}_1 x^i x^k - \Theta_1^i x^k + \phi_0^k \dot{\phi}_1 x^i - \Theta_1^i \phi_0^k] \\ & + \sum_k \gamma_0^{ik}(x^k - \phi_0^k) + \beta_0 w^{-1}(\phi_1 \phi_1 \dot{x}^i - \phi_1 \dot{\phi}_1 x^i + \Theta_1^i \phi_1) + u_0 w^{-1}(\phi_1 \phi_2 \dot{x}^i - \phi_2 \dot{\phi}_1 x^i + \Theta_1^i \phi_2) \\ & + \alpha_0 w^{-1}(-\phi_2 \phi_2 \dot{x}^i + \phi_2 \dot{\phi}_2 x^i + \Theta_2^i \phi_2) + \nu_0^i \phi_1 + \mu_0^i \phi_2, \end{aligned} \quad (7.21)$$

where

$$\Theta_A^i \equiv \phi_0^i \dot{\phi}_A - \phi_A \dot{\phi}_0^i. \quad (7.22)$$

By means of (3.3′) of Theorem 4.2 [or by use of Corollary 3.1.1 and (7.21)] we obtain the mapping functions

$$\begin{aligned} \xi^i = {} & \sum_k \alpha_0^k w^{-1}(\dot{\phi}_2 x^i x^k - \Theta_2^i x^k - \phi_0^k \dot{\phi}_2 x^i + \Theta_2^i \phi_0^k) + \sum_k \beta_0^k w^{-1}(\dot{\phi}_1 x^i x^k + \Theta_1^i x^k + \phi_0^k \dot{\phi}_1 x^i + \Theta_1^i \phi_0^k) \\ & + \sum_k \gamma_0^{ik}(x^k - \phi_0^k) - \beta_0 w^{-1}(\phi_1 \dot{\phi}_1 x^i - \Theta_1^i \phi_1) - u_0 w^{-1}(\phi_2 \dot{\phi}_1 x^i - \Theta_1^i \phi_2) + \alpha_0 w^{-1}(\phi_2 \dot{\phi}_2 x^i - \Theta_2^i \phi_2) \\ & + \nu_0^i \phi_1 + \mu_0^i \phi_2, \end{aligned} \quad (7.23)$$

$$\begin{aligned} \xi^0 = {} & \sum_k \alpha_0^k w^{-1}(x^k - \phi_0^k)\phi_2 - \sum_k \beta_0^k w^{-1}(x^k - \phi_0^k)\phi_1 - \beta_0 w^{-1}\phi_1 \phi_1 \\ & - u_0 w^{-1}\phi_1 \phi_2 + \alpha_0 w^{-1}\phi_2 \phi_2. \end{aligned} \quad (7.24)$$

The $N = n^2 + 4n + 3$ arbitrary constants γ_0^{ik}, α_0^k, β_0^k ν_0^i, μ_0^i, α_0, β_0, and u_0 that appear in the mapping functions (7.23) and (7.24) determine N infinitesimal symmetry mappings (3.1) and (3.2). These mapping define the complete N-parameter group of classical symmetries of the n-dimensional isotropic linear system (7.2) and have been obtained (in the same basis) by Lopez[8] by an alternative method.

In Theorem 4.2 and Corollary 4.2.2 it was necessary to distinguish the cases $n = 1$ and $n > 1$ when formulating the conditions for obtaining the $B^A(C)$'s that determine linear $Z^i(\dot{x}^i,x,t)$ and therefore classical symmetries. It is of interest to note, however, that if the mapping functions $\xi^i(x,t)$ in (7.23) and $\xi^0(x,t)$ in (7.24), which were based upon the $n > 1$ formulation for the isotropic systems (7.2), are formally evaluated for $n = 1$ the resulting mapping functions lead to eight generators ($X = \xi\partial_x + \xi^0\partial_t$) which by a simple basis change can be expressed in the same form as those [(5.24)–(5.31)] determined by the $n = 1$ formulation for the one-dimensional dynamical system (5.1).

The isotropic system (7.2) is a decoupled system; hence Corollary 4.2.3 is applicable. The $3n$ generators determined by use of the solution functions (7.6) of the isotropic system in (4.44) of Corollary 4.2.3 are seen by inspection to be those associated with the $3n$-parameters γ_0^{ii}, ν_0^i, and μ_0^i, $i = 1,\ldots,n$ which appear in the above-determined mapping functions ξ^i in (7.23) and ξ^0 in (7.24).

VIII. EXAMPLE: CLASSICAL SYMMETRIES OF THE DYNAMICAL SYSTEM $\ddot{x} - \dot{x}^2/x = 0$, $(n = 1)$

The theory developed in Sec. III will now be used to determine the classical symmetries admitted by the one-dimensional ($n = 1$) dynamical system

$$\ddot{x} - \dot{x}^2/x = 0. \tag{8.1}$$

Remark 8.1: By the coordinate transformation

$$x = e^y \tag{8.2}$$

the nonlinear equation (8.1) can be converted into the simple linear equation

$$\ddot{y} = 0. \tag{8.3}$$

Clearly, the classical symmetries of (8.3) may be readily found (with the appropriate notation change $x \to y$) as a special case of those derived in Sec. V for the general one-dimensional linear system (5.1); when obtained, these symmetries could by (8.2) be expressed in terms of the coordinate system of (8.1). However, *solely for illustration purposes*, we shall work directly with the nonlinear equation (8.1) to show how Theorem 3.1, Theorem 3.2, and Corollary 3.1.1 may be used to obtain the classical symmetries of a simple nonlinear equation. □

Remark 8.2: The dynamical equation (8.1) was used in Ref. 1 to illustrate the procedure for determining the characteristic functional structure of the auxiliary symmetry mapping function $Z^i(\dot{x},x,t)$ in (2.27). Use will be made of results of that calculation in the present illustration. □

Equation (8.1) has the solution

$$x = c_1 e^{c_2 t}, \quad c_1, c_2 = \text{const}, \tag{8.4}$$

from which it follows that

$$\dot{x} = c_1 c_2 e^{c_2 t}. \tag{8.5}$$

From (8.4) and (8.5) there is obtained functionally independent constants of motion [refer to (2.3)]

$$C_1(\dot{x},x,t) = xe^{-\dot{x}t/x} \stackrel{t}{=} c_1, \tag{8.6}$$

$$C_2(\dot{x},x,t) = \dot{x}/x \stackrel{t}{=} c_2. \tag{8.7}$$

For the dynamical system (8.1) the auxiliary symmetry condition [Eq. (2.15′) of Theorem 3.2] takes the form

$$\ddot{Z} - (2\dot{x}/x)\dot{Z} + (\dot{x}/x)^2 Z \stackrel{\circ}{=} 0. \tag{8.8}$$

From Sec. VII of Ref. 1 the solution to the partial differential equation obtained by the formal expansion of (8.8) is expressible in the form [refer to (2.27′) of Theorem 3.2]

$$Z = B^1(C_1,C_2)e^{C_2 t} + B^2(C_1,C_2)te^{C_2 t}, \tag{8.9}$$

where C_1, C_2 are given by (8.6) and (8.7), respectively, and where the functions $B^1(C_1, C_2)$, $B^2(C_1,C_2)$ are arbitrary.

From Sec. VII of Ref. 1 or by comparison of (8.9) and (2.27) it follows that the functions $g_A^i(C,t)$ [$\equiv g_A(C,t), n = 1$] have the form

$$g_1(C,t) = e^{C_2 t}, \quad g_2(C,t) = te^{C_2 t}. \tag{8.10}$$

We now determine the $B(C)$'s, so that Z in (8.9) will be linear in $\dot{x}$, as required for classical symmetries.

With reference to (8.4)–(8.7) and the discussion in the paragraph preceding Theorem 3.2 it is found that the functions $P_K(C,t)$ in (3.18) and $Q_K(C,t)$ in (3.19) are

$$P_1(C,t) = -te^{-C_2 t}, \tag{8.11}$$

$$P_2(C,t) = e^{-C_2 t}/C_1, \tag{8.12}$$

$$Q_1(C,t) = t^2 e^{-2C_2 t}/C_1, \tag{8.13}$$

$$Q_2(C,t) = 0. \tag{8.14}$$

The η's defined by (3.15)–(3.17), ($n = 1$) are evaluated by means of (8.10)–(8.14) and used in (3.12). The resulting equation may be expressed in the form

$$t^3 F_3 + t^2 F_2 + tF_1 + F_0 = 0, \tag{8.15}$$

where the F_α, $\alpha = 0,\ldots,3$ are functions of C_1 and the B's and their derivatives with respect to C_1 and C_2.

With reference to Remark 3.4, it follows that $F_\alpha = 0$, ($\alpha = 0,1,2,3$). This leads to the following conditions on the $B^A(C_1,C_2)$:

$$C_1^2 \frac{\partial^2 B^2}{\partial C_1 \partial C_1} - C_1 \frac{\partial B^2}{\partial C_1} + B^2 = 0, \tag{8.16}$$

$$C_1^2 \frac{\partial^2 B^1}{\partial C_1 \partial C_1} - 2C_1 \frac{\partial^2 B^2}{\partial C_1 \partial C_2} - C_1 \frac{\partial B^1}{\partial C_1} + 2\frac{\partial B^2}{\partial C_2} + B^1 = 0, \tag{8.17}$$

$$\frac{\partial^2 B^2}{\partial C_2 \partial C_2} - 2C_1 \frac{\partial^2 B^1}{\partial C_1 \partial C_2} + 2\frac{\partial B^1}{\partial C_2} = 0, \tag{8.18}$$

$$\frac{\partial^2 B^1}{\partial C_2 \partial C_2} = 0. \tag{8.19}$$

Equations (8.16)–(8.19) can be solved without difficulty to obtain

$$B^1 = \tfrac{1}{2}a_1C_2C_1\ln C_1 + a_2C_1C_2 + a_3C_1 + a_4C_1\ln C_1 + a_5C_1(\ln C_1)^2, \quad (8.20)$$

$$B^2 = \tfrac{1}{2}a_1C_1C_2^2 + a_6C_1C_2 + a_7C_1 + a_5C_2C_1\ln C_1 + a_8C_1\ln C_1. \quad (8.21)$$

The B's in (8.20) and (8.21) are used in (8.9). If the resulting equation is expressed in terms of $\dot{x}$, x, and t by means of (8.6) and (8.7) we obtain $Z(\dot{x},x,t)$, which has the desired linearity in $\dot{x}$. By means of (3.4) and (3.5) of Corollary 3.1.1 this linear Z leads to the following eight-parameter symmetry mapping functions:

$$\xi(x,t) = a_3x + a_4x\ln x + a_5x(\ln x)^2 + a_7tx + a_8tx\ln x, \quad (8.22)$$

$$\xi^0(x,t) = -\tfrac{1}{2}a_1\ln x - a_2 + a_4t + a_5t\ln x - a_6t + a_8t^2. \quad (8.23)$$

By Theorem 3.1 the mapping functions (8.22) and (8.23) determine the complete eight-parameter group of classical symmetry mappings for the dynamical equation (8.1).

With reference to Remark 8.1, we may express the mapping functions $\xi(x,t)$ in (8.22) and $\xi^0(x,t)$ in (8.23) in terms of the y coordinates to obtain $\bar{\xi}(y,t)$ and $\bar{\xi}^0(y,t)$, where

$$\bar{\xi}(y,t) = \frac{\partial y}{\partial x}\xi(x,t) = e^{-y}\xi[x(y),t], \quad (8.24)$$

$$\bar{\xi}^0(y,t) = \xi^0[x(y),t]. \quad (8.25)$$

As to be expected (and the reader may readily verify), the transformed functions $\bar{\xi}(y,t)$ and $\bar{\xi}^0(y,t)$ will be those obtained by specializing the mapping functions (5.21) and (5.22) of a general linear system (7.1) to the case of a free particle (8.3) by taking $\phi_1(t) = t$, $\phi_2(t) = 1$, $\phi_0(t) = 0$ and making the notational change $x \to y$.

IX. EXAMPLE: CLASSICAL SYMMETRIES OF THE DYNAMICAL SYSTEM $\ddot{x} + \dot{s}^2e^{\dot{x}/\dot{s}} - (\ddot{s}/\dot{s})\dot{x} = 0$

As our final illustration of Theorem 3.1, Theorem 3.2, and Corollary 3.1.1 we determine the classical symmetries of the class of one-dimensional *nonlinear* dynamical systems

$$\ddot{x} + \dot{s}^2e^{\dot{x}/\dot{s}} - (\ddot{s}/\dot{s})\dot{x} = 0, \quad s = s(t), \quad \ddot{s} \neq 0. \quad (9.1)$$

Equation (9.1) has solution

$$x = (s + c_1)[1 - \ln(s + c_1)] + c_2, \quad c_1,\ c_2 = \text{const}, \quad (9.2)$$

from which it follows that

$$\dot{x} = \dot{s}\ln(s + c_1). \quad (9.3)$$

By inversion of (9.2) and (9.3) we obtain the functionally independent constants of motion [refer to (2.3)]

$$C_1(\dot{x},x,t) = e^{-\dot{x}/\dot{s}} - s \overset{t}{=} c_1, \quad (9.4)$$

$$C_2(\dot{x},x,t) = -(1 + \dot{x}/\dot{s})e^{-\dot{x}/\dot{s}} + x \overset{t}{=} c_2. \quad (9.5)$$

The auxiliary symmetry condition [Eq. (2.15′) of Theorem 3.2] for the dynamical equation (9.1) is given by

$$\ddot{Z} + (\dot{s}e^{\dot{x}/\dot{s}} - \ddot{s}/\dot{s})\dot{Z} \overset{\circ}{=} 0. \quad (9.6)$$

Following the procedure described in Sec. II (refer to the Alternative Proof) we obtain an associated equation [of the form (2.25)]:

$$\ddot{z} + [\dot{s}(s + C_1)^{-1} - \ddot{s}\dot{s}^{-1}]\dot{z} \overset{\circ}{=} 0. \quad (9.7)$$

The solution to (9.7) is

$$z(C,t) = B_1(C_1,C_2)\ln(s + C_1) + B_2(C_1,C_2), \quad (9.8)$$

where B_1 and B_2 are arbitrary functions of C_1 and C_2. Hence the partial differential equation obtained by formal expansion of the auxiliary symmetry condition (9.6) has the solution

$$Z(\dot{x},x,t) = B_1[C_1(\dot{x},x,t),C_2(\dot{x},x,t)]\ln[s(t) + C_1(\dot{x},x,t)] + B_2[C_1(\dot{x},x,t),C_2(\dot{x},x,t)], \quad (9.9)$$

where B_1 and B_2 are arbitrary functions of the constants of motion C_1 in (9.4) and C_2 in (9.5).

Remark 9.1: As mentioned in Remark 3.6, the new procedure for obtaining classical symmetries introduced in this paper may, for certain problems, lead to classical symmetry solutions by inspection. In practice this may be accomplished by a judicious choice of the arbitrary constants of motion $B^A[C(\dot{x},x,t)]$ which occur in the general auxiliary symmetry mapping functions $Z^i(\dot{x},x,t)$ in (2.27), so that the resulting Z^i are linear in $\dot{x}^i$. As an illustration consider the dynamical system (9.1). To formulate a Z that is linear in $\dot{x}$ from the general $Z(\dot{x},x,t)$ function (9.9) we note by inspection from (9.4) that $\ln(s + C_1) = -\dot{x}/\dot{s}$ and hence the choice $B^1 = a_2 = \text{const}$, $B^2 = 0$ immediately gives $Z = -a_2\dot{x}/\dot{s}$, which results in the classical symmetry mapping functions $\xi = 0$, $\xi^0 = a_2\dot{s}^{-1}$. □

To obtain the complete group of classical symmetry mappings for the dynamical system (9.1) we now continue with the formal procedure for determining the functions $B^A(C)$, so that $Z(\dot{x},x,t)$ in (9.9) will be linear in $\dot{x}$.

By comparison of (9.9) with (2.27) it follows that the functions $g_A^1(C,t)$ [$\equiv g_A(C,t), n = 1$] have the form

$$g_1(C,t) = \ln[s(t) + C_1], \quad g_2 = 1. \quad (9.10)$$

With reference to (2.16)–(2.20) and the paragraph preceding Theorem 3.2, it follows from (9.2)–(9.5) that the functions $P_K(C,t)$ in (3.18) and $Q_K(C,t)$ in (3.19) take the form

$$P_1(C,t) = -\dot{s}^{-1}(s + C_1), \quad (9.11)$$

$$P_2(C,t) = -\dot{s}^{-1}(s + C_1)\ln(s + C_1), \quad (9.12)$$

$$Q_1(C,t) = \dot{s}^{-2}(s + C_1), \quad (9.13)$$

$$Q_2(C,t) = \dot{s}^{-2}(s + C_1)[1 + \ln(s + C_1)]. \quad (9.14)$$

The η's defined by (3.15)–(3.17) are evaluated by means of (9.10)–(9.14) and used in (3.12) to obtain an equation of the form

$$\lambda_1 + \lambda_2v + \lambda_3u + \lambda_4v^2 + \lambda_5uv + \lambda_6uv^2 + \lambda_7uv^3 = 0, \quad (9.15)$$

where λ_α, $\alpha = 1,\ldots,7$ are linear combinations of the first and second derivatives of the B's with respect to the C's (defined below) and

$$u \equiv s + C_1, \quad v \equiv \ln(s + C_1). \quad (9.16)$$

With reference to Remark 3.4, it is readily shown by successive partial differentiation of Eq. (9.15) with respect to t that $\lambda_\alpha = 0$, $\alpha = 1,...7$. The following equations for $B^A(C_1,C_2)$, $A = 1,2$ result:

$$\lambda_1 \equiv 2\frac{\partial B^1}{\partial C_1} + \frac{\partial B^2}{\partial C_1} + \frac{\partial B^2}{\partial C_2} = 0, \tag{9.17}$$

$$\lambda_2 \equiv \frac{\partial B^1}{\partial C_1} + 3\frac{\partial B^1}{\partial C_2} + \frac{\partial B^2}{\partial C_2} = 0, \tag{9.18}$$

$$\lambda_3 \equiv \frac{\partial^2 B^2}{\partial C_1\,\partial C_1} = 0, \tag{9.19}$$

$$\lambda_4 \equiv \frac{\partial B^1}{\partial C_2} = 0, \tag{9.20}$$

$$\lambda_5 \equiv \frac{\partial^2 B^1}{\partial C_1\,\partial C_1} + 2\frac{\partial^2 B^2}{\partial C_1\,\partial C_2} = 0, \tag{9.21}$$

$$\lambda_6 \equiv \frac{\partial^2 B^2}{\partial C_2\,\partial C_2} + 2\frac{\partial^2 B^1}{\partial C_1\,\partial C_2} = 0, \tag{9.22}$$

$$\lambda_7 \equiv \frac{\partial^2 B^1}{\partial C_2\,\partial C_2} = 0. \tag{9.23}$$

Equations (9.17)–(9.23) are easily solved to obtain

$$B^1 = a_1 C_1 + a_2, \tag{9.24}$$

$$B^2 = -a_1(C_1 + C_2) + a_3, \tag{9.25}$$

where a_1, a_2, and a_3 are arbitrary constants.

Use of B^1 in (9.24) and B^2 in (9.25) in (9.9) results in a linear Z of the form

$$Z(\dot{x},x,t) = -a_1(x-s) + a_3 - \dot{x}(-a_1 s\dot{s}^{-1} + a_2\dot{s}^{-1}). \tag{9.26}$$

By comparison of (9.26) with (3.3) one obtains the classical symmetry mapping functions

$$\xi = -a_1(x-s) + a_3, \tag{9.27}$$

$$\xi^0 = -a_1 s\dot{s}^{-1} + a_2\dot{s}^{-1}. \tag{9.28}$$

When used in (3.1) and (3.2) the mapping functions (9.27) and (9.28) define a three-parameter (complete) group of infinitesimal classical symmetry mappings of the dynamical system (9.1).

APPENDIX: CLASSICAL NOETHER SYMMETRIES ADMITTED BY THE DYNAMICAL EQUATION $\ddot{x} - R(t)\dot{x} - S(t) - T(t) = 0$

When the Lagrangian (5.32) [associated with the linear dynamical system (5.1)] is used in the Noether symmetry condition (5.33) the resulting equation (cubic in $\dot{x}$) leads to the following conditions on the Noether mapping functions $\xi(x,t)$, $\xi^0(x,t)$, and $\tau(x,t)$ ($\xi_x \equiv \partial\xi/\partial x$, etc.):

$$\xi^0_x = 0, \tag{A1}$$

$$\xi_x = \tfrac{1}{2}\xi^0_t - \tfrac{1}{2}R\xi^0 = 0, \tag{A2}$$

$$\xi_t + (\tfrac{1}{2}Sx^2 + Tx)\xi^0_x + W\tau_x = 0, \tag{A3}$$

$$(\tfrac{1}{2}Sx^2 + Tx)\xi^0_t + [\tfrac{1}{2}(\dot{S} - RS)x^2 + (\dot{T} - RT)x]\xi^0 + (Sx + T)\xi + W\tau_1 = 0, \tag{A4}$$

where $W(t)$, $R(t)$, $S(t)$, and $T(t)$ are defined by (5.3)–(5.6).

Following the procedure discussed in the paragraph containing Eq. (5.33) we shall require that the general classical mapping functions (5.21) and (5.22) satisfy the Noether symmetry equations (A1)–(A4).

Use of ξ^0 in (5.22) in (A1) shows that in order to satisfy the Noether conditions the parameters λ^1_{11} λ^2_{22} that appear in the general mapping functions (5.21) and (5.22) must satisfy

$$\lambda^1_{11} = \lambda^2_{22} = 0. \tag{A5}$$

If ξ in (5.21) and ξ^0 in (5.22) are simplified by means of (A5) and the resulting functions used in (A2) it is found [with use of (5.12)] that the parameters μ^1_1 μ^2_2 must satisfy the condition

$$\mu^1_1 + \mu^2_2 = 0. \tag{A6}$$

We may, by use of (A6), eliminate μ^2_2 in the above-described simplified ξ and ξ^0 to obtain $\xi(x,t)$ in (5.34) and $\xi^0(t)$ in (5.35).

Equations (A3) and (A4) remain to be solved for τ. Before considering these equations we first derive an identity involving M in (5.36) and its derivatives which will be useful in obtaining the solution of these two remaining equations. To obtain this identity consider the quadratic constant of motion

$$Q \equiv \mu^2_1 C_1^2 - 2\mu^1_1 C_1 C_2 - \mu^1_2 C_2^2 \tag{A7}$$

defined in terms of the linear constants of motion C_1 in (5.7) and C_2 in (5.8) of the dynamical system (5.1). By means of (5.7)–(5.11) Q in (A7) can be expressed in the form

$$Q = W^{-2}\{M\dot{x}^2 - \dot{M}x\dot{x} + \tfrac{1}{2}(\ddot{M} - R\dot{M} - 2SM)x^2 + (\phi_0\dot{M} - 2\dot{\phi}_0 M)\dot{x} + [\dot{\phi}_0\dot{M} - \phi_0(\ddot{M} - R\dot{M} - 2SM)]x + \tfrac{1}{2}\phi_0^2(\ddot{M} - R\dot{M} - 2SM) - \phi_0\dot{\phi}_0\dot{M} + \dot{\phi}_0^2 M\}. \tag{A8}$$

From the manner in which Q in (A7) was defined in terms of the constants of motion C_1 in (5.7) and C_2 in (5.8) it follows that (refer to Remark 2.2)

$$\dot{Q} \overset{\circ}{\equiv} 0. \tag{A9}$$

Use of (A8) in (A9) [with (5.12) to eliminate the $\dot{W}$ terms, (5.1) to eliminate $\ddot{x}$ terms, and (5.6) to eliminate any remaining functions T] leads to

$$[\dddot{M} - 3R\ddot{M} - (\dot{R} - 2R^2 + 4S)\dot{M} - 2(\dot{S} - 2RS)M](x - \phi_0)^2 \equiv 0. \tag{A10}$$

Since (A10) must hold for all solutions (5.2) we obtain the desired identity

$$\dddot{M} - 3R\ddot{M} - (\dot{R} - 2R^2 + 4S)\dot{M} - 2(\dot{S} - 2RS)M \equiv 0. \tag{A11}$$

We now return to the solution of (A3) and (A4). By use of (5.35) we may integrate (A3) to obtain

$$\tau(x,t) = -W^{-1}\xi_t x + g(t), \tag{A12}$$

where $g(t)$ is to be determined. Substitution of τ in (A12) into (A4) with use of (5.6), (5.34)–(5.37), and (A11) leads to

$$\dot{g}(t) = g_1(t) + g_2(t), \tag{A13}$$

where

$$g_1(t) \equiv W^{-2}(\ddot{\phi}_0 - R\dot{\phi}_0 - S\phi_0)(\tfrac{1}{2}\phi_0\dot{M} - \dot{\phi}_0 M), \tag{A14}$$

$$g_2(t) \equiv -W^{-1}(\ddot{\phi}_0 - R\dot{\phi}_0 - S\phi_0)N. \tag{A15}$$

To integrate (A13) we shall express $g_1(t)$ in (A14) and $g_2(t)$ in (A15) as total derivatives. Consider first $g_1(t)$. If to the rhs of (A14) we substract the well-chosen zero obtained by multiplying the lhs of (A11) by $W^{-2}\phi_0^2/4$, we find that g_1 is expressible in the form

$$g_1(t) = \frac{d}{dt}\left\{ -\frac{1}{2} W^{-2}\left[\frac{1}{2}\phi_0^2(\ddot{M} - R\dot{M} - 2SM) - \phi_0\dot{\phi}_0\dot{M} + \dot{\phi}_0^2 M\right]\right\}. \tag{A16}$$

Next consider $g_2(t)$. If in (A15) we make the substitution $SN = \ddot{N} - R\dot{N}$ [which follows from (5.37), (5.1), and (5.2)] and use is made of (5.12) we find g_2 is expressible in the form

$$g_2(t) = \frac{d}{dt}\left[W^{-1}(\phi_0\dot{N} - \dot{\phi}_0 N)\right]. \tag{A17}$$

It follows from (A13), (A16), and (A17) that

$$g(t) = -\tfrac{1}{2}W^{-2}\left[\tfrac{1}{2}\phi_0^2(\ddot{M} - R\dot{M} - 2SM) - \phi_0\dot{\phi}_0\dot{M} + \dot{\phi}_0^2 M\right] + W^{-1}(\phi_0\dot{N} - \dot{\phi}_0 N), \tag{A18}$$

where the constant of integration has been dropped.

The function $\tau(x,t)$ may now be determined by evaluation of (A12) by means of (5.34) and (A18) to obtain (5.38).

[1]G. H. Katzin and J. Levine, J. Math. Phys. **26**, 3080 (1985).
[2]A discussion of the characteristic functional structure of first-order systems of differential equations is given by G. H. Katzin and J. Levine, J. Math. Phys. **26**, 3100 (1985).
[3]Special symmetries admitted by first- or second-order systems of differential equations that have cyclic variables may be determined by use of the characteristic functional structure theory. See G. H. Katzin and J. Levine, J. Math. Phys. **27**, 1756 (1986).
[4]W. Sarlet and F. Cantrijn, SIAM Rev. **23**, 467 (1981).
[5]M. Lutzky, J. Phys. A: Math, Gen. **11**, 249 (1978).
[6]P. G. L. Leach, J. Math. Phys. **21**, 300 (1980).
[7]P. G. L. Leach, Research Report No. AM-79:05 (La Trobe Univ., Bundoora, Australia, 1979).
[8]A. G. Lopez, J. Math. Phys. **29**, 1097 (1988).
[9]Repeated indices are summed (lower case 1–n, upper case 1–2n) unless otherwise indicated or apparent.
[10]H. R. Lewis, Jr., Phys. Rev. Lett. **18**, 510 (1967); J. Math. Phys. **9**, 1976 (1968); Phys. Rev. **172**, 1313 (1968).
[11]M. Aguire and J. Krause, J. Math. Phys. **29**, 9 (1988).
[12]G. H. Katzin, J. Levine, and R. N. Sane, J. Math. Phys. **18**, 424 (1977).
[13]G. H. Katzin and J. Levine, J. Math. Phys. **18**, 1267 (1977).
[14]C. J. Eliezer and A. Gray, SIAM J. Appl. Math. **30**, 463 (1976).
[15]H. R. Lewis and P. G. L. Leach, J. Math Phys. **23**, 165 (1982).
[16]M. Lutzky, Phys. Lett. A **68**, 3 (1978).
[17]G. E. Prince and C. J. Eliezer, J. Phys. A: Math. Gen. **13**, 815 (1980).
[18]R. K. Colegrave and M. A. Mannan, J. Math. Phys. **29**, 1580 (1988).

The Helmholtz conditions in terms of constants of motion in classical mechanics

Francisco Pardo
Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Circuito Exterio, C.U., 04510 México, D.F., Mexico

(Received 18 October 1988; accepted for publication 25 January 1989)

The Helmholtz conditions are the necessary and sufficient conditions for a set of second-order differential equations to be equivalent to a variational principle. In this work an alternative approach to the inverse problem in classical mechanics is described. It is proven that the Helmholtz conditions can be transformed into a set of conditions for a nonsingular antisymmetric matrix whose entries are constants of motion of the problem in question.

I. INTRODUCTION

Classical mechanics deals with systems of particles and their equations of motion. The equations of motion will be considered as a set of second-order differential equations whose solutions determine the evolution of a given system of particles. When written as functions of time, these solutions will be called trajectories.

Although a set of differential equations determines a unique set of trajectories—its solutions—the converse is not true, that is, a set of trajectories does not determine uniquely a set of differential equations. For example, take the set of differential equations

$$\ddot{q}^1 = 0, \quad \ddot{q}^2 = 0. \tag{1.1}$$

The set of solutions to Eqs. (1.1) are the curves (with α_i, β_i constant)

$$q^1(t) = \alpha_1 t + \beta_1, \quad q^2(t) = \alpha_2 t + \beta_2. \tag{1.2}$$

However, the curves (1.2) are also solutions of

$$\ddot{q}^1 + \ddot{q}^2 = 0, \quad \ddot{q}^1 - \ddot{q}^2 = 0. \tag{1.3}$$

Basically, the sets of differential equations (1.1)–(1.3) are different; however, their sets of solutions coincide. This fact means that a given set of trajectories S can be determined by more than one set of differential equations and what is more, it could be that one does not need, in principle, a set of differential equations for determining such a set of trajectories. With this idea in mind, it is then natural to seek other methods of obtaining the set of trajectories. One method is the Hamilton principle,[1] which states that the set of trajectories S of a given system of particles are those curves that make stationary a certain functional.

In general, for a given set of trajectories S, this functional does not exist; if it does, it could happen that it is not unique. The problem of determining the existence of this functional is known as the inverse problem of the variational calculus, a problem whose study dates back to the past century.[2,3] The related problem of nonuniqueness is known as the problem of s-equivalent Lagrangians.[4–9]

The inverse problem of the variational calculus in classical mechanics can be studied using different, but equivalent approaches. For example, Sarlet[10] proves that the Helmholtz conditions, as written in Sec. II, can be transformed, in principle, into an infinite set of algebraic equations. Henneaux,[8] using a more geometrical method (differential forms), analyzes the inverse problem and shows that the Helmholtz conditions are in general very strong, that is, he proves that in general the Helmholtz conditions do not have a solution and if they do, this solution will in general be unique.

Hojman *et al.*[11–16] also study the inverse problem of the variational calculus using what they call the first-order approach to the inverse problem: Using this approach they find another method for constructing a Lagrangian as a linear combination of the lhs of the equations of motion.

In Sec. II we will present the inverse problem of the variational calculus in classical mechanics and exhibit the Helmholtz conditions, which are a set of algebraic and differential equations for a certain matrix W which plays the role of an integrating factor for the inverse problem. The existence of more than one matrix W implies the nonuniqueness of a variational principle for such systems, that is, the existence of s-equivalent Lagrangians.

In Sec. III we will use the approach used by Hojman *et al.*[11–16] In accordance with the ideas of Hojman *et al.*[11–16] we will show that it is possible to obtain a compact formula for building a Lagrangian L which can be thought of as an intermediate recipe between that of Engels[17] and Hojman *et al.*[11]

In Sec. IV we will use the Henneaux[8] approach to show that a variational principle for a given set of second-order differential equations exists if and only if there exists a certain nonsingular $2n \times 2n$ matrix such that all its entries are constants of motion. This matrix will satisfy some properties that are shown to be equivalent to the Helmholtz conditions. Using this approach to the inverse problem we will provide an alternative proof of the trace theorem.

In Sec. V we will exhibit some examples of the inverse problem related to the ideas of this work and in Sec. VI we will present the conclusions.

II. THE INVERSE PROBLEM AND THE HELMHOLTZ CONDITIONS

We consider a set of particles whose configuration space is R^N, with the Cartesian coordinates $\{q^i\}$. In what follows we make use of integration theorems which are valid in R^N; we will not treat the case where the configuration space has a topology different from R^N. A trajectory of the system is

 0022-2488/89/092054-08$02.50

defined by the functions $q^i(t)$, where t is the time. The components of the velocity are $\dot{q}^i = dq^i/dt$. The set of trajectories S that describes the evolution of a system of particles will be considered to consist of all the solutions of a given set of second-order differential equations.

The inverse problem of the calculus of variations in classical mechanics consists[4,18] in determining whether the set S of solutions of a set of second-order differential equations

$$F_i(t,q,\dot{q},\ddot{q}) = 0 \tag{2.1}$$

is also the set of stationary curves of a variational problem based on the integral

$$I[q] = \int_{t_1}^{t_2} L(t,q,\dot{q})dt. \tag{2.2}$$

The functional $I[q]$ will be referred to as the action integral and $L(t,q,\dot{q})$ will be referred to as the Lagrangian.

In other words, the inverse problem of the variational calculus in classical mechanics consists in determining the existence of an action integral such that its set of stationary curves $q^i(t)$ coincides with the set of solutions of a given set of second-order differential equations such as (2.1).

In this work we consider only nonsingular, second-order systems of differential equations, that is, systems of differential equations (2.1) such that the determinant of the matrix

$$\left|\left|\frac{\partial F_i}{\partial \ddot{q}^j}\right|\right| \tag{2.3}$$

is different from zero in the whole set of points where the F_i are defined. We will not consider the inverse problem for singular systems in this work, although its study is of great importance (take, for example, the case of gauge theories). However, from the point of view of the inverse problem, singular systems involve problems that fall out of the scope of this work.

It is a well-known result that the set of curves that make $I[q]$ stationary is precisely the set of solutions of the Euler–Lagrange equations.

$$E_i L \equiv \frac{d}{dt}\frac{\partial L}{\partial \dot{q}^i} - \frac{\partial L}{\partial q^i} = 0. \tag{2.4}$$

Consequently, the inverse problem can be stated as seeking the existence of a Lagrangian L such that its Euler–Lagrange equations have the set S as solutions.

The Euler–Lagrange equations are also a set of second-order differential equations; however, they have a very particular form. The result is that any set of differential equations with this very particular form is also a set of Euler–Lagrange equations for some Lagrangian L. This result means that the inverse problem can also be thought of as seeking the existence of a set of second-order differential equations whose solutions are S and have this very particular form.

In a neighborhood of a point where the determinant (2.3) is different from zero the equations of motion (2.1) can be written equivalently as

$$\ddot{q}^i = f^i(t,q,\dot{q}). \tag{2.5}$$

Equations (2.1) also have the set S as solutions. The functions $f^i(t,q,\dot{q})$ will be called forces (in lieu of the more cumbersome, but more proper "forces divided by masses").

It is possible to prove[6] that any other set of nonsingular second-order differential equations of the form

$$W_{ij}(\ddot{q}^i - \tilde{f}^i(t,q,\dot{q})) = 0, \tag{2.6}$$

where $W_{ij}(t,q,\dot{q})$ is a nonsingular matrix called the mass matrix, will also have the set S as solutions if and only if $\tilde{f}^i = f^i$. The inverse problem for this kind of system is then reduced to finding a Lagrangian $L(t,q,\dot{q})$ such that

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}^i} - \frac{\partial L}{\partial q^i} \equiv W_{ij}(\ddot{q}^j - f^j) \tag{2.7}$$

for some nonsingular matrix $W_{ij}(t,q,\dot{q})$. If such a Lagrangian exists, then

$$W_{ij} = \frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j} \tag{2.8}$$

and

$$W_{ij}f^j = \frac{\partial L}{\partial q^i} - \frac{\partial^2 L}{\partial \dot{q}^i \partial q^j}\dot{q}^j - \frac{\partial^2 L}{\partial \dot{q}^i \partial t}. \tag{2.9}$$

This form of considering the inverse problem is the most convenient for nonsingular systems. The following question thus arises: Is it possible to write a set of conditions for the existence of a variational system using only the forces? The answer is affirmative: The set of conditions can be written as a set of equations for the matrix W_{ij} in terms of the forces f^i. These equations are known as the Helmholtz conditions and can be written in matrix notation:

$$W = W^T, \tag{2.10a}$$

$$\frac{\partial W_{ij}}{\partial \dot{q}^k} = \frac{\partial W_{ik}}{\partial \dot{q}^j}, \tag{2.10b}$$

$$\frac{\bar{d}}{dt}W = WF + F^T W, \tag{2.10c}$$

$$WA_0 = A_0^T W, \tag{2.10d}$$

where the superscript T indicates the transposed matrix and

$$\frac{\bar{d}}{dt} \equiv \frac{\partial}{\partial t} + \dot{q}^i\frac{\partial}{\partial q^i} + f^i\frac{\partial}{\partial \dot{q}^i}; \tag{2.11}$$

also,

$$F^i_j \equiv -\tfrac{1}{2}\frac{\partial f^i}{\partial \dot{q}^j}; \quad E^i_j \equiv -\frac{\partial f^i}{\partial q^j}, \quad A_0 \equiv E - F^2 - \frac{\bar{d}F}{dt}. \tag{2.12}$$

The conditions (2.10) will have a solution W_{ij} if and only if there exists a variational system such as (2.2) with the set S as its set of stationary curves. In fact, if the matrix W_{ij} satisfies the Helmholtz conditions, then a Lagrangian L can be constructed using the formula shown in Engels[17] such that its Euler–Lagrange equations are $W_{ij}(\ddot{q}^j - f^j) = 0$:

$$L = -q^i\int_0^1 E_i(t,\tau q, \tau\dot{q}, \tau\ddot{q})d\tau + \frac{d}{dt}\int_0^1\int_0^1 \tau q^i\dot{q}^j W_{ij}(t,\tau q, \tau\tau'\dot{q})d\tau'\,d\tau, \tag{2.13}$$

where

$$E_i(t,q,\dot{q},\ddot{q}) \equiv W_{ij}(\ddot{q}^j - f^j). \tag{2.14}$$

Note that the first integral in (2.13) involves the accelerations: the reason for adding the second integral is to cancel the accelerations; it does not have any influence in the equations of motion since it is only a total derivative.

Given a Lagrangian L_1 whose Euler–Lagrange equations have S as solutions, it is always possible to build another Lagrangian L_2 which generates the same set S of solutions:

$$L_2 = \alpha L_1 + \frac{d}{dt} g(t,q), \tag{2.15}$$

where α is a nonzero numerical constant. The Euler–Lagrange equations obtained from the new Lagrangian L_2 are exactly α times those obtained from L_1.

It can be said that a Lagrangian L is essentially unique if any other Lagrangian L' whose Euler–Lagrange equations $E_i L' = 0$, also having the set S as solutions, is related to L through a relation such as (2.15).

Two Lagrangians L and L' can be said to be s equivalent if their Euler–Lagrange equations generate the same set of solutions and if they are not related through a relation such as (2.15).

With the above definitions and the kind of systems considered in this work, an s-equivalent Lagrangian L' will be one such that

$$\frac{d}{dt}\frac{\partial L'}{\partial \dot{q}^i} - \frac{\partial L'}{\partial q^i} = W'_{ij}(\ddot{q}^j - f^j), \tag{2.16}$$

with $W_{ij} \neq \alpha W'_{ij}$ and $\det(W'_{ij}) \neq 0$.

The Helmholtz conditions in their form (2.10) can be thought of as a set of conditions for the existence of a certain matrix W_{ij} when a set of forces f^i is given. Assume that a nonsingular set of second-order differential equations such as (2.1) is given and we are required to know if there exists a variational problem such as (2.2) that has S as the set of stationary curves. First, one transforms the differential equations to the form (2.5) and sees if conditions (2.10) have a solution W_{ij}. If this is the case, using formula (2.13) one can easily build the Lagrangian: This Lagrangian will be uniquely determined up to the addition of a total derivative.

Note that if a certain matrix W_{ij} is a solution of the Helmholtz conditions (2.10), then the matrix $W'_{ij} = \alpha W_{ij}$ (α constant) is also a solution. This property defines an equivalence class of the solutions of the Helmholtz conditions and the Helmholtz conditions have two or more solutions if these solutions pertain to different equivalence classes. A solution is unique if there exists only one class of equivalent matrices W_{ij} which are solutions of (2.10). From each solution of the Helmholtz conditions it is possible to build a Lagrangian; therefore, each class of equivalent solutions of the Helmholtz conditions also defines a class of equivalent Lagrangians related by (2.15). In the same way, a Lagrangian is unique if there exists only one class of equivalent Lagrangians. Two Lagrangians pertaining to two different classes of equivalence will be said to be s equivalent.

It is important to mention that conditions (2.10) do not always have a solution, that is, given a set of forces f^i, it is not always possible to find a matrix W_{ij} that satisfies (2.10). If the Helmholtz conditions have a solution, in general it will be unique (in the above sense).[8]

The trace theorem says that the trace of any power of a certain matrix M must be a constant of the motion; this matrix is formed when two Lagrangians L and L' are s equivalent in the sense defined above.[6,9]

Trace theorem: If the Lagrangians L and L' are s equivalent, then the trace of any power of the matrix

$$M = W'W^{-1} \tag{2.17}$$

is a constant of the motion.

It is assumed that the mass matrix W is nonsingular. The inverse of W is written as

$$W^{-1} = (W^{ij}), \tag{2.18}$$

so that

$$W^{ik}W_{kj} = \delta^i_j. \tag{2.19}$$

The matrix W' is the mass matrix defined from the Lagrangian L'.

The trace theorem does not use all the information contained in the Helmholtz conditions (2.10).[19] Therefore, the existence of a matrix that satisfies the trace theorem does not guarantee the existence of s-equivalent Lagrangians.

III. FIRST-ORDER APPROACH

In this section we will present some definitions for a system of nonsingular second-order differential equations in order to exhibit the first-order approach to the inverse problem and to obtain an alternative formula for building a Lagrangian L. As mentioned in Sec. II, a set of nonsingular second-order differential equations is characterized by the equations of motion

$$\ddot{q}^i = f^i(t,q,\dot{q}), \quad i = 1,...,n. \tag{3.1}$$

Hojman *et al.*[11–16] use the definitions given below in the analysis they make of the inverse problem, using what they call the first-order approach. The first-order approach to the inverse problem consists in transforming the set (3.1) of n second-order differential equations into a set of $2n$ first-order differential equations.

To accomplish this aim, it is necessary to define the variables x^a, $a = 1,...,2n$ by

$$x^i = q^i, \quad x^{n+1} = \dot{q}^i. \tag{3.2}$$

A new set of forces $\tilde{f}^a$ is also defined by

$$\tilde{f}^i = \dot{q}^i, \quad \tilde{f}^{n+i} = f^i. \tag{3.3}$$

Actually, Hojman *et al.*[12–16] use an extended system x^μ and $\tilde{f}^\mu(x^\nu)$ ($\mu,\nu = 0,1,...,2n$) defined by

$$x^0 \equiv t, \quad \tilde{f}^0 \equiv 1. \tag{3.4}$$

Therefore, the system of differential equations

$$\dot{x}^\mu = \tilde{f}^\mu(x^\nu) \tag{3.5}$$

is equivalent to (3.1).

Consider now a set of $2n+1$ functionally independent variables c^μ:

$$c^\alpha = c^\alpha(x^\mu), \quad c^0 \equiv x^0 = t, \quad \alpha = 0,1,...,2n \tag{3.6}$$

such that

$$\det\left(\frac{\partial c^\alpha}{\partial x^\mu}\right) \neq 0 \tag{3.7}$$

and chosen in such a way that

$$\frac{\bar{d}}{dt}c^{\alpha}=0, \tag{3.8}$$

where

$$\frac{\bar{d}}{dt}\equiv\tilde{f}^{\mu}\frac{\partial}{\partial x^{\mu}}, \tag{3.9}$$

that is, the variables c^{α} are a set of $2n$ functionally independent constants of motion. Hojman[16] proves that the system of differential equations (3.5) can be written in terms of the coordinates c^{α} equivalently as

$$\dot{c}^{\alpha}=g^{\alpha}(c^{\beta}), \tag{3.10}$$

where

$$g^{\alpha}=\frac{\partial c^{\alpha}}{\partial x^{\mu}}\tilde{f}^{\mu}, \tag{3.11}$$

that is,

$$g^{0}=1, \quad g^{a}=0. \tag{3.12}$$

Using the first-order approach Hojman *et al.*[11] proved that if a given set of nonsingular second-order differential equations such as (3.1) can be represented by a Lagrangian $L(t,q,\dot{q})$, then there exists another Lagrangian $\bar{L}(t,q,\dot{q},\ddot{q})$ which differs from L by a total time derivative and which can be written as a linear combination of the lhs of the equations of motion (3.1). In other words, if there exists a Lagrangian L for the set of differential equations (3.1), then there exists a Lagrangian $\bar{L}$ such that

$$\bar{L}=\mu_i(t,q,\dot{q})(\ddot{q}^i-f^i)=L(t,q,\dot{q})+\frac{d}{dt}g(t,q,\dot{q}). \tag{3.13}$$

Thus the Euler–Lagrange equations of $\bar{L}$ and L coincide. The functions μ_i satisfy

$$\frac{\partial\mu_i}{\partial\dot{q}_j}=\frac{\partial\mu_j}{\partial\dot{q}^i}, \tag{3.14}$$

$$\frac{\bar{d}}{dt}\left(\frac{\bar{d}}{dt}\mu_i+\mu_k\frac{\partial f^k}{\partial\dot{q}^i}\right)-\mu_k\frac{\partial f^k}{\partial q^i}=0, \tag{3.15}$$

and

$$\det(W_{ij})\neq 0, \tag{3.16}$$

where

$$W_{ij}\equiv\frac{\partial}{\partial\dot{q}^j}\left(\frac{\bar{d}}{dt}\mu_i+\mu_k\frac{\partial f^k}{\partial\dot{q}^i}\right)+\frac{\partial\mu_j}{\partial q^i}. \tag{3.17}$$

Hojman *et al.*[11] also proved that if μ_i is such that it satisfies relations (3.14)–(3.16), then the matrix W_{ij}, defined by (3.17), satisfies the Helmholtz conditions (2.10): Thus these relations can also be thought of as another equivalent form of the Helmholtz conditions.

Relation (3.14) implies that the accelerations can be canceled from $\bar{L}$: Let L' be given by

$$L'=\bar{L}-\frac{d}{dt}h(t,q,\dot{q}). \tag{3.18}$$

Thus requiring L' to be independent of the accelerations, $\partial L'/\partial\ddot{q}^i=0$, implies that the function h is such that

$$\frac{\partial h}{\partial\dot{q}^i}=\mu_i. \tag{3.19}$$

Relation (3.14) is precisely the integrability condition of Eq. (3.19), so that

$$h=\int_0^1 q^k\mu_k(t,q,\tau\dot{q})\,dr. \tag{3.20}$$

Relation (3.20) can be used in (3.18) to obtain

$$L'=\mu_i(\ddot{q}^i-f^i)-\frac{d}{dt}\int_0^1\dot{q}^k\mu_k(t,q,\tau\dot{q})\,d\tau$$

$$=-\frac{\bar{d}}{dt}\int_0^1\dot{q}^k\mu_k(t,q,\tau\dot{q})\,d\tau. \tag{3.21}$$

Equation (3.21) is another formula for building a Lagrangian which can be thought of as a combination of that by Engels[17] (see Sec. II) and that of Hojman *et al.*[11] [(3.13)]. Note that in (3.21), as in that of Hojman *et al*,[11] one needs to first find the functions μ_i that satisfy (3.14)–(3.16). In general this is a hard problem which is completely equivalent to finding a solution W_{ij} for the Helmholtz conditions (2.10).

IV. ANOTHER FORM OF THE HELMHOLTZ CONDITIONS

In this section we will use the Henneaux[8] approach to the inverse problem to show the existence of an alternative form of the Helmholtz conditions.

Definitions (3.2) and (3.3) are also used by Henneaux[8] in his study of the inverse problem. Henneaux shows that the set of differential equations (3.1) can be described equivalently by a variational principle such as

$$I=\int_{t_1}^{t_2}L(t,q,\dot{q})\,dt \tag{4.1}$$

if and only if there exists a nonsingular two-form σ which satisfies

$$\sigma_{ab}=0, \quad \text{for } n<a\leqslant 2n, \quad n<b\leqslant 2n, \tag{4.2a}$$

$$d\sigma=0, \tag{4.2b}$$

$$(\partial_t+\mathcal{L}_{\tilde{f}})\sigma=0, \tag{4.2c}$$

where $\mathcal{L}_{\tilde{f}}$ is the Lie derivative along the vector $\tilde{f}$:

$$\mathcal{L}_{\tilde{f}}\sigma_{ab}=\tilde{f}^c\sigma_{ab,c}+\tilde{f}^c_{,b}\sigma_{ac}+\tilde{f}^c_{,a}\sigma_{cb} \tag{4.3}$$

and

$$\partial_t\equiv\frac{\partial}{\partial t}, \quad {}_{,a}\equiv\partial_a\equiv\frac{\partial}{\partial x^a}. \tag{4.4}$$

Relations (4.2) are completely equivalent to the Helmholtz conditions (2.10), as will be proved in the following theorem.

Theorem 4.1: The Helmholtz conditions (2.1) have a nonsingular solution W_{ij} if and only if there exists a nonsingular two-form which satisfies relations (4.2).

Proof: Assume first that a given nonsingular two-form σ satisfies relations (4.2). Then the matrix W defined by

$$W_{ij}\equiv\sigma_{n+ij}=-\sigma_{jn+i} \tag{4.5}$$

will also be nonsingular and will satisfy the Helmholtz conditions (2.10). Now relation (4.2c) can be written as

$$\frac{\bar{d}}{dt}\sigma_{ab}=-\tilde{f}^c_{,a}\sigma_{cb}-\tilde{f}^c_{,b}\sigma_{ac}, \tag{4.6}$$

where

$$\frac{\bar{d}}{dt} \equiv \partial_t + \tilde{f}^a \partial_a. \tag{4.7}$$

Relation (4.6), written for $a = n + i$ and $b = n + j$, is the Helmholtz condition (2.10a):

$$0 = -\sigma_{in+j} - \sigma_{n+ij} = W_{ji} - W_{ij}, \tag{4.8}$$

where (4.2a) is used to cancel the lhs.

Relation (4.6) also implies the Helmholtz condition (2.10c). Since

$$\frac{\bar{d}}{dt}\sigma_{n+ij} - \frac{\bar{d}}{dt}\sigma_{in+j} = -\frac{\partial f^k}{\partial \dot{q}^i}\sigma_{n+kj} + \frac{\partial f^k}{\partial \dot{q}^j}\sigma_{in+k}, \tag{4.9}$$

(2.10a) and the above definition for W cause this equation to read exactly as (2.10c).

Now, in order to prove (2.10d), it is necessary to consider

$$\frac{\bar{d}}{dt}\sigma_{n+ij} + \frac{\bar{d}}{dt}\sigma_{in+j} = -2\sigma_{ij} - \frac{\partial f^k}{\partial \dot{q}^i}\sigma_{n+kj} - \frac{\partial f^k}{\partial \dot{q}^j}\sigma_{in+k}. \tag{4.10}$$

However, the lhs of Eq. (4.10) is zero. Thus

$$\sigma_{ij} = \frac{1}{2}\left(\frac{\partial f^k}{\partial \dot{q}^j}W_{ki} - \frac{\partial f^k}{\partial \dot{q}^i}W_{kj}\right). \tag{4.11}$$

Therefore, when $a = i$ and $b = j$, relation (4.6) can be written as

$$\frac{1}{2}\frac{\bar{d}}{dt}\left(\frac{\partial f^k}{\partial \dot{q}^j}W_{ki} - \frac{\partial f^k}{\partial \dot{q}^i}W_{kj}\right) = \frac{\partial f^k}{\partial q^j}W_{ki} - \frac{\partial f^k}{\partial q^i}W_{kj}. \tag{4.12}$$

As in Sec. II (2.10d) is obtained using (2.10c) in relation (4.12). It is now necessary to prove the Helmholtz condition (2.10b). Relation (4.2b) implies that

$$\sigma_{n+in+j,k} + \sigma_{n+jk,n+i} + \sigma_{kn+i,n+j} = 0, \tag{4.13}$$

where the first term is zero and the last two terms are exactly (2.10b).

The proof of Theorem 4.1 in the opposite direction is as follows: Assume that there exists a nonsingular $n \times n$ matrix W which satisfies the Helmholtz conditions (2.10). Then, using Engels'[9,17] formula (2.13) it is possible to build a Lagrangian L such that

$$W_{ij} = \frac{\partial^2 L}{\partial \dot{q}^i \partial \dot{q}^j} \tag{4.14}$$

and

$$W_{ij}f^j = \frac{\partial L}{\partial q^i} - \frac{\partial^2 L}{\partial t \partial \dot{q}^i} - \frac{\partial^2 L}{\partial \dot{q}^i \partial q^j}\dot{q}^j. \tag{4.15}$$

Now define the time-dependent one-form A in the $2n$-dimensional space by

$$A_a \equiv \frac{\partial L}{\partial \dot{x}^a} = \left(\frac{\partial L}{\partial \dot{q}^i}, 0\right). \tag{4.16}$$

Relation (4.15) can be rewritten more geometrically in terms of the one-form A as

$$(\partial_t + \mathcal{L}_{\tilde{f}})A_a = \frac{\partial L}{\partial x^a}. \tag{4.17}$$

Using the one-form A it is easy to define a two-form which satisfies relations (4.2): Let σ be the exterior derivative of A, $\sigma = dA$, that is, by components, the two-form σ is

$$\sigma_{ab} \equiv \frac{\partial A_b}{\partial x^a} - \frac{\partial A_a}{\partial x^b}. \tag{4.18}$$

It is clear that the definition of σ implies (4.2a) and (4.2b). Also, σ is nonsingular since by hypothesis W is nonsingular and $\det(\sigma) = \det(W^2)$. Then it is only necessary to prove that (4.2c) is satisfied. Relation (4.2c) is obtained by taking the exterior derivative of (4.17):

$$(\partial_t + \mathcal{L}_{\tilde{f}})\left(\frac{\partial A_b}{\partial x^a} - \frac{\partial A_a}{\partial x^b}\right) = 0. \tag{4.19}$$

Therefore, it has been proved that the Helmholtz conditions (2.10) and (4.2) are completely equivalent. In what follows, we will refer either to (2.10) or (4.2) as the Helmholtz conditions and we will say that the Helmholtz conditions have a solution if there exists a two-form σ that satisfies (4.2) for a given vector $\tilde{f}$.

Henneaux[8] uses relations (4.2) to prove that in general the Helmholtz conditions do not have a solution, that is, a given set of differential equations such as (3.1) in general cannot be represented by a variational principle such as (4.1). Henneaux also proves that if the Helmholtz conditions have a solution σ then in general it is unique.

As will be shown below, the Helmholtz conditions (2.10) or (4.2) can be written in another equivalent form based on a certain matrix formed only with constants of motion. Assume that the Helmholtz conditions have a solution for a given set of nonsingular second-order differential equations such as (3.1), that is, there exists a two-form σ which satisfies relations (4.2). Define the nonsingular antisymmetric $2n \times 2n$ matrix Ω by

$$\Omega_{ab} \equiv \frac{\partial x^c}{\partial c^a}\frac{\partial x^d}{\partial c^b}\sigma_{cd}, \tag{4.20}$$

where the variables c^a are any set of $2n$ functionally independent constants of motion of the system (3.1). Thus the nonsingular matrix Ω satisfies the properties

$$\Omega_{ab}\frac{\partial c^a}{\partial x^{n+i}}\frac{\partial c^b}{\partial x^{n+j}} = 0, \tag{4.21a}$$

$$\frac{\partial \Omega_{ab}}{\partial c^c} + \frac{\partial \Omega_{bc}}{\partial c^a} + \frac{\partial \Omega_{ca}}{\partial c^b} = 0, \tag{4.21b}$$

$$\frac{\bar{d}}{dt}\Omega_{ab} = 0. \tag{4.21c}$$

It is clear that (4.21a) and (4.21b) are a direct consequence of (4.2a) and (4.2b). In order to prove that (4.2c) implies (4.21c) it is enough to see that

$$\frac{\bar{d}}{dt}\Omega_{ab} = \frac{\partial x^c}{\partial c^a}\frac{\partial x^d}{\partial c^b}(\partial_t + \mathcal{L}_{\tilde{f}})\sigma_{cd}. \tag{4.22}$$

It is also clear that the inverse of the above statement is true, that is, if a given nonsingular antisymmetric matrix Ω is such that it satisfies (4.21), then there exists a nonsingular two-form σ which satisfies (4.2). Therefore, a necessary and suf-

ficient condition for the Helmholtz conditions [say (2.10) or (4.2)] to have a solution is that for any set of $2n$ functionally independent constants of motion of system (3.1) there exists a nonsingular antisymmetric matrix Ω which satisfies relations (4.21).

Thus relations (4.21) can also be thought of as the Helmholtz conditions since it has been proved that (4.2) and (4.21) are completely equivalent. In what follows we will refer to (4.21) as the Helmholtz conditions. The interesting aspect regarding Ω is that all of its entries are constants of motion. This fact will be used below to show that the Helmholtz conditions (4.21) can be reduced to a single condition related to the existence of a certain set of constants of motion.

It is possible to use the form of the Helmholtz conditions (4.21) to find the functions μ_i defined in Sec. III and from there build a Lagrangian $L(t,q,\dot{q})$ using (3.21). Note that relation (4.21b) implies that

$$\Omega_{ab} = \frac{\partial \gamma_a}{\partial c^b} - \frac{\partial \gamma_b}{\partial c^a}, \tag{4.23}$$

where the functions γ_a depend only on the constants of motion c^a, that is, $\bar{d}\gamma_a/dt = 0$. Then the functions μ_i defined by

$$\mu_i \equiv \gamma_a \frac{\partial c^a}{\partial \dot{q}^i} \tag{4.24}$$

satisfy relations (3.14)–(3.16). In effect, (3.14) is satisfied since

$$\frac{\partial \mu_i}{\partial \dot{q}^j} - \frac{\partial \mu_j}{\partial \dot{q}^i} = \Omega_{ab} \frac{\partial c^a}{\partial \dot{q}^i} \frac{\partial c^b}{\partial \dot{q}^j}. \tag{4.25}$$

To prove that (3.15) and (3.16) are also satisfied, note that

$$\frac{\bar{d}}{dt}\mu_i + \mu_k \frac{\partial f^k}{\partial \dot{q}^i} = -\gamma_a \frac{\partial c^a}{\partial \dot{q}^i}. \tag{4.26}$$

Therefore, (3.15) and (3.16) are satisfied and (3.21) can be used to build a Lagrangian L.

The Helmholtz conditions (4.21) can also be used to prove the trace theorem in a very easy way. Two s-equivalent Lagrangians L and L' imply the existence of two two-forms σ and σ' such that

$$\sigma_{ab} = \frac{\partial c^c}{\partial x^a}\frac{\partial c^d}{\partial x^b}\Omega_{cd}, \quad \sigma'_{ab} = \frac{\partial c^c}{\partial x^a}\frac{\partial c^d}{\partial x^b}\Omega'_{cd}, \tag{4.27}$$

where $\sigma \neq \sigma'$ if and only if $\Omega \neq \Omega'$. The trace theorem can be proved as Henneaux[8] does, using the two-forms σ and σ', since

$$\text{trace}((\sigma'\sigma^{-1})^s) = 2\,\text{trace}(M^s), \tag{4.28}$$

where M is defined as in Sec. II as $M = W'W^{-1}$. The matrices σ^{-1} and Ω^{-1} are defined in such a way that

$$\sigma_{ab}(\sigma^{-1})^{bc} = \Omega_{ab}(\Omega^{-1})^{bc} = \delta^c_a. \tag{4.29}$$

Relations (4.27) for σ and σ' imply

$$\sigma'\sigma^{-1} = \Omega'\Omega^{-1}. \tag{4.30}$$

Therefore,

$$2\,\text{trace}\,(M^s) = \text{trace}\,((\Omega'\Omega^{-1})^s) \tag{4.31}$$

and the trace of any power of M is a constant of motion since the matrix $\Omega'\Omega^{-1}$ is also a constant of motion. Another point that is interesting regarding the Helmholtz conditions (4.21) is the following: The Darboux theorem[20] says that at least locally there always exists a set of coordinates (in this case $2n$ functionally independent constants of motion) such that the matrix Ω_{ab} can be written as

$$\Omega_{ab} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{4.32}$$

where 1 stands for the $n \times n$ identity matrix. Obviously, the matrix (4.32) obeys the Helmholtz conditions (4.21b) and (4.21c). Therefore, the Darboux theorem[20] implies that the Helmholtz conditions (4.21) have a solution if and only if there exists a set of $2n$ functionally independent constants of motion of the system (3.1) such that

$$\Omega_{ab} \frac{\partial c^a}{\partial x^{n+i}} \frac{\partial c^b}{\partial x^{n+j}} = 0, \tag{4.33}$$

where Ω_{ab} is given by (4.32).

That is, the Helmholtz conditions in any of their previous forms have been changed to a single relation over a set of $2n$ functionally independent constants of motion. Then to say that the Helmholtz conditions have a solution is equivalent to saying that there exists a set of $2n$ functionally independent constants of motion which satisfies (4.33).

Note that the existence of two or more sets of constants of motion that satisfy (4.33) does not guarantee the existence of s-equivalent Lagrangians since these sets can be related by a canonical transformation: These transformations are known to preserve the form of Ω_{ab}. What does guarantee the existence of s-equivalent Lagrangians is the existence of Ω and Ω' such that both satisfy (4.21) and such that $\Omega' \neq \alpha\Omega$.

As a final remark about the Helmholtz conditions (4.21), note that if there exist two nonsingular antisymmetric matrices Ω and Ω' that satisfy these conditions, then the matrix $\bar{\Omega}$ defined by

$$\bar{\Omega} \equiv \alpha\Omega + \beta\Omega' \tag{4.34}$$

also satisfies the Helmholtz conditions. Therefore, if there exist two nonsingular solutions, then there exist infinitely many nonsingular solutions. In other words, the existence of two s-equivalent Lagrangians implies the existence of infinitely many s-equivalent Lagrangians.

V. EXAMPLES

In this section we will present some examples to show how the previous ideas can be applied.

A. Example 1

The first example is that of one-dimensional systems:

$$\ddot{q} = f(t,q,\dot{q}). \tag{5.1}$$

When integrated one-dimensional systems have two constants of motion c^1 and c^2, the Helmholtz conditions (4.21) always have a solution. In effect, the matrix

$$\Omega = \begin{pmatrix} 0 & -\alpha(c^1,c^2) \\ \alpha(c^1,c^2) & 0 \end{pmatrix} \tag{5.2}$$

is the most general solution to the Helmholtz conditions (4.21). In particular, if $\alpha = 1$, then

$$\Omega = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{5.3}$$

is a solution for any set of two functionally independent constants of motion c^1 and c^2. To build a Lagrangian, take, for example,

$$\gamma_1 = 0, \quad \gamma_2 = c^1. \tag{5.4}$$

The function μ is then

$$\mu = c^1 \frac{\partial c^2}{\partial \dot{q}} \tag{5.5}$$

and a good Lagrangian for one-dimensional systems is

$$L = -\frac{\bar{d}}{dt} \int_0^1 \dot{q}\mu(t,q,\tau\dot{q})d\tau. \tag{5.6}$$

Since the function α can be chosen in an infinite number of ways, the solution (5.2) actually is an infinite family of solutions. Therefore, one-dimensional systems always have infinitely many s-equivalent Lagrangians.

B. Example 2

The system described by the equations of motion

$$\ddot{x} = \dot{y}, \quad \ddot{y} = y \tag{5.7}$$

cannot be described by a variational method.[4] In effect, consider now the ideas of this work. This system can be completely integrated. A set of four functionally independent constants of motion for this system is

$$\begin{aligned} &c^1 = \tfrac{1}{2}(y + \dot{y})e^{-t}, \quad c^2 = \tfrac{1}{2}(y - \dot{y})e^{t}, \\ &c^3 = \dot{x} - y, \quad c^4 = \dot{y} - x + t(\dot{x} - y). \end{aligned} \tag{5.8}$$

The Helmholtz condition (4.21a) reads in this case as

$$\Omega_{34} - (\Omega_{13} + t\Omega_{14})e^{-t} + (\Omega_{23} + t\Omega_{24})e^{t} = 0. \tag{5.9}$$

Relation (5.9) must be satisfied for all times t. Therefore, assuming that (4.21c) is satisfied, relation (5.9) implies

$$\Omega_{34} = \Omega_{13} = \Omega_{14} = \Omega_{23} = \Omega_{24} = 0. \tag{5.10}$$

Thus Ω is singular and no variational problem exists for this system.

C. Example 3

As a last example consider the system described by the differential equations

$$\ddot{x} = \dot{y}, \quad \ddot{y} = -\dot{x} + y. \tag{5.11}$$

A convenient Lagrangian for this system is

$$L = \tfrac{1}{2}(\dot{x}^2 + \dot{y}^2) + \tfrac{1}{2}(x\dot{y} - y\dot{x}) + \tfrac{1}{2}y^2. \tag{5.12}$$

Using the ideas of this work it is possible to show that the Helmholtz conditions (4.21) have a single solution. A convenient set of four functionally independent constants of motion of this system is

$$\begin{aligned} &c^1 = \dot{x} - y, \quad c^2 = y - \dot{y}t - \tfrac{1}{2}(\dot{x} - y)t^2, \\ &c^3 = x - \dot{x}t + \tfrac{1}{6}(\dot{x} - y)t^3 + \tfrac{1}{2}\dot{y}t^2, \quad c^4 = \tfrac{1}{3}(\dot{y} - (\dot{x} - y)t). \end{aligned} \tag{5.13}$$

The Helmholtz condition (4.21a) for these constants of motion reads as

$$\begin{aligned} 0 = {} & \tfrac{1}{3}\Omega_{14} - t\Omega_{12} + \tfrac{1}{2}t^2(\Omega_{13} - \Omega_{24}) \\ & - (\tfrac{1}{12}t^4 + t^2)\Omega_{23} + (\tfrac{2}{5}t^3 - \tfrac{1}{3}t)\Omega_{34}. \end{aligned} \tag{5.14}$$

Relation (5.14) must be satisfied for all times t. Therefore, assuming that (4.21c) is satisfied, the relation (5.14) implies

$$\Omega_{12} = \Omega_{14} = \Omega_{23} = \Omega_{34} = 0 \tag{5.15}$$

and

$$\Omega_{13} = \Omega_{24}. \tag{5.16}$$

Therefore, the most general solution to the Helmholtz conditions (4.21a) and (4.21c) is

$$\Omega = \beta(c^1, c^2, c^3 c^4) \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{5.17}$$

The matrix (5.17) must also satisfy the Helmholtz condition (4.21b). However, (4.21b) implies that the function β is such that

$$\frac{\partial \beta}{\partial c^1} = \frac{\partial \beta}{\partial c^2} = \frac{\partial \beta}{\partial c^3} = \frac{\partial \beta}{\partial c^4} = 0. \tag{5.18}$$

Therefore, the only solution to the Helmholtz conditions is obtained when β is a number.

It is interesting to notice that the set of constants of motion (5.13) satisfies relation (4.33), with Ω as in (4.32), that is, if we had first checked (4.33), then we would have known that this system has a solution. However, it was necessary to use (4.21) in order to know that this solution is unique.

The functions γ for this system can be chosen as

$$\gamma_1 = 0, \quad \gamma_2 = 0, \quad \gamma_3 = c^1, \quad \gamma_4 = c^2. \tag{5.19}$$

Then a solution for the functions μ_i is

$$\begin{aligned} &\mu_1 = (\tfrac{1}{6}t^3 - t)c^1 - \tfrac{1}{3}tc^2, \\ &\mu_2 = \tfrac{1}{2}t^2 c^1 + \tfrac{1}{3}c^2. \end{aligned} \tag{5.20}$$

The Lagrangian for this system can be built using (3.21).

VI. CONCLUSIONS

In this work we have shown an alternative approach to the inverse problem. This consists mainly in an alternative, more geometric form of the Helmholtz conditions. The conditions were written here as a set of conditions (4.21) of a certain matrix whose entries are constants of motion. An interesting aspect of these conditions is that if one considers the Darboux theorem,[20] then the Helmholtz conditions (4.21b), (4.21c) are obeyed automatically. The remaining condition (4.21a) means that the Helmholtz conditions have a solution if and only if there exists a set of functionally independent constants of motion which satisfy (4.33). This result implies a completely different form of the Helmholtz conditions (4.2). Physically, it is important since it shows that the constants of motion of systems that can be described by a variational principle have a very particular form.

The Helmholtz conditions in the form (4.21) are an alternative approach for the study of the inverse problem: The interesting aspect of this form of the Helmholtz condi-

tions is that it is possible to change them to a single relation for the existence of a set of $2n$ functionally independent constants of motion.

In Sec. V we exhibited some examples to show how the ideas of this work can be applied. However, as shown with the last example (Sec. V C), relation (4.33) is not of great help in determining if the solution Ω is unique. It is necessary to use all the information in the Helmholtz conditions (4.21) to know that the solution was unique.

ACKNOWLEDGMENT

I would like to thank Dr. S. Hojman for his encouragement and helpful suggestions.

[1]H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA, 1980).
[2]H. Helmholtz, "Über die physikalische Bedeutung des Princips der kleinsten Wirkung," Journal für die reine und angewandte Mathematik. Berlin. **100**, 137 (1887).
[3]G. Darboux, *Leçons sur la théorie générale des surfaces* (Gauthier-Villars, Paris, 1891).
[4]J. Douglas, "Solution of the inverse problem of the calculus of variations," Trans. Am. Math. Soc. **50**, 71 (1941).
[5]D. G. Currie and E. J. Saletan, "q-equivalent particle Hamiltonians. I. The classical one-dimensional case," J. Math. Phys. **7**, 967 (1966).
[6]S. Hojman and H. Harleston, "Equivalent Lagrangians: Multidimensional case," J. Math. Phys. **22**, 1414 (1981).
[7]M. Henneaux, "On the inverse problem of the calculus of variations," J. Phys. A **15**, L93 (1982).
[8]M. Henneaux, "Equation of motion, commutation relations and ambiguities in the Lagrangian formalism," Ann. Phys. **140**, 1 (1982).
[9]S. Hojman and L. C. Shepley, "Lagrangianos equivalentes," Rev. Mex. Fis. **28**, 149 (1982).
[10]W. Sarlet "The Helmholtz conditions revisited. A new approach to the inverse problem of Lagrangian dynamics," J. Phys. A **15**, 1503 (1982).
[11]R. Hojman, S. Hojman, and J. Sheinbaum, "Shortcut for constructing any Lagrangian from its equations of motion," Phys. Rev. D **28**, 1333 (1983).
[12]S. Hojman and L. F. Urrutia, "First order approach to the inverse problem of the calculus of variation," Internal report, Inst. de Cíencías Nucleares UNAM (1981).
[13]S. Hojman and L. F. Urrutia, "On the inverse problem of the calculus of variations," J. Math. Phys. **22**, 1896 (1981).
[14]S. Hojman and J. Gómez, "First-order equivalent Lagrangians and conservation laws," J. Math. Phys. **25**, 1776 (1983).
[15]S. Hojman, L. Nuñez, A. Patiño, and H. Rago, "Symmetries and conserved quantities in geodesic motion," J. Math. Phys. **27**, 281 (1986).
[16]S. Hojman, "Symmetries of Lagrangians and of their equations of motion," J. Phys. A **17**, 2399 (1984).
[17]E. Engels, "A method for the computation of a Lagrangian within the context of the inverse problem of Newtonian mechanics," Hadronic J. **1**, 465 (1978).
[18]R. M. Santilli, *Foundations of Theoretical Mechanics I* (Springer, New York, 1978).
[19]F. Pardo O. and L. C. Shepley, "Helmholtz conditions and the trace theorem in classical mechanics," Rev. Mex. Fis. **32**, 245 (1986).
[20]V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).

Canonoid transformations in generalized mechanics

Julio M. Teixeira, Luiz J. Negri, and Luimar C. De Oliveira
Departamento De Física, Universidade Federal da Paraíba, João Pessoa, (PB), Brazil, 58.000

(Received 23 June 1988; accepted for publication 8 February 1989)

The problem of obtaining two Hamiltonian functions linked by a canonoid transformation is solved in the realm of generalized mechanics. By pairing the canonically conjugated pairs of variable a technique that also appears to be useful in solving many other problems in generalized mechanics is introduced.

I. INTRODUCTION

In a recent publication in this journal we derived the necessary and sufficient conditions for a canonoid transformation with respect to a given Hamiltonian.[1] In the present paper we consider the same problem in the realm of generalized mechanics.

As is well known, in generalized mechanics the Lagrangian and Hamiltonian functions depend on time derivatives of generalized coordinates higher than first order. After the original work of Ostrogadsky,[2] Borneas[3] laid the foundations of the formalism and at present the theory is enriched by the contributions of many authors (Ref. 4 and the others cited here are a brief list of such contributions).

In this paper we solve the problem of obtaining two generalized Hamiltonian functions related by a canonoid transformation using a technique that reduces generalized Hamiltonian mechanics to classical (with only first-order time derivatives) mechanics. Our procedure is based on the "pairing of the canonically conjugated pairs" maintaining the dimensionality of the phase space: It shares some features with a proposal by Riahi.[5] Our technique is suggested by the great similarity between the Lagrangian and Hamiltonian formalisms of higher-order mechanics and that of classical theory. Indeed, both theories are analogs in all relevant aspects: the definition of the canonically conjugated variables, the definition of the Poisson brackets, the canonical transformation theory, the the Hamilton–Jacobi equation, etc. Besides, it must also be remembered that the formalism of generalized mechanics goes over into classical mechanics when $s = 1$, where s is the order of the highest temporal derivative.

Motivated by this similarity we unify, after defining an "ξ family" of coordinates, the treatment of classical and generalized Hamiltonian formalisms; our main aim of obtaining "canonoidically" conjugated generalized Hamiltonians is easily reached. Also, remembering the close relation between canonical and canonoid transformations (a transformation that is canonoid with respect to any Hamiltonian function is a canonical transformation), we go a step further in our procedure and discuss how one can obtain canonically conjugated generalized Hamiltonians. To explicitly show the main features of our approach we exhibit a model for which we derive some standard results and parallel them with those of the corresponding generalized case.

II. GENERALIZED AND CLASSICAL DESCRIPTIONS

Let $L(t,q_k,\dot{q}_k,\ldots,\overset{(s)}{q}_k)$, $k = 1,2,\ldots,r$ be a Lagrangian function for some generalized system with $r \geqslant 1$ degrees of freedom. The associated Hamiltonian function is written in terms of $t, q_k, \dot{q}_k,\ldots, \overset{(s-1)}{q}_k$ and $p_{k/1}, p_{k/2},\ldots,p_{k/s}$, where $\overset{(n)}{q}_k$ and $p_{k/n+1}$ are the canonically conjugated variables[6–8] (when $s = 1$, $\overset{(s-1)}{q}_k \to \overset{(0)}{q}_k$, which, as usual, is defined as q_k).

Alternatively, we can write $H = H(\xi_\nu,t)$, $\nu = 1,2,\ldots, 2rs$ after introducing the following compact ξ notation:

$$
\begin{aligned}
&\xi_1 = q_1, \quad \xi_2 = \dot{q}_1,\ldots, \quad \xi_s = \overset{(s-1)}{q}_k,\\
&\xi_{s+1} = q_2, \quad \xi_{s+2} = \dot{q}_2,\ldots, \quad \xi_{2s} = \overset{(s-1)}{q}_k\\
&\quad\vdots \qquad\qquad \vdots \quad,\ldots, \quad \vdots\\
&\xi_{(r-1)s+1} = q_r, \quad \xi_{(r-1)s+2} = \dot{q}_r,\ldots, \quad \xi_{rs} = \overset{(s-1)}{q}_k,\\
&\xi_{rs+1} = p_{1/1}, \quad \xi_{rs+2} = p_{1/2},\ldots, \quad \xi_{s(r+1)} = p_{1/s},\\
&\xi_{(r+1)s+1} = p_{2/1}, \quad \xi_{(r+1)s+2} = p_{2/2},\ldots, \quad \xi_{s(r+2)} = p_{2/s},\\
&\quad\vdots \qquad\qquad \vdots \quad,\ldots, \quad \vdots\\
&\xi_{(2r-1)s+1} = p_{r/1}, \quad \xi_{(2r-1)s+2} = p_{r/2},\ldots,\xi_{2rs} = p_{r/s}.
\end{aligned}
\tag{1}
$$

The ξ notation (1) has the property of pairing the canonically conjugated pairs of variables, thus allowing the use of a more tractable form of the equations; also, it does not alter the dimensionality of the original phase space.

Now, formally, the $H(\xi_\nu,t)$ function can be regarded as the Hamiltonian function for some ordinary (not higher-order dependent) system. The canonically conjugated variables are ξ_m and ξ_{m+rs} $(m = 1,2,\ldots,rs)$ and the Hamiltonian equations of motion are conveniently written as[9,10]

$$
\dot{\xi}_\alpha = \gamma_{\alpha\beta}\frac{\partial H}{\partial \xi_\beta}, \tag{2}
$$

where the indices $\alpha, \beta, \nu, \ldots$ range from 1 to $2rs$ and

$$
\|\gamma_{\alpha\beta}\| = \begin{vmatrix} 0_{rs} & 1_{rs} \\ -1_{rs} & 0_{rs} \end{vmatrix}, \tag{3}
$$

 0022-2488/89/092062-03$02.50

so that

$$\gamma_{\alpha\beta}\gamma_{\alpha\nu} = \delta_{\beta\nu}, \tag{4a}$$

$$\gamma_{\alpha\beta} + \gamma_{\beta\alpha} = 0. \tag{4b}$$

Adopting this point of view, all that is needed is just to use what we learned from ordinary mechanics. To illustrate the procedure let us consider some pertinent applications.

Letting R and S be any two dynamical variables for a given generalized system we first switch over to the ξ notation, writing $R(\xi_\nu,t)$, $S(\xi_\nu,t)$. The Poisson bracket of R and S is then given by[9]

$$[R,S] = \frac{\partial R}{\partial \xi_\mu}\gamma_{\mu\nu}\frac{\partial S}{\partial \xi_\nu} \tag{5}$$

and all related results follow, for example, the Poisson theorem: If $R(\xi_\nu,t)$, $S(\xi_\nu,t)$ are constants of motion [that is, $\dot{R}=0$, $\dot{S}=0$ after using Eq. (2)], so is $[R,S]$.

Consider now an invertible transformation from the set $(\overset{(n)}{q_k}, P_{k/n})$ to another set where the variables are $(\overset{(n)}{Q_k}, P_{k/n})$. In the ξ notation this transformation is the map $\xi_\nu \to \eta_\nu(\xi_\mu)$, where η_ν stands for the new set of variables. To see whether this is a canonical transformation it suffices to verify whether or not there exists a nonzero constant z such that for all dynamical variables R, S we have $[R,S]^\eta = z[r,S]^\xi$, where the superscripts specify the coordinates to be used.[11] Also, an invertible transformation that preserves the form of the Hamiltonian equations for some given Hamiltonian function H (but not for every!) is called canonoid with respect to H. Recently, we established the necessary and sufficient conditions for such a transformation to exist in the realm of ordinary mechanics.[1] There had been no such definition nor any systematic procedure for obtaining two Hamiltonians linked by a canonoid transformation in generalized mechanics, but we can now do both based on the procedure depicted above. Indeed, we need only retain the same definition of canonoid transformation and after writing $H(\xi_\nu,t)$ we just follow the systematic procedure given in Ref. 1 to arrive at a new Hamiltonian, say $K(\eta_\nu,t)$, canonoidically conjugated to $H(\xi_\nu,t)$. Also, if we prefer to write everything in the original generalized notation, it can easily be done from definitions (1). In Sec. III we present an example in which we demonstrate the above results.

III. EXAMPLE

We consider the Lagrangian for a classical spinning particle[12,13]:

$$L = (m/2)[\dot{X}^2 - (1/\omega^2)\ddot{X}^2], \tag{6}$$

where $X=(q_1,q_2,q_3)$. The Hamiltonian function is

$$H = \sum_{j=1}^{3}\left(p_{j/1}\dot{q}_j - \frac{m}{2}\dot{q}_j^2 - \frac{\omega^2}{2m}p_{j/2}^2\right), \tag{7}$$

where the momenta are

$$p_{j/1} = m\dot{q}_j + (m/\omega^2)\overset{(3)}{q}_j, \tag{8a}$$

$$p_{j/2} = (m/\omega^2)\ddot{q}_j. \tag{8b}$$

Thus the canonical equations of motion are (for $j=1,2,3$)

$$\dot{q}_j = \dot{q}_j, \tag{9a}$$

$$\ddot{q}_j = -(\omega^2/m)p_{j/2}, \tag{9b}$$

$$\dot{p}_{j/1} = 0, \tag{9c}$$

$$\dot{p}_{j/2} = m\dot{q}_j - p_{j/1}. \tag{9d}$$

Now, we switch over to the ξ notation. Using definitions (1) for this case we obtain

$$H(\xi_\mu,t) = \xi_2\xi_7 + \xi_4\xi_9 + \xi_6\xi_{11} - (m/2)(\xi_2^2 + \xi_4^2 + \xi_6^2) - (\omega^2/2m)(\xi_8^2 + \xi_{10}^2 + \xi_{12}^2) \tag{10}$$

and the corresponding Hamiltonian equations (2) are

$$\dot{\xi}_A = \xi_{A+1}, \quad \text{for } A \text{ odd}, \tag{11a}$$

$$\dot{\xi}_A = -(\omega^2/m)\xi_{A+6}, \quad \text{for } A \text{ even}, \tag{11b}$$

$$\dot{\xi}_a = 0, \quad \text{for } a \text{ odd}, \tag{11c}$$

$$\dot{\xi}_a = m\xi_{a-6} - \xi_{a-1}, \quad \text{for } a \text{ even}, \tag{11d}$$

where, also for future convenience, we adopt the convention of using the indices A, B, C,... ranging from 1 to 6 and a, b, c,... ranging from 7 to 12. It is very simple to verify that Eqs. (11a)–(11d) are the same as Eqs. (9a)–(9d) as a result of definition (1). Note, also, that the problem of finding integrals of motion for Lagrangians, including higher-order derivatives studied by Constantellos,[13] can easily be solved in the present approach: In fact, starting from Eqs. (11a)–(11d) it is very simple to obtain these constants. We present some of the constants, expressing them in the ξ and the generalized notations:

$$K_a \equiv \xi_a = \text{const}, \quad \text{for } a \text{ odd} \leftrightarrow K_j = m\dot{q}_j + (m/\omega^2)\overset{(3)}{q}_j, \quad j=1,2,3,$$

$$M_a \equiv \xi_{a+2}^2 + (1/\omega^2)(m\xi_{a-4} - \xi_{a+1})^2 = \text{const}$$

$$\text{for } a \text{ even} \leftrightarrow M_j = (m^2/\omega^4)\left[q_j^2 + (1/\omega^2)\overset{(3)}{q}_j^2\right], \quad j=1,2,3,$$

$$N_a \equiv (\cos\omega t/\omega)(\xi_a - m\xi_{a-5}) - \xi_{a-5} - \xi_{a+1}\sin\omega t = \text{const},$$

$$\text{for } a \text{ odd} \leftrightarrow N_j = (m/\omega^2)\left[(\cos\omega t/\omega)\overset{(3)}{q}_k + \ddot{q}_j\sin\omega t\right], \quad j=1,2,3.$$

The Poisson theorem can also be used to generate other constants from the above set. For example,

$$[N_7, M_6] \equiv J = (2m/\omega)[(m\xi_2 - \xi_7)(\sin\omega t/\omega) - \xi_8\cos\omega t]$$

is a constant of motion which corresponds to

$$J = -(2m^2/\omega^3)\left[\overset{(3)}{q}_k(\sin\omega t/\omega) - \ddot{q}_1\cos\omega t\right]$$

in the original notation.

Now, consider the time-dependent invertible transformation

$$Q_j = p_{j/1}, \quad \dot{Q}_j = p_{j/2},$$
$$P_{j/1} = (t/m)p_{j/1} - q_j, \quad P_{j/2} = (t/m)p_{j/2} - \dot{q}_j, \tag{12}$$

with $j = 1,2,3$. In the ξ notation we have

$$\eta_A = \xi_{A+6}, \quad \eta_a = (t/m)\xi_a - \xi_{a-6}. \tag{13}$$

Hence, with R and S being any two dynamical variables it is simple to verify that

$$[R,S]^{\eta} = [R,S]^{\xi},$$

thus proving that Eqs. (13) define a canonical transformation. The generating function of this transformation is

$$F(\xi_\mu,t) = \sum_A \xi_A \xi_{A+6} - \sum_a \xi_a^2 \tag{14}$$

and the new Hamiltonian $K(\eta_\nu,t)$ has the form

$$\begin{aligned} K(\eta_\mu,t) = {} & (t/m)[\eta_1\eta_2 + \eta_3\eta_4 + \eta_5\eta_6 + m(\eta_2\eta_8 + \eta_4\eta_{10} + \eta_6\eta_{12})] \\ & - (1/2m)(t^2+\omega^2+1)(\eta_2^2 + \eta_4^2 + \eta_6^2) - (\eta_1\eta_8 + \eta_3\eta_{10} + \eta_5\eta_{12}) \\ & - (m/2)(\eta_8^2 + \eta_{10}^2 + \eta_{12}^2) - (1/2m)(\eta_1^2 + \eta_3^2 + \eta_5^2). \end{aligned} \tag{15a}$$

For completeness we also write the Hamiltonian function in terms of the set $\left(\overset{(n)}{Q}_k, P_{k/n}\right)$:

$$\begin{aligned} K = {} & t/m[Q_1\dot{Q}_1 + Q_2\dot{Q}_2 + Q_3\dot{Q}_3 + m(\dot{Q}_1P_{1/2} + \dot{Q}_2P_{2/2} + Q_3P_{3/2})] \\ & - (1/2m)(t^2+\omega^2+1)(\dot{Q}_1^2 + \dot{Q}_2^2 + \dot{Q}_3^2) \\ & - (Q_1P_{1/2} + Q_2P_{2/2} + Q_3P_{3/2}) - (m/2)(P_{1/2}^2 + P_{2/2}^2 + P_{3/2}^2) - (1/2m)(Q_1^2 + Q_2^2 + Q_3^2). \end{aligned} \tag{15b}$$

Finally, let us consider the problem of obtaining a Hamiltonian function which is related to the Hamiltonian given in Eq. (7) by a canonoid transformation. As pointed out in Sec. II, what is needed is to apply the systematic procedure we have developed previously[1] to $H(\xi_\nu,t)$ from Eq. (10). We shall not go into the details of the calculations to be performed; instead, we refer the reader to Ref. 1 on this subject. Thus it is not too difficult to obtain the following canonoid transformation with respect to $H(\xi_\nu,t)$:

$$\eta_A = \xi_A,$$
$$\eta_a = \xi_a + \xi_{a-6}\xi_{a-5}, \quad \text{for } a \text{ odd},$$
$$\eta_a = \xi_a + \tfrac{1}{2}\xi_{a-7}^2, \quad \text{for } a \text{ even}.$$

The new Hamiltonian function is

$$\begin{aligned} K(\eta_\mu,t) = {} & (\omega^2/2m)(\eta_1^2\,\eta_8 + \eta_3^2\,\eta_{10} + \eta_5^2\,\eta_{12}) - (\omega^2/8m)(\eta_1^4 + \eta_3^4 + \eta_5^4) - (\eta_1\,\eta_2^2 + \eta_3\,\eta_4^2 + \eta_5\,\eta_6^2) \\ & + (\eta_2\eta_7 + \eta_4\eta_9 + \eta_6\eta_{11}) - (m/2)(\eta_2^2 + \eta_4^2 + \eta_6^2) - (\omega^2/2m)(\eta_8^2 + \eta_{10}^2 + \eta_{12}^2) \end{aligned} \tag{16a}$$

and in terms of the set $\left(\overset{(n)}{Q}_k, P_{k/n}\right)$ we have

$$\begin{aligned} K = {} & (\omega^2/2m)(Q_1^2\,P_{1/2} + Q_2^2\,P_{2/2} + Q_3^2\,P_{3/2}) - (\omega^2/8m)(Q_1^4 + Q_2^4 + Q_3^4) - (\dot{Q}_1^2\,Q_1 + \dot{Q}_2^2\,Q_2 + \dot{Q}_3^2\,Q_3) \\ & + (\dot{Q}_1P_{1/1} + \dot{Q}_2P_{2/1} + \dot{Q}_3P_{3/1}) - (m/2)(\dot{Q}_1^2 + \dot{Q}_2^2 + \dot{Q}_3^2) - (\omega^2/2m)(P_{1/2}^2 + P_{2/2}^2 + P_{3/2}^2). \end{aligned} \tag{16b}$$

The Hamiltonian function (16a) and the one given by Eq. (7) describe the same system, namely, a classical spinning particle.

[1] L. J. Negri, L. C. de Oliveira, and J. M. Teixeira, J. Math. Phys. **28**, 2369 (1987).
[2] M. Ostrogradsky, Mém. Acad. St.-Pét. **6**, 385 (1850).
[3] M. Borneas, Am. J. Phys. **7**, 265 (1959); Nuovo Cimento **16**, 806 (1960).
[4] K. Boehm, J. Reine Angew. Math. **121**, 124 (1900); C. F. Hayes and J. M. Jankowski, Nuovo Cimento B **58**, 494 (1968); C. F. Hayes, J. Math. Phys. **10**, 1555 (1969); M. Borneas, Am. J. Phys. **40**, 248 (1972); H. Tesser, J. Math. Phys. **13**, 796 (1972); C. Ryan, *ibid.* **13**, 283 (1972); D. Anderson, *ibid.* **14**, 934 (1973); J. R. Ellis, J. Phys. A **8**, 496 (1975); A. L. Vanderbauwhede, Hadron. J. **1**, 1177 (1978); G. Cognola, L. Vanzo, and S. Zerbini, Lett. Nuovo Cimento **38**, 533 (1983); J. R. Farias and N. L. Teixeira, J. Phys. A **16**, 1517 (1983); M. Sz. Kirkovitz, Acta Math. Hung. **43**, 341 (1984); J. R. Farias, Hadron. J. **8**, 227 (1985); L. J. Negri and E. G. da Silva, Phys. Rev. D **33**, 2227 (1986).
[5] R. Riahi, Am. J. Phys. **40**, 383 (1972).
[6] J. G. Koestler and J. A. Smith, Am. J. Phys. **33**, 44 (1965).
[7] L. M. C. de Souza and P. R. Rodrigues, J. Phys. A **2**, 304 (1969).
[8] J. G. Kruger and D. K. Callebaut, Am. J. Phys. **36**, 557 (1968).
[9] E. J. Saletan and A. H. Cromer, *Theoretical Mechanics* (Wiley, New York, 1971).
[10] E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974).
[11] D. Currie and E. J. Salentan, Nuovo Cimento B **9**, 143 (1972).
[12] F. Reine, Lett. Nuovo Cimento **1**, 807 (1971).
[13] G. C. Constantelos, Nuovo Cimento B **21**, 279 (1974).

Bäcklund transformations for the Caudrey–Dodd–Gibbon–Sawada–Kotera equation and its λ-modified equation

W. L. Chan and Yu-kun Zheng
Department of Mathematics, The Chinese University of Hong Kong, Science Center, Shatin, N. T., Hong Kong

(Received 4 October 1988; accepted for publication 3 May 1989)

A λ-modified equation for the Caudrey–Dodd–Gibbon–Sawada–Kotera equation is introduced. A new Bäcklund transformation for this equation is derived from the invariance property of the scattering problem for the Caudrey–Dodd–Gibbon–Sawada–Kotera equation under a Crum transformation [Q. J. Math. **6**, 121 (1955)]. This, in turn, gives rise to a new Bäcklund transformation for the Caudrey–Dodd–Gibbon–Sawada–Kotera equation.

I. INTRODUCTION

One of the interesting fifth-order integrable nonlinear evolution equations is the Caudrey–Dodd–Gibbon–Sawada–Kotera equation[1,2] (CDGSKE). It is not a member of the Lax hierarchy of the Korteweg-de Vries equation and has some distinct properties, as reported in Ref. 3. The aim of the present article is to construct a new Bäcklund transformation (BT) for the CDGSKE. Other BT's have been found earlier.[2,4] Our study is prompted by the work of Crum.[5] The BT developed here can be considered as an extension and improvement of that of Strampp and Briz.[6] The article is organized as follows. In Sec. II the results on the BT of Ref. 6 for the CDGSKE are summarized. Its shortcoming is presented. A λ-modified CDGSKE (λ-mCDGSKE) is introduced. In Sec. III, the general solution of the scattering problem of the CDGSKE is found under the assumption that one of its solutions is known. A BT is derived, in Sec. IV, for the λ-mCDGSKE. This, in turn, gives rise to a BT for the CDGSKE in Sec. V.

II. THE CDGSKE AND THE λ-mCDGSKE

In Ref. 6 Strampp and Briz had studied the following CDGSKE[2]:

$$u_t + u_{xxxxx} + 5(uu_{xxx} + u_x u_{xx} + u^2 u_x) = 0. \tag{2.1}$$

It is an integrability condition of the scattering problem[3]

$$\psi_{xxx} + u\psi_x = \lambda\psi, \tag{2.2}$$

$$\psi_t = 6\lambda u\psi + (u_{xx} - u^2)\psi_x + (9\lambda - 3u_x)\psi_{xx}, \tag{2.3}$$

For convenience, here we have replaced that "$6u$" in Ref. 6 by u. By using bilinear operators, they showed that the scattering problem (2.2) and (2.3) are invariant under the following transformations:

$$\psi \to \psi' = 1/\psi, \tag{2.4}$$

$$\lambda \to \lambda' = -\lambda, \tag{2.5}$$

$$u \to u' = u + 6v_x, \tag{2.6}$$

where

$$v = \psi_x/\psi. \tag{2.7}$$

This means that if the function u in (2.6) is a solution of the CDGSKE (2.1), then function u', defined in (2.6) and (2.7), is also a solution of Eq. (2.1). and the function ψ', defined in (2.4), is a solution of (2.2) and (2.3) with λ' in (2.5). Their results can be summarized as follows:

(SB1): A necessary and sufficient condition of integrability for the scattering problem (2.2) and (2.3) is that the function u satisfies the CDGSKE (2.1).

(SB2): The scattering problem (2.2) and (2.3) is invariant under the transformations (2.4)–(2.7).

(SB3): If u is a solution of Eq. (2.1), then the function u', defined in (2.6) and (2.7), is also a solution of (2.1), that is (2.6) and (2.7) is a BT for (2.1).

Generally, in constructing a BT for an evolution equation, we naturally expect that the BT can be used repeatedly in generating infinitely many solutions of that equation. Unfortunately, the BT (2.6) and (2.7) can only be used effectively once, since if we make use of (2.6) and (2.7) twice, we will have

$$\psi' \to \psi'' = 1/\psi' = \psi, \tag{2.8}$$

$$v \to v' = \psi'_x/\psi' = -\psi_x/\psi = -v, \tag{2.9}$$

$$u' \to u'' = u' + 6v'_x = u + 6v_x - 6v_x = u, \tag{2.10}$$

$$\lambda' \to \lambda'' = -\lambda' = \lambda. \tag{2.11}$$

We are back to the original solution.

In this paper, we propose a method to overcome the above deficiency. Note that the function v, defined in (2.7), is in fact a solution of the following evolution equation:

$$v_t + [v_{xxxx} - (5/v)(v_x v_{xxx} - \lambda v_{xx} + \lambda^2) + (5/v^2)(v_x^2 v_{xx} - \lambda v_x^2) + v(8v_x^2 - 8vv_{xx} - 5\lambda v) + v^5]_x = 0. \tag{2.12}$$

This equation can be obtained by the following procedures: (a) Divide Eq. (2.3) by ψ and then take the derivative with respect to x and use (2.7); (b) divide Eq. (2.2) by ψ and use (2.7); and (c) solve for u from (b) and substitute this expression of u into the equation obtained in (a).

We call Eq. (2.12) the λ-modified CDGSKE (λ-mCDGSKE). Thus we have the following

Proposition 1: If ψ is a solution of the scattering problem (2.2) and (2.3), then the function v, defined by (2.7), is a solution of Eq. (2.12).

Substituting (2.9) and (2.5) into (2.12), one finds that (2.12) is invariant under these transformations. Our task is to find a BT

$$(v,\lambda) \to (v', -\lambda) \tag{2.13}$$

 0022-2488/89/092065-04$02.50

for (2.12) that excludes the transformation (2.9). We will discuss this problem in the following sections.

III. THE GENERAL SOLUTION FOR THE SCATTERING PROBLEM OF THE CDGSKE

We now try to find the general solution for the scattering problem (2.2) and (2.3), when one of its solutions is known. From (SB2), we know that it possesses a particular solution (2.4), corresponding to u' in (2.6) and λ' in (2.5); we denote this solution by φ:

$$\varphi = 1/\psi. \tag{3.1}$$

Assume that φ^* is another particular solution of (2.2) and (2.3). Let

$$\varphi^* = \varphi \int_{x_0}^{x} Q\,dx. \tag{3.2}$$

We want to determine the unknown function Q. Substituting (3.2) and u' and λ' into (2.2) and (2.3), we find that Q satisfies the following system of partial differential equations:

$$Q_{xx} - 3vQ_x + [3(v^2 - v_x) + u']Q = 0, \tag{3.3a}$$

$$Q_t = [B'_x - 2(vC')_x]Q + (B' - 2vC' + C'_x)Q_x + C'Q_{xx}, \tag{3.3b}$$

where

$$B' = u'_{xx} - u'^2, \tag{3.4a}$$

$$C' = 9\lambda' - 3u'_x. \tag{3.4b}$$

Assume that Q has been solved from (3.3a) and (3.3b); then we have two particular solutions (3.1) and (3.2) for (2.2) and (2.3), corresponding to u' and λ'.

It is well known that, for a differential equation of order three

$$y''' + ay'' + by' + cy = 0, \tag{3.5}$$

if two of its independent solutions y_1 and y_2 are known; then it possesses a general solution of the following form[7]:

$$y = \left[C_1 - C_3\int_{x_0}^{x} y_2(y_1y_2' - y_1'y_2)^{-2}\exp\left(-\int_{x_0}^{x} a\,dx\right)\right]y_1 + \left[C_2 + C_3\int_{x_0}^{x} y_1(y_1y_2' - y_1'y_2)^{-2} \times \exp\left(-\int_{x_0}^{x} a\,dx\right)dx\right]y_2, \tag{3.6}$$

where C_1, C_2, and C_3 are some constants. Therefore, if φ and φ^* are two known solutions of scattering problem (2.2) and (2.3), then comparing (2.2) to (3.5) and using (3.6) and (2.2), will possess the following general solution:

$$\psi' = (C_1 - C_3Z)\varphi + (C_2 + C_3W)\varphi^*, \tag{3.7}$$

where

$$Z = \int_{x_0}^{x} \varphi^*\Delta^{-2}\,dx, \tag{3.8}$$

$$W = \int_{x_0}^{x} \varphi\Delta^{-2}\,dx, \tag{3.9}$$

$$\Delta = \varphi\varphi_x^* - \varphi_x\varphi^*, \tag{3.10}$$

and C_1, C_2, and C_3 are some arbitrary functions of t and λ'. We now try to determine these functions, such that ψ' in (3.7) also satisfies Eq. (2.3) with u and λ being replaced by u' and λ' in (2.6) and (2.5), respectively. Substituting (3.7) into (2.3), and with φ and φ^* being solutions of (2.3), we get

$$(C_{1t} - C_{3t}Z - C_3Z_t)\varphi + (C_{2t} + C_{3t}W + C_3W_t)\varphi^* = C_3C'\Delta^{-1}. \tag{3.11}$$

Further, by using (3.10) and (2.3), we have

$$\frac{\partial}{\partial t}(\varphi^*\Delta^{-2}) = \frac{\partial}{\partial x}\{[(B' + C'_x)\varphi^* - C'\varphi_x^*]\Delta^{-2}\}, \tag{3.12}$$

$$\frac{\partial}{\partial t}(\varphi\Delta^{-2}) = \frac{\partial}{\partial x}\{[(B' + C'_x)\varphi - C'\varphi_x]\Delta^{-2}\}, \tag{3.13}$$

where B' and C' are the two functions defined in (3.4a) and (3.4b). Thus, by (3.8) and (3.9), we obtain

$$Z_t = \int_{x_0}^{x} \frac{\partial}{\partial t}(\varphi^*\Delta^{-2})dx = \{[(B' + C'_x)\varphi^* - C'\varphi_x^*]\Delta^{-2}\}\Big|_{x_0}^{x}, \tag{3.14}$$

$$W_t = \int_{x_0}^{x} \frac{\partial}{\partial t}(\varphi\Delta^{-2})dx = \{[(B' + C'_x)\varphi - C'\varphi_x]\Delta^{-2}\}\Big|_{x_0}^{x}. \tag{3.15}$$

Then, substituting (3.14) and (3.15) into (3.11), it leads to

$$\{C_{1t} + [((B' + C'_x)\varphi^* - C'\varphi_x^*)\Delta^{-2}]_{x=x_0}C_3\}\varphi + \{C_{2t} - [((B' + C'_x)\varphi - C'\varphi_x)\Delta^{-2}]_{x=x_0}C_3\}\varphi^* + (W\varphi^* - Z\varphi)C_{3t} = 0. \tag{3.16}$$

Now we choose C_1, C_2, and C_3 to satisfy the following system of ordinary differential equations:

$$C_{1t} + \{[(B' + C'_x)\varphi^* - C'\varphi_x^*]\Delta^{-2}\}_{x=x_0}C_3 = 0, \tag{3.17}$$

$$C_{2t} - \{[(B' + C'_x)\varphi - C'\varphi_x]\Delta^{-2}\}_{x=x_0}C_3 = 0, \tag{3.18}$$

$$C_{3t} = 0; \tag{3.19}$$

then the function (3.7) will satisfy Eq. (2.3). Solving (3.17)–(3.19) gives

$$C_1(t,\lambda') = \alpha - \gamma\int_{t_0}^{t}\{[(B' + C'_x)\varphi^* - C'\varphi_x^*]\Delta^{-2}\}_{x=x_0}\,dt, \tag{3.20}$$

$$C_2(t,\lambda') = \beta + \gamma\int_{t_0}^{t}\{[(B' + C'_x)\varphi - C'\varphi_x]\Delta^{-2}\}_{x=x_0}\,dt, \tag{3.21}$$

$$C_3(t,\lambda') = \gamma, \tag{3.22}$$

where λ' corresponds to that contained in C', and α, β, and γ are some arbitrary constants. Note that the two relations (3.12) and (3.13) imply that the following two differential forms:

$$\varphi^*\Delta^{-2}\,dx + [(B' + C'_x)\varphi^* - C'\varphi_x^*]\Delta^{-2}\,dt, \tag{3.23}$$

$$\varphi\Delta^{-2}\,dx + [(B' + C'_x)\varphi - C'\varphi_x]\Delta^{-2}\,dt, \tag{3.24}$$

are exact differentials. Therefore there exist two functions

$P^*(x,t,\lambda')$ and $P(x,t,\lambda')$, such that

$$P^*(x,t,\lambda') = \int_{(x_0,t_0)}^{(x,t)} \varphi^*\Delta^{-2}\,dx + [(B' + C'_x)\varphi^* - C'\varphi^*_x]\Delta^{-2}\,dt, \tag{3.25}$$

$$P(x,t,\lambda') = \int_{(x_0,t_0)}^{(x,t)} \varphi\Delta^{-2}\,dx + [(B' + C'_x)\varphi - C'\varphi_x]\Delta^{-2}\,dt. \tag{3.26}$$

Substituting (3.20)–(3.22) into (3.7) and using (3.25) and (3.26), we finally get the general solution of (2.2) and (2.3), corresponding to u' and λ' as follows:

$$\psi'(x,t,\lambda') = [\alpha - \gamma P^*(x,t,\lambda')]\varphi + [\beta + \gamma P(x,t,\lambda')]\varphi^*. \tag{3.27}$$

We state the result of this section in the following.

Proposition 2: If φ and φ^* are two known solutions of the scattering problem (2.2) and (2.3) corresponding to u' and λ', then (2.2) and (2.3) possess a general solution (3.27).

IV. THE BT FOR THE λ-mCDGSKE

In this section we will use the result obtained above to derive a BT (2.13) for the λ-mCDGSKE (1,12).

First we rewrite the formula (3.27). Denote

$$Q^* = \int_{x_0}^{x} Q\,dx, \tag{4.1}$$

where Q is the function defined in (3.3a) and (3.3b). By (3.27), (3.1), (3.2), and (4.1), we have

$$\psi' = \psi^{-1}\{\alpha - \gamma P^*(x,t,\lambda') + [\beta + \gamma P(x,t,\lambda')]Q^*\}, \tag{4.2}$$

where [by (3.25), (3.26), (3.2), and (3.1)]

$$P^*(x,t,\lambda') = \int_{(x_0,t_0)}^{(x,t)} \psi^3 Q^{*-2}\{Q^*\,dx + [(B' + C'_x)Q^* + C'(Q^*v - Q^*_x)]dt\}, \tag{4.3}$$

$$P(x,t,\lambda') = \int_{(x_0,t_0)}^{(x,t)} \psi^3 Q^{*-2}\{dx + (B' + C'_x + C'v)dt\}. \tag{4.4}$$

Note that (4.2) is the transformation formula for ψ, which is a generalization of formula (2.4). Next we denote

$$v' = \psi'_x/\psi', \tag{4.5}$$

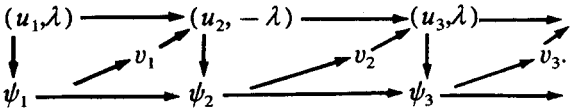
where ψ' is the function defined in (4.2). By the property of (2.4)–(2.7) and Proposition 2, the pair (v',λ') must satisfy Eq. (2.12). Substituting (4.2) into (4.5) gives

$$v' = v + \{\ln[\alpha - \gamma P^*(x,t,\lambda') + (\beta + \gamma P(x,t,\lambda'))Q^*]\}_x. \tag{4.6}$$

This is the transformation formula for function v, which is a generalization of formula (2.9) and is a BT for the λ-mCDGSKE (2.12).

V. THE BT FOR THE CDGSKE

The formulas (2.6), (4.2)–(4.4), and (4.6) now form a BT for the CDGSKE (2.1). In principle, by these formulas, starting from one known solution u_1 of (2.1), we can obtain three hierarchies of ψ's, v's, and u's. This procedure can be depicted by the following diagram:



(5.1)

We want to point out that the problem of integration of the CDGSKE (2.1) has not been reduced to quadrature. In iterating the BT, the solutions of (3.3a) and (3.4b) are required.

Example: Let us now give an example to show some partial success of the procedure of (5.1) and to see some of the difficulties that we will be facing.

Equation (2.1) obviously possesses the trivial solution

$$u_1 = 0. \tag{5.2}$$

Substituting (5.2) into (2.2) and (2.3) and writing $\lambda = \eta^3$, we get the following system of linear partial differential equation:

$$u_{xxx} = \eta^3\psi, \tag{5.3}$$

$$\psi_t = 9\eta^3\psi_{xx}. \tag{5.4}$$

This system of equations possesses the following solution:

$$\psi = \alpha\exp[-\tfrac{1}{2}\eta(x + 9\eta^4 t)]\sin(\sqrt{3}/2)\eta(x - 9\eta^4 t + \beta) + \gamma\exp\eta(x + 9\eta^4 t), \tag{5.5}$$

where α, β, and γ are some constants. For simplicity, we take $\gamma = 0$ in (5.5), so that

$$\psi_1 = \alpha\exp[-\tfrac{1}{2}\eta(x + 9\eta^4 t)]\sin(\sqrt{3}/2)\eta \times (x - 9\eta^4 t + \beta) \tag{5.6}$$

is the first generation of solutions of (5.3) and (5.4). Substituting (5.6) into (2.7), we get the first generation of solutions of Eq. (2.12),

$$v_1 = \tfrac{1}{2}\eta(\sqrt{3}\cot\omega\rho - 1), \tag{5.7a}$$

where

$$\omega = (\sqrt{3}/2)\eta, \quad \rho = x - 9\eta^4 t + \beta. \tag{5.7b}$$

Further, substituting (5.2) and (5.7a) into (2.6) gives the second generation of solutions of Eq. (2.1),

$$u_2 = -\tfrac{9}{2}\eta^2\csc^2\omega\rho. \tag{5.8}$$

To continue the procedure (5.1), we have to solve the function Q from (3.3a) and (3.4b). Substituting (5.7) and (5.8) into (3.3a) and (3.4b) we obtain the following system of differential equations:

$$Q_{xx} + \tfrac{3}{2}\eta(1 - \sqrt{3}\cot\omega\rho)Q_x - \tfrac{3}{2}\eta^2(1 + \sqrt{3}\cot\omega\rho)Q = 0, \tag{5.9}$$

$$Q_t = \tfrac{27}{4}\eta^5(9\sin^{-4}\omega\rho + 4\sqrt{3}\sin^{-3}\omega\rho\cos\omega\rho - 8\sin^{-2}\omega\rho)Q + \tfrac{9}{4}\eta^4(27\sin^{-4}\omega\rho - 6\sqrt{3}\sin^{-3}\omega\rho\cos\omega\rho - 30\sin^{-2}\omega\rho + 4\sqrt{3}\sin^{-1}\omega\rho\cos\omega\rho - 4)Q_x - \tfrac{9}{2}\eta^3(2 + 3\sqrt{3}\sin^{-3}\omega\rho\cos\omega\rho)Q_{xx}. \tag{5.10}$$

It is not obvious that they have an explicit solution. So the procedure could not go further. However, we have simplified the problem to that of basically solving a system of ordi-

nary differential equations. And, we believe that the subsequent third generation of solutions u_3 of (2.1) will be a new solution, that is, it does not coincide with u_2 in (5.8).

[1]K. Sawada and T. Kotera, Prog. Theor. Phys. **51**, 1355 (1974).
[2]R. K. Dodd and J. D. Gibbon, Proc. R. Soc. London Ser. **358**, 287 (1977).
[3]R. N. Aiyer, B. Fuchssteiner, and W. Oevel, J. Phys. A: Math. Gen. **19**, 3755 (1986).
[4]J. Satsuma and D. J. Kaup, J. Phys. Soc. Jpn. **43**, 692 (1977).
[5]M. M. Crum, Q. J. Math. **6**, 121 (1955).
[6]W.Strampp and K. H. Briz, Prog. Theor. Phys. **70**, 85 (1983).
[7]F. Brauer and J. A. Nohel, *Ordinary Differential Equations* (Benjamin, New York, 1967), p. 101.

Fokker–Wheeler–Feynman interactions without integrals

Paul Stephas
Department of Physics, Eastern Oregon State College, La Grande, Oregon 97850

(Received 4 April 1989; accepted for publication 10 May 1989)

The conserved quantities associated with Fokker–Wheeler–Feynman interactions between two particles are usually presented in terms of arbitrary times t_1 and t_2 for particles one and two, respectively; these conserved quantities involve integrals over the world lines of the particles. These integrals are evaluated so as to yield integral-free conserved quantities, with a resulting shift in focus to one particle at arbitrary time t_1 and the second particle at the t_1-related times of t_2 (retarded) and t_2 (advanced).

I. INTRODUCTION

Solutions for the relativistic two-body problem involving Fokker–Wheeler–Feynman interactions are not plentiful. These action-at-a-distance interactions are time symmetric, using half the sum of the retarded plus advanced potentials. Starting with either a least-action principle[1] or the equations of motion,[2,3] one can obtain conserved quantities in terms of arbitrary times t_1 for particle one and t_2 for the second particle; these quantities contain double integrals over the world lines of the particles. For example, the conserved four-momentum $\mathbf{P}$ for a massless vector interaction without radiation reaction between two point particles can be written as

$$\mathbf{P} = (\mathbf{p}_1 + q_1\mathbf{A}_1)_{\tau_1} + (\mathbf{p}_2 + q_2\mathbf{A}_2)_{\tau_2} + \mathbf{P}_I, \tag{1.1}$$

where $\mathbf{p}_j = m_j\mathbf{U}_j$ is the momentum of the jth particle with mass m_j and charge q_j moving with four-velocity $\mathbf{U}_j$ through a vector potential $\mathbf{A}_j$ due to the other particle, and $\mathbf{P}_I$ is the "interaction momentum" or "momentum in transit"[4] given by

$$\mathbf{P}_I = 2kq_1q_2c^{-3}\left(\int_{\tau_2}^{\infty} d\tau_2' \int_{-\infty}^{\tau_1} d\tau_1' - \int_{-\infty}^{\tau_2} d\tau_2' \int_{\tau_1}^{\infty} d\tau_1'\right)\mathbf{U}_1\cdot\mathbf{U}_2\mathbf{S}\delta'(\xi^2). \tag{1.2}$$

Here, τ_j is the proper time and $\mathbf{S}_j = (\vec{r}_j, ict_j)$ is the space-time position of the jth particle, $\mathbf{S} = \mathbf{S}_1 - \mathbf{S}_2$, $\xi^2 = (t_1 - t_2)^2 - R^2/c^2, \vec{R} = \vec{r}_1 - \vec{r}_2, R = |\vec{R}|, \delta'(\xi^2)$ is the derivative of the Dirac delta function with respect to its argument and $k = \mu_0/4\pi$ in SI units. Similar interaction momenta arise from scalar[5] and linear[6] potentials. The delta function and its derivative in the interaction terms, such as in Eq. (1.2), reduces the semi-infinite double integrals into integrals over finite segments of the world lines of the two particles, as illustrated in Fig. 1(a).

Alternatively, by integrating Eq. (1.2) first over the world line of particle one, the resulting interaction momentum contains integrals only over segments of the world line for particle two[2]:

$$\begin{aligned}\mathbf{P}_I = \tfrac{1}{2}ckq_1q_2\Bigg\{&\left.\frac{(\vec{R}, -iR)(1-\vec{\beta}_1\cdot\vec{\beta}_2)}{(R+\vec{R}\cdot\vec{\beta}_1)(R+\vec{R}\cdot\vec{\beta}_2)}\right|_{t_1,t_2^+} - \left.\frac{(\vec{R},iR)(1-\vec{\beta}_1\cdot\vec{\beta}_2)}{(R-\vec{R}\cdot\vec{\beta}_1)(R-\vec{R}\cdot\vec{\beta}_2)}\right|_{t_1,t_2^-}\\ &+\int_{t_2}^{t_2^+} dt_2'\left[\frac{1}{R+\vec{R}\cdot\vec{\beta}_1}\frac{d}{dt_1'}\frac{\mathbf{S}(1-\vec{\beta}_1\cdot\vec{\beta}_2)}{c(t_1'-t_2')-\vec{R}\cdot\vec{\beta}_1}\right]_{t_1'=t_2'-R/c}\\ &-\int_{t_2^-}^{t_2} dt_2'\left[\frac{1}{R-\vec{R}\cdot\vec{\beta}_1}\frac{d}{dt_1'}\frac{\mathbf{S}(1-\vec{\beta}_1\cdot\vec{\beta}_2)}{c(t_1'-t_2')-\vec{R}\cdot\vec{\beta}_1}\right]_{t_1'=t_2'+R/c}\Bigg\},\end{aligned} \tag{1.3}$$

where $t_2^{\pm} = t_1 \pm R/c$ and $\vec{\beta}_j = d\vec{r}_j/c\,dt_j$. Figure 1(b) illustrates the world-line segments which now enter into the calculation of a conserved quantity; although the advanced and retarded times $t_2^{\pm}$ are related to t_1, the time t_2 can be chosen independently of t_1. The residual integrals are still difficult to evaluate since, in principle, one must know the trajectories as a function of time; however, they have been evaluated for the vector interaction[7] as well as for the scalar[5] and linear[6] interactions for uniform concentric circular motions.

The integrals in Eq. (1.3) can be eliminated by using unphysical time-asymmetric interactions[8] where the two particles only interact for such times that particle one is at time t_1 and particle two is at time $t_2 = t_2^- = t_1 - R/c$, as illustrated in Fig. 1(c). Then only the second term in Eq. (1.3) is nonzero[2] and the resulting $\mathbf{P}_I$ must be multiplied by two. Time-asymmetric interactions led to solutions not only for uniform circular motions[9] but also for one-dimensional motions with vector[10] and Lorentz scalar[11] potentials.

In this paper I present the results of evaluating the integrals appearing in all time-symmetric interaction terms, such as in Eq. (1.3), so as to yield integral-free conserved quantities associated with the Lorentz group of transformations. This evaluation shifts the focus from arbitrary times t_1 and t_2 for the two particles, as illustrated in Fig. 1(b), to one particle at time t_1 and the second particle at the related times t_2^- and t_2^+, as illustrated by Fig. 1(d).

II. CONSERVED QUANTITIES

The integrands in all the interaction terms, such as in Eq. (1.3), were cast into exact differentials by using the

 0022-2488/89/092069-03$02.50

equations of motion, advanced/retarded time relations such as $ct_2^{\pm} = ct_1 \pm R$ and their derivatives, and the derivatives of other functions such as $d(R \pm \vec{R}\cdot\vec{\beta}_1)^{-1}/dt_2$. Lengthy and tedious calculations led to the following conserved energy W, momentum $\vec{P}$, angular momentum $\vec{L}$, and Lorentz momentum $\vec{L}_t$:

$$W = (w_1 + q_1\Phi_1)_{t_1} + \tfrac{1}{2}(w_2^+ + q_2\Phi_2^+)_{t_1,t_2^+} + \tfrac{1}{2}(w_2^- + q_2\Phi_2^-)_{t_1,t_2^-}$$
$$- \tfrac{1}{2}c^2kq_1q_2\left[\frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)R}{(R+\vec{R}\cdot\vec{\beta}_1)(R+\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^+} + \frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)R}{(R-\vec{R}\cdot\vec{\beta}_1)(R-\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^-}\right], \tag{2.1}$$

$$\vec{P} = (\vec{p}_1 + q_1\vec{A}_1)_{t_1} + \tfrac{1}{2}(\vec{p}_2^+ + q_2\vec{A}_2^+)_{t_1,t_2^+} + \tfrac{1}{2}(\vec{p}_2^- + q_2\vec{A}_2^-)_{t_1,t_2^-}$$
$$+ \tfrac{1}{2}ckq_1q_2\left[\frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)\vec{R}}{(R+\vec{R}\cdot\vec{\beta}_1)(R+\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^+} - \frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)\vec{R}}{(R-\vec{R}\cdot\vec{\beta}_1)(R-\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^-}\right], \tag{2.2}$$

which can be combined into

$$W - \vec{P}\cdot\vec{\beta}_1 c = (w_1 - c\vec{\beta}_1\cdot\vec{p}_1)_{t_1} + \tfrac{1}{2}[(w_2^+ + q_2\Phi_2^+) - c\vec{\beta}_1\cdot(\vec{p}_2^+ + q_2\vec{A}_2^+)]_{t_1,t_2^+}$$
$$+ \tfrac{1}{2}[(w_2^- + q_2\Phi_2^-) - c\vec{\beta}_1\cdot(\vec{p}_2^- + q_2\vec{A}_2^-)]_{t_1,t_2^-}, \tag{2.3}$$

$$\vec{L} = [\vec{r}_1\times(\vec{p}_1 + q_1\vec{A}_1)]_{t_1} + \tfrac{1}{2}[\vec{r}_2\times(\vec{p}_2^+ + q_2\vec{A}_2^+)]_{t_1,t_2^+} + \tfrac{1}{2}[\vec{r}_2\times(\vec{p}_2^- + q_2\vec{A}_2^-)]_{t_1,t_2^-}$$
$$+ \tfrac{1}{2}ckq_1q_2\left[\frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)\vec{r}_2\times\vec{r}_1}{(R+\vec{R}\cdot\vec{\beta}_1)(R+\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^+} - \frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)\vec{r}_2\times\vec{r}_1}{(R-\vec{R}\cdot\vec{\beta}_1)(R-\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^-}\right], \tag{2.4}$$
$$= \vec{r}_1\times\vec{P} - \tfrac{1}{2}[\vec{R}\times(\vec{p}_2^+ + q_2\vec{A}_2^+)]_{t_1,t_2^+} - \tfrac{1}{2}[\vec{R}\times(\vec{p}_2^- + q_2\vec{A}_2^-)]_{t_1,t_2^-}, \tag{2.5}$$

$$\vec{L}_t = [(\vec{p}_1 + q_1\vec{A}_1)ct_1 - \vec{r}_1(w_1 + q_1\Phi_1)/c]_{t_1} + \tfrac{1}{2}[(\vec{p}_2^+ + q_2\vec{A}_2^+)ct_2^+ - \vec{r}_2(w_2^+ + q_2\Phi_2^+)/c]_{t_1,t_2^+}$$
$$+ \tfrac{1}{2}[(\vec{p}_2^- + q_2\vec{A}_2^-)ct_2^- - \vec{r}_2(w_2^- + q_2\Phi_2^-)/c]_{t_1,t_2^-}$$
$$+ \tfrac{1}{2}ckq_1q_2\left[\frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)(\vec{r}_1R + \vec{R}ct_1)}{(R+\vec{R}\cdot\vec{\beta}_1)(R+\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^+} + \frac{(1-\vec{\beta}_1\cdot\vec{\beta}_2)(\vec{r}_1R - \vec{R}ct_1)}{(R-\vec{R}\cdot\vec{\beta}_1)(R-\vec{R}\cdot\vec{\beta}_2)}\bigg|_{t_1,t_2^-}\right], \tag{2.6}$$
$$= \vec{P}ct_1 - \vec{r}_1W/c + \tfrac{1}{2}[(w_2^+ + q_2\Phi_2^+)\vec{R}/c + (\vec{p}_2^+ + q_2\vec{A}_2^+)R]_{t_1,t_2^+}$$
$$+ \tfrac{1}{2}[(w_2^- + q_2\Phi_2^-)\vec{R}/c - (\vec{p}_2^- + q_2\vec{A}_2^-)R]_{t_1,t_2^-}. \tag{2.7}$$

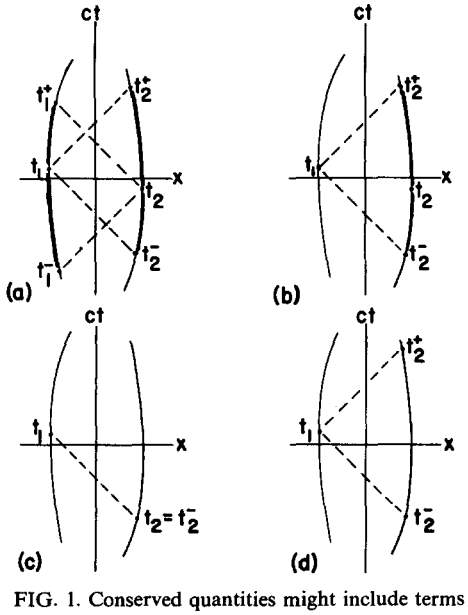


FIG. 1. Conserved quantities might include terms evaluated at t_1, t_2, $t_1^- = t_2 - R/c$, $t_1^+ = t_2 + R/c$, $t_2^- = t_1 - R/c$, $t_2^+ = t_1 + R/c$ plus integrals over sections of the world-lines indicated by heavy lines: (a) time-symmetric with integrals of the form in Eq. (1.2) with both t_1 and t_2 arbitrary; (b) time-symmetric with integrals of the form in Eq. (1.3) with both t_1 and t_2 arbitrary; (c) time-asymmetric with no integrals and $t_2 = t_2^-$ functionally related to t_1; (d) time-symmetric with no integrals and with both t_2^- and t_2^+ functionally related to t_1.

In the above equations the four-momenta are $\mathbf{p}_1 = (\vec{p}_1, iw_1/c)_{t_1}$ and $\mathbf{p}_2 = \tfrac{1}{2}(\vec{p}_2^-, iw_2^-/c) + \tfrac{1}{2}(\vec{p}_2^+, iw_2^+/c)$, and the Liénard–Wiechert vector potential $\mathbf{A}_j = (\vec{A}_j, i\Phi_j/c)$ is composed of retarded and advanced components, e.g., $\mathbf{A}_1(\vec{r}_1, t_1) = \tfrac{1}{2}(\mathbf{A}_1^+ + \mathbf{A}_1^-)$.

It is much easier and somewhat less tedious to verify that Eqs. (2.1)–(2.7) are conserved by differentiating them with respect to t_1 or t_2^- or t_2^+ and show that these derivatives are zero by virtue of the equations of motion, e.g., $dW/dt_1 = dW/dt_2^- = dW/dt_2^+ = 0$.

One notes that all these conserved quantities follow the same pattern. For example, the conserved energy W is composed of the canonical energy for particle one at observation time t_1, plus half the canonical energy for particle two at the retarded time t_2^- and at the advanced time t_2^+, plus an interaction potential energy again evaluated for (t_1, t_2^-) and (t_1, t_2^+). The integrals in Eq. (1.3) cancel the canonical energy for particle two at the arbitrary observation time t_2 of Eq. (1.1) and instead substitute half the values at the t_1-related times of t_2^- and t_2^+; compare Fig. 1(b) with Fig. 1(d).

The remaining conserved quantities associated with the full conformal group of transformations follow the same pattern, but consist of many terms. They can be combined with other conserved quantities, such as was done with $\vec{L}$ and $\vec{L}_t$ in Eqs. (2.5) and (2.7), to give the following compact forms

for the timelike component K_t and the spacelike component $\vec{K}$ of the conformal vector, and for the dilation scalar D:

$$D = \vec{r}_1 \cdot \vec{P} - t_1 W + m_1 c^2 \tau_1 + \tfrac{1}{2}\left[m_2 c^2 \tau_2^+ - w_2^+ (R + \vec{R}\cdot\vec{\beta}_2)/c\right]_{t_1, t_2^+} + \tfrac{1}{2}\left[m_2 c^2 \tau_2^- + w_2^- (R - \vec{R}\cdot\vec{\beta}_2)/c\right]_{t_1, t_2^-}, \tag{2.8}$$

$$K_t = (r_1^2 - c^2 t_1^2) W/c + 2\vec{r}_1 \cdot \vec{L}_t + 2 m_1 c^2 \int_{-\infty}^{\tau_1} d\tau_1' c t_1' + \left[m_2 c^2 \int_{-\infty}^{\tau_2^+} d\tau_2' c t_2' - t_2^+ w_2^+ (R + \vec{R}\cdot\vec{\beta}_2)\right]_{t_1, t_2^+} + \left[m_2 c^2 \int_{-\infty}^{\tau_2^-} d\tau_2' c t_2' + t_2^- w_2^- (R - \vec{R}\cdot\vec{\beta}_2)\right]_{t_1, t_2^-}, \tag{2.9}$$

$$\vec{K} = (r_1^2 - c^2 t_1^2)\vec{P} + 2\vec{r}_1 \times \vec{L} + 2 c t_1 \vec{L}_t + 2 m_1 c^2 \int_{-\infty}^{\tau_1} d\tau_1' \vec{r}_1 + \left[m_2 c^2 \int_{-\infty}^{\tau_2^+} d\tau_2' \vec{r}_2 - \vec{r}_2 w_2^+ (R + \vec{R}\cdot\vec{\beta}_2)/c\right]_{t_1, t_2^+} + \left[m_2 c^2 \int_{-\infty}^{\tau_2^-} d\tau_2' \vec{r}_2 + \vec{r}_2 w_2^- (R - \vec{R}\cdot\vec{\beta}_2)/c\right]_{t_1, t_2^-}. \tag{2.10}$$

III. DISCUSSION

The Schild[7] results for two particles moving in concentric circles were recovered using Eqs. (2.1), (2.2), (2.5), and (2.7); there was no advantage in using these equations over those containing integrals over the world lines since the trajectories $\vec{r}_1(t_1)$ and $\vec{r}_2(t_2)$ are predetermined for uniform circular motion. However, the possibility now exists that the more complicated motions related to the other conic sections can be computed by using these conserved quantities without integrals.

In this development of the Fokker–Wheeler–Feynman formalism, we have converted a manifestly nonlocal interaction between two particles at t_1 and t_2 which includes integrals, in general, over spacelike world lines, into a boundary value problem for which the positions and velocities for one particle at t_1 and the other particle located on the light cones of particle one determines the motion.[12] A well-defined boundary value problem is not restricted to the specification of initial conditions at some time t_1, but can also include the specification of initial conditions at times which are functionally related to t_1, such as $t_2^{\pm} = t_1 \pm R/c$.

[1] A. D. Fokker, Z. Phys. **58**, 386 (1929); J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. **21**, 425 (1949).
[2] P. Stephas, Am. J. Phys. **46**, 360 (1978).
[3] P. Stephas and H. C. von Baeyer, Phys. Rev. D **20**, 3155 (1979).
[4] H. Van Dam and E. P. Wigner, Phys. Rev. **142**, 838 (1966).
[5] C. M. Andersen and H. C. von Baeyer, Ann. Phys. (NY) **60**, 67 (1970).
[6] J. Weiss, J. Math. Phys. **27**, 1015, 1023 (1986).
[7] A. Schild, Phys. Rev. **131**, 2762 (1963).
[8] A. D. Fokker, Physica **9**, 33 (1929).
[9] B. Bruhns, Phys. Rev. D **8**, 2370 (1973).
[10] R. A. Rudd and R. N. Hill, J. Math. Phys. **11**, 2704 (1970).
[11] P. Stephas, Phys. Rev. D **31**, 319 (1985).
[12] P. Havas, *Causality and Physical Theories*, edited by W. B. Rolnick (American Institute of Physics, New York, 1974), p. 44.

Extended wave solutions in an integrable chiral model in (2+1) dimensions

Robert Leese
Centre for Particle Theory, University of Durham, Durham DH1 3LE, England

(Received 7 December 1988; accepted for publication 12 April 1989)

There is a modification of the SU(2) chiral model, which is integrable in (2 + 1) dimensions [J. Math. Phys. **29**, 386 (1988)]. In addition to localized lumps, it admits extended wave solutions, which move at constant velocity. The interaction of two waves causes each to experience a phase shift. In the interaction between a wave and a lump, the wave suffers no phase shift, but the lump changes shape.

I. INTRODUCTION

There are numerous examples of scalar field theories in (2 + 1) dimensions, which are integrable, including the Kadomtsev–Petviashvili and Davey–Stewartson equations. However, neither of these is Lorentz invariant in the sense of possessing an SO(2,1) symmetry acting on space-time. In fact, no example of a Lorentz invariant theory that is integrable in (2 + 1) dimensions is known, and it may well be that one does not exist.[1] A partial remedy is to take an SO(2,1) invariant model and modify it slightly, in such a way as to trade Lorentz invariance for integrability. One may still hope for a "generalized" Lorentz invariance, in the sense that the behavior of the soliton solutions, or of some restricted class of soliton solutions, is Lorentz invariant.

Ward[2] has studied a modification of the SU(2) chiral model in (2 + 1) dimensions, given by the field equation

$$(\eta^{\mu\nu} + V_\alpha \epsilon^{\alpha\mu\nu})\partial_\mu (J^{-1}\partial_\nu J) = 0. \tag{1.1}$$

Here J takes values in SU(2) and is thought of as a 2×2 matrix of functions of the space-time coordinates (t,x,y), sometimes also written (x^0,x^1,x^2). Greek letters are space-time indices, taking values 0,1,2, and ∂_μ denotes partial differentiation with respect to x^μ. The quantity $\epsilon^{\alpha\mu\nu}$ is the alternating tensor on three indices (with ϵ^{012} taken equal to + 1) and $\eta^{\mu\nu} = \mathrm{diag}(-1,+1,+1)$ is the (inverse) Minkowski metric. Finally, V_α is a constant vector in space-time.

Choosing $V_\alpha = (0,0,0)$ corresponds to the unmodified chiral model, which is Lorentz invariant but nonintegrable. Note that a nonzero V_α explicitly breaks Lorentz invariance by picking out a particular direction in space-time. A case of particular interest occurs when V_α is chosen to be a spacelike unit vector, i.e., $\eta^{\mu\nu}V_\mu V_\nu = +1$, since then the theory appears to be integrable.[1] Moreover, if $V_0 = 0$, then the theory possesses the same conserved energy-momentum vector as the unmodified chiral model, namely,

$$P_\mu = (-\delta^\alpha_\mu\delta^\beta_0 + \tfrac{1}{2}\eta_{\mu 0}\eta^{\alpha\beta})\mathrm{Tr}(J^{-1}J_\alpha J^{-1}J_\beta). \tag{1.2}$$

The corresponding energy density is

$$P_0 = -\tfrac{1}{2}\mathrm{Tr}((J^{-1}J_t)^2 + (J^{-1}J_x)^2 + (J^{-1}J_y)^2). \tag{1.3}$$

Here δ^α_μ is the Kronecker delta, Tr denotes the matrix trace, and $J_\alpha \equiv \partial_\alpha J$. It should be emphasized that P_0 is a positive-definite functional of the field J.

If $V_0 \neq 0$ then it is not at all clear that a conserved energy-momentum vector exists, and so from now on, in order to ensure integrability and a conserved energy, we shall take V_α to be a spacelike unit vector with $V_0 = 0$. To be specific, choose $V_\alpha = (0,1,0)$. Ward has shown that this model admits solitons, localized in two dimensions, which pass through each other without scattering or changing shape. It is the purpose of this paper to construct extended plane wave solutions and to investigate their interactions. Such waves are localized along the direction of motion, but have infinite spatial extent perpendicular to it.

In fact, one family of extended solutions may be exhibited immediately by noting that (1.1) is a generalization of the sine–Gordon (SG) equation in (1 + 1) dimensions. Consider a J of the form

$$J = \begin{pmatrix} \cos\frac{1}{2}\phi & e^{-2ix}\sin\frac{1}{2}\phi \\ -e^{2ix}\sin\frac{1}{2}\phi & \cos\frac{1}{2}\phi \end{pmatrix}, \tag{1.4}$$

where the field ϕ depends on y and t, but not on x. Then the field equation (1.1) with $V_\alpha = (0,1,0)$ is equivalent to the SG equation for ϕ:

$$\phi_{tt} - \phi_{yy} + 4\sin\phi = 0. \tag{1.5}$$

Furthermore, the energy density (1.3) becomes

$$P_0 = \tfrac{1}{4}(\phi_t^2 + \phi_y^2) + 4\sin^2\tfrac{1}{2}\phi, \tag{1.6}$$

which is precisely the energy density of the sine–Gordon theory. In other words, there are solutions that look like SG solitons living in the (1 + 1)-dimensional subspace spanned by (y,t), but spatially extended along the x axis. Note that while the J of Eq. (1.4) depends explicitly on x, the corresponding energy density (1.6) does not. This illustrates a general feature of extended wave solutions: although P_0 only depends on time together with one spatial coordinate (along the direction of motion), J is necessarily a function of all three space-time coordinates.

II. CONSTRUCTION OF SOLUTIONS

This section summarizes the general method for constructing multisoliton solutions of the field equation (1.1). The technique is a variation of the well-known "Riemann problem with zeros" (see, for example, Forgács *et al.*[3]), and full details are to be found in the paper by Ward.[2]

There are two ingredients to an n-soliton solution. First, a set of n complex numbers μ_k (k taking values from 1 to n), which must all be different and nonreal; second, for each k, a meromorphic function f_k of the linear combination

$$\omega_k = x + \tfrac{1}{2}\mu_k(t+y) + \tfrac{1}{2}\mu_k^{-1}(t-y). \tag{2.1}$$

 0022-2488/89/092072-06$02.50

Now form the two-component objects $m_a^k = (1, f_k)$, so that a takes values 1,2 with $m_1^k = 1$ and $m_2^k = f_k$. Then (the inverse of) a matrix J, which satisfies the field equation, is given by

$$(J^{-1})_{ab} = \frac{1}{\sqrt{\alpha}}\left(\delta_{ab} + \sum_{k,l} \frac{1}{\mu_k}(\Gamma^{-1})^{kl}\bar{m}_a^l m_b^k\right), \tag{2.2}$$

where

$$\Gamma^{kl} = \sum_{a=1}^{2} (\bar{\mu}_k - \mu_l)^{-1}\bar{m}_a^k m_a^l,$$

$$\alpha = \prod_{k=1}^{n} \frac{\bar{\mu}_k}{\mu_k},$$

and an overbar denotes complex conjugation.

Clearly, the expression for J becomes very complicated very quickly as n is increased. Fortunately, there is plenty of analysis that can be done while still taking n small. For the rest of this section, and the entire Sec. III, n will be equal to unity. Later on, a study of interactions (Secs. IV and V) will require $n = 2$.

To get a feel for the physical picture, first we shall investigate a simple family of lump solutions, very similar to those discussed by Ward. Consider $n = 1$, in which case solutions are specified by a complex number μ and a meromorphic function $f(\omega)$. Equation (2.2) simplifies to give

$$J = \frac{1}{|\mu|(1+|f|^2)}\begin{pmatrix} \mu + \bar{\mu}|f|^2 & (\mu - \bar{\mu})f \\ (\mu - \bar{\mu})\bar{f} & \bar{\mu} + \mu|f|^2 \end{pmatrix}. \tag{2.3}$$

Writing $\mu = me^{i\theta}$, the energy density becomes

$$P_0 = \frac{2(1+m^2)^2 \sin^2\theta}{m^2} \frac{|f'|^2}{(1+|f|^2)^2}, \tag{2.4}$$

where f' is the derivative of f as a function of ω. Keeping things simple, choose $f(\omega) = a\omega + c$, where $a \in \mathbb{R}$ and $c \in \mathbb{C}$. (One could generate a larger set of solutions by taking a also in $\mathbb{C}$, but this is a little more tricky to analyze and an unnecessary complication to introduce at this stage.) The factor $|f'|^2$ in the numerator of (2.4) becomes just a^2. So it is seen that the solution looks like a single lump located at the point where $f = a\omega + c = 0$. From (2.1), its velocity is computed to be

$$(v_x, v_y) = \left(\frac{-2m\cos\theta}{1+m^2}, \frac{1-m^2}{1+m^2}\right). \tag{2.5}$$

The parameters μ, a, and c have simple physical interpretations: μ specifies the soliton velocity via (2.5), c determines the position of the peak at time $t = 0$ and, finally, a fixes the ratio of the height of the lump to its width. Note that in the static case ($\mu = i$) one may easily integrate P_0 over x and y to obtain the total energy E. The result is

$$E \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_0\, dx\, dy = 8\pi, \tag{2.6}$$

which is independent of a.

III. EXTENDED WAVE SOLUTIONS

Now we shall set out to construct a family of extended wave solutions. Ward showed that taking f to be rational of degree N leads to a configuration with N peaks, which in the static case has energy $8N\pi$. An extended wave must have infinite energy and so no function of finite degree will do for f. The next candidate is some sort of exponential. Specifically, consider

$$f(\omega) = \exp(b\omega + c). \tag{3.1}$$

This leads to an energy density

$$P_0 = \frac{2(1+m^2)^2\sin^2\theta}{m^2} \frac{|b|^2|f|^2}{(1+|f|^2)^2}. \tag{3.2}$$

Here $\mu = me^{i\theta}$ as before. Note that P_0 only depends on c through its real part and so, without loss of generality, we can take $c \in \mathbb{R}$. However, b is, in general, complex. To see that (3.2) does indeed look like a wave, rewrite it as

$$P_0 = [(1+m^2)^2 \sin^2\theta/2m^2]\,|b|^2 \operatorname{sech}^2(\mathrm{Re}(b\omega) + c). \tag{3.3}$$

Note that P_0 is constant along each of the lines $\mathrm{Re}(b\omega) + c = \text{const}$. The wave front (i.e., the crest of the wave) lies along $\mathrm{Re}(b\omega) + c = 0$. For each value of t, this is the equation of a straight line in the xy plane. As t varies, the wave maintains its shape and simply moves at constant velocity.

To investigate this wave in more detail, write $b = |b|e^{i\alpha}$. Then the equation of the wave front may be written

$$Ax + By = Ct + D, \tag{3.4}$$

where

$$A = 2m\cos\alpha,$$
$$B = m^2\cos(\theta + \alpha) - \cos(\theta - \alpha),$$
$$C = -m^2\cos(\theta + \alpha) - \cos(\theta - \alpha),$$
$$D = -2mc/|b|.$$

The velocity may be readily calculated:

$$v_x = \frac{-2m\cos\alpha(\cos(\theta-\alpha) + m^2\cos(\theta+\alpha))}{m^4\cos^2(\theta+\alpha) + 2m^2(\sin^2\theta + \cos^2\alpha) + \cos^2(\theta-\alpha)},$$
$$v_y = \frac{\cos^2(\theta-\alpha) - m^4\cos^2(\theta+\alpha)}{m^4\cos^2(\theta+\alpha) + 2m^2(\sin^2\theta + \cos^2\alpha) + \cos^2(\theta-\alpha)}. \tag{3.5}$$

The speed v is given by

$$v^2 = 1 - \frac{4m^2\sin^2\theta}{m^4\cos^2(\theta+\alpha) + 2m^2(\sin^2\theta + \cos^2\alpha) + \cos^2(\theta-\alpha)} \tag{3.6}$$

and, although the integral of P_0 over all space is divergent, instead one can calculate the energy per unit length along the wave front, which turns out to be

$$\hat{E} = \frac{4\gamma|b||\sin^3\theta|(1+m^2)^2}{m^4\cos^2(\theta+\alpha)+2m^2(\sin^2\theta+\cos^2\alpha)+\cos^2(\theta-\alpha)}, \tag{3.7}$$

where $\gamma = (1-v^2)^{-1/2}$.

The question of classification of these solutions now arises. Clearly, c determines the position of the wave at time $t=0$. On the face of it, there are four other real parameters (m, θ, $|b|$, and α), which one might naively think could be chosen to fix the velocity (two parameters), the wave height, and wave width (one parameter each), all independently. If this were the case, it would support the conjecture of generalized Lorentz invariance. However, the following systematic study of some special cases shows that things are not quite so simple.

It is not difficult to pick out the solutions that look like static waves aligned along the coordinate axes. A wave lying on the x axis requires $\cos\alpha = 0$ and $m = 1$. Setting $k = ib\sin\theta$ leads to

$$P_0 = 2k^2\,\mathrm{sech}^2\,ky, \tag{3.8}$$

which is (if one parametrizes the solution using k and θ, rather than $|b|$ and θ) independent of θ. On the other hand, the conditions for a wave to lie on the y axis are $\cos\theta = \sin\alpha = 0$. Setting $k = |b|$ leads to

$$P_0 = [(1+m^2)^2/2m^2]k^2\,\mathrm{sech}^2\,kx. \tag{3.9}$$

In the latter case the height and width may be chosen independently, while in the former they are determined by a single parameter, with θ playing the role of an "internal" degree of freedom. Although these observations do not necessarily rule out a generalized Lorentz invariance [there may be other plane wave solutions, not generated by (3.1)], they make it seem unlikely.

The complete classification of waves generated by (3.1) appears to be difficult. Therefore it is useful to study subsets of solutions obtained by imposing some extra condition on the parameters. For example, if one requires

$$(1+m^2)\tan\alpha = (1-m^2)\tan\theta, \tag{3.10}$$

then the wave velocity (3.5) becomes

$$(v_x, v_y) = \left(-\frac{2m\cos\theta}{1+m^2}, \frac{1-m^2}{1+m^2}\right), \tag{3.11}$$

which matches the expression (2.5) for lump solutions, and the energy per unit length becomes

$$\hat{E} = 4\gamma|b||\sin^3\theta|. \tag{3.12}$$

In this scheme a physical interpretation becomes apparent: m and θ specify the velocity and $|b|$ determines both the height and width, with α fixed by (3.10). To make the situation even more transparent, one can replace $(m, |b|, \theta)$ with new parameters (k, ϕ, A), defined by

$$\begin{aligned} m &= [(1-k\sin\phi)/(1+k\sin\phi)]^{1/2}, \\ \cos\theta &= -k\cos\phi/(\sqrt{1-k^2\sin^2\phi}), \\ |b| &= A\sqrt{1-k^2\sin^2\phi}. \end{aligned} \tag{3.13}$$

Then the energy density becomes

$$P_0 = [2A^2(1-k^2)/(1-k^2\sin^2\phi)]\,\mathrm{sech}^2\,A(x\cos\phi + y\sin\phi - kt). \tag{3.14}$$

Now it is seen that k is the wave speed, ϕ is the angle of the direction of motion relative to the x axis, and A is the width. The height is then a simple function of k, ϕ, and A. Taking $\phi = \pi/2$ and $A = 2/\sqrt{1-k^2}$ leads to the extended SG waves mentioned earlier, namely,

$$P_0 = [8/(1-k^2)]\,\mathrm{sech}^2[2/(\sqrt{1-k^2})](y-kt). \tag{3.15}$$

Note that in this case, Eq. (3.12) reduces to $\hat{E} = 8\gamma$, confirming the relativistic behavior of SG solitons.

IV. WAVE–WAVE INTERACTIONS

It was seen in Sec. III that the classification of extended wave solutions is a far from trivial matter, and clearly a complete study of their interactions will be no simpler. Instead, we shall present an analysis of a few particular cases, pointing out the main features. It seems likely that the general case will be very similar.

For a two-soliton solution one takes $n = 2$ in the prescription of Sec. II. It turns out that the algebra is much simplified if μ_k is restricted to be pure imaginary, i.e., $\theta_k = \pi/2$. So as an example of a solution containing two waves, W_1 and W_2, consider

$$\mu_k = ip_k, \quad f_k(\omega_k) = \exp(b_k\omega_k + c_k), \tag{4.1}$$

where k takes values 1,2; p_k is real and

$$b_k = A_k[(1+p_k^2)\cos\phi_k - 2ip_k\sin\phi_k]. \tag{4.2}$$

Physically, $\phi_k \in [0,\pi]$ gives the direction of motion of each wave and the speed is $\sin\phi_k(1-p_k^2)/(1+p_k^2)$. The positive real parameter A_k fixes the width and height.

Even with the simplification of taking μ_k imaginary, the full expression for P_0 is rather complicated, but one can investigate the asymptotic behavior in the following sense. Recall that the equation of each wave front is $\mathrm{Re}(b_k\omega_k) + c_k = 0$. Taking the limits $\mathrm{Re}(b_1\omega_1) \to \pm\infty$ corresponds to moving far away from wave W_1 on either side. If at the same time $\mathrm{Re}(b_2\omega_2)$ is kept finite, then roughly speaking we are keeping our eyes fixed on W_2, but far away from W_1. To keep things in terms of f_k, note that $\mathrm{Re}(b_k\omega_k) \to +\infty$ im-

plies $|f_k| \to \infty$ and $\mathrm{Re}(b_k\omega_k) \to -\infty$ implies $|f_k| \to 0$.

Now let k' stand for "not k," so that $1' = 2$ and $2' = 1$. Then the asymptotics of the solution (4.1) may be summarized as follows:

$$|f_{k'}| \to \infty, \quad P_0 \sim \frac{2(p_1^2 - p_2^2)^2(p_k^2+1)^2 A_k^2((p_k^2+1)^2 - (p_k^2-1)^2 \sin^2\phi_k)|f_k|^2}{p_k^2(|f_k|^2(p_1-p_2)^2 + (p_1+p_2)^2)^2},$$
$$|f_{k'}| \to 0, \quad P_0 \sim \frac{2(p_1^2 - p_2^2)^2(p_k^2+1)^2 A_k^2((p_k^2+1)^2 - (p_k^2-1)^2 \sin^2\phi_k)|f_k|^2}{p_k^2(|f_k|^2(p_1+p_2)^2 + (p_1-p_2)^2)^2}. \tag{4.3}$$

The crucial point is the difference of sign in the denominators. It is not difficult to see that the essential behavior is

$$|f_{k'}| \to \infty, \quad P_0 \sim \mathrm{sech}^2(\mathrm{Re}(b_k\omega_k) + c_k - \gamma),$$
$$|f_{k'}| \to 0, \quad P_0 \sim \mathrm{sech}^2(\mathrm{Re}(b_k\omega_k) + c_k + \gamma), \tag{4.4}$$

where

$$\tanh\gamma = 2p_1p_2/(p_1^2 + p_2^2). \tag{4.5}$$

So the waves interact in a fairly simple way: each experiences a phase shift 2γ. The SG waves are present in the above solutions as the special case $\phi_k = \pi/2$, $A_k = 1/|p_k|$.

Figure 1 shows a snapshot of the energy density at time $t = 0$ for the following choice of parameters: $(p_1,A_1,\phi_1) = (1,2,\pi/2)$, $(p_2,A_2,\phi_2) = (2,1,\pi/4)$. The phase shift suffered by each wave is clearly visible. Note also the highly nonlinear superposition in the region of intersection.

One may ask whether internal parameters (which do not appear in the single wave energy density) can affect interactions. The answer is yes, as the following example will show.

Consider W_1 and W_2, both parallel to the x axis with W_2 stationary, i.e., $\phi_1 = \phi_2 = \pi/2$, $p_2 = 1$. Choose $p_1 = 1/\sqrt{2}$. Then in the above scheme, $\tanh\gamma = 2\sqrt{2}/3$. But now note that W_2 is equally well described by

$$\mu_2 = \exp(i\pi/4), \quad f_2(\omega_2) = \exp(-2\sqrt{2}iA_2\omega_2 + c_2).$$

Repeating the calculation (although μ_2 is not now pure imaginary, the parameters have been chosen to make the algebra as tractable as possible) one finds precisely the same asymptotic behavior but with a new phase shift γ' given by $\tanh\gamma' = \frac{2}{3}$.

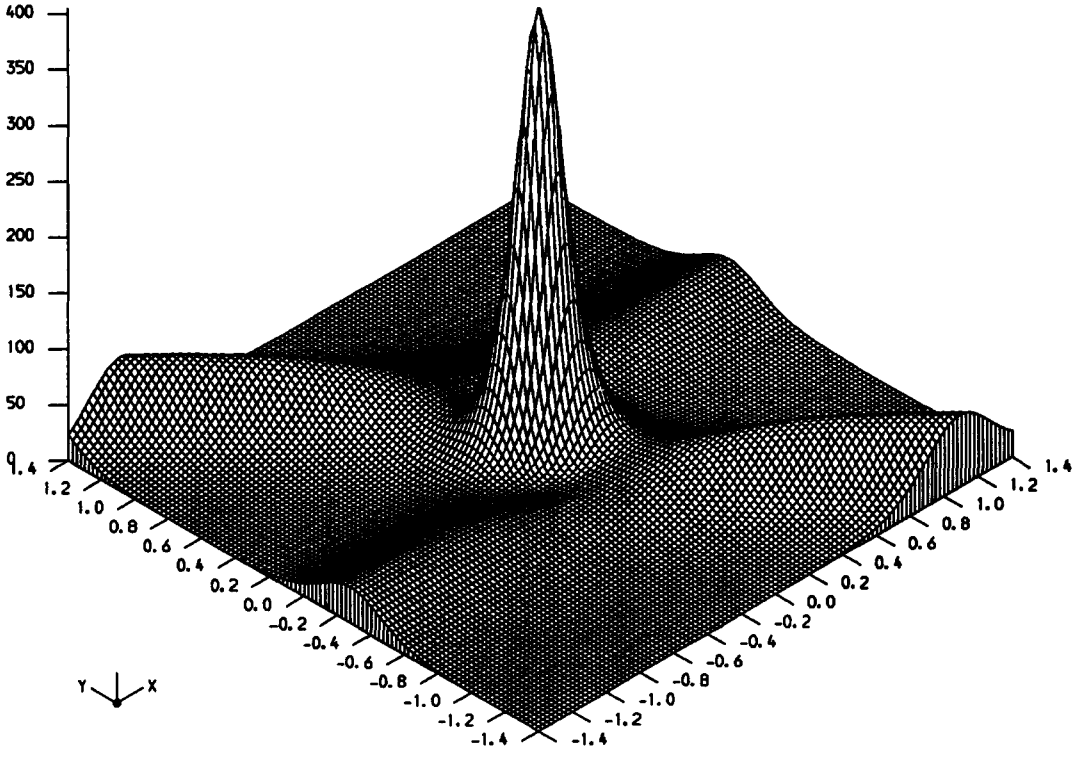


FIG. 1. A snapshot of the energy density for a two-wave interaction. The flatter wave is stationary along the x axis, and the taller one is moving across it at an angle of 45°.

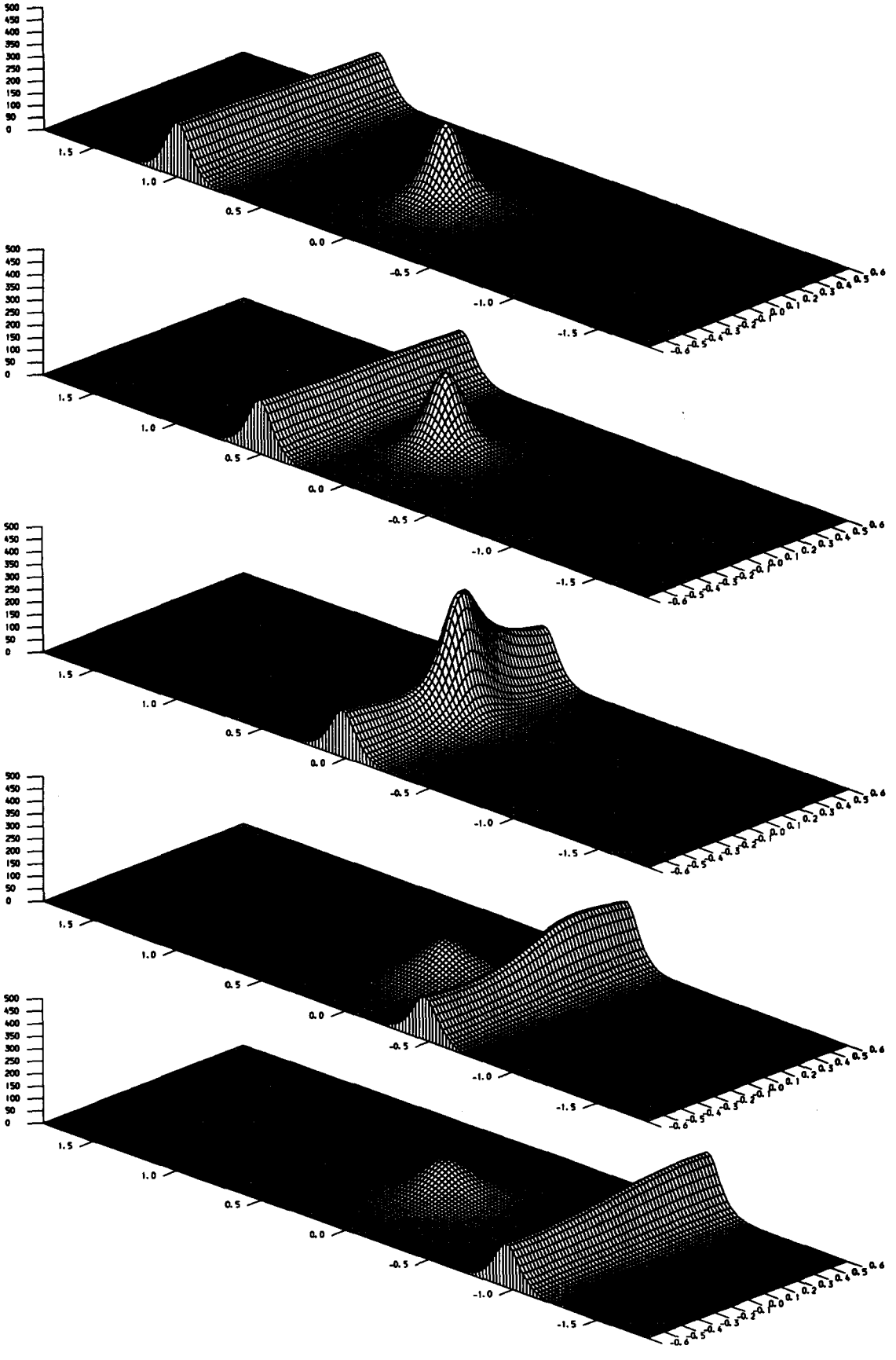


FIG. 2. A series of snapshots of the energy density for a wave–lump interaction. The lump is stationary at the origin and the wave is travelling parallel to the y axis. Time runs down the page in intervals of 0.5, starting at $t = -1.0$.

To sum up, as two waves interact, they do not change shape or velocity, but each has a phase shift across the region of intersection, which may be dependent upon internal parameters.

V. WAVE–LUMP INTERACTIONS

For simplicity, we shall only consider the case of a plane wave W_1 incident on a stationary lump L_2. In terms of the input to the two-soliton solution, take the same μ_1 and f_1 as in Sec. IV, but now with $\mu_2 = i$ and $f_2(\omega_2) = A_2\omega_2$ ($A_2 \in \mathbb{R}$ as in Sec. II). Again the effects of the interaction are revealed by the asymptotic behavior of P_0. Note that to look far away from L_2 in any direction, the relevant limit is $|f_2| \to \infty$. One finds the following:

$$
\begin{aligned}
&|f_1| \to \infty, \quad P_0 \sim \frac{8(p_1^2-1)^2 A_2^2}{(|f_2|^2(p_1-1)^2+(p_1+1)^2)^2}, \\
&|f_1| \to 0, \quad P_0 \sim \frac{8(p_1^2-1)^2 A_2^2}{(|f_2|^2(p_1+1)^2+(p_1-1)^2)^2}, \qquad (5.1) \\
&|f_2| \to \infty, \\
&P_0 \sim \frac{2A_1^2(p_1^4-1)^2((p_1^2+1)^2-(p_1^2-1)^2\sin^2\phi_1)|f_1|^2}{p_1^2(|f_1|^2(p_1-1)^2+(p_1+1)^2)^2}.
\end{aligned}
$$

The physical picture is this: the shape and velocity of the wave are the same long before and long after the collision, and it suffers no phase shift. The more remarkable feature is that the lump remains stationary, but changes its height by a factor

$$(p_1-1)^4/(p_1+1)^4.$$

Again the crucial point is the difference of sign in the denominators. A little care is needed at this stage, since it is not immediately clear which limit of $|f_1|$ corresponds to $t \to -\infty$ and which to $t \to +\infty$. The answer to this question depends on the size of p_1:

$$
\begin{aligned}
&\text{For } |p_1|<1, \quad \begin{cases} t \to -\infty \Rightarrow |f_1| \to \infty, \\ t \to +\infty \Rightarrow |f_1| \to 0; \end{cases} \\
&\text{for } |p_1|>1, \quad \begin{cases} t \to -\infty \Rightarrow |f_1| \to 0, \\ t \to +\infty \Rightarrow |f_1| \to \infty. \end{cases}
\end{aligned}
$$

So for $p_1 > 1$ or $-1 < p_1 < 0$ the lump decreases in height and for $p_1 < -1$ or $0 < p_1 < 1$ it increases in height. Figure 2 shows a series of snapshots taken at time intervals of 0.5, starting at $t = -1.0$, for the following parameters: $\phi_1 = \pi/2$, $A_1 = 0.1$, $A_2 = 5$, and $p_1 = 10$. In this case, W_1 is an SG wave. The lump decreases in height by a factor $(9/11)^4$ (≈ 0.45), but its total energy remains unchanged, equal to 8π.

Perhaps the most puzzling feature is the transverse asymmetric kink acquired by the wave as it squashes the lump, and which then gradually dies away. It could be that this is due somehow to the absence of Lorentz invariance. Alternatively, internal parameters at work may provide the explanation. In any event, the interaction seems quite unlike any occurring in other integrable models.

VI. CONCLUDING REMARKS

The modified SU(2) chiral model and the Kadomtsev–Petviashvili (KP) equation have several features in common. The latter also possesses "rational" solitons, which look like lumps and "exponential" solitons, which look like waves.[4] In both models two lumps pass through each other without scattering and two waves interact with a phase shift. However, the wave–lump interaction of Sec. V seems to have no analog in KP; compare, for example, with Fig. 9 of Ref. 4.

As a final remark, it might be interesting to consider letting the field J live in a noncompact Lie group such as $SL(2,\mathbb{R})$ or $SL(2,\mathbb{C})$. This would mean that the energy density is no longer positive definite, but should not rule out explicit construction of solutions. An $SL(2,\mathbb{R})$ model is expected to have embedded in it the KdV equation in $(1+1)$ dimensions, while taking J in $SL(2,\mathbb{C})$ will also include the nonlinear Schrödinger equation.[5] Maybe these models will exhibit a behavior closer to KP, since they are both, in some sense, generalizations of KdV, unlike the current SU(2) model.

ACKNOWLEDGMENTS

I am most grateful to Richard Ward for many useful discussions and also to the SERC for a Research Studentship.

[1] R. S. Ward, Nonlinearity **1**, 671 (1988).
[2] R. S. Ward, J. Math. Phys. **29**, 386 (1988).
[3] P. Forgács, Z. Horváth, and L. Palla, Nucl. Phys. B **229**, 77 (1983).
[4] N. C. Freeman, Adv. Appl. Mech. **20**, 1 (1980).
[5] L. J. Mason and G. A. J. Sparling, Phys. Lett. A **137**, 29 (1989).

On the entropic formulation of uncertainty for quantum measurements

Franklin E. Schroeck, Jr.
Institut für Theoretische Physik, Universität zu Köln, D-5000, Köln, West Germany, and Department of Mathematics, Florida Atlantic University, Boca Raton, Florida 33431

(Received 19 May 1988; accepted for publication 10 May 1989)

The formulation of the uncertainty principle based on entropy is given with the (noncommuting) observables described via positive operator-valued measures.

I. INTRODUCTION

In previous papers, Deutsch[1] and Partovi[2] have presented, as an alternative to the Heisenberg uncertainty relations, an analysis of uncertainty based on the semibounded nature of the sum of the entropies of a state ψ, relative to the measurement of two observables $\hat{A},\hat{B}$ with discrete spectra[1] (or with discrete partition of continuous spectra[2]):

$$U(\hat{A},\hat{B}\,|\psi) = S^A(\psi) + S^B(\psi),$$

$$S^A(\psi) = -\sum_a |\langle a|\psi\rangle|^2 \ln|\langle a|\psi\rangle|^2, \tag{1}$$

where $\{|a\rangle\}$ is an orthonormal set of eigenstates of spectral projectors of $\hat{A}$ forming a basis, and similarly for $\hat{B}$. In both analyses, however, the expression for uncertainty does not represent any uncertainty having to do with *simultaneous* measurement of $\hat{A}$ and $\hat{B}$, since in no case does a joint distribution for $\hat{A}$ and $\hat{B}$ enter their calculations.

We present here an analysis of such joint measurements and the corresponding bounds on uncertainty, based on the positive operator-valued measure (POV) or effect formalism by means of which the joint distribution for (unsharp) measurement of (noncommuting) observables is known to be realizable[3–5] and necessary. This analysis simplifies and clarifies the essential points in the previous analysis, as well as being general enough to lessen the gap known to exist between the Deutsch–Partovi lower bound and the actual greatest lower bound.[6]

II. GENERAL ANALYSIS

Let (X,Σ) be a measurable space, and let $\mathscr{H}$ denote the Hilbert space on which our quantum system is described. [We have in mind that X is the space of experimental outcomes. If one records data as points on a screen, then $X\subseteq\mathbb{R}^2$, for example; X may either be a continuum or a discrete set, or a little of each.] A POV, or effect valued measure, is a function A from Σ to the positive operators on $\mathscr{H}$ such that

$$\text{(a)}\quad A(\cup_i\Delta_i) = \Sigma_i A(\Delta_i), \tag{2a}$$

$\{\Delta_i\}$ any countable disjoint family of measurable sets;

$$\text{(b)}\quad A(X) = \mathbf{1}. \tag{2b}$$

One then has $A(\Delta)\leqslant\mathbf{1}$ for all $\Delta\in\Sigma$.

For any quantum (statistical) density operator ρ, $\mathrm{tr}(\rho A(\Delta))$ represents the probability for obtaining an experimental result in Δ.

If (X,Σ,μ) is a measure space [μ a measure on (X,Σ)] then A will be said to be absolutely continuous with respect to μ in the uniform sense (which we denote by $A\ll\mu$) iff there exists some constant $c>0$ such that

$$\|A(\Delta)\|\leqslant c\mu(\Delta), \quad \text{for all} \quad \Delta\in\Sigma. \tag{3}$$

The covariant POV's for Weyl systems which describe the unsharp simultaneous measurement of position and momentum[3] and the covariant POV's for simultaneous measurement of different spin $\frac{1}{2}$ components[3] satisfy absolute continuity since they are of the form

$$A(\Delta) = \int_\Delta T_x\, d\mu(x), \tag{4}$$

where T_x is a bounded operator-valued density. Furthermore, these POV's are derived from (covariant) instruments in the Davies–Lewis sense,[3,7] which from the analysis of Ozawa,[8] has an interpretation in terms of an interaction with a measuring device, so that one might experimentally interpret the meaning of these POV's. The set of observables simultaneously measurable in terms of the instrument are constructed from these POV's.[3,9] In the Weyl case, for example, (all functions of) the momentum and position operators may be so constructed.[3]

Ignoring any physical motivation, one may construct a joint POV for any finite collection of observables, commuting or not. In fact, there are an uncountable family of such POV's; so the existence of joint POV's is not in question.[10]

We shall pursue the simplest case here: We perform a countable partition $\{(\Delta_{ij})\,|\,(i,j)\in\mathbb{N}\times\mathbb{N}\}$ of X such that

$$\begin{aligned} &A(\cup_j\Delta_{ij}) = E^{\hat{A}}(\Delta_i), \quad \Delta_i = \cup_j\Delta_{ij},\\ &A(\cup_i\Delta_{ij}) = E^{\hat{B}}(\Delta'_j), \quad \Delta'_j = \cup_i\Delta_{ij}, \end{aligned} \tag{5}$$

define POV's describing unsharp measurement of marginal observables $\hat{A},\hat{B}$, respectively. The entropy for the entire measurement is then given by

$$S(\hat{A},\hat{B}\,|\rho) = -\Sigma_{i,j}P_{ij}\ln P_{ij}, \tag{6a}$$

where

$$P_{ij} = \mathrm{tr}(A(\Delta_{ij})\rho), \tag{6b}$$

and ρ is any density operator. Since, for each fixed i,

$$\Sigma_j P_{ij}\ln P_{ij} \leqslant \Sigma_j P_{ij}\ln(\Sigma_{j'}P_{ij'}),$$

we have

$$-\Sigma_{ij}P_{ij}\ln P_{ij} \geqslant -\Sigma_i P_i \ln P_i = S(\hat{A}\,|\rho),$$

where

$$P_i = \Sigma_j P_{ij} = \mathrm{tr}(\rho E^{\hat{A}}(\Delta_i)).$$

Similarly,

$$S(\hat{A},\hat{B}\,|\rho) \geqslant S(\hat{B}\,|\rho) = -\Sigma_j P'_j \ln P'_j,$$

$$P'_j = \mathrm{tr}(\rho E^{\hat{B}}(\Delta'_j)).$$

 0022-2488/89/092078-05$02.50

Thus

$$S(\hat{A},\hat{B}\,|\rho)\geqslant\tfrac{1}{2}[S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)]. \tag{7}$$

But, by the same reasoning, any convex combination of $S(\hat{A}\,|\rho)$ and $S(\hat{B}\,|\rho)$ gives a lower bound to $S(\hat{A},\hat{B}\,|\rho)$.

Also, from $\ln(x)\leqslant x-1$, we have, for $\Sigma'_{i,j}$ the sum over i,j such that $P_{ij}\neq 0$,

$$\begin{aligned}&-\Sigma_{i,j}P_{ij}\ln P_{ij}+\Sigma_{i,j}P_{ij}\ln(P_iP'_j)\\&\quad=\Sigma'_{i,j}P_{ij}\ln(P_iP'_j/P_{ij})\leqslant\Sigma'_{i,j}P_{ij}(P_iP'_j/P_{ij}-1)\\&\quad=\Sigma'_{i,j}(P_iP'_j-P_{ij})=1-1=0.\end{aligned}$$

Now

$$\begin{aligned}S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)&=-\Sigma_i(P_i)\ln P_i-\Sigma_jP'_j\ln P'_j\\&=-\Sigma_i(\Sigma_jP_{ij})\ln P_i-\Sigma_i(\Sigma_jP_{ij})\ln P'_j\\&=-\Sigma_{i,j}P_{ij}\ln P_iP'_j.\end{aligned} \tag{8}$$

From the previous inequality, then,

$$S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)\geqslant-\Sigma_{ij}P_{ij}\ln P_{ij}=S(\hat{A},\hat{B}\,|\rho).$$

In summary, we have

$$\begin{aligned}&S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)\geqslant S(\hat{A},\hat{B}\,|\rho)\\&\quad\geqslant\lambda S(\hat{A}\,|\rho)+(1-\lambda)S(\hat{B}\,|\rho),\quad\lambda\in[0,1].\end{aligned} \tag{7'}$$

For $\lambda=\frac{1}{2}$, this shows the entire relevance of the Deutsch–Partovi functional to the complete expression for entropy for joint measurement. The difference between $S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)$ and $S(\hat{A},\hat{B}\,|\rho)$ is termed the "correlation information." [11]

An alternative derivation of (7) may be made by beginning on the right-hand side of (7) and using, for each $(i_0j_0)\in\mathbb{N}\times\mathbb{N}$,

$$\ln[(\Sigma_jP_{i_0j})(\Sigma_iP_{ij_0})]\geqslant\ln[P^2_{i_0j_0}]$$

or diagrammatically, the probability squared of falling in the intersection of vertical and horizontal slices of X space is smaller than the product of the probabilities of falling in the slices. Now,

$$\begin{aligned}P_iP'_j&\leqslant\tfrac{1}{4}[P_i+P'_j]^2\\&=\tfrac{1}{4}\{\mathrm{tr}(\rho[E^{\hat{A}}(\Delta_i)+E^{\hat{B}}(\Delta'_j)])\}^2\\&\leqslant\tfrac{1}{4}\{\mathrm{tr}(\rho[E^{\hat{A}}(\Delta_i)+E^{\hat{B}}(\Delta'_j)]^2)\}\\&\leqslant\tfrac{1}{4}\|E^{\hat{A}}(\Delta_i)+E^{\hat{B}}(\Delta'_j)\|^2\\&=\tfrac{1}{4}\|A(\cup_{j'}\Delta_{ij'})+A(\cup_{i'}\Delta_{i'j})\|^2\\&=\tfrac{1}{4}\|A([\cup_{j'\neq j}\Delta_{ij'}]\cup[\cup_{i'}\Delta_{i'j}])+A(\Delta_{ij})\|^2\\&\leqslant\tfrac{1}{4}\|A(X)+A(\Delta_{ij})\|^2\\&\leqslant\tfrac{1}{4}\|\mathbf{1}+A(\Delta_{ij})\|^2\\&\leqslant\tfrac{1}{4}(1+\|A(\Delta_{ij})\|)^2,\end{aligned} \tag{9}$$

the second inequality coming from a convexity argument,[12] and the fourth inequality coming from the comparison of the union of a horizontal and a vertical slice of X with all of X. Thus, using (8), (9),

$$\begin{aligned}S(\hat{A},\hat{B}\,|\rho)&\geqslant\tfrac{1}{2}[S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)]\\&=-\tfrac{1}{2}\Sigma_{i,j}P_iP'_j\ln(P_iP'_j)\\&\geqslant-\tfrac{1}{2}\Sigma_{i,j}P_iP'_j\ln[\tfrac{1}{4}(1+\|A(\Delta_{ij})\|)^2]\\&=\Sigma_{i,j}P_iP'_j\ln[2/(1+\|A(\Delta_{ij})\|)].\end{aligned} \tag{10}$$

In the absolutely continuous case $A\ll\mu$, then

$$\|A(\Delta_{ij})\|\leqslant\min\{1,c\mu(\Delta_{ij})\}\equiv M_{ij},$$

and we obtain

$$\begin{aligned}S(\hat{A},\hat{B}\,|\rho)&\geqslant\tfrac{1}{2}[S(\hat{A}\,|\rho)+S(\hat{B}\,|\rho)]\\&\geqslant\Sigma_{i,j}P_iP'_j\ln(2/(1+M_{ij})).\end{aligned} \tag{11}$$

We state, for emphasis, that so far in the derivation, it has not mattered whether A was projection valued or only a POV. What is important is that we are dealing explicitly with a description of joint measurement of $\hat{A}$ and $\hat{B}$ from the very begining, and this gives the simple inequality (9), from which (11) follows. Furthermore, in the special case that $\hat{A},\hat{B}$ commute, all of our results apply. In the case in which $\hat{A},\hat{B}$ do not commute, the difference between (10) computed with projection-valued measures and (6) computed with the true POV measure, the correlation information could be labelled the "missing information," due to the joint measurement. This is interpreted as the loss of information due to the unsharpness of the POV measurement. However, since any single instrument yields only one measure A, either A is projection valued or not, but not both simultaneously. The missing information seems to be an idealization, not experimentally accessible.

III. COMPARATIVE EXAMPLES

Next we show that in the examples considered by Partovi, (11) provides a higher state independent lower bound than in the Partovi analysis.

The first example is for a measurement of angle and z component of angular momentum. Partovi is unclear on what (range of) total angular momentum value is being considered. We assume that J_z takes values only in $\{-m,-m+1,\ldots,m\}$. We also do not know in which sense "angle" is an observable, so there may be no physics in this. Assuming that there is some angle observable with spectrum on the circle, we then take the instrument measure space to be

$$X=\mathrm{circle}\times\{-m,-m+1,\ldots,m\},$$

partition the circle into equal segments of angle $\Delta\phi$:

$$\mu(\Delta_i)=\Delta\phi/2\Pi,$$

and, in order to be able to resolve down to a single value of the angular momentum,

$$\mu(\Delta'_j)=(2m+1)^{-1},$$

$$\Delta'_j=\mathrm{circle}\times\{j\},\quad j\in\{-m,\ldots,m\}.$$

Then

$$c\mu(\Delta_{ij})=(\Delta\phi/2\Pi)(2m+1)^{-1}.$$

Choosing $\Delta\phi$ so this is <1, (11) becomes

$$S(\hat{A},\hat{B}\,|\rho) \geqslant \tfrac{1}{2}\,[S(\hat{A}\,|\rho) + S(\hat{B}\,|\rho)]$$

$$\geqslant \Sigma_{i,j} P_i P'_j \ln\left(\frac{2}{1+(\Delta\phi/2\Pi)(2m+1)^{-1}}\right)$$

$$= \ln\left(\frac{2}{1+(\Delta\phi/2\Pi)(2m+1)^{-1}}\right). \quad (12)$$

For $(\Delta\phi/2\Pi)^{1/2} > (\Delta\phi/2\Pi)(2m+1)^{-1}$, (12) provides a larger lower bound that Partovi's lower bound "(6)." These constraints may be rewritten

$$(2m+1)^2 > \Delta\phi/2\Pi,$$

which is always satisfied if $(\Delta\phi/2\Pi)(2m+1)^{-1} < 1$, which we assumed in order to derive (11). Furthermore, in the situation in which these constraints on $\Delta\phi$ are not satisfied, we may use instead

$$\|A(\Delta\phi \times \Delta m)\| \leqslant 1$$

to obtain $S(\hat{A},\hat{B}\,|\rho) \geqslant \tfrac{1}{2}\,[S(\hat{A}\,|\rho) + S(\hat{B}\,|\rho)] \geqslant 0$.

The second example of Partovi was for position X, and momentum P. He specified "bins" to be uniformly of size $\Delta X, \Delta P$, respectively. From Ref. 3, we have (for one dimension)

$$\mu(\Delta X \times \Delta P) = (2\Pi\hbar)^{-1}\Delta X\,\Delta P, \quad c = 1.$$

Hence (11) reads

$$S(\hat{A},\hat{B}\,|\rho) \geqslant \tfrac{1}{2}\,[S(\hat{A}\,|\rho) + S(\hat{B}\,|\rho)]$$

$$\geqslant \Sigma_{i,j} P_i P'_j \ln\left[\frac{2}{1+\max\{\Delta X\Delta P/2\Pi\hbar, 1\}}\right]$$

$$= \begin{cases} \ln\dfrac{2}{1+\Delta X\Delta P/2\Pi\hbar}, & \dfrac{\Delta X\Delta P}{2\Pi\hbar} < 1 \\ 0, \quad \text{otherwise.} \end{cases} \quad (13)$$

Comparing with Partovi's equation (7), we again have achieved a higher lower bound whenever $\Delta X\Delta P/2\Pi\hbar < 1$. Furthermore, the bin widths $\Delta X, \Delta P$ here represent the true instrument bin widths, not theoretical estimates, variances, etc.

An analysis of (1) under joint measurement of momentum and position in the stochastic quantum mechanics formalism has been carried out by Grabowski[13] and Busch and Lahti.[14] The more general idea of using the POV formalism to describe entropy and information was proposed by Ingarden.[15] What is being presented here is a realization of the general ideas presented by Ingarden, and an extension of the analysis of Grabowski, Busch, and Lahti.

IV. OBTAINING NEW LOWER BOUNDS

The results (10), (11) were obtained in a derivation parallel to the analysis of Partovi so that (12) and (13) could be derived for comparison. Even at this level, the results obtained from absolutely continuous POV's are better and more easily derived since no eigenstates and eigenvalues need be constructed. Since, however, $S(\hat{A},\hat{B}\,|\rho)$ defined by (6) is the quantum entropy for joint measurement of $\hat{A},\hat{B}$ and not (1), we should and will analyze (6) directly. We may, for given $\hat{A},\hat{B}$ (a) find a lower bound for $S(\hat{A},\hat{B}\,|\rho)$, which is ρ independent and test to see if this lower bound is attainable; (b) find, by variational calculus, ρ, which optimizes $S(\hat{A},\hat{B}\,|\rho)$ and find the minimum value if desired; (c) vary the POV (the instrument), which gives $\hat{A},\hat{B}$ as marginals, and for fixed ρ find a ρ dependent lower bound; (d) other procedures.

We shall consider (a), (b) only. For (a) we have simply

$$S(\hat{A},\hat{B}\,|\rho) \geqslant \Sigma_{i,j} P_{ij} \ln\left[(\sup_{i'j'} P_{i'j'})^{-1}\right]$$

$$= \ln(\sup_{i,j} P_{ij})^{-1}, \quad (14)$$

which is a state-dependent lower bound. [One could ask which states achieve equality in this inequality.]

For $A \ll \mu$ we have, for $c\mu(\Delta_{ij}) < 1$,

$$P_{ij} = \mathrm{tr}(\rho A(\Delta_{ij})) \leqslant \mathrm{tr}(\rho c\mu(\Delta_{ij})) = c\mu(\Delta_{ij});$$

so (14) becomes

$$S(\hat{A},\hat{B}\,|\rho) \geqslant \ln(\sup_{i,j} c\mu(\Delta_{ij}))^{-1}, \quad (15)$$

a state-independent lower bound. For a uniform partition $\mu(\Delta_{ij}) = \text{constant} = K$ of the instrument indicator space X, this becomes

$$S(\hat{A},\hat{B}\,|\rho) \geqslant \ln(cK)^{-1}. \quad (16)$$

For an example of (16), for the position–momentum joint POV with bins of size $\Delta X, \Delta P$, this reads, for $\Delta X\Delta P < 2\Pi\hbar$,

$$S(X,P\,|\rho) \geqslant \ln\,[2\Pi\hbar/\Delta X\Delta P\,]. \quad (17)$$

We see from (17) that for choices $\Delta X\Delta P$ chosen arbitrary small, $\ln(2\Pi\hbar/\Delta X\Delta P)$ becomes arbitrarily large. Likewise in (16) choosing K arbitrarily small.

For nonuniform partitions of the instrument indicator space, we could use $P_{ij} \leqslant c\mu(\Delta_{ij})$ in (6) to obtain the state-independent bound

$$S(\hat{A},\hat{B}\,|\rho) \geqslant -\Sigma_{i,j} c\mu(\Delta_{ij}) \ln(c\mu(\Delta_{ij})), \quad (18)$$

which does not seem to have appeared in the literature.

Turning to process (b), we shall optimize $S(\hat{A},\hat{B}\,|\rho)$ by variations in ρ. The general theory[16,17] for this process then leads to equations for ρ. [The corresponding case for $S(\hat{A}\,|\rho) + S(\hat{B}\,|\rho)$ has already been solved and analyzed.[13,15]] We shall treat only one simple case, one in which the ranges of $\hat{A}$ and $\hat{B}$ are both doubletons, namely (unsharp) spin 1/2 observables oriented in different directions.

V. AN EXAMPLE

Let $\vec{\sigma} = (\sigma_1,\sigma_2,\sigma_3)$ be a set of (Pauli) spin matrices and $\vec{x} \in \mathbb{R}^3$. Set $\vec{x}\cdot\vec{\sigma} = x_1\sigma_1 + x_2\sigma_2 + x_3\sigma_3$. Then for $\|\vec{x}\| = 1$, $T_x = \frac{1}{2}(1 + \vec{x}\cdot\vec{\sigma})$ is a projection operator having eigenstates of spin in direction $\vec{x}$. The covariant POV for such a system under the group of rotations of the sphere is given by

$$A(\Delta) = \int_\Delta T_x\, d\mu(x), \quad (19)$$

where μ is the rotation invariant measure on the Stokes sphere normalized by $\mu(\text{sphere}) = 2$. Hence

$$A(\Delta) = \tfrac{1}{2}(\mu(\Delta)1 + \vec{r}\cdot\vec{\sigma}),$$

where

$$\vec{r} = \int_\Delta \vec{x}\, d\mu(x). \quad (20)$$

Let us now cut the Stokes sphere into four regions by equatorial planes with normals $\vec{N}_1, \vec{N}_2$, respectively,

$\vec{N}_1 \cdot \vec{N}_2 = \cos\theta$. Label the four regions counterclockwise by $\Delta_1, \Delta_2, \Delta_3, \Delta_4$, with Δ_1, Δ_3 "the rind of watermelon slices" of angular opening θ, and Δ_2, Δ_4 of angular opening $\Pi - \theta$. By symmetry of the measure μ, then from (19), the $A(\Delta_i)$ are of the form

$$A(\Delta_i) = (\tfrac{1}{2})(\mu(A_i)\mathbf{1} + \vec{r}_i \cdot \vec{\sigma}), \tag{21a}$$

$$\vec{r}_1 \cdot \vec{r}_2 = 0, \quad \vec{r}_1 = -\vec{r}_3, \quad \vec{r}_2 = -\vec{r}_4, \tag{21b}$$

and by direct integration

$$\begin{aligned} |\vec{r}_1| &= \frac{1}{2}\sin\frac{\theta}{2}, \\ |\vec{r}_2| &= \frac{1}{2}\sin\left(\frac{\Pi - \theta}{2}\right) = \frac{1}{2}\cos\frac{\theta}{2}. \end{aligned} \tag{21c}$$

(This is most easily computed in the coordinate system in which $\hat{z}$ is in the direction $\hat{N}_1 \times \hat{N}_2$, and the $\hat{x}$ and $\hat{y}$ axes are in planes bisecting the angles formed by $\hat{N}_1, \hat{N}_2$.) By symmetry, one also has $\mu(\Delta_1) = \mu(\Delta_3), \mu(\Delta_2) = \mu(\Delta_4), \Sigma\mu(\Delta_i) = 2$, $\mu(\Delta_1)/\mu(\Delta_2) = \theta/(\Pi - \theta)$. Thus

$$\mu(\Delta_1) = \theta/\Pi,$$

$$\mu(\Delta_2) = (\Pi - \theta)/\Pi.$$

We therefore obtain the POV:

$$\left\{\begin{aligned} A(\Delta_1) &= \frac{\theta}{2\Pi}\mathbf{1} + \frac{1}{4}\sin\frac{\theta}{2}\hat{r}_1 \cdot \vec{\sigma} \\ A(\Delta_2) &= \frac{\Pi - \theta}{2\Pi}\mathbf{1} + \frac{1}{4}\cos\frac{\theta}{2}\hat{r}_2 \cdot \vec{\sigma} \\ A(\Delta_3) &= \frac{\theta}{2\Pi}\mathbf{1} - \frac{1}{4}\sin\frac{\theta}{2}\hat{r}_1 \cdot \vec{\sigma} \\ A(\Delta_4) &= \frac{\Pi - \theta}{2\Pi}\mathbf{1} - \frac{1}{4}\cos\frac{\theta}{2}\hat{r}_2 \cdot \vec{\sigma} \end{aligned}\right\}, \tag{22}$$

with $\vec{r}_1 \cdot \vec{r}_2 = 0$.

The marginal observables corresponding to adjacent slices are

$$\left\{\begin{aligned} E^{\hat{A}}(\text{up}) &= A(\Delta_1) + A(\Delta_2) = \tfrac{1}{2}(\mathbf{1} + \tfrac{1}{2}\hat{N}_1 \cdot \vec{\sigma}) \\ E^{\hat{A}}(\text{down}) &= A(\Delta_3) + A(\Delta_4) = \tfrac{1}{2}(\mathbf{1} - \tfrac{1}{2}\hat{N}_1 \cdot \vec{\sigma}) \\ E^{\hat{B}}(\text{up}) &= A(\Delta_2) + A(\Delta_3) = \tfrac{1}{2}(\mathbf{1} + \tfrac{1}{2}\hat{N}_2 \cdot \vec{\sigma}) \\ E^{\hat{B}}(\text{down}) &= A(\Delta_4) + A(\Delta_1) = \tfrac{1}{2}(\mathbf{1} - \tfrac{1}{2}\hat{N}_2 \cdot \vec{\sigma}) \end{aligned}\right\}, \tag{23}$$

the desired observables for measurement of (unsharp) spin in directions $\hat{N}_1, \hat{N}_2$, respectively.

We now optimize the entropy for the POV A. In general,[15] if f is any differentiable function of n real variables, if A_i are bounded, self-adjoint operators, if ρ is any quantum density, and

$$\begin{aligned} F(\rho) &= f(\langle A_1\rangle_\rho, \ldots, \langle A_n\rangle_\rho), \\ \langle A_i\rangle_\rho &= \mathrm{Tr}(\rho A_i), \end{aligned} \tag{24}$$

then F is locally extremal only on states satisfying

$$\sum_{i=1}^{n} \partial_i f(\langle A_1\rangle_\rho, \ldots, \langle A_n\rangle_\rho)(A_i - \langle A_i\rangle_\rho \mathbf{1})\rho = 0. \tag{25}$$

Furthermore, the following must also hold:

$$\begin{aligned} &\sum_{i=1}^{n} \partial_i f(\langle A_1\rangle_\rho, \ldots, \langle A_n\rangle_\rho)\langle (A_k - \langle A_k\rangle_\rho \mathbf{1}) \\ &\quad \times (A_i - \langle A_i\rangle_\rho \mathbf{1})\rangle_\rho = 0, \end{aligned} \tag{26}$$

where either all $\partial_i f$ vanish or else

$$\det(\langle (A_k - \langle A_k\rangle_\rho \mathbf{1})(A_i - \langle A_i\rangle_\rho \mathbf{1})\rangle_\rho) = 0. \tag{27}$$

Extrema of F may also occur at the extrema of the numerical range of the $A_1, \ldots, A_n$.

For us here, $A_i = A(\Delta_i)$, and

$$F(\rho) = S(\hat{A}, \hat{B}\,|\rho) = -\sum_{k=1}^{4} P_k \ln P_k, \tag{28}$$

where we have abbreviated $P_k = \langle A(\Delta_k)\rangle_\rho$. Thus local extrema may occur under the condition

$$\sum_{k=1}^{4} (1 + \ln P_k)[A(\Delta_k) - P_k \mathbf{1}]\rho = 0. \tag{29a}$$

Since $\Sigma_{k=1}^4 A(\Delta_k) = \mathbf{1}$, then $\Sigma_{k=1}^4 P_k = 1$, and (29a) reduces to

$$\left[\sum_{k=1}^{4} (\ln P_k)(A(\Delta_k) - P_k \mathbf{1})\right]\rho = 0. \tag{29b}$$

As pointed out in Ref. 15, this is equivalent to solving

$$\sum_{k=1}^{4} [\, b_k A(\Delta_k) - \lambda \mathbf{1}\,]\rho = 0, \tag{30}$$

where $\{b_k\}$ and λ are constants, since these constants are then determined as functions of the covariances of the $A(\Delta_k)$ in state ρ as long as the determinant of the covariance matrix vanishes. Alternatively, we can check that

$$b_k = -\ln P_k, \quad \lambda = -\sum P_k \ln P_k, \tag{31}$$

where

$$P_k = \mathrm{Tr}(A(\Delta_k)\rho). \tag{32}$$

From (22), (30) becomes

$$[c_0\mathbf{1} + c_1\sigma_1 + c_2\sigma_2]\rho = 0, \tag{33}$$

where

$$c_0 = (b_1 - b_2 + b_3 - b_4)(\theta/2\Pi) + \tfrac{1}{2}(b_2 + b_4) - \lambda, \tag{34a}$$

$$c_1 = \tfrac{1}{4}(b_1 - b_3)\sin(\theta/2), \tag{34b}$$

$$c_2 = \tfrac{1}{4}(b_2 - b_4)\cos(\theta/2), \tag{34c}$$

$$\sigma_1 = \hat{r}_1 \cdot \vec{\sigma}, \quad \sigma_2 = \hat{r}_2 \cdot \vec{\sigma}. \tag{35}$$

Solutions to (33) are given by[17]

$$\rho = \tfrac{1}{2}\left[\mathbf{1} - \frac{c_1}{c_0}\sigma_1 - \frac{c_2}{c_0}\sigma_2\right] \tag{36}$$

under the necessary condition

$$c_0^2 = c_1^2 + c_2^2, \tag{37}$$

since under this constraint, (33) reads

$$\left[\mathbf{1} \pm (c_1/\sqrt{c_1^2 + c_2^2})\sigma_1 + (c_2/\sqrt{c_1^2 + c_2^2})\sigma_2\right]\rho = 0. \tag{33'}$$

This is equivalent to

$$T_{\pm z}\rho = 0, \tag{33''}$$

where

$$\hat{z} = (c_1/\sqrt{c_1^2 + c_2^2})\hat{r}_1 + (c_2/\sqrt{c_1^2 + c_2^2})\hat{r}_2;$$

so

$$\rho = T_{\mp z}. \tag{36'}$$

Thus local optimization occurs on a pure state.

From (22), (31), (32), and (36) we have

$$
\begin{aligned}
e^{-b_1} &= P_1 = \frac{\theta}{2\Pi} \mp \frac{c_1}{4c_0} \sin\frac{\theta}{2}, \\
e^{-b_2} &= P_2 = \frac{(\Pi-\theta)}{2\Pi} \mp \frac{c_2}{4c_0} \cos\frac{\theta}{2}, \\
e^{-b_3} &= P_3 = \frac{\theta}{2\Pi} \pm \frac{c_1}{4c_0} \sin\frac{\theta}{2}, \\
e^{-b_4} &= P_4 = \frac{(\Pi-\theta)}{2\Pi} \pm \frac{c_2}{4c_0} \cos\frac{\theta}{2},
\end{aligned}
\tag{38}
$$

which along with (31), (34), and (37) are to be solved simultaneously. Division of the first and third, and of the second and fourth equations in (38), and using (34) decouples these equations modulo equation (37). These two equations, after change of variables, may be put in the form

$$\exp(c_i t) = (1+t)/(1-t), \quad i = 1,2,$$

or equivalently

$$\tan^2\psi = \exp(-c_i \cos(2\psi)).$$

Equation (37) constrains the two resulting t_i to lie on an ellipse, but the problem is now susceptible to numerical solution, given θ.

Alternatively, we may use the procedure outlined in Ref. 16, namely, multiply (30) by $\{A(\Delta_i) - P_i\mathbf{1}\}$ and take the trace. This yields four equations in b_k with coefficients given by covariances:

$$C_{ij} b_i = 0,$$

$$C_{ij} = \mathrm{Tr}(\{A(\Delta_i) - P_i\mathbf{1}\}\{A(\Delta_j) - P_j\mathbf{1}\}\rho).$$

From (33), (34) we see that C_{ij} involves the b_k; the resulting equations are again transcendental.

In the case $\theta = \Pi/2$ by symmetry $c_1 = \pm c_2$ and z becomes simply

$$\hat{z} = \pm (1/\sqrt{2})(\hat{r}_1 \pm \hat{r}_2). \tag{33'''}$$

From this everything can be computed easily, for this special case.

The remaining extremal solutions occur on states corresponding to the boundary of the numerical range of the $A(\Delta_i)$. Now

$$
\begin{aligned}
A(\Delta_1) &= \frac{\theta}{2\Pi}\mathbf{1} + \frac{1}{4}\sin\frac{\theta}{2}\vec{r}_1\cdot\vec{\sigma} \\
&= \frac{\theta}{2\Pi}(T_{r_1} + T_{-r_1}) + \frac{1}{4}\sin\frac{\theta}{2}(T_{r_1} - T_{-r_1}) \\
&= \left[\frac{\theta}{2\Pi} + \frac{1}{4}\sin\frac{\theta}{2}\right] T_{r_1} \\
&\quad + \left[\frac{\theta}{2\Pi} - \frac{1}{4}\sin\frac{\theta}{2}\right] T_{-r_1}.
\end{aligned}
$$

Since any state ρ may be written in the form

$$\rho = \tfrac{1}{2}(1 + \vec{z}\cdot\vec{\sigma}), \quad \|\vec{z}\| \leqslant 1,$$

and, for $T_{\hat{u}}$ a projection,

$$
\begin{aligned}
\mathrm{Tr}(\rho T_{\hat{u}}) &= \mathrm{Tr}(T_{\hat{u}}\rho T_{\hat{u}}) = \tfrac{1}{2}(1 + \hat{z}\cdot\hat{u})\mathrm{Tr}(T_{\hat{u}}) \\
&= \tfrac{1}{2}(1 + \vec{z}\cdot\hat{u}),
\end{aligned}
$$

we obtain

$$
\begin{aligned}
\langle A(\Delta_1)\rangle &= \left[\frac{\theta}{2\Pi} + \frac{1}{4}\left(\sin\frac{\theta}{2}\right)\right]\mathrm{Tr}(\rho T_{r_1}) \\
&\quad + \left[\frac{\theta}{2\Pi} - \frac{1}{4}\left(\sin\frac{\theta}{2}\right)\right]\mathrm{Tr}(\rho T_{-r_1}) \\
&= \frac{\theta}{2\Pi} + \frac{1}{4}\left(\sin\frac{\theta}{2}\right)\vec{z}\cdot\hat{r}_1.
\end{aligned}
$$

Similarly,

$$\langle A(\Delta_2)\rangle\rho = \frac{\Pi-\theta}{2\Pi} + \frac{1}{4}\left(\cos\frac{\theta}{2}\right)\vec{z}\cdot\hat{r}_2,$$

$$\langle A(\Delta_3)\rangle\rho = \frac{\theta}{2\Pi} - \frac{1}{4}\left(\sin\frac{\theta}{2}\right)\vec{z}\cdot\hat{r}_1,$$

$$\langle A(\Delta_4)\rangle\rho = \frac{\Pi-\theta}{2\Pi} - \frac{1}{4}\left(\cos\frac{\theta}{2}\right)\vec{z}\cdot\hat{r}_2.$$

We may compute S as a function of $\vec{z}$ and optimize in the usual $\mathbb{R}^3$ form. For our present purposes, we see that the boundary of the numerical range of the $A(\Delta_i)$ is obtained only if $\|\vec{z}\| = 1$; i.e., ρ is again a pure state. Optimizing subject to this constraint leads to precisely the same transcendental equations obtained previously when one chooses a coordinate system such that $z_1 = \vec{z}\cdot\hat{r}_1$, $z_2 = \vec{z}\cdot\hat{r}_2$. (Recall $\hat{r}_1\cdot\hat{r}_2 = 0$.) In this system, z_3 must vanish in order to satisfy the optimization and the problem is a constrained two-dimensional problem, which we also leave for numerical computation once θ is given. The special case $\theta = \Pi/2$ can again be solved with ease.

ACKNOWLEDGMENTS

This research was done under a Deutsche Forschungsgemeinschaft Visiting Professorship. The author would like to thank the members of the Theoretical Physics Institute in Köln for the hospitality and collaboration which led to this and other research results.

[1] D. Deutsch, Phys. Rev. Lett. **50**, 631 (1983).
[2] M. H. Partovi, Phys. Rev. Lett. **50**, 1883 (1983); **54**, 373 (1985).
[3] F. E. Schroeck, Jr., J. Math. Phys. **22**, 2562 (1981); Found Phys. **12**, 79 (1982); **15**,677 (1985).
[4] P. Busch, J. Math. Phys. **25**, 1794 (1984); Int. J. Theor. Phys. **24**, 63 (1985); Phys. Rev. D **33**, 2253 (1986); Found Phys. **17**, 905 (1987).
[5] G. Ludwig, *Foundations of Quantum Mechanics* (Springer, New York, 1983, 1985).
[6] A. Prieur, Unscharfe Teilchen–Welle–Dualität, Diplomarbeit, Inst. für Theor. Phys. der Univ. zu Köln, 1987, §5.
[7] E. B. Davies and J. T. Lewis, Commun. Math. Phys. **17**, 239 (1970); E. B. Davies, J. Funct. Anal. **6**, 318 (1970).
[8] M. Ozawa, J. Math. Phys. **25**, 79 (1984).
[9] F. E. Schroeck, Jr., Int. J. Theor. Phys. **28**, 247 (1989).
[10] O. Abu-Zeid, Phys. Lett. A **125**, 162 (1987).
[11] See H. Everett in B. S. DeWitt and N. Graham, *The Many Worlds Interpretation of Quantum Mechanics* (Princeton U. P., Princeton, NJ, 1973), pp. 45–49.
[12] F. E. Schroeck, in *Symposium on the Foundations of Modern Physics*, edited by P. Lahti and P. Mittelstaedt (World Scientific, Singapore, 1985), pp. 573–590, Lemma 1.
[13] M. Grabowski, Rep. Math. Phys. **20**, 153 (1984).
[14] P. Busch and P. J. Lahti, J. Phys. A: Math. Gen. **20**, 899 (1987).
[15] R. S. Ingarden, Rep. Math. Phys. **10**, 43 (1976).
[16] P. Busch, T. P. Schonbek, and F. E. Schroeck, Jr., J. Math. Phys. **28**, 2866 (1987).
[17] F. E. Schroeck, Jr., in *Mathematical Foundations of Quantum Theory*, edited by A. R. Marlow (Academic, New York, 1978), pp. 299–327.

Bounds for the C^*-algebraic transition probability yield best lower and upper bounds to the overlap

Peter M. Alberti and Volker Heinemann
Karl-Marx-Universitat Leipzig, Sektion Mathematik und Naturwissenschaftlich-Theoretisches Zentrum, DDR-7010 Leipzig, Karl-Marx-Platz 10, German Democratic Republic

(Received 4 January 1989; accepted for publication 29 March 1989)

Bounds are proved for the C^*-algebraic transition probability $P_A(\omega,\nu)$ between the abstract ground state ν with respect to a symmetric subspace N of a unital C^* algebra A and a state ω with the restriction $\omega|N = \sigma|N$ to N for an arbitrarily given, but fixed state σ. A is assumed to be the unital C^*-algebra generated by N. The results are specified in the case where A is a subalgebra of a vN algebra in standard form and N is dimensionally finite. Under these assumptions, the relationships of the algebraic transition probability to the notion of the (square of the) overlap integral known in quantum physics are clearly established. The general results are used to treat the standard problem of finding upper and lower bounds to the overlap in a quantum mechanical context. The best bounds are found and their properties discussed.

I. INTRODUCTION

Let A be a C^* algebra with unit 1 and topological dual space A^* and a set of states $S(A) = \{\omega \in A^*: \omega(x^*x) \geqslant 0, \forall x \in A, \omega(1) = 1\}$. There exists a generalization of the notion of the quantum mechanical transition probability $|(\Psi,\Phi)|^2$ between the state vectors $\Psi, \Phi \in H$ to a general situation with two mixed states over some C^* algebra of observables [H is the Hilbert space of the quantum system considered and (,) is the scalar product; scalar products are linear in the second argument throughout this paper].

The C^*-algebraic transition probability in question was proposed by Uhlmann[1] and its definition is as follows. Let $\{\pi,H\}$ be a unital * representation of A on a Hilbert space H. For $\omega \in S(A)$ we define a set $S(\pi,\omega)$ as $S(\pi,\omega) = \{\Psi \in H: \omega(x) = (\Psi,\pi(x)\Psi), \forall x \in A\}$. The representation π is called ω,σ admissible if both $S(\pi,\omega)$ and $S(\pi,\sigma)$ are non void. The C^*-algebraic transition probability $P_A(\omega,\sigma)$ between ω, $\sigma \in S(A)$ is now given by

$$P_A(\omega,\sigma) = \sup\{|(\Psi,\Phi)|^2 : \Psi \in S(\pi,\omega), \Phi \in S(\pi,\sigma), \omega,\sigma \text{ admissible } \pi\}.$$

Here P is a mathematically well-investigated object; many important properties of the functor P are known. Especially, for the vector states $\omega_\Psi = (\Psi,(\cdot)\Psi)$ and $\omega_\Phi = (\Phi,(\cdot)\Phi)$ over $A = B(H)$ (the algebra of bounded linear operators over the Hilbert space H), one finds that $P_{B(H)}(\omega_\Psi,\omega_\Phi) = |(\Psi,\Phi)|^2$, where the latter gives support to a heuristic interpretation of P as a "generalized" transition probability if seen in the context of the algebraic approach to quantum statistical physics and quantum field theory. We note that P is only one of the possible generalizations; however, the functor P plays some distinguished role among all the other possible functor provided certain additional (but sound) conditions are imposed on a transition probability, cf. Refs. 2 and 3. For a short survey on generalized transition probabilities and the problems related to them we refer to Ref. 4 and the references therein.

The aim of this article is twofold. First, we shall derive some new formulas for P in a particular situation of states over a C^* algebra. Second, in specializing the context, the formulas derived are used in favor of thoroughly analyzing a standard exercise in quantum mechanics once again, although from a possibly new point of view. The class of problems under discussion is described best by one of the characteristic questions or exercises to be raised in this field: Give an evaluation of the accuracy with which a trial function Ψ approximates the wavefunction Φ (i.e., an eigenfunction to a nondegenerate eigenvalue of the Hamiltonian, e.g., the ground state) if only certain information on the system is available.

Characteristic examples of such information are the spectrum of the Hamiltonian (the energy spectrum, which is known by experience) and the expectation value and variance of the energy in the trial state (provided by a series of experiments when the system is prepared into the trial state). Note that the problem occurs since the true wavefunction Φ is not (exactly) known in many cases. Quantitatively, the overlap integral $S = |(\Psi,\Phi)|$ provides a criterion of accuracy. Hence, the problem is to find upper and lower bounds to S which read in terms of the information available. In this context, it is also of interest whether or not the possibly derived bounds appear to be the best ones under the suitable conditions to be imposed, etc. At this moment, let us be content with giving these few explanations. We no longer discuss the physical context; instead, for more details on this field of application the reader is referred to a quite detailed survey paper by Weinhold[5] and the textbook of Thirring[6] (Sec. 3.5 of Ref. 6). Especially, the references listed in Ref. 5 provide a good source to the background of the problem before 1970. The paper,[5] the textbook,[6] and a more recent research paper[7] demonstrate characteristic methods of attacking this interesting problem and provide some feeling for the widespread tools common in this field. The question as to whether or not generalized transition probabilities could provide some new insights into the accuracy problem was first raised by Thirring[8] in 1983 and then was forgotten for almost five years. In some sense; the applicative part of this paper might be taken as an answer (very late, indeed) to this question.

 0022-2488/89/092083-07$02.50

II. SOME FACTS ON GENERALIZED TRANSITION PROBABILITIES

In this section, we start placing at our disposal some facts about P which will be widely used throughout this investigation. Assume ω, $\sigma \in S(A)$ and $\{\pi,H\}$ to be an ω,σ-admissible * representation of the C^* algebra A over the Hilbert space H.

(i) For any $\Phi \in S(\pi,\sigma)$ we have $P_A(\omega,\sigma) = \sup\{|(\Psi,\Phi)|^2: \Psi \in S(\pi,\omega)\}$.

(ii) $P_A(\omega,\sigma) = \inf\{\omega(x)\sigma(x^{-1}): x \in A,\ x \geqslant 0,$ invertible$\}$.

(iii) When $B \subset A$ is a unital C^* subalgebra of A we find that $P_B(\omega|B,\sigma|B) \geqslant P_A(\omega,\sigma)$ [$\tau|B$ is the restriction of $\tau \in S(A)$ onto B].

(iv) When $u: A \to B$ is a * isomorphism of A onto the C^* algebra B and if ω', σ' are those states in $S(B)$ such that $\omega = \omega' \cdot u$ and $\sigma = \sigma' \cdot u$, we have $P_A(\omega,\sigma) = P_B(\omega',\sigma')$.

(v) Suppose A is a C^* algebra of bounded linear operators over some Hilbert space K. Assume ω, σ are states that have the normal extensions ω', σ' onto the vN algebra M generated by A, $M = A''$ [double commutant within $B(K)$]. Then $P_M(\omega',\sigma') = P_A(\omega,\sigma)$.

(vi) Let $p \in A$ be a minimal orthoprojection, i.e., $pAp = \mathbb{C}p$. Let $pxp = \nu(x)p$, with $\nu(x) \in \mathbb{C}$. Then the map $\nu: x \to \nu(x) \in \mathbb{C}$ defines a state $\nu \in S(A)$ with $\nu(p) = 1$ and we have $P_A(\omega,\nu) = \omega(p)$ for any state $\omega \in S(A)$.

The crucial assertions are (i) and (ii) and the remaining ones are more or less direct consequences of them. For proofs we refer to Corollary 1, Corollary 2, and Theorem 3 of Ref. 9. Especially, (vi) is a special case of a situation dealt with in Ref. 9 [cf. Eq. (8) of Ref. 9]. Since (vi) is a key result for all that follows, we shall include a derivation of this useful fact.

Proof of (vi): Suppose A,p, and ν as in the premises of (vi) and let $x_n \in A$ be defined as $x_n = p + np^{\perp}$ for any natural n. Then x_n is positive and invertible for any n. Hence, (ii) tells us that $P_A(\omega,\nu) \leqslant \omega(x_n^{-1})\nu(x_n) = \omega(x_n^{-1}) = \omega(p) + (\omega(p^{\perp})/n)$, $\forall n$, i.e., $P_A(\omega,\nu) \leqslant \omega(p)$ follows. Especially, whenever $\omega(p) = 0$, $P_A(\omega,\nu) = \omega(p) = 0$ also follows. In order to prove (vi), what remains is to show that $P_A(\omega,\nu) \geqslant \omega(p)$ in the case of $\omega(p) \neq 0$. Assume $\omega(p) \neq 0$ and let $\{\pi,H,\Omega\}$ be the ω–GNS (Gelfand–Naimark–Segal) representation of A, i.e., a cyclic representation π with $\pi(A)\Omega$ dense in H and $\omega(x) = (\Omega,\pi(x)\Omega)$, $\forall x \in A$. We define Φ as $\Phi = \omega(p)^{-1/2}\pi(p)\Omega$. Then Φ is a unit vector of H and $(\Phi,\pi(x)\Phi) = \omega(p)^{-1}(\Omega,\pi(pxp)\Omega) = \omega(p)^{-1}\omega(pxp) = \omega(p)^{-1}\nu(x)\omega(p) = \nu(x)$ for $\forall x \in A$. Hence, $\Phi \in S(\pi,\nu)$. Since $\Omega \in S(\pi,\omega)$, (i) tells us that $P_A(\omega,\nu) \geqslant |(\Omega,\Phi)|^2 = |(\Omega,\pi(p)\Omega)\omega(p)^{-1/2}|^2 = |\omega(p)^{1/2}|^2 = \omega(p)$ and the proof is complete. □

III. ALGEBRAIC CONSIDERATIONS

In this section we begin setting up the kinds of problems upon which the explained application will be based. Let N be a symmetric subspace of our unital C^* algebra A, i.e., whenever $x \in N$, $x^* \in N$ also. Henceforth it is assumed that A is generated by N and $\mathbb{1}$ as a C^* algebra, i.e., $A = C^*(\{\mathbb{1},N\})$ is supposed. In case A is a C^* algebra of bounded linear operators over some Hilbert space H, the enveloping vN algebra M of A, which is $M = A'' = \{N\}'' \subset B(H)$, will also be of interest. A state $\nu \in S(A)$ is said to be a ground state (of N) if $\nu(x^*x) = 0$, $\forall x \in N$ and in case we are over some Hilbert space a ground state ν is called normal ground state if ν has a normal extension onto M. In the latter case, the (uniquely determined) normal extension tacitly will also be named ν.

Lemma 3.1: For a given linear subspace of A there exists at most one ground state ν. If a ground state exists it is a multiplicative state, i.e., $\nu(xy) = \nu(x)\nu(y)$, $\forall x, y \in A$.

Proof: Suppose a ground state ν to exist. Then $I_\nu = \{x \in A: \nu(x^*x) = 0\}$ is a uniformly closed left ideal in A. Let $F(N)$ be the algebra of finite, complex linear combinations of products of finitely many elements of N. Clearly, the subset $B = \mathbb{C}\mathbb{1} + F(N)$ is a unital * subalgebra of A being uniformly dense in A. Moreover, $F(N) \subset I_\nu$. Let $x \in A$ and let $\lambda_n \in \mathbb{C}$, $y_n \in F(N)$ be chosen in such a way that $x = \lim_n \times(\lambda_n \mathbb{1} + y_n)$. Then $\nu(x) = \lim_n\{\lambda_n + \nu(y_n)\} = \lim_n \lambda_n$ since $y_n \in F(N)$ and ν is a ground state. Hence, $\{y_n\}$ is a sequence of elements with the limit $y = \lim_n (x_n - \lambda_n \mathbb{1}) = x - \nu(x)\mathbb{1}$.

Therefore, $A = \mathbb{C}\mathbb{1} + F(N)_{\mathrm{cl}}$, with $F(N)_{\mathrm{cl}} =$ norm closure of $F(N)$, and each $x \in A$ decomposes as $x = \nu(x)\mathbb{1} + y$ for some $y \in F(N)_{\mathrm{cl}}$. Suppose $y \in I_\nu$; $y = \nu(y)\mathbb{1} + y'$ is a decomposition of y, as previously explained. Since $I_\nu \subset \ker(\nu)$, $y = y' \in F(N)_{\mathrm{cl}}$, i.e., $I_\nu \subset F(N)_{\mathrm{cl}}$. According to the above mentioned $F(N) \subset I_\nu$, I_ν being closed gives $F(N)_{\mathrm{cl}} \subset I_\nu$. Hence, $I_\nu = F(N)_{\mathrm{cl}}$. The latter also has to hold for any possibly existing ground state ν'; hence, $I_\nu = I_{\nu'}$. For $x \in A$, which decomposes as $x = \nu(x)\mathbb{1} + y$ for some $y \in I_\nu$, we thus obtain $\nu'(x) = \nu(x) + \nu'(y) = \nu(x)$ since $y \in I_{\nu'}$ as well. The latter is true for any $x \in A$; hence, $\nu = \nu'$ has to be followed. Thus the ground state is unique if it exists. Finally, since I_ν is symmetric, it is even a two-sided ideal in A. According to $A = \mathbb{C}\mathbb{1} + I_\nu$, we have $\mathbb{C} \cong A/I_\nu$ and since $I_\nu = \ker(\nu)$ the multiplicative behavior finally becomes obvious. □

Let us now raise the main question of this paper. Suppose N has a ground state ν (which then is unique by Lemma 4.1) and assume another state $\sigma \in S(A)$ is given. What can be said about the transition probability $P_A(\omega,\nu)$ between ν and another state ω if it is known to us only that $\omega|N = \sigma|N$? Obviously, the problem formulated is an algebraic caricature of the sort of questions we have discussed briefly in Sec. II in a quantum mechanical context. A first answer to the question is given in the following form.

Theorem 3.2: Let ν be the ground state of N. For any $\sigma \in S(A)$ the following formula holds:

$$\sup\{P_A(\omega,\nu): \omega \in S(A),\ \omega|N = \sigma|N\} = \inf\{1 + \sigma(x): x \geqslant -\mathbb{1}, x \in N\}. \tag{1}$$

Proof: From the proof of Lemma 4.1 we know that $A = \mathbb{C}\mathbb{1} + F(N)_{\mathrm{cl}}$. Let $\{\pi,H\}$ be the universal representation of A over the Hilbert space H. Denote by ν' the unique normal state over the universal enveloping vN algebra $A^{**} = \pi(A)''$ such that $\nu' \cdot \pi = \nu$. Obviously, $\nu'|\pi(A)$ is the ground state for $\pi(N)$ over $\pi(A)$. Let p be the support projection of ν' within A^{**}. Since π is an * isometry, we have $\pi(A) = \mathbb{C}\mathbb{1} + F(\pi(N))_{\mathrm{cl}}$. Let F^{**} be the closure of $F(\pi(N))$ with respect to the weak or strong operator topology within

$B(H)$. Then it is easily seen that $A^{**} = \mathbb{C}\mathbf{1} + F^{**}$ and every $x \in A^{**}$ decomposes as $x = \nu'(x)\mathbf{1} + y$ with $y \in F^{**}$. Since $\nu'(x^*x) = 0$ for any $x \in \pi(N)$, $px^*xp = 0$ has to hold. Since $\pi(N)$ is symmetric, $px = xp = 0$, $\forall x \in \pi(N)$. The latter implies that $pF(\pi(N)) = F(\pi(N))p = \{0\}$. By closure we have $pF^{**} = F^{**}p = \{0\}$; hence, p is a central projection of A^{**} and A^{**} decomposes as $A^{**} = \mathbb{C}p + A^{**}p^{\perp}$, i.e.,

$$A^{**} = \mathbb{C}p + F^{**}, \tag{2}$$

where p is a minimal projection of A^{**} with $pxp = \nu'(x)p$, $\forall x \in A^{**}$. We are allowed to apply (vi), thus obtaining

$$P_{A^{**}}(\tau,\nu') = \tau(p), \ \forall\, \tau \in S(A^{**}). \tag{3}$$

Let us look at the symmetric linear subspace R of $\pi(A) \subset A^{**}$ which is spanned by $\mathbf{1}$ and $\pi(N)$. Then, because $\mathbf{1} \in R$ is an inner point of the positive cone with respect to the Hermitian part of A^{**} and since $\sigma'|R$, with σ' the unique normal state of A^{**} with $\sigma' \cdot \pi = \sigma$, is a positive linear form on R, the well-known Krein extension theorem (cf., e.g., Ref. 10) for the extension of positive linear forms applies in our situation, with the result that

$$\sup\{\mu(p): \mu \in S(A^{**}), \mu|R = \sigma'|R\} = \inf\{\sigma'(x): p \leqslant x, x \in R\}. \tag{4}$$

Let $x = x^*$ be an element of R. Then $x = \lambda\mathbf{1} + y$ for some $\lambda \in \mathbb{R}$ and $y = y^*$, $y \in \pi(N)$. It is easily recognized that $p \leqslant \lambda\mathbf{1} + y$ iff $\lambda \geqslant 1$ and $0 \leqslant \lambda\mathbf{1} + y$, i.e., $p \leqslant \lambda\mathbf{1} + y$ iff $(y/\lambda) \geqslant -\mathbf{1}$ and $\lambda \geqslant 1$. Our conclusion is now as follows:

$$\begin{aligned}\inf\{\sigma'(x):\ & p \leqslant x, x \in R\} \\ &= \inf \lambda\{1 + \sigma'(x):\ x \in \pi(N), x \geqslant -\mathbf{1}, \lambda \geqslant 1\} \\ &= \inf\{1 + \sigma'(x):\ x \geqslant -\mathbf{1}, x \in \pi(N)\} \\ &= \inf\{1 + \sigma(y):\ y \in N, y \geqslant -\mathbf{1}\},\end{aligned} \tag{5}$$

where we used that π is an * isometry. Now, by (iii) and (iv) we infer for $\forall\, \tau \in S(A^{**})$ that

$$P_A(\tau \cdot \pi, \nu) = P_{\pi(A)}(\tau|\pi(A), \nu'|\pi(A)) \geqslant P_A^{**}(\tau,\nu'). \tag{6}$$

For each $\omega \in S(A)$ with $\omega|N = \sigma|N$ we have $\omega'|\pi(N) = \sigma'|\pi(N)$, with ω' the unique normal state on A^{**} such that $\omega = \omega' \cdot \pi$. On the other hand, for any $\tau \in S(A^{**})$ with $\tau|\pi(N) = \sigma'|\pi(N)$ we have $\tau \cdot \pi \in S(A)$ with $\tau \cdot \pi|N = \sigma|N$. Therefore, we conclude from (3)–(6) that,

$$\begin{aligned}\sup\{P_A(\omega,\nu):\ & \omega \in S(A), \omega|N = \sigma|N\} \\ &\geqslant \inf\{1 + \sigma(y): y \in N, y \geqslant -\mathbf{1}\}.\end{aligned} \tag{7}$$

Let $y \in N$, $y \geqslant -\mathbf{1}$. Then $x(\epsilon) = \mathbf{1} + (1 - \epsilon)y$ is positive and invertible in A for any ϵ with $0 < \epsilon < 1$. From Lemma 3.1 we learned that a ground state is multiplicative; hence, $\nu(x(\epsilon)^{-1}) = \nu(x(\epsilon))^{-1} = 1$ as a result of $\nu(y) = 0$. It follows that $\omega(x(\epsilon))\nu(x(\epsilon)^{-1}) = 1 + (1 - \epsilon)\omega(y)$, $\forall \epsilon$ with $0 < \epsilon < 1$. Hence, $\lim_{\epsilon \downarrow 0} \omega(x(\epsilon))\nu(x(\epsilon)^{-1}) = 1 + \omega(y)$ and (ii) proves that $P_A(\omega,\nu) \leqslant 1 + \omega(y)$. Since the latter is true for any $y \in N$, $y \geqslant -\mathbf{1}$, we infer $P_A(\omega,\nu) \leqslant \inf\{1 + \sigma(y): y \in N, y \geqslant -\mathbf{1}\}$ provided that $\omega|N = \sigma|N$ is fulfilled. Hence,

$$\begin{aligned}\sup\{P_A(\omega,\nu):\ & \omega \in S(A), \omega|N = \sigma|N\} \\ &\leqslant \inf\{1 + \sigma(y): y \in N, y \geqslant -\mathbf{1}\}.\end{aligned}$$

According to the above equation and (7) we see that (1) is true. □

Let us assume now that N is a subspace of bounded linear operators over some Hilbert space. Then let us adopt the additional notions and notations concerning such a situation, as outlined at the beginning of this section. In line with this, we assume that a ground state ν exists and we suppose ν is normal. Then a support of ν within M exists. We call this projection p. Now, the relationships between N, ν, p, A, $M = A''$ are exactly the same as those we have seen between $\pi(N)$, ν', p, $\pi(A)$, and A^{**} in the last proof. Hence, we might take the latter as a model. Then by literally translating all formulas into the new situation, the repetition of the arguments successfully applied in the last proof will also work in the new situation at hand. The result is that p appears to be a central projection of M which is minimal in M, i.e., $M = \mathbb{C}p + Mp^{\perp}$, and $Mp^{\perp}$ is the weak or strong closure of $F(N)$ in $B(H)$. Moreover, $pxp = \nu(x)p$ for $x \in M$ and (vi) becomes applicable. Thus extension arguments may be applied in the same manner and at the end an analogon of (4) is obtained by saying that

$$\begin{aligned}\sup\{P_M(\omega,\nu):\ & \omega \in S(M), \omega|N = \sigma|N\} \\ &= \inf\{1 + \sigma(y): y \geqslant -\mathbf{1}, y \in N\},\end{aligned} \tag{8}$$

where σ might be thought of as a fixed state of A as well as a state over M. This follows since $N \subset A \subset M$: only the restriction onto N of σ figures in all the relevant formulas. Finally, Eq. (8), in view of Theorem 3.2, results in the following Corollary.

Corollary 3.3: For a normal ground state ν we have that

$$\begin{aligned}\sup\{P_M(\omega,\nu):\ & \omega \in S(M), \omega|N = \sigma|N\} \\ &= \inf\{1 + \sigma(y): y \geqslant -\mathbf{1}, y \in N\} \\ &= \sup\{P_A(\omega,\nu):\ \omega \in S(A), \omega|N = \sigma|N\}, \\ &\quad \text{for any } \sigma \in S(M).\end{aligned}$$

Now we are going to establish an assertion which is in some sense complementary to the preceding one, i.e., our basic assumption that ν be a normal ground state of N with support p in M. Let us define the set $\Gamma(N) = \{y \in N:$ $\operatorname{spec}(y) \setminus \{0\} \subset [1, \infty)$, y has full support in $N\} \cup \{\mathbf{1}\}$, where the support of N is the minimal projection q of the vN algebra M such that $qy = y$ for any $y \in N$. Hence, $y \in N$ has full support iff the support of the individual y equals this q.

Corollary 3.4: For each $\sigma \in S(M)$ the following formula holds:

$$\begin{aligned}\inf\{P_M(\omega,\nu):\ & \omega \in S(M), \omega|N = \sigma|N\} \\ &= \sup\{1 - \sigma(y): y \in \Gamma(N)\}.\end{aligned} \tag{9}$$

Proof: Proceeding further in accordance with the remarks we made in preparing the arguments for Corollary 3.3, we note that the consequences of the extension theorem have not been used exhaustively. Until now, we have been using the information on the maximal value $\mu(p)$, which can be attained on the support p of ν by a linear form $\mu \in S(M)$ with $\mu|N = \sigma|N$. Another direction of application provides the minimal value a positive linear form, with the same restriction on N as σ can take on p. This information reads as

$$\begin{aligned}\inf\{\mu(p):\ & \mu \in S(M), \mu|N = \sigma|N\} \\ &= \sup\{\sigma(x):\ p \geqslant x, x \in R\},\end{aligned} \tag{10}$$

with $R=[\mathbf{1},N]$. Now, let $x=x^*\in R$ and $x=\lambda\mathbf{1}+y$, with $\lambda\in\mathbb{R}$ and $y=y^*\in N$. Then $x\leqslant p$ iff $\lambda\leqslant 1$ and $y\leqslant -\lambda p^{\perp}$. Suppose first that $\lambda\leqslant 0$. Then $\sigma(y)\leqslant|\lambda|$ and $\sigma(x)\leqslant\lambda+|\lambda|=0$ follows. Note that for $y\in N$ with $y\geqslant p^{\perp}$ we see $\lambda(1-y)\leqslant\lambda p\leqslant p$ for any $\lambda\in[0,1]$. Therefore, in the case where

$$\{y\in N:\ y \text{ of full support, } \operatorname{spec}(y)\setminus\{0\}\subset[1,\infty)\}$$

is non void we may write that $\sup\{\sigma(x):\ p\geqslant x, x\in R\}=\sup\lambda\{1-\sigma(y):\ y\geqslant p^{\perp},\ 0\leqslant\lambda\leqslant 1, y\in N\}=\sup\{1-\sigma(y): y\in\Gamma(N)\}$. Referring to the definition of $\Gamma(N)$ above, we remark that $p^{\perp}$ is the support of N in M [since $Mp^{\perp}=$ weak closure of $F(N)$]. We note that $\mathbf{1}$ has been included into the definition in order to deal with the case where the set $\{y\in N: y$ has full support, $\operatorname{spec}(y)\setminus\{0\}\subset[1,\infty)\}$ is void. In the latter case, $\sup\{\sigma(x):\ p\geqslant x,\ x\in R\}$ amounts to zero, necessarily; this case can now be included formally by realizing $0=1-\sigma(\mathbf{1})$. Taking into account all these facts and respecting (vi), (10) will imply the asserted equation (9). □

IV. MINIMAX FORMULAS FOR THE "TRANSITION" INTO A NORMAL GROUND STATE OVER A FINITE-DIMENSIONAL SUBSPACE

In this section we wish to derive some consequences of the results obtained in the particularly important case when N is of finite dimension. In doing so we prepare for the applications we have in mind. Throughout this section we shall suppose that $H_1,\ldots,H_n\in B(H)$ are bounded, self-adjoint linear operators ($\neq 0$) over the Hilbert space H. We suppose that $N=[\{H_k\}]$ has a normal ground state ν with the support projection p. We note that for μ to be a ground state for N in this case it is sufficient to have $\mu(H_k^2)=0$, $\forall k$. The normal state space over the vN algebra $M=\{H_1,\ldots,H_n\}''$ be $S_0(M)$. We define the numerical range Γ of states with respect to the N generating family of operators as $\Gamma=\{\mathbf{t}\in\mathbb{R}^n:\exists\sigma\in S(M)$ with $\sigma(H_k)=t_k,\forall k\}$. Moreover, we define a function f on Γ by $f(\mathbf{t})=\inf\{1+\Sigma r_k t_k:\ \Sigma r_k H_k\geqslant -1\}$. The set of all $\mathbf{r}\in\mathbb{R}^n$ such that $\Sigma r_k H_k\geqslant -1$ will be named Λ.

Proposition 4.1: For any $\mathbf{t}\in\Gamma$ we have

$$\inf_{\epsilon>0}\sup\{P_M(\omega,\nu):\omega\in S_0(M),|\omega(H_k)-t_k|\leqslant\epsilon\}=f(\mathbf{t}). \tag{11}$$

Proof: Let $\mathbf{t}\in\Gamma$ be given and let $\sigma\in S(M)$ be such that $t_k=\sigma(H_k)$ for all k. Suppose $\epsilon>0$ is fixed. Then we may choose $\mathbf{r}\in\Lambda$ and $\delta(\epsilon)>0$, with $\epsilon>\delta(\epsilon)$ such that

$$|s_k-t_k|\leqslant\delta(\epsilon),\ \forall k \text{ implies } 1+\Sigma r_k s_k\leqslant f(\mathbf{t})+\epsilon. \tag{12}$$

The set of all states τ such that $\tau(H_k)=t_k$ for any k is non void by assumption on $\mathbf{t}$ and w^* compact. Hence, we find such τ with $\tau(p)=\sup\{\omega(p):\omega|N=\sigma|N\}$. As a result of Corollary 3.3 and (vi) we have

$$\tau(p)=P_M(\tau,\nu)=f(\mathbf{t})=f(\tau(H_1),\ldots,\tau(H_n)). \tag{13}$$

For $a>0$, let $K_a=\{\mu\in S_0(M):\ |\mu(H_k)-t_k|\leqslant a\}$. The normal state space is w^* dense within all states. Thus there is $\omega\in S_0(M)$ such that

$$|\tau(p)-\omega(p)|\leqslant\epsilon,\ \omega\in K_{\delta(\epsilon)}. \tag{14}$$

For $\mathbf{l}\in\Lambda$ we always have $p\leqslant\mathbf{1}+\Sigma l_k H_k$. Hence, by (12),

$$\alpha(p)\leqslant 1+\Sigma r_k\alpha(H_k)\leqslant f(\mathbf{t})+\epsilon,\ \forall\alpha\in K_{\delta(\epsilon)}. \tag{15}$$

In view of (13)–(15) we see that

$$f(\mathbf{t})-\epsilon\leqslant\tau(p)-\epsilon\leqslant\omega(p)\leqslant\sup\{\alpha(p):\alpha\in K_{\delta(\epsilon)}\}\leqslant f(\mathbf{t})+\epsilon. \tag{16}$$

We define a function z over $\mathbb{R}_+\setminus\{0\}$ by $z(r)=\sup\{\alpha(p):\ \alpha\in K_r\}$. Obviously, z decreases when r is tending to zero. From (16) we infer $|f(\mathbf{t})-z(\delta(\epsilon))|\leqslant 2\epsilon$ and since for $\epsilon\downarrow 0$, $\delta(\epsilon)\downarrow 0$, $f(\mathbf{t})=\inf_{\epsilon>0}z(\epsilon)$ has to be followed. However, the latter, with regard to (vi) and the definition of z, yields (11). □

Now, let a set Δ be given by

$$\Delta=\{\mathbf{r}\in\mathbb{R}^n:\ \Sigma r_k H_k \text{ has full support, } \operatorname{spec}(\Sigma r_k H_k)\setminus\{0\}\subset[1,\infty)\}.$$

In the case where Δ is non void, we define a function u on Γ by setting $u(\mathbf{t})=\sup\{1-\Sigma t_k r_k:\ \mathbf{r}\in\Delta\}$ for any $\mathbf{t}\in\Gamma$ and let $g=\max\{u,0\}$. In the case where Δ is void, $g=0$ is set on Γ. Then for $\sigma\in S(M)$, we see that

$$\sup\{1-\sigma(y):\ y\in\Gamma(N)\}=g(\mathbf{t}),\ \text{with } t_k=\sigma(H_k),\ \forall k. \tag{17}$$

Proposition 4.2: For any $\mathbf{t}\in\Gamma$ we have

$$\sup_{\epsilon>0}\inf\{P_M(\omega,\nu):\ \omega\in S_0(M),|\omega(H_k)-t_k|\leqslant\epsilon\}=g(\mathbf{t}). \tag{18}$$

Proof: In the case where $g=0$ the assertion follows from (17) and (9). Thus we may suppose henceforth that Δ is nonvoid. We will treat the case of $\mathbf{t}\in\Gamma$ with $u(\mathbf{t})>0$. By Corollary 3.4 and (vi) we have

$$\inf\{P_M(\omega,\nu):\ \omega\in S(M),\omega|N=\sigma|N\}=\inf\{\omega(p):\ \omega\in S(M),\ \omega|N=\sigma|N\}=u(\mathbf{t}).$$

Let $\epsilon>0$ be fixed. We can choose $\delta(\epsilon)>0$ and $\mathbf{r}\in\Delta$ such that $\epsilon>\delta(\epsilon)$ and $1-\Sigma r_k s_k\geqslant u(\mathbf{t})-\epsilon$ for any $\mathbf{s}$ with $|s_k-t_k|\leqslant\delta(\epsilon)$. Let τ be a state such that $\tau(p)=u(\mathbf{t})$ with $t_k=\tau(H_k)$, $\forall k$. Take a normal state ω such that $\omega\in K_{\delta(\epsilon)}$ (whose set is defined as in the proof of Proposition 4.1) and $|\omega(p)-\tau(p)|\leqslant\delta(\epsilon)$. It follows that $\omega(p)\geqslant u(\mathbf{s})\geqslant u(\mathbf{t})-\epsilon$ with $s_k=\omega(H_k)$. On the other hand, $\mu(p)\geqslant u(\mu(H_1),\ldots,\mu(H_n))$ holds for every state μ by (9) and the fact that $\sup\{1-\mu(y):\ y\in\Gamma(N)\}\geqslant u(\mu(H_1),\ldots,\mu(H_n))$. Thus

$$u(\mathbf{t})+\epsilon=\tau(p)+\epsilon\geqslant\omega(p)\geqslant\inf\{\mu(p):\mu\in K_{\delta(\epsilon)}\}\geqslant u(\mathbf{t})-\epsilon. \tag{19}$$

Let $z(r)=\inf\{\mu(p):\ \mu\in K_r\}$. Here $z(r)$ increases when r tends to zero. Since (19) can be derived for any $\epsilon>0$ and $\delta(\epsilon)\downarrow 0$ can be derived for $\epsilon\downarrow 0$, we see that $\sup_{\epsilon>0}\inf\{\mu(p):\ \mu\in K_\epsilon\}=u(\mathbf{t})=g(\mathbf{t})$ in this case. According to (vi) this will imply (18). In the remaining case with $g(\mathbf{t})=0$, by the definition of g, Corollary 3.4, and (17) we have

$$\inf\{P_M(\omega,\nu):\ \omega\in S(M),\ \omega|N=\sigma|N\}=\inf\{\omega(p):\omega\in S(M),\omega|N=\sigma|N\}=0.$$

Hence, we find $\tau\in S(M)$ with $\tau(H_k)=t_k$ for any k such that $\tau(p)=0$. Since the normal state space is w^* dense in $S(M)$, obviously $\inf\{\mu(p):\ \mu\in K_\epsilon\}=0$. Thus according to (vi), (18) follows in this case as well. □

V. PREPARATIONS FOR A PHYSICALLY RELEVANT SITUATION: BOUNDS FOR THE TRANSITION INTO A JOINT EIGENSTATE

Throughout this section we suppose M_0 to be a vN algebra over a Hilbert space H, with a cyclic and separating vector, i.e., M_0 is assumed to be of standard form. Let $H_1,...,H_n \in M_0$ be self-adjoint operators such that $\Phi \in H$, $\|\Phi\| = 1$ exists with $H_k\Phi = 0$, $\forall k$. Let $\Gamma(\{H_k\})$ be the joint numerical range of the family $\{H_k\}$, i.e., $\{\mathbf{t} \in \mathbb{R}^n : \exists \Psi \in H$, with $\|\Psi\| = 1$, $t_k = (\Psi, H_k\Psi)$, $\forall k\}$. By the assumptions on M_0, the Γ of Sec. V is obviously the closure of $\Gamma(\{H_k\})$.

Subsequently, the expectation value of an operator T with respect to the state vector $\Psi \in H$, $(\Psi, T\Psi)$ will be abbreviated as $\langle T \rangle_\Psi$. Then we have the following results.

Theorem 5.1: For any $\mathbf{t} \in \Gamma(\{H_k\})$,

$$\inf_{\epsilon>0} \sup\{|(\Psi,\Phi)|^2 : \Psi \in H, \|\Psi\| = 1, |\langle H_k \rangle_\Psi - t_k| \leqslant \epsilon\} = f(\mathbf{t}), \tag{20}$$

with the function f defined on the numerical range of the family $\{H_k\}$ by

$$f(\mathbf{t}) = \inf\left\{1 + \sum r_k t_k : \sum r_k H_k \geqslant -\mathbb{1}\right\}. \tag{21}$$

Proof: Because of $M = \{H_1,...,H_n\}'' \subset M_0$ and since M_0 is in standard form, every normal state over M can be realized by some vector of H, i.e., we have $S(\mathrm{id},\omega) \neq \phi$ for any $\omega \in S_0(M)$ (where id is the identical representation of M). By (i) it follows that for a normal state ω, $P_M(\omega,\nu) = \sup\{|(\Psi,\Phi)|^2 : \Psi \in H$, with $(\Psi, x\Psi) = \omega(x)$, $\forall x \in M\}$ for the normal state ν defined by $\nu(x) = (\Phi, x\Phi)$, $x \in M$. This then implies that $\sup\{P_M(\omega,\nu) : \omega \in S_0(M), |\omega(H_k) - t_k| \leqslant \epsilon\} = \sup\{|(\Psi,\Phi)|^2 : \Psi \in H, \|\Psi\| = 1,$ with $|\langle H_k \rangle_\Psi - t_k| \leqslant \epsilon, \forall k\}$. Since, by assumption on Φ, ν is the ground state of $N = [H_1,...,H_n]$ and is normal on M, Proposition 4.1 can be applied and (20) and (21) follow. □

As usual, let p be the support projection of ν in M. Two cases may occur: p is either a one-dimensional orthoprojection [the dimension refers to the standard dimension within $B(H)$] or not. In the latter case there is $\Psi \perp \Phi$ with $\Psi \in pH$, $\|\Psi\| = 1$. Assume that Θ is a unit vector of H. We associate with Θ another unit vector $\Theta' = P^\perp\,\Theta + \|p\Theta\|\Psi$. By construction and according to our assumptions, we then have to take notice of the facts that $\langle H_k \rangle_{\Theta'} = \langle H_k \rangle_\Theta$, $\forall k$ and $(\Theta',\Phi) = 0$. Hence, in the situation with $\dim(p) > 1$ we have

$$\sup_{\epsilon>0} \inf\{|(\Psi,\Phi)|^2 : \Psi \in H, \|\Psi\| = 1, |\langle H_k \rangle_\Psi - t_k| \leqslant \epsilon\} = 0 \tag{22}$$

for any $\mathbf{t} \in \Gamma(\{H_k\})$.

Assume now $\dim(p) = 1$. Then by (vi) and according to the assumptions of this section (note that p is the orthoprojection onto the one-dimensional subspace $[\Phi]$),

$$P_M(\omega,\nu) = \omega(p) = |(\Psi,\Phi)|^2 \tag{23}$$

for any $\omega \in S_0(M)$, where $\Psi \in S(\mathrm{id},\omega)$.

Since ν is a normal ground state of N, (22), (23), and Proposition 5.2 can be taken together to find the following result.

Theorem 5.2: For any $\mathbf{t} \in \Gamma(\{H_k\})$ one has

$$\sup_{\epsilon>0} \inf\{|(\Psi,\Phi)|^2 : \Psi \in H, \|\Psi\| = 1, |\langle H_k \rangle_\Psi - t_k| \leqslant \epsilon\} = g(\mathbf{t}), \tag{24}$$

with the function g defined on the numerical range of the family of operators $\{H_k\}$ by

$$g(\mathbf{t}) = \max\{0, u(\mathbf{t})\}, \tag{25}$$

with $u = 0$ provided that the set

$$\Delta = \{\mathbf{r} \in \mathbb{R}^n : \mathrm{spec}(\Sigma r_k H_k) \setminus \{0\} \subset [1,\infty),\ \mathrm{supp}(\Sigma r_k H_k)\ \max\}$$

is void or $\dim(p) > 1$ and $u(\mathbf{t}) = \sup\{1 - \Sigma r_k t_k : \mathbf{r} \in \Delta\}$ otherwise.

Note that the definition of u and thus, also, of g in (25) differs slightly from that introduced in Sec. V.

Remark 5.3: Let $H_1,...,H_n$ be self-adjoint operators of a vN algebra in standard form and suppose $\exists \Phi \in H$, $\|\Phi\| = 1$, with $H_k\Phi = 0$, $\forall k$, i.e., Φ is a joint eigenvector to the eigenvalue 0 of all the given operators. Then Theorems 5.1 and 5.2 tell us something about the accuracy within which a state $\Psi \in H$, $\|\Psi\| = 1$ approximates the joint eigenvector Φ:

$$g(\langle H_1 \rangle_\Psi,...,\langle H_n \rangle_\Psi) \leqslant |(\Psi,\Phi)|^2 \leqslant f(\langle H_1 \rangle_\Psi,...,\langle H_n \rangle_\Psi). \tag{26}$$

The bounds are functions over the joint numerical range of the family $\{H_k\}$ and (26) arises from Theorems 5.1 and 5.2 by specializing $t_k = \langle H_k \rangle_\Psi$, $\forall k$. Moreover, Theorems 5.1 and 5.2, by their very structure, also say in which sense f, g are the best possible (global) bounds one can find. In fact, by definition, f (resp., g) is upper (resp., lower) semicontinuously depending on $\mathbf{t}$ with respect to the relative $\mathbb{R}^n$ topology induced on the joint numerical range. This is easily seen since f (resp., g) appears to be an infimum (resp., supremum) of continuous functions over $\Gamma(\{H_k\})$, which is clearly shown by (21) and (25).

Assume f' is another (global) upper bound, i.e.,

$$|(\Theta,\Phi)|^2 \leqslant f'(\mathbf{t}), \tag{27}$$

$\forall \Theta \in H$, $\|\Theta\| = 1$, with $t_k = \langle H_k \rangle_\Theta$, $\forall k$. Suppose $f' \leqslant f$ and let f' be upper semicontinuous. Then if there are $\mathbf{t}$ such that $f'(\mathbf{t}) < f(\mathbf{t})$, we could find $\delta > 0$ for given $\epsilon < f(\mathbf{t}) - f'(\mathbf{t})$, with

$$|s_k - t_k| < \delta \text{ implies } f'(\mathbf{s}) < f(\mathbf{t}) - \epsilon. \tag{28}$$

Then for some $\delta' < \delta$ we have

$$\sup\{|(\Theta,\phi)|^2 : \Theta \in H, \|\Theta\| = 1, |\langle H_k \rangle_\Theta - t_k| \leqslant \delta'\} \geqslant f(\mathbf{t}).$$

Hence, there are $\Theta \in H$, $\|\Theta\| = 1$ such that $|(\Theta,\Phi)|^2 \geqslant f(\mathbf{t}) - \epsilon$, with $|\langle H_k \rangle_\Theta - t_k| \leqslant \delta' < \delta$. Putting $S_k = \langle H_k \rangle_\Theta$, by (28) it follows that $f'(\mathbf{s}) < f(\mathbf{t}) - \epsilon \leqslant |(\Theta,\Phi)|^2$, which contradicts (27). Hence, $f' = f$ has to hold, necessarily, whenever upper semicontinuity for an upper bound is imposed in the sense explained. Analogously, we might argue for the case of a lower semicontinuous lower (global) bound.

Remark 5.4: Note that the results of this section apply via a simple modification to any common eigenvector of a family $\{H_k\}$ of self-adjoint operators. In fact, if $H_k\Phi = \lambda_k\Phi$, $\forall k$, then we might draw our conclusions for the family $\{H_k - \lambda_k \mathbb{1}\}$.

Remark 5.5: We remark that the difficult part in establishing the best (global) bounds f, g in explicit form is the exercise to find, for given $\{H_k\}$, the sets $\{r\in\mathbb{R}^n: \Sigma r_k H_k \geqslant -1\}$ and $\{r\in\mathbb{R}^n: \text{spec}(\Sigma r_k H_k)\setminus\{0\} \subset [1,\infty), \text{supp}(\Sigma r_k H_k) \text{ max}\}$, respectively. In fact, in most cases of practical importance we have little information on the relative geometry of the set $\{H_k\}$. Hence, in many cases only subsets of the two sets mentioned can be isolated, strongly depending on the availability of the information on the set $\{H_k\}$. Thus we have to be satisfied with "bounds for the best bounds" in any situations. However, we note that there are also important cases where f, g can be calculated exactly (cf. the example given in Sec. VII A).

VI. APPLICATIONS AND DISCUSSION: BOUNDS TO THE OVERLAP

In this section, we wish to apply the results of Sec. VI in the sense explained in Sec. II. First, let us discuss briefly the relevance in applications of the assumption that $H_1,...,H_n$ be operators of a vN algebra in standard form.

One point of view in the algebraic approach to quantum physics says that in order to describe mathematically certain aspects of the system under consideration, the main object to start and deal with is an appropriately chosen quantum dynamical system with an invariant state, i.e., we are given $\{A,G,\tau,\omega\}$, where A is a unital C^* algebra; G is a (locally compact) group; τ is a strongly continuous action of G as a group of * automorphisms on A; and ω is a faithful state on A which is invariant under the action of G via τ, i.e., $\omega\cdot\tau_g = \omega$ for any $g\in G$. Then usually one considers the ω–GNS construction $\{\pi,H,\Omega\}$. There is an implementation of τ by a strongly continuous group of unitary operators $\{u_g\}$ such that $\pi(\tau_g(x)) = u_g\pi(x)u_g^*$ for $\forall g\in G$, $x\in A$, and $u_g\Omega = \Omega$. Of course, the choice of the dynamical system one starts with depends on the aspects to be dealt with.

As an example, let G be a one-parameter group, say $G = \mathbb{R}$, and let τ be the time evolution, e.g., then the generator of $\{u_t\}\subset B(H)$ should be interpretable as the Hamiltonian of the system. Of course, the most desirable case is the one where $\{u_t\}$ belongs to the vN algebra $\pi(A)''$, which is of standard form. For physical reasons, the Hamiltonian (which is a self-adjoint, unbounded linear operator over H, in general) should have a spectrum bounded below (in order to assure stability of the system described), with a stationary state of smallest energy (i.e., the ground state should be an eigenstate). However, just imposing the boundedness of the energy spectrum from below assures that even $\{u_t\}\subset\pi(A)''$ for the implementation of τ could have been chosen, via the Borchers–Arveson theorem (cf. Ref. 11), and the Hamiltonian might be thought of as a positive, linear, self-adjoint operator which is affiliated to $\pi(A)''$ and is the generator of $\{u_t\}$. This might justify the belief that the vN algebra in standard form, $M_0 = \pi(A)''$, is of some relevance as an algebra of certain (bounded) observables associated to the system under consideration. Adopting such a point of view, the results Sec. VI evidently should become applicable within $\pi(A)''$ in a natural manner.

A. Example

As an example, let us assume h is the Hamiltonian. Here h is affiliated with M_0, and we suppose, as above, that $h\geqslant 0$ and $h\Omega = 0$, i.e., Ω is supposed as a ground state of h for simplicity. Suppose for the moment that h is bounded. Then $h\in M_0$. We assume now that $H_1 = h$ and $H_2 = h^2$. Ω yields a normal ground state for the linear space $[h,h^2]$ in the sense of Sec. VI. Hence, we might ask for f, g. We suppose the spectrum of h, spec(h) to be known. According to Sec. V, what we have to do is locate the sets $\Lambda = \{r\in\mathbb{R}^2: \Sigma r_k H_k \geqslant -1\}$ and $\Delta = \{r\in\mathbb{R}^2: \text{spec}(\Sigma r_k H_k)\setminus\{0\}\subset[1,\infty), \text{supp}(\Sigma r_k H_k) \text{ max}\}$, respectively.

It is easy to see that $r\in\mathbb{R}^2$ belongs to Λ iff $P(r,\lambda) = r_2\lambda^2 + r_1\lambda + 1\geqslant 0$ for any $\lambda\in$ spec (h). Now, spec $(h) = \mathbb{R}_+\setminus\cup I_n$ with $I_n = (u_n, o_n)$ for $n = 0,1,2,...,$ denoting the holes in the spectrum [we suppose $I_0 = (\|h\|,\infty)$]. Note that a hole in the spectrum is a maximal open interval of the resolvent set and there is at most a countable number of such holes in a spectrum. The above condition amounts to the fact that $r\in\Lambda$ iff either $P(r,(\cdot))\geqslant 0$ on the whole $\mathbb{R}_+$ (and then the polynomial has either no real root, a real double root, or both real roots are in the negative part of the reals), or both the real roots of the polynomial $P(r,(\cdot))$ belong to the same hole I_n. We omit the details of the calculations (which have been carried out in Ref. 12), but give instead the result for f.

$$f(t_1,t_2) = 1 - [(u_n + o_n)/u_n o_n]t_1 + (1/u_n o_n)t_2, \text{ for } (t_2/t_1)\in I_n, \tag{29}$$

$$f(t_1,t_2) = 1, \text{ for } t_1 = 0 \text{ (which occurs iff } t_2 = 0), \tag{30}$$

$$f(t_1,t_2) = 1 - (t_1^2/t_2), \text{ otherwise}, \tag{31}$$

with $t_1 = \langle h\rangle$, $t_2 = \langle h^2\rangle$.

Note that in case there is an index such that $u_n = 0$, then for $t_1\neq 0$, automatically $(t_2/t_1)\geqslant o_n$ [in fact, then zero is isolated and $\langle h^2\rangle\geqslant o_n\langle h\rangle$ since o_n has to be the smallest nonzero spectral value of spec (h) in this case]. Hence, this index n, if it exists, cannot interfere in the definition of f via (29). We remark that the function f' given by

$$f'(t_1,t_2) = 1, \text{ for } t_1 = 0 \text{ (which occurs iff } t_2 = 0), \tag{32}$$

$$f'(t_1,t_2) = 1 - (t_1^2/t_2), \text{ otherwise} \tag{33}$$

on the numerical range is an upper bound which is known as the Cambet–Farnoux–Allard bound (cf. Refs. 5 and 6 for the derivation and original references). In the sense of Remark 6.3, (29)–(31) give the best global upper bound for S^2 if we know the full spectrum. The bound given by (32) and (33) gives the best upper bound if we have no information on the spectrum (besides the fact zero is an eigenvalue). In between these two extreme cases we find intermediate cases of approximations from above of the best bound via bounds constructed by means of inserting only the partial information we have (say, e.g., we know some gaps or holes) in the sense of Remark 5.5. The bound f given by (29)–(31) seems to be new. We notice that f depends on the norm of h in quite a weak sense only (there exists a hole with the lower bound $\|h\|$ and an upper bound ∞). Therefore, the formulas for the

upper bound can be also used for a densely defined, positive, self-adjoint operator h having zero as an eigenvalue provided that the test functions are taken from the domain of h^2. This, then, corresponds to the case with which we started (where h is unbounded, but affiliated with M_0).

Concerning the lower bound g, we remark that Δ is void if zero is not isolated point of the spectrum. Hence, by definition of g, in this case $g = 0$. The latter also occurs if the eigenvalue zero is degenerated (cf. Theorem 6.2). Hence, we might hope for a nontrivial bound to exist only in the case where zero is an isolated spectral value which corresponds to a nondegenerated eigenvalue. Suppose this case and let E be the first spectral value of h that is different from zero (i.e., we have a gap of size E). According to the result of some of the calculations in Ref. 12 we obtain the best lower bound g (cf. Remark 6.5) as

$$g = \max\{0,u\}, \text{ with } u(t_1,t_2) = 1 - (t_1/E) + (t_2 - Et_1)/E\,\|h\|. \tag{34}$$

The structure of g shows that for the lower bound the knowledge of the spectrum is of relevance only insofar as we have to know the gap and spectral radius. Hence, the details in the spectrum do not enter at all. Therefore, this bound is also the best bound if only E and $\|h\|$ are known. Just as in the case of the upper bound, we might also draw some conclusion for the unbounded case: Then $g'(\mathbf{t}) = \max\{0,1 - (t_1/E)\}$ is the best one can do. The same result also holds if h is bounded, but nothing is known about the spectral radius. Note that since $t_2 \geqslant Et_1$ holds for any $\mathbf{t}$ of the numerical range, (34) shows that g is really better than the bound g'. The bound g' is well-known and was derived in 1930 by Eckart (cf. Refs. 5 and 7 for detailed references).

B. Remark

Note that the assumptions of Sec. V can be taken for granted in each case of a family $\{H_k\}$ of mutually commuting bounded operators over some Hilbert space H. In fact, such a family can be thought of as being part of a maximally Abelian vN algebra over H. Such an algebra possesses a cyclic vector in any case, which by the maximality condition is also separating: The case we dealt with in Sec. VI A numbers among this particular class. Hence, the discussion at the beginning of this section is of relevance when $\{H_k\}$ does not consist of mutually commuting operators exclusively. In such a case, the task of locating Δ and Λ seems to be a rather big challenge and for the most part one should aim to calculate approximations for the best bounds (cf. the discussion of Remark 5.5). In contrast, in the commutative case it seems possible to make an atlas of the exact bounds f, g for many cases provided that the H_k's are given as the functions $F_k(H_0)$ of some (maybe, auxiliary) operator H_0. Once again, Sec. VI A can serve as an example in this direction [$H_0 = h, F_1(x) = x$, and $F_2(x) = x^2$], at the same time indicating the line of action. Finally, we note that the case of densely defined, unbounded positive self-adjoint linear operators affiliated to some Abelian vN algebra might also be dealt with in this way.

VII. CONCLUDING REMARK

In Ref. 12, we have dealt with the special example found in Sec. VII A. Parameter-dependent bounds for S^2 in terms of $\langle h \rangle$ and $\langle h^2 \rangle$ have been derived in Sec. VI A by using estimations and tools based on the methods demonstrated in Ref. 5. These bounds were optimized by suitably fitting the parameters in question. The result of this process was shown in (29)–(31) and (34), respectively. Consequently, in the context of Ref. 12 the final bounds seem to be best only with respect to the underlying basic assumptions.

ACKNOWLEDGMENTS

We acknowledge gratefully the discussions (relating upper bounds) in the Leipzig Seminar on Mathematical Physics, particularly those with J. Friedrich (KMU) in the spring of 1988. The arguments discussed at that time gave some evidence as to why the upper bound of the example should nevertheless be the best possible (provided that the spectrum of h is known). The idea to look at the accuracy problem from the point of view of generalized transition probabilities was suggested by W. Thirring in 1983. However, the results derived on the generalized transition probability (cf. Sec. III and Propositions 4.1 and 4.2) are of interest on their own, independent of this special application. We wish to thank H. Grosse (Vienna) for giving us some hints as to the relevant literature.

[1] A. Uhlmann, Rep. Math. Phys. **9**, 273 (1976).
[2] A. Uhlmann, Ann. Phys. (Leipzig) **42**, 524 (1985).
[3] P. M. Alberti, Wiss. Z. Karl-Marx-Univ. Leipzig MNR **34**, 572 (1985).
[4] P. M. Alberti and A. Uhlmann, in *Proceedings of the International Conference on Operator Algebras, Ideals and their Applications in Theoretical Physics*, edited by H. Baumgaertel, G. Lassner, A. Pietsch, and A. Uhlmann (Teubner, Leipzig, 1984), pp. 5–11.
[5] F. Weinhold, J. Math. Phys. **11**, 2127 (1970).
[6] W. Thirring, *Lehrbuch der Mathematischen Physik* (Springer, Wien, New York, 1979), Vol. 3.
[7] T. Hoffmann-Ostenhoff, M. Hoffmann-Ostenhoff, and G. Olbricht, J. Phys. A **9**, 27 (1976).
[8] W. Thirring, private communicatoin to PMA (Leipzig, 1983).
[9] P. M. Alberti, Lett. Math. Phys. **7**, 25 (1983).
[10] M. A. Neumark, *Normierte Algebren* (VEB Deutscher, Berlin, 1959).
[11] O. Bratteli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics* (Springer, New York, 1979, 1981), Vols. 1 and 2.
[12] V. Heinemann, *Geutekriterien fuer approximierte Wellenfunktionen* (Diplomarbeit, Leipzig, 1988).

On the structure of spatial infinity. I. The Geroch structure

Piotr T. Chruściel[a)]
Physics Department, Yale University, New Haven, Connecticut 06511

(Received 12 January 1989; accepted for publication 5 April 1989)

Theorems on uniqueness and "quasi-uniqueness" of the differentiable structure of pointwise singular conformal structures are derived. This allows the classification of all ambiguities in the differentiable structure of conformally completed asymptotic three-dimensional ends.

I. INTRODUCTION

A considerable amount of work on the behavior of the gravitational field at spatial infinity has been done by several authors[1-5] in order to understand how to overcome the apparent lack of definitional uniqueness of four-momentum in general relativity. An interesting framework has been developed by Geroch,[2,6] who adapted the conformal completion technique used in the description of null infinity to the spatial setting, the main idea underlying the construction being to replace "coordinate conditions" by "local differential geometry." In Geroch's framework, the problem of a possibly ill-defined energy momentum reappears in the possibility of existence of inequivalent conformal completions of spatial infinity. In this paper we show that there exists only the well-known three-parameter family of inequivalent—in a sense to be made precise—completions of a three-dimensional asymptotically flat end. The proof we give is almost a filling-in of necessary technicalities in the proof outlined in the Appendix of Ref. 7, the (mild) difficulties arising from the fact that standard results of the PDE theory do not apply here in such a simple manner as in Ref. 7, as a result of the essentially singular character of the conformally rescaled metric at i_0. We shall use two sets of conditions to be satisfied by the conformal completions, the first under which uniqueness of the structure holds, the second in which the three-parameter (n parameter in n dimensions) family of inequivalent completions appears. The conditions we shall use are in spirit those of Geroch,[2,6] though our conditions are (slightly) stronger than those originally introduced. It would be interesting to find out whether the results obtained here can be reproduced under the conditions of Ref. 2. It should be stressed that here we establish uniqueness or quasi-uniqueness of pointwise singular conformal structures rather than of conformal compactifications; a simpler proof of the latter problem, under slightly different conditions, can be obtained using the results of Refs. 4 or 5 (cf., e.g., Ref. 8). To be specific only the three-dimensional problem is considered; similar results hold in any dimension.

II. UNIQUE POINTWISE SINGULAR CONFORMAL STRUCTURES

Following Geroch we shall consider a one-point compactification $\overline{\Sigma} = \Sigma \cup \{i_0\}$ of the initial data (three-dimensional Riemannian) slice Σ, giving $\overline{\Sigma}$ the standard topology. Specifically, such a compactification will be obtained, e.g., by performing the inversion $x^i \to y^i = x^i/r(x)^2$ in some asymptotically flat coordinate system and adding the origin $y^i = 0$ to the set so obtained. This allows one to use the coordinates y^i to define the differentiable structure of $\overline{\Sigma}$ in a neighborhood of i_0. If one applies this procedure to a metric of the form

$$g_{ij} = \delta_{ij} + h_{ij}, \quad h_{ij} = O(r^{-1}), \tag{2.1}$$

the resulting conformally rescaled metric in coordinates y^i will only be Lipschitz continuous at i_0 rather than differentiable, therefore a natural procedure is to include in the atlas those coordinate transformations, the derivatives of which are merely Lipschitz continuous at i_0. (If one does not enlarge the, say C_k, atlas fixed by the coordinates y to the $A_{k,\alpha}$ atlas, as described below, one will end up with infinitely many conformally inequivalent completions of any asymptotically flat space-time, which seems to be quite a luxury. Another way of looking at this issue is that none of the $A_{k,\alpha}$ coordinate systems is singled out by physics, so that an $A_{k,\alpha}$ rather than a C_k structure on $\Sigma \cup \{i_0\}$ is physically appropriate.) To make things precise, let $B(R)$ be a ball of radius R and let us define the set $A_{k,\alpha}(B(R))$, $k \geqslant 1$, $\alpha \in (0,1]$, of functions on $B(R)$, such that $f \in C_1(B(R))$, f is C_k on $B(R) \setminus \{0\}$ and f satisfies

$$|\partial_i f(y) - \partial_i f(0)| \leqslant Cr(y)^{\alpha}, \ldots, |\partial_{i_1} \cdots \partial_{i_k} f| \leqslant Cr^{\alpha - k + 1}, \tag{2.2}$$

for some constant C. An $A_{k,\alpha}$ differentiable structure on $\overline{\Sigma}$ will be defined by a maximal atlas on $\overline{\Sigma}$ such that (a) $\overline{\Sigma} \setminus \{i_0\}$ is a C_k manifold and (b) in local coordinates, in a neighborhood of $i_0 = 0$, the transition functions are in $A_{k,\alpha}(B(R))$.

The $A_{k,\alpha}(\overline{\Sigma}, i_0)$ functions on $\overline{\Sigma}$ are defined in an obvious manner. We shall write $A_{k,\alpha}$ rather than $A_{k,\alpha}(\overline{\Sigma}, i_0)$, since confusion is unlikely to occur. A tensor field will be called of class $B_{l,\alpha}(\overline{\Sigma}, i_0)$ $(B_{l,\alpha})$, $l \geqslant 0$, if its components t_I in a map belonging to the $A_{k,\alpha}$ atlas, $k \geqslant l + 1$, where I is a set of covariant and contravariant indices, are C_l on Σ and satisfy

$$\forall y \in \mathcal{O} \setminus \{0\} \quad |t_I(y) - t_I(0)| \leqslant C' r(y)^{\alpha}, \ldots, |\partial_{i_1} \cdots \partial_{i_l} t_I| \leqslant C' r^{\alpha - l} \tag{2.3}$$

for some constant C', where $\mathcal{O}$ is a coordinate neighborhood

[a)] On leave of absence from the Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

 0022-2488/89/092090-04$02.50

of $0 = i_0$. (It is simple to check that if such a constant exists for one map, there will be some constant for any map from the $A_{k,\alpha}$ atlas.) We shall say that a Riemannian manifold (Σ, g) is asymptotically flat if there exists an $A_{k,\alpha}$, $k \geqslant 3$, Riemannian manifold $\bar{\Sigma} = \Sigma \cup \{i_0\}$ with $B_{k-1,\alpha}$ metric $\bar{g}$ and a function Ω: $\bar{\Sigma} \to \mathbb{R}^+ \cup \{0\}$, such that (1) on Σ we have $g_{ij} = \Omega^{-2}\bar{g}_{ij}$ and (2) $\Omega(i_0) = 0$, $\nabla\Omega(i_0) = 0$, $\forall i\, \partial_i \Omega \in A_{1,\alpha}$, and $\partial_i\, \partial_j \Omega(i_0) = 2\bar{g}_{ij}(i_0)$.

To prove our uniqueness results we shall need two auxiliary lemmas.

Lemma 2.1: Let $\alpha \in (0,1]$, let g be a metric in $B(r_0)$ (= ball of radius r_0), satisfying

$$|g_{ij}(x) - g_{ij}(0)| \leqslant Cr^{\alpha}, \quad |\partial_k g_{ij}| \leqslant Cr^{\alpha-1}, \ldots, \quad |\partial_{i_1} \cdots \partial_{i_l} g_{ij}| \leqslant Cr^{\alpha-l}, \tag{2.4}$$

$l \geqslant 1$, and let c: $B(r_0) \setminus \{0\} \to \mathbb{R}$ satisfy

$$|c| \leqslant Cr^{\alpha-2}, \quad |\partial_i c| \leqslant Cr^{\alpha-3}, \ldots, |\partial_{i_1} \cdots \partial_{i_k} c| \leqslant Cr^{\alpha-2-k}, \tag{2.5}$$

$k \geqslant 0$. There exists $0 < r_1 \leqslant r_0$ and a function f: $B(r_1) \setminus \{0\} \to \mathbb{R}$, a weak solution of

$$(\Delta_g + c) f = 0, \tag{2.6}$$

in $B(r_1) \setminus \{0\}$, satisfying $\frac{1}{4} \leqslant f \leqslant 4$.

Proof: Equation (2.6) for f is equivalent to the equation

$$(\Delta_{\bar{g}} + \bar{c})\bar{f} = 0, \tag{2.7}$$

where $\bar{g}_{ij} = u^{4/(n-2)} g_{ij}$, $\bar{f} = f/u$, and $\bar{c} = cu^{-4/(n-2)} + u^{-(n+2)/(n-2)} \Delta_g u$ (in dimension n). Let $u = 1 - ar^{\alpha}$, $a \in \mathbb{R}^+$. For $r < r_1$, small enough, one finds

$$\Delta_g u \leqslant -a\alpha(n - 2 + \alpha) r^{-2+\alpha}/2,$$

so that increasing a and decreasing r_1 if necessary one has (for $r \leqslant r_1$) $\bar{c} \leqslant 0$ and $u \geqslant \frac{1}{2}$, and the metric $\bar{g}$ satisfies inequalities of the form (2.4). The same calculation shows that for positive C_1 and C_2 large enough, again decreasing r_1 if necessary, the functions $u^+ = 2 - C_1 r^{\alpha}$ and $u^- = \frac{1}{2} + C_2 r^{\alpha}$ are supersolutions and subsolutions, respectively, for Eq. (2.7), and we have $u^+|_{S(r_1)} \geqslant 1$, $u^-|_{S(r_1)} \leqslant 1$, where $S(r_1)$ denotes a sphere of radius r_1. Let $\bar{f}_k$ be the sequence of solutions of (2.7) in $B(r_1) \setminus B(r_1/k)$ satisfying $\bar{f}_k|_{S(r_1)} = \bar{f}_k|_{S(r_1/k)} = 1$. By the comparison principle we have $u^- \leqslant \bar{f}_k \leqslant u^+$, wherever defined. By $C_{1,\epsilon}$ estimates, as given, e.g., in Ref. 9 and a standard diagonalization procedure one can extract a sequence f_{i_k} converging to a function $\bar{f} = \bar{f}_{\infty}$, a weak solution of (2.7) in $B(r_1) \setminus \{0\}$, satisfying $1/2 \leqslant \bar{f} \leqslant 2$. f is defined by $f \equiv u\bar{f}$.

Lemma 2.2: Under the hypotheses of Lemma 2.1, let f: $B(r_1) \setminus \{0\} \to \mathbb{R}$ be a weak solution of (2.6) in $B(r_1) \setminus \{0\}$ satisfying $c_0 \leqslant f \leqslant c_1$, $c_0, c_1 \in \mathbb{R}$. Then f satisfies (2.6) in $B(r_1)$ in a weak sense, and there exist constants $f(0)$ and C_2, such that

(a) $|f(x) - f(0)| \leqslant C_2 r^{\alpha}$,
$|\partial f| \leqslant C_2 r^{\alpha-1}, \ldots, |\partial_{i_1} \cdots \partial_{i_m} f| C_2 r^{\alpha-m}$, if $0 < \alpha < 1$,

(b) $|f(x) - f(0)| \leqslant C_2 r \ln r$,
$|\partial f| \leqslant C_2 \ln r, \ldots, |\partial_{i_1} \cdots \partial_{i_m} f| C_2 r^{\alpha-m} \ln r$, if $\alpha = 1$,

where $m = \min(l, k+1)$.

Proof: Because f is bounded from above and below the removable singularity theorem of Serrin[10] (Theorem 1) implies that f is a weak solution in $B(r_1)$ so that by Theorem 8 of Ref. 11, f is Hoelder continuous at 0 with some exponent ϵ. Performing an inversion, $r \to \rho = 1/r$, $f \to \bar{f} = rf$, one finds that $\bar{f}$ satisfies the following equation (It is not too difficult, using, e.g., the methods of Ref. 13, to prove the desired estimates directly without performing an inversion. We use the inversion argument for simplicity, to be able to use the classical results of Ref. 12):

$$\Delta_{\bar{g}} \bar{f} = O(\rho^{-3-\epsilon}), \quad \text{in } \mathbb{R}^n \setminus B(1/r_1),$$

where $\bar{g}$ is the conformally rescaled metric in $\mathbb{R}^n \setminus B(1/r_1)$, $\bar{g}_{ij} = \rho^4 g_{ij}$. An iterative application of the estimates of Meyers (Ref. 12, Lemma 5) (cf. also, e.g., Ref. 13) implies $\bar{f} = O(\rho^{-1-\alpha})$ for $0 < \alpha < 1$ or $\bar{f} = O(\rho^{-2} \ln \rho)$ for $\alpha = 1$, a standard scaling argument (cf., e.g., Ref. 12) gives the derivatives estimates, inverting back to the original variables our claims follow.

The idea of the proof of Proposition 2.3, which is the key result to prove our theorems, is due to Geroch (cf. the Appendix in Ref. 7).

Proposition 2.3: Let $0 < \alpha < 1$, and let g^1 and g^2 be two metrics in neighborhoods Ω_1 and Ω_2 of the origin conformally related to each other in $\Omega_i \setminus \{0\}$:

$$\text{for } x \neq 0, \ g^1_{ij}(x) = \phi^2(x) g^2_{kl}(y(x)) \frac{\partial y^k}{\partial x^i} \frac{\partial y^l}{\partial x^j}, \quad \phi(x) > 0;$$

$y(x)$ continuous in Ω_1, $y(0) = 0$, and $y(x) C_1$ for x different from 0, and let the metrics g^a satisfy the inequalities (2.4) with $l \geqslant 2$. The function ϕ can be extended to a continuous strictly positive function on Ω_1, satisfying

$$|\partial\phi| \leqslant Cr^{\alpha-1},$$

for some constant C.

Proof: Suppose first that $l \geqslant 3$. Lemma 2.1 implies the existence of functions ϕ_1 and ϕ_2, $0 < \frac{1}{4} < \phi_\alpha < 4$, such that

$$(\Delta_{g^a} - R_a/8)\phi_a = 0 \tag{2.8}$$

in suitable neighborhoods of the origins, where R_a denotes the Ricci scalar of g^a. (2.8) implies $\bar{R}_a = 0$, where $\bar{R}_a$ is the Ricci scalar of the conformally rescaled metric $\bar{g}^a_{ij} = \phi_a^4 g^a_{ij}$, and the estimates of Lemma 2.2 show that the rescaled metrics $\bar{g}^a_{ij}$ also satisfy inequalities of the form (2.4) with $l = 2$. Let $\psi = \phi_2^{-1} \phi^{1/2} \phi_1$. $\bar{R}_a = 0$ and $\bar{g}^1_{ij} = \psi^4 \bar{g}^2_{ij}$ imply

$$\forall x \in \Omega_1 \setminus \{0\} (\Delta_{\bar{g}^1} - \bar{R}_1/8)\psi = \Delta_{\bar{g}^1}\psi = 0,$$
$$\forall y \in \Omega_2 \setminus \{0\} (\Delta_{\bar{g}^2} - \bar{R}_2/8)(\psi^{-1}) = \Delta_{\bar{g}^2}(\psi^{-1}) = 0. \tag{2.9}$$

Since $\psi \geqslant 0$, a theorem by Serrin (Theorem 2, Ref. 10) implies that either ψ behaves as $1/r$ or ψ is bounded and satisfies (2.9) throughout Ω_1—whichever case occurs the maximum

principle applies and ψ is bounded in $\Omega_1\setminus\{0\}$ from below by the minimum of its values on $\partial\Omega_1$, which is strictly positive by construction of ϕ_a and by the hypothesis on ϕ. Similarly, ψ^{-1} either blows up as $1/r$ or is bounded from above and is bounded from below by a strictly positive constant. If ψ behaves like $1/r$ then ψ^{-1} must vanish at the origin, which leads to a contradiction; therefore both ψ and ψ^{-1} are bounded and Lemma 2.2 gives derivative estimates for ψ. Since g^1 is conformally equivalent to g^2 with a bounded conformal factor, a standard argument (cf., e.g., the proof of Lemma 1 in Ref. 5) implies $c^{-1}r(x)\leqslant r(y(x))\leqslant cr(x)$ for some constant c, so that from $|\nabla\phi_1|_{g_1}\leqslant Cr(x)^{\alpha-1}$, $|\nabla\phi_2|_{g_2}\leqslant Cr(y)^{\alpha-1}$ and $\phi=\psi^2\phi_1^{-2}\phi_2^2$ one obtains the derivatives estimate for ϕ. The case $l=2$ is obtained by an approximation argument.

Theorem 2.4 (Uniqueness of conformal $A_{k,\alpha}$ structure): Let (Σ_1,i_1,g_1), (Σ_2,i_2,g_2) be $A_{k,\alpha}$ manifolds with $B_{k-1,\alpha}$ metrics, $\alpha\in(0,1)$, $k\geqslant 3$; let Ψ be a continuous conformal mapping from Σ_1 to Σ_2, $\Psi(i_1)=i_2$, Ψ differentiable in $\Sigma_1\setminus\{i_1\}$. Then Ψ is $A_{k,\alpha}$.

Proof: In local coordinates in neighborhoods of i_1 and i_2 the hypotheses of Proposition 2.3 are satisfied, therefore the metric $\bar g^1_{ij}=\phi^{-2}g^1_{ij}$, ϕ as in Proposition 2.3, is of class $B_{1,\alpha}$. Replacing g^1 by $\bar g^1$ and $\lim_{r\to\infty}$ by $\lim_{r\to 0}$ in the proof of Lemma 1 of Ref. 5 shows that Ψ is $A_{2,\alpha}$. The formula for the transformation of the Ricci tensor under conformal changes of the metric can be written in the form

$$\phi^{-1}_{;ij}=\frac{1}{(n-2)}\{\phi^{-1}(R^1_{ij}-R^2_{ij})+2(n-2)\phi\phi_{,i}^{-1}\phi_{,j}^{-1}-(\Delta_{g_1}\phi^{-1}+(n-3)\phi|\nabla\phi^{-1}|^2_{g_1})g^1_{ij}\},\tag{2.10}$$

and we also have

$$\Delta_{g_1}\phi^{-1}=\{-\phi^{-3}R_2+\phi^{-1}R_1-(n-1)(n-4)\phi|\nabla\phi^{-1}|^2_{g_1}\}/2\{(n-1)\}.\tag{2.11}$$

In local coordinates x^i in a neighborhood of i_1, the right-hand side of (2.11) is bounded by $Cr^{\alpha-2}$, so that (2.10) implies that ϕ is twice differentiable outside the origin and

$$|\partial_i\partial_j\phi|\leqslant Cr^{\alpha-2}.\tag{2.12}$$

The right-hand side of the equation

$$\frac{\partial^2 y^i}{\partial x^k\partial x^l}=\bar\Gamma^{1m}_{kl}\frac{\partial y^i}{\partial x^m}-\Gamma^{2i}_{mn}\frac{\partial y^m}{\partial x^k}\frac{\partial y^n}{\partial x^l},\tag{2.13}$$

where $\bar\Gamma^{1m}_{kl}$ and Γ^{2i}_{mn} are the Christoffel symbols of the metrics $\phi^{-2}g^1_{ij}$ and g^2_{ij}, respectively, is in view of (2.12) differentiable with derivatives bounded by $Cr^{\alpha-2}$ for some constant C, which shows that ψ is $A_{3,\alpha}$. The equations obtained by differentiation of (2.10), (2.11), and (2.13) imply $\psi\in A_{k,\alpha}$ by induction.

III. THE LOGARITHMIC AMBIGUITIES

In Sec. II we have established uniqueness of $A_{k,\alpha}$ conformal structures, $\alpha\in(0,1)$. In Geroch's analysis one needs some more structure, which we introduce below, it should, however, be noted that the well-known condition $\alpha>\frac{1}{2}$ (Ref. 4, 5, and 14) is sufficient for a meaningful definition of energy in Geroch's framework (cf., e.g., Ref. 8). We shall say that a function f is of class $\bar A_k(\bar\Sigma,i_0)$ $(\bar A_k)$ if f is C_1 on $\bar\Sigma$ and if the following limits

$$\lim_{r\to 0}\partial_i\partial_j f(r\mathbf{n}),\ldots,\lim_{r\to 0}r^{k-2}\partial_{i_1}\cdots\partial_{i_k}f(r\mathbf{n})$$

exist, where $\mathbf{n}$ is any unit vector. As in Sec. II we can define an $\bar A_k$ differentiable structure and $\bar B_l$, $l\leqslant k-1$ tensors. The following theorem shows that there exists only a three-parameter family (parametrized by the vector $\bar C^i$) of inequivalent $\bar A_k$ structures compatible with a given conformal geometry.

Theorem 3.1 (Quasi-uniqueness of conformal $\bar A_k$ structures): Let (Σ_1,i_1,g_1), (Σ_2,i_2,g_2) be $\bar A_k$ manifolds, $k\geqslant 3$, with $\bar B_{k-1}$ metrics, let Ψ be a continuous conformal mapping from Σ_1 to Σ_2, $\Psi(i_1)=i_2$, Ψ of class C_1 in $\Sigma_1\setminus\{i_1\}$. Then there exist $\bar A_k$ charts $\{\bar x^i\}$, $\{\bar y^i\}$ in neighborhoods of i_1 and i_2 and a (constant) vector $\bar C^k$, such that, in local coordinates, Ψ takes the form

$$\bar y^i=\bar x^i+(\bar C^i\bar r^2-2\bar C^k\bar x^k\bar x^i)\ln\bar r,\quad \bar r=r(\bar x).$$

In particular, ψ is $\bar A_k$ if and only if $\bar C^i=0$.

Proof: For any $\alpha<1$, ψ is of class $A_{k,\alpha}$ by Theorem 2.4. There exist $\bar A_k$ coordinates $\{x^i\}$, $\{y^i\}$ in neighborhoods of i_1 and i_2, such that $x^i(i_1)=y^i(i_2)=0$, $g^1_{ij}(0)=g^2_{ij}(0)=\delta_{ij}$, $\phi(0)=1$. The estimates of Lemma 2.2, part (b) and Eqs. (2.10) and (2.11) yield an equation of the form

$$\Delta_{\text{flat}}\phi=\Delta_{\text{flat}}\phi-\Delta_{g_2}\phi+\Delta_{g_2}\phi=c(\theta,\varphi)/r+o(1/r),\tag{3.1}$$

for some function $c(\theta,\varphi)$, so that, e.g., an inversion argument, as in the proof of Lemma 2.2 together with the estimates of Ref. 12, lead to

$$\phi=1+C^ix^i\ln r+rg(\theta,\varphi)+o(r),\tag{3.2}$$

for some constants C^i and a function $g(\theta,\varphi)$, which implies

$$\bar\Gamma^{1m}_{kl}=(\delta^m_kC^l+\delta^m_lC^k-\delta^k_lC^m)\ln r+A^m_{kl}(\theta,\varphi)+o(1),\tag{3.3}$$

for some functions $A^m_{kl}(\theta,\varphi)$. (3.3) inserted in (2.13) gives

$$\frac{\partial}{\partial r}\frac{\partial y^k}{\partial x^i}=n^l\frac{\partial^2y^k}{\partial x^i\partial x^l}=(n^kC^i+\delta^k_iC^jn^j-n^iC^k)\ln r+A^k_i(\theta,\varphi)+o(1),\tag{3.4}$$

$n^i=x^i/r$, with some functions $A^k_i(\theta,\varphi)$. Twice integrating (3.4) along rays, rigidly rotating the coordinates x if necessary, $x^i\to\bar x^i=\omega^i_jx^j$, $\omega^i_j\in O(3)$, one obtains

$$y^k=\bar x^k+(2\bar C^j\bar x^j\bar x^k-\bar r^2\bar C^k)\ln\bar r+\bar r^2A^k(\theta,\varphi)+o(\bar r^2),\tag{3.5}$$

for some functions $A^k(\theta,\varphi)$. Let

$$\bar y^k=\bar x^k+(2\bar C^j\bar x^j\bar x^k-\bar r^2\bar C^k)\ln\bar r.$$

Then (3.5) yields

$$y^k=\bar y^k+r(\bar y)^2A^k(\theta,\varphi)+o(\bar r^2),\tag{3.6}$$

which is an $\bar A_k$ transformation and the theorem follows.

IV. CONCLUSIONS

Our results show that Geroch's description of spatial infinity is a fairly natural one, since the conformal $\overline{A}_k$ completions are "almost uniquely" defined by the asymptotic behavior of the metric: the only ambiguities consist of a three-parameter family of "logarithmic transformations." It is well known (cf., e.g., Ref. 15) that these transformations do not affect the numerical value of four-momentum. It must be stressed that the $A_{k,\alpha}$ conformal description of initial data slices is completely equivalent to the standard coordinate description (cf., e.g., Refs. 1,4,14,5), and the use of one or another seems to be a matter or personal taste, depending upon whether one prefers the symbol $\lim_{r\to 0}$ to the symbol $\lim_{r\to\infty}$.

ACKNOWLEDGMENTS

The author benefited from discussions with A. Ashtekar and R. Geroch. The friendly hospitality of the Center for Mathematical Analysis in Canberra, Australia during part of work on this paper is acknowledged.

This work was supported in part by NSF Grant No. PHY 8503072 to Yale University.

[1] R. Arnowitt, S. Deser, and C.W. Misner, "The Dynamics of General Relativity," in *Gravitation, an Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[2] R. Geroch, J. Math. Phys. **13**, 956 (1972).
[3] A. Ashtekar and R. Hansen, J. Math. Phys. **19**, 1542 (1978).
[4] R. Bartnik, Comm. Pure Appl. Math. **39**, 661 (1986).
[5] P. T. Chruściel, "Boundary Conditions at Spatial Infinity from a Hamiltonian Point of View," in *Topological Properties and Global Structure of Space-Time*, edited by P. G. Bergmann and V. de Sabbata (Plenum, New York, 1986), pp. 49–59.
[6] R. Geroch, in *Asymptotic Structure of Space-time*, edited by P. Esposito and L. Witten (Plenum, New York, 1976), p. 1.
[7] R. Geroch, J. Math. Phys. **11**, 2580 (1970).
[8] P. T. Chruściel, "On the Energy of the Gravitational Field at Spatial Infinity," in *Proceedings of the Canberra Miniconference on Mathematical General Relativity*, edited by R. Bartnik (ANU, Canberra, 1988).
[9] P. Tolskdorff, J. Diff. Eq. **51**, 126 (1984).
[10] J. Serrin, Acta Math. **113**, 219 (1965).
[11] J. Serrin, Acta Math. **111**, 247 (1964).
[12] N. Meyers, J. Math. Mech. **12**, 247 (1963).
[13] P. T. Chruściel, CMA ANU Preprint No. CMA-R27-88.
[14] N. O'Murchadha, J. Math. Phys. **27**, 2111 (1986).
[15] A. Ashtekar, Found. Phys. **15**, 419 (1985).

On the structure of spatial infinity. II. Geodesically regular Ashtekar–Hansen structures

Piotr T. Chruściel[a)]
Physics Department, Yale University, New Haven, Connecticut 06511

(Received 12 January 1989; accepted for publication 26 April 1989)

The ambiguities in the differentiable structure of Ashtekar–Hansen completions satisfying a geodesic condition are analyzed. The results obtained imply, in particular, uniqueness up to a four-parameter family of "logarithmic transformations" of completions of asymptotically flat space-times stationary "in a neighborhood of i_0."

I. INTRODUCTION

In a previous paper of this series[1] uniqueness up to a three-parameter family of logarithmic transformations of conformal one-point compactifications of three-dimensional asymptotically flat Riemannian manifolds was established. A corollary of that result is that within the Geroch framework[2] the four-momentum of an initial data set for Einstein equations is unambiguously defined. In general relativity, which is a four-dimensional theory *par excellence*, one hopes to assign a four-momentum p_μ or, say, its invariant square $m^2 = -\eta_{\alpha\beta}p^\alpha p^\beta$ to a four-dimensional set, rather than to a three-dimensional subset thereof. It has been shown in Ref. 3 that one can, in a meaningful way, associate an invariant mass m to a boost-type domain or, more generally, to a four-dimensional asymptotically flat end of a Lorentzian manifold defined by a collection of boost-type domains. This relatively satisfactory result suffers from the drawback that the somewhat arbitrary notion of the boost-type domain plays an essential role in the analysis. One would like to replace the statement that "two three-dimensional ends included in some boost-type domain of a vacuum space-time have the same mass" by something of the kind "two three-dimensional ends included in the same asymptotic region have the same mass," avoiding the use of some unnaturally preferred sets in some coordinate system as a primary concept of the construction. A reasonably natural setup in which one can define the notion of an asymptotic region has been proposed by Ashtekar and Hansen[4,5] who describe the behavior of the gravitational field at spatial infinity by means of conformal completions of asymptotically flat four-dimensional manifolds in which spatial infinity is represented by a point i_0. The existence of an Ashtekar–Hansen completion—or some variation thereof, as considered in this paper—adds useful information about the global causal structure of space-time to the standard coordinate notion of asymptotic flatness,[6,3,7] which seems difficult to describe in terms of asymptotically flat coordinates only. The main problem with the Ashtekar–Hansen completions is their potential nonuniqueness. In this paper we show that if a certain geodesic condition is satisfied by some completion, then there exists a four-parameter family of inequivalent completions only.

In Sec. II we introduce the notion of *weak conformal completions* and the *geodesic regularity* condition. We show that weak geodesically regular completions are unique. We also show that every completion of a *no-radiation* metric (in particular, of the Kerr metrics) is geodesically regular. In Sec. III we define *strong completions* (the conditions of this section are essentially those of Ashtekar and Hansen) and we show their uniqueness up to "logarithmic ambiguities" provided that geodesic regularity holds.

II. WEAK CONFORMAL COMPLETIONS

In order to give a motivation to the definitions of this section let us recall the fundamental result of Christodoulou and O'Murchadha[6] (the "boost theorem"): Given asymptotically flat data for general relativity (cf. Ref. 6 for the appropriate definition of asymptotic flatness) and given any "boost slope" $\theta < 1$ there exists a metric $g_{\mu\nu}$, solution of the vacuum Einstein equations, the evolution of the Cauchy data, and positive constants R and T such that $g_{\mu\nu}$ is defined for all x^μ belonging to the *boost-type domain* $\Omega_{\theta,R,T}$ (Ref. 8):

$$\Omega_{\theta,R,T} = \{x^\alpha : r \geqslant R, |x^0| \leqslant \theta r + T\},$$
$$\theta > 0, \quad R \geqslant 0, \quad T \in (-\infty, \infty],$$

with $g_{\mu\nu}$ satisfying

$$|g_{\mu\nu} - \eta_{\mu\nu}| \leqslant C(1+r)^{-\alpha}, \quad |\partial_\sigma g_{\mu\nu}| \leqslant C(1+r)^{-\alpha-1} \tag{1}$$

for some constants $C(\theta,R,T,g_{\mu\nu})$, $\alpha > 0$, where $\eta_{\mu\nu}$ is the Minkowski metric. For $x^\mu x_\mu > 0$ (Ref. 9) (signature $-+++$) let Φ denote the inversion $x^\mu \to y^\mu = x^\mu/(x^\alpha x_\alpha)$. It is simple to check that for $\theta < 1$ and $T > 0$ the set $\Phi(\Omega_{\theta,R,T})$ contains the "wedge" $W_{\theta,1/R}$ (Ref. 10):

$$W_{\theta,\epsilon} = \{y^\mu : r(y) < \epsilon, |y^0| < \theta r\}$$

and Eq. (1) gives, for $y^\mu \in W_{\theta,1/R}$,

$$g_{\mu\nu}\,dx^\mu\,dx^\nu = (y^\alpha y_\alpha)^{-2}\bar{g}_{\mu\nu}\,dy^\mu\,dy^\nu,$$
$$|\bar{g}_{\mu\nu} - \eta_{\mu\nu}| \leqslant C'(\theta,R,T,g_{\mu\nu})r(y)^\alpha, \tag{2}$$
$$\left|\frac{\partial \bar{g}_{\mu\nu}}{\partial y^\sigma}\right| \leqslant C'(\theta,R,T,g_{\mu\nu})r(y)^{\alpha-1}.$$

Equation (2) displays the expected behavior of the metric under an inversion which brings "spatial infinity" to a point $y^\alpha = 0$, say i_0. Following Ashtekar and Hansen[4,5] we shall ask for some more structure than what follows in a straightforward manner from the "boost theorem."

[a)] On leave of absence from the Institute of Mathematics of the Polish Academy of Sciences, Warsaw, Poland.

 0022-2488/89/092094-07$02.50

Definition 1: Let (M,g) be a space-time ($\equiv C^3$ four-dimensional manifold with a C^2 Lorentzian metric) and let $\mathscr{M}$ denote the disjoint union $M\cup\{i_0\}$ where i_0 is a point. We shall say that $(\mathscr{M},\bar{g},i_0)$ is a *weak α-completion* of M, $\alpha\in(0,1]$, if the following holds.

(i) In $\mathscr{M}$ there exists a coordinate system $\{y^\mu\}$, $y^\mu\in W_{1,\epsilon}\cup\{0\}$ such that $i_0=0$.

(ii) There exists a function Ω: $W_{1,\epsilon}\to\mathbb{R}^+$ such that the metric $\bar{g}_{\mu\nu}=\Omega^2 g_{\mu\nu}$ satisfies

$$\forall\theta<1,\quad \forall x\in W_{\theta,\epsilon}\ |\bar{g}_{\mu\nu}-\eta_{\mu\nu}|\leqslant C(\theta)r(y)^\alpha,$$
$$|\partial_\sigma \bar{g}_{\mu\nu}|\leqslant C(\theta)r(y)^{\alpha-1}.$$

(iii)

$$\lim_{y\to 0}\Omega=0,\quad \lim_{y\to 0}\partial_\alpha\Omega=0,\quad |\bar{\nabla}_\mu\bar{\nabla}_\nu\Omega-2\bar{g}_{\mu\nu}|\leqslant C(\theta)r(y)^\alpha$$

with some function $[0,1)\ni\theta\to C(\theta)<\infty$ (Ref. 11).

(iv) For all $p\in M$ there exists no timelike curve in $\mathscr{M}$ from p to i_0.

Here $\mathscr{M}$ shall be equipped with the natural topology induced by the topology of M and the coordinates y^α. We shall say that a weak conformal completion is *geodesically regular* if for every affinely parametrized spacelike geodesic Γ of the physical metric $g_{\mu\nu}$, $\Gamma=\{y^\alpha(s),s\geqslant s_0\}$, which extends to i_0, there exist constants $s_1\geqslant s_0$, $\Psi(\Gamma)<1$ such that for all $s\geqslant s_1$ we have[12]

$$|y^0(s)|\leqslant\Psi(\Gamma)r(y(s)).$$

We shall show that weak geodesically regular α-completions are unique for $0<\alpha<1$.

Lemma 1: Let $g_{\mu\nu}$ be a C_2 metric in a boost-type domain $\Omega_{\theta,R,T}$ satisfying

$$|g_{\mu\nu}-\eta_{\mu\nu}|\leqslant C(1+r)^{-\alpha},\quad |\partial_\sigma g_{\mu\nu}|\leqslant C(1+r)^{-\alpha-1} \tag{3}$$

with some constant C. There exist spacelike hypersurfaces $B^\pm\subset\Omega_{\theta,R,T}$ defined by

$$B^\pm=\{p\in\Gamma^\pm_{\mathbf{n}},\ \text{where } \Gamma^\pm_{\mathbf{n}}=\text{complete spacelike geodesic}[13]\ \text{satisfying } \mathbf{x}^\pm_{\mathbf{n}}(0)=R_1\mathbf{n},$$
$$\mathbf{n}\in S(1)[14],\ \frac{d\mathbf{x}^\pm_{\mathbf{n}}}{ds}(0)=\mathbf{n},\ (x^\pm_{\mathbf{n}})^0(0)=0,\ \lim_{s\to\infty}r(x^\pm_{\mathbf{n}}(s))=\infty,\ (x^\pm_{\mathbf{n}})^0(s)\geqslant 0\ \text{on}\ \Gamma^+_{\mathbf{n}},$$
$$(x^\pm_{\mathbf{n}})^0(s)\leqslant 0\ \text{on}\ \Gamma^-_{\mathbf{n}}\}$$

[in other words, $B^\pm$ are "sewn up" from geodesics starting from the sphere $\{r(x)=R_1,\ x^0=0\}$ which stay in $\Omega_{\theta,R,T}$ and remain either to the local future (with respect to the chronology of $\Omega_{\theta,R,T}$) or to the local past of $N^x_{R_1}\equiv\{x^\alpha: x^0=0, r(x)\geqslant R_1\}$]. The $B^\pm$ are graphs over $N^x_{R_1}$: $B^\pm=\{x^\mu: r(x)\geqslant R_1,\ x^0=w^\pm(\mathbf{x})\}$. Every future directed timelike curve starting at $N^x_{R_1}$ either remains entirely within the wedge $W^+=\{x^\mu: r(x)\geqslant R_1, 0\leqslant x^0\leqslant w^+(\mathbf{x})\}$ or meets B^+. Similarly, every past directed timelike curve starting at $N^x_{R_1}$ either remains entirely within the wedge $W^-=\{x^\mu: r(x)\geqslant R_1,\ w^-(\mathbf{x})\leqslant x^0\leqslant 0\}$ or meets B^-.

Proof: By Propositions B1 and B2 of Appendix B of Ref. 3 for R_1 sufficiently large the family of geodesics $\Gamma^\pm_{\mathbf{n}}$, $\mathbf{n}\in S(1)$ defined by $\mathbf{x}^\pm_{\mathbf{n}}(0)=R_1\mathbf{n}$, $\mathbf{n}\in S(1)$, $(d\mathbf{x}^\pm_{\mathbf{n}}/ds)(0)=\mathbf{n}$, $(x^\pm_{\mathbf{n}})^0(0)=0$, and $d(x^\pm_{\mathbf{n}})^0/ds(0)=\pm\theta_0\equiv\pm\min(\theta,1)/2$ will satisfy $\forall s\geqslant 0 (x^+_{\mathbf{n}})^0(s)\geqslant 0$, $(x^-_{\mathbf{n}})^0(s)\leqslant 0$, $(\mathbf{x}^\pm_{\mathbf{n}}/r)(d\mathbf{x}^\pm_{\mathbf{n}}/ds)(s)\geqslant\frac{1}{2}$, and $\theta_0/2\leqslant\pm d(x^\pm_{\mathbf{n}})^0/ds\leqslant 3\theta_0/2$. Let

$$A=\{\rho\geqslant R_1:\forall\mathbf{x}\in B(\rho)\setminus\mathrm{Int}[B(R_1)]\,\exists\mathbf{n}_\pm(\mathbf{x})\in S(1)\ \text{and}$$
$$\mathbf{x}^0_\pm(\mathbf{x}),\ |\mathbf{x}^0_\pm(\mathbf{x})|\leqslant\theta r(\mathbf{x})$$
$$+T\ \text{such that}\ (\mathbf{x}^0_\pm(\mathbf{x}),\mathbf{x})\in\Gamma^\pm_{\mathbf{n}_\pm(\mathbf{x})}\},$$

where $B(\rho)$ denotes a closed ball of radius ρ. A is nonempty because $R_1\in A$, A is closed by standard properties of solutions of differential equations, and openness of A follows from the implicit function theorem and the fact that $d\mathbf{x}^\pm_{\mathbf{n}}/ds$ is everywhere transversal to the spheres $S(\rho)$. This implies $A=\{\rho\in\mathbb{R}:\rho\geqslant R_1\}$, so that for every $\mathbf{x}\in\mathbb{R}^3\setminus B(R_1)$ there exists $p^\pm=(w^\pm(\mathbf{x})\equiv\mathbf{x}^0_\pm(\mathbf{x}),\mathbf{x})\in\Omega_{\theta,R,T}$ and two geodesics $\Gamma^\pm_{\mathbf{n}_\pm}$ which pass through $p^\pm$.

Lemma 2: Let $(\mathscr{M}_1,g_1,i_1)$, $(\mathscr{M}_2,g_2,i_2)$ be two weak α-completions of a space-time (M,g), $0<\alpha<1$; let x and y be the appropriate coordinate systems, $x^\alpha\in W_{1,\epsilon_1}\cup\{0\}$, $y^\alpha\in W_{1,\epsilon_2}\cup\{0\}$; let Φ denote the coordinate transformation $y^\alpha(x^\mu)$ wherever defined; suppose that Φ be differentiable; and define $N_{\epsilon_1}=\{x^\alpha: x^0=0, r(x)\leqslant\epsilon_1\}$. If Φ (N_{ϵ_1}) is contained in a wedge W_{θ,ϵ_2}, $\theta<1$, then there exists a Lorentz matrix Λ^μ_ν such that

$$y^\mu=\Lambda^\mu_\nu x^\nu+\zeta^\mu, \tag{4}$$

where ζ^μ satisfies

$$\forall\psi<1,\quad \forall x\in W_{\psi,\epsilon_1}: |\zeta^\mu|\leqslant C(\psi)r^{1+\alpha},$$
$$|\partial_\mu\zeta^\nu|\leqslant C(\psi)r^\alpha,\quad |\partial_\mu\partial_\nu\zeta^\rho|\leqslant C(\psi)r^{\alpha-1}. \tag{5}$$

Proof: The image by Φ of i_1 must be i_2, otherwise there would exist a timelike curve from some point $p\in M$ to i_1, contradicting point (iv) of Definition 1. Let $\hat{x}^\mu=x^\mu/(x^\alpha x_\alpha)$, $\hat{y}^\mu=y^\mu/(y^\alpha y_\alpha)$; the hypersurface $\hat{N}^x_{1/\epsilon_1}=\{\hat{x}^\mu: r(\hat{x})\geqslant 1/\epsilon_1, \hat{x}^0=0\}$ is asymptotically flat and is included in some boost-type domain of coordinates $\hat{y}^\mu$ with slope θ smaller than 1—the result follows from Theorem 1 and Point 2 of Corollary 2 of Ref. 3.

Lemmas 1 and 2 lead to the following theorem.

Theorem 1 (uniqueness of geodesically regular weak completions, $\alpha<1$): If a space-time (M,g) admits one geodesically regular weak α-completion $(\bar{M},\bar{g},\bar{\imath})$, $\alpha\in(0,1)$, then the following holds.

(i) All weak α-completions of (M,g) are geodesically regular.

(ii) All weak α-completions of (M,g) are related to each other by coordinate transformations of the form (4) and (5).

Proof: Let $(\tilde{M},\tilde{g},\tilde{\imath})$ be some completion of (M,g), let $\{x^\alpha\}$ be the appropriate coordinate system for $(\tilde{M},\tilde{g})$, let y^α be the coordinates of Definition 1 for $(\bar{M},\bar{g})$, and let Ψ denote the transformation $y^\alpha(x^\mu)$. From Lemma 1 applied to the physical metric $g_{\mu\nu}$ in coordinates $\hat{x}^\mu = x^\mu/(x^\alpha x_\alpha)$ one infers the existence of two hypersurfaces $B^\pm$ sewn up from the geodesics $\Gamma_{\mathbf{n}}^\pm$, $\mathbf{n}\in S(1)$. By geodesic regularity for all $\mathbf{n}\in S(1)$ there exist $\theta_{\mathbf{n}}^\pm$ such that $\Psi(\Gamma_{\mathbf{n}}^\pm)\subset W_{\theta_{\mathbf{n}}^\pm,\epsilon_2}$. By compactness of $S(1)$ we have $\sup_{\mathbf{n}\in S(1)}\theta_{\mathbf{n}}^\pm = \theta_0 < 1$.[15] Since $N_{1/R_1} = \{x^\alpha : x^0 = 0, r(x)\leqslant 1/R_1\}$ (R_1 given by Lemma 1) lies to the local future of $B^- \subset W_{\theta_0,\epsilon_2}$ and to the local past of $B^+ \subset W_{\theta_0,\epsilon_2}$ we must have $N_{1/R_1}\subset W_{\theta_0,\epsilon_2}$, so that we can apply Lemma 2 to conclude that $y(x)$ is of the form (4), which establishes point (ii). It is not too difficult to show that $x(y)$ must be of the form (4) as well, so that the image $\Psi(\Gamma)$ of any spacelike geodesic $\Gamma = \{y^\alpha(s), s\geqslant s_0\}\subset\Omega_{\theta(\Gamma),\epsilon_2}$ will be included in $\Psi(W_{\theta(\Gamma),\epsilon_2})\subset W_{\theta',\epsilon_1'}$ with some $\theta' < 1$ and point (i) ensues.

A metric shall be called a *no-radiation metric* if there exist coordinates $x^\alpha\in\Omega_{R,\infty}\equiv(-\infty,\infty)\times[\mathbb{R}^3\setminus B(R)]$ such that $g_{00}\leqslant -\epsilon$ for some positive ϵ, g_{ij} is a positive definite matrix with eigenvalues separated from zero, and (3) holds throughout $\Omega_{R,\infty}$ with some constant C. The Kerr metrics are no-radiation metrics in this sense, with $\alpha = 1$.

Proposition 1: Weak α-completions of no-radiation metrics, $\alpha > 0$, are geodesically regular.

Proof: By point (i) of Theorem 1 it is sufficient to show the existence of one geodesically regular completion. Let x^α be the coordinates satisfying (3) and let $\hat{x}^\mu = x^\mu/(x^\alpha x_\alpha)$. By Proposition B1 of Appendix B of Ref. 3 every spacelike geodesic meeting i_0 [i.e., such that $\hat{x}\to 0\Rightarrow r(x)\to\infty$] behaves asymptotically as follows:

$$x^\alpha(s) = \eta_\infty^\alpha s + o(s)$$

for some constant vector η_∞^α satisfying $\eta_{\mu\nu}\eta_\infty^\mu\eta_\infty^\nu > 0$, where s is an affine parameter, and we can normalize η_∞^α to satisfy $\eta_\infty^i\eta_\infty^i = 1 (\Rightarrow |\eta_\infty^0| < 1)$. We have

$$\hat{x}^\alpha(s) = [\eta_\infty^\alpha/(1-|\eta_\infty^0|^2)]s^{-1} + o(s^{-1}),$$

so that $\hat{x}^0(s) = \eta_\infty^0 r(\hat{x}(s)) + o[r(\hat{x}(s))]$ and for s large enough one obtains

$$|\hat{x}^0(s)|\leqslant\theta r(s),\quad \theta = (|\eta_\infty^0| + 1)/2 < 1.$$

III. STRONG CONFORMAL COMPLETIONS

In Sec. II we have investigated the structure of the set of conformal completions in which the metric is allowed to blow up as one approaches "what would be the light cone of i_0." As has been shown by Schmidt and Walker (cf. Appendix C of Ref. 4) much better behaved completions can be obtained for Kerr metrics. To justify our conditions on the conformally rescaled metric, to be presented later, let us recall the Schmidt–Walker coordinates for the Schwarzschild metric: For $s > 2m$ let $f(s)$ be defined by

$$f(s) = s + 2m\ln(s/2m - 1) \tag{6}$$

and let us set

$$\begin{aligned} &r^* = f(r),\quad r^* = [f(\hat{v}^{-1}) + f(\hat{w}^{-1})]/2,\\ &t = [f(\hat{v}^{-1}) - f(\hat{w}^{-1})]/2,\\ &\hat{r} = (\hat{v}+\hat{w})/2,\quad \hat{t} = (\hat{v}-\hat{w})/2, \end{aligned} \tag{7}$$

where r and t are the standard Schwarzschild coordinates

$$ds^2 = -(1-2m/r)dt^2 + (1-2m/r)^{-1}dr^2 + r^2 d\Omega^2. \tag{8}$$

In the coordinates $\hat{v}$, $\hat{w}$, $\hat{v}\hat{w} > 0$, the Schwarzchild metric takes the form

$$ds^2 = \frac{1}{(\hat{v}\hat{w})^2}\times\left\{\frac{(1-2m/r)}{(1-2m\hat{v})(1-2m\hat{w})}d\hat{v}\,d\hat{w} + (\hat{v}\hat{w}r)^2 d\Omega^2\right\}. \tag{9}$$

From Eqs. (6) and (7) one has

$$r = r^* - 2m\ln(r^*/2m) + \zeta^*,$$

with

$$\zeta^* = 2m\ln\left\{\frac{1+2m\ln[r/2m-1]/r}{1-2m/r}\right\} = O\left(\frac{\ln[r^*]}{r^*}\right)$$

for large r^*; therefore, from

$$\begin{aligned} r^* &= \hat{r}/\hat{v}\hat{w} + m\{\ln[(1-2m\hat{v})(1-2m\hat{w})]\\ &\quad - \ln[4m^2\hat{v}\hat{w}]\} \end{aligned}$$

one obtains, for small $\hat{r}$,

$$\begin{aligned} \zeta^* &= (\hat{v}\hat{w}/\hat{r})O(|\ln[\hat{r}]| + |\ln[\hat{v}\hat{w}]|)\\ &= O(\hat{r}|\ln[\hat{r}]|),\quad \text{for } |\hat{t}|\leqslant\hat{r} \end{aligned}$$

and one obtains

$$\omega\equiv\hat{v}\hat{w}r/\hat{r} = 1 + (m\hat{v}\hat{w}/\hat{r})\ln(\hat{v}\hat{w}/\hat{r}^2) + O(\hat{r}^2), \tag{10}$$

so that

$$(\omega-1)/\hat{r} = m(1-\hat{t}^2/\hat{r}^2)\ln(1-\hat{t}^2/\hat{r}^2) + O(\hat{r}).$$

The metric (9) can be written in the form

$$ds^2 = \Omega^{-2}\hat{g}_{\mu\nu}\,d\hat{x}^\mu\,d\hat{x}^\nu = \Omega^{-2}\,d\hat{s}^2,\quad \Omega = \hat{v}\hat{w} = \hat{r}^2 - \hat{t}^2,$$

$$d\hat{s}^2 = \frac{(1-2m\hat{v}\hat{w}/(\omega\hat{r}))}{(1-2m\hat{v})(1-2m\hat{w})}\left(-d\hat{t}^2 + \frac{(\hat{x}^i\,d\hat{x}^i)^2}{\hat{r}^2}\right) + \omega^2\left[d\hat{x}^2 + d\hat{y}^2 + d\hat{z}^2 - \frac{(\hat{x}^i\,d\hat{x}^i)^2}{\hat{r}^2}\right],$$

and we have $\lim_{|t|/\hat{r}}\omega = 1$, so that $\hat{g}_{\mu\nu}$ can be continuously extended to the set $\hat{t}^2\leqslant\hat{r}^2$, $0\leqslant\hat{r}\leqslant\hat{r}_0$ with some $\hat{r}_0$. The metric $\hat{g}_{\mu\nu}$ is of the form

$$\hat{g}_{\mu\nu} = \eta_{\mu\nu} + \hat{r}k_{\mu\nu}(\theta,\phi,\hat{t}/\hat{r}) + o(\hat{r}) \tag{11}$$

and we have

$$|\hat{g}_{\mu\nu} - \eta_{\mu\nu}|\leqslant C\hat{r} \tag{12}$$

for some constant C; the derivatives of $\hat{g}_{\mu\nu}$ are, however, not

bounded up to the light cone of $i_0 = \{x^\alpha = 0\}$: One finds

$$\partial\omega \sim \ln(1 - \hat{t}^2/\hat{r}^2),$$

where ω is defined in (10), so that one obtains

$$|\partial_\sigma \hat{g}_{\mu\nu}| \leqslant C_1(\hat{t}/\hat{r}),$$

with $C_1(\eta) = C[|\ln(1-\eta^2)| + 1]$ for some constant C. It must be emphasized that the singularity of $C_1(\eta)$ at $\eta = 1$ is rather mild in the sense that

$$\int_0^1 C_1(\eta)d\eta < \infty.$$

We also have

$$\hat{r}|\partial_\sigma\,\partial_\rho\,\hat{g}_{\mu\nu}| \leqslant C_2(\hat{t}/\hat{r}), \quad C_2(\eta) \sim (1-\eta^2)^{-1},$$

with

$$\int_0^\theta C_2(\eta)d\eta \leqslant C' C_1(\theta).$$

This analysis[16] motivates the following definition.

Definition 2: Here $(\bar{M},\bar{g},i_0)$ will be called a *strong completion* of a space-time (M,g) if $\bar{M}$ is the disjoint union $M \cup \{i_0\}$ and the following holds.

(i) For all $p \in M$ there exists no timelike curve in $\bar{M}$ from p to i_0.

(ii) There exists a coordinate system $x^\alpha \in W_{1,\epsilon_x} \cup \{0\}$ such that $i_0 = 0$ and the points $x^\alpha x_\alpha > 0$ correspond to points in M. On M there exists a function $\Omega: M \to \mathbb{R}$, $\Omega > 0$ such that $\bar{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}$ and there exists a constant C and nondecreasing functions $C_1, C_2: [0,1) \to \mathbb{R}^+$ such that

$$\begin{aligned} &\forall x \in \Omega\, W_{1,\epsilon_x}: |\bar{g}_{\mu\nu} - \eta_{\mu\nu}| \leqslant Cr, \quad |\partial_\sigma\, \bar{g}_{\mu\nu}| \leqslant C_1(|\eta|), \\ &r|\partial_\sigma\,\partial_\rho\,\bar{g}_{\mu\nu}| \leqslant C_2(|\eta|), \quad \eta \equiv t/r, \end{aligned} \tag{13}$$

with

$$\int_0^\theta C_2(\eta)d\eta \leqslant C_1(\theta), \quad \int_0^1 C_1(\eta)d\eta = \bar{C}_1 < \infty.$$

(iii) Ω satisfies

$$\lim_{x\to 0} \Omega = 0, \quad \lim_{x\to 0} \partial_\mu\, \Omega = 0, \text{ for } x \neq 0$$

$$|\partial_\mu\,\partial_\nu\,\Omega - 2\bar{g}_{\mu\nu}| \leqslant Cr.$$

(iv) For every $|\eta| < 1$ the limits

$$\begin{aligned} &\lim_{r\to 0} [\bar{g}_{\mu\nu}(t = \eta r, r, \theta, \phi) - \eta_{\mu\nu}]/r, \\ &\lim_{r\to 0} \partial_\sigma\, \bar{g}_{\mu\nu}(t = \eta r, r, \theta, \phi), \\ &\lim_{r\to 0} [r\,\partial_\alpha\,\partial_\beta\,\bar{g}_{\mu\nu}(t = \eta r, r, \theta, \phi)] \end{aligned} \tag{14}$$

exist and are continuous functions of (η,θ,ϕ), where θ and ϕ are standard spherical angles.

Let us note that (13) implies that the Ricci tensor $\bar{R}_{\alpha\beta}$ of $\bar{g}_{\mu\nu}$ satisfies

$$|\bar{R}_{\alpha\beta}| \leqslant C_{\mathrm{R}}(\eta) r^{-1},$$

$$\bar{C}_{\mathrm{R}}(\theta) = \int_0^{|\theta|} C_{\mathrm{R}}(\eta)d\eta, \quad \int_0^1 \bar{C}_{\mathrm{R}}(\theta)d\theta = \bar{\bar{C}}_{\mathrm{R}} < \infty$$

$$(C_{\mathrm{R}}(\eta) \leqslant c[C_2(|\eta|) + C_1^2(|\eta|)]; \text{ therefore, } \bar{C}_{\mathrm{R}}(\theta) \leqslant c[C_1(|\theta|) + C_1(|\theta|)\textstyle\int_0^{|\theta|} C_1(\eta)d\eta] = c[1 + \bar{C}_1]C_1(|\theta|)$$

with some numerical factor c). Our main result is the following theorem.

Theorem 2 (quasiuniqueness of geodesically regular strong completions): Let $(\mathscr{M}_1,g_1,i_1)$, $(\mathscr{M}_2,g_2,i_2)$ be two geodesically regular strong completions of a space-time (M,g) and let $\{x^\alpha\}$, $x \in W_{1,\epsilon_x} \cup \{0\}$ and $\{y^\alpha\}$, $y \in W_{1,\epsilon_y} \cup \{0\}$ be the appropriate coordinate systems in $\mathscr{M}_1$ and $\mathscr{M}_2$. There exists a Lorentz matrix Λ^μ_ν, a constant vector C^μ, and a constant C such that for $x \in W_{1,\epsilon_x}$,

$$\begin{aligned} &y^\alpha = \Lambda^\alpha_\beta x^\beta + (C^\alpha x^\mu x_\mu - 2x^\alpha x^\mu C_\mu)\ln r + \zeta^\alpha, \\ &|\zeta^\mu| \leqslant Cr^2, \quad |\partial_\nu\,\zeta^\mu| \leqslant Cr. \end{aligned} \tag{15}$$

Moreover, for all $|\eta| < 1$ the limits

$$\begin{aligned} &\lim_{r\to 0} [\zeta^\mu(t = r\eta, r, \theta, \phi) r^{-2}], \\ &\lim_{r\to 0} [\partial_\alpha\,\zeta^\mu(t = r\eta, r, \theta, \phi) r^{-1}], \\ &\lim_{r\to 0} \partial_\alpha\,\partial_\beta\,\zeta^\mu(t = r\eta, r, \theta, \phi), \\ &\lim_{r\to 0} [r\,\partial_\alpha\,\partial_\beta\,\partial_\gamma\,\zeta^\mu(t = r\eta, r, \theta, \phi)] \end{aligned}$$

exist and are continuous functions of (η,θ,ϕ).

Proof: By Theorem 1 there exists a Lorentz matrix Λ^μ_ν such that for $x \in N_{\epsilon_x} = \{x^\alpha : x^0 = 0, r \leqslant \epsilon_x\}$,

$$y^\alpha = \Lambda^\alpha_\beta x^\beta + \zeta^\alpha, \quad |\zeta^\alpha| \leqslant Cr^{2-\epsilon}, \quad |\partial_\beta \zeta^\alpha| \leqslant Cr^{1-\epsilon}, \tag{16}$$

with any $\epsilon > 0$ and, in fact, a straightforward extension of the estimates of Ref. 3 leads to

$$\begin{aligned} &\forall x \in N_{\epsilon_x}: \\ &\zeta^\mu = (C^\mu \mathbf{x}^2 - 2x^\mu \mathbf{C}\mathbf{x})\ln r + \mathring{\zeta}^\mu(\theta,\phi)r^2 + o(r^2), \\ &\partial_\nu \zeta^\alpha = \partial_\nu[(C^\alpha x^\mu x_\mu - 2x^\alpha x^\mu C_\mu)\ln r + \mathring{\zeta}^\alpha r^2] + o(r), \\ &\partial_\rho\,\partial_\sigma\,\zeta^\alpha = \partial_\rho\,\partial_\sigma[(C^\alpha x^\mu x_\mu - 2x^\alpha x^\mu C_\mu)\ln r \\ &\qquad + \mathring{\zeta}^\alpha r^2] + o(1) \end{aligned} \tag{17}$$

for some constant vector C^μ. By a slight abuse of notation let us denote by y^α the coordinates $\Lambda^\alpha_\beta y^\beta$, so that we can set $\Lambda^\mu_\nu = \delta^\mu_\nu$ in Eq. (16), and on N_{ϵ_x} we have

$$\begin{aligned} &\tau_{|t=0} \equiv y^0_{|t=0} = O(r^2 \ln r), \\ &r(y(x))_{|t=0} = r(x) + O(r^2 \ln r), \\ &\partial_\mu\, y^\alpha_{|t=0} = \delta^\alpha_\mu + O(r\ln r), \\ &\partial_\alpha\,\partial_\beta y^\mu_{|t=0} = O(\ln r). \end{aligned} \tag{18}$$

Now $g^1_{\mu\nu}$ and $g^2_{\mu\nu}$ are conformally related to each other, so that by definition, there exists a function Φ: $W_{1,\epsilon_x} \to \mathbb{R}^+$ such that

$$\forall x^\alpha, x^\alpha x_\alpha > 0: \Phi^2(x) g^1_{\mu\nu}(x) = g^2_{\alpha\beta}(y(x)) \frac{\partial y^\alpha}{\partial x^\mu} \frac{\partial y^\beta}{\partial x^\nu}. \tag{19}$$

Equations (13), (18), and (19) yield

$$\Phi = \left\{\frac{\det g^2_{\mu\nu}}{\det g^1_{\mu\nu}}\right\}^{1/8} \left\{\det\left(\frac{\partial y^\beta}{\partial x^\nu}\right)\right\}^{1/4} \tag{20}$$

$$\Rightarrow \Phi_{|t=0} = 1 - 2\mathbf{C}\mathbf{x}\ln r + \mathring{\Phi}(\theta,\phi)r + o(r),$$

$$\partial_\mu\, \Phi_{|t=0} = -2C_\mu \ln r + \chi_\mu(\theta,\phi) + o(1) \tag{21}$$

for some functions $\mathring{\Phi}(\theta,\phi)$, $\chi_\mu(\theta,\phi)$. The transformation law of the Christoffel symbols gives

$$\partial_\alpha\partial_\beta y^\mu = [\Gamma^{1\lambda}_{\alpha\beta} + \Phi^{-1}(\delta^\lambda_\alpha\partial_\beta\Phi + \delta^\lambda_\beta\partial_\alpha\Phi - g^1_{\alpha\beta}g_1^{\lambda\sigma}\partial_\sigma\Phi)]\frac{\partial y^\mu}{\partial x^\lambda} - \Gamma^{2\mu}_{\rho\sigma}(y(x))\frac{\partial y^\rho}{\partial x^\alpha}\frac{\partial y^\sigma}{\partial x^\beta}, \tag{22}$$

where Γ^1 and Γ^2 are the Christoffel symbols of the metrics $g^1_{\mu\nu}$ and $g^2_{\mu\nu}$. The formula for the transformation of the Ricci tensor under conformal transformations reads as

$$\frac{\partial^2\Phi}{\partial x^\mu\partial x^\nu} = \Gamma^{1\lambda}_{\mu\nu}\frac{\partial\Phi}{\partial x^\lambda} + \frac{1}{2}\left\{\Phi\left[R^{\,1}_{\mu\nu} - R^{\,2}_{\alpha\beta}(y(x))\frac{\partial y^\alpha}{\partial x^\mu}\frac{\partial y^\beta}{\partial x^\nu}\right] + 4\Phi^{-1}\frac{\partial\Phi}{\partial x^\mu}\frac{\partial\Phi}{\partial x^\nu} + \left[\frac{1}{6}\Phi^3R_2(y(x)) - \frac{1}{6}\Phi R_1 - \Phi^{-1}g_1^{\alpha\beta}\frac{\partial\Phi}{\partial x^\alpha}\frac{\partial\Phi}{\partial x^\beta}\right]g^1_{\mu\nu}\right\}. \tag{23}$$

By (18) and (21) one can find $\epsilon_0 \leqslant \min(\epsilon_x, e^{-1})$ (e is the Euler number) small enough so that for all $x \in N_{\epsilon_0}$ we have

$$\tfrac{3}{4} \leqslant \Phi_{|t=0} \leqslant \tfrac{5}{4}, \quad 3r(x)/4 \leqslant r(y(x))_{|t=0} \leqslant 5r(x)/4,$$
$$\left|\frac{\partial y^\sigma}{\partial x^\beta} - \delta^\sigma_\beta\right|_{|t=0} \leqslant \frac{1}{4}, \quad \left|\frac{\partial\tau}{\partial t} - \frac{\tau(y(x))}{r(y(x))}\frac{\partial r(y)}{\partial t}\right|_{|t=0} \geqslant \frac{3}{4}.$$

Let

$$C_\Phi \equiv \sup_{\mu,x\in N_{\epsilon_0}} |[\ln r]^{-1}\partial_\mu\Phi|,$$

for $\eta\in[0,1)$,

$$C_{R_1}(\eta) \equiv \sup_{\mu,\nu,x\in W_{\eta,\epsilon_x}} r(x)|R^{\,1}_{\mu\nu}(x)| \quad \text{and} \quad C_{R_1}(-\eta) \equiv C_{R_1}(\eta),$$

for $\eta\in[0,1)$,

$$C_{R_2}(\eta) \equiv \sup_{\mu,\nu,y\in W_{\eta,\epsilon_y}} r(y)|R^{\,2}_{\mu\nu}(y)| \quad \text{and} \quad C_{R_2}(-\eta) \equiv C_{R_2}(\eta),$$

$$C_{g_1} \equiv \sup_{\mu,\nu,x\in W_{1,\epsilon_x}} |g^1_{\mu\nu}(x)| + \sup_{\mu,\nu,x\in W_{1,\epsilon_x}} |g_1^{\mu\nu}(x)|,$$

$$C_{g_2} \equiv \sup_{\mu,\nu,y\in W_{1,\epsilon_y}} |g^2_{\mu\nu}(y)| + \sup_{\mu,\nu,y\in W_{1,\epsilon_y}} |g_2^{\mu\nu}(y)|,$$

for $a = 1,2$,

$$\overline{C}_{R_a}(\theta) = \int_0^{|\theta|} C_{R_a}(\eta)d\eta, \quad \overline{\overline{C}}_{R_a} = \int_0^1 \overline{C}_{R_a}(\eta)d\eta\,(<\infty).$$

Let $\Omega \subset W_{1,\epsilon_0}$ be the set of points such that

$$\begin{aligned}
&|\partial_\mu\Phi| < \widetilde{C}_\Phi(x) + C_\Phi\ln(1/r),\\
&\tfrac{1}{2} < \Phi < 2,\\
&r(x)/2 < r(y(x)) < 3r(x)/2,\\
&\left|\frac{\partial y^\mu}{\partial x^\nu} - \delta^\mu_\nu\right| < \frac{1}{2},\\
&\left|\frac{\partial\tau}{\partial t} - \frac{\tau(y(x))}{r(y(x))}\frac{\partial r(y)}{\partial t}\right| > \frac{1}{2},
\end{aligned} \tag{24}$$

where

$$\begin{aligned}
\widetilde{C}_\Phi(x) :&= (2 + 2^4 3^{-1}C^2_{g_1})\overline{C}_{R_1}(\eta) + (2^4 3^3 + 2^7 C_{g_1}C_{g_2})\overline{C}_{R_2}(\tilde\eta(x)) + 1 =: a\overline{C}_{R_1}(\eta)\\
&\quad + b\overline{C}_{R_2}(\tilde\eta(x)) + 1, \quad \tilde\eta(x) \equiv \tau(x)/r(y(x)).
\end{aligned}$$

Let $\Omega_1' \subset \Omega$ be the set of points x^α such that the curve $[0,t]\ni s\to(s,\mathbf{x})$ is included in Ω. We have $N_{\epsilon_0}\subset\Omega_1'$; let Ω_1 be the connected component of Ω_1' which contains N_{ϵ_0}. We shall show that there exists $0<\epsilon_1\leqslant\epsilon_0$ such that $\Omega_1\cap W_{1,\epsilon_1}$ is closed in W_{1,ϵ_1}. By (23) we have, for $x^\alpha\in\Omega_1$,

$$\begin{aligned}
\left|\frac{\partial}{\partial t}\frac{\partial\Phi}{\partial x^\mu}\right| \leqslant{}& [(1 + 2^3 3^{-1}C^2_{g_1})C_{R_1}(\eta) + (2^3 3^2 + 2^6 3^{-1}C_{g_1}C_{g_2})C_{R_2}(\tilde\eta)]r(x)^{-1}\\
&+ (2^2 + 2^4 C^2_{g_1})\left\{\widetilde{C}_\Phi(x) + C_\Phi\ln\left[\frac{1}{r(x)}\right]\right\}^2\\
&+ 2^3 3 C_{g_1}C^x_1(\eta)\left\{\widetilde{C}_\Phi(x) + C_\Phi\ln\left[\frac{1}{r(x)}\right]\right\},
\end{aligned}$$

where C^x_1 is the function C_1 in Eq. (13) for the coordinate system x^α, extended to negative η by $C^x_1(-\eta)\equiv C^x_1(\eta)$. From

$$\left|\frac{\partial\Phi}{\partial x^\mu}(t,\mathbf{x})\right| \leqslant \left|\frac{\partial\Phi}{\partial x^\mu}(0,\mathbf{x})\right| + \int_0^t \left|\frac{\partial^2\Phi}{\partial t\,\partial x^\mu}(s,\mathbf{x})\right| ds,$$

and from

$$\int_0^t C_{R_1}\left(\frac{s}{r}\right)ds = r\overline{C}_{R_1}\left(\frac{t}{r}\right),$$

$$\begin{aligned}
&\int_0^t C_{R_2}\left[\frac{\tau(s,\mathbf{x})}{r(y(s,\mathbf{x}))}\right]ds\\
&\quad = \int_0^{\tilde\eta(x)} C_{R_2}(\eta)\frac{r(y)}{\partial\tau/\partial t - [\tau/r(y)][\partial r(y(x))/\partial t]}d\eta\\
&\quad \leqslant 3r(x)\int_0^{\tilde\eta(x)} C_{R_2}(\eta)d\eta\\
&\quad = 3r(x)\overline{C}_{R_2}\left(\frac{\tau(x)}{r(y(x))}\right),
\end{aligned}$$

$$\begin{aligned}
&\int_0^t\left[\widetilde{C}_\Phi(s,\mathbf{x}) + C_\Phi\ln\left(\frac{1}{r}\right)\right]^2 ds\\
&\quad \leqslant 2\int_0^t\left[\widetilde{C}^2_\Phi(s,\mathbf{x}) + C^2_\Phi\ln^2\left(\frac{1}{r}\right)\right]ds\\
&\quad \leqslant 2\widetilde{C}_\Phi(x)(a\overline{\overline{C}}_{R_1} + 3b\overline{\overline{C}}_{R_2} + 1)r(x) + 2C^2_\Phi r\ln^2\left(\frac{1}{r}\right)
\end{aligned}$$

we have used the facts that $\widetilde{C}_\Phi$ is nondecreasing along the curves $[0,t)\ni s\to(s,\mathbf{x})$, and $|t|<r$ in W_{1,ϵ_1}],

$$\int_0^t \widetilde{C}_\Phi(s,\mathbf{x})C^x_1\left(\frac{s}{r}\right)ds \leqslant r\overline{C}^x_1\widetilde{C}_\Phi(x), \quad \overline{C}^x_1 \equiv \int_0^1 C^x_1(\eta)d\eta,$$

one obtains

$$\begin{aligned}
&|\partial_\mu\Phi(t,\mathbf{x})|\\
&\quad \leqslant C_\Phi\ln\left(\frac{1}{r}\right) + (2^3 + 2^5 C^2_{g_1})\epsilon_1\ln^2\left(\frac{1}{\epsilon_1}\right)C^2_\Phi\\
&\qquad + 2^3 3 C_{g_1}\overline{C}^x_1\epsilon_1\ln\left(\frac{1}{\epsilon_1}\right)C_\Phi + \left\{\frac{1}{2} + [(2^3 + 2^5 C^2_{g_1})(a\overline{\overline{C}}_{R_1} + 3b\overline{\overline{C}}_{R_2} + 1)\right.
\end{aligned}$$

$$+ 2^3 3 C_{g_1} \overline{C}_1^x]\epsilon_1 \Big\} \widetilde{C}_\Phi(x)$$

$$\leqslant C_\Phi \ln\left(\frac{1}{r}\right) + \frac{1}{8} + \frac{5\widetilde{C}_\Phi(x)}{8}$$

$$\leqslant C_\Phi \ln\left(\frac{1}{r}\right) + \frac{3\widetilde{C}_\Phi(x)}{4}$$

for ϵ_1 small enough. This gives

$$|\Phi(t,\mathbf{x}) - 1| \leqslant |\Phi(0,\mathbf{x}) - 1| + \int_0^t |\partial_t \Phi(s,\mathbf{x})| ds$$

$$\leqslant \frac{1}{4} + (a\overline{\overline{C}}_{R_1} + 3b\overline{\overline{C}}_{R_2} + 1)\epsilon_1$$

$$+ C_\Phi \epsilon_1 \ln\left(\frac{1}{\epsilon_1}\right) \leqslant \frac{3}{8},$$

decreasing ϵ_1 if necessary. One shows in a similar way that none of the inequalities in (24) can saturate for $x^\alpha \in \Omega_1 \cap W_{1,\epsilon_1}$ if ϵ_1 is small enough, so that $\Omega_1 \cap W_{1,\epsilon_1}$ is both open and closed in W_{1,ϵ_1}; therefore, $\Omega_1 \cap W_{1,\epsilon_1} = W_{1,\epsilon_1}$. It is not too difficult to show from (22) and (24) that Φ and $\partial y^\mu / \partial x^\nu$ uniformly tend to 1 and δ^μ_ν at i_0 and (23) yields

$$\lim_{r\to 0}\left[\frac{\partial \Phi}{\partial x^\lambda}(t = r\eta, r, \theta, \phi) - \frac{\partial \Phi}{\partial x^\lambda}(0, r, \theta, \phi)\right]$$

$$= \lim_{r\to 0} \int_0^{\eta r} \frac{\partial^2 \Phi}{\partial t\, \partial x^\lambda} ds =: A_\lambda(\eta, \theta, \phi), \tag{25}$$

with some continuous functions $A_\lambda(\eta,\theta,\phi)$, so that from (21) and (25) one has

$$\frac{\partial \Phi}{\partial x^\lambda}(t = r\eta, r, \theta, \phi) = -2C_\lambda \ln r + \widetilde{A}_\lambda(\eta,\theta,\phi) + o(1)$$

for some continuous functions $\widetilde{A}_\lambda$ and (17) and (22) imply

$$\partial_\beta y^\alpha = \delta^\alpha_\beta + 2(C^\alpha x_\beta - C_\beta x^\alpha - \delta^\alpha_\beta C^\mu x_\mu)$$

$$\times \ln r + r A^\alpha_\beta(\theta,\phi,\eta) + o(r),$$

with some continuous functions A^α_β; a straightforward analysis establishes our remaining claims.

Two completions differing by a transformation of the form (15) with $C^\mu = 0$ can be considered as equivalent. Theorem 2 and Proposition 1 imply[17] the following corollary.

Corollary 1: Strong conformal completions of no-radiation space-times are unique up to the four-parameter family of transformations (15).

It may be of some relevance to note that the logarithmic transformations (15) are not the ones given in Appendix 1 of Ref. 18: The latter introduce singularities in $g_{\mu\nu}$-$\eta_{\mu\nu}$ at the light cone of i_0, while (15) do not.

It is natural to ask about the group properties of the transformations (15) since r is not Lorentz invariant: Under a Lorentz transformation $y^\alpha = \Lambda^\alpha_\beta x^\beta$ we have

$$r(y) = \sqrt{\Lambda^i_\mu x^\mu \Lambda^i_\nu x^\nu}$$

$$= r(x)\sqrt{\Lambda^i_\mu (x^\mu/r)\Lambda^i_\nu (x^\nu/r)} =: f(\eta,\theta,\phi) r(x),$$

so that

$$\ln r(y) = \ln r(x) + \ln f$$

and the $\ln f$ terms can be absorbed in ζ^μ, which shows that a composition of transformations (15) is still of the form (15).

Let us finally note the existence of a set of coordinates for the Schwarzschild metric in which the metric is slightly worse behaved at i_0 than in (11) and (12); however, for $r \neq 0$ the first derivatives of the metric do not blow up as one approaches the light cone of i_0. Let us set

$$r^* = f(r), \quad r^* = \tfrac{1}{2}[\tilde{f}(\tilde{v}^{-1}) + \tilde{f}(\tilde{w}^{-1})],$$

$$t = \tfrac{1}{2}[\tilde{f}(\tilde{v}^{-1}) - \tilde{f}(\tilde{w}^{-1})], \quad \tilde{r} = \tfrac{1}{2}[\tilde{v} + \tilde{w}],$$

$$\tilde{t} = \tfrac{1}{2}[\tilde{v} - \tilde{w}], \quad \tilde{f}(s) = s + 4m \ln(s/2m),$$

with r, t as in (8) [this choice of $\tilde{f}$ cancels these terms in $r(\tilde{v},\tilde{w})$ which exhibit the worst behavior at the light cone of i_0]. One obtains

$$ds^2 = \widetilde{\Omega}^{-2} d\tilde{s}^2, \quad \widetilde{\Omega} = \tilde{v}\tilde{w}, \quad d\tilde{s}^2 = \tilde{g}_{\mu\nu}\, d\tilde{x}^\mu\, d\tilde{x}^\nu,$$

$$|\tilde{g}_{\mu\nu} - \eta_{\mu\nu}| \leqslant \widetilde{C}\tilde{r} \ln \tilde{r}, \quad |\partial_\sigma \tilde{g}_{\mu\nu}| \leqslant \widetilde{C} \ln \tilde{r}$$

for some constant $\widetilde{C}$. An unpleasant feature of these coordinates is the logarithmic blowing up of $(\tilde{g}_{\mu\nu} - \eta_{\mu\nu})/\tilde{r}$ at $\tilde{r} = 0$; however, we have

$$\tilde{g}_{\mu\nu} = \eta_{\mu\nu} + \tilde{r}[\tilde{h}^1_{\mu\nu}(\tilde{\eta},\theta,\phi)\ln \tilde{r}$$

$$+ \tilde{h}^2_{\mu\nu}(\tilde{\eta},\theta,\phi)] + o(\tilde{r}), \quad \tilde{\eta} = \tilde{t}/\tilde{r},$$

which is of simple and tractable form.

IV. CONCLUSIONS

We have shown uniqueness "up to logarithmic ambiguities" of Ashtekar–Hansen[4,5] completions satisfying a geodesic condition. We conjecture that the geodesic regularity condition is unnecessary in the case of strong completions and that it cannot be removed without losing quasiuniqueness of the weak completions. It may be of some interest to mention that in Theorems 1 and 2 the geodesic regularity hypothesis may be replaced by the probably much weaker condition that there exists a spacelike geodesic Γ extending to i_0 such that $\Gamma \subset W^y_{\theta_y,\epsilon_y}$, $\Gamma \subset W^x_{\theta_x,\epsilon_x}$ with some $0 \leqslant \theta_x$, $\theta_y < 1$, where $W^y_{\theta_y,\epsilon_y}$ and $W^x_{\theta_x,\epsilon_x}$ are appropriate y and x coordinate wedges. This condition does not, however, characterize some completion, but pairs of completions. It is likely that a proof of quasiuniqueness of strong conformal completions without any further conditions can be obtained by showing that such geodesics always exist.

It must be stressed that the Kerr family of metrics exhausts the up-until-now known set of vacuum Einstein metrics admitting strong completions of spatial infinity, so that the results of Sec. III cover all actually known physically relevant examples: The metrics recently constructed by Cutler and Wald[19] or Christodoulou[20] are "Schwarzschild in a neighborhood of i_0," so clearly our theorems apply. It seems rather difficult to guess whether there exists some sufficiently large class of vacuum space-times admitting weak or strong completions, the existence of which would justify the need of a search for more general results than those presented here.

An important consequence of our results is that one can assign an invariant mass parameter (cf., e.g., Ref. 21) to every vacuum space-time admitting strong or weak geodesi-

cally regular completions, $\alpha > \frac{1}{2}$—this complements the results presented in Ref. 3.

ACKNOWLEDGMENTS

Useful discussions with A. Ashtekar are acknowledged. The author enjoyed the friendly hospitality of the Center for Mathematical Analysis, Canberra, Australia, during part of the work on this paper.

This work was supported in part by NSF Grant No. PHY-8503072 to Yale University.

[1]P. T. Chruściel, J. Math. Phys. **30**, 2090 (1989).
[2]R. Geroch, J. Math. Phys. **13**, 956 (1972); also in *Asymptotic Structure of Space-Time*, edited by P. Esposito and L. Witten (Plenum, New York, 1976).
[3]P. T. Chruściel, Commun. Math. Phys. **120**, 233 (1988).
[4]A. Ashtekar and R. O. Hansen, J. Math. Phys. **19**, 1542 (1978).
[5]A. Ashtekar, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. 2.
[6]D. Christodoulou and N. O'Murchadha, Commun. Math. Phys. **80**, 27 (1981).
[7]N. O'Murchadha, J. Math. Phys. **27**, 2111 (1986).
[8]Latin indices run from 1–3 and Greek indices run from 0–3, $r(x) \equiv \{\Sigma_i (x^i)^2\}^{1/2}$.
[9]Throughout, $x^\alpha x_\alpha \equiv \eta_{\mu\nu} x^\mu x^\nu$, $y^\alpha y_\alpha \equiv \eta_{\mu\nu} y^\mu y^\nu$.
[10]Note that with this definition $0 \notin W_{\theta,1/R}$.
[11]$C(\theta)$ can blow up to ∞ as θ goes to 1.
[12]A related stronger condition would be the requirement that there exists a parametrization $u(s)$ such that the limit $\lim_{s\to\infty} dy^\mu(s(u))/du = \eta^\mu$ exists, with $\eta_{\mu\nu}\eta^\mu\eta^\nu > 0$.
[13]By complete geodesic we mean a geodesic defined for all $s \geqslant s_0$ for some s_0 where s is an affine parameter.
[14]$S(r)$ = sphere of radius r.
[15]It is simple, although a little tedious, to show that the assignment $\mathbf{n} \to \theta_{\mathbf{n}}^{\pm}$ can be made in a continuous way.
[16]A similar analysis (cf. Appendix C of Ref. 4) shows that the Kerr metrics also satisfy the requirements of the definition below.
[17]It should be borne in mind that some no-radiation metrics may admit no strong conformal completions—Corollary 1 classifies all of them if one exists. It has been pointed out to the author by B. Schmidt that it is likely that every no-radiation metric satisfying vacuum Einstein equations is a Kerr metric (which admits strong completions), but no field equations are assumed in Theorem 2 or Proposition 1.
[18]A. Ashtekar, in *General Relativity and Gravitation*, edited by B. Bertotti, F. de Felice, and A. Pascolini (Reidel, Dordrecht, 1984).
[19]C. Cutler and R. M. Wald, Class. Quantum Grav. **6**, 453 (1989); the Cutler–Wald construction of global solutions of Einstein–Maxwell equations can be generalized to Einstein–Yang–Mills equations (R. Bartnik, private communication).
[20]D. Christodoulou, Commun. Math. Phys. **105**, 337 (1986); **106**, 587 (1986); **109**, 591 (1987); **109**, 613 (1987). In these papers existence of global solutions is established only in the interior of the forward light cone of a point. It has, however, been pointed out to the author by D. Christodoulou that it is easy to prove that time symmetric, spherically symmetric, Schwarzschildian outside a compact set Cauchy data on a spacelike hypersurface can be evolved to a set that contains a complete light cone of the origin. If the norm of these data is small the norm of the resulting data on the light cone will be small as well, thus yielding, by the results proved in the above cited papers, a global solution of Einstein equations for a metric interacting with a scalar field.
[21]A. Ashtekar, Found. Phys. **15**, 419 (1985); P. T. Chruściel, in *Proceedings of the Canberra Miniconference on Mathematical General Relativity*, edited by R. Bartnik, Yale preprint YCTP-G2-89, to appear.

Two-sided conformally recurrent four-dimensional Riemannian manifolds

J. F. Plebański[a] and M. Przanowski[b]
Departamento de Física, Centro de Investigación y de Estudios Avanzados del I.P.N., Apartado Postal 14-740, 07000 México, D.F., Mexico

(Received 6 July 1988; accepted for publication 29 March 1989)

Two-sided conformally recurrent four-dimensional Riemannian manifolds are defined and analyzed from the point of view of the Petrov–Penrose classification. All two-sided conformally recurrent four-dimensional Riemannian manifolds of types $D\otimes D$, $N\otimes N$, and $N\otimes[\,-\,]$ are given.

I. INTRODUCTION

Let M be a C^∞ four-dimensional real differentiable manifold or a four-dimensional complex analytic differentiable manifold endowed with a C^∞ real or holomorphic, respectively, metric g.

Then (M,g) is said to be a two-sided conformally recurrent Riemannian manifold if there exist spinors $r_{A\dot B}$ and $\dot r_{A\dot B}$ such that

$$\nabla_{E\dot F}C_{ABCD}=r_{E\dot F}C_{ABCD}\,, \tag{1.1a}$$

$$\nabla_{E\dot F}C_{A\dot B\dot C\dot D}=\dot r_{E\dot F}C_{A\dot B\dot C\dot D}\,, \tag{1.1b}$$

and, moreover,

$$\sum_{A,B,C,D}(|C_{ABCD}|+|C_{A\dot B\dot C\dot D}|)\neq 0, \tag{1.1c}$$

where C_{ABCD} and $C_{A\dot B\dot C\dot D}$ are "undotted" and "dotted" Weyl spinors, respectively. (In what follows "$f\neq 0$" means "f is nowhere vanishing.") If (M,g) is a space-time of Einsteinian general relativity, i.e., a four-dimensional Riemannian manifold with metric g of the Lorentzian signature, then our definition contains the complex recurrent spaces defined by McLenaghan and Leroy.[1] The cited authors have found and analyzed in detail all two-sided conformally recurrent Riemannian manifolds with Lorentzian metrics. It is worth pointing out that the space-times of *pp* waves are two-sided conformally recurrent Riemannian manifolds.[1–4]

The purpose of our paper is to examine all two-sided conformally recurrent Riemannian manifolds. It may seem to be an "academic problem" only. But it is not so because of great interest in the complex relativity and gravitational instantons; and, as will be shown, there are some essentially new solutions of the type $N\otimes[\,-\,]$ that define the complex (or ultrahyperbolic) space-times. The formalism we use is the spinorial one, which seems to be the most convenient for our purpose. The spinorial formalism for all four-dimensional real or complex Riemannian manifolds has been amply described in Refs. 5 and 6. Here we will only outline the basic facts.

The four-dimensional Riemannian manifolds can be classified according to the scheme

"complex relativity" (CR):

M complex analytic, g holomorphic;

"real relativity" (RR):

M real, g real of

$$\text{signature}\begin{cases}(+\,+\,+\,-)\text{ or }(-\,-\,-\,+)\text{: HR},\\(+\,+\,-\,-)\text{: UR},\\(+\,+\,+\,+)\text{ or }(-\,-\,-\,-)\text{: ER},\end{cases}$$

where HR, UR, or ER stand for "hyperbolic relativity," "ultrahyperbolic relativity," or "Euclidean relativity," respectively. In all cases we postulate the metric in the form of

$$g=-\tfrac{1}{2}g_{A\dot B}\underset{s}{\otimes}g^{A\dot B},\quad A=1,2,\quad \dot B=\dot 1,\dot 2, \tag{1.2}$$

where $g^{A\dot B}$ are one-forms constituting the components of the canonical spinor-valued one-form on M. Spinorial indices are to be manipulated according to the scheme

$$\Psi_1=\Psi^2,\quad \Psi_2=-\Psi^1;\quad \Psi_{\dot 1}=\Psi^{\dot 2},\quad \Psi_{\dot 2}=-\Psi^{\dot 1}. \tag{1.3}$$

The one-forms $g^{A\dot B}$ in CR are holomorphic; in RR (in general) they are complex valued and additionally endowed with the following properties under complex conjugation:

$$\begin{aligned}&\text{HR: }\ \overline{g^{A\dot B}}=g^{B\dot A}\ \text{ or }\ \overline{g^{A\dot B}}=-g^{B\dot A},\ \text{resp.},\\&\text{UR: }\ \overline{g^{A\dot B}}=g^{A\dot B},\\&\text{ER: }\ \overline{g^{A\dot B}}=-g_{A\dot B}\ \text{ or }\ \overline{g^{A\dot B}}=g_{A\dot B},\ \text{resp.}\end{aligned} \tag{1.4}$$

Components of contravariant spinors of the first rank are subject to the transformations

$$\Psi^{A'}=l^{A'}{}_{A}\Psi^{A},\quad \Psi^{\dot A'}=l^{\dot A'}{}_{\dot A}\Psi^{\dot A}, \tag{1.5}$$

with

$$\begin{aligned}&\text{CR: }\ \|l^{A'}{}_{A}\|,\|l^{\dot A'}{}_{\dot A}\|\in \mathrm{SL}(2;C);\\&\text{HR: }\ \|l^{A'}{}_{A}\|\in \mathrm{SL}(2;C),\quad \|l^{\dot A'}{}_{\dot A}\|=\overline{\|l^{A'}{}_{A}\|};\\&\text{UR: }\ \|l^{A'}{}_{A}\|,\|l^{\dot A'}{}_{\dot A}\|\in \mathrm{SL}(2;R);\\&\text{ER: }\ \|l^{A'}{}_{A}\|,\|l^{\dot A'}{}_{\dot A}\|\in \mathrm{SU}(2).\end{aligned}$$

The first Cartan structure equations read

$$Dg^{A\dot B}=dg^{A\dot B}+\Gamma^{A}{}_{C}\wedge g^{C\dot B}+\Gamma^{\dot B}{}_{\dot C}\wedge g^{A\dot C}=0, \tag{1.6}$$

where D is the exterior covariant differentiation and $\Gamma^{A}{}_{B}$ and $\Gamma^{\dot A}{}_{\dot B}$ are components of connection one-forms for the vector bundles of undotted or dotted, respectively, contra-

[a] On leave of absence from the University of Warsaw, Warsaw, Poland.
[b] Permanent address: Instytut Fizyki, Politechnika tódzka, Wólczańska 219, 93-005 Łódź, Poland.

 0022-2488/89/092101-09$02.50

variant spinors of the first rank. The one-forms $\Gamma^{A}{}_{B}$ and $\Gamma^{\dot{A}}{}_{\dot{B}}$ satisfy the conditions

$$\Gamma^{A}{}_{A}=0=\Gamma^{\dot{A}}{}_{\dot{A}}\,. \tag{1.7}$$

The second Cartan structure equations read

$$R^{A}{}_{B}=d\Gamma^{A}{}_{B}+\Gamma^{A}{}_{C}\wedge\Gamma^{C}{}_{B},\quad R^{\dot{A}}{}_{\dot{B}}=d\Gamma^{\dot{A}}{}_{\dot{B}}+\Gamma^{\dot{A}}{}_{\dot{C}}\wedge\Gamma^{\dot{C}}{}_{\dot{B}}, \tag{1.8}$$

where $R^{A}{}_{B}$ and $R^{\dot{A}}{}_{\dot{B}}$ are the components of curvature two-forms of $\Gamma^{A}{}_{B}$ and $\Gamma^{\dot{A}}{}_{\dot{B}}$, respectively.

One has the following decomposition:

$$\begin{aligned}R_{AB}&=-\tfrac{1}{2}C_{ABCD}S^{CD}+(R/24)S_{AB}+\tfrac{1}{2}C_{AB\dot{C}\dot{D}}S^{\dot{C}\dot{D}},\\ R_{\dot{A}\dot{B}}&=-\tfrac{1}{2}C_{\dot{A}\dot{B}\dot{C}\dot{D}}S^{\dot{C}\dot{D}}+(R/24)S_{\dot{A}\dot{B}}+\tfrac{1}{2}C_{CD\dot{A}\dot{B}}S^{CD},\end{aligned} \tag{1.9}$$

where

$$S^{AB}:=\tfrac{1}{2}\epsilon_{\dot{C}\dot{D}}g^{A\dot{C}}\wedge g^{B\dot{D}},\quad S^{\dot{A}\dot{B}}:=\tfrac{1}{2}\epsilon_{CD}g^{C\dot{A}}\wedge g^{D\dot{B}} \tag{1.10}$$

constitute a basis for self-dual and anti-self-dual, respectively, two-forms

$$\left(\|\epsilon_{AB}\|:=\begin{vmatrix}0&1\\-1&0\end{vmatrix}=:\|\epsilon_{\dot{A}\dot{B}}\|\right);$$

$C_{ABCD}=C_{(ABCD)}$ and $C_{\dot{A}\dot{B}\dot{C}\dot{D}}=C_{(\dot{A}\dot{B}\dot{C}\dot{D})}$ are the Weyl spinors; R is the curvature scalar; and $C_{AB\dot{C}\dot{D}}=C_{(AB)\dot{C}\dot{D}}=C_{AB(\dot{C}\dot{D})}$ is the spinor image of the traceless Ricci tensor.

These objects have the following properties with respect to the complex conjugation:

$$\begin{aligned}&\text{CR: } C_{ABCD},C_{\dot{A}\dot{B}\dot{C}\dot{D}},C_{AB\dot{C}\dot{D}},R \text{ complex};\\ &\text{HR: } C_{\dot{A}\dot{B}\dot{C}\dot{D}}=\overline{C_{ABCD}},\quad \overline{C_{AB\dot{C}\dot{D}}}=C_{CD\dot{A}\dot{B}},\quad \overline{R}=R;\\ &\text{UR: } C_{ABCD},C_{\dot{A}\dot{B}\dot{C}\dot{D}},C_{AB\dot{C}\dot{D}},R \text{ real};\\ &\text{ER: } \overline{C_{ABCD}}=C^{ABCD},\quad \overline{C_{\dot{A}\dot{B}\dot{C}\dot{D}}}=C^{\dot{A}\dot{B}\dot{C}\dot{D}},\\ &\qquad \overline{C_{AB\dot{C}\dot{D}}}=C^{AB\dot{C}\dot{D}},\quad \overline{R}=R.\end{aligned} \tag{1.11}$$

One then introduces the "spinorial gradient $\nabla_{A\dot{B}}$" as acting on spinors via

$$D\Psi^{\cdots}_{\cdots}=-\tfrac{1}{2}g^{A\dot{B}}\nabla_{A\dot{B}}\Psi^{\cdots}_{\cdots}\,. \tag{1.12}$$

With the help of this operation we can express the Bianchi identities $DR^{A}{}_{B}=0=DR^{\dot{A}}{}_{\dot{B}}$ in the form of

$$\begin{aligned}&\nabla^{C\dot{D}}C_{AC\dot{B}\dot{D}}+\tfrac{1}{8}\nabla_{A\dot{B}}R=0,\quad \nabla^{E}{}_{\dot{A}}C_{BCDE}+\nabla_{(B}{}^{\dot{E}}C_{CD)\dot{A}\dot{E}}=0,\\ &\nabla_{A}{}^{\dot{E}}C_{B\dot{C}D\dot{E}}+\nabla^{E}{}_{(B}C_{|EA|\dot{C}\dot{D})}=0,\end{aligned} \tag{1.13}$$

and the Ricci identities for one-index spinors in the form of

$$\begin{aligned}&\tfrac{1}{2}\nabla^{E}{}_{(\dot{C}}\nabla_{|E|\dot{D})}\Psi^{A}=\Psi^{E}C^{A}{}_{ECD}\,,\\ &\tfrac{1}{2}\nabla_{(\dot{C}}{}^{\dot{E}}\nabla_{D)\dot{E}}\Psi^{A}=\Psi^{E}\{-C^{A}{}_{ECD}+(R/12)\epsilon_{E(\dot{C}}\delta^{A}{}_{D)}\},\end{aligned} \tag{1.14}$$

and

$$\begin{aligned}&\tfrac{1}{2}\nabla_{(\dot{C}}{}^{\dot{E}}\nabla_{D)\dot{E}}\Psi^{\dot{A}}=\Psi^{\dot{E}}C_{CD}{}^{\dot{A}}{}_{\dot{E}},\\ &\tfrac{1}{2}\nabla^{E}{}_{(\dot{C}}\nabla_{|E|\dot{D})}\Psi^{\dot{A}}=\Psi^{\dot{E}}\{-C^{\dot{A}}{}_{\dot{E}C\dot{D}}+(R/12)\epsilon_{\dot{E}(\dot{C}}\delta^{\dot{A}}{}_{\dot{D})}\}.\end{aligned} \tag{1.15}$$

Knowing (1.14) and (1.15) one can easily find the Ricci identities for any spinor $\Psi^{A\cdots\dot{B}\cdots}$, applying (1.14) and (1.15) to each index in the additive manner.

The above-presented basic facts of the spinorial formalism are essential for our further considerations.

In Sec. II we find the Petrov–Penrose types of two-sided conformally recurrent four-dimensional Riemannian manifolds.

In Sec. III we consider the types $D\otimes D$ and $N\otimes N$.

In Sec. IV we present the integration of the type $N\otimes[-]$.

II. THE PETROV–PENROSE CLASSIFICATION OF TWO-SIDED CONFORMALLY RECURRENT RIEMANNIAN MANIFOLDS

We assume that the four-dimensional Riemannian manifold (M,g) is two-sided conformally recurrent and in the condition (1.1c) we assume, for definiteness,

$$\sum_{A,\dot{B},C,D}|C_{ABCD}|\neq 0. \tag{2.1}$$

Acting with $\frac{1}{2}\nabla^{E}{}_{(\dot{P}}\nabla_{|E|\dot{Q})}$ on C^{ABCD}, employing (1.1a) and the first of Eqs. (1.14), one obtains

$$\begin{aligned}&C^{ABCD}\omega_{\dot{P}\dot{Q}}=4C^{S(ABC}C^{D)}{}_{S\dot{P}\dot{Q}},\\ &\omega_{\dot{P}\dot{Q}}=\omega_{(\dot{P}\dot{Q})}:=\tfrac{1}{2}\nabla^{S}{}_{(\dot{P}}r_{|S|\dot{Q})}\,.\end{aligned} \tag{2.2}$$

Similarly, acting with $\frac{1}{2}\nabla_{(P}{}^{\dot{E}}\nabla_{Q)\dot{E}}$ on C^{ABCD}, using (1.1a) and the second of Eqs. (1.14), one has

$$\begin{aligned}-C^{ABCD}\omega_{PQ}&=4[C^{S(ABC}C^{D)}{}_{SPQ}\\ &\quad+(R/12)C^{(ABC}{}_{(P}\delta^{D)}{}_{Q)}],\\ \omega_{PQ}&=\omega_{(PQ)}:=\tfrac{1}{2}\nabla_{(P}{}^{\dot{E}}r_{Q)\dot{E}}\,.\end{aligned} \tag{2.3}$$

Define the one-form

$$r:=-\tfrac{1}{2}r_{A\dot{B}}g^{A\dot{B}}\,. \tag{2.4}$$

Then we have

$$dr=\tfrac{1}{2}(\omega_{AB}S^{AB}+\omega_{\dot{A}\dot{B}}S^{\dot{A}\dot{B}}). \tag{2.5}$$

From (2.5) it follows that

$$dr=0\Leftrightarrow\omega_{AB}=0=\omega_{\dot{A}\dot{B}}\,. \tag{2.6}$$

We now intend to prove that $\omega_{PQ}=0$.

Using (1.1a) one easily finds the following formulas:

$$d\overset{2}{C}=2\overset{2}{C}r,\quad d\overset{3}{C}=3\overset{3}{C}r, \tag{2.7}$$

where

$$\overset{2}{C}:C^{AB}{}_{CD}C^{CD}{}_{AB},\quad \overset{3}{C}:=C^{AB}{}_{CD}C^{CD}{}_{EF}C^{EF}{}_{AB}\,.$$

For Petrov–Penrose type I, $|\overset{2}{C}|+|\overset{3}{C}|\neq 0$, and for types II and D, $\overset{2}{C}\neq 0$; therefore $dr=0$ [by (2.7)] and consequently $\omega_{PQ}=0$ [by (2.6)]. For type N, with $C_{ABCD}=k_{A}k_{B}k_{C}k_{D}$, (2.3) obviously reduces to

$$-k^{A}k^{B}k^{C}k^{D}\omega_{PQ}=(R/3)k^{A}k^{B}k^{(C}k_{(P}\delta^{D)}{}_{Q)}\,. \tag{2.8}$$

Contracting this with k_{D} one obtains $R=0$, which, used back in (2.8), implies $\omega_{PQ}=0$.

Therefore, one has

$$\text{type N: } R=0,\quad \omega_{PQ}=0. \tag{2.9}$$

It remains only to examine type III. We show that this type is altogether incompatible with condition (2.3). Indeed, substituting into (2.3) $C^{ABCD}=k^{(A}k^{B}k^{C}l^{D)}$, with $k^{A}l_{A}\neq 0$, and then contracting (2.3) with $l_{A}l_{B}l_{C}l_{D}$ one has

$$0=-\tfrac{1}{2}(k^{A}l_{A})^{5}k_{(P}l_{Q)}+(R/12)(k^{A}l_{A})^{3}l_{P}l_{Q}\,. \tag{2.10}$$

Contracting (2.10) with $k^P l^Q$ we obtain $\frac{1}{4}(k^A l_A)^7 = 0$, which contradicts the assumption $k^A l_A \neq 0$.

We thus conclude as follows: type III is "forbidden"; for types I, II, D, and N, if they are admissible, one has

$$\omega_{PQ} = 0. \tag{2.11}$$

With (2.11), formula (2.3) reduces to

$$C^{S(ABC}C^{D)}{}_{SPQ} + (R/12)C^{AB(C}{}_{(P}\delta^{D)}{}_{Q)} = 0. \tag{2.12}$$

Setting $D = P$, and then contracting with ϵ^{QD} and lowering the indices C and D, one obtains

$$C^{AB}{}_{PS}C^{PS}{}_{CD} + (R/12)C^{AB}{}_{CD} - \tfrac{1}{3}\overset{2}{C}\delta^{A}{}_{(C}\delta^{B}{}_{D)} = 0. \tag{2.13}$$

But we also have the Hamilton–Cayley equation[7,8]

$$C^{AB}{}_{PQ}C^{PQ}{}_{RS}C^{RS}{}_{CD} - \tfrac{1}{2}\overset{2}{C}C^{AB}{}_{CD} - \tfrac{1}{3}\overset{3}{C}\delta^{A}{}_{(C}\delta^{B}{}_{D)} = 0. \tag{2.14}$$

From (2.13) and (2.14) it follows that

$$\overset{2}{C} = 6(R/12)^2, \quad \overset{3}{C} = -6(R/12)^3. \tag{2.15}$$

Hence the invariant $\Delta := \frac{1}{2}(\overset{2}{C})^3 - 3(\overset{3}{C})^2$ vanishes. From this fact and from the fact that the minimal polynomial for $C^{AB}{}_{CD}$ is of order 2 [see (2.13)] it follows that the undotted Weyl spinor must be either of type D or N. In the case of type D, $\overset{2}{C} \neq 0$ and consequently by (2.15), $R \neq 0$.

The exterior covariant differentiation of (2.13), with the use of (1.1a) and (2.13), gives

$$dR - Rr = 0, \tag{2.16}$$

so that, in the case of type D,

$$r = d \ln R. \tag{2.17}$$

Similar considerations concerning the dotted Weyl spinor $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ satisfying condition (1.1b) lead to the final conclusion: If a four-dimensional Riemannian manifold (M,g) is two-sided conformally recurrent, then it must be one of the following Petrov–Penrose types:

CR,HR,UR,ER: $D \otimes D$, $R \neq 0$, $r = d \ln R = \dot{r}$; (2.18a)

CR,HR,UR: $N \otimes N$, $R = 0$; (2.18b)

CR,UR,ER: $D \otimes [\,-\,]$, $R \neq 0$, $r = d \ln R$; (2.18c)

CR,UR: $N \otimes [\,-\,]$, $R = 0$. (2.18d)

Cases (2.18a) and (2.18b) have been analyzed in detail for HR in Ref. 1 with the use of the bivector formalism.

The results can be easily generalized on all possible four-dimensional Riemannian manifolds and we consider this problem in the next section using the spinorial formalism. Then we find all $N \otimes [\,-\,]$ two-sided conformally recurrent spaces.

Up to now we have not succeeded in integrating the type $D \otimes [\,-\,]$. It seems to be a rather hard problem. The work on it is underway.

III. THE INTEGRATION OF TYPES $D \otimes D$ AND $N \otimes N$

We intend to integrate types $D \otimes D$ and $N \otimes N$ for the case of CR and then find the corresponding RR cases taking suitable real slices.

A. The type $D \otimes D$

From (2.12) and its dotted version it follows that

$$C_{ABCD} = \tfrac{1}{8}Rf_{(AB}f_{CD)}, \quad C_{\dot{A}\dot{B}\dot{C}\dot{D}} = \tfrac{1}{8}Rf_{(\dot{A}\dot{B}}f_{\dot{C}\dot{D})}, \tag{3.1}$$

with

$$f_{AB} = f_{(AB)}, \quad f_{\dot{A}\dot{B}} = f_{(\dot{A}\dot{B})}, \quad \tfrac{1}{2}f_{AB}f^{AB} = -1 = \tfrac{1}{2}f_{\dot{A}\dot{B}}f^{\dot{A}\dot{B}}. \tag{3.2}$$

One then easily finds that, as a consequence of (1.1a), (1.1b), and (2.18a),

$$Df_{AB} = 0, \quad Df_{\dot{A}\dot{B}} = 0. \tag{3.3}$$

Now as from (2.18a) one has $R \neq 0, dr = 0 = d\dot{r}$, and formulas (2.2) and (2.6) with their dotted versions give

$$f^{S(A}f^{BC}C^{D)}{}_{SPQ} = 0, \quad C_{PQ}{}^{\dot{S}(\dot{A}}f^{\dot{B}\dot{C}}f^{\dot{D})}{}_{\dot{S}} = 0. \tag{3.4}$$

Simple analysis of (3.4) leads to the formula

$$C_{\dot{A}\dot{B}\dot{C}\dot{D}} = \tfrac{1}{8}Pf_{AB}f_{\dot{C}\dot{D}}, \tag{3.5}$$

where P is some scalar.

Using the freedom of the $\mathrm{SL}(2;C) \times \dot{\mathrm{S}}\mathrm{L}(2;C)$ gauges we can always choose the spinorial frame so that

$$f_{11} = 0 = f_{22}; \quad f_{12} = \epsilon, \quad \epsilon = \pm 1; \quad f_{\dot{1}\dot{1}} = 0 = f_{\dot{2}\dot{2}}; \quad f_{\dot{1}\dot{2}} = \dot{\epsilon}, \quad \dot{\epsilon} = \pm 1. \tag{3.6}$$

From (3.3) with (3.6) one easily infers that in the present gauge

$$\Gamma_{11} = 0 = \Gamma_{22}, \quad \Gamma_{12} \neq 0, \quad \Gamma_{\dot{1}\dot{1}} = 0 = \Gamma_{\dot{2}\dot{2}}, \quad \Gamma_{\dot{1}\dot{2}} \neq 0. \tag{3.7}$$

(Γ_{12} and $\Gamma_{\dot{1}\dot{2}}$ must be nontrivial because R_{AB} and $R_{\dot{A}\dot{B}}$ are nontrivial.)

Assuming (3.7), the first structure equations (1.6) reduce to

$$\begin{aligned} &dg^{1\dot{1}} + (-\Gamma_{12} - \Gamma_{\dot{1}\dot{2}}) \wedge g^{1\dot{1}} = 0, \\ &dg^{1\dot{2}} + (-\Gamma_{12} + \Gamma_{\dot{1}\dot{2}}) \wedge g^{1\dot{2}} = 0, \\ &dg^{2\dot{2}} + (\Gamma_{12} + \Gamma_{\dot{1}\dot{2}}) \wedge g^{2\dot{2}} = 0, \\ &dg^{2\dot{1}} + (\Gamma_{12} - \Gamma_{\dot{1}\dot{2}}) \wedge g^{2\dot{1}} = 0. \end{aligned} \tag{3.8}$$

The second structure equations (1.8) now reduce to the only effective conditions

$$\begin{aligned} d\Gamma_{12} &= -(R/8)S^{12} + \epsilon\dot{\epsilon}(P/8)S^{\dot{1}\dot{2}}, \\ d\Gamma_{\dot{1}\dot{2}} &= -(R/8)S^{\dot{1}\dot{2}} + \epsilon\dot{\epsilon}(P/8)S^{12}. \end{aligned} \tag{3.9}$$

The structure (3.8) and (3.9) is now easily integrable.

As $g^{A\dot{B}} \wedge dg^{A\dot{B}} = 0$ (no summation over A and B), there exist local coordinates $(\eta,\tilde{\eta},\xi,\tilde{\xi})$ and the functions A, $\tilde{A}$, B, and $\tilde{B}$ such that

$$\begin{aligned} g^{1\dot{1}} &= \sqrt{2}A\,d\eta, \quad g^{1\dot{2}} = \sqrt{2}B\,d\xi, \\ g^{2\dot{2}} &= -\sqrt{2}\tilde{A}\,d\tilde{\eta}, \quad g^{2\dot{1}} = \sqrt{2}\tilde{B}\,d\tilde{\xi}. \end{aligned} \tag{3.10}$$

This fed back into (3.8) gives as a consequence that there exist functions C, $\tilde{C}$, D, and $\tilde{D}$ such that

$$\begin{aligned} \Gamma_{12} + \Gamma_{\dot{1}\dot{2}} &= d \ln A + 2C\,d\eta, \\ \Gamma_{12} + \Gamma_{\dot{1}\dot{2}} &= -d \ln \tilde{A} - 2\tilde{C}\,d\tilde{\eta}, \\ \Gamma_{12} - \Gamma_{\dot{1}\dot{2}} &= d \ln B + 2D\,d\xi, \\ \Gamma_{12} - \Gamma_{\dot{1}\dot{2}} &= -d \ln \tilde{B} - 2\tilde{D}\,d\tilde{\xi}. \end{aligned} \tag{3.11}$$

The consistence of these relations requires

$$A\tilde{A}=\Psi^{-2},\quad C=\partial_\eta \ln\Psi,\quad \tilde{C}=\partial_{\tilde{\eta}}\ln\Psi,$$
$$B\tilde{B}=\Phi^{-2},\quad D=\partial_\xi \ln\Phi,\quad \tilde{D}=\partial_{\tilde{\xi}}\ln\Phi, \tag{3.12}$$

where Φ and Ψ are the functions of two variables only:

$$\Phi=\Phi(\xi,\tilde{\xi}),\quad \Psi=\Psi(\eta,\tilde{\eta}). \tag{3.13}$$

We thus infer at this point that the metric has the general form of [see (1.2)]

$$g=2\Phi^{-2}(\xi,\tilde{\xi})d\xi \underset{s}{\otimes} d\tilde{\xi}+2\Psi^{-2}(\eta,\tilde{\eta})d\eta \underset{s}{\otimes} d\tilde{\eta}. \tag{3.14}$$

Without any loss of generality we can set $AB=\Psi^{-1}\Phi^{-1}=\tilde{A}\tilde{B}$ in (3.10). Then from (3.11) and (3.9) one obtains

$$R=-4[\Psi^2(\ln\Psi)_{,\eta,\tilde{\eta}}+\Phi^2(\ln\Phi)_{,\xi\tilde{\xi}}]\neq 0,$$
$$P=4\epsilon\dot{\epsilon}[\Psi^2(\ln\Psi)_{,\eta\tilde{\eta}}-\Phi^2(\ln\Phi)_{,\xi\tilde{\xi}}]. \tag{3.15}$$

We now can give a plausible geometric interpretation of the results obtained.

All two-sided conformally recurrent four-dimensional Riemannian manifolds of the type $D\otimes D$ are just the Cartesian product of two two-dimensional Riemannian manifolds

$$g=g_1+g_2,$$

$$g_1:=2\Phi^{-2}(\xi,\tilde{\xi})d\xi \underset{s}{\otimes} d\tilde{\xi},\quad g_2:=2\Psi^{-2}(\eta,\tilde{\eta})d\eta \underset{s}{\otimes} d\tilde{\eta}, \tag{3.16}$$

with the curvature scalars

$$R_1=-4\Phi^2(\ln\Phi)_{,\xi\tilde{\xi}},\quad R_2=-4\Psi^2(\ln\Psi)_{,\eta\tilde{\eta}}, \tag{3.17}$$

respectively. Therefore (3.15) just says

$$R=R_1+R_2\neq 0,\quad P=\epsilon\dot{\epsilon}(R_1-R_2). \tag{3.18}$$

Now according to Ruse[9] we define a recurrent n-dimensional Riemannian manifold to be one that is nonflat and for which there exists a vector field r_α, such that the covariant differential of the curvature tensor $R_{\alpha\beta\gamma\delta}$ satisfies the condition

$$R_{\alpha\beta\gamma\delta;\varepsilon}=r_\varepsilon R_{\alpha\beta\gamma\delta},\quad \alpha,\beta,\gamma,\delta,\varepsilon=1,\ldots,n. \tag{3.19}$$

If, in particular, $r_\varepsilon=0$, the Riemannian manifold is symmetric. It can be shown that (3.19) is equivalent to the following relations:

$$C_{\alpha\beta\gamma\delta;\varepsilon}=r_\varepsilon C_{\alpha\beta\gamma\delta},\quad C_{\alpha\beta;\varepsilon}=r_\varepsilon C_{\alpha\beta},\quad R_{;\varepsilon}=r_\varepsilon R, \tag{3.20}$$

where $C_{\alpha\beta\gamma\delta}$ and $C_{\alpha\beta}$ are the Weyl tensor and the traceless Ricci tensor, respectively. Therefore, every conformally nonflat recurrent Riemannian manifold is also conformally recurrent,[10] i.e.,

$$C_{\alpha\beta\gamma\delta;\varepsilon}=r_\varepsilon C_{\alpha\beta\gamma\delta}, \tag{3.21}$$

and every symmetric Riemannian manifold is conformally symmetric,[11] i.e., $C_{\alpha\beta\gamma\delta;\varepsilon}=0$.

In the case of $n=4$ we can state that every recurrent space with $C_{\alpha\beta\gamma\delta}\neq 0$ is two-sided conformally recurrent.

For the type considered, $D\otimes D$, one easily finds that the following statements are equivalent:

$$R=\text{const}\Leftrightarrow R_1=\text{const and } R_2=\text{const}$$
$$\Leftrightarrow (M,g)\text{ is conformally symmetric}$$
$$\Leftrightarrow (M,g)\text{ is symmetric.}$$

Now we would like to state that the derived metric (3.14) covers all two-sided conformally recurrent four-dimensional Riemannian manifolds $D\otimes D$ if some suitable restrictions are imposed on the local coordinates η, $\tilde{\eta}$, ξ, and $\tilde{\xi}$ and the functions $\Psi(\eta,\tilde{\eta})$ and $\Phi(\xi,\tilde{\xi})$. Namely, one has

CR: $\eta,\tilde{\eta},\xi,\tilde{\xi}$ complex, Ψ,Φ holomorphic;

HR: $\eta,\tilde{\eta}$ real, ξ complex, $\tilde{\xi}=\bar{\xi}$,
Ψ,Φ real $(+\ +\ +\ -)$ or pure imaginary
$(-\ -\ -\ +)$;

UR: $\eta,\tilde{\eta},\xi,\tilde{\xi}$ real, Ψ,Φ real;

ER: η,ξ complex, $\tilde{\eta}=\bar{\eta}$, $\tilde{\xi}=\bar{\xi}$,
Ψ,Φ real $(+\ +\ +\ +)$ or pure imaginary
$(-\ -\ -\ -)$.

It is worthwhile to notice that according to (3.5) the spinor $C_{ABC\dot{D}}$, which enters as the source into the Einstein equations, has an algebraic structure compatible with the structure of the source of the nonlinear electrodynamics of Born–Infeld type.[12] The consequences of this observation will be examined elsewhere.

In the special case $R=\text{const}\Rightarrow P=\text{const}$, the metric (3.14) is commonly interpreted as the Bertotti–Robinson solution[13,14] with the cosmological constant $\lambda\neq 0$.

The same case can be, however, reinterpreted as some solution to the Einstein and Born–Infeld equations with the electromagnetic field covariantly constant.

B. The type $N\otimes N$

In this case we have

$$C_{ABCD}=\rho k_A k_B k_C k_D,\quad C_{\dot{A}\dot{B}\dot{C}\dot{D}}=\dot{\rho}k_{\dot{A}}k_{\dot{B}}k_{\dot{C}}k_{\dot{D}},$$
$$\rho\neq 0,\quad \dot{\rho}\neq 0,\quad \sum_A |k_A|\neq 0,\quad \sum_{\dot{A}}|k_{\dot{A}}|\neq 0. \tag{3.22}$$

From (1.1a), (1.1b), and (3.22) one finds

$$(D+d\ln\rho-\tfrac{1}{4}r)k_A=0,\quad (D+d\ln\dot{\rho}-\tfrac{1}{4}\dot{r})k_{\dot{A}}=0. \tag{3.23}$$

Consequently, the null vector field $k_\mu=g^{A\dot{B}}{}_\mu k_A k_{\dot{B}}$ appears to be recurrent, i.e.,

$$Dk_\mu=\sigma k_\mu,\quad \mu=1,\ldots,4, \tag{3.24}$$

with $\sigma=\frac{1}{4}(r+\dot{r})-d(\ln\rho+\ln\dot{\rho})$. The spin tensor $g^{A\dot{B}}{}_\mu$, $\mu=1,\ldots,4$, is defined by the formula, $g^{A\dot{B}}=g^{A\dot{B}}{}_\mu\, dx^\mu$, where $\{x^\mu\}$, $\mu=1,\ldots,4$, is a local coordinate system. Conversely, if a null vector field k_μ, $\mu=1,\ldots,4$, is recurrent, i.e., it satisfies the condition (3.24) for some one-form σ, then the spinor fields k_A and $k_{\dot{A}}$ defined by the relation $k_\mu=g^{A\dot{B}}{}_\mu k_A k_{\dot{B}}$ are recurrent:

$$Dk_A=sk_A,\quad Dk_{\dot{A}}=\dot{s}k_{\dot{A}}, \tag{3.25}$$

where s and $\dot{s}$ are one-forms such that $s+\dot{s}=\sigma$. Therefore, using arguments similar to those in Sec. IV [see Eq. (4.28)] one arrives at the following statement: If (M,g) is two-sided

conformally recurrent Riemannian manifold of the type $N\otimes N$ then the quadruple Debever–Penrose vector field[8] is recurrent; conversely, if a null vector field k_μ on Riemannian manifold (M,g) of the type $N\otimes N$ is recurrent then (M,g) is two-sided conformally recurrent and k_μ is the quadruple Debever–Penrose vector field. In the case of HR this statement is given in Ref. 1. [Remark: One can easily prove that for UR or HR, if (M,g) admits a recurrent null vector field then it also admits a real recurrent null vector field.] The HR metrics admitting the existence of a recurrent null vector field were examined by Walker[15] and by Debever and Cahen.[16] The general $N\otimes N$ metric of this type ($\Leftrightarrow$ the general $N\otimes N$ metric of a two-sided conformally recurrent Riemannian manifold) for the case of HR was found by McLenaghan and Leroy.[1] Analogous considerations lead to the conclusion that, in the case of CR, the general metric of Riemannian manifold admitting a recurrent null vector field is of the following Debever–Spelkens[1,17] form:

$$g = 2\,dv \underset{s}{\otimes} (du + C\,dv) + 2q^{-2}(dz + p\,dv) \underset{s}{\otimes} (d\tilde{z} + \tilde{p}\,dv), \tag{3.26}$$

where v, u, z, and $\tilde{z}$ are local coordinates, $C = C(v,u,z,\tilde{z})$, $q = q(v,z,\tilde{z})$, $p = p(v,z,\tilde{z})$, and $\tilde{p} = \tilde{p}(v,z,\tilde{z})$ are arbitrary holomorphic functions. The one-forms $g^{A\dot{B}}$ defining the metric (3.26) according to (1.2) can be taken in the form

$$g^{1\dot{1}} = \sqrt{2}(du + C\,dv), \quad g^{1\dot{2}} = \sqrt{2}q^{-1}(d\tilde{z} + \tilde{p}\,dv),$$
$$g^{2\dot{2}} = -\sqrt{2}\,dv, \quad g^{2\dot{1}} = \sqrt{2}q^{-1}(dz + p\,dv). \tag{3.27}$$

Then from the first and second Cartan structure equations (1.6)–(1.10) with $g^{A\dot{B}}$ given by (3.27) we obtain

$$\Gamma_{11} = 0,$$
$$\Gamma_{12} = -\tfrac{1}{2}(C_{,u} + F - \widetilde{F})dv - \tfrac{1}{2}(\ln q)_{,z}\,dz + \tfrac{1}{2}(\ln q)_{,\tilde{z}}\,d\tilde{z},$$
$$\Gamma_{22} = q^{-1}[\tilde{p}(F + \widetilde{F}) - p\tilde{p}_{,z} - q^2 C_{,z}]dv - q^{-1}\tilde{p}_{,z}\,dz + q^{-1}(F + \widetilde{F})d\tilde{z}; \tag{3.28}$$
$$C_{1111} = 0, \quad C_{1112} = 0, \tag{3.29a}$$
$$C_{1122} = -\tfrac{1}{3}[\tfrac{1}{2}C_{,uu} + q^2(\ln q)_{,z\tilde{z}}], \tag{3.29b}$$
$$C_{1222} = q\{-\tfrac{1}{2}[C_{,u} + F - \widetilde{F} - (\ln q)_{,v}]_{,z} - \tilde{p}(\ln q)_{,z\tilde{z}}\}, \tag{3.29c}$$
$$C_{2222} = \{\tilde{p}[2\widetilde{F} + (\ln q)_{,v}] - p\tilde{p}_{,z} - q^2 C_{,z} + \tilde{p}_{,v}\}_{,z} + 2\tilde{p}_{,z}[\tfrac{1}{2}C_{,u} + F - \widetilde{F} - (\ln q)_{,v}] - 2\tilde{p}^2(\ln q)_{,z\tilde{z}}; \tag{3.29d}$$

$$R = 12C_{1122},$$
$$C_{11\dot{A}\dot{B}} = 0,$$
$$C_{12\dot{1}\dot{1}} = 0,$$
$$C_{12\dot{1}\dot{2}} = \tfrac{1}{2}[\tfrac{1}{2}C_{,uu} - q^2(\ln q)_{,z\tilde{z}}],$$
$$C_{12\dot{2}\dot{2}} = q\{\tfrac{1}{2}[C_{,u} + F - \widetilde{F} + (\ln q)_{,v}]_{,\tilde{z}} - p(\ln q)_{,z\tilde{z}}\},$$
$$C_{22\dot{1}\dot{1}} = 0,$$
$$C_{22\dot{1}\dot{2}} = q\{\tfrac{1}{2}[C_{,u} + \widetilde{F} - F + (\ln q)_{,v}]_{,z} - \tilde{p}(\ln q)_{,z\tilde{z}}\},$$
$$C_{22\dot{2}\dot{2}} = p\{[\widetilde{F} - F + (\ln q)_{,v}]_{,z} - 2\tilde{p}(\ln q)_{,z\tilde{z}}\} + [\tilde{p}(F + \widetilde{F}) - p\tilde{p}_{,z} - q^2 C_{,z}]_{,\tilde{z}} - 2q\{q^{-1}[\tilde{p}(F + \widetilde{F}) - p\tilde{p}_{,z} - q^2 C_{,z}]\}_{,\tilde{z}} + q[q^{-1}(F + \widetilde{F})]_{,v} + (F + \widetilde{F})(C_{,u} + F - \widetilde{F}); \tag{3.30}$$

where

$$F := p(\ln q)_{,z} - \tfrac{1}{2}(\ln q)_{,v} - \tfrac{1}{2}p_{,z},$$
$$\widetilde{F} := \tilde{p}(\ln q)_{,\tilde{z}} - \tfrac{1}{2}(\ln q)_{,v} - \tfrac{1}{2}\tilde{p}_{,\tilde{z}}. \tag{3.31}$$

Now $\Gamma_{A\dot{B}}$ and $C_{A\dot{B}C\dot{D}}$ can be found from (3.28) and (3.29), respectively, by the following interchanges:

$$z \leftrightarrow \tilde{z}, \quad p \leftrightarrow \tilde{p}, \quad F \leftrightarrow \widetilde{F}.$$

Formulas (3.26)–(3.31) define the general four-dimensional complex Riemannian manifold admitting a recurrent null vector field.

We intend now to specialize them for the type $N\otimes N$. In this case one has

$$C_{1122} = 0 = C_{1\dot{1}\dot{2}\dot{2}}, \tag{3.32}$$
$$C_{1222} = 0 = C_{1\dot{2}\dot{2}\dot{2}}, \tag{3.33}$$
$$C_{2222} \neq 0 \neq C_{2\dot{2}\dot{2}\dot{2}}. \tag{3.34}$$

From (3.29c) and its dotted version, with (3.33) assumed, $C_{,uuz} = 0 = C_{,uu\tilde{z}}$. Hence

$$C = -\varepsilon k^2(v)u^2 + l(v,z,\tilde{z})u + m(v,z,\tilde{z}), \tag{3.35}$$

where $\varepsilon = -1, 0, +1$ and $k(v)$, $l(v,z,\tilde{z})$, and $m(v,z,\tilde{z})$ are as yet arbitrary holomorphic functions.

Then from (3.29b) with (3.32) and (3.35) one has

$$qq_{,z\tilde{z}} - q_{,z}q_{,\tilde{z}} = \varepsilon k^2(v). \tag{3.36}$$

This is the Liouville equation for q with the general solution

$$q = k(v)(1 + \varepsilon\varphi\tilde{\varphi})/(\varphi_{,z}\tilde{\varphi}_{,\tilde{z}})^{1/2}, \tag{3.37}$$

where $\varphi = \varphi(z)$ and $\tilde{\varphi} = \tilde{\varphi}(\tilde{z})$ are arbitrary holomorphic functions. Without any loss of generality we can set $\varphi = z$ and $\tilde{\varphi} = \tilde{z}$. Thus (3.37) is of the form

$$q = k(v)(1 + \varepsilon z\tilde{z}). \tag{3.38}$$

Then Eqs. (3.33) and (3.29c) and its dotted version, with C and q given by (3.35) and (3.38), respectively, lead to the following conclusion (compare Ref. 1): There exists a function $\alpha(v,z,\tilde{z})$ such that performing the coordinate transformation preserving the form of the metric (3.26),

$$z \mapsto z, \quad \tilde{z} \mapsto \tilde{z}, \quad v \mapsto v, \quad u \mapsto u + \alpha(v,z,\tilde{z}), \tag{3.39}$$

one obtains that

$$p = p(v,z), \quad \tilde{p} = \tilde{p}(v,\tilde{z}), \tag{3.40}$$
$$l = \tfrac{1}{2}(p_{,z} + \tilde{p}_{,\tilde{z}}) - \varepsilon(1 + \varepsilon z\tilde{z})^{-1}(\tilde{z}p + z\tilde{p}). \tag{3.41}$$

Finally, the general metric of a two-sided conformally recurrent four-dimensional complex Riemannian manifold of type $N\otimes N$ is of the form

$$g = 2\,dv \underset{s}{\otimes} [du + (-\varepsilon k^2 u^2 + lu + m)] + 2k^{-2}(1 + \varepsilon z\tilde{z})^{-2}(dz + p\,dv) \underset{s}{\otimes} (d\tilde{z} + \tilde{p}\,dv), \tag{3.42}$$

where $\varepsilon = -1,0,+1$; $k = k(v)$, $m = m(v,z,\tilde{z})$, $p = p(v,z)$ and $\tilde{p} = \tilde{p}(v,\tilde{z})$ are arbitrary holomorphic functions; and $l = l(v,z,\tilde{z})$ is defined by (3.41).

Then the nonvanishing components of the curvature are

$$C_{2222} = -k^2(1+\varepsilon z\tilde{z})[\tfrac{1}{2}u(1+\varepsilon z\tilde{z})p_{,zzz} + 2\varepsilon\tilde{z}m_{,z} + (1+\varepsilon z\tilde{z})m_{,zz}]; \tag{3.43}$$

$$C_{12\dot{1}\dot{2}} = -\varepsilon k^2,$$

$$C_{12\dot{2}\dot{2}} = k(1+\varepsilon z\tilde{z})l_{,\tilde{z}},$$

$$C_{22\dot{1}\dot{2}} = k(1+\varepsilon z\tilde{z})l_{,z},$$

$$C_{22\dot{2}\dot{2}} = pl_{,z} + \tilde{p}l_{,\tilde{z}} - l_{,v} + k^2(1+\varepsilon z\tilde{z})^2 m_{,z\tilde{z}} - (\ln k)_{,vv} + 2\varepsilon k k_{,v}u + (l+\ln k)_{,v}(\ln k)_{,v}. \tag{3.44}$$

[$C_{\dot{2}\dot{2}\dot{2}\dot{2}}$ can be found from (3.43) by the replacements $z \to \tilde{z}$, $\tilde{z} \to z$, and $p \to \tilde{p}$.]

Assuming that

$$C_{12\dot{1}\dot{2}} = C_{12\dot{2}\dot{2}} = C_{22\dot{1}\dot{2}} = 0, \tag{3.45}$$

one can bring the metric (3.42) to the form

$$g = 2\,dv \underset{s}{\otimes} (du + m\,dv) + 2\,dz \underset{s}{\otimes} d\tilde{z}, \tag{3.46}$$

where $m = m(v,z,\tilde{z})$ is an arbitrary holomorphic function. This is the Robinson metric for the space of plane-fronted waves with parallel rays[1,2,4] (*pp* waves).

Then (3.46) is the vacuum metric iff $C_{2\dot{2}\dot{2}\dot{2}} = 0$, i.e., $m = H(v,z) + \tilde{H}(v,\tilde{z})$, where $H(v,z)$ and $\tilde{H}(v,\tilde{z})$ are arbitrary holomorphic functions.

One can easily find (compare Ref. 1) the following special cases of the metric (3.42): (i) the general conformally recurrent space of type $N \otimes N$,

$$g = 2\,dv \underset{s}{\otimes} [du + (f+n)dv] + 2\,dz \underset{s}{\otimes} d\tilde{z}, \tag{3.47}$$

where $f = f(v,z+\tilde{z})$ and $n = n(v,1/i(z-\tilde{z}))$ are arbitrary holomorphic functions; (ii) the general conformally symmetric space of type $N \otimes N$,

$$g = 2\,dv \underset{s}{\otimes} [du + (z^2 + \tilde{z}^2 + ez\tilde{z})dv] + 2\,dz \underset{s}{\otimes} d\tilde{z}, \tag{3.48}$$

where $e = e(v)$ is an arbitrary holomorphic function; and (iii) the general symmetric space of the type $N \otimes N$, where g is of the form (3.48) with $e = \text{const}$.

Finally, one can verify that all formulas concerning two-sided conformally recurrent complex Riemannian manifolds hold true in the cases HR and UR if the following restrictions are imposed:

HR: u,v real, z complex, $\tilde{z} = \overline{z}$,
the functions C,l,m,f,n,e are real,
the functions $\tilde{p} = \overline{p}$, $\tilde{H} = \overline{H}$,
the functions q and k are real (+ + + −)
or pure imaginary (− − − +);

UR: all coordinates and functions are real.

IV. THE INTEGRATION OF THE TYPE $N \otimes [-]$

Two-sided conformally recurrent Riemannian manifolds of the type $N \otimes [-]$ are allowed in CR and UR only [(2.18d)]. As before we consider the case of CR and then we find the metric for UR as a suitable real slice.

Now one has

$$C_{ABCD} = \rho k_A k_B k_C k_D, \quad \rho \neq 0, \quad |k_1| + |k_2| \neq 0,$$
$$C_{\dot{A}\dot{B}\dot{C}\dot{D}} = 0, \tag{4.1}$$

with ρ to be chosen as convenient ($\rho \neq 0$).

From (1.1a) and (4.1) we obtain

$$(D + d\ln\rho - \tfrac{1}{4}r)k_A = 0. \tag{4.2}$$

Therefore, if one defines a recurrent undotted spinor field to be a spinor field k'_A such that

$$\sum_A |k'_A| \neq 0$$

and

$$(D - s')k'_A = 0 \tag{4.3}$$

for some one-form s', then (4.2) says that our space admits k_A as a recurrent undotted spinor field. This space generalizes the one with a recurrent null vector field (see Sec. III B). Therefore it is of interest to explore the consequences of the existence of a recurrent undotted spinor field without any further assumptions, and only later on to specialize the results obtained.

We intend to proceed this way.

Define $g^{\dot{B}} := k'_A g^{A\dot{B}}$; with (4.3) assumed, because of $Dg^{A\dot{B}} = 0$, we have $Dg^{\dot{B}} = s' \wedge g^{\dot{B}}$, or explicitly

$$dg^{\dot{B}} + \Gamma^{\dot{B}}{}_{\dot{C}} \wedge g^{\dot{C}} = s' \wedge g^{\dot{B}}. \tag{4.4}$$

Using the Frobenius theorem one infers from (4.4) that there exist functions $q^{\dot{B}}$ and $L^{\dot{B}}{}_{\dot{C}}$ such that

$$g^{\dot{B}} = L^{\dot{B}}{}_{\dot{C}}\,dq^{\dot{C}}, \quad dq^{\dot{1}} \wedge dq^{\dot{2}} \neq 0, \quad \det\|L^{\dot{B}}{}_{\dot{C}}\| \neq 0. \tag{4.5}$$

Writing $L^{\dot{B}}{}_{\dot{C}} = Ll^{\dot{B}}{}_{\dot{C}}$, $\det\|l^{\dot{B}}{}_{\dot{C}}\| = 1$, $L \neq 0$, it is clear that we can adopt the $SL(2;C)$ gauge in a specific manner so that

$$k'_A g^{A\dot{B}} = g^{\dot{B}} = L\,dq^{\dot{B}}. \tag{4.6}$$

Using then the multiplicative ambiguity in the definition of the recurrent undotted spinor field (4.3) we set $k_A := L^{-1}k'_A$, reducing (4.6) to

$$k_A g^{A\dot{B}} = dq^{\dot{B}}. \tag{4.7}$$

Then from (4.3) one has

$$(D - s)k_A = 0, \quad s := s' - d\ln L. \tag{4.8}$$

In the next step we use the freedom of the $SL(2;C)$ gauge, choosing it so that

$$k_A = (1/\sqrt{2})\delta^2{}_A. \tag{4.9}$$

From (4.7) and (4.9) one finds

$$g^{2\dot{B}} = \sqrt{2}\,dq^{\dot{B}}. \tag{4.10}$$

Consequently,

$$S^{22} = g^{2\dot{1}} \wedge g^{2\dot{2}} = 2\,dq^{\dot{1}} \wedge dq^{\dot{2}}, \tag{4.11}$$

and thus we have a null string defined by S^{22} that is a simple, self-dual, and closed two-form.[5,18] Moreover, this string is

nonexpanding. Indeed, in the gauge (4.9), condition (4.8) takes the form of

$$-\Gamma^2{}_A = s\delta^2{}_A \Rightarrow \Gamma_{12} = -s, \quad \Gamma_{11} = 0. \tag{4.12}$$

The condition $\Gamma_{11} = 0$ assures us that the null string defined by S^{22} is nonexpanding.

Consequently, with (4.10), the $Dg^{2\dot{B}} = 0$ equations amount to

$$-s \wedge dq^{\dot{B}} + \Gamma^{\dot{B}}{}_{\dot{C}} \wedge dq^{\dot{C}} = 0; \tag{4.13}$$

and the $Dg^{1\dot{B}} = 0$ equations reduce to

$$dg^{1\dot{B}} + s \wedge g^{1\dot{B}} + \Gamma^{\dot{B}}{}_{\dot{C}} \wedge g^{1\dot{C}} - \sqrt{2}\Gamma_{22} \wedge dq^{\dot{B}} = 0. \tag{4.14}$$

Then, (4.13) wedged with $dq^{\dot{A}}$ amounts to

$$(s\epsilon^{\dot{A}\dot{B}} + \Gamma^{\dot{A}\dot{B}}) \wedge dq^{\dot{1}} \wedge dq^{\dot{2}} = 0$$

$$\Rightarrow s \wedge dq^{\dot{1}} \wedge dq^{\dot{2}} = 0 \text{ and } \Gamma_{\dot{A}\dot{B}} \wedge dq^{\dot{1}} \wedge dq^{\dot{2}} = 0. \tag{4.15}$$

This being so, we infer from (4.14) that

$$dq^{\dot{1}} \wedge dq^{\dot{2}} \wedge dg^{1\dot{B}} = 0, \tag{4.16}$$

and therefore there exist functions $p^{\dot{B}}$ and $Q'^{\dot{B}}{}_{\dot{C}}$ such that

$$g^{1\dot{B}} = \sqrt{2}(dp^{\dot{B}} + Q'^{\dot{B}}{}_{\dot{C}} dq^{\dot{C}}). \tag{4.17}$$

As

$$0 \neq g^{1\dot{1}} \wedge g^{1\dot{2}} \wedge g^{2\dot{1}} \wedge g^{2\dot{2}} = 4\, dq^{\dot{1}} \wedge dq^{\dot{2}} \wedge dp^{\dot{1}} \wedge dp^{\dot{2}}, \tag{4.18}$$

$\{q^{\dot{A}}, p^{\dot{B}}\}$ constitutes a chart for the studied structure. Now decomposing $Q'_{\dot{A}\dot{B}}$ according to

$$Q'_{\dot{A}\dot{B}} = Q_{\dot{A}\dot{B}} + \epsilon_{\dot{A}\dot{B}} Q, \quad Q_{\dot{A}\dot{B}} = Q_{(\dot{A}\dot{B})}, \tag{4.19}$$

one easily finds that

$$g = -\tfrac{1}{2} g_{AB} \underset{s}{\otimes} g^{A\dot{B}} = 2\, dq^{\dot{A}} \underset{s}{\otimes} (dp_{\dot{A}} + Q_{\dot{A}\dot{B}}\, dq^{\dot{B}}). \tag{4.20}$$

Therefore, without any loss of generality we can put $Q = 0$ in (4.19); thus $Q'_{\dot{A}\dot{B}} = Q_{\dot{A}\dot{B}} = Q_{(\dot{A}\dot{B})}$.

Gathering this all together, we have arrived at the conclusion that if a complex four-dimensional Riemannian manifold (M,g) admits a recurrent undotted spinor field k'_A then one can choose a function L and the $\mathrm{SL}(2;C) \times \dot{\mathrm{SL}}(2;C)$ gauge in such a manner that $L^{-1}k'_A = :k_A = (1/\sqrt{2})\delta^2{}_A$ and the tetrad is of the form

$$g^{2\dot{B}} = \sqrt{2}\, dq^{\dot{B}}, \quad g^{1\dot{B}} = \sqrt{2}(dp^{\dot{B}} + Q^{\dot{B}}{}_{\dot{C}}\, dq^{\dot{C}}) \tag{4.21}$$

for some chart $\{q^{\dot{A}}, p^{\dot{B}}\}$, where $Q_{\dot{A}\dot{B}} = Q_{(\dot{A}\dot{B})}$ are functions of the variables $q^{\dot{A}}$ and $p^{\dot{B}}$. Moreover, in this gauge

$$\Gamma_{11} = 0 \quad \text{and} \quad \Gamma_{12} = -s, \tag{4.22}$$

where s is the recurrence one-form defined by (4.8). The first condition of (4.22) means exactly that the null string defined by $S^{22} = g^{2\dot{1}} \wedge g^{2\dot{2}} = 2\, dq^{\dot{1}} \wedge dq^{\dot{2}}$ is nonexpanding. The metric g of the manifold considered is given by (4.20). [Notice that all these facts hold true in the case of UR with the only difference being that one takes the $\mathrm{SL}(2;R) \times \dot{\mathrm{SL}}(2;R)$ gauge and all functions and coordinates are real.]

With the tetrad (4.21) and $\{q^{\dot{A}}, p^{\dot{B}}\}$ being a chart we can directly specialize the results of Ref. 18, writing that article's (2.16)–(2.18) in the case of $\Phi = 1$. Thus we arrive at the following results:

$$\Gamma_{11} = 0,$$

$$-s = \Gamma_{12} = \tfrac{1}{2}\partial_{\dot{A}} Q^{\dot{A}\dot{B}}\, dq_{\dot{B}},$$

$$\Gamma_{22} = \eth_{\dot{A}} Q^{\dot{A}\dot{B}}\, dq_{\dot{B}}; \tag{4.23}$$

$$\Gamma_{\dot{A}\dot{B}} = \partial_{(\dot{A}} Q_{\dot{B})\dot{C}}\, dq^{\dot{C}}; \tag{4.24}$$

$$C_{1111} = 0,$$

$$C_{1112} = 0,$$

$$C_{1122} = -\tfrac{1}{6}\partial_{\dot{A}}\partial_{\dot{B}} Q^{\dot{A}\dot{B}},$$

$$C_{1222} = -\tfrac{1}{2}\partial^{\dot{A}}\eth^{\dot{B}} Q_{\dot{A}\dot{B}} = -\tfrac{1}{2}\eth^{\dot{A}}\partial^{\dot{B}} Q_{\dot{A}\dot{B}},$$

$$C_{2222} = -\eth^{\dot{A}}\eth^{\dot{B}} Q_{\dot{A}\dot{B}}; \tag{4.25}$$

$$C_{\dot{A}\dot{B}\dot{C}\dot{D}} = -\partial_{(\dot{A}}\partial_{\dot{B}} Q_{\dot{C}\dot{D})}; \tag{4.26}$$

$$R = -2\partial_{\dot{A}}\partial_{\dot{B}} Q^{\dot{A}\dot{B}},$$

$$C_{11\dot{A}\dot{B}} = 0,$$

$$C_{12\dot{A}\dot{B}} = -\tfrac{1}{2}\partial_{(\dot{A}}\partial^{\dot{C}} Q_{\dot{B})\dot{C}},$$

$$C_{22\dot{A}\dot{B}} = -\partial_{(\dot{A}}\eth^{\dot{C}} Q_{\dot{B})\dot{C}}; \tag{4.27}$$

where

$$\partial_{\dot{A}} := \frac{\partial}{\partial p^{\dot{A}}}, \quad \eth^{\dot{A}} := \frac{\partial}{\partial q_{\dot{A}}} + Q^{\dot{A}\dot{B}}\partial_{\dot{B}},$$

$$\partial^{\dot{A}} := \epsilon^{\dot{A}\dot{B}}\partial_{\dot{B}} = \frac{\partial}{\partial p_{\dot{A}}}.$$

We now use the results obtained above for the case of two-sided conformally recurrent space of type $N \otimes [-]$. We have found [see (4.1) and (4.2)] that every two-sided conformally recurrent space $N \otimes [-]$ admits a recurrent undotted spinor field. Conversely, if a space of the type $N \otimes [-]$ admits a recurrent undotted spinor field then the space is two-sided conformally recurrent. Indeed, writing the exterior covariant differential DC_{ABCD} in the gauge (4.9), remembering that, by (4.25), $C_{1111} = 0 = C_{1112}$ and therefore also $C_{1122} = 0 = C_{1222}$, and then using (4.12), one obtains

$$DC_{ABCD} = (4s + d \ln C_{2222}) C_{ABCD}. \tag{4.28}$$

Hence the space is two-sided conformally recurrent. Moreover, the recurrent spinor field appears to be a quadruple Penrose spinor field for C_{ABCD}.

Specializing Eqs. (4.25) and (4.26) in the case of $N \otimes [-]$ we obtain a set of differential conditions:

$$(C_{1122} = 0 \Leftrightarrow R = 0) \Leftrightarrow \partial_{\dot{A}}\partial_{\dot{B}} Q^{\dot{A}\dot{B}} = 0, \tag{4.29}$$

$$C_{1122} = 0 \Leftrightarrow \partial^{\dot{A}}\eth^{\dot{B}} Q_{\dot{A}\dot{B}} \equiv \eth^{\dot{A}}\partial^{\dot{B}} Q_{\dot{A}\dot{B}} = 0, \tag{4.30}$$

$$C_{\dot{A}\dot{B}\dot{C}\dot{D}} = 0 \Leftrightarrow \partial_{(\dot{A}}\partial_{\dot{B}} Q_{\dot{C}\dot{D})} = 0, \tag{4.31}$$

$$C_{2222} = -\eth^{\dot{A}}\eth^{\dot{B}} Q_{\dot{A}\dot{B}} \neq 0. \tag{4.32}$$

Similar to in Ref. 18, we infer first from (4.29) the existence of a spinor $A_{\dot{A}}$ such that

$$Q_{\dot{A}\dot{B}} = \partial_{(\dot{A}} A_{\dot{B})}. \tag{4.33}$$

Then, condition (4.31) takes the form

$$\partial_{(\dot{A}}\partial_{\dot{B}}\partial_{\dot{C}} A_{\dot{D})} = 0. \tag{4.34}$$

Acting on (4.34) with $\partial^{\dot{D}}$ we obtain

$$\partial_{\dot{A}}\partial_{\dot{B}}\partial_{\dot{C}}\partial^{\dot{D}} A_{\dot{D}} = 0. \tag{4.35}$$

Acting on (4.34) with $\partial_{\dot{E}}$, employing the identity

$\partial_{\dot E}A_{\dot D} = -\epsilon_{\dot E\dot D}\,\partial_{\dot F}A^{\dot F} + \partial_{\dot D}A_{\dot E}$, and then (4.35), one finds the equation

$$\partial_{\dot A}\partial_{\dot B}\partial_{\dot C}\partial_{\dot D}A_{\dot E} = 0. \tag{4.36}$$

From (4.36) it follows that $A_{\dot A}$ must be a polynomial of order $\leqslant 3$ with respect to $p^{\dot B}$.

Then one can easily verify that the most general $A_{\dot A}$ that satisfies (4.34) has the form

$$A_{\dot A} = \tfrac{1}{2}\epsilon_{\dot A\dot D}a_{\dot B\dot C}p^{\dot B}p^{\dot C}p^{\dot D} + (\tfrac{1}{2}\rho_{\dot A\dot B\dot C} + \epsilon_{\dot A\dot B}b_{\dot C})p^{\dot B}p^{\dot C} + (\sigma_{\dot A\dot B} + c\epsilon_{\dot A\dot B})p^{\dot B} + \tau_{\dot A}, \tag{4.37}$$

where $a_{\dot B\dot C} = a_{(\dot B\dot C)}$, $\rho_{\dot A\dot B\dot C} = \rho_{(\dot A\dot B\dot C)}$, $\sigma_{\dot A\dot B} = \sigma_{(\dot A\dot B)}$, $b_{\dot C}$, c, and $\tau_{\dot A}$ are the functions of variables $q^{\dot A}$.

Substituting $A_{\dot A}$ given by (4.37) into (4.33) we obtain

$$Q_{\dot A\dot B} = a_{\dot C(\dot A}\epsilon_{\dot B)\dot D}p^{\dot C}p^{\dot D} + (\rho_{\dot A\dot B\dot C} + b_{(\dot A}\epsilon_{\dot B)\dot C})p^{\dot C} + \sigma_{\dot A\dot B}. \tag{4.38}$$

Now, only Eq. (4.30) and inequality (4.32) remain to be satisfied. Simple manipulations show that the equation $\delta^{\dot A}\partial^{\dot B}Q_{\dot A\dot B} = 0$ with $Q_{\dot A\dot B}$ given by (4.38) is equivalent to three equations for the functions of $q^{\dot A}$ only:

$$\frac{\partial a_{\dot A\dot B}}{\partial q_{\dot B}} + (\rho_{\dot B\dot C\dot A} + b_{(\dot B}\epsilon_{\dot C)\dot A})a^{\dot B\dot C} = 0, \tag{4.39}$$

$$\frac{\partial b_{\dot A}}{\partial q_{\dot A}} + \frac{4}{3}\sigma_{\dot A\dot B}a^{\dot A\dot B} = 0. \tag{4.40}$$

We intend to solve these equations but first we would like to examine the transformations that leave the form of metric (4.20) unchanged. The general transformation of this type is

$$\tilde q^{\dot A} = \tilde q^{\dot A}(q^{\dot B}), \quad \tilde p_{\dot A} = T^{-1\dot B}{}_{\dot A}(p_{\dot B} - \xi_{\dot B}), \qquad T^{\dot A}{}_{\dot B} := \frac{\partial \tilde q^{\dot A}}{\partial q^{\dot B}}, \quad \xi_{\dot A} = \xi_{\dot A}(q^{\dot B}), \tag{4.41}$$

with $T := \det\|T^{\dot A}{}_{\dot B}\| \neq 0$ (compare Ref. 18).

One finds that (4.41) leaves the form of metric (4.20) unchanged:

$$g = 2\,d\tilde q^{\dot A} \underset{s}{\otimes} (d\tilde p_{\dot A} + \tilde Q_{\dot A\dot B}\,d\tilde q^{\dot B}),$$

$$\tilde Q_{\dot A\dot B} = \tilde a_{\dot C(\dot A}\epsilon_{\dot B)\dot D}\tilde p^{\dot C}\tilde p^{\dot D} + (\tilde\rho_{\dot A\dot B\dot C} + \tilde b_{(\dot A}\epsilon_{\dot B)\dot C})\tilde p^{\dot C} + \tilde\sigma_{\dot A\dot B},$$

where

$$\begin{aligned}
\tilde a_{\dot A\dot B} &= TT^{-1\dot C}{}_{\dot A}T^{-1\dot D}{}_{\dot B}a_{\dot C\dot D},\\
\tilde\rho_{\dot A\dot B\dot C} &= TT^{-1\dot D}{}_{\dot A}T^{-1\dot E}{}_{\dot B}T^{-1\dot F}{}_{\dot C}[\rho_{\dot D\dot E\dot F} + a_{(\dot D\dot E}\xi_{\dot F)} - T^{\dot H}{}_{(\dot D,\dot E}T^{-1}{}_{\dot F)\dot H}],\\
\tilde b_{\dot A} &= T^{-1\dot B}{}_{\dot A}[b_{\dot B} + \tfrac{4}{3}a_{\dot B\dot C}\xi^{\dot C} + \tfrac{2}{3}(\ln T)_{,\dot B}],\\
\tilde\sigma_{\dot A\dot B} &= T^{-1\dot C}{}_{\dot A}T^{-1\dot D}{}_{\dot B}[\sigma_{\dot C\dot D} + (\rho_{\dot C\dot D\dot E} + b_{(\dot C}\epsilon_{\dot D)\dot E})\xi^{\dot E} + a_{\dot E(\dot C}\epsilon_{\dot D)\dot F}\xi^{\dot E}\xi^{\dot F} + \xi_{(\dot C,\dot D)}]
\end{aligned} \tag{4.42}$$

(the symbol "$_{,\dot A}$" denotes $\partial/\partial q^{\dot A}$).

Now we are going to utilize the transformations (4.41) to simplify the final form of our metric.

First consider the case

(i) $a^{\dot A\dot B}a_{\dot A\dot B} \neq 0$.

We can choose $\tilde q^{\dot A} = \tilde q^{\dot A}(q^{\dot B})$, $\xi_{\dot B} = 0$ in (4.41) so that (we omit the overtildes)

$$a_{\dot 1\dot 1} = a_{\dot 2\dot 2} = 0, \quad a_{\dot 1\dot 2} \neq 0 \tag{4.43}$$

(this problem resembles the problem of characteristic curves in the theory of partial differential equations). With (4.43) fixed one has

$$\sigma_{\dot A\dot B}a^{\dot A\dot B} = -2\sigma_{\dot 1\dot 2}a_{\dot 1\dot 2}. \tag{4.44}$$

Now, from (4.42) we conclude that there exists functions $\xi_{\dot A} = \xi_{\dot A}(q^{\dot B})$ with $T^{\dot A}{}_{\dot B} = \delta^{\dot A}{}_{\dot B}$ such that

$$\sigma_{\dot 1\dot 2} = \sigma_{\dot 2\dot 2} = 0. \tag{4.45}$$

Hence, by (4.44), it follows that

$$\sigma_{\dot A\dot B}a^{\dot A\dot B} = 0, \tag{4.46}$$

and then from Eq. (4.40) with (4.46) one infers that there exists a function $b = b(q^{\dot A})$ such that

$$b_{\dot A} = b_{,\dot A}. \tag{4.47}$$

Substituting (4.43) and (4.47) into Eqs. (4.39) we find

$$\rho_{\dot 1\dot 1\dot 2} = \frac{\partial[\frac{1}{2}(b - \ln a_{\dot 1\dot 2})]}{\partial q^{\dot 1}}, \qquad \rho_{\dot 2\dot 2\dot 1} = -\frac{\partial[\frac{1}{2}(b - \ln a_{\dot 1\dot 2})]}{\partial q^{\dot 2}}. \tag{4.48}$$

Then one has $Q_{\dot A\dot B}$ defined by (4.38) in the form

$$\begin{aligned}
Q_{\dot 1\dot 1} &= Ap^2{}_{\dot 1} + B_{,\dot 1}p_{\dot 1} + Cp_{\dot 2} + D,\\
Q_{\dot 1\dot 2} &= (\ln A)_{,(\dot 1}p_{\dot 2)},\\
Q_{\dot 2\dot 2} &= -Ap^2{}_{\dot 2} + Ep_{\dot 1} + B_{,\dot 2}p_{\dot 2},
\end{aligned} \tag{4.49}$$

where

$$A := a_{\dot 1\dot 2}, \quad B := \tfrac{3}{2}b - \tfrac{1}{2}\ln a_{\dot 1\dot 2}, \quad C := -\rho_{\dot 1\dot 1\dot 1}, \quad D := \sigma_{\dot 1\dot 1}, \quad E := \rho_{\dot 2\dot 2\dot 2}$$

[A, B, C, D, and E are arbitrary functions ($A \neq 0$) of $q^{\dot A}$].

Consider now the case

(ii) $a^{\dot A\dot B}a_{\dot A\dot B} = 0, \quad \sum_{\dot A,\dot B=\dot 1}^{\dot 2}|a_{\dot A\dot B}| \neq 0.$

In this case we can choose $\tilde q^{\dot A} = \tilde q^{\dot A}(q^{\dot B})$, $\xi_{\dot B} = 0$, in (4.41) such that

$$a_{\dot 2\dot 2} = a_{\dot 1\dot 2} = 0, \quad a_{\dot 1\dot 1} \neq 0. \tag{4.50}$$

With (4.50) fixed one has

$$\sigma_{\dot A\dot B}a^{\dot A\dot B} = \sigma_{\dot 2\dot 2}a_{\dot 1\dot 1}. \tag{4.51}$$

Then we choose the transformation (4.41) with $T^{\dot A}{}_{\dot B} = \delta^{\dot A}{}_{\dot B}$ and $\xi_{\dot A} = \xi_{\dot A}(q^{\dot B})$ such that (4.45) holds. Therefore, by (4.51), $\sigma_{\dot A\dot B}a^{\dot A\dot B} = 0$ and from Eq. (4.40) it follows that there exists a function $b = b(q^{\dot A})$ such that (4.47) holds. Consequently, from Eqs. (4.39) one finds

$$\rho_{\dot 2\dot 2\dot 2} = 0, \quad \rho_{\dot 2\dot 2\dot 1} = \frac{\partial[b - \ln a_{\dot 1\dot 1}]}{\partial q^{\dot 2}}. \tag{4.52}$$

Finally we obtain $Q_{\dot A\dot B}$ in the form

$$\begin{aligned}
Q_{\dot 1\dot 1} &= Ap_{\dot 1}p_{\dot 2} + Bp_{\dot 1} + Cp_{\dot 2} + D,\\
Q_{\dot 1\dot 2} &= \frac{A}{2}p^2{}_{\dot 2} + \frac{\partial(\frac{3}{2}b - \ln A)}{\partial q^{\dot 2}}p_{\dot 1} + \left(\frac{3}{2}b_{,\dot 1} - B\right)p_{\dot 2},\\
Q_{\dot 2\dot 2} &= \frac{\partial \ln A}{\partial q^{\dot 2}}p_{\dot 2},
\end{aligned} \tag{4.53}$$

where $A:=-a_{\dot{1}\dot{1}}$, $B:=\rho_{\dot{1}\dot{1}\dot{2}}+b_{,\dot{1}}$, $C:=-\rho_{\dot{1}\dot{1}\dot{1}}$, $D:=\sigma_{\dot{1}\dot{1}}$, and b are arbitrary functions of $a^{\dot{A}}$.

Finally, consider the case

(iii) $a_{\dot{A}\dot{B}}=0$.

Then (4.46) holds and $b_{\dot{A}}$ is of the form (4.47). From formulas (4.42) it follows that one can choose the transformation (4.41) with $\xi_{\dot{A}}=0$ such that

$$b=0 \quad \text{and} \quad \rho_{\dot{2}\dot{2}\dot{2}}=0. \tag{4.54}$$

With (4.54) fixed we can also perform the transformation (4.41) of the form $\xi_{\dot{A}}=\xi_{\dot{A}}(q^{\dot{B}})$, $T^{\dot{A}}{}_{\dot{B}}=\delta^{\dot{A}}{}_{\dot{B}}$, in such a manner that $\sigma_{\dot{A}\dot{B}}$ satisfies (4.45).

Thus we arrive at the following formulas for $Q_{\dot{A}\dot{B}}$:

$$\begin{aligned} &Q_{\dot{1}\dot{1}}=Bp_{\dot{1}}+Cp_{\dot{2}}+D, \quad Q_{\dot{1}\dot{2}}=Ep_{\dot{1}}-Bp_{\dot{2}}, \\ &Q_{\dot{2}\dot{2}}=-Ep_{\dot{2}}, \end{aligned} \tag{4.55}$$

where $B:=\rho_{\dot{1}\dot{1}\dot{2}}$, $C:=-\rho_{\dot{1}\dot{1}\dot{1}}$, $D:=\sigma_{\dot{1}\dot{1}}$, and $E:=\rho_{\dot{2}\dot{2}\dot{1}}$ are arbitrary functions of $q^{\dot{A}}$.

Notice that the expressions (4.55) can be obtained from (4.53) by a limiting transition consisting of setting $b=\epsilon b'$ and $A=\epsilon A'$ with $\epsilon=\text{const}\to 0$.

Having $Q_{\dot{A}\dot{B}}$ for our three cases [formulas (4.49), (4.53), and (4.55)] we can find $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ from (4.27).

Straightforward computations give the following.

For (i),

$$\begin{aligned} &C_{11\dot{A}\dot{B}}=0, \\ &C_{12\dot{1}\dot{1}}=C_{12\dot{2}\dot{2}}=0, \\ &C_{12\dot{1}\dot{2}}=-A, \\ &C_{22\dot{1}\dot{1}}=-A_{,\dot{1}}p_{\dot{1}}+2ACp_{\dot{2}}+C_{,\dot{2}}-CB_{,\dot{2}}-\tfrac{1}{2}B_{,\dot{1}}(\ln A)_{,\dot{1}} \\ &\qquad +\tfrac{1}{4}[(\ln A)_{,\dot{1}}]^2+\tfrac{1}{2}C(\ln A)_{,\dot{2}}-\tfrac{1}{2}(\ln A)_{,\dot{1}\dot{1}}, \\ &C_{22\dot{1}\dot{2}}=2A_{,[\dot{1}}p_{\dot{2}]}+\tfrac{1}{2}(\ln A)_{,\dot{1}\dot{2}} \\ &\qquad -\tfrac{1}{4}(\ln A)_{,\dot{1}}(\ln A)_{,\dot{2}}-B_{,\dot{1}\dot{2}}+EC, \\ &C_{22\dot{2}\dot{2}}=A_{,\dot{2}}p_{\dot{2}}-\tfrac{1}{2}(\ln A)_{,\dot{2}\dot{2}}+\tfrac{1}{2}E(\ln A)_{,\dot{1}}+\tfrac{1}{4}[(\ln A)_{,\dot{2}}]^2 \\ &\qquad +E_{,\dot{1}}-\tfrac{1}{2}B_{,\dot{2}}(\ln A)_{,\dot{2}}-EB_{,\dot{1}} \end{aligned} \tag{4.56}$$

([$\dot{1}\dot{2}$] denotes the antisymmetrization over indices $\dot{1},\dot{2}$).

For (ii),

$$\begin{aligned} &C_{11\dot{A}\dot{B}}=0, \\ &C_{12\dot{1}\dot{2}}=C_{12\dot{2}\dot{2}}=0, \\ &C_{12\dot{1}\dot{1}}=A, \\ &C_{22\dot{1}\dot{1}}=(\tfrac{3}{2}Ab_{,\dot{1}}-A_{,\dot{1}}-2AB)p_{\dot{2}}+(\tfrac{3}{2}Ab_{,\dot{2}}-A_{,\dot{2}})p_{\dot{1}} \\ &\qquad +C_{,\dot{2}}-\tfrac{3}{2}b_{,\dot{1}\dot{1}}+B_{,\dot{1}}+(\tfrac{3}{2}b_{,\dot{1}}-B)(\tfrac{3}{2}b_{,\dot{1}}-2B) \\ &\qquad +C[\tfrac{3}{2}b_{,\dot{2}}-2(\ln A)_{,\dot{2}}], \\ &C_{22\dot{1}\dot{2}}=(A_{,\dot{2}}-\tfrac{3}{2}Ab_{,\dot{2}})p_{\dot{2}}-B_{,\dot{2}}+\tfrac{3}{2}b_{,\dot{1}\dot{2}}-(\ln A)_{,\dot{2}\dot{1}} \\ &\qquad -(\tfrac{3}{2}b_{,\dot{1}}-B)[\tfrac{3}{2}b_{,\dot{2}}-(\ln A)_{,\dot{2}}], \\ &C_{22\dot{2}\dot{2}}=(\ln A)_{,\dot{2}\dot{2}}-\tfrac{3}{2}b_{,\dot{2}\dot{2}} \\ &\qquad -[\tfrac{3}{2}b_{,\dot{2}}-(\ln A)_{,\dot{2}}][2(\ln A)_{,\dot{2}}-\tfrac{3}{2}b_{,\dot{2}}]. \end{aligned} \tag{4.57}$$

For (iii), the spinor $C_{\dot{A}\dot{B}\dot{C}\dot{D}}$ in this case can be obtained from (ii) by the limiting transition $b\to 0$, $A\to 0$, $(\ln A)_{,\dot{2}}\to -E$. The result reads

$$\begin{aligned} &C_{11\dot{A}\dot{B}}=0, \\ &C_{12\dot{A}\dot{B}}=0, \\ &C_{22\dot{1}\dot{1}}=C_{,\dot{2}}+B_{,\dot{1}}+2B^2+2EC, \\ &C_{22\dot{1}\dot{2}}=-B_{,\dot{2}}+E_{,\dot{1}}+BE, \\ &C_{22\dot{2}\dot{2}}=-E_{,\dot{2}}+2E^2. \end{aligned} \tag{4.58}$$

The expressions for C_{2222} are involved and we do not give them here. But it is evident that in general C_{2222} does not vanish.

In summary, we have found all two-sided conformally recurrent four-dimensional Riemannian manifolds of the type $N\otimes[\ -\ ]$. In the case of CR the coordinates $q^{\dot{A}}$ and $p^{\dot{B}}$ are complex and all functions considered are holomorphic; in the case of UR, $q^{\dot{A}}$ and $p^{\dot{B}}$ are real and all functions are real. For (i) and (ii) the metric (4.20) contains five arbitrary functions of variables $q^{\dot{A}}$ [formulas (4.49) or (4.53), respectively]; for (iii) the metric contains four arbitrary functions of variables $q^{\dot{A}}$ [the formulas (4.55)].

ACKNOWLEDGMENTS

One of us (M.P.) is grateful to all members of Departamento de Física, Centro de Investigación del I.P.N., México, D.F., and especially to Dr. A. Zepeda, Dr. A. García Díaz, and Dr. B. Mielnik for warm hospitality during his stay at the Centro.

The assistance of the Secretario Académico, Dr. E. Campesino, is appreciated.

The work of M.P. was supported in part by the CONACyT, Ciudad Universitaria 04515, México, D.F., and by the Centro de Investigación y de Estudios Avanzadas del I.P.N.

[1] R. G. McLenaghan and J. Leroy, Proc. R. Soc. London Ser. A **327**, 229 (1972).
[2] J. Ehlers and W. Kundt, in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).
[3] G. S. Hall, J. Phys. A **7**, L42 (1974).
[4] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of the Einstein's Field Equations*, edited by E. Schmutzer (Deutscher Verlag der Wisenschaften, Berlin, 1980), p. 347.
[5] J. F. Plebański and K. Rózga, J. Math. Phys. **25**, 1930 (1984).
[6] J. F. Plebański and M. Przanowski, Acta Phys. Pol. B **19**, 805 (1988).
[7] J. F. Plebański, "Spinors, tetrads and forms," Centro de Investigación y de Estudios Avanzados del I.P.N., México, D.F., unpublished monograph, 1974.
[8] M. Przanowski and J. F. Plebański, Acta Phys. Pol. B **10**, 485 (1979).
[9] H. S. Ruse, Proc. London Math. Soc. **53**, 13 (1951). See, also, H. S. Ruse, A. G. Walker, and T. J. Wilmore, *Harmonic Spaces* (Edizioni Cremonese, Roma, 1961).
[10] T. Adati and T. Miyazawa, Tensor **18**, 348 (1967).
[11] M. C. Chaki and B. Gupta, Ind. J. Math. **5**, 113 (1963).
[12] J. F. Plebański, *Lectures on Non-linear Electrodynamics* (Niels Bohr Institute Nordita, Copenhagen, 1968).
[13] B. Bertotti, Phys. Rev. **116**, 1331 (1959).
[14] I. Robinson, Bull. Acad. Pol. Sci., Ser. Math. Phys. **7**, 352 (1959).
[15] A. G. Walker, Q. Il. Math. **1**, 69, 147 (1950).
[16] R. Debever and M. Cahen, Bull. Acad. R. Belg. Cl. Sci. **47**, 491 (1961).
[17] R. Debever and J. Spelkens, Bull. Acad. R. Belg. Cl. Sci. **54**, 116 (1968).
[18] J. D. Finley, III and J. F. Plebański, J. Math. Phys. **17**, 2207 (1976).

Nonstatic charged spheres admitting a conformal Killing vector

H. Rago[a)]
Departmento de Fisica, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela and Laboratorio de Fisica Teórica, Departamento de Fisica, Facultad de Ciencias, Universidad de Los Andes, Mérida 5101, Venezuela

(Received 12 February 1988; accepted for publication 5 April 1989)

Exact, nonstatic, spherically symmetric solutions of the Einstein–Maxwell equations are found for self-gravitating charged spheres under the assumption of the existence of a conformal Killing vector. Solutions are matched to the Reissner–Nordstrom metric and it is found that as a consequence of the junction conditions, the material must be anisotropic. The radius of the sphere for static distributions corresponds to the radius of the unstable null circular orbit of the Reissner–Nordstrom geometry.

I. INTRODUCTION

Exact analytic solutions of the Einstein–Maxwell equations for charged spheres have been studied in both static and dynamic cases since the discussion by Bonnor[1] on the equilibrium of charged dust.[2–7] Besides assuming spherical symmetry, several authors[8–11] make the further assumption that space-time admits a conformal Killing vector ξ, in order to aid in the solution of the Einstein field equations, that is,

$$L_\xi g_{\alpha\beta} = \Psi g_{\alpha\beta}; \tag{1}$$

where the left-hand side is the Lie derivative of the metric tensor $g_{\alpha\beta}$ and Ψ is an arbitrary function of space-time coordinates. This symmetry, which is the simplest generalization of a homothetic motion (Ψ = constant), allows the integration of the field equation in the nonstatic case for neutral spheres[8] and in the static case for charged ones.[9,10]

In this paper we show how the analytical integration of the Einstein–Maxwell field equations under the above assumption may be accomplished, leading to solutions representing different cases of self-gravitating charged spheres. The solutions are matched to the Reissner–Nordstrom metric and the junction conditions give the evolution of the boundary surface. We find an oscillating distribution and another expanding asymptotically to a static sphere lying outside the singularity. In the neutral fluid limit $q = 0$, known results can be recovered.

The discussion is organized as follows. The field equations and the implications of conformal motion are presented in Sec. II. Section III includes the matching conditions and the evolution of the boundary. Finally in Sec. IV the interior solutions are proposed in terms of an arbitrarily chosen function and, as an example, a particular model is considered.

II. THE FIELD EQUATIONS

If comoving coordinates $x^\alpha = (t,r,\theta,\varphi)$ are adopted, the lines element is given by

$$ds^2 = e^\nu dt^2 - e^\lambda dr^2 - R^2(d\theta^2 + \sin^2\theta\, d\varphi^2) \tag{2}$$

where ν, λ, and R are unknown functions of the Lagrangian coordinates comoving with the fluid r, and t.

Consider the Einstein–Maxwell system

$$R^\alpha_\beta - \tfrac{1}{2}\delta^\alpha_\beta R = 8\pi T^\alpha_\beta = 8\pi(M^\alpha_\beta + E^\alpha_\beta), \tag{3}$$

$$F^{\alpha\beta}_{;\beta} = 4\pi j^\alpha \tag{4}$$

$$F_{\alpha\beta,\sigma} + F_{\beta\sigma,\alpha} + F_{\sigma\alpha,\beta} = 0 \tag{5}$$

where $M_{\alpha\beta}$, the energy tensor for anisotropic matter, is given by

$$M_{\alpha\beta} = (\rho + p_\perp)u_\alpha u_\beta - p_\perp g_{\alpha\beta} + (p_r - p_\perp)\chi_\alpha\chi_\beta \tag{6}$$

and $u^\alpha = e^{-\nu/2}\delta^\alpha_0$ is the four-velocity, χ^α a unitary vector in the radial direction, and p_r and $p_\perp$ are the radial and tangential stresses, respectively. The electromagnetic energy tensor $E_{\alpha\beta}$, in terms of the Maxwell field tensor $F_{\alpha\beta}$, takes its usual expression

$$E^\alpha_\beta = [1/(4\pi)][F^{\alpha\lambda}F_{\lambda\beta} + \tfrac{1}{4}\delta^\alpha_\beta F_{\mu\nu}F^{\nu\mu}]. \tag{7}$$

Finally, j^α is the electric current vector which is proportional to the four-velocity u^α.

For a spherically symmetric charge distribution, F^{01} will be the only surviving component of $F^{\alpha\beta}$, and thus Eq. (5) is satisfied identically. Furthermore, since we are using comoving coordinates only j^0 is nonvanishing; hence Eq. (4) is integrated to give

$$F^{01} = Q(r)e^{-1/2(\nu+\lambda)}/R^2, \tag{8}$$

where

$$Q(r) = \int_0^r 4\pi R^2 j^0 e^{1/2(\nu+\lambda)}\, dr \tag{9}$$

represents the constant electric charge inside a sphere of radius r.

It can be shown (see Ref. 8 for details with a slightly different notation) that the assumption that space-time admits a conformal Killing vector

$$L_\alpha g_{\alpha\beta} = \xi_{\alpha;\beta} + \xi_{\beta;\alpha} = \Psi g_{\alpha\beta}, \tag{10}$$

restricted by demanding that the vector field ξ is orthogonal to the velocity (i.e., $\xi_\alpha u^\alpha = 0$), implies that the metric functions can be written as

$$e^\nu = e^{\lambda - f(t)}, \tag{11}$$

$$R^2 = e^{\lambda - f(t)}/\omega^2, \tag{12}$$

[a)] Postal Address: Apartado 32, Ipostel La Hechicera, Merida, Venezuela.

 0022-2488/89/092110-04$02.50

where $f(t)$ is a function of t and ω is a positive constant. Moreover, using (11) and (12) the field equation $R^0_1 = 0$ can be integrated to obtain

$$e^{-\lambda/2} = h(r) + g(t), \tag{13}$$

where $h(r)$ and $g(t)$ are unknown dimensionless functions of their arguments. Thus the line element (2) can be expressed in the form

$$ds^2 = R^2(t,r)[\omega^2 dt^2 - e^f \omega^2 dr^2 - d\Omega^2] \tag{14}$$

with $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\varphi^2$ and

$$R(t,r) = e^{-f(t)/2}/\omega[h(r) + g(t)]. \tag{15}$$

Taking into account this line element and Eqs. (6)–(8), Einstein equations become

$$8\pi\rho + Q^2/R^4 = -(3h'^2 - 3\dot{g}^2 e^f) + 2e^{-\lambda/2}(h'' + \dot{g}\dot{f}e^f) + e^{-\lambda}e^f(\dot{f}^2/4 + \omega^2), \tag{16}$$

$$8\pi p_r - Q^2/R^4 = 3h'^2 - 3\dot{g}^2 e^f + e^{-\lambda/2}e^f(2\ddot{g} - \dot{g}\dot{f}) + e^{-\lambda}e^f(\ddot{f} - \dot{f}^2/4 - \omega^2), \tag{17}$$

$$8\pi p_\perp - Q^2/R^4 = 3h'^2 - 3\dot{g}^2 e^f - 2e^{-\lambda/2}(h'' - \ddot{g}e^f) + e^{-\lambda}e^f\ddot{f}/2, \tag{18}$$

and the electric field intensity defined as $E = (-F^{01}F_{01})^{1/2}$ results:

$$E(r,t) = Q(r)/R^2(r,t). \tag{19}$$

Dots and primes denote hereafter differentiation with respect to t and r, respectively.

III. THE JUNCTION CONDITIONS AND THE EVOLUTION OF THE BOUNDARY

Let the boundary of the distribution be a timelike three-space denoted by Σ. The interior space-time V^-, with coordinates $x^\alpha_- = (t,r,\theta,\varphi)$ and metric given by Eq. (14), is to be matched across Σ to the exterior Reissner–Nordstrom spacetime V^+ with coordinates $x^\alpha_+ = (T,R,\theta,\varphi)$ and metric given by

$$ds^2_+ = \left(1 - \frac{2M}{R} + \frac{q^2}{R^2}\right)dT^2 - \left(1 - \frac{2M}{R} + \frac{q^2}{R^2}\right)^{-1} dR^2 - R^2 d\Omega^2. \tag{20}$$

The intrinsic metric to Σ can be given as

$$ds^2_\Sigma = g_{ij}\, d\xi^i\, d\xi^j = d\tau^2 - S^2(\tau)d\Omega^2, \tag{21}$$

where the Gaussians coordinates ξ^i are τ, θ, and φ.

Following Israel,[12] we will demand continuity of the intrinsic metric so that when approaching Σ from V^+ or V^-, we have

$$[q_{ij}] = 0, \tag{22}$$

where $[a] \equiv a^+ - a^-$. The second condition imposed on Σ is the continuity of its extrinsic curvature or second fundamental form

$$[K_{ij}] = 0, \tag{23}$$

where the three tensor K_{ij} is given in terms of the unit space-like normal vectors $n^\pm{}_\alpha$, as

$$K^\pm_{ij} = -n^\pm_\alpha\left[\frac{\partial^2 x^\alpha_\pm}{\partial\xi^i\partial\xi^j} + \dot{\Gamma}^\alpha_{\mu\nu}\frac{\partial x^\mu_\pm}{\partial\xi^i}\frac{\partial x^\nu_\pm}{\partial\xi^j}\right]. \tag{24}$$

In writing Eq. (23) the absence of a thin mass shell is assumed; also we will assume the continuity of the electric field across Σ, thus there is not surface concentration of charge and as a consequence, $Q(r_0) = q$. It can be shown[13] from Eqs. (22) and (23) and the Gauss–Codazzi contracted equations that

$$[T_{\alpha\beta}n^\alpha n^\beta] = 0, \tag{25}$$

$$[T_{\alpha\beta}n^\alpha n^\beta] = 0. \tag{26}$$

(i) The interior space-time. For the interior spacetime V^- the junction condition (22) gives

$$R(t,r_0) = S(\tau), \tag{27}$$

$$R(t,r_0)\omega = \frac{d\tau}{dt}, \tag{28}$$

where r_0 is the value of the radial coordinate at Σ. The equation of the surface is

$$f^-(r,t) = r - r_0 = 0, \tag{29}$$

hence, the unit normal vector is

$$n^-_\alpha = R(t,r_0)\omega e^{f/2}\delta^1{}_\alpha \tag{30}$$

and the unit tangent vector is

$$u^\alpha_- = \frac{dt}{d\tau}\delta^\alpha_0. \tag{31}$$

Using Eq. (24), the only nonvanishing components of the extrinsic curvature are calculated to be

$$K^-_{\tau\tau} = -n^-_\alpha\left[\frac{du^\alpha_-}{d\tau} + \Gamma^\alpha_{\mu\nu}u^\mu_- u^\nu_-\right] = \frac{e^{-f/2}}{\omega R^2}R'\Big|_\Sigma \tag{32}$$

and

$$K^-_{\theta\theta} = \frac{1}{\sin^2\theta}K^-_{\varphi\varphi} = -n^-_\alpha\,\Gamma^\alpha_{\theta\theta} = \frac{e^{-f/2}}{\omega}R'\Big|_\Sigma. \tag{33}$$

(ii) The exterior space-time. The equation of the surface Σ is now given by

$$f^+(T,R) = R - R_b(T) = 0. \tag{34}$$

The unit normal vector and the four-velocity are

$$n^+_\alpha = -\frac{dR}{d\tau}\delta^0_\alpha + \frac{dT}{d\tau}\delta^1_\alpha, \tag{35}$$

$$u^\alpha_+ = \frac{dT}{d\tau}\delta^\alpha_0 + \frac{dR}{d\tau}\delta^\alpha_i. \tag{36}$$

The junction condition (22) gives now

$$R_b(T) = S(\tau), \tag{37}$$

$$\left(1 - \frac{2M}{R} + \frac{q^2}{R^2}\right)\left(\frac{dT}{d\tau}\right)^2 - \left(1 - \frac{2M}{r} + \frac{q^2}{R^2}\right)^{-1}\left(\frac{dR}{d\tau}\right)^2 = 1. \tag{38}$$

The nonvanishing components of the extrinsic curvature are

$$K^+_{\tau\tau} = -n^+_\alpha\frac{Du^\alpha}{D\tau}, \tag{39}$$

$$K^+_{\theta\theta} = \frac{1}{\sin^2\theta}K^+_{\varphi\varphi} = \frac{dT}{d\tau}\left(1 - \frac{2M}{R} + \frac{q^2}{R^2}\right)R. \tag{40}$$

The junction condition $K^{-}_{\theta\theta} = K^{+}_{\theta\theta}$ gives

$$\frac{e^{-f/2}}{\omega} R'\Big|_{\Sigma} = \frac{dT}{d\tau}\left(1 - \frac{2M}{R_b} + \frac{Q^2}{R_b^2}\right) R_b. \tag{41}$$

Using Eq. (15) to bring R into the form $R'|_{\Sigma} = -h'(r_0)R_b^2$ and taking into account Eqs. (28) and (38), we obtain finally

$$\dot{R}_b^2/\omega^2 = 2MR_b - R_b^2 + h'^2(r_0)R_b^4 - q^2. \tag{42}$$

Equation (42) can also be obtained from the continuity of the effective gravitational mass $m(t,r)$, defined by the relation

$$2m(t,r) = (g_{\theta\theta})^{3/2} R^{\theta\varphi}_{\theta\varphi}, \tag{43}$$

where $R^{\theta\varphi}_{\theta\varphi}$ is the mixed angular component of the Riemann–Christoffel curvature tensor. In fact, the continuity of $K_{\theta\theta}$ implies the continuity of the mass function.[13]

On the other hand, Eq. (25) gives in our case the continuity of the radial pressure cross Σ,

$$[p_r] = 0. \tag{44}$$

Using the field equation (17) and taking into account that $p_r^+ \equiv 0$ and $Q(r_0) = q$ (no surface concentration of charge), we obtain

$$(2\ddot{R}_b R_b/\omega^2) - (\dot{R}_b^2/\omega^2) = 3h'^2(r_0)R_b^4 - R_b^2 + q^2. \tag{45}$$

Equation (42) shows that the boundary of the distribution evolves slower than in the neutral case $q = 0$. Notice that if the model is static or has a static limit ($\ddot{R} = \dot{R} = 0$), Eqs. (42) and (45) imply that the radius of the distribution will be given by

$$R_{st} = \tfrac{1}{2}[3M \pm (9M^2 - 8q^2)^{1/2}], \tag{46}$$

which corresponds to the radius of the unstable null circular orbits in the exterior Reissner–Nordstrom space-time.[14] Finally, observe that Eq. (45) is just the time derivative of Eq. (42); hence, the evolution of the boundary will be known once we succeed in integrating this equation.

General solutions of Eq. (42) are expressed in terms of elliptical functions. However, it is also possible to obtain solutions in terms of elementary functions for very particular forms of $H'(r_0)$. We consider two cases.

(i) If the function $h(r)$ is such that $h'(r_0) = 0$, then the first solution becomes

$$R_b(t) = M - (M^2 - q^2)^{1/2} \cos \omega t, \tag{47}$$

which represents a sphere whose surface oscillates between its gravitational radius $R^+ = M + (M^2 - q^2)^{1/2}$ and the inner horizon $R^- = M - (M^2 - q^2)^{1/2}$, as seen by an observer comoving with the matter. For an exterior observer, the boundary is not oscillating but propagating forward inside the event horizon. Note that for the extremal case $q/M = \pm 1$, the distribution is static with a radius equal to M.

(ii) If the function $h(r)$ is such that

$$h'^2(r_0) = (MR_{st} - q^2)/R_{st}^4, \tag{48}$$

the integration of Eq. (45) can be carried out to give

$$R_b(t) = R_{st}\left[1 - \frac{A}{2 + \cosh\sqrt{kA}\,\omega(t - t_0) + [q^2 h'^2(r_0)/2k^2]e^{-\sqrt{kA}\,\omega(t - t_0)}}\right], \tag{49}$$

where

$$k = (MR_{st} - q^2)/R_{st}^2$$
$$A = 3 - q^2 h'^2(r_0)/k^2.$$

This solution represents an expanding boundary which tends asymptotically to a radius given by Eq. (46). For the critical charge–mass relation $q/M = \pm \frac{8}{9}$, the models are static since $A = 0$ Eq. (49) with a radius equal to $3M/2$. Note that in both cases for $q = 0$, known results for neutral fluid can be recovered.[8]

IV. THE SOLUTIONS

In this section we present some particular model with boundaries evolving according to the solutions found above. We select for simplicity $g(t) = 0$, such that from Eq. (15) we have

$$R_b(t) = e^{-f(t)/2}/wh(r_0). \tag{50}$$

Thus the surface equations (42) and (45) in terms of $f(t)$ are given by

$$e^f(\ddot{f} - \dot{f}^2/4 - \omega^2) = -3d^2 - q^2\omega^4 h^2(r_0)e^{2f}, \tag{51}$$

$$\dot{f}^2/4 + \omega^2 = 2M\omega^3 h(r_0)e^{f/2} + d^2 e^{-f} - q^2 w^4 h^2(r_0)e^f, \tag{52}$$

where $d^2 = h'^2(r_0)/h^2(r_0)$.

Using Eqs. (50)–(52), the field equations (16)–(19) with $g(t) = 0$ become

$$8\pi p_r = 3h'^2(r) - 3d^2h^2(r) + [h^2(r)/h^4(r_0)R_b^4] \times [Q^2(r)h^2(r) - q^2h^2(r_0)], \tag{53}$$

$$8\pi p_\perp = 3h'^2(r) - 2h(r)h''(r) - d^2h^2(r) + Mh^2(r)/R_b^3 h^2(r_0) - [h^2(r)/h^4(r_0)R_b^4] \times [Q^2(r)h^2(r) + q^2 h_0^2(r_0)], \tag{54}$$

$$8\pi p = 3h'^2(r) + 2h(r)h''(r) + d^2h^2(r) + 2Mh^2(r)/R_b^3 h^2(r_0) - [h^2(r)/h^4(r_0)R_b^4] \times [Q^2(r)h^2(r) + q^2h^2(r_0)], \tag{55}$$

$$E(r,t) = Q(r)h^2(r)/R_b^2 h^2(r_0). \tag{56}$$

Note that after making $p_r = p_\perp$, we get from Eqs. (53) and (54)

$$\frac{h''(r)}{h(r)} - d^2 = \frac{M}{2R_b^3 h(r_0)^2} - \frac{Q(r)^2}{R_b^4}\frac{h(r)^2}{h(r_0)^4}.$$

This equation is contradictory as the left-hand side depends only on r whereas the right-hand side depends on time. Thus local isotropic pressure is forbidden and, consequently, the

material must be anisotropic, a peculiarity shared with the neutral case.[8]

Note also that the models are completely specified for a given charge distribution, by giving explicitly the function $h(r)$. This function is not completely arbitrary since the energy density and the stresses depend upon it. Therefore, after giving explicitly a function $h(r)$, it must be verified if the energy tensor satisfies the energy conditions in order to obtain physically reasonable models.

As an example consistent with case (ii), we consider the following choice:

$$Q(r) = q(r/r_0)^3 \tag{57}$$

and

$$h(r) = c^2/r^2. \tag{58}$$

Without loss of generality we can take $c = r_0$, and from Eq. (48) after some manipulations we have

$$r_0^2 = 4R_{\rm st}^2(1 - 2M/R_{\rm st} + q^2/R_{\rm st}^2)^{-1}. \tag{59}$$

Substituting Eqs. (57) and (58) in Eqs. (53)–(56), we obtain for the matter variables and the electric field

$$8\pi p_r = \left(\frac{r_0}{r}\right)^2\left[\frac{12}{r^2} - \frac{q^2}{R_b^4}\right]\left[\left(\frac{r_0}{r}\right)^2 - 1\right], \tag{60}$$

$$8\pi p_\perp = \frac{M}{R_b^3}\left(\frac{r_0}{r}\right)^4 - \frac{4r_0^2}{r^4} - \frac{q^2}{R_b^4}\left(\frac{r_0}{r}\right)^4\left[\left(\frac{r}{r_0}\right)^2 + 1\right], \tag{61}$$

$$8\pi p = \frac{2M}{R_b^3}\left(\frac{r_0}{r}\right)^4 + \frac{4r_0^2}{r^4} - \frac{q^2}{R_b^4}\left(\frac{r_0}{r}\right)^4\left[\left(\frac{r}{r_0}\right)^2 + 1\right], \tag{62}$$

$$E(r,t) = (q/R_b^2)(r_0/r), \tag{63}$$

and the line element becomes

$$ds^2 = R_b^2(r/r_0)^4[\omega^2\,dt^2 - (\omega^2\,dr^2/R_b^2) - d\Omega^2]. \tag{64}$$

As can be seen, the weak energy condition $T_{\alpha\beta}u^\alpha u^\beta \geqslant 0$ is satisfied for the above choice of $h(r)$. In fact, this inequality is just

$$8\pi p + Q^2/R^4 \geqslant 0,$$

but using Eq. (16) with $g(t) = 0$, this is equivalent to

$$-3h'^2 + 2hh'' + he^f(\dot{f}^2/4 + 1) \geqslant 0.$$

For $h(r) \propto r^{-2}$, the two first terms cancel out and the third one is always positive. Moreover, by inspection of Eqs. (61) and (62) we see that $\rho \geqslant p_\perp$. Also, it can be shown that $\rho \geqslant p_r$.

As $t \to \infty$, the distribution evolves towards the static situation described by Eqs. (61)–(63) substituting the function $R_b(t)$ by its limit $R_{\rm st}$ as given by Eq. (46). In particular, the line element becomes

$$ds^2 = R_{\rm st}^2(r/r_0)^4[\omega^2\,dt^2/R_{\rm st}^2 - d\Omega^2]. \tag{65}$$

We can introduce Schwarzschild-like coordinates (T,R) through the transformation

$$r = 2(1 - 2M/R_{\rm st} + q^2/R_{\rm st}^2)^{-1/2}\sqrt{R_{\rm st}R}, \tag{66}$$

$$t = (1 - 2M/R_{\rm st} + q^2/R_{\rm st}^2)(T/\omega R_{\rm st}), \tag{67}$$

which brings the line element (65) to the form

$$ds^2 = \left(1 - \frac{2M}{R_{\rm st}}\frac{q^2}{R_{\rm st}^2}\right)\frac{R^2}{R_{\rm st}^2}\,dT^2 - \left(1 - \frac{2M}{R_{\rm st}} + \frac{q^2}{R_{\rm st}^2}\right)^{-1} \times (R/R_{\rm st})\,dR^2 - R^2\,d\Omega^2,$$

which clearly shows that both the interior and the exterior metrics join smoothly at the boundary surface. Finally, it should be noted that space-time is not regular at the origin, and the energy density and the pressure diverge as R^{-2} (in Schwarzschild-like coordinates); thus this type of solutions must be considered as the outer envelope of a well-behaved central core.

ACKNOWLEDGMENTS

The author has benefited from the warm hospitality of the Facultad de Ciencias de la Universidad Central de Venezuela and fruitful discussions with Dr. L. Herrera. He also expresses his appreciation to Dr. C. Mendoza for his critical reading of the manuscript, as well as to the referee who made some interesting observations.

This work has been partially supported by the C.D.C.H.T.-U.L.A. under project C-187-82.

[1] W. B. Bonnor, Mon. Not. R. Astron. Soc. **129**, 443 (1965).
[2] Y. P. Sha and P. C. Vaidya, Ann. Inst. Henry Poincare **VI**, 219 (1967).
[3] M. C. Faulkes, Can. J. Phys. **47**, 1989 (1969).
[4] J. Bekenstein, Phys. Rev. D **4**, 2185 (1971).
[5] B. Mashoon and M. Hossein-Partovi, Phys. Rev. D **20**, 2455 (1979).
[6] R. A. Sussman, J. Math. Phys. **28**, 1118 (1987).
[7] A. Banerjee, N. Chakrabarty, and S. B. Dutta Choudhury, Acta Phys. Pol. B **7**, 675 (1976).
[8] L. Herrera and J. Ponce de Leon, J. Math. Phys. **26**, 2018 (1985).
[9] L. Herrera and J. Ponce de Leon, J. Math. Phys. **26**, 2302 (1985).
[10] K. D. Krori, P. Borgohain, K. Das, and A. Sarma, Can. J. Phys. **64**, 58 (1986).
[11] C. C. Dyer, G. C. McVittie, and L. M. Dattes, Gen. Relativ. Gravit. **19**, 887 (1987).
[12] W. Israel, Nuovo Cimento B **44**, 1 (1966); **48**, 463 (1966).
[13] K. Lake, "Some notes on the propagation of discontinuities in solutions to Einstein equations," in Proceedings of V Scola de Cosmologia y Gravitacao, Rio de Janeiro, Brazil, 1987, edited by M. Novello (World Scientific, Singapore, 1988).
[14] A. Armenti, Nuovo Cimento B **25**(2), 442 (1975).

Spin-$\frac{3}{2}$ perturbations of algebraically special solutions of the Einstein–Maxwell equations

G. F. Torres del Castillo
Departamento de Física Matemática, Instituto de Ciencias de la Universidad Autónoma de Puebla, 72000 Puebla, Mexico and Departamento de Física, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Apartado Postal 14-740, 07000 Mexico, D. F., Mexico

(Received 6 June 1988; accepted for publication 29 March 1989)

The equations for the spin-$\frac{3}{2}$ perturbations of the solutions of the Einstein–Maxwell equations given by the linearized O(2) extended supergravity are considered. It is shown that for each geodetic and shear-free principal null direction of the background electromagnetic field there exists a gauge-invariant quantity made out of the spin-$\frac{3}{2}$ field that satisfies a decoupled equation. In the case of type-D solutions with a nonsingular aligned electromagnetic field it is explicitly shown that the decoupled equations associated with the two principal null directions admit separable solutions and the separated functions obey certain differential relations.

I. INTRODUCTION

In recent years there has been considerable progress in the study of linear perturbations of gravitational fields and test massless fields of spin-0, $\frac{1}{2}$, and 1, especially in the case of the algebraically special solutions of the Einstein vacuum field equations, where, by means of the procedure introduced by Teukolsky,[1] one obtains a decoupled equation for a component of the field that gives useful information about the behavior of the perturbation. When there exists a suitably aligned background electromagnetic or neutrino field a similar treatment is applicable and one obtains a decoupled system of equations (see Ref. 2 and the references cited therein). Among the algebraically special solutions, type-D metrics are distinguished because of their separability properties[3–7]; for each massless perturbation of spin greater than zero there exist two decoupled equations which can be solved by separation of variables.

For a spin-$\frac{3}{2}$ field the usual massless free-field equations on a curved background have integrability conditions which severely restrict the possible solutions. On the other hand, supergravity theory gives a consistent system of equations for a spin-$\frac{3}{2}$ massless field coupled to a gravitational field, which reduces to Einstein's vacuum field equations when the spin-$\frac{3}{2}$ field vanishes. The supergravity field equations, linearized with respect to the spin-$\frac{3}{2}$ field about an algebraically special solution of the Einstein vacuum field equations, lead to decoupled equations that, in the case of type-D metrics, are solvable by separation of variables.[8–10]

In a similar way, the O(2) extended supergravity field equations[11] give a consistent coupling of an O(2) doublet of spin-$\frac{3}{2}$ fields, electromagnetism, and gravity in such a way that when the spin-$\frac{3}{2}$ fields vanish, one recovers the Einstein–Maxwell equations. Therefore, the O(2) extended supergravity field equations, when linearized with respect to the spin-$\frac{3}{2}$ fields about a solution with vanishing spin-$\frac{3}{2}$ fields, give consistent equations for an O(2) doublet of spin-$\frac{3}{2}$ test massless fields on a solution of the Einstein–Maxwell equations.[12] In this approximation the back reaction of the spin-$\frac{3}{2}$ fields on the background solution is neglected and the supersymmetry transformations affect only the spin-$\frac{3}{2}$ fields. By considering the case of a type-D solution of the Einstein–Maxwell equations, where the principal null directions of the electromagnetic field are aligned with those of the conformal curvature, transforming to zero certain components constructed from the spin-$\frac{3}{2}$ field by means of a supersymmetry transformation (which in the case of the Kerr–Newman solution can be made only if the electric charge is different from the mass parameter), Aichelburg and Güven[12] showed that there exist two decoupled equations for the spin-$\frac{3}{2}$ perturbations, which in the Kerr–Newman background admit separable solutions.

In the present paper we extend Aichelburg and Güven's results[12] by showing that in a solution of the Einstein–Maxwell equations such that one (single or double) principal null direction of the electromagnetic field is geodetic and shear-free (and hence, the conformal curvature is algebraically special, but not necessarily type-D) there exists a decoupled equation for a gauge-invariant quantity constructed from the spin-$\frac{3}{2}$ perturbation. Moreover, the derivation given here does not invoke the supersymmetry transformations and applies without any further condition. We also show that in all type-D solutions of the Einstein–Maxwell equations such that the principal null directions of the electromagnetic field are aligned with those of the conformal curvature, the two decoupled equations are solvable by separation of variables and the separated functions satisfy differential relations of the Teukolsky–Starobinsky type found in the case of type-D vacuum metrics (see Ref. 7 and the references therein).

In Sec. II, following Ref. 12, the equations for the spin-$\frac{3}{2}$ perturbations and the effect of the supersymmetry transformations are written in spinor notation. In Sec. III, following Teukolsky's procedure,[1] we show that under suitable conditions there exists a component made out of the spin-$\frac{3}{2}$ perturbation, invariant under the supersymmetry transformations, which obeys a decoupled equation. In Sec. IV, making use of the explicit form of type-D solutions of the Einstein–Maxwell equations with a nonsingular aligned electromagnetic

 0022-2488/89/092114-06$02.50

field,[13,14] we show that in these backgrounds the decoupled equations are solvable by separation of variables and the separated functions are related through certain differential operators. Spinor formalism and Newman–Penrose notation are used throughout the paper (see, e.g., Ref. 15).

II. PRELIMINARIES

The equations for an O(2) doublet of test massless spin-$\frac{3}{2}$ fields on a curved background with an electromagnetic field can be expressed as[11,12]

$$\nabla_{AB'}\psi^{jA}{}_{CD'} + i\sqrt{2}\epsilon^{jk}\varphi^{A}{}_{C}\psi^{k}{}_{D'B'A} = \nabla_{CD'}\psi^{jA}{}_{AB'} \tag{1}$$

($j,k = 1,2$), where φ_{AB} is the electromagnetic spinor, ϵ^{jk} is the usual Levi-Civita symbol, and $\psi^{j}{}_{A'B'C} = \overline{\psi^{j}{}_{ABC'}}$. Equivalently, Eq. (1) can be written in the form (cf. Ref. 10)

$$H^{j}{}_{ABC} = H^{j}{}_{(ABC)} \tag{2a}$$

$$H^{j}{}_{AB'C'} = 0, \tag{2b}$$

where

$$H^{jA}{}_{BC} \equiv \nabla_{(B|S'|}\psi^{jA}{}_{C)}{}^{S'} - i\sqrt{2}\epsilon^{jk}\varphi^{A}{}_{(B}\psi^{k}{}_{|S'|}{}^{S'}{}_{C)} \tag{3}$$

and

$$H^{jA}{}_{B'C'} \equiv \nabla_{R(B'}\psi^{jAR}{}_{C')} - i\sqrt{2}\epsilon^{jk}\varphi^{A}{}_{R}\psi^{k}{}_{(B'C')}{}^{R}. \tag{4}$$

(The parentheses denote symmetrization on the indices enclosed and the indices between the vertical bars are excluded from the symmetrization.)

Equations (1) and (2) are invariant under the supersymmetry transformations

$$\psi^{j}{}_{ABC'} \to \psi^{j}{}_{ABC'} + \nabla_{BC'}\alpha^{j}{}_{A} - i\sqrt{2}\epsilon^{ik}\varphi_{AB}\alpha^{k}{}_{C'}, \tag{5}$$

where $\alpha^{j}{}_{A}$ is a pair of arbitrary spinor fields provided that the Einstein–Maxwell equations, without cosmological constant, are satisfied:

$$\Phi_{ABC'D'} = 2\varphi_{AB}\varphi_{C'D'}, \tag{6a}$$

$$\Lambda = 0, \tag{6b}$$

$$\nabla^{A}{}_{C'}\varphi_{AB} = 0. \tag{6c}$$

In fact, using the Ricci identities one finds that under transformation (5), the fields $H^{j}{}_{ABC}$ and $H^{j}{}_{AB'C'}$ transform according to

$$H^{j}{}_{ABC} \to H^{j}{}_{ABC} - \Psi_{ABC}{}^{D}\alpha^{j}{}_{D} + 2\Lambda\epsilon_{A(B}\alpha^{j}{}_{C)} + i\sqrt{2}\epsilon^{jk}\alpha^{k}{}_{S'}\nabla_{(B}{}^{S'}\phi_{C)A} \tag{7a}$$

and

$$H^{j}{}_{AB'C'} \to H^{j}{}_{AB'C'} - (\Phi_{A}{}^{D}{}_{B'C'} - 2\varphi_{A}{}^{D}\varphi_{B'C'})\alpha^{j}{}_{D} + i\sqrt{2}\epsilon^{jk}\alpha^{k}{}_{(B'}\nabla^{R}{}_{C')}\varphi_{RA}, \tag{7b}$$

which shows that the invariance of Eqs. (2) under the supersymmetry transformations requires the fulfillment of Eqs. (6).

The spinor fields $H^{j}{}_{ABC}$ are closer to the usual gauge-invariant description of the massless fields than the fields $\psi^{j}{}_{ABC'}$. Making use of Eqs. (1), (3), (6), and the Ricci identities one obtains that the fields $H^{j}{}_{ABC}$ satisfy

$$\nabla^{AR'}H^{j}{}_{ABC} = \Psi_{ABCD}\psi^{jADR'} + i\sqrt{2}\epsilon^{jk}\psi^{k}{}_{S'}{}^{R'A}\nabla_{B}{}^{S'}\varphi_{AC}. \tag{8}$$

The integrability conditions of Eq. (8), obtained by applying $\nabla^{B}{}_{R'}$ to both sides of this equation and using the Ricci identities; the Bianchi identities; and Eqs. (2b), (3), and (6) are satisfied identically. In contrast, the usual massless free-field equations for spin-$\frac{3}{2}$ lead to the Buchdahl–Plebański constraints (see, for example, Ref. 15). It is a remarkable fact that several cancellations depend crucially on the fulfillment of the background field equations (6).

III. DECOUPLED EQUATIONS

In this section we shall assume that (at least) one of the principal null directions of the background electromagnetic field is tangent to a geodetic and shear-free null congruence. This implies that the conformal curvature is algebraically special, with its multiple principal null direction also tangent to the geodetic and shear-free null congruence. If l_A denotes the geodetic and shear-free principal spinor of φ_{AB}, then using the facts that $l^{A}l^{B}\nabla_{AC'}l_{B} = 0$ and $l^{A}l^{B}l^{C}\Psi_{ABCD} = 0$, one finds that $l^{A}l^{B}l^{C}H^{j}{}_{ABC}$ are invariant under the transformation (7a).

Despite the complexity of the foregoing equations it turns out that with the only restriction of the existence of a spinor field l_A as above, there exists a decoupled equation for each gauge-invariant component $l^{A}l^{B}l^{C}H^{j}{}_{ABC}$ ($j = 1,2$). In a frame such that $\varphi_0 = 0$ and $\kappa = 0 = \sigma$, it follows that $\Psi_0 = 0 = \Psi_1$; from Eq. (8) one finds that the derivatives of the gauge-invariant component $H^{j}{}_{000}$ are given by

$$\begin{aligned}(\bar{\delta} - 3\alpha + \pi)H^{j}{}_{000} - (D - \epsilon - 3\rho)H^{j}{}_{001} &= \Psi_2\psi^{j}{}_{000'} + i2\sqrt{2}\epsilon^{jk}\varphi_1(\rho\psi^{k}{}_{1'0'0} - \tau\psi^{k}{}_{0'0'0}), \\ (\Delta - 3\gamma + \mu)H^{j}{}_{000} - (\delta - \beta - 3\tau)H^{j}{}_{001} &= \Psi_2\psi^{j}{}_{001'} + i2\sqrt{2}\epsilon^{jk}\varphi_1(\rho\psi^{k}{}_{1'1'0} - \tau\psi^{k}{}_{0'1'0}),\end{aligned} \tag{9}$$

where we have made use of the Maxwell equations

$$(D - 2\rho)\varphi_1 = 0, \quad (\delta - 2\tau)\varphi_1 = 0. \tag{10}$$

On the other hand, from Eq. (3) one finds that

$$H^{j}{}_{000} = (D - 2\epsilon + \bar{\epsilon} - \bar{\rho})\psi^{j}{}_{001'} - (\delta - 2\beta - \bar{\alpha} + \bar{\pi})\psi^{j}{}_{000'} \tag{11}$$

and from Eq. (1),

$$\begin{aligned}(\bar{\delta} - 2\alpha - \bar{\beta} + \pi)\psi^{j}{}_{000'} - (D - \bar{\epsilon} - \rho)\psi^{j}{}_{010'} + \rho\psi^{j}{}_{100'} &= i\sqrt{2}\epsilon^{jk}\varphi_1\psi^{k}{}_{0'0'0}, \\ (D + 2\epsilon - \bar{\epsilon} - \rho)\psi^{j}{}_{110'} - (\bar{\delta} - \bar{\beta} + \pi)\psi^{j}{}_{100'} - \pi\psi^{j}{}_{010'} + \lambda\psi^{j}{}_{000'} &= i\sqrt{2}\epsilon^{jk}(\varphi_1\psi^{k}{}_{0'0'1} - \varphi_2\psi^{k}{}_{0'0'0}).\end{aligned} \tag{12}$$

By applying $(\delta - 2\beta - \bar{\alpha} - 3\tau + \bar{\pi})$ to the first equation in (9) and $(D - 2\epsilon + \bar{\epsilon} - 3\rho - \bar{\rho})$ to the second and subtracting, the terms with $H^{j}{}_{001}$ cancel (cf. Ref. 1). Using Eqs. (6a), (10), and (11); the complex conjugates of Eqs. (12); and

$$(D - 3\rho)\Psi_2 = 2\rho\Phi_{11}, \quad (\delta - 3\tau)\Psi_2 = -2\tau\Phi_{11} + 2\rho\Phi_{12},$$

$$(D - \epsilon - \bar{\epsilon} - \rho)\rho = 0, \quad (\delta - \beta - \bar{\alpha} - \tau)\rho = -\tau\bar{\rho},$$

$$(D - \epsilon + \bar{\epsilon} - \rho)\tau = \rho\bar{\pi}, \quad (\delta - \beta + \bar{\alpha} - \tau)\tau = \rho\bar{\lambda},$$

which follow from the Bianchi and Ricci identities, respectively; one obtains

$$[(D-2\epsilon+\bar{\epsilon}-3\rho-\bar{\rho})(\Delta-3\gamma+\mu)$$
$$-(\delta-2\beta-\bar{\alpha}-3\tau+\bar{\pi})$$
$$\times(\delta-3\alpha+\pi)-\Psi_2]H^j{}_{000}=0. \quad (13)$$

Equation (13) is equivalent to the decoupled equation found in Ref. 12, which was obtained by assuming that the background solution of the Einstein–Maxwell equations is of type D and the principal null directions of the electromagnetic field are aligned with the double principal null directions of the curvature, with the further assumption that by means of the supersymmetry transformation (7a), the noninvariant components of $H^j{}_{ABC}$ may be transformed to zero (which is not always possible). The derivation of (13) does not make use of the supersymmetry transformations and applies for all the algebraically special solutions of the Einstein–Maxwell equations such that the multiple principal null direction of the curvature is geodetic, shear-free, and coincides with one of the principal null directions of the electromagnetic field. In particular, this shows that several conclusions obtained by Aichelburg and Güven[12] concerning the spin-$\frac{3}{2}$ perturbations of Kerr–Newman black holes are actually valid for all values of the parameters. It is a remarkable fact that in the form given above, Eq. (13) does not contain the electromagnetic field explicitly, but only through the metric. Furthermore, Eq. (13) has exactly the form of the decoupled equation found in the case of the spin-$\frac{3}{2}$ perturbations of an algebraically special vacuum spacetime[10] and hence, it corresponds to make $s=\frac{3}{2}$ in Teukolsky's equation.[1] As we shall show explicitly in Sec. IV, another remarkable feature of Eq. (13) is that it is separable in all the type-D solutions of the Einstein-Maxwell equations such that the principal null directions of the electromagnetic field coincide with those of the conformal curvature.

IV. SPIN-$\frac{3}{2}$ PERTURBATIONS OF TYPE-D ELECTROVAC SPACE-TIMES

In a type-D solution of the Einstein–Maxwell equations with an algebraically general aligned electromagnetic field one can choose the spin frame in such a way that the only nonvanishing components of Ψ_{ABCD} and φ_{AB} are Ψ_2 and φ_1, respectively; then from the Maxwell equations it follows that

$$\varphi_1=\tfrac{1}{2}(e+ig)\phi^2, \quad (14)$$

where e and g are real constants, interpretable as electric and magnetic charges, and ϕ is a function such that

$$\rho=D\ln\phi,\ \tau=\delta\ln\phi,\ \pi=-\bar{\delta}\ln\phi,\ \mu=-\Delta\ln\phi. \quad (15)$$

Since the cosmological constant is required to be equal to zero, it turns out that both principal null directions must be geodetic and shear-free,[16] i.e.,

$$\kappa=\sigma=\lambda=\nu=0. \quad (16)$$

The remaining spin coefficients can be expressed in the form

$$\epsilon=D\ln\zeta,\ \beta=\delta\ln\zeta,\ \alpha=-\bar{\delta}\ln\xi,\ \gamma=-\Delta\ln\xi, \quad (17)$$

where ζ and ξ are some functions. The integrability conditions on ζ and ξ are satisfied as a consequence of (16) and the fact that Ψ_2 is the only nonvanishing component of the Weyl spinor. Owing to the existence of two geodetic and shear-free principal spinors of the electromagnetic field, which are taken as the spin frame, the components $H^j{}_{000}$ and $H^j{}_{111}$ are invariant under the transformation (7a) and satisfy the decoupled equation (13) and

$$[(\Delta+2\gamma-\bar{\gamma}+3\mu+\bar{\mu})(D+3\epsilon-\rho)$$
$$-(\bar{\delta}+2\alpha+\bar{\beta}+3\pi-\bar{\tau})$$
$$\times(\delta+3\beta-\tau)-\Psi_2]H^j{}_{111}=0, \quad (18)$$

respectively.

All the type-D solutions of the Einstein–Maxwell equations with an aligned electromagnetic field possess (at least) two commuting Killing vectors. The (two-dimensional) orbits generated by this isometry group are called null or nonnull according to whether the metric induced on the orbits is singular or not.[13,14] With respect to a (local) coordinate system $\{x,y,u,v\}$ such that ∂_u and ∂_v are the two commuting Killing vectors, Eqs. (13) and (18) admit solutions with a dependence in the variables u and v of the form

$$e^{i(ku+lv)}, \quad (19)$$

where k and l are separation constants. In order to determine the dependence of the decoupled components of the spin-$\frac{3}{2}$ field on the coordinates x and y we shall consider the nonnull and null orbit solutions separately.

A. Non-null orbit solutions

The metric corresponding to a non-null orbit solution can be written in the form

$$ds^2=(\phi\bar{\phi})^{-1}\left\{\frac{Q}{(p_1q_2-p_2q_1)^2}(p_2\,du-p_1\,dv)^2-\frac{dy^2}{Q}\right.$$
$$\left.-\frac{P}{(p_1q_2-p_2q_1)^2}(q_2\,du-q_1\,dv)^2-\frac{dx^2}{P}\right\}, \quad (20)$$

where ϕ is defined by (14); $p_1=p_1(x)$ and $q_1=q_1(y)$ are polynomials of degree not greater than 2; p_2 and q_2 are constants; and $P=P(x)$ and $Q=Q(y)$ are polynomials of degree not greater than 4 that contain the parameters e and g, as well as other arbitrary parameters corresponding to mass, NUT parameter, acceleration, and angular momentum per unit mass. The tangent vectors

$$D=\partial_y+(1/Q)(q_1\,\partial_u+q_2\,\partial_v),$$
$$\Delta=-\tfrac{1}{2}\phi\bar{\phi}Q(\partial_y-(1/Q)(q_1\,\partial_u+q_2\,\partial_v)),$$
$$\delta=(P/2)^{1/2}\bar{\phi}(\partial_x+(i/P)(p_1\,\partial_u+p_2\,\partial_v)), \quad (21)$$
$$\bar{\delta}=(P/2)^{1/2}\phi(\partial_x-(i/P)(p_1\,\partial_u+p_2\,\partial_v))$$

form a null tetrad such that D and Δ are double principal null directions of the conformal curvature.

Acting on functions with a dependence of the form (19), the tetrad vectors can be replaced according to

$$D\to\mathscr{D}_0,\ \Delta\to-\tfrac{1}{2}\phi\bar{\phi}Q\mathscr{D}_0^{\dagger},$$
$$\delta\to(1/\sqrt{2})\bar{\phi}\mathscr{L}_0^{\dagger},\ \bar{\delta}\to(1/\sqrt{2})\phi\mathscr{L}_0, \quad (22)$$

where

$$\mathscr{D}_n\equiv\partial_y+iq/Q+nQ^{(1)}/Q=Q^{-n}\mathscr{D}_0Q^n,$$
$$\mathscr{D}_n^{\dagger}\equiv\partial_y-iq/Q+nQ^{(1)}/Q=Q^{-n}\mathscr{D}_0^{\dagger}Q^n,$$

TABLE I. Expression of the functions that determine the null tetrad. The metrics denoted as $G-S$ and $D-M$ correspond to the null orbit solutions. Here a, b, m, n, e, g, γ_0, and ϵ_0 are arbitrary constants.

Metric	$p(x)$	$q(y)$	$P(x)$	$Q(y)$
$gR-N$	$k+l$	k	$1-\epsilon_0x^2$	$\epsilon_0y^2-2my^3+(e^2+g^2)y^4$
g^*R-N	$k+l$	k	$-\epsilon_0x^2+2nx^3-(e^2+g^2)x^4$	$1+\epsilon_0y^2$
gC	$k+l$	k	$b-\epsilon_0x^2-2mx^3-(e^2+g^2)x^4$	$-b+\epsilon_0y^2-2my^3+(e^2+g^2)y^4$
$C\tilde{B}(+)$	$l+2akx$	$-k(a^2+y^2)$	$1-\epsilon_0x^2$	$e^2+g^2-2my+\epsilon_0(y^2-a^2)$
$C\tilde{B}(-)$	$k(a^2+x^2)$	$-l-2aky$	$-e^2-g^2+2nx-\epsilon_0(x^2-a^2)$	$1+\epsilon_0y^2$
$G-S$	$k(a^2+x^2)$	$-l-2aky$	$-e^2-g^2+2nx$	$\cdots$
CA	$l+kx^2$	$l-ky^2$	$b-g^2+2nx-\epsilon_0x^2$	$b+e^2-2my+\epsilon_0y^2$
$D-M$	$l+kx^2$	$l-ky^2$	$-e^2-g^2+2nx$	$\cdots$
$P-D$	$l+kx^2$	$l-ky^2$	$-g^2+\gamma_0+2nx-\epsilon_0x^2+2mx^3-(e^2+\gamma_0)x^4$	$e^2+\gamma_0-2my+\epsilon_0y^2-2ny^3+(g^2-\gamma_0)y^4$

$$\mathscr{L}_n\equiv\sqrt{P}\,(\partial_x+p/P+nP^{(1)}/2P)=P^{-n/2}\mathscr{L}_0P^{n/2},\tag{23}$$

$$\mathscr{L}_n^\dagger\equiv\sqrt{P}\,(\partial_x-p/P+nP^{(1)}/2P)=P^{-n/2}\mathscr{L}_0^\dagger P^{n/2},$$

$$p(x)\equiv p_1(x)k+p_2l,\quad q(y)\equiv q_1(y)k+q_2l,\tag{24}$$

and $f^{(k)}$ denotes the kth derivative of f with respect to its argument.

The polynomials p, q, P, and Q are listed in Table I, following Ref. 13 with some slight changes in notation (see, also, Ref. 14). For each specific metric, the spin coefficients can be computed using Eqs. (15) and (17) and the expressions listed in Table II. Some of the metrics given in Table I can be obtained from others also given there by setting some of the parameters equal to zero and making a coordinate transformation or by means of limiting transitions; nevertheless, these particular branches are included in order to simplify the application of the results derived here to the specific cases avoiding irrelevant parameters. The Kerr–Newman metric is a special case of the CA metric given in Table I, if one takes $b=a^2$, $g=0=n$, $\epsilon_0=1$. In terms of the Boyer–Lindquist coordinates, $y=r$, $x=-a\cos\theta$, $u=-t+a\varphi$, and $v=\varphi/a$; therefore, the separation constants k and l correspond to $k=-\omega$, $l=a(m+a\omega)$, where m is an integer.

Using Eqs. (15)–(17), (22), and (23), together with the expressions given in Tables I and II, a straightforward computation shows that Eqs. (13) and (18) admit separable solutions of the form given in Table III, with the one-variable functions $R_{\pm3/2}(y)$ and $S_{\pm3/2}(x)$ obeying the ordinary differential equations

$$[Q\mathscr{D}_{-1/2}\mathscr{D}_0^\dagger-2iq^{(1)}+Q^{(2)}/6]Q^{3/2}R_{+3/2}=AQ^{3/2}R_{+3/2},\tag{25}$$

$$[Q\mathscr{D}_{-1/2}^\dagger\mathscr{D}_0+2iq^{(1)}+Q^{(2)}/6]R_{-3/2}=AR_{-3/2},\tag{26}$$

and

$$[\mathscr{L}_{-1/2}^\dagger\mathscr{L}_{3/2}+2p^{(1)}+P^{(2)}/6]S_{+3/2}=-AS_{+3/2},\tag{27}$$

$$[\mathscr{L}_{-1/2}\mathscr{L}_{3/2}^\dagger-2p^{(1)}+P^{(2)}/6]S_{-3/2}=-AS_{-3/2},\tag{28}$$

where A is a separation constant. By means of the commutation relations

$$Q\mathscr{D}_{1-s}\mathscr{D}_0^\dagger=Q^{s+1}\mathscr{D}_{1+s}^\dagger\mathscr{D}_0Q^{-s}-2iq^{(1)}+sQ^{(2)},$$

$$\mathscr{L}_{1-s}\mathscr{L}_s^\dagger=\mathscr{L}_{1+s}^\dagger\mathscr{L}_{-s}-2p^{(1)}+sP^{(2)},\tag{29}$$

which follow from the definitions (23), Eqs. (25)–(28) can be summarized by the master equations

$$[Q\mathscr{D}_{1+s}^\dagger\mathscr{D}_0-(2s+1)iq^{(1)}+(s+1)(2s+1)Q^{(2)}/6]R_s=AR_s,\tag{30}$$

$$[\mathscr{L}_{1+s}\mathscr{L}_{-s}^\dagger+(2s+1)p^{(1)}+(s+1)(2s+1)P^{(2)}/6]S_s=-AS_s,\tag{31}$$

where s now takes the values $\frac{1}{2}$ or $-\frac{1}{2}$. Furthermore, Eqs. (23), (25), and (26) show that $Q^{3/2}R_{+3/2}$ and $R_{-3/2}$ satisfy complex-conjugate equations.

Equations (30) and (31) have exactly the same form as that found for perturbations of spins-$s=0$, $\frac{1}{2}$, 1, $\frac{3}{2}$, and 2 in the non-null orbit type-D vacuum metrics[7]; the only difference comes from the presence of the parameters e and g in

TABLE II. Expression of the only nonvanishing component of the conformal curvature and of the functions that determine the spin coefficients.

Metric	Ψ_2	ϕ	ζ	ξ
$gR-N$	$[-m+(e^2+g^2)\bar{\phi}]\phi^3$	y	$P^{1/4}y^{-1}$	$P^{1/4}Q^{1/2}$
g^*R-N	$[-in+(e^2+g^2)\bar{\phi}]\phi^3$	$-ix$	$P^{1/4}x^{-1}$	$P^{1/4}Q^{1/2}$
gC	$[-m+(e^2+g^2)(y-x)]\phi^3$	$x+y$	$P^{1/4}(x+y)^{-1}$	$P^{1/4}Q^{1/2}$
$C\tilde{B}(+)$	$[-(m+i\epsilon_0a)+(e^2+g^2)\bar{\phi}]\phi^3$	$(y+ia)^{-1}$	$P^{1/4}$	$P^{1/4}Q^{1/2}(y+ia)^{-1}$
$C\tilde{B}(-)$	$[-(in+\epsilon_0a)+(e^2+g^2)\bar{\phi}]\phi^3$	$(a+ix)^{-1}$	$P^{1/4}$	$P^{1/4}Q^{1/2}(a+ix)^{-1}$
$G-S$	$[-in+(e^2+g^2)\bar{\phi}]\phi^3$	$(a+ix)^{-1}$	$P^{1/4}$	$P^{1/4}(a+ix)^{-1}$
CA	$[-(m+in)+(e^2+g^2)\bar{\phi}]\phi^3$	$(y+ix)^{-1}$	$P^{1/4}$	$P^{1/4}Q^{1/2}(y+ix)^{-1}$
$D-M$	$[-in+(e^2+g^2)\bar{\phi}]\phi^3$	$(y+ix)^{-1}$	$P^{1/4}$	$P^{1/4}(y+ix)^{-1}$
$P-D$	$\{-(m+in)+(e^2+g^2)[(1+xy)/(y-ix)]\}\phi^3$	$(1-xy)/(y+ix)$	$P^{1/4}(1-xy)^{-1}$	$P^{1/4}Q^{1/2}(y+ix)^{-1}$

TABLE III. Decoupled components in terms of the separated functions.

Metric	H^{j}_{000}	H^{j}_{111}
$C\tilde{B}(+), C\tilde{B}(-)$	$e^{i(ku+lv)}R_{+3/2}S_{+3/2}$	$-(2)^{-3/2}\phi^3 e^{i(ku+lv)}R_{-3/2}S_{-3/2}$
$G-S, CA, D-M$		
$gR-N, g^*R-N, gC$	$\phi e^{i(ku+lv)}R_{+/32}S_{+3/2}$	$-(2)^{-3/2}\phi^4 e^{i(ku+lv)}R_{-3/2}S_{-3/2}$
$P-D$	$(1-xy)e^{i(ku+lv)}R_{+3/2}S_{+3/2}$	$-(2)^{-3/2}(1-xy)\phi^3 e^{i(ku+lv)}R_{-3/2}S_{-3/2}$

the fourth-order polynomials P and Q. Therefore, from Eqs. (25) and (26) it follows that[7]

$$Q^{3/2}(\mathscr{D}_0)^3 R_{-3/2} = CQ^{3/2}R_{+3/2},$$
$$Q^{3/2}(\mathscr{D}_0^\dagger)^3 Q^{3/2}R_{+3/2} = \bar{C}R_{-3/2}, \tag{32}$$

where C is a constant (taking into account that $Q^{3/2}R_{+3/2}$ and $R_{-3/2}$ satisfy complex-conjugate equations). By substituting the first equation (32) into the second, we obtain $Q^{3/2}(\mathscr{D}_0^\dagger)^3 Q^{3/2}(\mathscr{D}_0)^3 R_{-3/2} = |C|^2 R_{-3/2}$. Then by commuting the differential operators and using Eq. (26) we find[7]

$$|C|^2 = A^3 + \{-\tfrac{1}{6}[QQ^{(4)} - Q^{(1)}Q^{(3)} + \tfrac{1}{2}(Q^{(2)})^2] + 4(q^{(1)})^2 - 8qq^{(2)}\}A + \tfrac{4}{3}\{\tfrac{1}{2}q^2Q^{(4)} - qq^{(1)}Q^{(3)}$$
$$+ (qq^{(2)} + (q^{(1)})^2)Q^{(2)} - 3q^{(1)}q^{(2)}Q^{(1)} + 3(q^{(2)})^2Q\} + \tfrac{1}{36}\{-2QQ^{(2)}Q^{(4)} + \tfrac{3}{2}(Q^{(1)})^2Q^{(4)} + Q(Q^{(3)})^2$$
$$- Q^{(1)}Q^{(2)}Q^{(3)} + \tfrac{1}{3}(Q^{(2)})^3\}. \tag{33}$$

Similarly, the functions $S_{\pm 3/2}$ obey the relations

$$\mathscr{L}_{-1/2}\mathscr{L}_{1/2}\mathscr{L}_{3/2}S_{+3/2} = -BS_{-3/2}, \quad \mathscr{L}^\dagger_{-1/2}\mathscr{L}^\dagger_{1/2}\mathscr{L}^\dagger_{3/2}S_{-3/2} = BS_{+3/2}, \tag{34}$$

where by adjusting the phases of $R_{\pm 3/2}$ and $S_{\pm 3/2}$, B is a real constant. By substituting one of the equations (34) into the other, commuting the differential operators, and using Eq. (31) one obtains

$$B^2 = A^3 + \{-\tfrac{1}{6}[PP^{(4)} - P^{(1)}P^{(3)} + \tfrac{1}{2}(P^{(2)})^2] - 4(p^{(1)})^2 + 8pp^{(2)}\}A + \tfrac{4}{3}\{\tfrac{1}{2}p^2P^{(4)} - pp^{(1)}P^{(3)} + (pp^{(2)} + (p^{(1)})^2)P^{(2)}$$
$$- 3p^{(1)}p^{(2)}P^{(1)} + 3(p^{(2)})^2P\} + \tfrac{1}{36}\{2PP^{(2)}P^{(4)} - \tfrac{3}{2}(P^{(1)})^2P^{(4)} - P(P^{(3)})^2 + P^{(1)}P^{(2)}P^{(3)} - \tfrac{1}{3}(P^{(2)})^3\}. \tag{35}$$

The values of the constants B^2 and $|C|^2$ differ because of the presence of a background electromagnetic field; in fact, we find that

$$|C|^2 = B^2 + |4(e+ig)h|^2, \tag{36}$$

where h is given by

$$\phi^{-1}(\rho\Delta + \mu D - \tau\bar{\delta} - \pi\delta)e^{i(ku+lv)} = he^{i(ku+lv)} \tag{37}$$

in terms of the Killing vector field $\phi^{-1}(\rho\Delta + \mu D - \tau\bar{\delta} - \pi\delta)$. [Compare with Eq. (41) of Ref. 7.]

B. Null orbit solutions

In the case of a null orbit solution the metric can be written in the form

$$ds^2 = (\phi\bar{\phi})^{-1}\left\{\frac{2\,dy(p_2\,du - p_1\,dv)}{p_1q_2 - p_2q_1} - \frac{P}{(p_1q_2 - p_2q_1)^2}(q_2\,du - q_1\,dv)^2 - \frac{dx^2}{P}\right\}, \tag{38}$$

where $p_1 = p_1(x)$ and $q_1 = q_1(y)$ are polynomials of degree not greater that 2, p_2 and q_2 are constants, and $P = P(x)$ is a polynomial of degree not greater than 4. There exist only two different branches of the null orbit solutions, which are given in Table I (denoted as $G-S$ and $D-M$). A null tetrad for the metric (38) is determined by the vector fields

$$D = 2(q_1\,\partial_u + q_2\,\partial_v), \quad \Delta = -\tfrac{1}{2}\phi\bar{\phi}\,\partial_y,$$
$$\delta = (P/2)^{1/2}\bar{\phi}(\partial_x + (i/P)(p_1\,\partial_u + p_2\,\partial_v)), \tag{39}$$
$$\bar{\delta} = (P/2)^{1/2}\phi(\partial_x - (i/P)(p_1\,\partial_u + p_2\,\partial_v)).$$

Acting on functions with a dependence on u and v of the form (19), the tetrad vectors (39) can be replaced according to

$$D \to 2iq, \ \Delta \to -\tfrac{1}{2}\phi\bar{\phi}\,\partial_y, \ \delta \to (1/\sqrt{2})\bar{\phi}\mathscr{L}_0^\dagger, \ \bar{\delta} \to (1/\sqrt{2})\phi\mathscr{L}_0, \tag{40}$$

with $q = q(y)$ defined in (24) and $\mathscr{L}_0$ and $\mathscr{L}_0^\dagger$ given by Eq. (23). The spin coefficients and curvature can be obtained with the aid of Eqs. (15)–(17) and Table II.

The decoupled equations (13) and (18) admit separable solutions of the form given in Table III, where the functions $R_{\pm 3/2}(y)$ and $S_{\pm 3/2}(x)$ satisfy

$$[2iq\,\partial_y - (2s-1)iq^{(1)}]R_s = AR_s \tag{41}$$

and Eq. (31), respectively. In the present case the functions $R_{\pm 3/2}$ obey the following relations:

$$(2iq)^3 R_{-3/2} = ER_{+3/2}, \quad (\partial_y)^3 R_{+3/2} = FR_{-3/2}, \tag{42}$$

where E and F are constants. The value of the product EF can be obtained by substituting the first equation in (42) into the second and using Eq. (41). The result amounts to the rhs of Eq. (33), setting Q equal to zero, i.e.,

$$EF = A^3 + 4\{(q^{(1)})^2 - 2qq^{(2)}\}A. \tag{43}$$

Equation (36) also holds in this case if $|C|^2$ is replaced by EF.

V. CONCLUSIONS

The results of this paper show that many of the regularities found in the study of test massless fields in algebraically special vacuum space-times also apply to the case of spin-$\frac{3}{2}$ perturbations of certain solutions of the Einstein–Maxwell equations if one employs the linearized equations of the $O(2)$ extended supergravity. Equation (36), applicable to the case of type-D solutions with a non-null aligned electromagnetic field, is similar to that obtained for the gravitational perturbations of type-D vacuum metrics. Therefore, specifically for the Kerr–Newman solution, it is possible to have perturbations with one of the two components H^j_{000} or H^j_{111} equal to zero, which corresponds to having $|C|^2$ or EF equal to zero (cf. Ref. 17).

[1]S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).
[2]G. F. Torres del Castillo, Class. Quant. Gravit. **5**, 649 (1988).
[3]B. Carter, Commun. Math. Phys. **10**, 280 (1968).
[4]A. L. Dudley and J. D. Finley, III, J. Math. Phys. **20**, 311 (1979).
[5]N. Kamran and R. G. McLenaghan, in *Gravitation and Geometry: a Volume in Honor of I. Robinson*, edited by W. Rindler and A. Trautman (Bibliopolis, Naples, 1987).
[6]G. F. Torres del Castillo, J. Math. Phys. **29**, 971 (1988).
[7]G. F. Torres del Castillo, J. Math. Phys. **29**, 2078 (1988).
[8]R. Güven, Phys. Rev. D. **22**, 2327 (1980).
[9]N. Kamran, J. Math. Phys. **26**, 1740 (1985).
[10]G. F. Torres del Castillo, J. Math. Phys. **30**, 446 (1989).
[11]S. Ferrara and P. van Nieuwenhuizen, Phys. Rev. Lett. **37**, 1669 (1976).
[12]P. C. Aichelburg and R. Güven, Phys. Rev. D **24**, 2066 (1981).
[13]A. García Díaz, J. Math. Phys. **25**, 1951 (1984).
[14]R. Debever, N. Kamran, and R. G. McLenaghan, J. Math. Phys. **25**, 1955 (1984).
[15]R. Penrose and W. Rindler, *Spinors and Space-time* (Cambridge U.P., Cambridge, 1984), Vol. 1.
[16]A. García Díaz and J. F. Plebański, J. Math. Phys. **23**, 123 (1982).
[17]S. Chandrasekhar, Proc. R. Soc. London Ser. A **392**, 1 (1984).

An analogy of the charge distribution on Julia sets with the Brownian motion

Artur O. Lopes[a)]
Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742

(Received 29 December 1988; accepted for publication 26 April 1989)

A way to compute the entropy of an invariant measure of a hyperbolic rational map from the information given by a Ruelle–Perron–Frobenius operator of a generic Holder-continuous function will be shown. This result was motivated by an analogy of the Brownian motion with the dynamical system given by a rational map and the maximal measure. In the case the rational map is a polynomial, then the maximal measure is the charge distribution in the Julia set. The main theorem of this paper can be seen as a large deviation result. It is a kind of Donsker–Varadhan formula for dynamical systems.

I. INTRODUCTION

We will show an interesting analogy of the Brownian motion on $\mathbb{R}^n$ with the maximal measure of a hyperbolic rational map (the quotient of two polynomials) on the complex plane. In this context, the Ruelle–Perron–Frobenius operator plays the role of the semigroup (at time $t = 1$) associated with the infinitesimal generator of a diffusion process. We will show these results in Sec. IV of this paper. First in Sec. II and Sec. III we will explain carefully the concepts that we want to relate.

We believe it is worthwhile to present all the considerations that motivate the main theorem of this paper.

We refer the reader to Walters,[1] Mañe,[2] and Ruelle[3] for general results about ergodic theory and thermodynamic formalism, and we refer to Varadhan[4] for results about diffusions and large deviation properties of stochastic differential equations. Another source of references for the latter subject is Freidlin–Wentzell,[5] but here we will follow the more concise version of Varadhan.

All analogies presented here are based on some results presented in Ref. 6 about relations of the pressure, entropy, free energy, and large deviation. In Refs. 7–9, results related to the theorems in Ref. 6 are also obtained.

II. DIFFUSION AND BROWNIAN MOTION

Here we will follow the nice presentation of the main ideas about diffusion that appeared in Varadhan.[4]

There are many cases where solutions to problems are expressed as an integral over a space of functions. A simple example below is the (simplified) version of the Feynman–Kac formula that expresses the solution of the equation

$$\frac{\delta u}{\delta t} = \frac{1}{2}\Delta u + v(x)u\,, \quad u(0,x) = 1\,, \tag{2.1}$$

as the function space integral

$$u(t,x) = E_x\left\{\exp\int_0^t v(x(s))\,ds\right\}, \tag{2.2}$$

where E_x refers to the expectation with respect to Brownian motion on $\mathbb{R}^n$, starting from the point x in $\mathbb{R}^n$ at time $t = 0$.

Denote $E_x\{\exp\int_0^t v(x(s))ds\}$ by $\alpha(t)$, $t\in\mathbb{R}$ and consider the limit

$$\lambda = \lim_{t\to\infty}(1/t)\log\alpha(t)\,. \tag{2.3}$$

When $v(x)$ is periodic with period 1 in each variable, we can visualize, via the spectral theorem for $\frac{1}{2}\Delta + v$ on the n torus, that the above limit exists and is also the largest eigenvalue of $\frac{1}{2}\Delta + v$.

Now, by the variational principle, we have

$$\lambda = \lim_{t\to\infty}(1/t)\log\alpha(t)$$

$$= \sup_{\substack{\phi\in L_2(T^n)\\ \|\phi\|^2=1}}\left[\int_{T^n} v(x)\phi^2(x)\,dx - \frac{1}{2}\int_{T^n}|\nabla\phi|^2\,dx\right]. \tag{2.4}$$

This last expression can be interpreted in the following way: $\int_{T_n} v(x)\phi^2(x)dx$ is the potential term, that is, the term where the action of the external potential $v(x)$ appears.

The other term is a kind of inertial term. If there is no external potential v, that is $v = 0$, then we just notice the solution given by the regular Brownian motion.

Making analogy with classical mechanics, we can say the first term corresponds to potential energy and the second term to kinetic energy. Hamilton's principle of least action claims that motions of a mechanical system coincide with the extremal of a functional related to the difference of kinetic and potential energy. In the case that we are in a Riemannian manifold, and there is no potential energy, the trajectories are geodesics.

Now, let us return to diffusions. For a Markov process with infinitesimal generator L and domain D, consider the semigroup T_t corresponding to L; then the deviation function (see Ref. 4, Sec. 13) is

$$I(v) = \lim_{t\to\infty} -\frac{1}{t}\inf_{u\in B^+}\left\{\int\log\frac{T_t u(x)}{u(x)}\,dv(x)\right\}. \tag{2.5}$$

Here v is any probability measure on the state space X of the Markov process and B^+ is the space of continuous bounded positive functions.

We also have

$$I(v) = -\inf_{\substack{u\in D\\ \inf_x u(x)>0}}\int_X\left(\frac{Lu}{u}\right)(x)\,dv(x)\,. \tag{2.6}$$

Note that $L = \log T_1$.

III. THE MAXIMAL MEASURE AND THE PRESSURE

In this section we will explain the main reason to consider the pressure (sometimes called topological pressure) and the Ruelle–Perron–Frobenius operator.

[a)] Permanent address: Instituto Mat-UFRGS, Porto Alegre, Brazil.

 0022-2488/89/092120-05$02.50

Now we will follow the beautiful and simple motivation of the subject presented in Bowen.[10]

Consider a physical system with possible states $1,2,...,m$ and the energies of these states are $E_1,E_2,...,E_m$, respectively. Suppose our system is put in contact with a much larger "heat source," which is at temperature T. Energy is thereby allowed to pass between the original system and the heat source, and the temperature T of the source remains constant, as it is so much larger than our system. As the energy of our system is not considered fixed, the array of the states can occur. It has been known from statistical mechanics for a long time that the probability P_j that the state j occurs and is given by the Gibbs distribution:

$$P_j = \frac{e^{-BE_j}}{\sum_{i=1}^{m} e^{-BE_i}}, \quad j\in\{1,2,...,m\}, \tag{3.1}$$

where $B = 1/kT$ and k is a physical constant.

A mathematical formulation of the above consideration in a variational way can be obtained in the following way: consider

$$\tilde{F}(p_1,p_2,...,p_m) = \sum_{i=1}^{m} -p_i \log p_i - \sum_{i=1}^{m} p_i BE_i, \tag{3.2}$$

defined over the simplex in $\mathbb{R}^m$, given by

$$\left\{(p_1,p_2,...,p_m):\ p_i \geqslant 0,\ i\in\{1,2,...,m\} \text{ and } \sum_{i=1}^{m} p_i = 1\right\}.$$

Using Lagrange multipliers, it is easy to show that the maximum of $\tilde{F}$ in the simplex is obtained for

$$P_j = \frac{e^{-BE_j}}{\sum_{i=1}^{m} e^{-BE_i}}, \quad j\in\{1,2,...,m\},$$

in accordance with 3.1.

The quantity $H(p_1,p_2,...,p_m) = \sum_{i=1}^{m} -p_i \log p_i$ is called entropy of the distribution $(p_1,p_2,...,p_m)$. Denote $-\sum_{i=1}^{m} p_i E_i$ as the average energy $E(p_1,p_2,...,p_m)$.

Then we can say that the Gibbs distribution maximizes

$$H(p_1,p_2,...,p_m) - BE(p_1,p_2,...,p_m). \tag{3.3}$$

The expression $BE - H$ is called, in this context, free energy (in fact, there exist several different concepts in mathematics and physics also called free energy.)

Therefore we can say that nature minimizes free energy.

In the absence of the heat source, that is $E = 0$, nature maximizes entropy. In this case the Gibbs state is the most random probability, namely, $P_j = 1/m$, $j\in\{1,2,...,m\}$. Again, using analogy with classical mechanics, E plays the role of potential energy and H plays the role of kinetic energy.

Now, let us return to Gibbs measures. Generalizing the above considerations, Ruelle proposed the above model: consider the one-dimensional lattice $\mathbb{Z}$. Here one has for each integer a physical system with possible states $1,2,...,m$. A configuration of the system consists of assigning an $x_i\in\{1,2,...,m\}$ for each $i\in\mathbb{Z}$.

Thus a configuration is a point

$$\mathbf{x} = \{x_i\}_{i\in Z} \in \prod_{\mathbf{Z}} \{1,2,...,m\} = \Sigma_m.$$

Considering now the space Σ_m, the shift map

$$\sigma:\ \underset{\{x_i\}_{i\in Z}}{\Sigma_m} \to \underset{\{x_{i+1}\}_{i\in Z}}{\Sigma_m},$$

and $M(\sigma)$, the space of probabilities v, such that for any Borel set A

$$v(A) = v(\sigma^{-1}(A)),$$

we have the well-known Bernoulli shift model.

A continuous function $\phi: \Sigma_m \to \mathbb{R}$, in this setting, plays the role of the energy.

The problem here is to find a way to obtain the Gibbs distribution in the one-dimensional lattice in a similar way as how it was obtained before, in the beginning of Sec. III. Note that it is natural to consider just probabilities $p\in M(\sigma)$, because there is no natural reason to consider a certain distinguished point of the lattice as the origin in $\mathbb{Z}$.

Given a certain continuous function $\phi: \Sigma_m \to \mathbb{R}$ (as we said before will play the role of the energy), consider the following variational problem:

$$\sup_{p\in M(\sigma)} \left\{h(p) + \int \phi(z)\, dp(z)\right\}, \tag{3.4}$$

where $h(p)$ is the entropy of the probability p[1,2].

Denote such supremum by $P(\phi)$, the pressure associated with ϕ. It is natural to ask which properties have a probability p_ϕ that eventually attain such supremum value.

The above setting was proposed by Ruelle. In fact, he was able to find a certain ϕ, such that the above p_ϕ is exactly the Gibbs state for the one-dimensional lattice that with other procedures people in physics already knew a long time ago.[10]

Now, given the above setting, then following Ruelle and Bowen,[3,10–12] consider the below variational problem: given a rational map F of degree d in the complex plane and a Holder-continuous function ϕ on $\mathbb{C}$, consider

$$\sup_{p\in M(F)} \left\{h(p) + \int \phi(z)\, dp(z)\right\} = P(\phi), \tag{3.5}$$

where $h(p)$ is the entropy of the probability p and $M(F)$ is the set of probabilities, such that for any Borel set A,

$$p(A) = p(F^{-1}(A)),$$

$$p(\mathbb{C}) = 1.$$

The support of such measures in $M(F)$ will be contained always in the Julia set.[13,14]

When $\phi = 0$, there always exists a unique measure μ of maximal entropy.[15,16] We will call this measure the maximal measure. The entropy of such measure is $\log d$.

In the case where the rational map is a complex polynomial, the maximal measure is the charge distribution in the Julia set.[13,15] If the rational map is not a polynomial, the maximal measure is not the charge distribution in the Julia set.[14]

The results presented here are for the maximal measure μ. If one considers F a polynomial then, as we said before, our result is, in fact, for the charge distribution in the Julia

set. This last measure is also called the harmonic measure seen from ∞.

In any case, given ϕ, the value $P(\phi)$ will be called the pressure of the function ϕ.

There exists a very interesting way, developed by Ruelle, to obtain the above value $P(\phi)$.[10,11,3]

If a measure v satisfies $h(v) + \int \phi(z)dv(z) = P(\phi)$, that is, v attains the supremum of the above-mentioned variational problem, it is natural to call such measure a Gibbs state. In the case where F is hyperbolic and ϕ is Holder continuous, there always exists such a Gibbs state and it is unique.[10,3]

We will denote J the Julia set of F.

Consider $0 < \delta < 1$ and denote $\mathbf{F}$ the space of δ-Holder-continuous real-valued functions in J with the metric

$$\|g\| = \|g\|_0 + \sup_{x \neq y} \frac{|g(x) - g(y)|}{|x-y|^{\delta}},$$

where $\| \ \|_0$ is the usual supreme norm and $|\ |$ is the modulus.

Consider now the linear operator on $\mathbf{F}$, L_ψ: $\mathbf{F} \to \mathbf{F}$, given by

$$L_\psi(\Phi(z)) = \sum_{i=1}^{d} e^{\psi(x_i(z))}\Phi(x_i(z)), \tag{3.6}$$

where ψ is considered fixed, $\Phi \in \mathbf{F}$, and $x_i(z)$, $i \in (1,2,...,d)$ are the d solutions (counted with multiplicity) of $F(x) = z$.

In the literature this operator is called the Ruelle–Perron–Frobenius operator associated with $\psi \in \mathbf{F}$.[10,12,13]

The conjugate of L_ψ, denoted by $L_\psi{}^*$, acts on the space of signed measures, and is defined by taking a measure p to the $q = L_\psi{}^*(p)$, the unique one such that for any continuous function Φ,

$$\int \Phi(z)\, dq(z) = \int L_\psi(\Phi)(z)dp(z). \tag{3.7}$$

Theorem 1[11,12]: Let F be a hyperbolic rational map and $\psi: J \to \mathbb{R}$ Holder continuous. Then there exist $h: J \to \mathbb{R}$ ($h \in \mathbf{F}$), a probability v (not necessarily invariant) and $\lambda > 0$, such that

(1) $\int h(z)dv(z) = 1$;

(2) $L_\psi(h) = \lambda h$;

(3) $L_\psi^*(v) = \lambda v$;

(4) $\left\| \left| \lambda^{-n} L_\psi^n(\Phi) - h \int \Phi(z)\, dv(z) \right| \right\|_0 \to 0$,

$\forall \Phi \in \mathbf{F}$;

(5) h is the unique positive eigenfunction of L_ψ (up to multiplication by scalars);

(6) The probability $u = hv \in M(F)$ satisfies $h(u) + \int \psi(z)du(z) = P(\psi)$ and is the unique solution of the variational problem (3.5). Therefore u is the Gibbs state for ψ;

(7) $P(\psi) = \log \lambda$;

(8) λ is the largest eigenvalue of L_ψ.

The above theorem is proved in Refs. 11 and 12.

Given a probability $p \in M(F)$, consider the limit

$$\lim_{r \to 0} \frac{p(F(B(z,r)))}{p(B(z,r))}.$$

The above limit exists by the Radon–Nykodin theorem for z,p-almost everywhere.

Let us call

$$J(z) = \lim_{r \to 0} \frac{p(F(B(z,r)))}{p(B(z,r))}, \tag{3.8}$$

for z,p-almost everywhere, point $z \in J$, the Jacobian of the measure p. In fact, $h(p) = -\int \log J(z)dp(z)$ (Ref. 17).

In Ref. 6 it is shown that the set of probabilities in $M(F)$, such that the Jacobian is Holder continuous and never zero, is dense in $M(F)$.

Note that $J(z) \leqslant 1$ for $z \in J$, and also $\Sigma_{i=1}^{d} J(x_i(z)) = 1$ if $p \in M(F)$.

From this fact, it follows easily from Theorem 1 that, if p has Jacobian Holder continuous, then

$$P(\log J) = h(p) + \int \log J(z)dp(z) = 0. \tag{3.9}$$

Using the notation of Theorem 1, we also have $h(z) = 1 \forall z \in J$, $\lambda = 1$, and $u = v = p$.

We will use these results later in this paper.

The Gibbs measure for $\log J$ is sometimes referred to as a "g measure."

Theorem 2[6]: Let F be a hyperbolic rational map and ψ; then

$$P(\psi) = \lim_{n \to \infty} n^{-1} \log \int \exp\left(\sum_{j=0}^{n-1} \psi(F^j(z))\right) d\mu(z) + \log d, \tag{3.10}$$

where μ is the maximal measure.

Denote $\delta(z)$, the Dirac measure, with mass one in the point $z \in J$.

In Ref. 6 the above result was used to prove the following theorem.

Theorem 3[6]: Let F be a hyperbolic rational map of degree d and μ the maximal measure, then for any open convex set G of I continuity[6] G contained in the set of probabilities with support in the Julia set, $G \cap M(F) \neq \phi$, we have the limit

$$\lim_{n \to \infty} \frac{1}{n} \log \mu\left\{ z \in J \,\middle|\, \frac{1}{n} \sum_{j=0}^{n-1} \delta(F^j(z)) \in G \right\} \tag{3.11}$$

exists and is equal to

$$-\inf_{v \in G \cap M(F)} \{\log d - h(v)\}. \tag{3.12}$$

Therefore $\log d - h(v)$ is a deviation function for the process (F,μ). In fact, this deviation function is the Legendre transform of the pressure minus $\log d$.[6]

IV. BROWNIAN MOTION AND CHARGE DISTRIBUTION: AN ANALOGY

Suppose F is a polynomial. Therefore all considerations made before can be applied for the charge distribution in the Julia set because, in this case, this measure is equal to μ, the maximal measure.

We have that for $\psi \in \mathbf{F}$ and v a continuous function, the

expression $\Sigma_{j=0}^{n-1}\psi(F^j(z))$ can be considered a discrete time analogous version of $\int_0^t v(x(s))ds$. In this way

$$P(\psi) - \log d = \lim_{n\to\infty} \frac{1}{n} \log\left(\int \exp\left(\sum_{j=0}^{n-1} \psi(F^j(z))\right) d\mu(z)\right) \tag{4.1}$$

is analogous to

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \log E_x\left\{\exp \int_0^t v(x(s))\, ds\right\}, \tag{4.2}$$

where μ plays the role of the Brownian motion.

As we can see, respectively, in Sec. III and Sec. II, ψ and v play the role of a potential energy.

Now from Sec. II the semigroup T_t associated with the diffusion $\frac{1}{2}\Delta + v$ is such that T_1 has the largest eigenvalue e^{λ}. Observe that it also follows from Theorem 1 that $e^{P(\psi)}$ is the largest eigenvalue of L_ψ. Therefore $(1/d)L_\psi$ is analogous to T_1.

Finally, trying to find some kind of analogy with the last part of Sec. II [remember that L is analogous to $\log(d^{-1}L_\psi)$], then we have the following theorem.

Theorem 4: Consider F a hyperbolic rational map of degree d and $v \in M(F)$ a probability with Jacobian $J(z)$ Holder continuous that never vanishes, then

$$\log d - h(v) - \int \psi(z)dv(z) = -\inf_{u\in B^+}\left\{\int \log \frac{(d^{-1}L_\psi)u(z)}{u(z)}\, dv(z)\right\}. \tag{4.3}$$

We will make some remarks before the proof of this theorem.

Remark 1: The function I of the end of Sec. II is a large deviation function, as can be seen in Sec. 13.[4] By the other way, as we see in Sec. II, in the case $\psi = 0$, then in Theorem 3, the value $\log d - h(v)$ is a deviation function for the process (F,μ). Therefore the above theorem also represents an analogy with a diffusion process. Here we cannot consider a limit as t goes to zero:

$$I(v) = \lim_{t\to 0} -\frac{1}{t} \inf_{u\in B^+}\left\{\int \log \frac{T_t\, u(x)}{u(x)}\, dv(x)\right\}, \tag{4.4}$$

because the time n is discrete. Therefore here in Theorem 4, we consider $n = 1$.

It is usual in large deviation theory to suppose the deviation function is defined in a certain dense set with good properties. In the case of hyperbolic rational maps, this good set is the set of measures with Jacobian Holder continuous (see Ref. 6). Therefore the assumption about v in the setting of large deviation is a mild assumption.

Now we will prove the main theorem of this paper.

Proof of Theorem 4: Consider first $u(z) = e^{-\psi(z)}J(z)$; then

$$\begin{aligned}\log \frac{L_\psi(u(z))}{u(z)} &= \log \frac{\Sigma_{i=1}^d e^{\psi(x_i(z))}u(x_i(z))}{u(z)} \\ &= \log \frac{\Sigma_{i=1}^d (e^{\psi(x_i(z))}e^{-\psi(x_i(z))}J(x_i(z)))}{u(z)} \\ &= \log \sum_{i=1}^d J(x_i(z)) - \log e^{-\psi(z)}J(z) \\ &= \psi(z) - \log J(z).\end{aligned} \tag{4.5}$$

Therefore, for such u,

$$\begin{aligned}\int \log \frac{L_\psi u(z)}{u(z)}\, dv(z) &= \int \psi(z)dv(z) - \int \log J(z)dv(z) \\ &= \int \psi(z)dv(z) + h(v).\end{aligned}$$

Finally

$$\int \log \frac{(d^{-1}L_\psi)u(z)}{u(z)}\, dv(z) = \int \psi(z)dv(z) + h(v) - \log d. \tag{4.6}$$

Now we will show that for any positive continuous function u, we have

$$\int \log \frac{L_\psi u(z)}{u(z)}\, dv(z) \geqslant \int \psi(z)dv(z) + h(v). \tag{4.7}$$

We can , of course, suppose that instead of a general $u \in B^+$, we have $u(z)e^{-\psi(z)}J(z)$, because $e^{-\psi(z)}$ and $J(z)$ are nonzero by hypothesis.

Therefore

$$\begin{aligned}&L_\psi(u(z)) \\ &= \sum_{i=i}^d e^{\psi(x_i(z))}e^{-\psi(x_i(z))}J(x_i(z))u(x_i(z)) \\ &= \sum_{i=1}^d J(x_i(z))u(x_i(z)) = L_{\log J}u(z).\end{aligned} \tag{4.8}$$

In this case we have

$$\begin{aligned}\log \frac{L_\psi(u(z)e^{-\psi(z)}J(z))}{u(z)e^{-\psi(z)}J(z)} &= \log L_{\log J}u(z) \\ &\quad - \log u(z) + \psi(z) - \log J(z).\end{aligned}$$

From this, it follows that

$$\begin{aligned}\int \log \frac{L_\psi(u(z)e^{-\psi(z)}J(z))}{u(z)e^{-\psi(z)}J(z)}\, dv(z) &= \int \log L_{\log J}u(z)dv(z) - \int \log u(z)dv(z) \\ &\quad + \int \psi(z)dv(z) + h(v).\end{aligned}$$

Therefore all we have to prove is that

$$\int \log L_{\log J}u(z)dv(z) - \int \log u(z)dv(z) \geqslant 0. \tag{4.9}$$

Remember now that from (3) in Theorem 1, $L^*_{\log J}(v) = v$, therefore

$$\int \log u(z)dv(z) = \int L_{\log J} \log u(z)dv(z).$$

If we are able to show that

$$\log L_{\log J}u(z) \geqslant L_{\log J} \log u(z) \tag{4.10}$$

for any $z \in J$, then (4.9) follows.

This last inequality means

$$\log \sum_{i=1}^{d} J(x_i(z))u(x_i(z)) \geqslant \sum_{i=1}^{d} J(x_i(z))\log u(x_i(z)). \quad (4.11)$$

As $\Sigma_{i=1}^{d} J(x_i(z)) = 1$ for $z \in J$, the last inequality follows from the fact that log is a concave function and u is positive.

This is the end of the proof of Theorem 4.

V. CONCLUSION

Theorem 4 gives a way to compute the entropy of the measure v as an information obtained from the Ruelle–Perron–Frobenius operator of a Holder-continuous function ψ.

ACKNOWLEDGMENTS

We would like to thank F. Ledrapier for a suggestion in the proof of Theorem 4.

This research was partially supported by AFOSR.

[1]P. Walters, *An Introduction to Ergodic Theory* (Springer, Berlin, 1982).

[2]R. Mañe, *Ergodic Theory and Differentiable Dynamics* (Springer, Berlin, 1987).

[3]D. Ruelle, *Thermodynamic Formalism* (Addison–Wesley, Reading, MA, 1978).

[4]S. R. S. Varadhan, "Large deviations and applications," CBMS-NSF Regional Conf. Series in Appl. Math. (1984).

[5]M. I. Friedlin and A. D. Wentzell, *Random Perturbation of Dynamical Systems* (Springer, Berlin, 1984).

[6]A. Lopes, "Entropy and large deviation," University of Maryland preprint.

[7]M. Denker, "Large deviation and the pressure function," University of Göttingen preprint.

[8]S. Orey and S. Pelikan, "Large deviation principles for stationary processes," University of Minnesota preprint.

[9]S. Orey and S. Pelikan, "Deviations of trajectory averages and the defect in Pesin's formula for Anosov diffeomorphisms," University of Minnesota preprint.

[10]R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Difeomorphisma, Lecture Notes in Mathematics* (Springer, Berlin, 1975), Vol. 470.

[11]D. Ruelle, "Repellers for real analytic maps," Erg. Theory Dyn. Syst. **2**, 99 (1982).

[12]M. Pollicott, "A complex Ruelle–Perron–Frobenius operator and two counter examples," Erg. Theory Dyn. Syst. **4**, 135 (1984).

[13]H. Brolin, "Invariant sets under iteration of rational function," Ark. Mat. (Band G) **6**, 103 (1966).

[14]A. Lopes, "Equilibrium measures for rational maps," Erg. Theory Dyn. Syst. **6**, 393 (1986).

[15]A. Freire, A. Lopes, and R. Mañé, "An invariant measure for rational maps," Bol. Soc. Brasil. Mat. **14**, 45 (1983).

[16]V. Lubitsh, "Entropy properties of rational endomorphisms on the Riemann sphere," Erg. Theory Dyn. Syst. **3**, 351 (1983).

[17]R. Mañe, *On the Hausdorff Dimension of Invariant Probabilities of Rational Maps, Lecture Notes in Mathematics* (Springer, Berlin, 1988), Vol. 1331.

Fock space description of simple spinors

Paolo Budinich and Andrzej Trautman[a)]
International School for Advanced Studies, 34014 Trieste, Italy

(Received 2 February 1989; accepted for publication 29 March 1989)

Cartan's *simple*—often called *pure*—spinors corresponding to even-dimensional complex vector spaces are defined in terms of the associated maximal totally null planes. Their geometrical properties are derived and described using notions familiar to physicists: Dirac and Weyl spinors, gamma matrices, tensors formed bilinearly from pairs of spinors, and creation and annihilation operators of Fermi states. A new theorem characterizes a simple spinor ϕ by the properties of the vector $t_\psi B\gamma^\mu\phi$, where ψ is an arbitrary spinor and B is the matrix connecting the gamma matrices with their transposes. The Cartan constraint equations on the components of simple spinors are given a new, geometrically transparent derivation based on the action on simple spinors of a maximal Abelian subgroup of the group Spin.

I. INTRODUCTION

During the winter semester of 1935–1936, Elie Cartan gave a course of lectures at the Sorbonne on spinors. The notes of these lectures, taken by Mercier, were published in 1938.[1] In 1966 an English translation by Streater appeared.[2] In the lectures, Cartan presented a new approach to spinors, associated with a vector space of n dimensions, based on their intimate relation to totally *null* (Cartan used the word *isotropic*) subspaces of maximal dimension. In fact, Cartan showed that for $n > 6$ not all spinors correspond to maximal, totally null subspaces; he called *simple* those that do and described their properties. Characteristically of Cartan, the lectures combine a depth and originality of ideas with only rough outlines of the proofs. The importance of the lectures was recognized by another outstanding mathematician, Chevalley, whose book[3] connects Cartan's ideas with the approach to spinors presented by Brauer and Weyl[4] and based on Clifford algebras. Chevalley's emphasis is on algebra: Spinors are identified with elements of a minimal left ideal of the Clifford algebra and most of the theorems are proved without restrictions on the basic field (it may be of characteristic 2, for example). The generality of Chevalley's exposition made his book difficult to use by physicists; there is a readable account of parts of it by Benn and Tucker.[5]

Following Chevalley, most of the authors of publications on spinors in English replace the adjective *simple* by *pure*. However, here we shall use the original Cartan expression: Otherwise, we would have to accept that the Dirac spinor is impure (cf. Proposition 3).

Weyl spinors and the related null geometrical elements are known to play an important role in general relativity,[6] twistor theory,[7,8] and optical geometry.[9] Recent work on fundamental interactions and their unification makes essential use of geometries of more than four dimensions. For this reason, nontrivial simple spinors—which occur for $n > 6$—now have more chance of becoming relevant to physics than they had at the time of the appearance of Cartan's lectures. Further remarks on this subject can be found in our recent publications.[10,11]

In this paper we present a straightforward and explicit description of simple spinors and their principal properties. A new theorem characterizes a simple spinor ϕ in terms of the properties of the vector bilinear in ϕ and another spinor which need not be simple. Our approach is based on the observation—which can be traced back to Brauer and Weyl—that to a complex vector space of dimension $2m$ or $2m + 1$ there corresponds a spinor space S of complex dimension 2^m representable as the Fock space of m Fermi states. We show that every simple spinor can be used to define the "vacuum state" in S and then all eigenstates of the occupation number operators are also simple.

The study of simple spinors may be considered as a preliminary to the problem of "classification of spinors,"[12] which consists in finding the orbits of the Spin groups in spinor spaces, computing their stabilizers, and exhibiting the generators of the ring of invariants of the representation. Simple spinors correspond to the orbit of the lowest dimension. The classification problem is difficult and very little is known for $n > 14$. We hope our approach will also shed light on this problem.

We restrict ourselves here to complex vector spaces of even dimension $2m$. It is easy to extend our considerations to complex odd-dimensional spaces as well as to real spaces with a scalar product of signature (m,m) and $(m + 1,m)$. Other signatures require a subtler study because in those cases, the dimension of the maximal totally null subspaces is less than m.

II. PRELIMINARIES: NOTATION, CLIFFORD ALGEBRAS, AND SPINORS[13]

Let V be a complex vector space of dimension $2m$ ($m = 1,2,...$) with a scalar product g. The Clifford algebra $\mathrm{Cl}(g)$ admits a faithful and irreducible representation in a complex spinor space S of dimension 2^m. To alleviate the notation we identify V with its image in $\mathrm{Cl}(g)$. Moreover, since the representation $\mathrm{Cl}(g) \to \mathrm{End}\, S$ is an isomorphism (of algebras), we can also identify $\mathrm{Cl}(g)$ with End S. Therefore, the same letter u denotes a vector, an element of $\mathrm{Cl}(g)$, and an endomorphism of S; if $u, v \in V$, then

a) Permanent address: Instytut Fizyki Teoretycznej, Uniwersytet Warszawski, Hoza 69, 00-681 Warszawa, Poland.

 0022-2488/89/092125-07$02.50

$$uv + vu = 2g(u,v).$$

The vector space V admits a (generalized) orthonormal basis $(\gamma_1,...,\gamma_{2m})$ such that

$$\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 2g_{\mu\nu},$$

where

$$g_{\mu\mu} = (-1)^{\mu+1},$$

$$g_{\mu\nu} = 0 \quad \text{for} \quad \mu \neq \nu \quad (\mu,\nu = 1,...,2m).$$

We shall write $\gamma^\mu = \Sigma_\nu g_{\mu\nu}\gamma_\nu$, so that $\gamma^\mu\gamma_\nu + \gamma_\nu\gamma^\mu = 2\delta^\mu_\nu$.

From the orthonormal basis one can construct a null basis $(n_1,...,n_m,p_1,...,p_m)$ by putting

$$n_\alpha = \tfrac{1}{2}(\gamma_{2\alpha-1} - \gamma_{2\alpha}), \quad p_\alpha = \tfrac{1}{2}(\gamma_{2\alpha-1} + \gamma_{2\alpha}),$$

so that

$$n_\alpha n_\beta + n_\beta n_\alpha = 0, \quad p_\alpha p_\beta + p_\beta p_\alpha = 0 \tag{1}$$

and

$$n_\alpha p_\beta + p_\beta n_\alpha = \delta_{\alpha\beta}, \tag{2}$$

where $\alpha,\beta = 1,...,m$. The null vectors (n_α) span a maximal totally null (MTN) subspace N of V:

$$N = \text{span}\{n_1,...,n_m\}.$$

Similarly,

$$P = \text{span}\{p_1,...,p_m\}.$$

is also a MTN subspace, $N \cap P = \{0\}$, and there is a decomposition of V into a direct sum

$$V = N \oplus P.$$

Conversely, given a pair (N,P) of MTN subspaces of V such that $N \cap P = \{0\}$, one can find a basis $(n_1,..,n_m)$ of N and a basis $(p_1,...,p_m)$ of P such that Eqs. (1) and (2) hold. Every vector u admits a unique decomposition

$$u = n + p,$$

where $n \in N$ and $p \in P$. Writing

$$n = \sum_{\alpha=1}^{m} x_\alpha n_\alpha, \quad p = \sum_{\alpha=1}^{m} y_\alpha p_\alpha$$

one can express the fundamental quadratic form of V as

$$g(u,u) = x_1y_1 + \cdots + x_my_m.$$

The orthogonal group $O(g)$ acts transitively on the set of all MTN subspaces of V; this set has a natural structure of complex manifold of dimension $m(m-1)/2$.

Assume now that V has a preferred orientation and the basis (γ_μ) agrees with the orientation. The volume element

$$\Gamma = \gamma_1\gamma_2\cdots\gamma_{2m}$$

can be expressed in terms of the null basis as

$$\Gamma = [n_1,p_1][n_2,p_2]\cdots[n_m,p_m], \tag{3}$$

where square brackets denote commutators. Note, also, that Γ changes sign when the orientation of V is reversed. Since $\Gamma^2 = I$ (the unit automorphism of S), the eigenvalues of Γ are 1 and -1; the corresponding eigenvectors are *Weyl spinors* of positive and negative *helicity*, respectively. There is the decomposition

$$S = S_+ \oplus S_-,$$

where

$$S_\pm = \{\phi \in S \mid \Gamma\phi = \pm\phi\}.$$

The transposed endomorphisms (matrices) ${}^t\gamma_\mu$ define a representation of $\text{Cl}(g)$ in the space S^* dual to S. Since $\text{Cl}(g)$ is simple, there is an isomorphism B: $S \to S^*$ such that

$${}^t\gamma_\mu = B\gamma_\mu B^{-1}. \tag{4}$$

One shows that

$${}^tB = (-1)^{m(m-1)/2}B \tag{5}$$

and

$${}^t\Gamma B = (-1)^m B\Gamma. \tag{6}$$

If $\phi \in S$ is a "contravariant" and $\psi^* \in S^*$ a "covariant" spinor, then $\langle\psi^*,\phi\rangle$ is the evaluation ("contraction") of ψ^* on ϕ. The isomorphism of End S with $S \otimes S^*$ makes it possible to consider $\phi \otimes \psi^*$ as the endomorphism of S such that $(\phi \otimes \psi^*)(\psi) = \langle\psi^*,\psi\rangle\phi$ for every $\psi \in S$. Clearly,

$$\text{Tr}(\phi \otimes \psi^*) = \langle\psi^*,\phi\rangle, \quad \text{Tr}\, I = 2^m.$$

If $A \in \text{End}\, S$, then

$$A \circ (\phi \otimes \psi^*) = (A\phi) \otimes \psi^*. \tag{7}$$

The isomorphism B defines a bilinear map

$$S \times S \ni (\psi,\phi) \to \langle B\psi,\phi\rangle \in \mathbb{C},$$

which is invariant with respect to the action of the group[14] $\text{Pin}(g)$: If u is a unit vector, then

$$\langle Bu\psi,u\phi\rangle = \langle B\psi,\phi\rangle. \tag{8}$$

With every pair of spinors ψ, ϕ one can associate a sequence $B_k(\psi,\gamma)$, $k = 0,1,2,...,2m$ of multivectors over V: Their components with respect to an orthonormal basis (γ_μ) are given by

$$B_k^{\mu_1\cdots\mu_k}(\psi,\phi) = \langle B\psi,\gamma^{\mu_k}\cdots\gamma^{\mu_1}\phi\rangle, \tag{9}$$

with

$$1 \leqslant \mu_1 < \cdots < \mu_k \leqslant 2m. \tag{10}$$

With the understanding that the product of an empty sequence of the gammas is the unit automorphism I, Eq. (9) makes sense for $k = 0$ and $B_0(\psi,\phi) = \langle B\psi,\phi\rangle$. The set of all products $\gamma_{\mu_1}\cdots\gamma_{\mu_k}$, where $k = 0,1,...,2m$ and the indices satisfy (10), is a basis of End S. Therefore, there is a decomposition[15]

$$\phi \otimes B\psi = 2^{-m}\sum_k B_k(\psi,\phi), \tag{11}$$

where

$$B_k(\psi,\phi) = \sum_{(10)} B_k^{\mu_1\cdots\mu_k}\gamma_{\mu_1}\cdots\gamma_{\mu_k},$$

which is proved by noting that the trace of the product

$$\gamma^{\nu_l}\cdots\gamma^{\nu_1}\gamma_{\mu_1}\cdots\gamma_{\mu_k}, \quad \text{where} \quad \nu_1 < \cdots < \nu_l,$$

is 2^m for $k = \ell, \mu_1 = \nu_1,...,\mu_k = \nu_k$ and zero otherwise.

The symmetry properties (4) and (5) imply

$$B_k(\phi,\psi) = (-1)^{(k(k-1)+m(m-1))/2}B_k(\psi,\phi), \tag{12}$$

so that $B_k(\phi,\phi) = 0$ for $m \equiv 0,1$ and $k = 2,3$ or $m \equiv 2,3$ and $k = 0,1 \pmod 4$. Equation (6) implies

$$B_k(\Gamma\psi,\phi) = (-1)^{m-k}B_k(\psi,\Gamma\phi). \tag{13}$$

Therefore, if ψ and ϕ are Weyl spinors, then

$$B_k(\psi,\phi) = 0 \quad \text{if the helicities of } \psi \text{ and } \phi \text{ are } \begin{cases}\text{equal and } m-k \text{ is odd}\\ \text{opposite and } m-k \text{ is even}\end{cases} \tag{14}$$

In particular, if ϕ is a Weyl spinor, then

$$B_k(\phi,\phi) = 0 \quad \text{for} \quad m-k \equiv 1,2,3 (\text{mod } 4). \tag{15}$$

Since

$$\Gamma\gamma^1\cdots\gamma^k = (-1)^k\gamma^{k+1}\cdots\gamma^{2m},$$

there is a convenient way of defining *Hodge duality* of multivectors by means of the gammas. For every k-vector F with the components $F^{\mu_1\cdots\mu_k}$ with respect to the orthonormal basis (γ_μ), one defines its *dual* to be the $(2m-k)$-vector $^*F$ with the components given by

$$\sum {}^*F^{\nu_{k+1}\cdots\nu_{2m}}\gamma_{\nu_{k+1}}\cdots\gamma_{\nu_{2m}}$$

$$= \sum F^{\mu_1\cdots\mu_k}\Gamma\gamma_{\mu_1}\cdots\gamma_{\mu_k}, \tag{16}$$

where the sums are taken over all strictly increasing sequences of the indices. Since $\Gamma^2 = I$ one has $^{**}F = F$.

Replacing ϕ by $\Gamma\phi$ in Eq. (11) and using (7) and (16) one obtains

$$^*B_k(\psi,\phi) = B_{2m-k}(\psi,\Gamma\phi). \tag{17}$$

Therefore, if ϕ is a Weyl spinor $\Gamma\phi = \pm\phi$, then

$$^*B_k(\psi,\phi) = \pm B_{2m-k}(\psi,\phi) \tag{18}$$

and, in particular,

$$^*B_m(\psi,\phi) = \pm B_m(\psi,\phi). \tag{19}$$

The only essential component of $B_{2m}(\psi,\phi)$ is the pseudoscalar $\langle B\psi,\Gamma\phi\rangle$.

Let F be a k-vector and $u = u^\mu\gamma_\mu \in V$; then the *contraction* $u \lrcorner F$ of u with F is the $(k-1)$-vector with the components

$$(u \lrcorner F)^{\mu_2\cdots\mu_k} = \sum_{\mu\nu} g_{\mu\nu}u^\nu F^{\mu\mu_2\cdots\mu_k}$$

and the *exterior product* $u \wedge F$ is the $(k+1)$ vector obtained from the tensor product $u \otimes F$ by the alternating map ("antisymmetrization over all indices"). The isomorphism (of vector spaces) $\mathrm{Cl}(g) = \Lambda V$ leads to the following useful formula[16]:

$$uF = u \lrcorner F + u \wedge F, \tag{20}$$

where uF is the Clifford product of u and F.

Computing $u\phi \otimes B\psi$, where $u \in V$ and using Eqs. (7), (11), and (20) one obtains

$$B_k(\psi,u\phi) = u \lrcorner B_{k+1}(\psi,\phi) + u \wedge B_{k-1}(\psi,\phi) \tag{21}$$

for $k = 0, 1,\ldots,2m$; it is understood that B_{-1} and B_{2m+1} are zero.

III. DEFINITION OF SIMPLE SPINORS AND AN EXAMPLE

The vector space *associated* with a spinor $\phi \in S$,

$$M(\phi) = \{u \in V \mid u\phi = 0\}, \tag{22}$$

depends only on the direction of ϕ. For $\phi \neq 0$ this vector space is totally null: If $u, v \in M(\phi)$, then $g(u,v) = 0$.

Definition: A nonzero spinor is said to be simple if its associated totally null vector space is maximal.

In other words, if V is $2m$-dimensional, then simplicity of ϕ is equivalent to $\dim M(\phi) = m$. To see that simple spinors exist in every dimension, consider a MTN subspace N of V and a basis $(n_1,\ldots,n_m)$ of N. Since the representation of $\mathrm{Cl}(g)$ in S is faithful, there exists a spinor χ such that

$$\omega = n_1 n_2 \cdots n_m \chi \tag{23}$$

is nonzero; then $M(\omega) = N$ and ω is simple. On the other hand, not all spinors are simple, as may be seen from the following example, which is familiar to physicists.

Example: If V is four-dimensional, $m = 2$, then S is also:

$$^tB = -B, \quad B\Gamma = {}^t\Gamma B.$$

Let $\phi = \phi_+ + \phi_-$ be the decomposition of a Dirac spinor ϕ into its Weyl components, $\phi_\pm = \frac{1}{2}(I \pm \Gamma)\phi$.

For every $u \in V$, the condition $u\phi = 0$ is equivalent to $u\phi_+ = 0$ and $u\phi_- = 0$; this shows that

$$M(\phi) = M(\phi_+) \cap M(\phi_-).$$

To determine the spaces $M(\phi_\pm)$, consider the endomorphisms $\phi_\pm \otimes B\phi_\pm$. From (15) it follows that only the term with $k = 2$ is present in the decomposition (11):

$$\phi_\pm \otimes B\phi_\pm = F_\pm, \quad \text{where} \quad F_\pm = \tfrac{1}{4}B_2(\phi_\pm,\phi_\pm) \tag{24}$$

and (15) and (19) imply

$$\Gamma F_\pm = {}^*F_\pm = \pm F_\pm. \tag{25}$$

Similarly,

$$\phi_\pm \otimes B\phi_\mp = \tfrac{1}{2}(1 \pm \Gamma)k, \tag{26}$$

where k is a vector. Using (21) for $\psi = \phi = \phi_\pm$ one obtains that

$$u\phi_\pm = 0 \quad \text{is equivalent to} \quad u \lrcorner F_\pm = 0 \quad \text{and} \quad u \wedge F_\pm = 0. \tag{27}$$

In particular, computing $k\phi_\pm$ and using (26) and $\langle B\phi_\pm,\phi_\pm\rangle = 0$ one obtains

$$k\phi_\pm = 0. \tag{28}$$

Equation (28) implies that the two-forms F_+ and F_- are decomposable and have the null vector k as their common eigenvector. The spaces $M(\phi_+)$ and $M(\phi_-)$ coincide if and only if both ϕ_+ and ϕ_- are zero. If $\phi_+ \neq 0$ and $\phi_- \neq 0$, then the intersection

$$M(\phi_+) \cap M(\phi_-) = \mathbb{C}k \tag{29}$$

is one-dimensional and the Dirac spinor $\phi_+ + \phi_-$ is not simple. If $\phi_+ \neq 0$ and $\phi_- = 0$, the $M(\phi) = M(\phi_+)$ is two-dimensional and the spinor $\phi = \phi_+$ is simple: similarly, the same holds when the roles of ϕ_+ and ϕ_- are interchanged.

Parenthetically, we should mention that in Minkowski space, which can be defined by embedding $\mathbb{R}^4$ in $\mathbb{C}^4$, so that $(\gamma_1,\gamma_2,\gamma_3,\gamma_4)$ are replaced by $(\gamma_1,i\gamma_2,\gamma_3,\gamma_4)$, the null vector k can be made real by taking the charge conjugate of ϕ_+ for ϕ_-. With the direction of such a vector there is an associated

two-dimensional space of "null electromagnetic fields"[8,9,17]. Remembering that in signature (3,1), the square of the dual is $-id$, one can write

$$F_{\pm} = f \mp i^* f,$$

where the null two-form f, representing the electromagnetic field, is now real. The totally null subspaces $M(\phi_+)$ and $M(\phi_-)$ are complex conjugate to each other and $\mathbb{R}k$ is their only real direction. The spinorial representation of null electromagnetic waves, contained in Eq. (24), gives rise to various extensions and generalizations to gravitation and other fields. They play a major role in the Penrose program and the Newman–Penrose formalism for the treatment of algebraically special metrics,[8] as well as in the treatment of null strings.[18] There is also a related spinorial form of the Enneper–Weierstrass formula for solutions of the equation for minimal surfaces and its extension to strings.[19]

IV. A FOCK REPRESENTATION BASIS IN THE SPACE OF SPINORS

The vectors (endomorphisms of S) n_α and p_α ($\alpha = 1,...,m$) defined in Sec. II fulfill the anticommutation relations (1) and (2), which are identical to those satisfied by the annihilation and creation operators of states subject to Fermi statistics. This observation—made previously by Brauer and Weyl[4]— can be used to construct a convenient basis in S.

Given the decomposition of V into a direct sum of MTN subspaces N and P and the corresponding null basis $(n_1,...,n_m,p_1,...,p_m)$ one can find a nonzero spinor ω of the form (23) and interpret it as the "vacuum state": It is annihilated by all the "operators" $n_1,...,n_m$. By acting on ω with products of the "creation operators" p_α corresponding to all sequences (μ_i) subject to (10) one obtains a collection of 2^m spinors:

$$\begin{aligned} &\omega; \quad p_1\omega,...,p_m\omega; \\ &p_1p_2\omega,...,p_{m-1}p_m\omega; \\ &\cdots; \quad p_1\cdots p_m\omega. \end{aligned} \tag{30}$$

The collection (30) is linearly independent, as may be easily shown using (1) and (2) only. All the spinors occurring in the sequence (30) are simple, e.g.,

$$M(p_{k+1}p_{k+2}\cdots p_m\omega) = \operatorname{span}\{n_1,...,n_k,p_{k+1},p_{k+2},...,p_m\}$$

is totally null and of dimension m. The basis (30) can be used to show that the direction of a simple spinor is determined by the associated MTN subspace. Indeed, let $\omega' \in S$ be such that $M(\omega') = M(\omega) = N$ and let

$$\omega' = \xi_0\omega + \xi_1p_1\omega + \cdots + \xi_{2\cdots m}p_2\cdots p_m\omega + \xi_{12\cdots m}p_1p_2\cdots p_m\omega.$$

Multiplying both sides of the above equation by $n_1n_2\cdots n_m$ yields $\xi_{12\cdots m} = 0$; multiplying it next by $n_2\cdots n_m$ leads to $\xi_{2\cdots m} = 0$, etc. This proves the following proposition.

Proposition 1: There is a natural, one-to-one correspondence between the set of all MTN subspaces of V and the set of directions of all simple spinors.

Every simple spinor ω can be taken to be the first, "vacuum" or "standard spinor"[2] of the sequence (30). The spinor ω determines only the subspace $N = M(\omega)$: There is still considerable freedom in choosing the complementary MTN subspace P and the null basis adapted to the decomposition $V = N \oplus P$.

Proposition 2: Let ω and ϕ be linearly independent simple spinors. There is a basis in S of the form (30) such that ϕ is one of the basis vectors, other than ω.

Indeed, it follows from Proposition 1 that $N = M(\omega) \neq M(\phi) = N'$. Let

$$k = \dim N \cap N';$$

then $0 \leqslant k < m$ and there exists a basis $(n_1,...,n_m)$ of N adapted to the subspace $N \cap N'$ in the sense that

$$N \cap N' = \operatorname{span}\{n_1,...,n_k\}.$$

Similarly, there is basis $(n'_1,...,n'_m)$ of N' such that $n'_1 = n_1,...,n'_k = n_k$. The matrix of solar products

$$g(n_a,n'_b), \quad \text{where} \quad a,b = k+1,...,m$$

is nonsingular: Otherwise there would be a vector $\lambda_{k+1}n'_{k+1} + \cdots + \lambda_m n'_m$ different from zero, null and orthogonal to N, but not in N. This would contradict the maximality of N among totally null subspaces. Therefore, by a linear transformation of the vectors n'_a ($a = k+1,...,m$) one can achieve $g(n_a,n'_b) = \frac{1}{2}\delta_{ab}$. One can now put

$$p_{k+1} = n'_{k+1},...,p_m = n'_m$$

and complete the sequence $(p_{k+1},...,p_m)$ to a basis $(p_1,...,p_m)$ of a MTN subspace P complementary to N and such that relations (1) and (2) hold. Since now

$$M(\phi) = \operatorname{span}\{n_1,...,n_k,p_{k+1},...,p_m\},$$

it is clear that after rescaling,

$$\phi = p_{k+1}\cdots p_m\omega. \tag{31}$$

Proposition 3: Simple spinors are Weyl.

The proof is straightforward: Since $[n_1,p_1]n_1 = n_1$ and $[n_1,p_1]p_1 = -p_1$, etc., it is clear that with Γ given by (3) and ϕ by (31), one has

$$\Gamma\phi = (-1)^{m-k}\phi.$$

Choosing the orientation in V so that Γ—rather than $-\Gamma$—is of the form (3) is equivalent to assigning a positive helicity to ω. Since parallel spinors have equal helicity, this notion is transferred, via Proposition 1, to MTN subspaces of V. We should also mention that if $n > 6$ there are Weyl spinors that are not simple. For example, for $n = 8$, the spinor $\phi = (1 + p_1p_2p_3p_4)\,\omega$ is not simple because $u\phi = 0$ implies $u = 0$.

If u is a unit vector and ϕ is a simple spinor, then $u\phi$ is also simple and of opposite helicity to ϕ. Indeed,

$$M(u\phi) = uM(\phi)u, \quad \Gamma u\phi = -u\Gamma\phi.$$

Remembering that the $\operatorname{Pin}(g)$ group consists of products of unit vectors and $\operatorname{Spin}(g)$ is its subgroup, consisting of products of even sequences of such vectors, one arrives at once at the following proposition.

Proposition 4: The group $\operatorname{Pin}(g)$ acts transitively on the set of directions of all simple spinors. The group $\operatorname{Spin}(g)$ acts

transitively on each of the two sets of directions of simple spinors of equal helicity.

Indeed, Eq. (31) can be written as

$$\phi = (n_{k+1} + p_{k+1}) \cdots (n_m + p_m)\omega,$$

where each factor $n_\alpha + p_\alpha$ is a unit vector; their product belongs to Spin(g) whenever $m - k$ is even, i.e., whenever ω and ϕ are of equal helicity. Otherwise, the product belongs to Pin(g), but not to Spin(g), and the spinors ω and ϕ are of opposite helicity.

The helicities of the spinors ω and ϕ related by (31) are equal or opposite depending on whether $m - k$ is even or odd. Therefore, the dimension k of $M(\omega) \cap M(\phi)$ is even if and only if m is even and the helicities are equal or m is odd and the helicities are opposite. Thus, for example, the complementary MTN subspaces N and P are of equal helicity if and only if m is even. One also proves the following proposition.

Proposition 5 (Ref. 3, Proposition III. 1.12): If ω and ϕ are linearly independent simple spinors, then $\omega + \phi$ is simple if and only if

$$\dim M(\omega) \cap M(\phi) = m - 2;$$

then

$$M(\omega) \cap M(\phi + \omega) = M(\omega) \cap M(\phi).$$

The "if" part of the proposition is immediate: Taking ϕ to be of the form (31) with $k = m - 2$ one has

$$M(\phi + \omega) = \text{span}\{n_1,...,n_{m-2}, n_{m-1} + p_m, n_m - p_{m-1}\}.$$

The *Lie algebra* spin(g) of the group Spin(g) can be identified[20] with the subspace $[V,V]$ of Cl(g) spanned by all the commutators $[u,v]$ where $u, v \in V$. Every MTN subspace P defines the subalgebra

$$[V,P] = \{A \in \text{spin}(g) \mid A\psi = \lambda\psi, \lambda \in \mathbb{C}\}, \tag{32}$$

where ψ is a simple spinor such that $P = M(\psi)$. If (p_α) is a basis in P, then

$$A \in [V,P] \leftrightarrow A = \sum_{\alpha=1}^{m} [v_\alpha, p_\alpha],$$

where $v_\alpha \in V$. The commutator subalgebra of $[V,P]$ is the Abelian Lie algebra $[P,P]$ consisting of all those elements of spin(g) that annihilate ψ. If

$$A = \sum_{\alpha<\beta} A_{\alpha\beta} p_\alpha p_\beta \in [P,P],$$

where $A_{\alpha\beta} \in \mathbb{C}$, then the element

$$a = \exp A = \prod_{\alpha<\beta} (1 + A_{\alpha\beta} p_\alpha p_\beta) \in \text{Spin}(g) \tag{33}$$

leaves invariant all elements of P: If $p \in P$, then $apa^{-1} = p$. If N is a MTN subspace complementary to P, then the Abelian subalgebra $[P,P]$ is complementary, as a vector subspace, to $[V,N]$ in spin(g). The subgroup of Spin(g) corresponding to the subalgebra $[V,N]$ leaves the direction of ω invariant; the subgroup corresponding to $[P,P]$ "moves" ω, but its action is not transitive on the set of directions of simple spinors of equal helicity. Indeed, if ϕ and ϕ' are two such spinors and there is an element (33) such that $\phi' = \text{const}\, a\phi$, then $M(\phi') = aM(\phi)a^{-1}$. Since $p \in P$ implies $p = apa^{-1}$ one obtains

$$M(\phi') \cap P = M(\phi) \cap P$$

as a necessary condition for the existence of a. This condition is also sufficient, as asserted in the following proposition.

Proposition 6: Let N and P be two complementary MTN subspaces of V and ϕ a simple spinor such that $M(\phi) \cap P$ is k-dimensional. There then exists an element a of the form (33) such that

$$\phi = \lambda a p_1 \cdots p_k \omega, \tag{34}$$

where $\lambda \in \mathbb{C}$, $\lambda \neq 0$, $M(\omega) = N$, and the vectors $(p_1,...,p_k)$ constitute a basis of $M(\phi) \cap P$.

Proof: Let $(p_1,...,p_m)$ be a basis of P adapted to $M(\phi) \cap P$, i.e., such that

$$M(\phi) \cap P = \text{span}\{p_1,...,p_k\}.$$

Let $(n_1,...,n_m)$ be a basis of N such that Eqs. (2) hold and consider the MTN subspaces

$$N_k = \text{span}\{p_1,...,p_k, n_{k+1},...,n_m\},$$

$$P_k = \text{span}\{n_1,...,n_k, p_{k+1},...,p_m\}.$$

Clearly, $N_k \cap P_k = \{0\}$, but, also, $M(\phi) \cap P_k = \{0\}$ because if $u \in M(\phi)$, then $g(u,p_\alpha) = 0$ for $\alpha = 1,...,k$. Therefore, if $u \in M(\phi) \cap P_k$, then $u \in \text{span}\{p_{k+1},...,p_m\} \subset P$, so that $u \in M(\phi) \cap P$, but $M(\phi) \cap P \cap P_k = \{0\}$. Complete now the basis $(p_1,...,p_k)$ of $M(\phi) \cap P$ to a basis $(p_1,...,p_k, n'_{k+1},...,n'_m)$ of M(ϕ), so that

$$n_\alpha n'_\kappa + n'_\kappa n_\alpha = 0, \quad \text{for} \quad \alpha = 1,...,k, \quad \kappa = k+1,...,m; \tag{35}$$

$$n'_\kappa p_\lambda + p_\lambda n'_\kappa = \delta_{\kappa\lambda}, \quad \text{for } \kappa,\lambda = k+1,...,m. \tag{36}$$

Writing $n'_\kappa = n_\kappa - v_\kappa$; computing the scalar products $g(v_\kappa,p)$, $p \in P$ and $g(v_\kappa,n_\alpha)$, $\alpha = 1,...,k$; and using Eqs. (2), (35), and (36) one finds that $v_\kappa \in \text{span}\{p_{k+1},...,p_m\}$, i.e., there is a matrix $(A_{\kappa\lambda})$ such that

$$v_\kappa = \sum_\lambda A_{\kappa\lambda} p_\lambda. \tag{37}$$

The vectors n'_κ belong to an MTN subspace, $g(n'_\kappa, n'_\lambda) = 0$. Therefore,

$$g(n_\kappa, v_\lambda) + g(v_\kappa, n_\lambda) = 0$$

and the matrix $(A_{\kappa\lambda})$ is antisymmetric. Let

$$A = \sum_{k+1 \leq \kappa < \lambda \leq m} A_{\kappa\lambda} p_\kappa p_\lambda; \tag{38}$$

then

$$n_\kappa A - A n_\kappa = \sum_\lambda A_{\kappa\lambda} p_\lambda.$$

Therefore, if a is given by (3) with A determined from (37) and (38), then

$$a p_\alpha a^{-1} = p_\alpha \, (\alpha = 1,...,k),$$

$$a n_\kappa a^{-1} = n'_\kappa \, (\kappa = k+1,...,m).$$

In other words, the element a of the group Spin(g) transforms N_k into $M(\phi)$ preserving $N_k \cap M(\phi) = P \cap M(\phi)$. Since

$$N_k = M(p_1 \cdots p_k \omega),$$

the simple spinor ϕ is proportional to $a p_1 \cdots p_k \omega$, as claimed.

Remark 1: The element (38) of Cl(g) is nilpotent: There exists an integer l such that $2l \leqslant m-k+2$ and $A^l = 0$. In particular, if $k = m-2$, then

$$A^2 = 0.$$

In this case A is proportional to $p_{m-1}p_m$, $\exp tA = 1 + tA$ and the straight line in S,

$$t \to (1+tA)p_1 \cdots p_{m-2}\omega, \quad 0 \leqslant t \leqslant 1,$$

connects the simple spinor $\phi = p_1 \cdots p_{m-2}\,\omega$ with the simple spinor $\phi = \psi$, where $\psi \sim p_1 \cdots p_m \omega$. This sheds some light on the property of simple spinors described in Proposition 5.

V. MULTIVECTORS ASSOCIATED WITH SIMPLE SPINORS

The decomposition formula (11) associates with a pair of spinors a sequence of multivectors; they provide useful information, often with a clear geometrical interpretation. This is especially so when one of the spinors—or both—are simple.

For the sequel we need the following useful result.

Lemma (Ref. 3, Proposition III. 2.4): If ω and ψ are simple spinors, then

$$M(\omega) \cap M(\psi) \neq \{0\} \leftrightarrow B_0(\psi,\omega) = 0. \tag{39}$$

Indeed, if $u \in M(\omega) \cap M(\psi)$ and $u \neq 0$, then there is a vector $v \in V$ such that $uv + vu = 1$ and since $u\omega = u\psi = 0$, one obtains

$$B_0(\psi,\omega) = \langle B\psi, uv\omega\rangle = \langle {}^t uB\psi, v\omega\rangle = \langle Bu\psi, v\omega\rangle = 0.$$

Conversely, if $M(\omega) \cap M(\psi) = \{0\}$, then one can take $N = M(\omega)$, $P = M(\psi)$ and construct a null basis $(n_1,...,n_m, p_1,...,p_m)$ of V and the "Fock basis" (30) of S. If ϕ is any spinor from the sequence (30) other than $\psi = p_1 \cdots p_m \omega$, then $M(\omega) \cap M(\phi) \neq \{0\}$ and thus $B_0(\phi,\omega) = 0$. Since B is an isomorphism, the form B_0 is nondegenerate and therefore, $B_0(\psi,\omega) \neq 0$; this completes the proof of the lemma.

We may now prove the following proposition.

Proposition 7: A necessary and sufficient condition for a spinor $\omega \neq 0$ to be a simple and have N as its associated MTN subspace is that the vector $B_1(\phi,\omega)$ belongs to N for every spinor ϕ.

Indeed, let ω be a simple spinor $N = M(\omega)$ and let ϕ be any spinor. For $k = 0$ the recurrence relation (21) gives

$$u \lrcorner B_1(\phi,\omega) = B_0(\phi, u\omega). \tag{40}$$

If $u \in N$, then $u\omega = 0$ and (40) yields $u \lrcorner B_1(\phi,\omega) = 0$. Therefore, the vector $B_1(\phi,\omega)$ is orthogonal to N and as such, contained in N. Conversely, let $B_1(\phi,\omega)$ belong to N for every $\phi \in S$. If $u \in N$, then $B_0(\phi, u\omega) = 0$ for every ϕ; therefore, $u\omega = 0$, i.e., $N = M(\omega)$ and ω is simple.

Can every element of N be represented as $B_1(\phi,\omega)$ with a suitable choice of ϕ? To show that this is so, we first observe that the lemma implies

$$\text{if } \psi = p_1 \cdots p_m \omega, \text{ then } \langle B\psi,\omega\rangle \neq 0. \tag{41}$$

To prove that the map $S \to N$ given by $\phi \to B_1(\phi,\omega)$ is surjective we consider the spinors $\phi_\alpha = n_\alpha \psi$ and notice that by virtue of (40),

$$p_\alpha \lrcorner B_1(\phi_\beta,\omega) = \langle B\psi,\omega\rangle \delta_{\alpha\beta}, \quad (\alpha,\beta = 1,...,m),$$

and the collection of vectors $B_1(\phi_\alpha,\omega)$, $\alpha = 1,...,m$ constitutes a basis of N.

The characterization of simple spinors contained in Proposition 7 seems to be new.

The following proposition is derived from the work by Cartan.[2]

Proposition 8: If ω is a simple spinor, then $B_k(\omega,\omega) = 0$ for $k \neq m$ and the m-vector $B_m(\omega,\omega)$ is proportional to the product $n_1 \cdots n_m$ of the vectors constituting a basis of $M(\omega)$.

Indeed, if ω is simple, then it is a Weyl spinor and (15) gives $B_k(\omega,\omega) = 0$ unless $k \equiv m \pmod 4$. If $n \in M(\omega)$, then $n\omega = 0$ and (21) yields

$$n \lrcorner B_{k+2}(\omega,\omega) + n \wedge B_k(\omega,\omega) = 0. \tag{42}$$

If $k \equiv m \pmod 4$, then $B_{k+2}(\omega,\omega) = 0$ and (42) implies

$$n \wedge B_k(\omega,\omega) = 0, \quad \text{for every } n \in N. \tag{43}$$

A similar argument also gives

$$n \lrcorner B_k(\omega,\omega) = 0, \quad \text{for every } n \in N.$$

The only nonzero solution of (43) is for $k = m$. Therefore,

$$\omega \otimes B\omega = n_1 \cdots n_m \langle B\omega, p_m \cdots p_1 \omega\rangle, \tag{44}$$

where the numerical coefficient is determined by putting $\phi = \psi = \omega$ in (11), multiplying on the lhs by $p_m \cdots p_1$, taking the trace of both sides, and noting that $\mathrm{Tr}(p_m \cdots p_1 n_1 \cdots n_m) = 1$.

Chevalley proved also the converse[21] of Proposition 8: If $\omega \neq 0$ is a Weyl spinor and $B_k(\omega,\omega) = 0$ for $k \neq m$, then ω is simple. Therefore, Eq. (44), with the provision that $\omega \neq 0$ is a Weyl spinor, provides a definition of simple spinors which is equivalent to the one based on the maximality of the associated totally null plane $M(\omega)$ given by (22).

It is now easy to see that all Weyl spinors in spaces of dimension $n \leqslant 6$ are simple. Indeed, it follows from (15) that for every Weyl spinor ω corresponding to a space of dimension $2m \leqslant 6$ one has

$$B_k(\omega,\omega) = 0, \quad \text{for } k = 0,1,...,m-1.$$

In dimension (7 and) 8 one encounters the first quadratic constraint on simple spinors, namely $B_0(\omega,\omega) = 0$. In higher dimensional spaces there is a sequence of such constraints, namely

$$B_k(\omega,\omega) = 0,$$

where $k \equiv m \pmod 4$ and $k < m$. Since, for Weyl spinors, ${}^*B_k(\omega,\omega) = \pm B_{2m-k}(\omega,\omega)$, it is enough to consider the constraints for $k < m$.

Consider now a linearly independent pair of simple spinors ω and ϕ. According to Proposition 2, one can find a null basis $(n_1,...,n_m,p_1,...,p_m)$ of V such that

$$M(\omega) = \mathrm{span}\{n_1,...,n_m\},$$

$$M(\phi) = \mathrm{span}\{n_1,...,n_k,p_{k+1},...,p_m\},$$

and

$$\phi = p_{k+1} \cdots p_m \omega,$$

where $k = \dim M(\omega) \cap M(\phi)$. Since

$$p_m n_1 \cdots n_m = \tfrac{1}{2}(-1)^{m-1} n_1 \cdots n_{m-1}(1 + [p_m, n_m])$$

one easily proves by induction that $p_{k+1}\cdots p_m n_1\cdots n_m$ is proportional to $n_1\cdots n_k$ + multivectors of degrees $k+2$, $k+4,\ldots,2m-k$. By multiplying $\omega\otimes B\omega$ with $p_{k+1}\cdots p_m$ on the left and using the last observation one arrives at the following proposition.

Proposition 9: If ω and ϕ are simple spinors, then the dimension of the intersection $M(\omega)\cap M(\phi)$ is the least integer k such that $B_k(\omega,\phi)\neq 0$. The multivector $B_k(\omega,\phi)$ is then proportional to the product of the vectors of a basis of the intersection. This proposition generalizes Proposition 8 and the lemma.

Remark 2: The Abelian subgroup $G(P)$ of Spin(g) corresponding to the subalgebra $[P,P]$ of spin(g) is of dimension $m(m-1)/2$, equal to that of the manifold Σ of directions of simple spinors of one helicity coinciding, say, with that of ω: Its action on Σ is not transitive. For example, if P is a complementary subspace to the MTN subspace $M(\omega)$, then the direction dir ψ of the spinor $\psi=p_1\cdots p_m\omega$ is left invariant by $G(P)$. However, this action is "almost transitive"[22] in the sense that the orbit of $G(P)$ containing dir ω is an open submanifold of Σ and its complement is a submanifold Σ_1 of lower dimension. Indeed, according to Proposition 6 and the lemma, the direction of a simple spinor ϕ does not belong to the orbit containing dir ω if and only if it satisfies the homogeneous equation $B_0(\psi,\phi)=0$ defining the submanifold Σ_1 of Σ. Therefore, a simple spinor ϕ, of the same helicity as ω, can be said to be in a generic position with respect to P if $B_0(\psi,\phi)\neq 0$; there then exists an antisymmetric matrix $(A_{\alpha\beta})$, $\alpha,\beta=1,..,m$ and a number $\xi_0\neq 0$ such that

$$\phi=\xi_0(\exp A)\omega, \tag{45}$$

where

$$A=\sum_{\alpha<\beta}A_{\alpha\beta}p_\alpha p_\beta.$$

(Note that in the expansion of exp A only a finite number of terms are different from 0.) On the other hand, every spinor can be expressed in terms of the basis (30), as was already done in Sec. IV:

$$\phi=\sum_{k=0}^{m}\sum\xi_{\alpha_1\cdots\alpha_k}p_{\alpha_1}\cdots p_{\alpha_k}\omega, \tag{46}$$

where the second sum is over all the sequences (α_i) such that

$$1\leqslant\alpha_1<\cdots<\alpha_k\leqslant m \tag{47}$$

and the term corresponding to $k=0$ is $\xi_0\omega$. Comparing (45) and (46) one obtains

$$\xi_{\alpha_1\cdots\alpha_k}=0,\quad \text{for } k \text{ odd},$$

a condition resulting from the fact that ω and ϕ have equal helicities, and

$$\xi_{\alpha_1\alpha_2}=\xi_0 A_{\alpha_1\alpha_2},\quad \text{for } k=2. \tag{48}$$

Taking (33) into account gives

$$\phi=\xi_0\prod_{\alpha<\beta}\left(1+\frac{\xi_{\alpha\beta}}{\xi_0}\right)p_\alpha p_\beta\omega.$$

By comparing the other terms with even k one obtains

$$l!\sum\xi_{\alpha_1\cdots\alpha_{2l}}p_{\alpha_1}\cdots p_{\alpha_{2l}}=\xi_0 A^l, \tag{49}$$

where the sum is over the sequences (α_i) satisfying (47) with $k=2l$. Since $2l\leqslant m$, there is only a finite set of such relations. Using (48) to eliminate A one obtains a sequence of constraints on the components ξ of a generic simple spinor ϕ of the same helicity as ω; these constraints coincide with the set of equations (a) in Sec. 92 of Cartan's lectures.[1,2] For example, for $l=2$ one obtains

$$\xi_0=\xi_{\alpha_1\alpha_2\alpha_3\alpha_4}=\xi_{\alpha_1\alpha_2}\xi_{\alpha_3\alpha_4}+\xi_{\alpha_1\alpha_3}\xi_{\alpha_4\alpha_2}+\xi_{\alpha_1\alpha_4}\xi_{\alpha_2\alpha_3}.$$

If m is even, then for $l=m/2$ the constraint (49) relates $\xi_{1\ldots m}$ to the Pfaffian of the antisymmetric matrix $(\xi_{\alpha\beta})$; with a suitable generalization of the notion of the Pfaffian one can extend this observation to other values of l; see Ref. 23.

We hope to have convinced the reader that simple spinors are simple indeed.

ACKNOWLEDGMENTS

One of us (AT) thanks the International School for Advanced Studies and the International Centre for Theoretical Physics for their hospitality during his stay in Trieste.

This research was supported in part by the Polish Research Program CPBP 01.03.

[1]E. Cartan, *Leçons sur Théorie des Spineurs I & II* (Hemann, Paris, 1938).
[2]E. Cartan, *The Theory of Spinors*, English transl. by R. F. Streater (Hermann, Paris, 1966).
[3]C. Chevalley, *The Algebraic Theory of Spinors* (Columbia U.P., New York, 1954).
[4]R. Brauer and H. Weyl, Am. J. Math. **57**, 425 (1935).
[5]I. M. Benn and R. W. Tucker, *An Introduction to Spinors and Geometry with Applications in Physics* (Hilger, Bristol, 1987).
[6]R. Penrose, Ann. Phys. (NY) **10**, 71 (1960).
[7]R. Penrose, J. Math. Phys. **8**, 345 (1967).
[8]R. Penrose and W. Rindler, *Spinors and Space-time* (Cambridge U.P., Cambridge, 1986), Vol. 2.
[9]I. Robinson and A. Trautman, in *Proceedings of the Conference on New Theories in Physics*, Kazimierz 1988, edited by Z. Ajduk, S. Pokorski, and A. Trautman (World Scientific, Singapore, 1989).
[10]P. Budinich, Phys. Rep. 137, 35 (1986); P. Budinich and A. Trautman, Lett. Math. Phys. **11**, 315 (1986).
[11]P. Budinich and A. Trautman, *The Spinorial Chessboard*, Trieste Notes in Physics (Springer, Berlin, 1988).
[12]J. Igusa, Am. J. Math. **92**, 997 (1970); V. L. Popov, Trans. Moscow Math. Soc. **1**, 181 (1980).
[13]This introductory material is presented in many books; see, for example, Ref. 5, Chap. 2; Ref. 8, Appendix; Ref. 11, Chap. 6; W. Greub, *Multilinear Algebra* (Springer, Berlin, 1988).
[14]M. F. Atiyah, R. Bott, and A. Shapiro, Topology Suppl. 1 **3**, 3 (1964); H. Baum, *Spin-Strukturen und Dirac-Operatoren über Pseudoriemannscher Mannigfaltigkeiten* (Teubner, Leipzig, 1981); L. Dabrowski, *Group Actions on Spinors* (Bibliopolis, Naples, 1988).
[15]K. M. Case, Phys. Rev. **97**, 810 (1955) and earlier papers cited therein.
[16]Reference 3, Proposition II. 1.6.
[17]L. Silberstein, Philos. Mag. **23**, 790 (1912); I. Robinson, J. Math. Phys. **2**, 290 (1961).
[18]J. F. Plebanski and S. Hacyan, J. Math. Phys. **16**, 2403 (1975); J. Plebanski and K. Rozga, J. Math. Phys. **25**, 1930 (1984).
[19]W. T. Shaw, Class. Quant. Gravit. **2**, L 113 (1985); P. Budinich, Commun. Math. Phys. **107**, 455 (1986); P. Budinich and M. Rigoli, Nuovo Cimento B **102**, 609 (1988).
[20]See, e.g., Ref. 5, Sec. 2.4.
[21]Reference 3, Proposition III. 3.2.
[22]P. Furlan and R. Rączka, J. Math. Phys. **26**, 3021 (1985).
[23]E. R. Caianiello and A. Giovannini, Lett. Nuovo Cimento **34**, 301 (1982).

Dirac equation in external vector fields: Separation of variables

German V. Shishkin and Víctor M. Villalba[a)]
Department of Theoretical Physics, Byelorussian State University, Minsk 220080, Union of Soviet Socialist Republics

(Received 17 May 1988; accepted for publication 5 April 1989)

The method of separation of variables in the Dirac equation in the external vector fields is developed through the search for exact solutions. The essence of the method consists of the separation of the first-order matricial differential operators that define the dependence of the Dirac bispinor on the related variables, but commutation of such operators with the operator of the equations or between them is not assumed. This approach, which is perfectly justified in the presence of gravitational fields, permits one to prove rigorous theorems about necessary and sufficient conditions on the field functions that allow one to separate variables in the Dirac equation. In analogous investigations by other authors [Bagrov *et al.*, *Exact solutions of Relativistic Wave Equations* (Nauka, Novosibirsk, 1982)] for electromagnetic fields an essential demand related to the operators that define the dependence of the bispinor on the separated variables is the demand for the commutation of a complete set of operators between them or with the operators of the Dirac equation. For this reason a series of possibilities that do not satisfy this demand escape the attention of these other authors. The present work liquidates this gap, solving the problem for external vector fields in general.

I. INTRODUCTION

The most important instrument for investigating the spin-$\frac{1}{2}$ particle is the well known Dirac equation that presents a system of four differential equations with first-order partial derivatives. We do not have any universal method for solving such systems today and the creation of such a method is undoubtedly a mathematically important problem. The physical actuality of the problem is evident in view of the wide role of the Dirac equation in modern physics.

For a long time the set of exact solutions of the Dirac equation has been limited, because of mathematical difficulties, to the cases: free electrons, electrons in the Coulomb field, electrons in the constant magnetic field (Zeeman effect), and electrons in the field of plane monochromatic electromagnetic wave. There has been interest in the exact solutions of the Dirac equation because of the necessity of analysis of synchrotronic radiation.

Now we have a large set of exact solutions of Dirac equation. The general feature of such works (by various authors) is the desire first to separate variables, because the separation of variables reduces the system of equations, with partial derivatives, to a system of well studied ordinary differential equations.

First, we want to note the well known symmetry approach successfully used in the investigations of single differential equations with partial derivatives (see the excellent handbook of Miller[1]).

Although Miller and his co-authors began the investigations of the problem of exact solutions of the Dirac equation not long ago,[2] the rigorous demonstration of the fact that the second-order symmetry operators of the Dirac equation in the absence of fields always must be products of the first-order symmetry operators of the same equation seems to be very significant. Note that in Ref. 2 the complete set of operators are deduced.

The problem of separation of variables using the point of view of the complete set of operators of the Dirac equation commuting with the operator of the equation in the presence of the electromagnetic fields has been studied in detail by Bagrov *et al.* (see Ref. 3 and references therein). In these investigations two mutually complementary approaches may be selected.

(1) First, the authors investigate the possibility of separation of variables in the Klein–Gordon–Fock (KGF) equation in the external electromagnetic fields. Thus, the authors try to separate variables in the Dirac equation, for all the classes of field symmetries which allow separation in the KGF equation. However, it should be noticed that the squared Dirac equation is not equivalent to the KGF equation in the presence of external fields because the squared Dirac equation contains the first derivatives of field functions that are absent in the corresponding KGF equation, i.e., authors have assumed that corresponding derivatives are equal to zero. It does not appear to be an accident that Miller's theorem on the decomposition of the second-order operator in the form of the product of the first-order operators mentioned above has been demonstrated only for the free case.

(2) The sets of first-order operators are searched for directly in the Dirac equation.

A remarkable contribution to the above-mentioned investigations is the work of Cook,[4] where the author studies the possibilities of separation of variables in the Dirac equation in the presence of a scalar field. Cook uses the method of Stackel's spaces. This method has been used before successfully in the investigations of the classical Hamilton–Jacobi equation. In view of the characteristics of the Dirac equation, Cook uses his own modification of Stackel's space method. Cook obtains a set of orthogonal curvilinear coordinates, where the separation of variables in the Dirac equation in the presence of scalar fields is possible.

One of the serious problems in modern physics is the investigation of the Dirac equation in gravitational fields.

[a)] Permanent address: Centro de Física, Instituto Venezolano de Investigaciones Científicas (IVIC) Apdo 21827, Caracas 1020-A, Venezuela.

 0022-2488/89/092132-11$02.50

Concerning the separation of variables in the general covariant Dirac equation, we have to acknowledge the pioneer work of Brill and Wheeler,[5] where the Dirac equation is studied in the central symmetric gravitational field and where the separation of variables is fulfilled. Chandrasekhar and other authors[6–10] have considered the problem of separation of variables in the Dirac equation in the gravitational Kerr's field. In contrast to Brill and Wheeler, who, taking into account the diagonality of the metric, work in the simplest normal diagonal tetrad, Chandrasekhar must deal with a nondiagonal metric. Using the Newman–Penrose method and *a priori* knowledge of the symmetry of the problem (Kerr's field) and also the appropriate isotropic tetrad, Chandrasekhar has separated variables in such a complex geometrical situation. In fact, he has used the widely known generalized approach to the separation of variables based on the idea of R separability of Miller.[1]

We should note, too, that all the above-mentioned authors also use very complex curvilinear coordinates except the well known orthogonal coordinates (Cartesian, cylindrical, and spherical). In particular, Cook[4] in the scalar field, Bagrov *et al.*[3] in the electromagnetic field, and Chandrasekhar in the gravitational Kerr's field successfully use very exotic oblate spheroidal coordinates (see also Ref. 2). It is curious why in such general approaches (see again Refs. 2–4 and 11) the separation of variables in the Dirac equation has not been fulfilled in the prolate spheroidal coordinates and in simple coordinates, for example, elliptic cylindrical and parabolic cylindrical coordinates. In the present paper this problem is solved in general.

One of us has proposed a new approach to the problem of separation of variables in the Dirac equation.[12] It is the method of the complete set of operators, but in contrast to a number of authors the commutation of such operators with the operators of the equation or between them is not assumed. At first we have used this method successfully with gravitational fields with diagonal metrics and all the gravitational fields of such kind, allowing the separation of variables in the Dirac equation, have been found. Here this method is developed for the Dirac equation in external vector fields in Cartesian and general orthogonal curvilinear coordinates. Such an approach allows us to enumerate all the vector fields for which partial or complete separation of variables is possible. Naturally we hope that the results of other authors, in particular, those mentioned above, must be partial cases in our investigation.

Planck's constant, the speed of light, and the electron charge have been equated to unity throughout.

II. CARTESIAN COORDINATES

It is advisable at first to consider the problem of separation of variables in the Dirac equation in the vector fields in Cartesian coordinates. Allowing us to find a wide class of fields with "flat" symmetry admitting the separation on one hand and to prepare the basis for the corresponding investigation in curvilinear coordinates on the other hand.

The Dirac equation in Cartesian coordinates takes the form

$$\{\gamma^i(\partial_i - iA_i) + \gamma^j(\partial_j - iA_j) + \gamma^m(\partial_m - iA_m) + \gamma^n(\partial_n - iA_n) + m_0\}\Psi = 0. \tag{2.1}$$

Here A_i, A_j, A_m, and A_n are components of the vector potential. We use the following commutation relations for Dirac's matrices:

$$[\gamma^k,\gamma^l]_+ = 2I\eta^{kl}, \quad \eta^{kl} = \mathrm{diag}(1,1,1,1), \tag{2.2}$$

i.e., all Dirac's matrices are Hermitian. Thus one of the variables is imaginary; we do not concretize it for generality. One of the components of the vector potential must be redetermined to within imaginary unity, correspondingly. Moreover, for mathematical generality we do not require that the vector potential components satisfy the Lorentz condition. However, this may be taken into account in concrete applications.

The separation of variables in Eq. (2.1) is possible by multiplication of its operator by any matrix on the right hand according to the scheme

$$\{H\}\Psi \Rightarrow \{H\}\Gamma\Gamma^{-1}\Psi \Rightarrow (\hat{K}_\alpha + \hat{K}_\beta)\Phi, \quad \Psi = \Gamma\Phi, \tag{2.3}$$

where α and β are groups of separated variables and $\{H\}$ is the operator of Eq. (2.1). If the variables are separated we have

$$(\hat{K}_\alpha + \hat{K}_\beta)\Phi = 0, \tag{2.4}$$

where $\hat{K}_\alpha$ and $\hat{K}_\beta$ are operators depending only on their own variables.

If one of the operators $\hat{K}_\alpha$ or $\hat{K}_\beta$ depends only on one space-time variable, two possibilities exist. The corresponding operator does not include the mass term and therefore contains only one Dirac matrix. In this case after multiplication of Eq. (2.1) by the corresponding Dirac matrix, we obtain that the operators $\hat{K}_\alpha$ and $\hat{K}_\beta$ commute. In our scheme such multiplication is included in the matrix Γ. Analogously if the operator depending on the one variable includes the mass term. The operators $\hat{K}_\alpha$ and $\hat{K}_\beta$ in Eq. (2.4) will commute after multiplication of Eq. (2.1) by the other three Dirac matrices. In the situation where the operators $\hat{K}_\alpha$ and $\hat{K}_\beta$ both depend on two space-time variables (each on its own variables) again one of the operators (without a mass term) contains two Dirac matrices and again the operators $\hat{K}_\alpha$ and $\hat{K}_\beta$ in (2.4) will commute.

Thus in our scheme for separation of variables in the Dirac equation it is necessary that

$$\hat{K}_\alpha\hat{K}_\beta - \hat{K}_\beta\hat{K}_\alpha = 0. \tag{2.5}$$

Of course, for arbitrary operators, for example, with higher-order derivatives or with more complicated matrix algebra, the conditions (2.3)–(2.5) may not be fulfilled. In general, there may be a situation where it is possible to have an operator dependent on only one variable commuting with the operator of the whole equation.

In order to reach the complete separation of variables in Eq. (2.1) we have two possibilities.

(a) We can separate x^i from x^j, x^m, x^n, then separate x^j from x^m, x^n, and at last separate x^m from x^n.

(b) It is possible the other way, namely, we can separate x^i, x^j from x^m, x^n and then separate x^i from x^j and x^m from x^n.

The indices do not have absolute character as a result of covariant consideration. With both possibilities the arbitrary variables may be taken as the first in the scheme of separation of variables.

A. Separation according to the scheme

$x^j,x^l,x^m,x^n \Rightarrow x^j;x^l,x^m x^n \Rightarrow x^j;x^l;x^m,x^n \Rightarrow x^j;x^l;x^m;x^n$

(1) Let us consider the separation of variables in Eq. (2.1) according to scheme (2.3) given $\alpha = x^j$, $\beta = x^l$, x^m, and x^n. Then, instead of (2.3), we have

$$\{H\}\Psi \rightrightarrows \{H\}\Gamma\Gamma^{-1}\Psi \rightrightarrows (\hat{K}_i + \hat{K}_{jmn})\Phi, \quad \Psi = \Gamma\Phi, \tag{2.6}$$

i.e.,

$$(\hat{K}_i + \hat{K}_{jmn})\Phi = 0. \tag{2.7}$$

Assuming that the operators $\hat{K}_i$ and $\hat{K}_{jmn}$ are responsible only for their own variables, we need

$$\hat{K}_i\hat{K}_{jmn} - \hat{K}_{jmn}\hat{K}_i = 0. \tag{2.8}$$

From the explicit form of Eq. (2.1) we can see that the separation of variables according to the scheme (2.7) is possible only if the components of the vector potential separate their variables in an analogous way, namely,

$$A_k = \tilde{A}_k(x^j) + B_k(x^l,x^m,x^n), \quad k = j,l,m,n. \tag{2.9}$$

Indeed even if one of the operators $\hat{K}_i$ or $\hat{K}_{jmn}$ contains a function where x^j is not separated from x^l, x^m, x^n we shall not satisfy requirement (2.8).

Taking into account (2.1) and (2.9) we can write the most general forms of the operators $\hat{K}_i$ and $\hat{K}_{jmn}$:

$$\hat{K}_i = \{\gamma^j(\partial_j - i\tilde{A}_j) - i\gamma^l\tilde{A}_l - i\gamma^m\tilde{A}_m - i\gamma^n\tilde{A}_n + \alpha m_0\}\Gamma, \tag{2.10}$$

$$\hat{K}_{jmn} = \{-i\gamma^j B_j + \gamma^l(\partial_l - iB_l) + \gamma^m(\partial_m - iB_m) + \gamma^n(\partial_n - iB_n) + \beta m_0\}\Gamma, \tag{2.11}$$

where $\alpha + \beta = 1$.

Substituting (2.10) in (2.8) and taking into account only nontrivial commutators we have the matrix equation system

$$\begin{aligned}
&(\gamma^j\Gamma\gamma^l - \gamma^l\Gamma\gamma^j) = 0, && (\gamma^j\Gamma\gamma^m - \gamma^m\Gamma\gamma^j) = 0,\\
& && (\gamma^j\Gamma\gamma^n - \gamma^n\Gamma\gamma^j) = 0,\\
&(\gamma^l\Gamma\gamma^m - \gamma^m\Gamma\gamma^l)\tilde{A}_l = 0, && (\gamma^j\Gamma - \Gamma\gamma^j)\beta = 0,\\
&(\gamma^l\Gamma\gamma^m - \gamma^m\Gamma\gamma^l)\tilde{A}_l B_m = 0, && (\gamma^j\Gamma - \Gamma\gamma^j)\beta\tilde{A}_j = 0,\\
&(\gamma^l\Gamma\gamma^n - \gamma^n\Gamma\gamma^l)\tilde{A}_l = 0, && (\gamma^l\Gamma - \Gamma\gamma^l)\beta\tilde{A}_l = 0,\\
&(\gamma^l\Gamma\gamma^n - \gamma^n\Gamma\gamma^l)\tilde{A}_l B_n = 0, && (\gamma^m\Gamma - \Gamma\gamma^m)\beta\tilde{A}_m = 0,\\
&(\gamma^m\Gamma\gamma^l - \gamma^l\Gamma\gamma^m)\tilde{A}_m = 0, && (\gamma^n\Gamma - \Gamma\gamma^n)\beta\tilde{A}_n = 0,\\
&(\gamma^m\Gamma\gamma^l - \gamma^l\Gamma\gamma^m)\tilde{A}_m B_l = 0, && (\Gamma\gamma^j - \gamma^j\Gamma)\alpha B_j = 0,\\
&(\gamma^m\Gamma\gamma^n - \gamma^n\Gamma\gamma^m)\tilde{A}_m = 0, && (\Gamma\gamma^j - \gamma^j\Gamma)\alpha = 0,\\
&(\gamma^m\Gamma\gamma^n - \gamma^n\Gamma\gamma^m)\tilde{A}_m B_n = 0, && (\Gamma\gamma^l - \gamma^l\Gamma)\alpha B_l = 0,\\
&(\gamma^n\Gamma\gamma^l - \gamma^l\Gamma\gamma^n)\tilde{A}_n = 0, && (\Gamma\gamma^m - \gamma^m\Gamma)\alpha = 0,\\
&(\gamma^n\Gamma\gamma^l - \gamma^l\Gamma\gamma^n)\tilde{A}_n B_l = 0, && (\Gamma\gamma^m - \gamma^m\Gamma)\alpha B_m = 0,\\
&(\gamma^n\Gamma\gamma^m - \gamma^m\Gamma\gamma^n)\tilde{A}_n = 0, && (\Gamma\gamma^n - \gamma^n\Gamma)\alpha = 0,\\
&(\gamma^n\Gamma\gamma^m - \gamma^m\Gamma\gamma^n)\tilde{A}_n B_m = 0, && (\Gamma\gamma^n - \gamma^n\Gamma)\alpha B_n = 0.
\end{aligned} \tag{2.12}$$

The matrix equations of this system have two solutions:

$$\text{(a)}\quad \Gamma = \gamma^j, \quad \text{(b)}\quad \Gamma = \gamma^j\gamma^m\gamma^n. \tag{2.13}$$

Taking into account all of the system (2.12) we have two possibilities for separation of the variables:

(a) $\Gamma = \gamma^j$,

$$A_j = \tilde{A}_j + B_j, \quad A_l = B_l, \quad A_m = B_m, \quad A_n = B_n; \tag{2.14}$$

$$\hat{K}_j = \partial_j - i\tilde{A}_j,$$

$$\hat{K}_{jmn} = \{-i\gamma^j B_j + \gamma^l(\partial_l - iB_l) + \gamma^m(\partial_m - iB_m) + \gamma^n(\partial_n - iB_n) + m_0\}\gamma^j; \tag{2.15}$$

(b) $\Gamma = \gamma^j\gamma^m\gamma^n$,

$$A_j = \tilde{A}_j, \quad A_l = B_l, \quad A_m = B_m, \quad A_n = B_n; \tag{2.16}$$

$$\hat{K}_j = \{\gamma^j(\partial_j - i\tilde{A}_j) + m_0\}\gamma^j\gamma^m\gamma^n,$$

$$\hat{K}_{jmn} = \gamma^j(\partial_j - iB_j) + \gamma^m(\partial_m - iB_m) + \gamma^n(\partial_n - iB_n)\}\gamma^j\gamma^m\gamma^n. \tag{2.17}$$

It may be demonstrated by reverse consideration that the conditions (2.14) or (2.16) are also sufficient for the separation of variables (2.7) in Eq. (2.1). Notice that the conditions (2.16) are a particular case of conditions (2.14).

Now we can formulate the following theorem.

Theorem 1: In order to separate the variables according to the scheme

$$x^j,x^l,x^m,x^n \Rightarrow x^j;x^l,x^m,x^n,$$

it is necesssary and sufficient that the components of the vector potential satisfy conditions (2.14).

Remark: It may be that the vector field satisfies both conditions (2.14) and (2.16). Then there may be variants in the separation of x^j from x^l, x^m, x^n.

(2) Considering that the separation of x^j from x^l, x^m, x^n is fulfilled already let us consider the following step of separation, i.e., the separation of x^l from x^m, x^n in the operator $\hat{K}_{jmn}$.

Now Eq. (2.7) may be rewritten in the form of the problem on eigenvalues and eigenvectors:

$$\hat{K}_j\Phi = -\hat{K}_{jmn}\Phi = k^j\Phi, \tag{2.18}$$

where k^j is an eigenvalue of operators. For the next step of separation we have the scheme

$$(\hat{K}_{jmn} + k^j)\Phi \Rightarrow (\hat{K}_{jmn} + k^j)\Lambda\Lambda^{-1}\Phi \Rightarrow (\hat{K}_j + \hat{K}_{mn})\Theta, \quad \Phi = \Lambda\Theta, \tag{2.19}$$

$$(\hat{K}_j + \hat{K}_{mn})\Theta = 0. \tag{2.20}$$

Assuming that the separation of x^l from x^m, x^n takes place we necessarily have

$$\hat{K}_j\hat{K}_{mn} - \hat{K}_{mn}\hat{K}_j = 0. \tag{2.21}$$

Taking into account the explicit form of the operator $\hat{K}_{jmn}$ [(2.15) or (2.17)], in order to have the separation of x^j from x^m, x^n it is necessary that

$$B_k(x^l,x^m,x^n) = \tilde{B}_k(x^j) + C_k(x^m,x^n), \quad k = j,l,m,n. \tag{2.22}$$

According to (2.15)–(2.19) we obtain the most general form for the operators

$$\hat{K}_j = \{ -i\gamma^i \tilde{B}_i + \gamma^j(\partial_j - iB_j) - i\gamma^m \tilde{B}_m - i\gamma^n \tilde{B}_n + \alpha m_0 + \delta\gamma^i k_i \}\gamma^j \Lambda,$$
$$\hat{K}_{mn} = \{ -i\gamma^i C_i - i\gamma^j C_j + \gamma^m(\partial_m - iC_m) + \gamma^n(\partial_n - iC_n) + \beta m_0 + \epsilon\gamma^i k_i \}\gamma^i \Lambda, \quad (2.23)$$
$$\alpha + \beta = 1, \quad \delta + \epsilon = 1. \quad (2.24)$$

As in the previous step of separation, from (2.21) taking into account the explicit form of the operators (2.23) we obtain a complex system of matrix equations analogous to (2.12) which admits the solutions:

$$\text{(a)}\ \Lambda = \gamma^j, \quad \text{(b)}\ \Lambda = \gamma^i\gamma^j, \quad \text{(c)}\ \Lambda = \gamma^m\gamma^n, \quad \text{(d)}\ \Lambda = \gamma^i\gamma^m\gamma^n. \quad (2.25)$$

We have the following possibilities for separation of variables (2.20):

$$\text{(a)}\ \Lambda = \gamma^j, \quad B_i = \tilde{B}_i, \quad B_j = \tilde{B}_j, \quad B_m = C_m, \quad B_n = C_n; \quad (2.26)$$
$$\hat{K}_j = \{ -i\gamma^i \tilde{B}_i + \gamma^j(\partial_j - i\tilde{B}_j) + m_0 + \gamma^i k_i \}\gamma^i\gamma^j,$$
$$\hat{K}_{mn} = \{\gamma^m(\partial_m - iC_m) + \gamma^n(\partial_n - iC_n)\}\gamma^i\gamma^j; \quad (2.27)$$

$$\text{(b)}\ \Lambda = \gamma^i\gamma^j, \quad B_i = C_i, \quad B_j = \tilde{B}_j + C_j, \quad B_m = C_m, \quad B_n = C_n; \quad (2.28)$$
$$\hat{K}_j = \partial_j - i\tilde{B}_j,$$
$$\hat{K}_{mn} = \{ -i\gamma^i C_i - i\gamma^j C_j + \gamma^m(\partial_m - iC_m) + \gamma^n(\partial_n - iC_n) + m_0 + \gamma^i k_i \}\gamma^j; \quad (2.29)$$

$$\text{(c)}\ \Lambda = \gamma^m\gamma^n, \quad B_i = C_i, \quad B_j = \tilde{B}_j, \quad B_m = C_m, \quad B_n = C_n; \quad (2.30)$$
$$\hat{K}_j = \{\gamma^j(\partial_j - i\tilde{B}_j) + m_0\}\gamma^i\gamma^m\gamma^n,$$
$$\hat{K}_{mn} = \{ -i\gamma^i C_i + \gamma^m(\partial_m - iC_m) + \gamma^n(\partial_n - iC_n) + \gamma^i k_i \}\gamma^i\gamma^m\gamma^n; \quad (2.31)$$

$$\text{(d)}\ \Lambda = \gamma^i\gamma^m\gamma^n, \quad B_i = \tilde{B}_i, \quad B_j = \tilde{B}_j, \quad B_m = C_m, \quad B_n = C_n; \quad (2.32)$$
$$\hat{K}_j = \{ -i\gamma^i \tilde{B}_i + \gamma^j(\partial_j - i\tilde{B}_j) + \gamma^i k_i \}\gamma^m\gamma^n,$$
$$\hat{K}_{mn} = \{\gamma^m(\partial_m - iC_m) + \gamma^n(\partial_n - iC_n) + m_0\}\gamma^m\gamma^n. \quad (2.33)$$

The possibilities (a)–(d) are connected with the operator $\hat{K}_{jmn}$ (2.15). Yet one possibility follows from (2.17):

$$\text{(e)}\ B_j = \tilde{B}_j + C_j, \quad B_m = C_m, \quad B_n = C_n; \quad (2.34)$$
$$\hat{K}_j + (\partial_j - i\hat{B}_j)\gamma^m\gamma^n\Lambda,$$
$$\hat{K}_{mn} = \{ -i\gamma^j C_j + \gamma^m(\partial_m - iC_m) + \gamma^n(\partial_n - iC_n) - \gamma^j\gamma^m\gamma^n k_i \}\gamma^j\gamma^m\gamma^n\Lambda. \quad (2.35)$$

The conditions of separability of variables are deduced from (2.21), i.e., they are necessary. By reverse consideration it may be demonstrated that these conditions are sufficient, too. So we have the following theorem [note that conditions (2.26) and (2.32) are equivalent and (2.30) and (2.34) are particular cases of (2.28)].

Theorem 2: In order to separate the variables in (2.1) according to the scheme

$$x^i,x^j,x^m,x^n \Rightarrow x^i;x^j,x^m,x^n \Rightarrow x^i;x^j;x^m,x^n,$$

it is necessary and sufficient that the components of the vector potential satisfy conditions (2.14), (2.28) or (2.14), (2.26).

Remark: The variants are possible for stronger conditions.

(3) Before beginning the separation of x^m from x^n in the operator $\hat{K}_{mn}$ notice that the operator $\hat{K}_{mn}$ contains these variables symmetrically independent of the variant of the previous steps of separation of variables (x^i and x^j from x^m,x^n). Therefore we shall write only the results that are not identical relative to the change $m \rightleftarrows n$. Since the procedure to separate x^m from x^n is not different than previously given, we will note only the main points and write the results.

Thus we accept the scheme

$$(\hat{K}_{mn} + k^j)\Theta \Rightarrow (\hat{K}_{mn} + k^j)\Sigma\Sigma^{-1}\Theta \Rightarrow (\hat{K}_m + \hat{K}_n)\Omega, \quad \Theta = \Sigma\Omega, \quad (2.36)$$
$$(\hat{K}_m + \hat{K}_n)\Omega = 0, \quad (2.37)$$
$$\hat{K}_m\hat{K}_n - \hat{K}_n\hat{K}_m = 0. \quad (2.38)$$

Here k^j is an eigenvalue of the operator $\hat{K}_j$.

Combining the requirement (2.38) with each possible definition of the operator $\hat{K}_{mn}$ we have all the possibilities for separation (2.17).

It may be seen from (2.17) and (2.38) that

$$C_m = \tilde{C}_m(x^m) + \tilde{D}_m(x^n), \quad C^n = \tilde{D}_n(x^n); \quad (2.39)$$
$$\hat{K}_m = \gamma^m(\partial_m - i\tilde{C}_m)\gamma^i\gamma^j\Sigma, \quad (2.40)$$
$$\hat{K}_n = \{ -i\gamma^m \tilde{D}_m + \gamma^n(\partial_n - i\tilde{D}_n) - \gamma^i\gamma^j k_i \}\gamma^i\gamma^j\Sigma;$$
$$\Sigma = \gamma^m,\gamma^i\gamma^m,\gamma^j\gamma^m,\gamma^i\gamma^j\gamma^m. \quad (2.41)$$

Analogously from (2.29) and (2.38) we have

$$\text{(a)}\ C_i = \tilde{D}_i(x^n), \quad C_j = \tilde{C}_j(x^m), \quad C_m = \tilde{C}_m(x^m), \quad C_n = \tilde{D}_n(x^n); \quad (2.42)$$
$$\hat{K}_m = \{ -i\gamma^j \tilde{C}_j + \gamma^m(\partial_m - i\tilde{C}_m) + m_0 + \gamma^j k_j \}\gamma^j\gamma^m,$$
$$\hat{K}_n = \{ -i\gamma^i \tilde{D}_i + \gamma^n(\partial_n - i\tilde{D}_n) + \gamma^i k_i \}\gamma^j\gamma^m; \quad (2.43)$$

(b) $C_i = \tilde{C}_i(x^m), \quad C_j = \tilde{C}_j(x^m),$
$C_m = \tilde{C}_m(x^m), \quad C_n = \tilde{D}_n(x^n);$ (2.44)

$\hat{K}_m = \{ -i\gamma^i\tilde{C}_i - i\gamma^j\tilde{C}_j + \gamma^m(\partial_m - i\tilde{C}_m)$
$+ \gamma^i k^{\,i} + \gamma^j k^{\,j}\}\gamma^i\gamma^j\gamma^m,$ (2.45)

$\hat{K}_n = \{ -\gamma^n(\partial_n - i\tilde{D}_n) - m_0\}\gamma^i\gamma^j\gamma^m;$

(c) $C_i = \tilde{D}_i(x^n), \quad C_j = \tilde{D}_j(x^n),$
$C_m = \tilde{C}_m(x^m) + \tilde{D}_m(x^n), \quad C_n = D_n(x^n);$ (2.46)

$\hat{K}_m = \partial_m - iC_m,$

$\hat{K}_n = \{ -i\gamma^i\tilde{D}_i - i\gamma^j\tilde{D}_j - i\gamma^m\tilde{D}_m + \gamma^n(\partial_n - i\tilde{D}_n)$
$+ m_0 + \gamma^i k^{\,i} + \gamma^j k^{\,j}\}\gamma^m;$ (2.47)

(d) $C_i = \tilde{C}_i(x^m), \quad C_j = \tilde{C}_j(x^m),$
$C_m = \tilde{C}_m(x^m), \quad C_n = \tilde{D}_n(x^n);$ (2.48)

$\hat{K}_m = -\{ -i\gamma^i\tilde{C}_i + \gamma^m(\partial_m - i\tilde{C}_m) + m_0$
$+ \gamma^i k^{\,i}\}\gamma^i\gamma^m,$ (2.49)

$\hat{K}_n = -\{ -i\gamma^j\tilde{D}_j + \gamma^n(\partial_n - i\tilde{D}_n) + \gamma^j k^{\,j}\}\gamma^i\gamma^m.$

It follows from (2.31) and (2.38) that

$C_i = \tilde{C}_i(x^m), \quad C_j = \tilde{C}_j(x^m),$
$C_m = \tilde{C}_m(x^m), \quad C_n = \tilde{D}_n(x^n);$ (2.50)

$\hat{K}_m = \{ -i\gamma^i\tilde{C}_i - i\gamma^j\tilde{C}_j + \gamma^m(\partial_m - i\tilde{C}_m)$
$+ \gamma^i k^{\,i} - \gamma^i\gamma^j\gamma^m k^{\,j}\}\gamma^j\gamma^m\gamma^n\Sigma,$ (2.51)

$\hat{K}_n = (\partial_n - i\tilde{D}_n)\gamma^i\gamma^m\Sigma;$

$\Sigma = \gamma^n, \gamma^j\gamma^n, \gamma^i\gamma^m, \gamma^i\gamma^j\gamma^m.$ (2.52)

We can see from (2.33) and (2.38) that

$C_m = \tilde{C}_m(x^m), \quad C_n = \tilde{C}_n(x^m) + \tilde{D}_n(x^n);$ (2.53)

$\hat{K}_m = \{\gamma^m(\partial_m - i\tilde{C}_m) - i\gamma^n\tilde{C}_n\}\gamma^m\gamma^n\Sigma,$ (2.54)

$\hat{K}_n = \{\gamma^n(\partial_n - i\tilde{D}_n) + m_0 - \gamma^m\gamma^n k^{\,j}\}\gamma^m\gamma^n\Sigma;$

$\Sigma = \gamma^n, \gamma^j\gamma^n, \gamma^i\gamma^m, \gamma^i\gamma^j\gamma^m.$ (2.55)

At last we have from (2.35) and (2.38)

$C_j = \tilde{C}_j(x^m); \quad C_m = \tilde{C}_m(x^m),$
$C_n = \tilde{C}_n(x^m) + \tilde{D}_n(x^n);$ (2.56)

$\hat{K}_m = \{ -i\gamma^j\tilde{C}_j + \gamma^m(\partial_m - i\tilde{C}_m) - i\gamma^n C_n$
$- \gamma^j\gamma^m\gamma^n k^{\,i} + \gamma^j k^{\,j}\}\gamma^j\Sigma,$ (2.57)

$\hat{K}_n = \gamma^n(\partial_n - i\tilde{D}_n)\gamma^j\Sigma;$

$\Sigma = \gamma^m, \gamma^j\gamma^m, \gamma^j\gamma^n, \gamma^j\gamma^j\gamma^n.$ (2.58)

Taking the least strict and different conditions and convincing ourselves of their sufficiency we can formulate the following theorem.

Theorem 3: In order to separate all the variables in Eq. (2.1) according to the scheme

$x^i, x^j, x^m, x^n \Rightarrow x^i; x^j, x^m, x^n$
$\Rightarrow x^i; x^j; x^m, x^n \Rightarrow x^i; x^j; x^m; x^n,$

it is necessary and sufficient that the components of the vector potential satisfy conditions (2.14), (2.28), (2.46), or (2.14), (2.26), (2.46), or (2.14), (2.26), (2.42) or (2.14), (2.26), (2.44).

Remark: In the case of more strict conditions the variants are possible.

B. Separation according to the scheme $x^i, x^j, x^m, x^n \Rightarrow x^i, x^j; x^m, x^n \Rightarrow x^i; x^j; x^m; x^n$

Let us consider the separation of variables in Eq. (2.1) according to the scheme

$\{H\}\Gamma\Gamma^{-1}\Psi \Rightarrow (\hat{K}_{ij} + \hat{K}_{mn})\Phi, \quad \Psi = \Gamma\Phi,$ (2.59)

$(\hat{K}_{ij} + \hat{K}_{mn})\Phi = 0,$ (2.60)

$\hat{K}_{ij}\hat{K}_{mn} - \hat{K}_{mn}\hat{K}_{ij} = 0.$ (2.61)

Notice that in this separation we have symmetry relative to the change $i,j \rightleftarrows m,n$. Later in the separation within x^i and x^j (or within x^m and x^n) symmetry such as $i \rightleftarrows j$ (or $m \rightleftarrows n$) takes place. Therefore from now on we shall write only results that are not identical relative to the above-mentioned changes.

In order to provide the commutation requirement (2.61)

$A^{\,k}(x^i, x^j, x^m, x^n) = V_k(x^i, x^j) + W_k(x^m, x^n),$
$k = i, j, m, n,$ (2.62)

must hold.

Writing the most general form of the operators $\hat{K}_{ij}$ and $\hat{K}_{mn}$ and taking into account the commutator (2.61), analogously to our previous discussions, we have

$A_i = V_i, \quad A_j = V_j, \quad A_m = W_m, \quad A_n = W_n;$ (2.63)

$\hat{K}_{ij} = \{\gamma^i(\partial_i - iV_i) + \gamma^j(\partial_j - iV_j) + m_0\}\gamma^i\gamma^j,$
$\hat{K}_{mn} = \{\gamma^m(\partial_m - iW_m) + \gamma^n(\partial_n - iW_n)\}\gamma^i\gamma^j.$ (2.64)

After the change $i,j \rightleftarrows m,n$ we have the identical variant.

Now we are led to Theorem 4.

Theorem 4: In order to separate the variables into pairs in Eq. (2.1) according to the scheme

$x^i, x^j, x^m, x^n \Rightarrow x^i, x^j; x^m, x^n,$

it is necessary and sufficient that the components of the vector potential satisfy conditions (2.63).

The variables x^i, x^j are separated according to the scheme

$(\hat{K}_{ij} + k)\Phi \Rightarrow (\hat{K}_{ij} + k)\Lambda\Lambda^{-1}\Gamma$
$\Rightarrow (\hat{K}_i + \hat{K}_j)\Theta, \quad \Phi = \Lambda\Theta,$ (2.65)

$(\hat{K}_i + \hat{K}_j)\Theta = 0,$ (2.66)

$\hat{K}_i\hat{K}_j - \hat{K}_j\hat{K}_i = 0.$ (2.67)

Here k is an eigenvalue of the operator $\hat{K}_{mn}$.

After this separation we have

$V_i = \tilde{A}_i(x^i) + \tilde{B}_i(x^j), \quad V_j = \tilde{B}_j(x^j);$ (2.68)

$\hat{K}_i = \gamma^i(\partial_i - i\tilde{A}_i)\Lambda,$
$\hat{K}_j = \{ -i\gamma^i\tilde{B}_i + \gamma^j(\partial_j - \tilde{B}_j) + m_0 - \gamma^i\gamma^j k\}\Lambda;$ (2.69)

$\Lambda = \gamma^j, \gamma^i\gamma^m\gamma^n, \gamma^i\gamma^m, \gamma^i\gamma^n.$ (2.70)

We must remember about variants connected with the symmetry relative to the changes $i,j \rightleftarrows m,n$, $i \rightleftarrows j$.

Analogously

$$(\hat{K}_{mn} - k)\Phi \Rightarrow (\hat{K}_{mn} - k)\Sigma\Sigma^{-1}\Phi \Rightarrow (\hat{K}_m + \hat{K}_n)\Omega, \quad \Phi = \Sigma\Omega, \tag{2.71}$$

$$(\hat{K}_m + \hat{K}_n)\Omega = 0, \tag{2.72}$$

$$\hat{K}_m\hat{K}_n - \hat{K}_n\hat{K}_m = 0; \tag{2.73}$$

$$W_m = \tilde{C}_m(x^m) + \tilde{D}_m(x^n), \quad W_n = \tilde{D}_n(x^n); \tag{2.74}$$

$$\hat{K}_m = \gamma^m(\partial_m - i\tilde{C}_m)\Sigma, \quad \hat{K}_n = \{-i\gamma^m\tilde{D}_m + \gamma^n(\partial_n - i\tilde{D}_n) + \gamma^i\gamma^j k\}\Sigma; \tag{2.75}$$

$$\Sigma = \gamma^m, \gamma^i\gamma^m, \gamma^j\gamma^m, \gamma^i\gamma^j\gamma^m. \tag{2.76}$$

Again we remember the symmetric-variants (the changes $i,j \rightleftarrows m,n$, $m \rightleftarrows n$).

Combining the results (2.63), (2.68), and (2.74) we have the following theorem.

Theorem 5: In order to separate all the variables in Eq. (2.1) according to the scheme

$$x^i,x^j,x^m,x^n \Rightarrow x^i,x^j;x^m,x^n \Rightarrow x^i;x^j;x^m;x^n,$$

it is necessary and sufficient that the components of the vector potential satisfy conditions (2.63), (2.68), and (2.74).

Analyzing the conditions of Theorems 3 and 5 while taking into account the equivalence of the variables we can see that the conditions of Theorem 3 contain conditions of Theorem 5. So if the conditions of Theorem 5 are fulfilled we have additional possibilities for separation in comparison with Theorem 3 connected with pair separation in the first step. The mathematical realization of the pair separation is simpler and therefore more preferable. Taking into account the least strict conditions on the components of the vector potential we can formulate the general theorem.

Theorem 6: In order to separate all the variables in Eq. (2.1) it is necessary and sufficient that the components of the vector potential satisfy conditions (2.14), (2.28), (2.46) or (2.14), (2.26), (2.46) or (2.14), (2.26), (2.42) or (2.14), (2.26), (2.44).

The explicit form of the operator connected with the concrete set of conditions takes into account a concrete way of separating variables. In the case of stronger conditions on the vector potential we obtain weaker conditions on the operators, and then the different variants of separation of variables in the equation investigated [(2.1)] are possible.

III. CURVILINEAR COORDINATES

When we write the Dirac equation in curvilinear coordinates, Lame's coefficient functions appear. More precisely, we find that each partial derivative is multiplied by a linear combination of "flat" Dirac matrices where the coefficients are related to Lame's functions. Therefore we can consider each linear combination as a matrix which depends on the curvilinear coordinates. Such a situation introduces some additional limitations into the problem of separation of variables, but as before the requirements of partial or complete separability of variables in the Dirac equation allow us to find the conditions on the components of vector potential and to select the concrete types of orthogonal curvilinear coordinates allowing such separations. It has been shown in Ref. 13 that the linear combination of Dirac matrices before a concrete curvilinear derivative through any unitary transformation may be reduced to one constant matrix with any functional coefficient. Together with the transition to a new unknown function this allows us to simplify the Dirac equation as much as possible. We will consider that such a procedure for simplification of the Dirac equation is already fulfilled. However, we should notice that the unitary transformation used here for simplicity leads us to the framework of the natural Cartesian gauge of a space triad and when we solve the simplified equation we cannot require that the wave function be single-valued with respect to the variables connected with the above-mentioned unitary transformation. Therefore after solving the simplified equation it is natural to do the inverse unitary transformation of solutions, and afterward the correct boundary conditions may be imposed.

Investigating the Dirac equation in the vector fields on the subject of separation of variables in the curvilinear coordinates we select two groups of coordinates including all the known concrete orthogonal curvilinear coordinates.

A. Coordinates μ, ν, z, $\tau = it$ [$x = f(\mu,\nu)$, $y = g(\mu,\nu)$]

Here the curvilinear coordinates are introduced on the XY plane. In particular, the well known cylindrical coordinates are of this kind.

The Dirac equation in the coordinates μ, ν, z, and τ takes the form

$$\left\{\frac{f_\mu\gamma^1 + g_\mu\gamma^2}{(f_\mu^2 + g_\mu^2)^{1/2}}\left(\frac{\partial_\mu}{(f_\mu^2 + g_\mu^2)^{1/2}} - iA_\mu\right) + \frac{f_\nu\gamma^1 + g_\nu\gamma^2}{(f_\nu^2 + g_\nu^2)^{1/2}}\left(\frac{\partial_\nu}{(f_\nu^2 + g_\nu^2)^{1/2}} - iA_\nu\right) + \gamma^z(\partial_z - iA_z) + \gamma^4(\partial_\tau - iA_4) + m_0\right\}\Psi = 0. \tag{3.1}$$

Here f_μ, f_ν, g_μ, and g_ν are the curvilinear derivatives of the functions f and g on the corresponding variables. After some unitary transformation through the operator $\hat{S}$ and transition to the new unknown function (see Ref. 13) according to

$$\gamma \rightarrow \hat{S}^{-1}\gamma\hat{S}, \quad \Psi = [\hat{S}/(f_\mu^2 + g_\mu^2)^{1/4}]\tilde{\Psi}, \quad \hat{\zeta} = \exp\{-\tfrac{1}{2}\varphi\gamma^1\gamma^2\}, \quad \varphi = \varphi(\mu,\nu), \tag{3.2}$$

we have

$$\left\{\gamma^\mu\left(\frac{\partial_\mu}{(f_\mu^2 + g_\mu^2)^{1/2}} - iA_\mu\right) + \gamma^\nu\left(\frac{\partial_\nu}{(f_\nu^2 + g_\nu^2)^{1/2}} - iA_\nu\right) + \gamma^z(\partial_z - iA_z) + \gamma^4(\partial_\tau - iA_4) + m_0\right\}\tilde{\Psi} = 0. \tag{3.3}$$

All the matrices γ are now constant. Naturally, the components of the vector potential are determined here through curvilinear variables.

From the orthogonality of curvilinear variables we have

$$f_\mu = g_\nu, \quad f_\nu = -g_\mu; \tag{3.4}$$

$$\{f_\mu^2 + g_\mu^2\}^{1/2} = \{f_\nu^2 + g_\nu^2\}^{1/2} = \mathcal{F}(\mu,\nu) = \mathcal{F}. \tag{3.5}$$

Since the variables z and τ are not "spoiled" by curvilinearity we can separate them from μ, ν and later separate between them corresponding to the conditions of Sec. II.

If the conditions of Theorems 1 and 2 are fulfilled we separate first τ and then z (due to the symmetry $z \rightleftarrows \tau$ the sequence may be reversed) and combining requirements (2.14), (2.26) or (2.14), (2.28) we have all the possibilities for such a separation:

(a)
$$A_\mu = C_\mu(\mu,\nu), \quad A_\nu = C_\nu(\mu,\nu), \quad A_z = \tilde{B}_z(z), \quad A_4 = \tilde{A}_4(\tau) + \tilde{B}_4(z); \tag{3.6}$$

$$\hat{K}_{\mu\nu} = \{\gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu)\}\gamma^A\gamma^z; \tag{3.7}$$

(b)
$$A_\mu = C_\mu(\mu,\nu), \quad A_\nu = C_\nu(\mu,\nu), \quad A_z = \tilde{B}_z(z) + C_z(\mu,\nu), \quad A_4 = \tilde{A}_4(\tau) + C_4(\mu,\nu); \tag{3.8}$$

$$\hat{K}_{\mu\nu} = \{-i\gamma^A C_4 - i\gamma^z C_z + \gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu) + m_0 + \gamma^A\epsilon\}\gamma^z; \tag{3.9}$$

(c)
$$A_\mu = C_\mu(\mu,\nu), \quad A_\nu = C_\nu(\mu,\nu), \quad A_z = \tilde{B}_z(z), \quad A_4 = \tilde{A}_4(\tau) + C_4(\mu,\nu); \tag{3.10}$$

$$\tilde{K}_{\mu\nu} = \{-i\gamma^A C_4 + \gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu) + \gamma^A\epsilon\}\gamma^A\gamma^\mu\gamma^\nu; \tag{3.11}$$

(d)
$$A_\mu = C_\mu(\mu,\nu), \quad A_\nu = C_\nu(\mu,\nu), \quad A_z = \tilde{B}_z(z), \quad A_4 = \tilde{A}_4(\tau) + C_4(\mu,\nu); \tag{3.12}$$

$$\hat{K}_{\nu\mu} = \{\gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu) + m_0\}\gamma^\mu\gamma^\nu; \tag{3.13}$$

(e)
$$A_\mu = C_\mu(\mu,\nu), \quad A_\nu = C_\nu(\mu,\nu), \quad A_z = \tilde{B}_z(z) + C_z(\mu,\nu), \quad A_4 = \tilde{A}_4(\tau); \tag{3.14}$$

$$\hat{K}_{\mu\nu} = \{-i\gamma^z C_z + \gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu) - \gamma^z\gamma^\mu\gamma^\nu\epsilon\}\Gamma, \tag{3.15}$$

$$\Gamma = \gamma^z, \gamma^A\gamma^z, \gamma^\mu\gamma^\nu, -\gamma^A\gamma^\mu\gamma^\nu. \tag{3.16}$$

Here ϵ is an eigenvalue of the operator responsible for the variable separated first, for example, $\hat{K}_\tau$.

If the requirements of Theorem 5 are fulfilled on the variables τ, z, separating first τ, z from μ, ν and using the conditions (2.62), (2.68), we have

$$A_\mu = C_\mu(\mu,\nu), \quad A_\nu = C_\nu(\mu,\nu), \quad A_z = \tilde{B}_z(z), \quad A_4 = \tilde{A}_4(\tau) = \tilde{B}_4(z). \tag{3.17}$$

There are two possibilities for the operator $\hat{K}_{\mu\nu}$:

$$\hat{K}_{\mu\nu} = \{\gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu)\}\gamma^z\gamma^A \tag{3.18}$$

or

$$\hat{K}_{\mu\nu} = \{\gamma^\mu(\partial_\mu/\mathcal{F} - iC_\mu) + \gamma^\nu(\partial_\nu/\mathcal{F} - iC_\nu) + m_0\}\gamma^\mu\gamma^\nu. \tag{3.19}$$

Also, it should be noticed that conditions (3.6), (3.12), and (3.17) coincide and conditions (3.10) and (3.14) are included in (3.8).

The Lame's function $\mathcal{F}$, appearing in the expression for the operator $\hat{K}_{\mu\nu}$, depends on two variables, and, in general, the separation of μ and ν is impossible. The separation of μ and ν in Eq. (3.3) is possible if the function $\mathcal{F}$ depends on one variable only, as for example, for the cylindrical variables

$$x = \rho\cos\varphi, \quad y = \rho\sin\varphi \Rightarrow (f_\varphi^2 + g_\varphi^2)^{1/2} = \rho. \tag{3.20}$$

Because of the symmetry between μ and ν, without loss of generality for the separation μ,ν we take

$$\mathcal{F} = \tilde{\mathcal{F}}(\nu). \tag{3.21}$$

The presence of the function $\mathcal{F}$ requires the introduction of some functional factor G in the scheme of separation that must be defined. So we now take the scheme

$$G(\hat{K}_{\mu\nu} + k)\Sigma\Sigma^{-1}\Phi \Rightarrow (\hat{K}_\mu + \hat{K}_\nu)\Omega, \quad \Phi = \Sigma\Omega. \tag{3.22}$$

Here k is an eigenvalue of the operator $\hat{K}_{\mu\nu}$ if we use (3.7), (3.9), (3.13), and (3.15) or is an eigenvalue of the operator $\hat{K}_{\tau z}$ if we use (3.18) and (3.19). In both cases the separation is possible if

$$A_\mu = \tilde{C}_\mu(\mu)/\tilde{\mathcal{F}}(\nu) + \tilde{D}_\mu(\nu), \quad A_\nu = \tilde{D}_\nu(\nu). \tag{3.23}$$

We do not write the operators $\hat{K}_\mu$ and $\hat{K}_\nu$. They may be taken from the operators $\hat{K}_m$ and $\hat{K}_n$ (Sec. II) after the changes $i\to 4$, $j\to z$, $m\to\mu$, $n\to\nu$, $k^i\to\epsilon$, and $k^j\to k^z$, or else the operator $\hat{K}_n$ must be multiplied on $\tilde{\mathcal{F}}(\nu)$. Because of the specification of the scheme (3.22) only the variants (2.40), (2.47), (2.51), and (2.57) are realized. So we have the following theorem.

Theorem 7: In order to separate all the variables in Eq. (3.3) under condition (3.21) it is necessary and sufficient that the components of the vector potential satisfy conditions (3.8), (3.23) or (3.6), (3.23).

Remark: The variants are possible for stronger conditions on the components A_μ, A_ν (see Sec. II).

B. Coordinates μ, ν, φ, τ, $[x = f(\mu,\nu)\cos\varphi, y = f\sin\varphi, z = g(\mu,\nu)]$

The Dirac equation in the variables μ, ν, and φ takes the form

$$\left\{\frac{f_\mu\cos\varphi\,\gamma^1 + f_\mu\sin\varphi\,\gamma^2 + g_\mu\gamma^3}{(f_\mu^2 + g_\mu^2)^{1/2}}\left(\frac{\partial_\mu}{(f_\mu^2 + g_\mu^2)^{1/2}} - iA_\mu\right) + \frac{f_\nu\cos\varphi\,\gamma^1 + f_\nu\sin\varphi\,\gamma^2 + g_\nu\gamma^3}{(f_\mu^2 + g_\mu^2)^{1/2}}\left(\frac{\partial_\nu}{(f_\mu^2 + g_\mu^2)^{1/2}} - iA_\nu\right) + (-\sin\varphi\,\gamma^1 + \cos\varphi\,\gamma^2)\left(\frac{\partial_\varphi}{f} - iA_\varphi\right) + \gamma^A(\partial_\tau - iA_4) + m_0\right\}\Psi = 0. \tag{3.24}$$

After the unitary transformation and transition to the new unknown function, according to[13]

$$\gamma \to \hat{S}^{-1}\gamma\hat{S}, \quad \tilde{\Psi} = \frac{\hat{S}\alpha}{(f_\mu^2 + g_\mu^2)^{1/4} f^{1/2}}\Psi,$$
$$\hat{S} = \exp\left\{-\frac{\chi}{2}\gamma^1\gamma^2\right\}\exp\left\{\frac{\theta}{2}\gamma^1\gamma^3\right\}, \tag{3.25}$$
$$\chi = \chi(\mu,\nu), \quad \theta = \theta(\mu,\nu),$$
$$\alpha = \tfrac{1}{2}(\gamma^1\gamma^2 + \gamma^2\gamma^3 + \gamma^3\gamma^1 + I),$$

we have the simplest form of the Dirac equation:

$$\left\{\gamma^\mu\left(\frac{\partial_\mu}{(f_\mu^2 + g_\mu^2)^{1/2}} - iA_\mu\right) + \gamma^\nu\left(\frac{\partial_\nu}{(f_\mu^2 + g_\mu^2)^{1/2}} - iA_\nu\right) + \gamma^\varphi\left(\frac{\partial_\varphi}{f} - iA_\varphi\right) + \gamma^4(\partial_\tau - iA_4) + m_0\right\}\tilde{\Psi} = 0. \tag{3.26}$$

Now, as before, the matrices γ are constant and Hermitian.

Here only the variable τ has an unspoiled form and therefore it is natural first to separate τ from μ, ν, φ, i.e., now we follow the scheme

$$\tau,\varphi,\mu,\nu \Rightarrow \tau;\varphi;\mu,\nu \Rightarrow \tau;\varphi;\mu,\nu \Rightarrow \tau;\varphi;\mu;\nu.$$

Then according to (2.14)–(2.17) we have the following possibilities of separation of τ from μ, ν, φ:

(a) $$A_\mu = B_\mu(\mu,\nu,\varphi), \quad A_\nu = B_\nu(\mu,\nu,\varphi),$$
$$A_\varphi = B_\varphi(\mu,\nu,\varphi), \quad A_4 = \tilde{A}_4(\tau) + B_4(\mu,\nu,\varphi); \tag{3.27}$$

$$\hat{K}_\tau = \partial_\tau - i\tilde{A}_4,$$
$$\hat{K}_{\mu\nu\varphi} = \left\{-i\gamma^4 B_4 + \gamma^\mu\left(\frac{\partial_\mu}{\mathscr{F}} - iB_\mu\right) + \gamma^\nu\left(\frac{\partial_\nu}{\mathscr{F}} - iB_\nu\right) + \gamma^4\left(\frac{\partial_\varphi}{f} - iA_\varphi\right) + m_0\right\}\gamma^4; \tag{3.28}$$

(b) $$A_\mu = B_\mu(\mu,\nu,\varphi), \quad A_\nu = B_\nu(\mu,\nu,\varphi),$$
$$A_\varphi = B_\varphi(\mu,\nu,\varphi), \quad A_4 = \tilde{A}_4(\tau); \tag{3.29}$$
$$\hat{K}_\tau = \{\gamma^4(\partial_\tau - i\tilde{A}_4) + m_0\}\gamma^\mu\gamma^\nu\gamma^\varphi,$$
$$\hat{K}_{\mu\nu\varphi} = \left\{\gamma^\mu\left(\frac{\partial_\mu}{\mathscr{F}} - iB_\mu\right)\gamma^\nu\left(\frac{\partial_\nu}{\mathscr{F}} - iB_\nu\right) + + \gamma^\varphi\left(\frac{\partial_\varphi}{f} - iA_\varphi\right)\right\}\gamma^\mu\gamma^\nu\gamma^\varphi. \tag{3.30}$$

The complete separation is possible only if

$$(f_\mu^2 + g_\mu^2)^{1/2} = \tilde{\mathscr{F}}(\nu), \quad f(\mu,\nu) = \tilde{\mathscr{F}}(\nu)\tilde{F}(\mu). \tag{3.31}$$

As above we remember the possibilities of symmetrical variants ($\mu \rightleftarrows \nu$).

We have from (3.28)

(a.a) $$B_\mu = C_\mu(\mu,\nu), \quad B_\nu = C_\nu(\mu,\nu),$$
$$B_\varphi = \tilde{B}_\varphi(\varphi)/f + C_\varphi(\mu,\nu), \quad B_4 = C_4(\mu,\nu); \tag{3.32}$$

$$\hat{K}_\varphi = \partial_\varphi - i\tilde{B}_\varphi,$$
$$\hat{K}_{\mu\nu} = f\{-i\gamma^4 C_4 + \gamma^\mu(\partial_\mu/\tilde{\mathscr{F}} - iC_\mu) + \gamma^\nu(\partial_\nu/\tilde{\mathscr{F}} - iC_\nu) - i\gamma^\varphi C_\varphi + m_0 + \gamma^4\epsilon\}\gamma^\varphi; \tag{3.33}$$

(a.b) $$C_\mu = \tilde{C}_\mu(\mu)/\tilde{\mathscr{F}}, \quad C_\nu = \tilde{D}_\nu(\nu),$$
$$C_\varphi = \tilde{C}_\varphi(\mu)/\tilde{\mathscr{F}}, \quad C_4 = \tilde{D}_4(\nu); \tag{3.34}$$
$$\hat{K}_\mu = \{\gamma^\mu(\partial_\mu - i\tilde{C}_\mu) - i\gamma^\varphi\tilde{C}_\varphi + \gamma^\varphi k^\varphi/F\}\gamma^\nu\gamma^4; \tag{3.35}$$

$$\hat{K}_\nu = \tilde{\mathscr{F}}\{-i\gamma^4\tilde{D}_4 + \gamma^\nu(\partial_\nu/\tilde{\mathscr{F}} - \tilde{D}_\nu) + m_0 + \gamma^4\epsilon\}\gamma^\nu\gamma^4. \tag{3.36}$$

From (3.30) it follows that

(b.a) $$B_\mu = C_\mu(\mu,\nu), \quad B_\nu = C_\nu(\mu,\nu),$$
$$B_\varphi = \tilde{B}_\varphi(\varphi)/f + C_\varphi(\mu,\nu); \tag{3.37}$$
$$\hat{K}_\varphi = (\partial_\varphi - i\tilde{B}_\varphi)\gamma^\mu\gamma^\nu\Lambda,$$
$$\hat{K}_{\mu\nu} = f\{\gamma^\mu(\partial_\mu/\tilde{\mathscr{F}} - iC_\mu) + \gamma^\nu(\partial_\nu/\tilde{\mathscr{F}} - iC_\nu) - i\gamma^\varphi C_\varphi - \gamma^\mu\gamma^\nu\gamma^\varphi\epsilon\}\gamma^\mu\gamma^\nu\Lambda; \tag{3.38}$$
$$\Lambda = \gamma^\varphi, \gamma^4\gamma^\varphi, \gamma^\mu\gamma^\nu, \gamma^\varphi\gamma^\mu\gamma^\nu; \tag{3.39}$$

(b.b) $$C_\mu = \tilde{C}_\mu(\mu)/\tilde{\mathscr{F}} + \tilde{D}_\mu(\nu), \quad C_\nu = \tilde{D}_\nu(\nu),$$
$$C_\varphi = \tilde{D}_\varphi(\nu); \quad k^\varphi = 0; \tag{3.40}$$
$$\hat{K}_\mu = \gamma^\mu(\partial_\mu - i\tilde{C}_\mu)\Sigma,$$
$$\hat{K}_\nu = \tilde{\mathscr{F}}\{-i\gamma^\mu\tilde{D}_\mu + \gamma^\nu(\partial_\nu/\tilde{\mathscr{F}} - i\tilde{D}_\nu) - i\gamma^\varphi\tilde{D}_\varphi + \gamma^\mu\gamma^\nu\gamma^\varphi\epsilon\}\Sigma; \tag{3.41}$$
$$\Sigma = \gamma^\nu, \gamma^\nu\gamma^\varphi, \gamma^\nu\gamma^4, \gamma^\nu\gamma^\varphi\gamma^4. \tag{3.42}$$

Other variants of the scheme of separation with pairwise separation as the first step are identical to the above or lead to weaker conditions on the vector potential. Using the separation scheme outlined above, we encounter the following cases.

(a) $$A_\mu = C_\mu(\mu,\nu)/\tilde{\mathscr{F}}, \quad A_\nu = B_\nu(\nu,\tau),$$
$$A_\varphi = C_\varphi(\mu,\nu)/\tilde{\mathscr{F}}, \quad A_4 = B_4(\nu,\tau); \tag{3.43}$$
$$\hat{K}_{\nu\tau} = \tilde{\mathscr{F}}\{\gamma^\nu(\partial_\nu/\tilde{\mathscr{F}} - iB_\nu) + \gamma^4(\partial_\tau - iB_4) + m_0\}\gamma^\nu\gamma^4,$$
$$\hat{K}_{\mu\varphi} = \{\gamma^\mu(\partial_\mu - iC_\mu) + \gamma^\varphi(\partial_\varphi/\tilde{F} - iC_\varphi)\}\gamma^\nu\gamma^4; \tag{3.44}$$

(b) $$B_\nu = \tilde{B}_\nu(\nu), \quad B_4 = \tilde{A}_4(\tau) + \tilde{B}_4(\nu); \tag{3.45}$$
$$\hat{K}_\tau = (\partial_\tau - i\tilde{A}_4)\gamma^4\gamma^\varphi\gamma^\mu\gamma^\nu,$$
$$\hat{K}_\nu = \{\gamma^\nu(\partial_\nu/\tilde{\mathscr{F}} - i\tilde{B}_\nu) - i\gamma^4\tilde{B}_4 - \gamma^\nu\gamma^4 k^{\mu\varphi}/\tilde{\mathscr{F}} + m_0\}\gamma^\mu\gamma^\nu\gamma^\varphi; \tag{3.46}$$

(c) $$C_\mu = \tilde{C}_\mu(\mu), \quad C_\varphi = \tilde{C}_\varphi(\mu) + \tilde{D}_\varphi(\varphi)/\tilde{F}; \tag{3.47}$$

$$\hat{K}_\varphi = \gamma^\varphi(\partial_\varphi - i\tilde{D}_\varphi)\gamma^\nu\gamma^4\Lambda,$$
$$\hat{K}_\mu = \tilde{F}\{\gamma^\mu(\partial_\mu - i\tilde{C}_\mu) - i\gamma^\varphi\tilde{C}_\varphi + \gamma^\nu\gamma^4 k^{\mu\varphi}\}\gamma^\nu\gamma^4\Lambda; \tag{3.48}$$

$$\Lambda = \gamma^{\varphi}, \quad \gamma^{A}\gamma^{\varphi}\gamma^{\nu}. \tag{3.49}$$

Notice that conditions (3.29), (3.37), and (3.40) are included in (3.27), (3.32), and (3.34). Thus we have the following theorem.

Theorem 8: In order to separate all the variables in Eq. (3.26) under conditions (3.31) it is necessary and sufficient that the components of the vector potential satisfy conditions (3.27), (3.32), (3.34) or (3.43), (3.45), (3.47).

IV. SPECIAL CASES OF SEPARATION

As has been shown in Sec. III, the presence of Lame's functions connected with the curvilinear coordinates leads to additional difficulties in the separation of variables in the Dirac equation and the requirements (3.21) and (3.31) must be satisfied in order to separate all the variables in Eqs. (3.3) and (3.26), respectively. Therefore the results of Sec. III do not contain the oblate spheroidal coordinates successfully used by other authors.[2–4,11] As will be shown in this section the separation of variables in the Dirac equation may be realized even if the conditions (3.21) and (3.31) are not satisfied by means of some additional similarity transformation. It is advisable first to consider the most complex orthogonal curvilinear coordinates.

A. Coordinates μ, ν, φ, τ, $[x=f(\mu,\nu)\cos\varphi, y=f\sin\varphi, z=g(\mu,\nu)]$

The free Dirac equation in the coordinates μ, ν, φ, and τ in the diagonal gauge tetrad has the form

$$\left\{\gamma^{\mu}\frac{\partial_{\mu}}{h_1} + \gamma^{\nu}\frac{\partial_{\nu}}{h_2} + \gamma^{\varphi}\frac{\partial_{\varphi}}{h_3} + \gamma^{A}\partial_{\tau} + m_0\right\}\tilde{\Psi} = 0. \tag{4.1}$$

As Lame's functions for all the known orthogonal curvilinear coordinates depend on a maximum of two variables the unknown additional similarity transformation must depend, in general, on the same two variables. So we make the transformation

$$\gamma \to \hat{S}^{-1}\gamma\hat{S}, \quad \Phi = \hat{S}^{-1}\Psi, \quad \hat{S}\hat{S}^{-1} = I; \tag{4.2}$$

$$\hat{S} = \exp(\gamma^{A}\gamma^{\varphi}\theta_1/2)\exp(\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A}\theta_2/2); \tag{4.3}$$

$$\theta_1 = \theta_1(\mu,\nu), \quad \theta_2 = \theta_2(\mu,\nu). \tag{4.4}$$

It is easy to see that

$$\gamma^{\mu}\hat{S}^{-1}\partial_{\mu}\hat{S} + \gamma^{\nu}\hat{S}^{-1}\partial_{\nu}\hat{S} = 0, \tag{4.5}$$

$$\frac{\partial}{\partial\nu}\theta_1 = \frac{\partial}{\partial\mu}\theta_2, \quad \frac{\partial}{\partial\mu}\theta_1 = -\frac{\partial}{\partial\nu}\theta_2. \tag{4.6}$$

Equation (4.1), after such a transformation, takes the form

$$\left\{\gamma^{\mu}\frac{\partial_{\mu}}{h} + \gamma^{\nu}\frac{\partial_{\nu}}{h} + \exp(\gamma^{A}\gamma^{\varphi}\theta_1)\left(\gamma^{\varphi}\frac{\partial_{\varphi}}{h_3} + \gamma^{A}\partial_{\tau}\right) + m_0\exp(\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A}\theta_2)\right\}\Phi = 0. \tag{4.7}$$

Here we have taken into account that, because of the orthogonality of coordinates, $h_1 = h_2 = h$.

The separation of variables in Eq. (4.7) is possible if

$$\exp(\gamma^{A}\gamma^{\varphi}\theta_1) = (a + b\gamma^{A}\gamma^{\varphi})/h, \tag{4.8}$$

$$\exp(\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A}) = (c + d\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A})/\hbar, \tag{4.9}$$

$$a^2 - b^2 = c^2 + d^2 = h^2. \tag{4.10}$$

Here a, b, c, and d are functions of separated variables.

Indeed now we have from (4.7)

$$\left\{\gamma^{\mu}\partial_{\mu} + \gamma^{\nu}\partial_{\nu} + (a + b\gamma^{A}\gamma^{\varphi})\left(\gamma^{\varphi}\frac{\partial_{\varphi}}{h_3} + \gamma^{A}\partial_{\tau}\right) + m_0(c + d\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A})\right\}\Phi = 0, \tag{4.11}$$

and the separation is possible if

$$h_3 = ab, \tag{4.12}$$

where each of the functions a and b depends only on one variable (μ or ν).

In the case when the Dirac equation contains the vector potential, in order not to disturb the conditions of separability it is necessary to require

$$A_{\mu} = A_{\nu} = 0. \tag{4.13}$$

Further we shall study Eq. (4.11), introducing into it the vector potential taking into account (4.12) and (4.13). Then we have

$$\{\gamma^{\mu}\partial_{\mu} + \gamma^{\nu}\partial_{\nu} + ((1/b) + (1/a)\gamma^{A}\gamma^{\varphi})\gamma^{\varphi}\partial_{\varphi} + i(a + b\gamma^{A}\gamma^{\varphi})(\gamma^{\varphi}A_{\varphi} + \gamma^{A}A_4) + (a + b\gamma^{A}\gamma^{\varphi})\gamma^{A}\partial_{\tau} + m_0(c + d\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A})\}\Phi = 0. \tag{4.14}$$

Demanding

$$(a + b\gamma^{A}\gamma^{\varphi})\gamma^{A}A_{\varphi} + (a + b\gamma^{A}\gamma^{\varphi})\gamma^{A}A_4 = \gamma^{\varphi}(aA_{\varphi} - bA_4) + \gamma^{A}(bA_{\varphi} + aA_4) = \gamma^{\varphi}\xi + \gamma^{A}\zeta; \tag{4.15}$$

$$aA_{\varphi} - bA_4 = \xi, \quad bA_{\varphi} + aA_4 = \zeta; \tag{4.16}$$

$$A_{\varphi} = (a\xi + b\zeta)/(a^2 + b^2), \quad A_4 = (a\zeta - b\xi)/(a^2 + b^2), \tag{4.17}$$

Dirac equation takes the form

$$\{\gamma^{\mu}\partial_{\mu} + \gamma^{\nu}\partial_{\nu} + \gamma^{\varphi}(\partial_{\varphi}/b + b\partial_{\tau} - i\xi) + \gamma^{A}(\partial_{\varphi}/a + a\partial_4 - i\zeta) + m_0(c + d\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A})\}\Phi = 0, \tag{4.18}$$

which allows the complete separation of variables in the following cases:

(a) $a = a(\mu), \quad b = b(\nu), \quad c = c(\mu), \quad d = d(\nu);$ (4.19)

$$\hat{A} = \{\gamma^{\mu}\partial_{\mu} + \gamma^{A}(\partial_{\varphi}/a + a\partial_{\tau} - i\zeta) + m_0 c\}\Gamma,$$

$$\hat{B} = \{\gamma^{\nu}\partial_{\nu} + \gamma^{\varphi}(\partial_{\varphi}/b + b\partial_{\tau} - i\xi) + m_0 d\gamma^{\mu}\gamma^{\nu}\gamma^{\varphi}\gamma^{A}\}\Gamma; \tag{4.20}$$

$$\Gamma = \gamma^{\mu}\gamma^{A}, \quad [\hat{A},\hat{B}]_{-} = 0; \tag{4.21}$$

(b) $a = a(\nu), \quad b = b(\mu), \quad c = c(\mu), \quad d = d(\nu);$ (4.22)

$$\hat{A} = \{\gamma^\mu \partial_\mu + \gamma^\varphi(\partial_\varphi/b + b\,\partial_\tau - i\xi) + m_0 c\}\Gamma,$$
$$\hat{B} = \{\gamma^\nu \partial_\nu + \gamma^A(\partial_\varphi/a + a\,\partial_\tau - i\zeta) + m_0 d\gamma^\mu\gamma^\nu\gamma^\varphi\gamma^A\}\Gamma; \quad (4.23)$$

$$\Gamma = \gamma^\mu\gamma^\varphi, \quad [\hat{A},\hat{B}]_- = 0; \quad (4.24)$$

(c) $$a = a(\mu), \quad b = b(\nu), \quad c = c(\nu), \quad d = d(\mu); \quad (4.25)$$

$$\hat{A} = \{\gamma^\mu \partial_\mu + \gamma^A(\partial_\varphi/a + a\,\partial_\tau - i\xi) + m_0 d\gamma^\mu\gamma^\nu\gamma^\varphi\gamma^A\}\Gamma,$$
$$\hat{B} = \{\gamma^\nu \partial_\nu + \gamma^\varphi(\partial_\varphi/b + b\,\partial_\tau - i\xi) + m_0 c\}\Gamma; \quad (4.26)$$

$$\Gamma = \gamma^\nu\gamma^\varphi, \quad [\hat{A},\hat{B}]_- = 0; \quad (4.27)$$

(d) $$a = a(\nu), \quad b = b(\mu), \quad c = c(\nu), \quad d = d(\mu); \quad (4.28)$$

$$\hat{A} = \{\gamma^\mu \partial_\mu + \gamma^\varphi(\partial_\varphi/b + b\,\partial_\tau - i\xi) + m_0 d\gamma^\mu\gamma^\nu\gamma^\varphi\gamma^A\}\Gamma,$$
$$\hat{B} = \{\gamma^\nu \partial_\nu + \gamma^A(\partial_\varphi/a + a\,\partial_\tau - i\zeta) + m_0 c\}\Gamma; \quad (4.29)$$

$$\Gamma = \gamma^\varphi\gamma^A, \quad [\hat{A},\hat{B}]_- = 0. \quad (4.30)$$

Two well known systems of coordinates allow us to realize these conditions of separation.

Namely, (b) is realized in the oblate spheroidal coordinates that have been used by Cook in the scalar fields,[4] by Bagrov *et al.* in the electromagnetic fields,[3] and by Chandrasekhar in the gravitational Kerr's field.[5,6]

Case (c) is realized in the prolate spheroidal coordinates.

Thus we have

(b) $$x = \alpha \sin\mu \cosh\nu \cos\varphi,$$
$$y = \alpha \sin\mu \cosh\nu \sin\varphi, \quad (4.31)$$
$$z = \alpha \cos\mu \sinh\nu;$$

$$h_1 = h_2 = \alpha(\cosh^2\nu - \sin^2\mu)^{1/2},$$
$$h_3 = \alpha \sin\mu \cosh\nu; \quad (4.32)$$

$$a = \alpha\cosh\nu, \quad b = \alpha\sin\mu,$$
$$c = \alpha\cos\mu, \quad d = \alpha\sinh\nu; \quad (4.33)$$

$$\theta_1 = \operatorname{arctanh}((\sin\mu)/(\cosh\mu)),$$
$$\theta_2 = \arctan((\sinh\nu)/(\cos\mu)); \quad (4.34)$$

$$\hat{A} = \{\gamma^\mu \partial_\mu + \gamma^\varphi(\partial_\varphi/\sin\mu + \alpha\sin\mu\,\partial_\tau - i\xi) + m_0\alpha\cos\mu\}\gamma^\mu\gamma^\varphi,$$

$$\hat{B} = \{\gamma^\nu \partial_\nu + \gamma^A(\partial_\varphi/\cosh\nu + \alpha\cosh\nu\,\partial_\tau - \mu i\zeta) + m_0\alpha\sinh\nu\,\gamma^\mu\gamma^\nu\gamma^\varphi\gamma^A\}\gamma^\mu\gamma^\varphi; \quad (4.35)$$

(c) $$x = \alpha\cos\mu\sinh\nu\cos\varphi,$$
$$y = \alpha\cos\mu\sinh\nu\sin\varphi, \quad (4.36)$$
$$z = \alpha\sin\mu\cosh\nu;$$

$$h_1 = h_2 = \alpha(\sinh^2\nu + \cos^2\mu)^{1/2},$$
$$h_3 = \alpha\cos\mu\sinh\nu; \quad (4.37)$$

$$a = \alpha\cos\mu, \quad b = i\alpha\sinh\nu,$$
$$c = \alpha\cosh\nu, \quad d = i\alpha\sin\mu; \quad (4.38)$$

$$\theta_1 = i\arctan((\sinh\nu)/(\cos\mu)),$$
$$\theta_2 = i\operatorname{arctanh}((\sin\mu)/(\cosh\nu)); \quad (4.39)$$

$$\hat{A} = \{\gamma^\mu \partial_\mu + \gamma^A(\partial_\varphi/\cos\mu + \alpha\cos\mu\,\partial_\tau - i\xi) + im_0\sin\mu\,\gamma^\mu\gamma^\nu\gamma^\varphi\gamma^A\alpha\}\gamma^\nu\gamma^\varphi,$$
$$\hat{B} = \{\gamma^\nu \partial_\nu + \gamma^\varphi(\partial_\varphi/(i\sinh\nu) + i\alpha\sinh\nu - i\xi) + m_0\alpha\cosh\nu\}\gamma^\nu\gamma^\varphi. \quad (4.40)$$

The cases (a) and (d) are equivalent to (b) and (c) in the sense of the change $\mu \rightleftarrows \nu$. Finally the same results may be deduced to do the redefining of θ_1 and θ_2 in the sense of the symmetry $\theta_1 \rightleftarrows \theta_2$.

Returning to (4.16) we notice the configurations of the vector fields A_φ and A_4 allows the complete separation of variables in the Dirac equation according to (4.19)–(4.40).

B. Coordinates $\mu, \nu, z, \tau, [x=f(\mu,\nu), y=g(\mu,\nu)]$

The Dirac equation in the absence of fields in the cylindrical orthogonal coordinates μ, ν, and z in the diagonal tetrad gauge takes the form

$$\{(\gamma^\mu/h)\partial_\mu + (\gamma^\nu/h)\partial_\nu + \gamma^z\partial_z + \gamma^A\partial_\tau + m_0\}\tilde{\Psi} = 0. \quad (4.41)$$

Here the pairwise separation of variables is trivial:

$$\{((\gamma^\mu/h)\partial_\mu + (\gamma^\nu/h)\partial_\nu)\gamma^z\gamma^A\}\Phi = -k\Phi, \quad (4.42)$$
$$\{(\gamma^z\partial_z + \gamma^A\partial_\tau + m_0)\gamma^z\gamma^A\}\Phi = k\Phi, \quad (4.43)$$
$$\Phi = \gamma^z\gamma^A\tilde{\Psi}, \quad (4.44)$$

where k is an eigenvalue of the operators.

If we introduce the vector potential, further separation is possible only under the conditions

$$A_z = A_z(\tau), \quad A_4 = 0 \quad \text{or} \quad A_z = 0, \quad A_4 = A_4(z). \quad (4.45)$$

Because of the separation between μ and ν in (4.42) we can again perform the similarity transformation (4.2)–(4.4) because here we again have Lame's functions depending on μ and ν. Then we have

$$\{((\gamma^\mu/h)\partial_\mu + (\gamma^\nu/h)\partial_\nu)\gamma^z\gamma^A + k\exp(\gamma^\mu\gamma^\nu\gamma^z\gamma^A\theta_2)\}\Phi = 0. \quad (4.46)$$

In order to separate μ and ν in (4.46) we must require

$$\exp(\gamma^\mu\gamma^\nu\gamma^z\gamma^A\theta_2) = (a + b\gamma^\mu\gamma^\nu\gamma^z\gamma^A)/h, \quad (4.47)$$
$$a = a(x^i), \quad b = b(x^j), \quad x^i \neq x^j. \quad (4.48)$$

The separation is realized if

(a) $$a = a(\mu), \quad b = b(\nu); \quad (4.49)$$
$$\hat{A} = (\gamma^\mu\gamma^z\gamma^A\partial_\mu + ak)\gamma^\nu\gamma^z, \quad (4.50)$$
$$\hat{B} = (\gamma^\nu\gamma^z\gamma^A\partial_\nu + bk\gamma^\mu\gamma^\nu\gamma^z\gamma^A)\gamma^\nu\gamma^z; \quad (4.51)$$
$$[\hat{A},\hat{B}]_- = 0; \quad (4.52)$$

(b) $$a = a(\nu), \quad b = b(\mu); \quad (4.53)$$

$$\hat{A} = (-\gamma^4 \partial_\mu + bk\gamma^z\gamma^4), \tag{4.54}$$

$$\hat{B} = (\gamma^\mu\gamma^\nu\gamma^4 \partial_\nu + ak\gamma^\mu\gamma^z), \tag{4.55}$$

$$[\hat{A},\hat{B}] = 0. \tag{4.56}$$

Requirements (a) are satisfied by the parabolic cylindrical and elliptic cylindrical coordinates. Namely,

(a.a) $$x = (\mu^2 - \nu^2)/2, \quad y = \mu\nu; \tag{4.57}$$

$$a = \mu, \quad b = \nu; \tag{4.58}$$

$$\theta_1 = \tfrac{1}{2}\ln(\nu^2 + \mu^2), \quad \theta_2 = \arctan\left(\frac{\nu}{\mu}\right); \tag{4.59}$$

$$\hat{A} = \gamma^\mu\gamma^4\gamma^\nu \partial_\mu + k\mu\gamma^\nu\gamma^z, \quad \hat{B} = -\gamma^4 \partial_\nu + k\nu\gamma^4\gamma^\mu. \tag{4.60}$$

(a.b) $$x = \alpha \sin\mu \cosh\nu, \quad y = \alpha \cos\mu \sinh\nu; \tag{4.61}$$

$$a = \alpha\cos\mu, \quad b = \alpha \sinh\nu; \tag{4.62}$$

$$\theta_1 = \arctan\left(\frac{\sinh\nu}{\cos\mu}\right), \quad \theta_2 = \operatorname{arctanh}\left(\frac{\sin\mu}{\cosh\nu}\right); \tag{4.63}$$

$$\hat{A} = \gamma^\mu\gamma^4\gamma^z \partial_\mu + k\alpha\cos\mu\gamma^\nu\gamma^z, \quad \hat{B} = -\gamma^4 \partial_\nu + \alpha k \sinh\nu\gamma^4\gamma^\mu. \tag{4.64}$$

Case (b), after the redetermination of variables μ and ν in the sense of symmetry $\mu \rightleftarrows \nu$, again leads to the same results.

V. DISCUSSION

The results of Secs. II–IV contain all the possibilities for separation of variables in the Dirac equation in the presence of vector fields in the framework of the method of noncommuting first-order matrix differential operators. All the results known to us using the first-order operators may be found in our scheme. Indeed the first-order symmetry operators of Ref. 2 are equivalent to ours within unitary transformation if the external field is removed in our results. The similarity of the results of Secs. II and III to the corresponding results of Ref. 4 can be seen. The same may be noted about a series of other results.

Notice that in the detailed analysis of the separation of the variables in the free Dirac equation[2] the equivalence of information contained in the second- and first-order operators has been demonstrated. In the presence of vector fields according to our consideration it is evident from (2.8) that

$$[\hat{K}_\alpha^{(1)},\hat{K}_\beta^{(1)}] \equiv 0 \Rightarrow [\hat{K}_\alpha^{(2)},\hat{K}_\beta^{(2)}] \equiv 0, \quad \hat{K}_\alpha^{(2)} = \hat{K}_\alpha^2, \quad \hat{K}_\beta^{(2)} = \hat{K}_\beta^2. \tag{5.1}$$

However, if we have

$$[\hat{K}_\alpha^{(2)},\hat{K}_\beta^{(2)}]_- = 0, \quad K_\alpha^{(2)} = \hat{K}_{\alpha 1}^{(1)}\hat{K}_{\alpha 2}^{(1)}, \quad \hat{K}_\beta^{(2)} = \hat{K}_{\beta 1}^{(1)}\hat{K}_{\beta 2}^{(1)}, \tag{5.2}$$

it is not obligatory that

$$[\hat{K}_{\alpha i}^{(1)},\hat{K}_{\beta j}^{(1)}]_- = 0, \quad (i,j = 1,2). \tag{5.3}$$

The upper indices in parentheses denote the differential order of operators.

For example, for the KGF equation in the absence of fields we have

$$(\partial_i^2 + \partial_j^2 + \partial_m^2 + \partial_n^2 + m_0^2)\Psi = 0, \tag{5.4}$$

$$[\partial_k^2,\partial_l^2]_- = 0, \quad k,l = i,j,m,n. \tag{5.5}$$

For the corresponding Dirac equation we have

$$(\gamma^i \partial_i + \gamma^j \partial_j + \gamma^m \partial_m + \gamma^n \partial_n + m_0)\Psi = 0, \tag{5.6}$$

$$[\gamma^k \partial_k, \gamma^l \partial_l] \neq 0. \tag{5.7}$$

However, notice that the dynamical information in both (5.4) and (5.6) is identical if we neglect proper degrees of freedom in the case (5.6). In general, the introduction of fields may disturb this identity. Thus the separation of variables in the Dirac equation in the presence of fields by means of the second-order matrix differential operators requires a special investigation.

The method proposed here may be useful for the investigations of separation of variables in the more general systems of differential equations with partial derivatives in the search for exact solutions.

[1] W. Miller, Jr., *Symmetry and Separation of Variables* (Addison–Wesley, London, 1977).
[2] E. G. Kalnins, W. Miller, Jr., and G. C. Williams, J. Math. Phys. **27**, 1893 (1986).
[3] V. G. Bagrov, D. M. Gitman, I. M. Ternov, V. R. Khalilov, and V. N. Shapovalov, *Exact Solutions of Relativistic Wave Equations* (Nauka, Novosibirsk, 1982), in Russian.
[4] A. H. Cook, Proc. R. Soc. London Ser. A **383**, 247 (1982).
[5] D. Brill and J. Wheeler, Rev. Mod. Phys. **29**, 465 (1957).
[6] S. Chandrasekhar, *The Mathematical Theory of Black Holes* (Oxford U.P., Oxford, 1983).
[7] S. Chandrasekhar, Proc. R. Soc. London Ser. A **349**, 571 (1976).
[8] B. Carter and R. G. McLenaghan, Phys. Rev. D **19**, 1093 (1979).
[9] A. L. Dudley and J. D. Finlay, III, J. Math. Phys. **20**, 311 (1979).
[10] R. Güven, Proc. R. Soc. London Ser. A **356**, 465 (1977).
[11] V. N. Shapovalov, V. G. Bagrov, and G. G. Ekle, *Complete Sets and Separation of Variables in Dirac Equation* (article deposited in All Union Institute of Scientific and Technical Information–VINITI, Moscow, 1976), No. 405–75, in Russian.
[12] I. E. Andrushkevich and G. V. Shishkin, Theor. Math. Phys. **70**, 204 (1987).
[13] V. A. Fock, *Basis of Quantum Mechanics* (Gostekhizdat, Leningrad, 1932), in Russian.

The delta expansion in zero dimensions

Hing Tong Cho[a] and Kimball A. Milton[a]
Department of Physics and Astronomy, The University of Oklahoma, Norman, Oklahoma 73019 and Department of Physics, The Ohio State University, Columbus, Ohio 43210

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 24 January 1989; accepted for publication 3 May 1989)

The recently introduced δ-expansion (or logarithmic-expansion) technique for obtaining nonperturbative information about quantum field theories is reviewed in the zero-dimensional context. There, it is easy to study questions of analytic continuation that arise in the construction of the Feynman rules that generate the δ series. It is found that for six- and higher-point Green's functions, a cancellation occurs among the most divergent terms, and that divergences that arise from summing over an infinite number of internal lines are illusory. The numerical accuracy is studied in some detail: The δ series converges inside a circle of radius one for positive bare mass squared, and diverges if the bare mass squared is negative, but in all cases, low-order Padé approximants are extremely accurate. These general features are expected to hold in higher dimensions, such as four.

I. INTRODUCTION

In a recent series of papers[1-9] we have developed a new *artificial* perturbation technique that can be applied to quantum field theory. The technique relies upon the introduction of an artificial perturbation parameter δ, which describes the exponent of the interaction term. We have established that the Green's functions of the theory can be expressed as series in powers of δ. Moreover, in some model field theories that we have explored, the δ expansion has a finite radius of convergence. This is to be contrasted with the conventional weak-coupling series, which is known to have zero radius of convergence. Even when the δ expansion is only asymptotic or when one needs information outside the circle of convergence, one can extract accurate information from the δ expansion using Padé approximants. Another important feature of the δ expansion is that it does not force a polynomial dependence on the physical coupling constant, as does the weak-coupling expansion, but instead allows a functional dependence on this and other physical parameters of the theory that can be highly nontrivial. The principal disadvantage of the δ expansion is that, even in finite order in the expansion parameter δ, it produces sums over infinite classes of graphs. In order to perform practical calculations in the δ expansion it is essential that one have techniques at hand for evaluating these infinite sums. Techniques for performing these sums have been partly developed in Refs. 2 and 6. We will further elaborate upon these techniques in this paper.

Recently we have developed[5] a new and simplified formulation of the δ expansion that does not require the explicit introduction of the provisional Lagrangian in terms of which the theory was originally formulated. Although quite equivalent to the original scheme, a set of Feynman rules for the δ expansion can be derived that requires a knowledge of only the original Lagrangian and is entirely analogous to the conventional weak-coupling Feynman rules for field theory.

The basic structure of the δ expansion (Feynman diagrams, symmetry factors, summation constraints, etc.) remains unchanged, as one considers field theories in various dimensions. Therefore we expect the techniques and insights developed in the study of zero-dimensional field theory to have a large carryover to more realistic field theories in higher dimensions. For example, the δ-expansion sums for the n-point functions are divergent for $n \geqslant 6$. These sums are Borel summable and we will exhibit techniques for evaluating them. The divergent nature of these sums arises from the forms of the vertices in the Feynman rules for the δ expansion and graph-counting arguments that are independent of the dimensionality of the field theory. We will also see that these divergences are illusory, and arise from the analytic continuation to the simplified Feynman rules. Of course, in higher dimensions, true divergences emerge from the integrals over closed loops; these divergences must be removed by renormalization.[7,8]

We are interested in the self-interacting scalar field theory in d dimensions, defined by the following generating functional:

$$Z = \int D\varphi \exp\left\{\int\left[-\frac{1}{2}\partial_\mu\varphi\,\partial^\mu\varphi - \frac{1}{2}\mu^2\varphi^2 - \lambda M^2\varphi^2(M^{2-d}\varphi^2)^p + J\varphi\right]d^dx\right\}, \tag{1.1}$$

where μ is the bare mass, λ is the bare coupling constant, and M is an arbitrary scale mass, introduced so that the coupling constant is dimensionless. For positive values of the bare-

[a] Permanent address: Department of Physics and Astronomy, The University of Oklahoma, Norman, Oklahoma 73019.

 0022-2488/89/092143-08$02.50

mass squared, the usual way of attacking this field theory is by the weak-coupling expansion, producing a series in powers of λ. Weak-coupling perturbation theory gives an asymptotic expansion for the Green's functions of the field theory.

The δ expansion is obtained by replacing the exponent p in (1.1) by δ, which we regard as a small parameter, and expanding the functional integral in powers of δ:

$$Z=\int D\varphi \exp\left\{\int\left[-\frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi-\frac{1}{2}(\mu^2+2\lambda M^2)\varphi^2 - \lambda M^2\varphi^2[\delta\ln(M^{2-d}\varphi^2)+\cdots]\right]d^dx\right\}. \tag{1.2}$$

In the case when the bare-mass term is negative, $\mu^2\to-\mu^2$, the weak-coupling expansion does not exist because the unperturbed theory is undefined. The δ expansion, however, is well defined as an asymptotic series and can be used to give very accurate results. When the bare mass vanishes, the weak-coupling expansion is also undefined, but the delta expansion remains well defined, and is convergent. This independence of the δ expansion, from constraints imposed by the values of the bare mass, is a powerful advantage.

In this paper we will further illustrate the features of the δ expansion by fully exploring its application in the simple case of scalar field theory in zero dimensions. For this case the generating functional for the field theory reduces to an ordinary integral, which we write as

$$Z\left(\lambda,\frac{\mu^2}{M^2},\delta,J\right)=\frac{m}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}dx\exp\left\{-\frac{\mu^2x^2}{2} - \lambda M^2x^2(M^2x^2)^\delta+Jx\right\}, \tag{1.3}$$

where

$$m^2=\mu^2+2\lambda M^2. \tag{1.4}$$

The normalization has been chosen for convenience so that $Z(J=\delta=0)=1$.

The integral (1.3) converges for all real values of δ if μ^2 and λ are positive. For λ negative the integral (1.3) is undefined, indicating that Z has a cut singularity along the negative real λ axis and the expansion of Z as a series in λ has a zero radius of convergence.

II. DELTA EXPANSION

To obtain the δ expansion we will use the Feynman graph techniques developed in Ref. 5 and summarized here. The rules are an abbreviation of, but equivalent to, the original provisional Lagrangian technique described in Refs. 1 and 2. Although an infinite class of graphs is summed in each order, the rules for calculating any given diagram are similar to the standard rules for a scalar field theory, except for the specific form of the vertex factor and the mass term in the propagator. (For the zero-dimensional theory treated in this paper, the propagator reduces to a simple inverse mass factor).

To obtain the expansion of $G^{(2n)}$, the $2n$-point one-particle-irreducible (1PI) Green's function, to order δ^k, we include all 1PI diagrams with up to k vertices, $2n$ external lines, and *any number of internal lines*. To obtain the contribution of a given j-vertex diagram, to the order δ^k, one uses the power series expansion of each vertex factor $v_{2l}(\delta)$ in the diagram and retains, in the expansion of the full diagram, all terms of order δ^k or lower. For the theory defined in (1.3) the vertex factor for vertices with $2l$ external legs is given by[5]

$$v_{2l}(\delta)=\frac{2^{l+1}}{\sqrt{\pi}}\lambda M^2(m^2)^{l-1}(\delta+1)\delta(\delta-1)\cdots(\delta+2-l)\times\Gamma\left((\delta+\frac{3}{2}\right)\left(\frac{2M^2}{m^2}\right)^\delta=\sum_{k=1}^{\infty}\frac{\delta^k}{k!}v_{2l}^{(k)}. \tag{2.1}$$

Here the "loop integral" that is associated with each closed loop has reduced, because we are in zero dimensions, to the simple propagator factor m^{-2}, where m^2 is defined in (1.4). The expansion of the "interaction" term, $\lambda M^2x^2(M^2x^2)^\delta$, in (1.3) about the point $\delta=0$, introduces a term λM^2x^2, which combines with the "free Lagrangian" to produce a shift of the bare-mass term $\mu^2x^2/2$ by this amount. Therefore the "propagator" that enters the δ-expansion calculations is $(\mu^2+2\lambda M^2)^{-1}$.

We shall calculate the $2n$-point functions for the zero-dimensional field theory defined by (1.3) for $n=0,1,2,3$, through second order in the δ expansion, by summing the δ-expansion Feynman diagrams. These results will be compared with the analytic results obtained by directly evaluating the integrals that result from expanding (1.3) in powers of J and δ. In all these calculations we include only the one-particle-irreducible graphs.

A. Zero-point function

The Feynman graphs that contribute to the δ expansion of the zero-point function through order δ^2 are given in Fig. 1. The contribution of the graph in Fig. 1(a) is given by

$$-\delta v_0^{(1)}-\frac{\delta^2}{2}v_0^{(2)} = -\frac{\delta\lambda M^2}{m^2}\left\{\psi\left(\frac{3}{2}\right)+\ln\left(\frac{2M^2}{m^2}\right)\right\} - \frac{\delta^2}{2}\frac{M^2}{m^2}\left\{\psi'\left(\frac{3}{2}\right)+\left[\psi\left(\frac{3}{2}\right)+\ln\left(\frac{2M^2}{m^2}\right)\right]^2\right\}. \tag{2.2}$$

The diagrams in Fig. 1(b) contribute

$$\frac{1}{2}\sum_{l=1}^{\infty}\delta^2\frac{1}{(m^2)^{2l}}\frac{1}{(2l)!}[v_{2l}^{(1)}]^2 = \frac{\delta^2\lambda^2M^4}{m^4}\left[\psi\left(\frac{3}{2}\right)+1+\ln\left(\frac{2M^2}{m}\right)\right]^2 + \frac{\delta^2\lambda^2M^4\sqrt{\pi}}{2m^4}\sum_{l=2}^{\infty}\frac{\Gamma(l-1)}{l(l-1)\Gamma(l+\frac{1}{2})}. \tag{2.3}$$

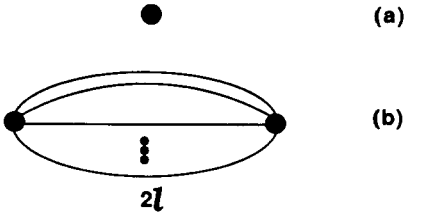


FIG. 1. Feynman graphs that contribute to the zero-point function through $O(\delta^2)$. Here, the vertices are obtained by expanding $v_{2l}(\delta)$ [Eq. (2.1)] out to the requisite order in δ.

To evaluate the sum that appears in the above expression we recognize the ratio of gamma functions that appears in the sum as a beta function,

$$\Gamma(l-1)/\Gamma(l+\tfrac{1}{2}) = B(l-1,\tfrac{3}{2})/\Gamma(\tfrac{3}{2}), \tag{2.4}$$

and use the integral representation for the beta function

$$B(a,b) = \int_0^1 dx(1-x)^{a-1}x^{b-1}, \quad \mathrm{Re}(a,b) > 0. \tag{2.5}$$

Then we separate the factor $l(l-1)$ by partial fractions, perform the sums under the integral, and evaluate the integral to obtain

$$\sum_{l=2}^{\infty} \frac{\Gamma(l-1)}{l(l-1)\Gamma(l+\frac{1}{2})} = \frac{2}{\pi}\left[\frac{3}{2}\psi'\left(\frac{3}{2}\right) - 1\right].$$

The same technique is used to evaluate all the sums that appear in the expressions for the two-point and four-point functions. In the six-point and higher functions the δ expansion produces sums that are formally divergent but Borel summable. These are treated below by a variation of this technique.

Our final expression for the zero-point function through order δ^2 is

$$\begin{aligned}\ln Z = & -\frac{\delta\lambda M^2}{m^2}\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right)\right] \\ & -\frac{\delta^2\lambda M^2}{2m^2}\left\{\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m}\right)\right]^2 + \psi'\left(\frac{3}{2}\right)\right\} \\ & +\frac{\delta^2\lambda^2 M^4}{m^4}\left\{\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right]^2 \right. \\ & \left. +\frac{3}{2}\psi'\left(\frac{3}{2}\right) - 1\right\}. \end{aligned} \tag{2.6}$$

Note that this reduces to the case treated in Ref. 2, when $M^2 = 1$, $\lambda = 1$, and $m^2 = 2$.

B. Two-point function

The δ-expansion Feynman diagrams that contribute to the two-point function through order δ^2 are given in Fig. 2. The graph in Fig. 2(a) is of order δ^0 and its contribution is just the propagator term m^{-2}. The graph in Fig. 2(b) contributes

$$\begin{aligned} & -\delta v_2^{(1)} - \frac{\delta^2}{2} v_2^{(2)} \\ & = -2\delta\lambda M^2\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] \\ & \quad -\lambda M^2\delta^2\left\{\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right]^2 \right. \\ & \quad \left. +\psi'\left(\frac{3}{2}\right) - 1\right\}. \end{aligned} \tag{2.7}$$

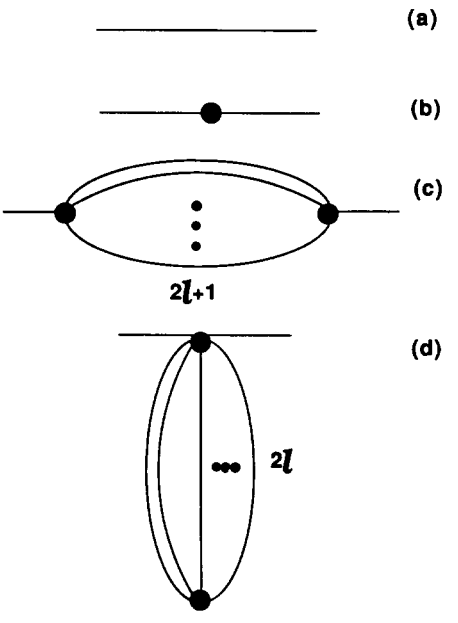


FIG. 2. Feynman graphs that contribute to the two-point function through $O(\delta^2)$.

The contribution of the graphs in Fig. 2(c) is given by

$$\begin{aligned} & \sum_{l=1}^{\infty} \frac{1}{(m^2)^{2l+1}} \frac{1}{(2l+1)!}[v_{2l+2}^{(1)}]^2 \\ & = \frac{2\delta^2\lambda^2 M^4}{m^2}\sqrt{\pi}\sum_{l=2}^{\infty}\frac{\Gamma(l-1)}{(l-1)\Gamma(l+\frac{1}{2})} \\ & = \frac{4\delta^2\lambda^2 M^4}{m^2}\psi'\left(\frac{3}{2}\right). \end{aligned} \tag{2.8}$$

(The sum here is given in Ref. 3.) The contributions of the graphs in Fig. 2(d) are

$$\begin{aligned} & \frac{1}{2(m^2)^2} v_4^{(1)} v_2^{(1)} + \sum_{l=2}^{\infty} \frac{1}{(2l)!(m^2)^{2l}} v_{2l+2}^{(1)} v_{2l}^{(1)} \\ & = \frac{4\delta^2\lambda^2 M^4}{m^2}\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] \\ & \quad -\frac{2\delta^2\lambda^2 M^4}{m^2}\sqrt{\pi}\sum_{l=2}^{\infty}\frac{\Gamma(l-1)}{l\Gamma(l+\frac{1}{2})} \\ & = \frac{4\delta^2\lambda^2 M^4}{m^2}\left\{\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] \right. \\ & \quad \left. -1 + \frac{1}{2}\psi'\left(\frac{3}{2}\right)\right\}. \end{aligned} \tag{2.9}$$

The final expression for the δ expansion of the two-point function through order δ^2 is the sum of (2.7)–(2.9):

$$\begin{aligned} G^{(2)} = & -2\delta\lambda M^2\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] \\ & -\delta^2\lambda M^2\left\{\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right]^2 + \psi'\left(\frac{3}{2}\right) - 1\right\} \\ & +\frac{4\delta^2\lambda^2 M^4}{m^2}\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right) + \frac{3}{2}\psi'\left(\frac{3}{2}\right)\right]. \end{aligned} \tag{2.10}$$

C. Four-point function

The δ-expansion graphs that contribute to the four-point function through order δ^2 are given in Fig. 3. The diagram in Fig. 3(a) contributes

$$-\delta v_4^{(1)} - \frac{\delta^2}{2!} v_4^{(2)} = -4\delta\lambda M^2 m^2 - 4\delta^2\lambda M^2 m^2 \times\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right]. \tag{2.11}$$

The graphs in Fig. 3(b) contribute

$$\sum_{l=1}^{\infty} \frac{4\delta^2}{(m^2)^{2l+1}(2l+1)!} v_{2l+2}^{(1)} v_{2l+4}^{(1)} = -2^4\delta^2\lambda^2 M^4\sqrt{\pi}\sum_{l=1}^{\infty}\frac{\Gamma(l)}{\Gamma(l+\frac{3}{2})} = -2^6\delta^2\lambda^2 M^4. \tag{2.12}$$

The contribution of the graphs in Fig. 3(c) is

$$\sum_{l=1}^{\infty} \frac{3\delta^2}{(m^2)^{2l}(2l)!}\left[v_{2l+2}^{(1)}\right]^2 = 12\delta^2\lambda^2 M^4\sqrt{\pi}\sum_{l=1}^{\infty}\frac{\Gamma(l)}{l\Gamma(l+\frac{1}{2})} = 12\delta^2\lambda^2 M^4\psi'\left(\frac{1}{2}\right). \tag{2.13}$$

The graphs in Fig. 3(d) contribute

$$\sum_{l=1}^{\infty} \frac{\delta^2}{(m^2)^{2l}(2l)!} v_{2l}^{(1)} v_{2l+4}^{(1)} = -8\delta^2\lambda^2 M^4\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] + 4\delta^2\lambda^2 M^4\sqrt{\pi}\sum_{l=2}^{\infty}\frac{\Gamma(l-1)}{\Gamma(l+\frac{1}{2})} = -8\delta^2\lambda^2 M^4\left\{\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] - 2\right\}. \tag{2.14}$$

The final expression fo the four-point function through order δ^2 is obtained as the sum of (2.11)–(2.14):

$$G^{(4)} = -4\delta\lambda M^2 m^2 - 4\delta^2\lambda M^2 m^2\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] - \delta^2\lambda^2 M^4 8\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right) + 1 - \frac{3}{2}\psi'\left(\frac{3}{2}\right)\right]. \tag{2.15}$$

D. Six-point function

The δ-expansion graphs that contribute to the six-point function through order δ^2 are given in Fig. 4. The contribution of the graph in Fig. 4(a) is

$$-\delta v_6^{(1)} - \frac{\delta^2}{2} v_6^{(2)} = 8\delta\lambda M^2 m^4 + 8\delta^2\lambda M^2 m^4 \times\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right)\right]. \tag{2.16}$$

The graphs in Fig. 4(b) contribute

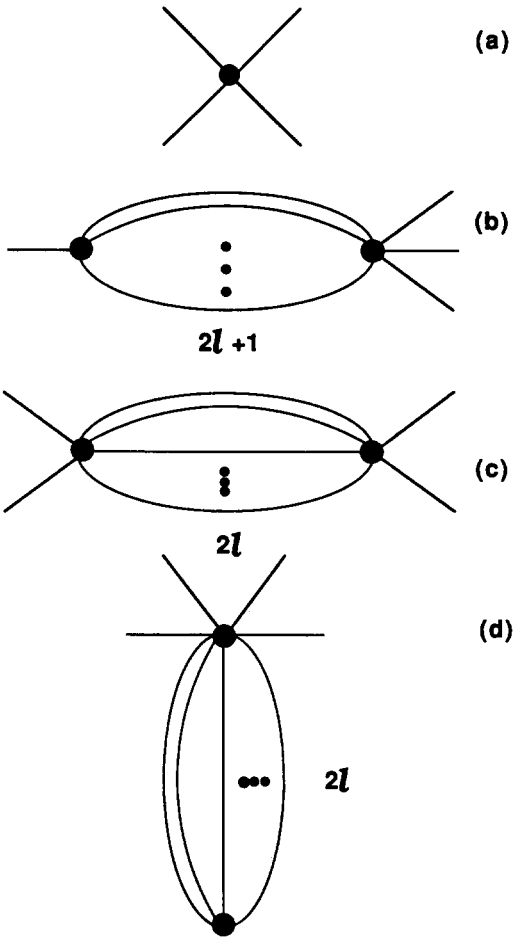


FIG. 3. Feynman graphs that contribute to the four-point function through $O(\delta^2)$.

$$\frac{\delta^2}{2(m^2)^2} v_2^{(1)} v_8^{(1)} + \sum_{l=2}^{\infty}\frac{\delta^2}{(m^2)^{2l}(2l)!} v_{2l}^{(1)} v_{2l+6}^{(1)} = 2^5\delta^2\lambda^2 M^4 m^2\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] - 2^3\delta^2\lambda^2 M^4 m^2\sqrt{\pi}\sum_{l=2}^{\infty}\frac{(l+1)\Gamma(l-1)}{\Gamma(l+\frac{1}{2})} = 2^5\delta^2\lambda^2 M^4 m^2\left\{\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] - 1\right\}. \tag{2.17}$$

We note that the summand in the expression above behaves asymptotically as $l^{-1/2}$, so the sum is divergent. There are two ways to address this potentially embarrassing problem. First, as will be seen below, the other graphs that contribute to the six-point function in this order also produce divergent sums. The divergences exactly cancel so that the sum of all the δ-expansion graphs to order δ^2 is finite. Second, we can formally sum the divergent expression by using the beta-function technique and ignore the divergence. The (finite) result is the final expression given above.

The graphs in Fig. 4(c) contribute

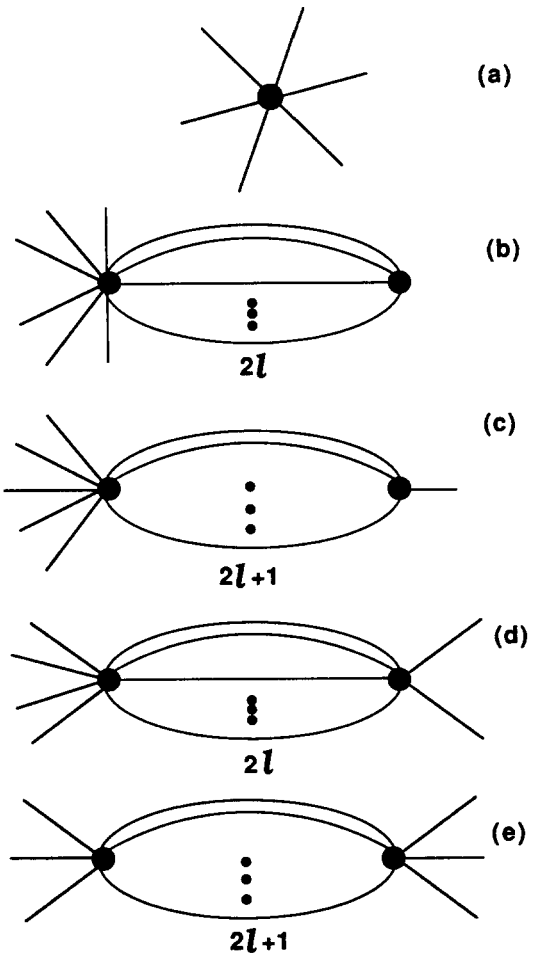


FIG. 4. Feynman graphs that contribute to the six-point function through $O(\delta^2)$.

$$\sum_{l=1}^{\infty} \frac{6\delta^2}{(m^2)^{2l+1}(2l+1)!} v_{2l+2}^{(1)} v_{2l+6}^{(1)} = 48\delta^2\lambda^2 M^4 m^2 \sqrt{\pi} \sum_{l=1}^{\infty} \frac{\Gamma(l)(l+1)}{\Gamma(l+\frac{3}{2})} = 0. \tag{2.18}$$

Again, the final expression above is the result of ignoring the divergence of the series and formally evaluating the sum using the beta-function technique. The graphs in Fig. 4(d) yield

$$\begin{aligned}\sum_{l=1}^{\infty} \frac{15}{(m^2)^{2l}(2l)!} v_{2l+2}^{(1)} v_{2l+4}^{(1)} &= -15\cdot 2^2\delta^2\lambda^2 M^4 m^2 \sqrt{\pi} \sum_{l=1}^{\infty} \frac{\Gamma(l)}{\Gamma(l+\frac{1}{2})} \\ &= 15\cdot 2^4\delta^2\lambda^2 M^4 m^2,\end{aligned} \tag{2.19}$$

while the graphs in Fig. 4(e) give

$$\begin{aligned}\sum_{l=1}^{\infty} \frac{10}{(m^2)^{2l+1}(2l+1)!} [v_{2l+4}^{(1)}]^2 &= 10\cdot 2^3\delta^2\lambda^2 M^4 m^2 \sqrt{\pi} \sum_{l=1}^{\infty} \frac{\Gamma(l+1)}{\Gamma(l+\frac{3}{2})} \\ &= -10\cdot 2^5\delta^2\lambda^2 M^4 m^2.\end{aligned} \tag{2.20}$$

It is easy to see that when the sums that appear in (2.17)–(2.20) are combined, the result is a finite expression.[10] The result is the same as the sum of the analytically continued results found above. Our final expression for the delta expansion of the six-point function through order δ^2 is

$$G^{(6)} = 8\delta\lambda M^2 m^4 + 8\delta^2\lambda M^2 m^4 \left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right)\right] + 2^4\delta^2\lambda^2 M^4 m^2 \left\{2\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right)\right] - 5\right\}. \tag{2.21}$$

III. DIRECT EXPANSION OF THE GENERATING FUNCTION

We now verify that the Feynman rules for the δ expansion used in Sec. II correctly produce an expansion in powers of δ of the connected, one-particle-irreducible Green's functions of the theory defined by the generating function $Z(\lambda,\mu^2/M^2,\delta,J)$. To do this we first recall that $\ln Z$ contains only the connected graphs of the theory and that the connected Green's functions are defined by

$$\Gamma^{(n)} = \left(\frac{\partial}{\partial J}\right)^n \ln Z \Big|_{J=0}. \tag{3.1}$$

The resulting expression for $\Gamma^{(n)}$, however, contains one-particle-reducible graphs that must be removed. Finally, we must remove the external propagator factors to obtain a form that can be compared directly with the results for $G^{(n)}(\lambda,m^2,M^2,\delta)$ obtained in Sec. II.

We first expand Z in powers of J through all orders and in powers of δ through second order. The result is

$$\begin{aligned}Z(J) = \frac{1}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(\sqrt{2}J/m)^{2n}}{(2n)!} \Bigg\{ &\Gamma\left(n+\frac{1}{2}\right) \\ &- \delta\left(\frac{2\lambda M^2}{m^2}\right)\Gamma\left(n+\frac{3}{2}\right)F(n) \\ &- \frac{1}{2}\delta^2\left(\frac{2\lambda M^2}{m^2}\right)\Gamma\left(n+\frac{3}{2}\right) \\ &\times\left[F^2(n) + \psi'\left(n+\frac{3}{2}\right)\right] \\ &+ \frac{1}{2}\delta^2\left(\frac{2\lambda M^2}{m^2}\right)\Gamma\left(n+\frac{5}{2}\right) \\ &\times\left[F^2(n+1) + \psi'\left(n+\frac{5}{2}\right)\right] + \cdots\Bigg\},\end{aligned} \tag{3.2}$$

where $F(n) = \ln(2M^2/m^2) + \psi(n+\frac{3}{2})$. Because $Z(\lambda,\mu^2/M^2,\delta,J)$ contains only even powers of J, only even derivatives of Z survive at $J=0$. One finds the following expressions for $\Gamma^{(n)}$ for $n=2,4,6$ in terms of derivatives of Z at $J=0$:

$$\Gamma^{(2)} = Z''(0)/Z(0), \tag{3.3a}$$

$$\Gamma^{(4)} = \frac{Z''''(0)}{Z(0)} - 3\left(\frac{Z''(0)}{Z(0)}\right)^2, \tag{3.3b}$$

$$\Gamma^{(6)} = \frac{Z^{(6)}(0)}{Z(0)} - 15\frac{Z''(0)}{Z(0)}\left[\frac{Z''''(0)}{Z(0)} - 2\left(\frac{Z''(0)}{Z(0)}\right)^2\right]. \tag{3.3c}$$

Inserting the expansion (3.2) for $Z(J)$ into the expressions (3.3) one obtains the connected Green's functions of the theory as follows. The two-point function is

$$\Gamma^{(2)} = \frac{1}{m^2} - \frac{\delta}{m^2}\left(\frac{2\lambda M^2}{m^2}\right)\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] - \frac{\delta^2}{2m^2}\left(\frac{2\lambda M^2}{m^2}\right)\left[\left(\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right)^2 + \psi'\left(\frac{3}{2}\right) - 1\right] + \frac{\delta^2}{m^2}\left(\frac{2\lambda M^2}{m^2}\right)^2\left[\left(\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right)^2 + \psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right) + \frac{3}{2}\psi'\left(\frac{3}{2}\right)\right]. \tag{3.4}$$

The first term above is simply the bare propagator. In each of the remaining terms there is an extraneous factor of m^{-4} corresponding to the two external propagators. These propagator factors we remove. Then, we note that the coefficient of δ is exactly the same as the coefficient of δ in (2.10), which corresponds to a contribution of the graph in Fig. 2(b). This same term occurs squared in the coefficient of δ^2 in (3.4), where it represents the iteration of the graph in Fig. 2(b) and is a one-particle-reducible term. When this term is removed along with the external propagator factors, (3.4) agrees exactly with the expression for $G^{(2)}$ in (2.10) obtained from the δ expansion.

From (3.2) and (3.3b) the expression for the four-point function is

$$\Gamma^{(4)} = -\frac{2\delta}{m^4}\left(\frac{2\lambda M^2}{m^2}\right) - \frac{2\delta^2}{m^4}\left(\frac{2\lambda M^2}{m^2}\right)\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right)\right] + \frac{6\delta^2}{m^4}\left(\frac{\lambda 2M^2}{m^2}\right)^2\left[\psi\left(\frac{3}{2}\right) + 1 + \ln\left(\frac{2M^2}{m^2}\right) + \frac{1}{2}\psi'\left(\frac{3}{2}\right)\right]. \tag{3.5}$$

There are four extraneous factors of m^{-2} to be removed, corresponding to the four external propagators. Thus $m^8\Gamma^{(4)}$ must agree with the expression for $G^{(4)}$ given in (2.15) when the one-particle-reducible terms are removed. Equation (3.5) contains four such terms, each given by the graph in Fig. 5 with the propagator modification on a different leg. The required modification corresponds to the product of the contributions from the graphs in Figs. 2(b) and 3(a), which are given in (2.7) and (2.11), respectively. When these modifications to (3.5) are made, we obtain exactly the δ-expansion result for $G^{(4)}$, given in (2.15).

From (3.2) and (3.3c) the six-point function is

$$\Gamma^{(6)} = \frac{4\delta}{m^6}\left(\frac{2\lambda M^2}{m^2}\right) + \frac{4\delta^2}{m^6}\left(\frac{2\lambda M^2}{m^2}\right)\left[\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right)\right] - \frac{4\delta^2}{m^6}\left(\frac{2\lambda M^2}{m^2}\right)^2\left[4\left(\psi\left(\frac{3}{2}\right) + \ln\left(\frac{2M^2}{m^2}\right)\right) + 1\right]. \tag{3.6}$$

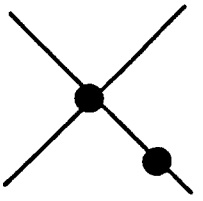


FIG. 5. One-particle-reducible graphs in $\Gamma^{(4)}$.

To remove the six external propagator factors, we consider $m^{12}\Gamma^{(6)}$. From this expression it is necessary to remove the one-particle-reducible terms corresponding to the graphs in Fig. 6. The graph in Fig. 6(a) is the square of the graph in Fig. 3(a) divided by m^2. Its contribution is given by the square of the expression in (2.11) with a combinatorial factor of 6!/3!3!2! corresponding to the various ways of rearranging the external legs. The graph in Fig. 6(b) is the product of the graphs in Figs. 2(b) and 4(a). Its contribution is given by the product of the expressions in (2.7) and (2.16) divided by m^2 with a factor of six for the six possible propagator insertions. When these modifications are made, (3.6) yields an expression for $G^{(6)}$ that agrees exactly with that in (2.21) obtained from the δ expansion.

This completes our demonstration that, for the zero-dimensional field theory defined by (1.3), the Feynman rules for the δ expansion produce exactly the correct expressions for the n-point functions of the theory for $n = 0, 2, 4$, and 6 through order δ^2, for any value of δ. The nontrivial nature of this verification lies principally in the analytic continuation implicit in the Feynman rules, which becomes especially apparent in the six-point function, where the individual classes of graphs diverge. The calculation given above does not address the question of for what values of δ, if any, the δ expansion converges or yields an asymptotic series for the Green's functions of the theory. This issue was addressed in Ref. 2 where we demonstrated, *inter alia*, the convergence of the δ expansion for simple model field theories. We will discuss this issue further in the following section.

IV. CONVERGENCE OF THE δ EXPANSION

In Ref. 2 we studied the convergence of the δ expansion of the generating function Z in (1.3) for zero bare mass,

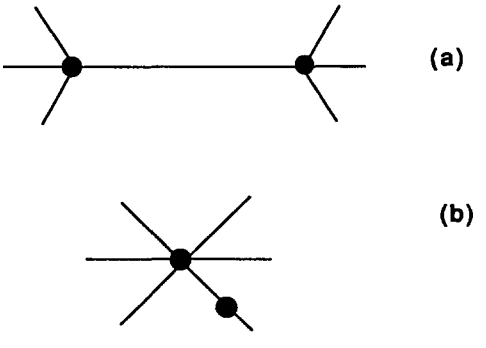


FIG. 6. One-particle-reducible graphs in $\Gamma^{(6)}$.

$\mu^2 = 0$. We found that the series in powers of δ converged inside a circle of radius 1 about the origin, and that both inside and outside this circle, low-order Padé approximants offered a spectacularly good numerical approximation. It was easy to discuss the analytic structure in that case, because a closed form exists for Z:

$$Z(\lambda,0,\delta,0) = (2/\sqrt{\pi})\Gamma[(3+2\delta)/(2+2\delta)]. \quad (4.1)$$

Here we will generalize this discussion to the $\mu \neq 0$ situation, with particular emphasis on what happens when $\mu^2 < 0$.

In a variety of field theory contexts it is interesting to consider the case when the sign of the bare-mass squared term, $\mu^2\varphi^2/2$ in (1.1) is reversed. The quantum potential for a physical system may have more than one minimum. This may give rise to spontaneous symmetry breaking. The study of phase transitions is closely related to the study of the transitions between multiple minima. It is well known that this problem cannot be attacked by weak-coupling techniques. This is well illustrated by the zero-dimensional theory defined by (1.3), where the weak coupling expansion does not exist then, because the unperturbed ($\lambda = 0$) integral does not converge. The negative mass-squared case has been successfully attacked by a variety of nonperturbative techniques, such as the introduction of kink solutions or pseudoparticles.[11] These nonperturbative solutions exhibit an essential singularity in λ at the point $\lambda = 0$, which is, of course, the reason that the weak-coupling solution fails.

With negative mass squared, the generating function (1.3) becomes

$$Z = \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left\{\frac{1}{2}\mu^2x^2 - \lambda M^2x^2(M^2x^2)^\delta + Jx\right\}dx, \quad (4.2)$$

where $m^2 = 2\lambda M^2 - \mu^2$. In Sec. II we pointed out that the lowest term of the δ expansion shifts the bare mass by an amount $2\lambda M^2$, so that the propagator in the δ expansion has a mass term given by $m^2 = -\mu^2 + 2\lambda M^2$. Recall, also, that the parameter M is a totally arbitrary scale[12] mass introduced for convenience so that the coupling λ is dimensionless. Therefore M can always be chosen so that m^2 is positive. (Such a choice of M will not alter the theory at, say, $\delta = 1$, but changes the δ expansion.) Because of this flexibility, the δ expansion is a useful perturbation technique for the "double-well" problem. We will demonstrate this by calculating the δ expansion for zero-dimensional theories defined by (4.2) and comparing the results with explicit numerical integration of (4.2).

We remark here that at $\delta = 0$ the integral (4.2) is convergent (for sufficiently large M), but this is not so for $\delta < 0$. Therefore we must expect that Z, defined by (4.2), has singularities on the negative δ axis, with $\delta = 0$ as a limit point. Accordingly, the δ expansion for (4.2) will not yield a convergent series. Nevertheless we will see that the (asymptotic) series produced by the δ expansion defines Padé approximants that give very accurate numerical results.

First, however, let us turn to the numerical accuracy of the δ expansion for $\mu^2 > 0$. Without loss of generality we may set $M^2 = 1$ and evaluate

TABLE I. The eight-term power series, [4,4] Padé approximant, and the exact value of $Z(\delta)$ [Eq. (4.3)] for $\mu^2 = 1$ and $\lambda = 1$.

δ	$p_8(\delta)$	$pd_{[4,4]}(\delta)$	$Z(\delta)$
0.5	1.046 31	1.046 30	1.046 30
1.0	1.077 19	1.074 36	1.074 36
2.0	1.810 47	1.106 47	1.106 49
5.0	745.176	1.142 53	1.142 85

$$Z(\delta) = m\left(\frac{2}{\pi}\right)^{1/2} \int_0^\infty \exp\left[-\frac{\mu^2}{2}x^2 - \lambda x^{2(1+\delta)}\right]dx. \quad (4.3)$$

In Table I we show representative values of $Z(\delta)$ for the case of $\mu^2 = 1, \lambda = 1$. These are to be compared with the eight-term power series obtained by expanding (4.3) in powers of δ, $p_8(\delta)$, and the [4,4] Padé approximant obtained from that series, $pd_{[4,4]}(\delta)$. It will be seen that the power series is convergent for $|\delta| < 1$ and divergent for $|\delta| > 1$. However, the [4,4] Padé approximant is spectacularly good for all positive values of δ. (It is off by only -0.03% at $\delta = 5$.)

The corresponding results for $\mu^2 < 0$ are quite different. In Table II we show the same quantities for the case $\mu^2 = -1, \lambda = 1$. It is apparent, as expected, that the power series no longer converges for any value of δ. However, the [4,4] Padé approximant remains excellent: in spite of the series being hopelessly divergent, the Padé is off by less than 1.5% at the large value of $\delta = 5$.

It is useful to examine the "potential" that appears in (4.2) in the case $\delta = 1$. That potential is

$$V(x) = -\tfrac{1}{2}\mu^2x^2 + \lambda(M^2x^2)^2. \quad (4.4)$$

The depth of the potential minima is given by

$$D = \mu^4/16\lambda M^2, \quad (4.5)$$

and the location of the minima is $\pm x_0$, where

$$x_0^2M^2 = (1/4\lambda)(\mu^2/M^2). \quad (4.6)$$

Thus the situation illustrated in Table II is that for a shallow well of depth $D = \frac{1}{16}$. The contrast with the case of a deep well, say with $D = 1$, is brought out in Table III, where we compare $Z(\delta)$ with a six-term and a seven-term power series, $p_6(\delta)$, and $p_7(\delta)$, respectively, and with the corresponding [3,3] and [3,4] Padés, $pd_{[3,3]}(\delta)$ and $pd_{[3,4]}(\delta)$.

TABLE II. The eight-term power series, [4,4] Padé approximant, and the exact value of $Z(\delta)$ [Eq. (4.3)] for $\mu^2 = -1$ and $\lambda = 1$.

δ	$p_8(\delta)$	$pd_{[4,4]}(\delta)$	$Z(\delta)$
0.1	0.94808	0.947 90	0.947 90
0.5	137.697	0.883 88	0.883 81
1.0	40109.3	0.873 23	0.872 53
2.0	$1.107\,61\times10^7$	0.883 34	0.879 74
5.0	$1.774\,23\times10^{10}$	0.918 30	0.905 17

TABLE III. The six-term power series, seven-term power series, [3,3] Padé approximant, [3,4] Padé approximant, and the exact value of $Z(\delta)$ [Eq. (4.3)] for $\mu^2 = -16$ and $\lambda = 16$.

δ	$p_6(\delta)$	$p_7(\delta)$	$pd_{[3,3]}(\delta)$	$pd_{[3,4]}(\delta)$	$Z(\delta)$
0.5	2.306 87	2.305 65	2.305 70	2.305 88	2.306 30
1.0	4.337 33	4.184 00	4.233 69	4.245 68	4.280 78
1.5	7.799 45	5.187 09	6.469 74	6.595 55	7.008 06
2.0	15.518	−4.033 26	7.967 41	8.490 63	10.495 54

There the series gives a quite good *asymptotic* approximation, leaving only a relatively small improvement for the Padé approximants.

V. CONCLUSIONS

In this paper we have treated the δ or logarithmic approximation in detail for the simple situation of zero space-time dimension. We believe that many of the features illustrated here have general validity, and will enable new light to be shed on problems in four- and higher-dimensional field theories. Let us summarize those features here.

(1) The terms in the δ series may be readily computed using standard Feynman rules on graphs derived from a "provisional Lagrangian," or, more directly, from graphs having an infinity of vertices, and an unlimited number of internal lines.

(2) The sums over the number of internal lines may be carried out in closed form.

(3) Although the line sums diverge in general for each graph, this divergence is illusory, and at least one cancellation occurs among the most divergent terms.

(4) A simple analytic continuation based on a beta-function representation in any case yields the correct result.

(5) When the bare mass μ satisfies $\mu^2 \geqslant 0$, the δ series converges for $|\delta| < 1$.

(6) When the bare mass μ satisfies $\mu^2 < 0$, the δ series diverges for all δ.

(7) In all cases, low-order Padé approximants are extremely accurate.

ACKNOWLEDGMENTS

We thank Carl Bender for many useful conversations during the course of this work.

This work was supported in part by the U.S. Department of Energy.

[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).
[4] C. M. Bender and K. A. Milton, Phys. Rev. D **38**, 1310 (1988).
[5] S. S. Pinksy and L. M. Simmons, Jr., Phys. Rev. D **38**, 2518 (1988); N. Brown, *ibid.* **38**, 723 (1988).
[6] Integral representations for the zero-, two-, and four-point functions in any number of dimensions are given in C. M. Bender and H. F. Jones, J. Math. Phys. **29**, 2659 (1988).
[7] C. M. Bender and H. J. Jones, Phys. Rev. D **38**, 2526 (1988); I. Yotsuyanagi, Phys. Rev. D **39**, 485 (1989).
[8] H. T. Cho, K. A. Milton, J. Cline, S. S. Pinsky, and L. M. Simmons, Jr., "Triviality of monomial Higgs potentials," Ohio State University preprint No. DOE/ER/01545-420; C. M. Bender, K. A. Milton, S. S. Pinsky, and L. M. Simmons, Jr., "δ expansion for a quantum field theory in the nonperturbative regime," Ohio State University preprint No. DOE/ER/01545-421.
[9] C. M. Bender, K. A. Milton, S. S. Pinsky, and L. M. Simmons, Jr., J. Math. Phys. **30**, 1447 (1989).
[10] A similar cancellation of the most divergent sums occurs in order δ^2 for the eight- and higher-point functions. However, this is not sufficient to insure convergence, as these higher-point Green's functions are successively more divergent. However, these divergences are illusory, and the correct answer, as obtained by the direct use of the provisional Lagrangian (Refs. 1 and 2), is obtained by analytic continuation, as with the six-point function.
[11] E. Gildener and A. Patrascioiu, Phys. Rev. D **16**, 423 (1977), and references therein.
[12] H. F. Jones and M. Monoyios, Imperial College Preprint No. TH/87-88/21.

Highly excited stable bound states in strongly coupled lattice gauge theories

Sabino José Ferreira, Ricardo S. Schor, and Michael L. O'Carroll
Departmento de Fisica, ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

(Received 29 December 1988; accepted for publication 12 April 1989)

The mass spectrum of pure three-dimensional SU(2) lattice gauge theory in the adjoint representation is studied. An exact analysis in an approximate model predicts the existence of highly excited bound states. The existence of this bound state in a high-spin state is proved and a convergent expansion for its mass is obtained.

I. INTRODUCTION

The investigation of the particle spectrum in lattice gauge field theories is by now a subject in itself, the goal being to understand quantitatively the rich spectroscopy of continuum QCD. Most of the effort in this direction is done through numerical calculations; see Ref. 1 for recent advances in the field.

Lattice gauge field theories, on the other hand, are also quite suitable for a theoretical study of the particle spectrum *per se*, where the emphasis is in understanding the formation of bound states and resonances not necessarily connected with real particles found in nature. This is because these theories are simple to rigorously construct—because of the lattice cutoff—yet they possess a rich structure of stable and unstable particles.

Basically, there have been two routes used in the rigorous study of the particle spectrum in lattice theories. One of them is based on methods reminiscent of continuum field theories (Euclidean subtractions).[2–6] The other uses statistical mechanical methods (random surfaces).[7–11] In this paper, we follow the former approach.

In Ref. 4 we pointed out that the transfer matrix of a lattice gauge theory acting on gauge invariant states can be explicitly diagonalized if the chromomagnetic interaction is turned off. The resulting spectrum of particles is very rich, and is expected to persist (up to the removal of degeneracies) in the original theory in the strong coupling regime. Unfortunately, we are still very far from a general proof of this statement. What has been verified[3,4] is that it holds for the spectrum associated to the two lightest particles.

In this paper, we go a step further and consider the persistence of the spectrum up to the three lightest particles. Thus we study one of the simplest models, where two bound states of glueballs are expected to exist, namely, a lattice gauge field with the Wilson action in the adjoint representation of the gauge group SU(2), and restrict ourselves to three Euclidean dimensions to further simplify combinatorial estimates. Nevertheless we should stress that the methods used here work in any dimension.

The proof of the existence of the highly excited bound states follows the general strategy of Refs. 2–5, as remarked before. In particular, we use the Z(4) symmetry of the theory associated to successive rotations R of $\pi/2$ around an axis of the lattice to decompose the space of states into four subspaces, $H = \Sigma_{i=1}^{4} H_i$, each transforming according to the irreducible representations of Z(4). These associate to the abstract group $\{R^0 = 1, R, R^2, R^3\}$, respectively, $\{1,1,1,1\}$, $\{1,-1,1,-1\}$, $\{1,i,-1,-i\}$, and $\{1,-i,-1,+i\}$. Here H_1 corresponds to spin-0, since its states are rotation invariant. The lightest particle (the basic glueball) has mass $\sim -4\log\beta$ and lies in H_1^2. Note that β is related to the coupling constant g by $\beta = 1/2g^2$, see Sec. II. The basic glueball has the approximate wave function $\chi(g_p)$ (which becomes exact when the chromomagnetic interaction is turned off), where χ is the character of the representation of the gauge group in the Wilson action and g_p is the oriented product of the group elements around the boundary of the plaquette P, see Fig. 1. Next, we find two nearly degenerate bound states of the basic glueball, with masses $\sim -6\log\beta$, living in the subspaces H_1 and H_2, with approximate wave functions $\chi(g_w)$, where w is the window shown in Fig. 2.[3,5] Depending on the gauge group and its representation in the Wilson action, these are the only strongly bounded states of glueballs. An example of such a situation is SU(2) in the fundamental representation.[4] We remark that there might be weakly bounded states, with masses $\sim -8\log\beta$, but these have not yet been investigated.

If we consider an SU(2) gauge theory with Wilson action in the adjoint representation, then we expect an additional strongly bounded state with mass $\sim -7\log\beta$, whose approximate wave function is not a loop, see Fig. 3. Using the bond assignment shown in Fig. 3, the wave function is

$$\sigma = \begin{pmatrix} 1 & 1 & 1 \\ i_1 & i_3 & i_5 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ i_2 & i_4 & i_6 \end{pmatrix} \mathrm{U}(g_1 g_2 g_3)_{i_1 i_2} \times \mathrm{U}(g_4)_{i_3 i_4} \mathrm{U}(g_5 g_6 g_7)_{i_5 i_6},$$

where U(g) is the spin 1 representation of SU(2) and $\begin{pmatrix} 1 & 1 & 1 \\ i & j & k \end{pmatrix}$ is a 3-j Wigner coefficient.[12]

In this paper, we prove the existence of this excitation by showing its presence in the subspace H_2. We also show that the excitation is absent in the subspaces H_3 and H_4. It is also expected to be present in H_1, but we were unable to prove this. The reason is because our method is based on analyticity properties of the subtracted Euclidean Green's function, a process that introduces spurious poles, whose relations to

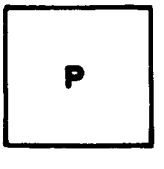


FIG. 1. Plaquette P.

 0022-2488/89/092151-05$02.50

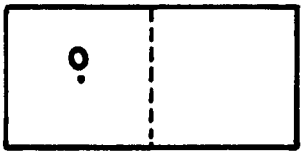


FIG. 2. Window w around the origin.

the mass spectrum become increasingly more complicated and difficult to analyze. Specifically, these relations have been established only for the two lowest mass groups.[4] Now, as we know, in H_2 there is only one excitation below $\sim -7\log\beta$, whereas in H_1 there are two.

The organization of this paper is as follows. In Sec. II we define the model, fix notations, recall results obtained before,[2-5] and state the new results obtained in the present work. The proofs are deferred to Sec. III.

II. DEFINITIONS AND MAIN RESULTS

Our model is defined in the finite volume $\Lambda \subset Z^3$ by the Boltzmann factor

$$e^{A_\Lambda} = \exp\left\{\beta \sum_{P\subset\Lambda} \chi(g_p)\right\}, \tag{2.1}$$

where the sum is over all nonoriented plaquettes P in Λ, g_p is the oriented product of SU(2) group elements along the boundary of P, and χ is the character of the adjoint representation of SU(2), which is a real representation. The expectation of a function ϕ of bound variables is

$$\langle\phi\rangle_\Lambda(\beta) = \frac{\int \phi e^{A_\Lambda}\, dg_\Lambda}{\int e^{A_\Lambda}\, dg_\Lambda}, \tag{2.2}$$

where dg_Λ is the product of Haar measures, one for each bond in Λ. For β sufficiently small, the limit of (2.2) as $\Lambda \to Z^3$ exists[13] and is denoted by $\langle\phi\rangle$. If ϕ and ψ depend only on a finite number of bond variables, we define their truncated correlation function $G_{\phi\psi}(x)$, $x\in Z^3$ by

$$G_{\phi\psi}(x) = \langle\bar\phi\psi(x)\rangle - \langle\bar\phi\rangle\langle\psi\rangle, \tag{2.3}$$

where $\psi(x)$ is the function ψ translated by the lattice point x. It is well known[13] that $G_{\phi\psi}(x)$ clusters exponentially for small beta.

The lattice quantum field theory associated with the action in (2.1) is obtained through the Feynman–Kac formula.[4] Thus the physical Hilbert space H with inner product $(\cdot,\cdot)_H$ is composed of gauge invariant functions supported in the time-zero plane of Z^3. The energy H and momentum

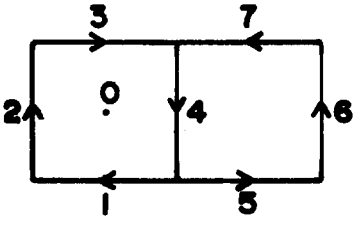


FIG. 3. Wave function σ for excited state.

$\mathbf{P} = (P_1, P_2)$ operators are related to the expectations (2.2) by

$$\langle\bar\phi\psi(x)\rangle = (\phi, e^{-H|x_0|} e^{i\mathbf{P}\cdot\mathbf{x}}\psi)_H, \tag{2.4}$$

where $x = (x_0,\mathbf{x})$. By the spectral theorem and uniqueness of the vacuum (which follows from the cluster property) we have

$$\tilde G_{\phi\psi}(p_0) = (2\pi)^3 \int_{(0,\infty)} \int_{(-\pi,\pi]^2} \frac{\sinh\lambda_0}{\cosh\lambda_0 - \cos p_0} \times \delta(\lambda)\, d(\phi, E(\lambda_0,\lambda)\psi)_H, \tag{2.5}$$

where

$$\tilde G_{\phi\psi}(p_0) = \sum_x G_{\phi\psi}(x) e^{ip_0x_0}. \tag{2.6}$$

Thus the singularities of $\tilde G_{\phi\psi}$ coincide with the energy spectrum at zero momentum (plotted on the imaginary axis).

Let $R_\mathbf{a}$ denote a lattice rotation by $\pi/2$ along an axis parallel to the time (vertical) direction and passing through the point $(0,\mathbf{a})\in Z^3$, which is the center of a horizontal plaquette. We use the same notation for rotations of configurations of gauge fields and functions of configurations. From the invariance of the action, it is easy to show[3] that

$$\hat G_{R_\mathbf{a}\phi, R_\mathbf{b}\psi}(x_0) = \hat G_{\phi,\psi}(x_0), \tag{2.7}$$

where

$$\hat G_{\phi\psi}(x_0) = \sum_\mathbf{x} G_{\phi\psi}(x_0,\mathbf{x}). \tag{2.8}$$

Define

$$\begin{aligned} P_\mathbf{a}^{(1)} &= \tfrac14(1 + R_\mathbf{a} + R_\mathbf{a}^2 + R_\mathbf{a}^3),\\ P_\mathbf{a}^{(2)} &= \tfrac14(1 - R_\mathbf{a} + R_\mathbf{a}^2 - R_\mathbf{a}^3),\\ P_\mathbf{a}^{(3)} &= \tfrac14(1 + iR_\mathbf{a} - R_\mathbf{a}^2 - iR_\mathbf{a}^3),\\ P_\mathbf{a}^{(4)} &= \tfrac14(1 - iR_\mathbf{a} - R_\mathbf{a}^2 + iR_\mathbf{a}^3). \end{aligned} \tag{2.9}$$

We have $P_\mathbf{a}^{(i)} P_\mathbf{a}^{(j)} = \delta_{ij} P_\mathbf{a}^{(i)}$ and $\Sigma_i P_\mathbf{a}^{(i)} = 1$. Also, $\hat G_{P_\mathbf{a}^{(i)}\phi,\psi}(x_0) = \hat G_{\phi, P_\mathbf{b}^{(i)}\psi}(x_0)$, implying the selection rule referred to in the Introduction. Thus the spectrum of the energy operator at zero momentum can be analyzed in each sector H_i separately, where $H_i = \overline{H_i^{(0)}}$ [closure in the inner product (2.4)] and $H_i^{(0)}$ consist of functions ϕ depending only on a finite number of bond variables at time zero, such that $\phi = P_\mathbf{a}^{(i)}\phi$ for some $a\in Z^2$.

The results presented below are valid for beta sufficiently small; the constant ϵ appearing in the theorems depends on β and $\lim_{\beta\downarrow 0}\epsilon(\beta) = 0$.

Theorem 2.1: If $\phi\in H_i^{(0)}$, $i = 3,4$, then $\tilde G_{\phi\phi}(p_0)$ is analytic on $|\mathrm{Re}\, p_0| \leqslant \pi$, $|\mathrm{Im}\, p_0| \leqslant -8(1-\epsilon)\log\beta$.

Theorem 2.2: If $\phi\in H_2^{(0)}$, $\tilde G_{\phi\phi}(p_0)$ is meromorphic on $|\mathrm{Re}\, p_0| \leqslant \pi$, $|\mathrm{Im}\, p_0| \leqslant -8(1-\epsilon)\log\beta$ with possible poles (independent of ϕ) at $p_0 = -6i\log\beta + r_1(\beta)$ and $p_0 = -7i\log\beta + r_2(\beta)$, where $r_i(\beta)$ are analytic at $\beta = 0$. These poles are present if $\phi = P_\mathbf{a}^{(2)}\chi(g_w)$ or $\phi = P_\mathbf{a}^{(2)}\sigma$.

Theorem 2.3: If $\phi\in H_1^{(0)}$, $\tilde G_{\phi\phi}(p_0)$ is meromorphic on $|\mathrm{Re}\, p_0| \leqslant \pi$, $|\mathrm{Im}\, p_0| \leqslant -7(1-\epsilon)\log\beta$, with possible poles (independent of ϕ) $p_0 = -4i\log\beta + r_3(\beta)$ and $p_0 = -6i\log\beta + r_4(\beta)$, where $r_j(\beta)$ are analytic at $\beta = 0$.

These poles occur if $\phi = P_a^{(1)}\chi(g_p)$ or $\phi = P_a^{(1)}\chi(g_w)$ and are given by a convergent expansion in β.

Our new results are contained in Theorems 2.1 and 2.2. Theorem 2.3 was already established.[3,5] As mentioned in the Introduction, we conjecture the existence of a pole $\tilde{G}_{\phi\phi}(p_0)$ at $p_0 \simeq -7i\log\beta$ when $\phi = P_a^{(1)}\sigma$.

The proofs of Theorems 2.1 and 2.2 will be given in the next section.

III. TECHNICAL ESTIMATES AND PROOFS

As in Refs. 3 and 4, the proofs of our main results follow after establishing exponential decay properties of suitable correlation functions. We introduce a lattice approximation to the action with complex parameters $\{w_q\},z$ and periodic boundary conditions in the spatial direction, given by

$$A_\Lambda = \sum_q w_q \sum_{P\in P_q''} \chi(g_p) + z \sum_{P\in P^\perp} \chi(g_p), \tag{3.1}$$

where P_q'' denote the plaquettes parallel to the time direction (x_0) between the planes $x_0 = q$ and $x_0 = q+1$, and $P^\perp$ are the plaquettes perpendicular to the time direction. The average of a function $\phi(g)$ with respect to the Gibbs factor, defined by (3.1), is

$$\langle\phi\rangle_\Lambda(\{w_q\},z) = \frac{1}{Z_\Lambda}\int \phi e^{A_\Lambda}\, dg_\Lambda, \tag{3.2}$$

where Z_Λ is such that $\langle 1\rangle_\Lambda = 1$. From the polymer expansion,[13] $\langle\phi\rangle_\Lambda$ is analytic in all variables $\{w_q\},z$ in a sufficiently small neighborhood of the origin (independent of Λ); setting all w_q,z equal to β, the thermodynamic limit exists and is translation invariant. Given ϕ and ψ of finite support, there is a constant m_0 independent of $\{w_q\}$, z and Λ, such that

$$|\langle\phi(x)\psi(y)\rangle_\Lambda - \langle\phi(x)\rangle_\Lambda\langle\psi(y)\rangle_\Lambda| \leqslant C_{\phi\psi}e^{-m_0|x-y|}, \tag{3.3}$$

where $\phi(x)$ is ϕ translated by $x\in Z^3$ and $C_{\phi\psi}$ depends only on ϕ and ψ.

As in Sec. II, we define

$$G_{\phi\psi}(x,y;\Lambda) = \langle\bar{\phi}(x)\psi(y)\rangle_\Lambda - \langle\bar{\phi}(x)\rangle_\Lambda\langle\psi(y)\rangle_\Lambda \tag{3.4}$$

and

$$\hat{G}_{\phi\psi}(x_0,y_0;\Lambda) = \sum_{y\in Z^2} G_{\phi\psi}(x,y;\Lambda). \tag{3.5}$$

Notice that (3.3) implies $|\hat{G}_{\phi\psi}(x_0,y_0;\Lambda)| \leqslant C'_{\phi\psi}e^{-m_0|x_0-y_0|}$. It is useful to represent the truncated correlation (3.4) in terms of duplicate variables

$$G_{\phi\psi}(x,y;\Lambda) = \frac{1}{2Z_\Lambda^2}\int (\bar{\phi}(x) - \bar{\phi}'(x)) \times(\psi(y) - \psi'(y))e^{(A_\Lambda + A'_\Lambda)}\, dg_\Lambda\, dg'_\Lambda, \tag{3.6}$$

where, e.g., ψ' is ψ at g'. The basic result on the analytic structure of $G_{\phi\psi}(x,y;\Lambda)$ is given in Theorem 1 below. We let χ be the elementary plaquette function centered around the origin in the x_1,x_2 plan (Fig. 1), and χ_h,χ_v the elementary rectangular loop functions in the x_1,x_2 plane with long axis along x_1,x_2, respectively, and with the origin 0, as in Fig. 2. We make similar definitions for σ_h,σ_v (Fig. 3).

Theorem 3.1: Let $x_0 \leqslant q < y_0$, then

$$\begin{aligned} G_{\phi\psi}(x,y;\Lambda) = {} & \sum_{t_0=q} G_{\phi\chi}(x,t;\Lambda)G_{\chi\psi}(t+\hat{e}_0,y;\Lambda)\left[c_{44}w_q^4 + c_{45}w_q^5 + c_{46}w_q^6 + c_{47}w_q^7\right] \\ & + \sum_{j=h,v}\sum_{t_0=q} G_{\phi\chi_j}(x,t;\Lambda)G_{\chi_j\psi}(t+\hat{e}_0,y;\Lambda)\left[c_{66}w_q^6 + c_{67}w_q^7\right] \\ & + \sum_{j=h,v}\sum_{t_0=q} G_{\phi\sigma_j}(x,t;\Lambda)G_{\sigma_j\psi}(t+\hat{e}_0,y;\Lambda)c_{77}w_q^7 + R_{\phi\psi}^{q,8}(x,y;\Lambda), \end{aligned} \tag{3.7}$$

where the c's are combinatorial constants, and c_{44}, c_{66}, and c_{77} are positive,

$$\frac{\partial^n}{\partial w_q^n} R_{\phi,\psi}^{q,8}(x,y,\Lambda)\bigg|_{w_q=0} = 0, \quad \text{for } 0\leqslant n\leqslant 7,$$

and the G's on the right-hand side are evaluated at $w_q = 0$.

Proof: The proof is an extension of one developed in Ref. 3, so that we will be very brief. The idea is to expand the term proportional to w_q in both the numerator and denominator of (3.6). The coefficients of the resulting terms can then be calculated explicitly using the Peter–Weyl orthogonality theorem. For example, the term proportional to c_{44} comes from integrating the vertical bonds of the form $[(q,x_1,x_2),(q+1,x_1,x_2)]$ when four plaquette variables, parallel to the time direction, are disposed along the sides of a cube. The c_{45} term corresponds to five plaquettes occupying the four sides parallel to the time direction of a cube, and similarly for c_{46} and c_{47}. For c_{66}, six plaquettes are disposed along the sides parallel to the time direction of an elementary parallelepiped and c_{67} corresponds to seven such plaquettes. Finally, in c_{77} seven plaquettes are positioned along the seven faces parallel to the time direction of a parallelepiped with one face at its center. □

The finite volume correlation functions $G_{\phi\psi}(x,y;\Lambda)$ with complex parameters $\{w_q\},z$ still have the Z(4) symmetry corresponding to successive rotations of $\pi/2$ along the time axis. With the same notation as in Sec. II, we have

$$\hat{G}_{\phi\psi}(x_0,y_0;\Lambda) = \hat{G}_{R_a\phi,R_b\psi}(x_0,y_0;\Lambda) \tag{3.8}$$

and

$$\hat{G}_{P_a^{(i)}\phi,\psi}(x_0,y_0;\Lambda) = \hat{G}_{\phi,P_b^{(i)}\psi}(x_0,y_0;\Lambda). \tag{3.9}$$

Taking the partial Fourier transform (at zero spatial momentum) (3.5) of (3.7) and denoting the new constants again by $c_{44},\ldots,c_{77}$, we have the following.

Corollary 3.2: Let $x_0 \leqslant q \leqslant y_0$. Then

$$\hat{G}_{\phi\psi}(x_0,y_0;\Lambda)=\hat{G}_{\phi\chi}(x_0,q;\Lambda)\hat{G}_{\chi\psi}(q+1,y_0,\Lambda)\,[c_{44}w_q^4+c_{45}w_q^5+c_{46}w_q^6+c_{47}w_q^7]$$
$$+\sum_{i=1}^{2}\hat{G}_{\phi\chi_i}(x_0,q;\Lambda)\hat{G}_{\chi_i\psi}(q+1,y_0;\Lambda)\,[c_{66}w_q^6+c_{67}w_q^7]$$
$$+\sum_{i=1}^{2}\hat{G}_{\phi\sigma_i}(x_0,q;\Lambda)\hat{G}_{\sigma_i\psi}(q+1,y_0;\Lambda)c_{77}w_q^7+\hat{R}^{q,8}_{\phi\psi}(x_0,y_0;\Lambda), \tag{3.10}$$

where c_{44}, c_{66}, and c_{77} are positive,

$$\frac{\partial^n}{\partial w_q^n}\hat{R}^{q,8}_{\phi\psi}(x_0,y_0;\Lambda)\Big|_{w_q=0}=0,\quad \text{for } 0\leqslant n\leqslant 7,$$

and the $\hat{G}$'s on the right-hand side are evaluated at $w_q=0$. Also,

$$\chi_i=P^{(i)}_{\mathbf{a}=0}\chi_h \text{ and } \sigma_i=P^{(i)}_{\mathbf{a}=0}\sigma_h. \qquad \square$$

We remark that the sums in (3.10) only involve χ_i,σ_i for $i=1,2$. This is because $\hat{G}_{\phi\chi_i}(x_0,y_0;\Lambda)=\hat{G}_{\phi\sigma_i}(x_0,y_0;\Lambda)=0$ when $i=3,4$, as a result of translation invariance in the spatial direction.

Proof of Theorem 2.1: If $\phi=P^{(i)}_{\mathbf{a}}\phi$ for $i=3,4$, then because of (3.9), $\hat{G}_{\phi\chi}=\hat{G}_{\phi\chi_i}=\hat{G}_{\phi\sigma_i}=0$. From Correlation 3.2, we see that for $x_0\leqslant q<y_0$,

$$\frac{\partial^n}{\partial w_q^n}\hat{G}_{\phi\phi}(x_0,y_0;\Lambda)\Big|_{w_q=0}=0$$

for $0\leqslant n\leqslant 7$. Now, as remarked before, $\hat{G}_{\phi\phi}(x_0,y_0;\Lambda)$ is analytic in $\{w_q\},z$ on $|w_q|,|z|<\beta_0$ (β_0 is fixed and is given by the polymer expansion) and uniformly bounded there, say $|\hat{G}_{\phi\phi}(x_0,y_0;\Lambda)|\leqslant C_{\phi\phi}$. From the maximum modulus theorem, it follows that

$$|\hat{G}_{\phi\phi}(x_0,y_0;\Lambda)|\leqslant C_{\phi\phi}\prod_{x_0\leqslant q<y_0}|w_q/\beta_0|^8.$$

Setting all w_q,z equal to β, $0<\beta<\beta_0$ and taking $\Lambda\to Z^3$ we get $|\tilde{G}_{\phi\phi}(x_0)|\leqslant C_{\phi\phi}(\beta/\beta_0)^{8|x_0|}$, proving the analyticity of $\tilde{G}_{\phi\phi}(p_0)$ on $|\mathrm{Re}\,p_0|\leqslant\pi$, $|\mathrm{Im}\,p_0|<-8\log(\beta/\beta_0)$. $\square$

We now begin the proof of Theorem 2.2. The existence of the pole near $(0,-6i\log\beta)$ was already established[3-5] by exhibiting the corresponding zero of $\tilde{\Gamma}_{\chi_2\chi_2}(p_0)\equiv-\tilde{G}_{\chi_2\chi_2}(p_0)$. Here, we look for an additional zero of $\tilde{\Gamma}_{\chi_2\chi_2}(p_0)$ near $(0,-7i\log\beta)$. To this end, introduce the function

$$\tilde{F}_{\sigma_2\sigma_2}(p_0)=\tilde{G}_{\sigma_2\sigma_2}(p_0)+\tilde{G}_{\sigma_2\chi_2}(p_0)\times\tilde{\Gamma}_{\chi_2\chi_2}(p_0)\tilde{G}_{\chi_2\sigma_2}(p_0). \tag{3.11}$$

Theorem 3.3: $\tilde{F}_{\sigma_2\sigma_2}(p_0)$ is analytic on $|\mathrm{Re}\,p_0|\leqslant\pi$, $|\mathrm{Im}\,p_0|\leqslant-7(1-\epsilon)\log\beta$.

Proof: Introduce the finite volume approximation $\hat{F}_{\sigma_2\sigma_2}(x_0,y_0;\Lambda)$ with parameters $\{w_q\},z$ by (in matrix notation)

$$\hat{F}_{\sigma_2\sigma_2}(\Lambda)=\hat{G}_{\sigma_2\sigma_2}(\Lambda)+\hat{G}_{\sigma_2\chi_2}(\Lambda)\hat{\Gamma}_{\chi_2\chi_2}(\Lambda)\hat{G}_{\chi_2\sigma_2}(\Lambda). \tag{3.12}$$

By a direct calculation we have, if $x_0\leqslant q<y_0$,

$$\frac{\partial^n}{\partial w_q^n}\hat{\Gamma}_{\chi_2\chi_2}(x_0,y_0,\Lambda)\Big|_{w_q=0}=0,\quad \text{for } 0\leqslant n\leqslant 6$$

and

$$\frac{\partial^6}{\partial w_q^6}\hat{\Gamma}_{\chi_2\chi_2}(x_0,y_0;\Lambda)\Big|_{w_q=0}=6!c_{66}\delta(x_0,q)\delta(q+1,y_0).$$

Such calculations are lengthy but straightforward, and were explained in detail in Ref. 3. Here and in the sequel, we just present the final results. Using this result and Corollary 3.2 in Leibniz's formula,

$$\frac{\partial^m}{\partial w_q^m}\hat{F}_{\sigma_2\sigma_2}(\Lambda)=\frac{\partial^m}{\partial w_q^m}\hat{G}_{\sigma_2\sigma_2}(\Lambda)+\sum_{n_1+n_2+n_3=m}\times\frac{m!}{n_1!n_2!n_3!}\frac{\partial^{n_1}}{\partial w_q^{n_1}}\hat{G}_{\sigma_2\chi_2}(\Lambda)\times\frac{\partial^{n_2}}{\partial w_q^{n_2}}\hat{\Gamma}_{\chi_2\chi_2}(\Lambda)\cdot\frac{\partial^{n_3}}{\partial w_q^{n_3}}\hat{G}_{\chi_2\sigma_2}(\Lambda), \tag{3.13}$$

we show that

$$\frac{\partial^m}{\partial w_q^m}\hat{F}_{\sigma_2\sigma_2}(x_0,y_0;\Lambda)\Big|_{w_q=0}=0$$

if $x_0\leqslant q<y_0$ and $0\leqslant m\leqslant 6$. The proof now follows from the reasoning used in Theorem 2.1 above. $\square$

The next result shows that, in fact, $\tilde{F}_{\sigma_2\sigma_2}(p_0)$ has a pole near $p_0=-7i(1-\epsilon)\log\beta$.

Therorem 3.4: $\tilde{F}^{-1}_{\sigma_2\sigma_2}(p_0)$ is analytic on $|\mathrm{Re}\,p_0|\leqslant\pi$, $0\leqslant\mathrm{Im}\,p_0\leqslant-8(1-\epsilon)\log\beta$ and has precisely one zero at $p_0=ip_{\sigma_2}=i[-7\log\beta+r_{\sigma_2}(\beta)]$, where $r_{\sigma_2}(\beta)$ is analytic at $\beta=0$.

Proof: Taking the seventh derivative of $\hat{F}_{\sigma_2\sigma_2}(\Lambda)$ in (3.12), we get after a lengthy calculation

$$\frac{\partial^7}{\partial w_q^7}\hat{F}_{\sigma_2\sigma_2}(x_0,y_0,\Lambda)\Big|_{w_q=0}=c_{77}\hat{F}_{\sigma_2\sigma_2}(x_0,q,\Lambda)\hat{F}_{\sigma_2\sigma_2}(q+1,y_0;\Lambda)\,|_{w_q=0}.$$

This result, together with (3.13), implies that

$$\frac{\partial^m}{\partial w_q^m}\hat{F}^{-1}_{\sigma_2\sigma_2}(x_0,y_0;\Lambda)\Big|_{w_q=0},\quad \text{for } 0\leqslant m\leqslant 6$$

and

$$\frac{\partial^7}{\partial w_q^7}\hat{F}^{-1}_{\sigma_2\sigma_2}(x_0,y_0;\Lambda)\Big|_{w_q=0}=c_{77}\delta(x_0,q)\delta(q+1,y_0)=0 \quad \text{if } y_0\geqslant x_0+2.$$

The analyticity of $\tilde{F}^{-1}_{\sigma_2\sigma_2}(p_0)$ follows then as before. Now, from the structure of the derivatives, the infinite volume $\hat{F}^{-1}_{\sigma_2\sigma_2}(0,x_0)\equiv\hat{F}^{-1}_{\sigma_2\sigma_2}(x_0)$ with all $\{w_q\},z$ equal to the same (complex) β has the form

$$\hat{F}^{-1}_{\sigma_2\sigma_2}(x_0=0)=\sum_{n=0}^{\infty}c_n\beta^n,\quad c_0>0,$$

$$\hat{F}^{-1}_{\sigma_2\sigma_2}(|x_0|=1)=\sum_{n=7}^{\infty} d_n\beta^n,\quad d_7=-\frac{1}{7!}c_{77}<0,$$

$$|\hat{F}^{-1}_{\sigma_2\sigma_2}(x_0)|\leqslant k_1(k_2|\beta|)^{8|x_0|},\quad \text{if } |x_0|\geqslant 2.$$

Thus

$$\tilde{F}^{-1}_{\sigma_2\sigma_2}(p_0)=(c_0+d_7\beta^7e^{-ip_0})+\left(\sum_{n=1}^{\infty}c_n\beta^n+d_7\beta^7e^{+ip_0}\right)$$
$$+\left(\sum_{n=8}^{\infty}d_n\beta^n\right)(e^{ip_0}+e^{-ip_0})$$
$$+\sum_{|x_0|=2}^{\infty}F^{-1}_{\sigma_2\sigma_2}(|x_0|)(e^{ip_0|x_0|}+e^{-ip_0|x_0|})$$
$$=g(p_0)+h(p_0),$$

where $g(p_0)=c_0+d_7\beta^7e^{-ip_0}$. It is easy to see that on the boundary ∂R of

$$R=\{p_0:\ |\mathrm{Re}\,p_0|\leqslant\pi, 0\leqslant \mathrm{Im}\,p_0\leqslant -8(1-\epsilon)\log\beta\},$$
$$|g(p_0)|\geqslant C_0/2.$$

Also, for small β, $|h(p_0)|<C_0/2$. It follows from Rouche's theorem that $\tilde{F}^{-1}_{\sigma_2\sigma_2}(p_0)$ has a unique zero inside R. Finally, letting

$$H(w=c_0+d_7\beta^7e^{-ip_0};\beta)=\tilde{F}^{-1}_{\sigma_2\sigma_2}(p_0;\beta),$$

the analytic implicit function theorem gives an analytic function $w(\beta)$ around $\beta=0$, such that $H(w(\beta),\beta)\equiv 0$ and $w(0)=0$. The proof of the theorem is complete. □

Let

$$\tilde{L}_{\chi_2\sigma_2}=\tilde{\Gamma}_{\chi_2\chi_2}\tilde{G}_{\chi_2\sigma_2}\tilde{F}^{-1}_{\sigma_2\sigma_2},\quad \tilde{L}_{\sigma_2\chi_2}=\tilde{F}^{-1}_{\sigma_2\sigma_2}\tilde{G}_{\sigma_2\chi_2}\tilde{G}_{\chi_2\chi_2}$$

and

$$\tilde{M}=\tilde{\Gamma}_{\chi_2\chi_2}\tilde{G}_{\sigma_2\sigma_2}\tilde{F}^{-1}_{\sigma_2\sigma_2}.$$

Multiplying (3.11) by $\tilde{\Gamma}_{\chi_2\chi_2}\tilde{F}^{-1}_{\sigma_2\sigma_2}$ we obtain

$$\tilde{\Gamma}_{\chi_2\chi_2}=\tilde{M}+\tilde{L}_{\chi_2\sigma_2}\tilde{F}_{\sigma_2\sigma_2}\tilde{L}_{\sigma_2\chi_2}. \tag{3.14}$$

The basic results about $\tilde{M}$ and $\tilde{L}$ are given in the following.

Theorem 3.5: (a) $\tilde{M}$ and $\tilde{L}$ are analytic on $|\mathrm{Re}\,p_0|\leqslant\pi$, $0\leqslant\mathrm{Im}\,p_0\leqslant -8(1-\epsilon)\log\beta$, (b) $\tilde{L}(ip_{\sigma_2})\neq 0$, and (c) $\tilde{M}(p_0)\neq 0$ for $|\mathrm{Re}\,p_0|\leqslant\pi$, $-6(1+\epsilon)\log\beta\leqslant\mathrm{Im}\,p_0\leqslant -8(1-\epsilon)\log\beta$.

Proof: (a) This proof follows from the explicit calculations

$$\frac{\partial^m}{\partial w_q^m}\tilde{L}_{\chi_2\sigma_2}(x_0,y_0;\Lambda)\bigg|_{w_q=0}=0,\quad \text{if } x_0\leqslant q<y_0$$
$$\text{and}\quad 0\leqslant m\leqslant 7, \tag{3.15}$$

and

$$\frac{\partial^m}{\partial w_q^m}\hat{M}(x_0,y_0;\Lambda)\bigg|_{w_q=0}=0,\quad \text{if } x_0\leqslant q<y_0$$
$$\text{and}\quad 0\leqslant m\leqslant 5,$$
$$\frac{\partial^m}{\partial w_q^m}\hat{M}(x_0,y_0;\Lambda)\bigg|_{w_q=0}=\alpha_m\delta(x_0,q)\delta(q+1,y_0)$$
$$\text{for } m=6,7, \tag{3.16}$$

where $\alpha_m\neq 0$ are numerical constants.

(b) Here, we use the fact [which goes beyond (3.15)] that $\hat{L}_{\chi_2\sigma_2}(x_0,y_0;\Lambda)=O(\beta^9)$ when we set all w_q,z equal to β. The mechanism responsible for this is the same as in Ref. 3. It implies that for $p_0=i\rho_2+r$, with $|r|\leqslant 1$,

$$|\tilde{L}_{\chi_2\sigma_2}(p_0)-\hat{L}_{\chi_2\sigma_2}(x_0=0)|=O(\beta^2).$$

Since $|\hat{L}_{\chi_2\sigma_2}(x_0=0)|\geqslant\text{const}$, β, the result follows.

(c) Write

$$|\tilde{M}(p_0)-2\hat{M}(|x_0|=1)\cos p_0|$$
$$\leqslant|\hat{M}(x_0=0)|+2\sum_{|x_0|\geqslant 2}|\hat{M}(|x_0|)|\cos p_0|x_0||.$$

From (3.16), $|\hat{M}(x_0)|\leqslant\gamma_1\beta^{8|x_0|}$ for some constant γ_1, and also $|\hat{M}(|x_0|=1)|\geqslant\gamma_2\beta^6$. Now, if $|\mathrm{Im}\,p_0|\leqslant -8(1-\epsilon)\log\beta$, we have $|\cos p_0|x_0||\leqslant e^{-8(1-\epsilon)|x_0|\log\beta}$ and

$$\sum_{|x_0|\geqslant 2}|\hat{M}(|x_0|)||\cos p_0|x_0||\leqslant\gamma_3\beta^{16\epsilon}.$$

Also, if $\mathrm{Im}\,p_0\geqslant -6(1+\epsilon)\log\beta$, then $|\hat{M}(|x_0|=1)\times\cos p_0|\geqslant\frac{1}{2}\gamma_2\beta^{-6\epsilon}$. Since $|\hat{M}(x_0=0)|\leqslant\gamma_4$, the proof is complete.

From (3.14) and Theorem 3.5, we see that $\tilde{\Gamma}_{\chi_2\chi_2}$ is analytic on $|\mathrm{Re}\,p_0|\leqslant\pi$, $0\leqslant\mathrm{Im}\,p_0\leqslant -8(1-\epsilon)\log\beta$, except for a pole at $p_0=ip_{\sigma_2}$. For $-6(1+\epsilon)\log\beta\leqslant\mathrm{Im}\,p_0\leqslant -8(1-\epsilon)\log\beta$, rewrite (3.14) as

$$\tilde{\Gamma}_{\chi_2\chi_2}=\tilde{M}\tilde{F}_{\sigma_2\sigma_2}(\tilde{F}^{-1}_{\sigma_2\sigma_2}+\tilde{L}_{\chi_2\sigma_2}\tilde{M}^{-1}\tilde{L}_{\sigma_2\chi_2}). \tag{3.17}$$

In this form, we can identify a zero of $\tilde{\Gamma}_{\chi_2\chi_2}$ in the above region as a zero of $\tilde{F}^{-1}_{\sigma_2\sigma_2}+\tilde{L}_{\chi_2\sigma_2}\tilde{L}_{\sigma_2\chi_2}\tilde{M}^{-1}$.

Theorem 3.6: Note that $\tilde{\Gamma}_{\chi_2\chi_2}(p_0)$ has precisely one zero in the region $|\mathrm{Re}\,p_0|\leqslant\pi$, $-6(1+\epsilon)\log\beta\leqslant\mathrm{Im}\,p_0\leqslant -8(1-\epsilon)\log\beta$, at $p_0=im=i[-7\log\beta+r(\beta)]$, where $r(\beta)$ is analytic at $\beta=0$.

Proof: The proof uses the same arguments developed in the proof of Theorem 3.4, with the additional remark that from Theorem 3.5, $|\tilde{L}_{\chi_2\sigma_2}\tilde{L}_{\sigma_2\chi_2}\tilde{M}^{-1}|\leqslant\text{const}\,\beta^{(\text{const}\,\epsilon)}$. □

Theorem 3.6 establishes the most important result stated in Theorem 2.2. The other statements now follow by a standard argument, which is presented in detail at the end of Sec. 4 of Ref. 3.

[1] *International Symposium on Field Theory on the Lattice*, Nucl. Phys. B, Proc. Suppl. **4** (1988).
[2] R. Schor, Nucl. Phys. B **222**, 71 (1983).
[3] R. Schor, Commun. Math. Phys. **92**, 369 (1984).
[4] R. Schor and M. O'Carroll, Commun. Math. Phys. **103**, 569 (1986).
[5] M. O'Carroll, J. Math. Phys. **26**, 2342 (1985).
[6] J. C. Barata and K. Fredenhagen, Commun. Math. Phys. **113**, 403 (1987).
[7] J. Bricmont and J. Fröhlich, Nucl. Phys. B **251**, 517 (1985).
[8] J. Bricmont and J. Fröhlich, Commun. Math. Phys. **98**, 553 (1985).
[9] J. Bricmont and J. Fröhlich, Nucl. Phys. B **280**, 385 (1987).
[10] P. A. Marchetti, Commun. Math. Phys. **117**, 501 (1988).
[11] C. King and J. Fröhlich, Nucl. Phys. B **290**, 157 (1987).
[12] E. Wigner, *Group Theory* (Academic, New York, 1959).
[13] E. Seiler, *Lecture Notes in Physics* (Springer, Berlin, 1982), Vol. 159.

Cluster expansion in terms of knots in gauge theories with finite non-Abelian gauge groups

K. Szlachànyi
II. Institut für Theoretische Physik der Universität Hamburg, Luruper Chaussee 149, D-2000 Hamburg 50, Federal Republic of Germany

P. Vecsernyès
Central Research Institute for Physics, Budapest, P.O. Box 49, H-1525 Budapest 114, Hungary

(Received 18 January 1989; accepted for publication 5 April 1989)

The cluster expansion is developed in lattice gauge theories with finite gauge groups in $d \geqslant 3$ dimensions in which the clusters are connected $(d-2)$-dimensional complexes, i.e., connected $(d-2)$ surfaces that can branch along $(d-3)$ cells. The interaction between them has a knot theoretical interpretation. It is a many-body interaction, depending on the type of knot they form together. For small enough gauge coupling g analyticity of the correlation functions in the variable $\exp(-1/g^2)$ is proven.

I. INTRODUCTION

In this paper we give a reformulation of lattice gauge theory in terms of purely geometrical, gauge invariant objects: the sets of plaquettes where the field strength can be nontrivial. The coconnected components of these plaquette sets will be called vortices. The interaction between the vortices depends both on the gauge group and on the topology of how each vortex is embedded in the complement of the others in space-time. This advocates the importance of knot theoretical[1] considerations in gauge theory, at least in the case of finite gauge groups when the vortices can be dilute. This is indeed true in the weak coupling regime where we can construct a convergent expansion using knots. But it may also give new insight on the structure of the phase(s) at stronger coupling where the vortex knots "percolate."

There is an interesting theorem of Zeeman,[2] which suggests also that knot theory should be relevant to gauge theory. According to this theorem the knotting of n-dimensional spheres in d-dimensional space is always trivial if $n \neq d-2$. Therefore, with their $(d-2)$-dimensional vortex sheets, gauge theories are associated with precisely the nontrivial knotting problem.

In constructing a cluster expansion the first step is to find a procedure that splits any configuration into parts (clusters) in such a way that the different clusters are located far enough from each other in order to be considered as independent excitations with respect to the Boltzmann measure. In our case the configurations are gauge equivalence classes and the problem is how to split a gauge equivalence class in space-time. For Abelian gauge theories the answer has been well known for a long time: the clusters are gauge equivalence classes with field strengths F having coconnected support, i.e., the set $P = \operatorname{Supp} F$ of plaquettes p where $F_p \equiv U(\partial p) \neq 1$ can be walked by using a nearest neighbor plaquette-cube–plaquette-cube$\cdots$walk, which hits only the plaquettes of P. That these clusters are independent, i.e., that they factorize the Boltzmann weight, is a consequence of the following properties of Abelian gauge theories.

(i) An Abelian gauge equivalence class can be uniquely characterized by its field strength configuration F; (ii) if F is a field strength configuration with support P and P' is a coconnected component of P, then the restriction F' of F to P' is again an allowed field strength configuration; and (iii) if $F_1,\ldots,F_n$ are field strengths with supports that are pairwise not coconnected, then their (plaquettewise) product is again a field strength.

This allows one to rewrite the system as a (dilute for weak coupling) gas of coconnected plaquette sets with only hard core pair interaction.[3]

If the gauge group G is non-Abelian one faces the difficulty that even property (i) breaks down. On the one hand, only the conjugacy class $[F_p]$ of F_p is gauge invariant; on the other hand, one can construct examples showing that different gauge equivalence classes may possess identical $[F]$ configurations. That is, $[F]$ alone does not determine a unique gauge equivalence class. One has to use a nonlocal characterization of the gauge equivalence classes: for each closed curve C starting at a fixed base point, the parallel transport $U(C)$ along C should be given.

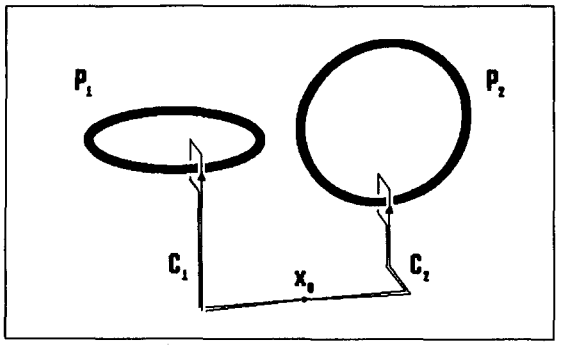
The notion of coconnectedness is also not satisfactory for non-Abelian theories. It does not give rise to independent clusters; in fact, there is an action at a distance between the coconnected components of $P = \operatorname{Supp} F$. Consider the three-dimensional example shown in Fig. 1 where the dual of P consists of two circles. If these circles are not linked with each other [Fig. 1(a)] then arbitrary values of the "magnetic" fluxes $U(C_1), U(C_2) \in G \setminus \{1\}$ are possible. However, as we shall see later, if P_1 and P_2 are linked [Fig. 1(b)], then $U(C_1)$ and $U(C_2)$ have to commute, otherwise the corresponding gauge equivalence class does not exist.

That this type of knotting interaction is a genuine many-body interaction can be seen from the following example. There exists a P (Fig. 2) consisting of n coconnected components $P_1,\ldots,P_n$, such that for any $i = 1,\ldots,n$, $P \setminus P_i$ is completely unknotted but P constitutes a single knot in the sense that no part of P can be pulled out from the rest without intersecting it.

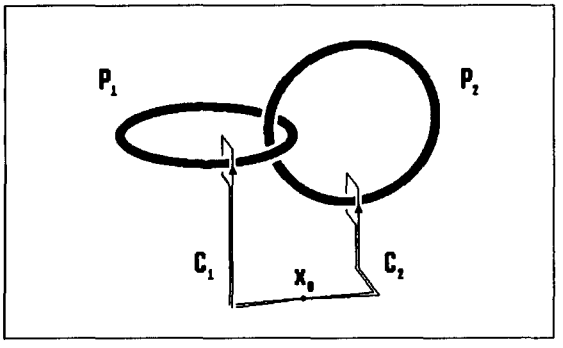
II. GAUGE THEORY AS A GAS OF KNOTTED MAGNETIC VORTICES

Let G be a finite group and $\mathscr{C}$ and $\mathscr{G}$ be the sets of G-valued functions with finite support on the lattice links and

 0022-2488/89/092156-04$02.50



(a)



(b)

FIG. 1. (a) The fundamental group of the complement of the two unlinked circles is the free group with two generators: $\mathcal{F}(P_1 \cup P_2) = \langle a_1, a_2 | - \rangle$. (b) If they are linked, then the fundamental group is the Abelian group with two generators: $\mathcal{F}(P_1 \cup P_2) = \langle a_1, a_2 | a_1 a_2 a_1^{-1} a_2^{-1} \rangle$.

lattice sites, respectively. Here $\mathcal{G}$ acts as a gauge group on the configuration space $\mathcal{C}$. Let Ω^1 be the one-skeleton of the lattice, that is, the union of links as a subset of $\mathbb{R}^d$ and denote by $\mathcal{F}$ the fundamental group $\pi_1(\Omega^1)$ of Ω^1 with base point $x_0 = \infty$. One can show that there is a one-to-one correspondence between the $\mathcal{G}$ orbits in $\mathcal{C}$ (the gauge equivalence classes) and the homomorphisms $\nu \in \mathrm{Hom}(\mathcal{F}, G)$ from the fundamental group $\mathcal{F}$ to the gauge group G.

We denote by Supp ν the set of plaquettes p, such that if C_p is a plaquette curve around p then $\nu(C_p) \neq 1$. ("Plaquette curves" are the closed curves like C_1 and C_2 in Fig. 1.) If Supp $\nu = P$ is fixed then ν can be considered a homomorphism from $\mathcal{F}(P) = \pi_1(\Omega^2 \setminus P)$ to G, where Ω^2 denotes the

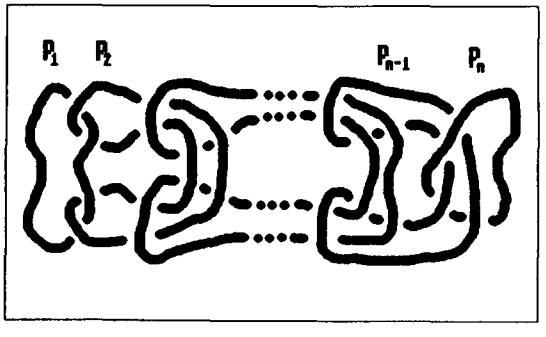


FIG. 2. A knot of n circles, which is irreducible in the sense that it has only trivial subknots.

two-skeleton of the lattice. Since any plaquette set can be decomposed into coconnected components, one can rewrite the partition function as

$$Z = \sum_{\Gamma \subset \mathcal{P}_c} \Phi(\Gamma), \quad \Phi(\Gamma) = \Phi_{\mathrm{hc}}(\Gamma)\Phi_k(\Gamma),$$

$$\Phi_{\mathrm{hc}}(\Gamma) = \prod_{\substack{P,P' \in \Gamma \\ P \neq P'}} \delta_{P \sim P'}, \tag{1}$$

$$\Phi_k(\Gamma) = \sum_{\substack{\nu \in \mathrm{Hom}(\mathcal{F}(P_\Gamma), G) \\ \mathrm{Supp}\, \nu = P_\Gamma}} e^{-S(\nu)}.$$

Here $\mathcal{P}_c$ denotes the set of coconnected finite plaquette sets, $P_\Gamma = \cup\{P \mid P \in \Gamma\}$, and $P \sim P'$ means that P and P' are not coconnected.

In order to enlighten the nature of the interaction between the plaquette sets of Γ, we note that if $\mathcal{F}(P_\Gamma)$ is a free product $\mathcal{F}(P_{\Gamma_1}) * \cdots * \mathcal{F}(P_{\Gamma_n})$ for an appropriate partition $\{\Gamma_1, \ldots, \Gamma_n\}$ of Γ, then $\Phi_k(\Gamma)$ factorizes $\Phi_k(\Gamma) = \Phi_k(\Gamma_1) \cdots \Phi_k(\Gamma_n)$. This corresponds to decomposing Γ in such a way that the parts $\Gamma_1, \ldots, \Gamma_n$ are completely unknotted, that is, they can be separated from each other arbitrarily far away without producing any intersection. Among such decompositions there exists a unique finest one, which we call the fundamental partition of Γ. For example, if $\Gamma = \{P_1, P_2\}$ with P_1 and P_2 being the circles of Fig. 1(a) or 1(b), then the fundamental partition of Γ is $\{P_1, P_2\}$ or $\{P_1 \cup P_2\}$, respectively. As a matter of fact, in case of Fig. 1(a), $\mathcal{F}(P_1 \cup P_2)$ is the free group generated by a_1 and a_2, where a_1 and a_2 are the homotopy classes of C_1 and C_2, respectively, therefore $\mathcal{F}(P_1 \cup P_2) = \mathbb{Z} * \mathbb{Z}$. On the other hand, $\mathcal{F}(P_1) = \mathcal{F}(P_2) = \mathbb{Z}$. In the case of Fig. 1(b), $\mathcal{F}(P_1 \cup P_2)$ is again generated by a_1 and a_2, but now a relation emerges: $a_1 a_2 a_1^{-1} a_2^{-1} = 1$. Thus $\mathcal{F}(P_1 \cup P_2) = \mathbb{Z}^2 \neq \mathbb{Z} * \mathbb{Z}$. The subsets $\Gamma \subset \mathcal{P}_c$, the fundamental partition of which consists of one element, namely, P_Γ, will be called knots [e.g., the ones on Figs. 1(b) and 2]. These knots are the clusters that we were looking for. The many-body interaction $\Phi_k(\Gamma)$ can then be interpreted as a knotting interaction.

There is a nice explanation of why the knotting interaction is absent in Abelian gauge theories. Using the Wirtinger presentation[1] of the fundamental group, one finds that all the relations that connect generators belonging to different coconnected components are generated by commutators; therefore, they are mapped to the unit element by any $\nu \in \mathrm{Hom}(\mathcal{F}(P), G)$.

III. THE CLUSTER EXPANSION

Formula (1) represents the weak coupling gauge theory as a gas of vortex lines (sheets, etc.), i.e., a gas of the plaquette sets $P \in \mathcal{P}_c$, with the two-body hard core interaction Φ_{hc} and the many-body knotting interaction Φ_k. In such a system one is interested in calculating the correlation functions

$$\rho(\Gamma)=\frac{1}{Z}\sum_{\substack{\Gamma'\subset\mathscr{P}_c\\ \Gamma'\supset\Gamma}}\Phi(\Gamma'),\quad \Gamma\subset\mathscr{P}_c. \tag{2}$$

These correlation functions satisfy the Kirkwood–Salsburg type of equation

$$\rho(\Gamma)=\Phi(\Gamma)+\sum_{\Gamma'\subset\mathscr{P}_c}\mathbf{K}(\Gamma,\Gamma')\rho(\Gamma'), \tag{3}$$

where the kernel $\mathbf{K}$ has the form

$$\mathbf{K}(\Gamma,\Gamma')=\Phi(\Gamma)\left\{\sum_{\substack{\Gamma_0\subset\Gamma\\ \Gamma_0\neq\phi}}(-1)^{|\Gamma_0|}\delta_{\Gamma_0\Gamma'}+\sum_{\Gamma_1\subset\mathscr{P}_c}\mathbf{K}_0(\Gamma,\Gamma_1)\times\sum_{\Gamma_0\subset\Gamma}(-1)^{|\Gamma_0|}\delta_{\Gamma_0\cup\Gamma_1\Gamma'}\right\},$$

where

$$\mathbf{K}_0(\Gamma,\Gamma_1)=\delta_{\Gamma\cap\Gamma_1,\phi}(1-\delta_{\Gamma_1,\phi})\times\sum_{\Gamma_0\subset\Gamma_1}(-1)^{|\Gamma_1\setminus\Gamma_0|}\frac{\Phi(\Gamma\cup\Gamma_0)}{\Phi(\Gamma)\Phi(\Gamma_0)}.$$

Using standard methods[4] one can show that the iterative solution of Eq. (3) converges if there is an upper bound on the generator function of the number of knots of the form

$$\sum_{\substack{\Gamma\in\mathscr{K}\\ p_0\in P_\Gamma}}e^{-b\|\Gamma\|}<F(b),\quad b\geqslant b_0. \tag{4}$$

Here $\mathscr{K}$ denotes the set of knots, $\|\Gamma\|=\Sigma_{P\in\Gamma}|P|$ is the "length" of Γ, and p is a fixed plaquette.

By analyzing the infinite hierarchy to be found inside the knots we introduced the notion of "the order of a knot" and then proved recursively in the order that an upper bound $F(b)=F(0)\exp\{-b(2d-2)\}$ satisfies (4) (e.g., $b_0=5.87$ for $d=3$).

Our main result[5] can be formulated in the following way. Let the action have the form

$$S(U)=\beta\sum_p\chi(dU_p),$$

where χ is a positive linear combination of characters. Let

$$\Delta=\min_{g\in G\setminus\{1\}}[\chi(1)-\chi(g)].$$

Then there exists a constant $\beta_0(d)$, depending only on the dimension d such that for arbitrary finite gauge group G the cluster expansion converges for $\beta>[\ln(|G|-1)+\beta_0(d)]/\Delta$ [e.g., $\beta_0(3)=6.08$]. This implies exponential clustering and analyticity of the correlation functions in the variable $\exp(-\beta)$.

IV. OUTLOOK

There are several possible applications of this weak coupling cluster expansion. One of them would be the construction of charged sectors with non-Abelian gauge charge in the weak coupling regime. Physically this would correspond to a discrete example of what is called the free gluon in QCD. Operators creating the free "gluons" can be constructed in these discrete models if the dimension $d\geqslant 4$ using the nonlocal fields introduced in Ref. 6. By studying such a model one hopes to get an answer to the important question: in what sense can a non-Abelian gauge charge be observed?

In the "free charge" phase the gas of knots is dilute and the average size of the knots is finite. So it seems natural to suppose that a confinement transition occurs when the knots percolate. However, one can imagine at least three types of percolation: (A) the probability that a plaquette belongs to an infinite coconnected component is positive, (B) there is no percolation in the sense of A, but the probability that a plaquette belongs to an infinite knot of infinite order (e.g., a closed chain whose links are closed chains whose links are...) is positive, and (C) the same as case (B) but with an infinite knot having finite order (e.g., a long chain made of infinitely many finite circles). It requires further study as to which one of the above types of percolations is responsible for confinement and which one describes a possible intermediate phase.

At the end let us mention how gauge symmetry breaking could be explained in terms of the knotting of vortices. Let us call the "flux group" of a knot the subgroup of G one obtains as the image of the fundamental group under a homomorphism ν. This is the group generated by the fluxes flowing through the various parts of the knot (holonomy group). There are two mechanisms that increase the flux group. The longer the vortex the higher the probability of finding a knot on it.[7] A knotted vortex, however, has more flux degrees of freedom. For example, in three dimensions, if G is the tetrahedron group A_4, an unknotted circle can have only $Z(2)$ (in three cases) or $Z(3)$ (in eight cases) as its flux group, while a circle with a trefoil knot yields $Z(2)$ or $Z(3)$ in 11 cases and A_4 in 24 cases. If the gauge group is larger, let us say S_4, then the trefoil knot seems to be not "very knotted," in the sense that its flux group is a proper subgroup of S_4 in 71 cases while it is equal to S_4 in only 24 cases. (This is only an entropy argument, so the energy functional may modify these probabilities to some extent.) The other mechanism is the branching of vortices that obviously increases the flux group. At strong coupling, where one has an (A) type of percolation, this latter mechanism will bring about the effect that the flux group of the infinite cluster is G with probability 1. The mechanism that works against these two and decreases the flux group is the linking of vortices. As an example, Fig. 1 shows linking implying constraints between the fluxes of different vortices. Suppose that there is a phase with a (C) type of percolation, where the typical vortex is not very knotted, so its flux group (with highest probability) is a proper subgroup H of G or one of its conjugates. Then the linkings in the infinite cluster may give rise to strong flux correlations, even at large distances. Up to a global G symmetry the system will behave as if it were a gauge theory with gauge group H.

ACKNOWLEDGMENTS

We have benefitted very much from discussions with A. Szücs. One of us (K.S.) was supported by the Alexander von Humboldt Foundation.

[1] J. Stillwell, *Classical Topology and Combinatorial Group Theory*, Graduate texts in mathematics (Springer, New York, 1980), Vol. 72; R. H. Crowell and R. H. Fox, *Introduction to Knot Theory*, Graduate texts in mathemat-

ics (Springer, New York, 1963), Vol. 57.
[2]E. C. Zeeman, Ann. Math. **78**, 501 (1963).
[3]E. Seiler, *Gauge Theories as a Problem of Constructive Quantum Field Theory and Statistical Mechanics, Lecture Notes in Physics* (Springer, Berlin, 1982), Vol. 159.
[4]D. Ruelle, *Statistical Mechanics* (Benjamin, New York, 1969).
[5]K. Szlachányi and P. Vecsernyès, unpublished.
[6]K. Szlachányi, Commun. Math. Phys. **108**, 319 (1987).
[7]D. W. Sumners and S. G. Whittington, J. Phys. A: Math. Gen. **21**, 1689 (1988).

String wave equations in Polyakov's path integral framework

Luiz C. L. Botelho
Departamento de Física, Universidade Federal do Pará, Belem, Para, Brazil

(Received 3 January 1989; accepted for publication 29 March 1989)

String wave equations in Polyakov's path integral framework for string quantization are studied. This analysis is applied to formally solve the quantum chromodynamic [SU(∞)] (bosonic) contour wave equation by means of a self-avoiding string theory possessing intrinsic fermionic degrees of freedom.

I. INTRODUCTION

In the Feynman path integral formulation for (first) quantization of a physical system,[1] the central object is the transition amplitude for the system evolution from a prescribed initial state to a prescribed final state. Its explicit expression is given by the continuous sum over all system trajectories connecting these states and weighted by the classical system action. This quantization procedure does not rely on the conventional operator Heisenberg–Schrödinger formulation of quantum mechanics. However, for most of the physical systems analyzed up to the present time, the formal equivalence between these two alternatives is implemented by showing that the above-mentioned Feynman transition amplitude satisfies the associated wave equation obtained from the operator approach.

The purpose of this paper is to describe a simple procedure for writing string wave equations directly from the Feynman path integral for the covariant bosonic and fermionic string transition amplitude presented by Polyakov some years ago.[2] In Sec. II we present our ideas in the simple case of covariant particle dynamics. The reason for writing wave equations in the Polyakov path integral is that it may shed some light on the role of the Liouville conformal freedom degree in the string quantization below the critical dimension. This study is presented in Sec. III. Another more important motivation is that the quantum chromodynamic [SU(∞)] (bosonic) contour average satisfies a closed stringlike evolution equation.[3] With a general procedure for writing string wave equations directly from the string path integral, the search for its (string) solutions becomes a simple and transparent task. This analysis is presented in Sec. IV. Finally in Sec. V we deduce a kind of Dirac–Ramond–Marshall string wave equation by extending the bosonic path integral formalism to the fermionic case.

II. THE WAVE EQUATION IN COVARIANT PARTICLE DYNAMICS

In the covariant description of a relativistic bosonic particle,[4] the particle trajectory is described by two degrees of freedom: the usual vector position $X_\mu(\zeta)$, with $0<\zeta<1$, and an additional one-dimensional metric $e(\zeta)$. The parameter ζ describes the evolution of the system and the particle trajectory $X_\mu(\zeta)$ does not change its orientation in space-time $[X_\mu(\zeta)\neq X_\mu(\zeta'),\ \zeta\neq\zeta']$ (see Ref. 1).

The covariant classical action for this particle, moving under the influence of an external potential $V(x)$, is given by

$$S[X_\mu(\zeta),V(x)]=\int_0^1\left(\frac{1}{2}\frac{\dot{X}_\mu(\zeta)^2}{e(\zeta)}+\frac{1}{2}m^2e(\zeta)+e(\zeta)V[X_\mu(\zeta)]\right), \tag{1}$$

where m^2 is the particle mass.

Following Feynman, the transition amplitude for which a particle initial state $(X_\mu^{\rm in},e^{\rm in})$ propagates to a final state $(X_\mu^{\rm out},e^{\rm out})$ is given explicitly by the path integral:

$$G[(X_\mu^{\rm out},e^{\rm out});(X_\mu^{\rm in},e^{\rm in})]$$

$$=\int_{\substack{X_\mu(0)=X_\mu^{\rm in}\\ X_\mu(1)=X_\mu^{\rm out}}}d\mu[X_\mu(\zeta)]\int_{\substack{e(0)=e^{\rm in}\\ e(1)=e^{\rm out}}}d\mu[e(\zeta)]\times\exp\{-S[X_\mu(\zeta),V(x)]\}. \tag{2}$$

Here the covariant Feynman measures $d\mu[e(\zeta)]$ and $d\mu[X_\mu(\zeta)]$ are, respectively, defined as the volume element of the covariant functional metrics

$$\|\delta e\|^2=\int_0^1(\delta e\delta e)(\zeta)d\zeta$$

and

$$\int_0^1 e(\zeta)(\delta X_\mu\cdot\delta X_\mu)d\zeta.$$

It is possible to evaluate explicitly the above transition amplitude in the proper-time gauge $e(\zeta)=$ const, thus producing the (Euclidean) Green's function of the Klein–Gordon operator in the presence of the external potential $V(x)$.

An alternative way to obtain the above result is by closely following Feynman,[1] and by considering the identity that results by making variations of the intrinsic metric at the end-point trajectory. Since a gauge exists where $e(\zeta)$ can be fixed as the trajectory proper-time parameter, we expect that this identity should produce a covariant wave equation that (in the proper time gauge) reduces to the usual Klein–Gordon equation (see the Appendix of Ref. 1).

As a consequence of the invariance under functional translations of the functional measure $d\mu[e(\zeta)]$, we show that the following relation holds true:

$$0=\int_{\substack{X_\mu(0)=X_\mu^{\rm in}\\ X_\mu(1)=X_\mu^{\rm out}}}d\mu[X(\zeta)]\int_{\substack{e(0)=e^{\rm in}\\ e(1)=e^{\rm out}}}d\mu[e(\tau)]\times\exp\{-S[X_\mu(\zeta),V(x)]\}. \tag{3}$$

 0022-2488/89/092160-08$02.50

By considering the boundary $\bar{\zeta}\to 0$ in Eq. (3) we show that transition amplitude, Eq. (2), satisfies the identity

$$\int_{\begin{pmatrix} X_\mu(0)=X_\mu^{\rm in} \\ X_\mu(1)=X_\mu^{\rm out}\end{pmatrix}} d\mu[X_\mu(\zeta)] \times \int_{\begin{pmatrix} e(0)=e^{\rm in} \\ e(1)=e^{\rm out}\end{pmatrix}} d\mu[e(\zeta)]\exp\{S[X_\mu(\zeta),V(x)]\} \lim_{\zeta\to 0^+}\left(\Pi_\mu^2(\bar{\zeta})-\frac{1}{2}m^2-V(x_\mu(\bar{\zeta}))\right), \tag{4}$$

where $\Pi_\mu(\zeta)=\dot{X}_\mu(\zeta)/e(\zeta)$ denotes the classical canonical momentum of the covariant particle.

In order to translate the path integral constraint equation (4) into an operator statement, we have to use the covariant Heisenberg commutation relation

$$[\Pi_\mu(\zeta),X_\nu(\zeta')] = -[i/e(\zeta)']\delta(\zeta'-\zeta)\delta_{\mu\nu} \quad (i=\sqrt{-1}),$$

which in the Schrödinger representation is given explicitly by

$$\Pi_\mu(\zeta) = -\frac{i}{e(\zeta)}\frac{\delta}{\delta X_\mu(\zeta)}.$$

After fixing the particle proper-time gauge [since Eq. (4) is invariant under the group of the trajectories reparametrization] and taking into account that the particle trajectory does not self-intersect in "time" [$X_\mu(\zeta)\neq X_\mu(\zeta')$ if $\zeta\neq\zeta'$], we finally obtain that Eq. (4) reduces to the Klein–Gordon wave equation in the presence of the external potential $V(x)$, namely,

$$(-\Box_{X^{\rm in}}+\tfrac{1}{2}m^2-V(X^{\rm in}))G(X^{\rm out};X^{\rm in})=0. \tag{5}$$

It is instructive to point out that by considering functional variations of the functional metric $d\mu[X_\mu(\zeta)]$ we obtain constraints without dynamical content that are associated to the invariance of the theory under the action of the space-time translation Poincaré group.

III. THE WAVE EQUATION IN THE COVARIANT BOSONIC STRING DYNAMICS

The basic object in the Polyakov approach[2,5] for the string covariant quantization (in the trivial topological sector) is that the following transition amplitude for an initial string state

$$C^{\rm in}=\{(X_\mu^{\rm in}(\sigma),e^{\rm in}(\sigma));\quad 0\leqslant\sigma\leqslant 1\}$$

propagates to a final string state $[C^{\rm out}=\{(X_\mu^{\rm out}(\sigma),e^{\rm out}(\sigma))\}]$

$$G[e^{\rm out};c^{\rm in}]=\int d\mu[g_{ab}]d\mu[\phi_\mu]e^{-I_0(g_{ab},\phi_\mu)}, \tag{6}$$

where the covariant string action is given by

$$I_0(g_{ab},\phi_\mu)=\int_D(\tfrac{1}{2}\sqrt{g}g^{ab}\partial_a\phi^\mu\partial_b\phi^\mu+\mu_0^2\sqrt{g})(\sigma,\zeta)d\sigma\,d\zeta. \tag{7}$$

The string surface parameter domain is taken to be the rectangle $D=\{(\sigma,\zeta),\ 0\leqslant\sigma\leqslant 1,\ 0\leqslant\zeta\leqslant T\}$. The functional measures $d\mu[g_{ab}]$ and $d\mu[\phi_\mu]$ are defined over all cylindrical quantum surfaces without holes and handles having as a boundary the string end configurations $\{C^{\rm in};C^{\rm out}\}$; i.e., $\phi_\mu(\sigma,0)=X_\mu^{\rm in}(\sigma)$ and $\phi_\mu(\sigma,T)=X_\mu^{\rm out}(\sigma)$. The intrinsic metric $\{g_{ab}(\sigma,\zeta)\}$ (which, roughly, plays the role of the covariant string proper-time parameter) can be chosen to satisfy the conformal gauge

$$g_{ab}(\sigma,\zeta)=\exp\beta(\sigma,\zeta)\delta_{ab}$$

and the initial end-point boundary condition $e^{\rm in}(\sigma)=\exp(\beta(\sigma,0))$.

At this point a fundamental difference appears between the string and particle case (Sec. I). In the last case it is always possible to fix the proper-time gauge $e(\zeta)=\text{const}=1$, where the intrinsic metric decouples from the dynamical description of the theory. This result reveals itself in the form of the associated wave equation [Eq. (5), Sec. II], where it does not have any functional dependence on the intrinsic metric. This decoupling phenomenon will not happen in the string case due to the conformal anomaly of the theory[2,5] unless it is canceled. Further, the associated string wave equation will depend on the intrinsic Liouville field at the boundary $\beta(\sigma,0)=\beta^{\rm in}(\sigma)$, as we will show explicitly below.

Let us now proceed as in the particle case by considering the following identity related to the integrand invariance under translations in the conformal factor $\beta(\sigma,\zeta)$ functional space $[g_{ab}(\sigma,\zeta)=\exp(\beta(\sigma,\zeta)\delta_{ab})]$ in the string propagator Eq. (6):

$$\int D[\beta(\sigma,\zeta)]\exp\left\{-\frac{26}{48\pi}\int_D\left(\frac{1}{2}(\partial_a\beta)^2+\frac{1}{2}\mu_R^2e^\beta\right)\right\} \lim_{\bar{\zeta}\to 0^+}\left(e^{-\beta(\bar{\sigma},\bar{\zeta})}\frac{\overset{\leftrightarrow}{\delta}}{\delta\beta(\bar{\sigma},\bar{\zeta})}\delta_{ab}\right)F(\phi_\mu,g_{ab}), \tag{8}$$

where

$$F(\phi_\mu,g_{ab})=\int d\mu[\phi_\mu]\exp(-I_0(\phi_\mu,g_{ab})) \tag{9}$$

denotes the pure string vector position term in Eq. (6).

It is worthwhile to remark that this procedure for deducing a dynamical (wave) equation is the two-dimensional analog of that used to write the Wheeler–DeWitt equation four-dimensional quantum gravity from the path integral expression for the universe propagator.[6]

The variation associated to the Faddeev–Popov term is given by

$$\int D[\beta(\sigma,\zeta)]\left\{-\frac{26}{28\pi}\int_D\left(\frac{1}{2}(\partial_a\beta)^2+\frac{1}{2}\mu^2e^\beta\right)\times(\sigma,\zeta)d\sigma\,d\zeta\right\}\frac{26}{24\pi}(R(e^\beta)+\mu^2)(\bar{\sigma},\bar{\zeta})F(\phi_\mu,g_{ab}), \tag{10}$$

where $R(e^\beta)=-(e^{-\beta}\Delta\beta)(\sigma,\zeta)$ denotes the scalar of curvature associated to the metric $g_{ab}(\sigma,\zeta)=\exp(\beta(\sigma,\zeta))\delta_{ab}$.

The $\delta/\delta\beta(\bar{\sigma},\bar{\zeta})$ functional derivative of the term $F(\phi_\mu,g_{ab}=e^\beta\delta_{ab})$ is more subtle since the covariant functional measure $d\mu[\phi_\mu]$ [see Eq. (9) of Ref. 2] depends in a nontrivial way on the conformal factor $\beta(\sigma,\zeta)$ as a consequence of its definition as the functional volume element associated to the covariant functional metric

$$||\delta\phi^{\mu}|| = \int_D (e^{\beta}\delta\phi^{\mu}\delta\phi^{\mu})(\sigma,\zeta)d\sigma\, d\zeta. \tag{11}$$

Its evaluation proceeds in the following way:

$$d\mu[\phi^{\mu},(e^{\delta h+\beta})\delta_{ab}] - d\mu[\phi^{\mu},e^{\beta}\delta_{ab}]$$
$$\overset{\text{def}}{=} \frac{\delta}{\delta\beta} d\mu[\phi^{\mu},e^{\beta}\delta_{ab}] + O(h^2). \tag{12}$$

Since, as a consequence of Eq. (11), we have the result

$$d\mu[\phi^{\mu},e^{\delta h+\beta}\delta_{ab}] = d\mu[e^{\delta h/2}\phi^{\mu},e^{\beta}\delta_{ab}], \tag{13}$$

and the effect of the functional string vector position measure under a conformal scale was evaluated exactly by Fujikawa [see Eqs. (38) and (39) in Ref. 5],

$$d\mu[e^{\delta h/2}\phi^{\mu},e^{\beta}\delta_{ab}]$$
$$= \exp\left\{\frac{D}{48\pi}\int_D \frac{1}{2}(\partial_a\beta)^2 + \frac{1}{2}\mu^2 e^{\beta})\delta h\right\} d\mu[\phi^{\mu},e^{\beta}\delta_{ab}], \tag{14}$$

we thus have the following result by taking $h(\sigma,\zeta) = \epsilon\delta(\sigma-\bar{\sigma})\delta(\zeta-\bar{\zeta})$ and considering the linear term in ϵ:

$$\frac{\delta}{\delta\beta(\bar{\sigma},\bar{\zeta})} d\mu[\phi^{\mu},e^{\beta}\delta_{ab}]$$
$$= \frac{1}{\epsilon}\lim_{\epsilon\to 0^+}(d\mu[\phi^{\mu},e^{\delta h+\beta}\delta_{ab}] - d\mu[\phi^{\mu},e^{\beta}\delta_{ab}])$$
$$= (D/24\pi)(R(e^{\beta(\bar{\sigma},\bar{\zeta})}) + \mu^2)\times d\mu[\phi^{\mu},e^{\beta}\delta_{ab}]. \tag{15}$$

Finally the term $[\delta/\delta\beta(\bar{\sigma},\bar{\zeta})]\, I_0(\phi^{\mu},g_{ab} = e^{\beta}\delta_{ab})$ is given by the diagonal component of the string energy momentum tensor:

$$\left(e^{-\beta}\frac{\delta}{\delta\beta}\right)(I_0(\phi^{\mu},g_{ab} = e^{\beta}\delta_{ab}))(\bar{\sigma},\bar{\zeta})$$
$$= ((\partial_{\zeta}\phi^{\mu})^2 - (\partial_{\sigma}\phi^{\mu})^2)(\bar{\sigma},\bar{\zeta}). \tag{16}$$

By grouping together Eqs. (10), (15), and (16), we obtain that the string transition amplitude in the conformal gauge satisfies the dynamic constraint

$$O = \int d\mu[g_{ab}]|_{g_{ab}=e^{\beta}\delta_{ab}} \int d\mu[\phi_{\mu}]$$
$$\times\exp(-I_0(g_{ab},\phi_{\mu}))\left\{\frac{26-D}{48\pi}\lim_{\bar{\zeta}\to 0^+}(R((\bar{\sigma},\bar{\zeta})) + \mu^2)\right.$$
$$\left. + (\tfrac{1}{2}\Pi^{\text{in}}_{\mu}(\bar{\sigma})^2 - \tfrac{1}{2}|X'^{\mu}_{\text{in}}(\bar{\sigma})|^2)\right\}, \tag{17}$$

where $d\mu[g_{ab}]|_{g_{ab}=e^{\beta}\delta_{ab}}$ means that the functional measure over the intrinsic metric field $\{g_{ab}(\bar{\sigma},\bar{\zeta})\}$ is defined in the conformal gauge,

$$\Pi^{\text{in}}_{\mu}(\bar{\sigma}) = \lim_{\bar{\zeta}\to 0^+}\partial_{\bar{\zeta}}\phi_{\mu}(\bar{\sigma},\bar{\zeta})$$

denotes the string canonical momentum and

$$X'^{\mu}_{\text{in}}(\bar{\sigma}) = \lim_{\bar{\zeta}\to 0^+}\partial_{\bar{\sigma}}\phi_{\mu}(\bar{\sigma},\bar{\zeta}).$$

In order to translate the above path integral relation into a wave equation form,[7] we introduce covariant string commutation relations[8]

$$[\Pi^{\text{in}}_{\mu}(\bar{\sigma}),X^{\nu}(\bar{\sigma}')] = [1/\hbar^{(D)}][1/e^{\text{in}}(\bar{\sigma})]\delta(\bar{\sigma}-\bar{\sigma}'), \tag{18}$$

with $\hbar^{(D)}$ being the Planck constant in the physical space-time R^D. Using the Schrödinger representation for this commutation relation,

$$\Pi^{\text{in}}_{\mu}(\sigma) = \frac{i}{\hbar^{(D)}e^{\text{in}}(\sigma)}\frac{\delta}{\delta X^{\text{in}}_{\mu}(\sigma)}, \tag{19}$$

we can express Eq. (17) in the following form, which generalizes the usual $D = 26$ Nambu–Virasoro wave equation[7]:

$$\left\{-\frac{1}{2}\frac{e^{-2\beta_{\text{in}}(\sigma)}}{(\hbar^D)^2}\frac{\delta^2}{\delta X^{\text{in}}_{\mu}(\sigma)\delta X^{\text{in}}_{\mu}(\sigma)} - \frac{1}{2}|X'^{\text{in}}(\sigma)|^2\right.$$
$$\left. + \frac{26-D}{24\pi}\left(-\frac{1}{2}\Pi^{\text{in}}_{\beta}(\sigma)^2 - \frac{1}{2}\beta'_{\text{in}}(\sigma)^2 + \frac{1}{2}\mu^2 e^{\beta_{\text{in}}(\sigma)}\right)\right\}$$
$$\times G((X^{\text{in}}_{\mu}(\sigma),e^{\beta_{\text{in}}(\sigma)});(X^{\text{out}}_{\mu}(\sigma),e^{\beta_{\text{out}}(\sigma)})) = 0, \tag{20}$$

where we have written the conformal contribution in Eq. (17) in the Polyakov proposed Liouville Hamiltonian,[2] with

$$\Pi^{\text{in}}_{\beta}(\sigma) = \lim_{\zeta\to 0^+}\partial_{\zeta}\beta(\sigma,\zeta)$$

being the canonical momentum associated with the Liouville field $\beta(\sigma,\zeta)$ at the boundary. We note that it has the following representation:

$$\Pi^{\text{in}}_{\beta}(\sigma) = \frac{i}{\hbar^{(2)}}\frac{\delta}{\delta\beta_{\text{in}}(\sigma)}. \tag{21}$$

Here $\hbar^{(2)}$ now denotes the Planck constant associated with two-dimensional string space-time D.

It is worth mentioning that the dynamical status acquired by the metric $g_{ab}(\sigma,\zeta) = \exp(\beta(\sigma,\zeta))\delta_{ab}$ in Eq. (20) induced pure quantum gravity in D as a result of the dynamical breaking of the complete diffeomorphism ground of the action in Eq. (7), denoted by $G_{\text{diff}}(D)$, to the subgroup $G_{\text{diff}}(D)/G_{\text{Weil}}(D)_{\text{diff}}$, where $G_{\text{Weil,diff}}(D)$ is the subgroup of $G_{\text{diff}}(D)$ that acts on the metric field as a Weil scaling.

As a consequence of these remarks we can see that only at $D = 26$ can we choose the proper time string gauge $g_{ab}(\sigma,\zeta) = \delta_{ab}$ in an analogous way as in covariant particle dynamics (see Sec. II), since now the invariance of the theory under $G_{\text{diff}}(D)$ is preserved by quantization.

IV. A STRING SOLUTION FOR THE QCD [SU(∞)] BOSONIC CONTOUR AVERAGE EQUATION

There are several compelling arguments for the existence of a string representation for quantum chromodynamics (QCD) at the 't Hooft large number of colors. One of these arguments is that the QCD [SU(∞)] covariant loop average with an additional intrinsic global SO(M) flavor group (see Appendix A),

$$W_{ik}[\mathscr{C}_{X(-\pi),X(\pi)}]$$
$$= \frac{1}{N_c}\left\langle T_r^{\text{color}}\exp\left(i\oint_{\mathscr{C}_{X(-\pi),X(\pi)}} A_{\mu}(X_{\mu}(\sigma))\frac{dX_{\mu}(\sigma)}{e(\sigma)}\right)\right\rangle, \tag{22}$$

satisfies the following (formal) stringlike contour equation[3] [$e(\sigma) = 1$]:

$$\frac{\delta^{(2)}}{\delta X_\mu(\sigma)\delta X_\mu(\sigma)} W_{ik}[\mathscr{C}_{X(-\pi),X(\pi)}]$$
$$= \lambda_0^2 \oint_{\mathscr{C}_{X(-\pi),X(\pi)}} X'_\mu(\sigma)\delta^{(D)}(X_\mu(\sigma)-X_\mu(\bar\sigma))X'_\mu(\bar\sigma)$$
$$\times (W_{ij}[\mathscr{C}_{X(-\pi),X(\sigma)}]W_{jk}[\mathscr{C}_{X(\sigma),X(\pi)}])$$
$$-\gamma^2|X'_\mu(\sigma)|^2 W_{ik}[\mathscr{C}_{X(-\pi),X(\pi)}], \tag{23}$$

where the contour integral $\oint_{\mathscr{C}_{X(-\pi),X(\pi)}}$ means that the coincident $\sigma=\bar\sigma$ does not contribute for the integrand (Cauchy principal value).

It is thus conjectured that some sort of string propagator should solve Eq. (23) in some sense. Our aim in this section is to present an interacting string theory with an intrinsic fermionic structure that possesses as a string wave equation (in our proposed framework of Sec. III) Eq. (23) with a fixed flavor group SO(22).

Let us start our analysis by describing the covariant string action of our proposed QCD [SU(∞)] string:

$$S[\phi_\mu(\sigma,\zeta),\psi_{(k)}(\sigma,\zeta),g_{ab}(\sigma,\zeta)]$$
$$= S_0[\phi_\mu(\sigma,\zeta),g_{ab}(\sigma,\zeta)]$$
$$+ S_1[\psi_{(k)}(\sigma,\zeta),g_{ab}(\sigma,\zeta)]$$
$$+ S_{\rm int}[\phi_\mu(\sigma,\zeta),\psi_{(k)}(\sigma,\zeta),g_{ab}(\sigma,\zeta)], \tag{24}$$

where

$$S_0[\phi_\mu(\sigma,\zeta),g_{ab}(\sigma,\zeta)]$$
$$= \tfrac{1}{2}\left(\int_D (\sqrt{g}g_{ab}\,\partial_a\phi^\mu\,\partial_b\phi^\mu)(\sigma,\zeta)d\sigma\,d\zeta\right), \tag{25a}$$

$$S_1[\psi_{(k)}(\sigma,\zeta),g_{ab}(\sigma,\zeta)]$$
$$= \tfrac{1}{2}\int_D (\sqrt{g}\bar\psi_{(k)}\gamma_a(\sigma,\zeta)\partial_a\psi_{(k)})(\sigma,\zeta)d\sigma\,d\zeta, \tag{25b}$$

$$S_{\rm int}[\phi_\mu(\sigma,\zeta),\psi_{(k)}(\sigma,\zeta),g_{ab}(\sigma,\zeta)]$$
$$= \beta\left(\int_D d\sigma\,d\zeta\,\sqrt{g}(\bar\psi_{(k)}\psi_{(k)})\hat T^{\mu\nu}(\phi_\mu)(\sigma,\zeta)\right.$$
$$\left.\times\left(\int_D \sqrt{g(\bar\sigma,\bar\zeta)}\,\delta^{(D)}(\phi_\mu(\sigma,\zeta)-\phi_\mu(\bar\sigma,\bar\zeta))\right.\right.$$
$$\left.\left.\times\hat T^{\mu\nu}(\phi_\mu(\bar\sigma,\bar\zeta))\right)d\bar\sigma\,d\bar\zeta\right). \tag{25c}$$

The notation is as follows: The bosonic degrees of freedom are $\{\phi_\mu(\sigma,\zeta),g_{ab}(\sigma,\zeta)\}$ as in Sec. II. Additionally we introduce a set of intrinsic two-dimensional Weyl spinors in the string surface and belonging to the SO(M) fundamental representation. They are denoted by $\{\psi_{(k)}(\sigma,\zeta),\ k=1,...,M\}$. We impose on them the Neumann boundary condition

$$\lim_{\zeta\to 0^+}\partial_\sigma\psi_{(k)}(\sigma,\zeta)=0.$$

The bosonic $\{\psi_k(\sigma,\zeta),g_{ab}(\sigma,\zeta)\}$ string sector interacts with fermionic $\{\psi_{(k)}(\sigma,\zeta)\}$ sector through a self-avoiding interaction involving the surface orientation tensor

$$\hat T^{\mu\nu}(\phi_\mu(\sigma,\zeta)) = (\epsilon^{ab}\partial_a\phi^\mu\,\partial_b\phi^\nu/\sqrt{h})(\sigma,\zeta),$$
$$h=\det h_{ab},\quad h_{ab}=\partial_a\phi^\mu\,\partial_b\phi^\nu,$$

and an attractive ($\beta<0$) delta function potential supported at the self-intersecting lines of the string surface. These nontrivial self-intersections are supposed to arise at those submanifolds where $X_\mu(\sigma,\zeta)=X_\mu(\sigma',\zeta')$ with $\sigma\neq\sigma'$ for every $\zeta\in[0,T]$. We notice that self-intersections of the form $X_\mu(\sigma,\zeta)=X_\mu(\sigma,\zeta)$ with $\zeta\neq\zeta'$ arise only in the case where the string surface possesses holes and handles, which is not the case here.

After having described our string theory, we consider the following O(M) string transition amplitude[9]:

$$Z_{kl}[\mathscr{C}_{X(-\pi),X(\pi)}] = \int d\mu[g_{ab}]d\mu[\phi_\mu]$$
$$\times d\mu[\psi_{(k)}](\psi_{(k)}(-\pi,0)\bar\psi_l(\pi,0))$$
$$\times\exp\{-S[\phi_\mu,\psi_{(k)},g_{ab}]\}. \tag{26}$$

In order to write the wave function equation associated with the above string Green's function, in the physical spacetime R^4, we proceed as in Sec. III by considering the analogous identity of Eq. (8), namely,

$$\int d\mu[g_{ab}]d\mu[\phi_\mu]d\mu[\psi_{(k)}](\psi_{(k)}(-\pi,0)\bar\psi_{(l)}(\pi,0))\exp\{-S[\phi_\mu,\psi_{(k)},g_{ab}]\}$$
$$\times(\tfrac{1}{2}\Pi_\mu^{\rm in}(\sigma)^2-\tfrac{1}{2}|X'_\mu(\sigma)|^2+\lim_{\tau\to 0^+}(\bar\psi_{(k)}\gamma_1\partial_\sigma\psi_{(k)})(\sigma,\zeta))$$
$$=\frac{\beta}{2}\int_{-\pi}^{\pi} d\bar\sigma\,X'^\mu(\sigma)\delta^{(D)}(X_\mu(\sigma)-X_\mu(\bar\sigma))X'^\mu(\bar\sigma)\int d\mu[g_{ab}]d\mu[\phi_\mu]d\mu[\psi_{(k)}]$$
$$\times\left(\sum_{(p)=1}^{22}\psi_{(p)}\bar\psi_{(p)}\right)(\sigma,0)(\psi_{(k)}(-\pi,0)\bar\psi_{(l)}(\pi,0))\exp\{-S[\phi_\mu,\psi_{(k)},g_{ab}]\}. \tag{27}$$

Our choice of the intrinsic "flavor" group to be SO(22) is dictated by the fact that the QCD [SU(∞)] string should preserve the full invariance under the diffeomorphism group and this happens only in the case where the conformal anomaly of the theory vanishes (see Sec. III). Since, in our proposed theory ($D=4$), the anomalous term is proportional to $[26-(D+M)]/24\pi$ we see that only for $M=22$ can we preserve the above-mentioned symmetry.

We thus can rewrite Eq. (27) in the form

$$\left(-\frac{1}{2}\frac{\delta^{(2)}}{\delta X_\mu(\sigma)\delta X_\mu(\sigma)}-\frac{1}{2}|X'_\mu(\sigma)|^2\right)Z_{kl}[\mathscr{C}_{X(-\pi),X(\pi)}]$$

$$= \frac{\beta}{2} \int_{-\pi}^{\pi} d\bar{\sigma}\, X'_{\mu}(\sigma)\delta^{(D)}(X_{\mu}(\sigma) - X_{\mu}(\bar{\sigma}))X'_{\mu}(\bar{\sigma})(Z_{kp}[\mathscr{C}_{X(-\pi),X(\sigma)}]Z_{pl}[\mathscr{C}_{X(\sigma),X(\pi)}]), \tag{28}$$

where we have used the string measure factorization properties

$$\int \prod_{\substack{-\pi<\beta<\pi \\ 0<\zeta<T \\ 1<k<22}} (d\psi_{(k)}(\beta,\zeta))(\psi_{(k)}(-\pi,0)\bar{\psi}_{(l)}(\pi,0))(\psi_{(p)}(\sigma,0)\bar{\psi}_{(p)}(\sigma,0))\exp\{-S[\phi_{\mu},\psi_{(k)},g_{ab}]\}$$

$$= \int \prod_{\substack{-\pi<\beta<\sigma \\ 0<\zeta<T \\ 1<k<22}} (d\psi_{(k)}(\beta,\zeta))(\psi_{(k)}(-\pi,0)\bar{\psi}_{(p)}(\sigma,0))\exp\{-S^{(1)}[\phi_{\mu},\psi_{(k)},g_{ab}]\}$$

$$\times \int \prod_{\substack{\sigma<\beta<\pi \\ 0<\zeta<T \\ 1<k<22}} (d\psi_{(k)}(\beta,\zeta))(\psi_{(p)}(\sigma,0)\bar{\psi}_{(l)}(\pi,0))\exp\{-S^{(2)}[\phi_{\mu},\psi_{(k)},g_{ab}]\} \tag{29a}$$

and

$$\int \left(\prod_{\substack{-\pi<\beta<\pi \\ 0<\tau<T}} d\phi^{\mu}(\beta,\zeta)|_{\phi^{\mu}(\beta,0)=\mathscr{C}_{X(\pi),X(-\pi)}} \right) \exp\left\{ -\frac{1}{2} \int_{D_{[-\pi,\pi]\times[0,T]}} (\partial_a \phi^{\mu})^2 \right\}$$

$$= \int \left(\prod_{\substack{-\pi<\beta<\sigma \\ 0<\zeta<T}} d\phi^{\mu}(\beta,\zeta)|_{\phi^{\mu}(\beta,0)=\mathscr{C}_{X(-\pi),X(\sigma)}} \right) \exp\left\{ -\frac{1}{2} \int_{D_{[-\pi,\sigma]\times[0,T]}} (\partial_a \phi^{\mu})^2 \right\}$$

$$\times \int \left(\prod_{\substack{\sigma<\beta<\pi \\ 0<\zeta<T}} d\phi^{\mu}(\beta,\zeta)|_{\phi^{\mu}(\beta,0)=\mathscr{C}_{X(\sigma),X(\pi)}} \right) \exp\left\{ -\frac{1}{2} \int_{D_{[\sigma,\pi]\times[0,T]}} (\partial_a \phi^{\mu})^2 \right\}. \tag{29b}$$

Here

$$\left(\prod_{\substack{-\pi<\beta<\pi \\ 0< \zeta<T}} d\phi^{\mu}(\beta,\zeta)|_{\phi^{\mu}(\beta,0)=\mathscr{C}_{X(-\pi)X(\pi)}} \right)$$

means that the functional integration is done with the boundary condition $\phi^{\mu}(\beta,0) = \mathscr{C}_{X(-\pi)X(\pi)}$.

We remark that these factorization properties hold true only in the case that the split string surfaces $\phi_{\mu}(D_{[-\pi,\sigma]\times[0,T]})$ and $\phi_{\mu}(D_{[\sigma,\pi]\times[0,T]})$ possess the same topology as in our case of trivial topology.

Let us now identify the string wave equation [Eq. (28)] with the QCD [SU(∞)] contour average equation [Eq. (23)]. The first step is to identify the SU(∞) gauge coupling constant λ_0^2 with the string interaction coupling $-\beta$. Second, we make the identification of the constant $-\gamma^2$ (the Euclidean gluon condensate—see Appendix A) with the Regge slope parameter $1\pi\alpha'$, which was adjusted to unity in our study.

After these coupling constant identifications we see that the Euclidean self-suppressing string theory should represent Euclidean QCD [SU(∞)] in the gauge invariant observable algebra (color singlet currents, espectrum, etc.).

V. THE NEVEU–SCHWARZ STRING WAVE EQUATION

Let us start by considering the open fermionic string action in a D-dimensional Euclidean space-time[10] ($\mu = 1,...D$, $(A) = 1,2$, $a = 1,2$):

$$S[\phi_{\mu}(\sigma,\zeta),\psi_{\mu}(\sigma,\zeta),e_a^{(A)}(\sigma,\zeta),\chi_a(\sigma,\zeta)]$$

$$= \int_D d\sigma\, d\zeta\, e(\sigma,\zeta)\left[\frac{1}{2}\partial_a\phi^{\mu}\,\partial_b\phi^{\mu}\,g^{ab} + \frac{1}{2} i\psi_{\mu}(\gamma\partial)\psi_{\mu}\right.$$

$$\left. -\frac{1}{2}F^2 - \frac{1}{2} i(\chi_a\gamma^b\gamma^a\psi^{\mu})\left(\partial_b\phi^{\mu} - \frac{1}{4} i\chi_b\psi^{\mu}\right)\right](\sigma,\zeta) + \text{boundary terms}. \tag{30}$$

Here the fermionic string is characterized by two (external) fields: the usual bosonic vector position $\phi^{\mu}(\sigma,\zeta)$ and the Majorana spinor $\psi^{\mu}(\sigma,\zeta)$ describing the string Lorentz spin. The presence of the vierbein $e_a^{(A)}(\sigma,\zeta)$ and of the two-dimensional vector Majorana spinor $\chi_a(\sigma,\zeta)$ together with the auxiliary scalar field $F(\sigma,\zeta)$ ensures, respectively, the action's invariance under general Lorentz and coordinate transformations together with the world-sheet local supersymmetric transformations.

Following Polyakov the (formal) fermionic string propagator is given by the following path integral connecting the initial C^{in} string state to a final string state C^{out}:

$$G[C^{\text{out}};C^{\text{in}}] = \int d\mu[\phi^{\mu},\psi^{\mu},e_a^{(A)},\chi_a]\exp\{-S[\phi^{\mu},\psi^{\mu},e_a^{(A)},\chi_a]\} \tag{31}$$

(here the boundary terms were absorbed in $G[C^{\rm out};C^{\rm in}]$).

In order to write dynamical wave equations we exploit the invariance under translations in the superconformal factor $(\varphi(\sigma,\xi);\zeta(\sigma,\xi))$ functional space of the fermionic string propagator [Eq. (31)]

$$[g_{ab}(\sigma,\xi)=\exp(2\varphi(\sigma,\xi)\delta_{ab}),\quad \chi_a(\sigma,\xi)=\gamma_a^{(B)}\zeta_{(B)}(\sigma,\xi)],$$

which produces the following identities:

$$\int d\mu[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]e^{-S[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]}\left(-\frac{\delta S[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]}{\delta\varphi(\bar\sigma,\bar\xi)}\right)=\int\left(\frac{\delta}{\delta\varphi(\bar\sigma,\bar\xi)}d\mu[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]\right)e^{-S[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]} \tag{32a}$$

and

$$\int d\mu[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]e^{-S[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]}\left(-\frac{\delta S[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]}{\delta\zeta_{(B)}(\bar\sigma,\bar\xi)}\right)=\int\left(\frac{\delta}{\delta\zeta_{(B)}(\bar\sigma,\bar\xi)}d\mu[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]e^{-S[\phi^\mu,\psi^\mu,e_a^{(A)},\chi_a]}\right). \tag{32b}$$

By noting that the fermionic string is defined at the quantum level only at $D=10$ (the so-called Neveu–Schwarz string) or at $D\to-\infty$,[11] we will consider $D=10$, which means that the functional measure variations in the right-hand side of Eqs. (32a) and (32b) vanish. In the superconformal gauge and using the Euclidean identity $\gamma_{(A)}\gamma_{(B)}=i\epsilon_{(A)(B)}\gamma_s$, we rewrite Eqs. (32a) and (32b) as

$$\int d\mu[\phi^\mu,\psi^\mu]e^{-S[\phi^\mu,\psi^\mu]}\frac{1}{2}((\partial_\xi\phi^\mu)^2-(\partial_\sigma\phi^\mu)^2+\psi^\mu\gamma_{(1)}\partial_\sigma\psi^\mu)(\bar\sigma,\bar\xi)=0, \tag{33a}$$

$$\int d\mu[\phi^\mu,\psi^\mu]e^{-S[\phi^\mu,\psi^\mu]}\frac{1}{2}(1+\gamma_5)\psi^\mu(\bar\sigma,\bar\xi)(\partial_\xi\phi^\mu-\partial_\sigma\phi^\mu)(\bar\sigma,\bar\xi)=0, \tag{33b}$$

respectively.

In order to translate the above-written string path integral identities into a wave equation form we take its boundary limit $\bar\xi\to0^+$ and translate the result into an operator equation by using the Schrödinger quantum representation

$$\lim_{\xi\to0^+}\partial_\xi\phi^\mu(\sigma,\xi)\Leftrightarrow\frac{i}{\hbar^{(D)}}\frac{\delta}{\delta\phi^\mu_{\rm in}(\sigma)}, \tag{34a}$$

$$\lim_{\xi\to0^+}\partial_\sigma\phi^\mu(\sigma,\xi)\Leftrightarrow\phi'^\mu_{\rm in}(\sigma), \tag{34b}$$

$$\lim_{\xi\to0^+}\psi^\mu(\sigma,\xi)\Leftrightarrow\Gamma^\mu_{\rm in}(\sigma). \tag{34c}$$

Here the quantum $C^{\rm in}$ string state in the operator framework is characterized by the coordinates $(\Gamma^\mu_{\rm in}(\sigma),\phi^\mu_{\rm in}(\sigma))$ where the $\Gamma^\mu_{\rm in}(\sigma)$ are string valued Dirac matrices obeying the space-time anticommuting relations[8]

$$\{\Gamma^\mu_{(A),{\rm in}}(\sigma),\Gamma^\nu_{(B),{\rm in}}(\sigma')\}=2\delta(\sigma-\sigma')\delta_{\mu\nu}\delta_{(A),(B)}.$$

By noting that the Neveu–Schwarz string fermion field $\psi^\mu(\sigma,\tau)$ satisfies the Neumann condition

$$\lim_{\tau\to0^+}\partial_\sigma\psi^\mu(\sigma,\tau)=0,$$

we obtain a fermionic string wave equation

$$D^{(\pm)}_{C^{\rm in}}G[C^{\rm in},C^{\rm out}]=0, \tag{35}$$

where

$$D^{(\pm)}_{C^{\rm in}}=\frac{1}{2}(1+\gamma_5)\left(\frac{i}{\hbar^{(D)}}\Gamma^\mu_{\rm in}\frac{\delta}{\delta\phi^\mu_{\rm in}}-\Gamma^\mu_{\rm in}\phi'^\mu_{\rm in}\right)(\sigma). \tag{36}$$

It is instructive to remark that in Eq. (35) the same $\Gamma^\mu_{\rm in}(\sigma)$ used in the momenta operator is also used in the string length factor $\phi'^{\rm in}_\mu(\sigma)$, opposite to the earlier proposed Ramond–Marshall fermionic string wave equation[8] where two different sets of $\Gamma^\mu(\sigma)$ matrices are used.

Finally we note that the formal anticommutator $\{D^{(\pm)}_{C^{\rm in}}(\sigma);D^{(\pm)}_{C^{\rm in}}(\sigma)\}$ is equal to the bosonic

$$-\frac{1}{2}\frac{\delta^{(2)}}{\delta\phi^{\rm in}_\mu(\sigma)\delta\phi^{\rm in}_\mu(\sigma)}-\frac{1}{2}|\phi'^{\rm in}_\mu(\sigma)|^2$$

string wave D'Alembertian since we have preserved the superdiffeomorphism group of the theory, which, in turn, manifests itself in the following constraint imposed in the physical Hilbert space of Neveu–Schwarz string states:

$$\left(\phi'^\mu_{\rm in}(\sigma)\frac{\delta}{\delta\phi^\mu_{\rm in}(\sigma)}\right)G[C^{\rm in},C^{\rm out}]=0. \tag{37}$$

ACKNOWLEDGMENT

This work was supported by CNPq, Brazil.

APPENDIX A: THE QCD (SU(∞)) BOSONIC CONTOUR AVERAGE

The basic dynamical variable in the loop space formulation for Euclidean QCD [SU(∞)] is the amplitude for a quark loop propagating in the quantum (confining) vacuum of a pure Yang–Mills field, since at the 't Hooft limit for a large number of colors the second-quantized quark matter effective action reduces to the quark first-quantized action, namely,[9]

$$\lim_{\substack{(g^2N_c)\,\text{fixed}\\ N_c\to\infty}}(\det(i\partial_\mu(\partial_\mu+A_\mu)))=\int d^DX\left(\sum_{\substack{\mathscr{C}_{X(\pi),X(-\pi)}\\ X(\pi)=X(-\pi)=X}}\langle{\rm Tr}\,U[\mathscr{C}_{X(\pi),X(-\pi)}]\rangle\right), \tag{A1}$$

where

$$U[\mathscr{C}_{X(-\pi),X(\pi)}] = P\left\{\exp\int_{-\pi}^{\pi} d\sigma\, A_\mu(X_\mu(\sigma))\frac{dX_\mu(\sigma)}{e(\sigma)}\right\} \tag{A2}$$

denotes the covariant Wu–Yang phase factor defined by the closed (covariant) quark trajectory

$$\mathscr{C}_{X(-\pi),X(\pi)} = \{(X_\mu(\sigma),e(\sigma));\ -\pi\leqslant\sigma\leqslant\pi\}$$

and representing the interaction of the pair with the Yang–Mills external field $A_\mu(x)$. The notation $\langle\ \rangle$ means the quantum average defined by the Yang–Mills functional integral at $N_c\to\infty$ (planar graphs).

In order to deduce a closed contour functional equation for the amplitude inside Eq. (A2), we remark the validity of the classical second-order functional derivatives results[3] $[e(\sigma)=1]$

$$\begin{aligned}
&\lim_{\sigma\to\sigma'}\frac{\delta^2}{\delta x_\mu(\sigma)\delta x_\mu(\sigma')}(\mathrm{Tr}\,U[\mathscr{C}_{X(-\pi),X(\pi)}])\\
&\quad=\lim_{\sigma\to\sigma'}\delta(\sigma-\sigma')\mathrm{Tr}((\nabla_\mu F_{\mu\nu})(x(\sigma))X'^{\nu}(\sigma)(U[\mathscr{C}_{X(\sigma),X(\pi)}]U[\mathscr{C}_{X(-\pi)X(\sigma)}]))\\
&\qquad+\lim_{\sigma\to\sigma'}\theta(\sigma-\sigma')\mathrm{Tr}(U[\mathscr{C}_{X(-\pi),X(\sigma')}]F_{\alpha\beta}(X(\sigma'))X'^{\beta}(\sigma')U[\mathscr{C}_{X(\sigma'),X(\sigma)}]F_{\alpha\rho}(X(\sigma))X'^{\rho}(\sigma)U[\mathscr{C}_{X(\sigma),X(\pi)}])\\
&\qquad+\lim_{\sigma'\to\sigma}\theta(\sigma'-\sigma)\mathrm{Tr}(\text{above written expression with }\sigma\text{ exchanged by }\sigma').
\end{aligned}\tag{A3}$$

By using that $\theta(\sigma'-\sigma)=\frac{1}{2}$ if $\sigma=\sigma'$ and imposing the loop periodicity property

$$U[\mathscr{C}_{X(a),X(a+2\pi)}] = U[\mathscr{C}_{X(-\pi),X(\pi)}]\quad(-\pi\leqslant a\leqslant\pi), \tag{A4}$$

we can finally rewrite Eq. (A3) in the loop invariant form

$$\begin{aligned}
&\frac{\delta^2}{\delta X_\mu(\sigma)\delta X_\mu(\sigma)}\mathrm{Tr}\,U[\mathscr{C}_{X(-\pi),X(\pi)}]\\
&\quad=\mathrm{Tr}((\nabla_\mu F_{\mu\nu})(X(\sigma))X'^{\nu}(\sigma)(\mathrm{Tr}U[\mathscr{C}_{X(-\pi),X(\pi)}]))\\
&\qquad+\mathrm{Tr}(F_{\alpha_\beta}(X(\sigma))X'^{\beta}(\sigma)F^{\alpha\rho}(X(\sigma))X'_\rho(\sigma)(\mathrm{Tr}U[\mathscr{C}_{X(-\pi),X(\pi)}])).
\end{aligned}\tag{A5}$$

In order to write the (unrenormalized) quantum analogous loop equation, we take the quantum ($N_c\to\infty$) average of both sides of Eq. (4a) and observe the quantum results

$$\begin{aligned}
&\langle\mathrm{Tr}(\nabla_\mu F_{\mu\nu})(X(\sigma))X'^{\nu}(\sigma)\mathrm{Tr}U[\mathscr{C}_{X(-\pi),X(\pi)}]\rangle\\
&\quad=\lambda_0^2\oint_{X(-\pi),X(\pi)}X'_\mu(\sigma)\delta^{(D)}(X_\mu(\sigma)-X_\mu(\bar\sigma))X'_\mu(\bar\sigma)\langle\mathrm{Tr}\,U[\mathscr{C}_{X(-\pi),X(\sigma)}]\rangle\langle\mathrm{Tr}\,U[\mathscr{C}_{X(\sigma),X(\pi)}]\rangle
\end{aligned}\tag{A6}$$

and

$$\begin{aligned}
&\langle\mathrm{Tr}(F_{\alpha\beta}(X(\sigma))X'^{\beta}(\sigma)F^{\alpha\beta}((\sigma))X'_\rho(\sigma)U[\mathscr{C}_{X(-\pi),X(\pi)}])\rangle\\
&\quad=\left(\int d^D x\langle\mathrm{Tr}(F_{\alpha\beta}F^{\alpha\beta})(x)\rangle\right)|X'(\sigma)|^2\langle\mathrm{Tr}\,U[\mathscr{C}_{X(-\pi),X(\pi)}]\rangle.
\end{aligned}\tag{A7}$$

Equation (A7) was obtained by supposing the very existence of confining in QCD [SU(N)] for any value of the color parameter N signaled by the (formal) nonvanishing gauge invariant SU(N) gluon condensate in R^D:

$$\int d^D x\langle\mathrm{Tr}(F_{\alpha\beta}F^{\alpha\beta})(x)\rangle = -\gamma^2. \tag{A8}$$

By making the assumption that confining persists at $N_c\to\infty$ we obtain the QCD [SU(∞)] loop wave equation [Eq. (23)] in the proper-time gauge $e(\sigma)=1$.

APPENDIX B: THE β TERM

In this Appendix we present the calculations leading to the β term in Eq. (27).

Therefore let us consider the boundary value of the following quantity:

$$\begin{aligned}
&\lim_{\tau\to0^+}\int_D d\bar\sigma\, d\bar\xi\,\hat T_{\mu\nu}(\phi_\mu(\sigma,\xi))\\
&\quad\times\delta^{(0)}(\phi_\mu(\sigma,\zeta)-\phi_\mu(\bar\sigma,\bar\xi)\hat T_{\mu\nu}(\bar\sigma,\bar\xi)).
\end{aligned}\tag{B1}$$

We can evaluate Eq. (B1) by taking into account the following results.

First, formally

$$\begin{aligned}
&\lim_{\tau\to0^+}\delta^{(D)}(\phi_\mu(\sigma,\zeta)-\phi_\mu(\bar\sigma,\bar\xi))\\
&\quad=\lim_{\tau\to0^+}\delta^{(D)}(\phi_\mu(\sigma,\zeta)-\phi_\mu(\bar\sigma,\bar\xi))\delta(\zeta-\bar\xi)\\
&\quad=\delta^{(D)}(X_\mu(\sigma)-X_\mu(\bar\sigma))\delta(\bar\xi),
\end{aligned}\tag{B2}$$

since our topologically trivial string surface does not possess self-intersections in the intrinsic string time variable ζ, which in turn, is related to the nonexistence of handles and

holes in the string world sheet.

Second, in the asymptotic limit $\zeta \to 0^+$ the string surface has the behavior

$$\lim_{\tau \to 0^+} \phi_\mu(\sigma,\zeta) = \lim_{\tau \to 0^+} X_\mu(\sigma)(1+\zeta),$$

since the string surface is a homotopical (contractible) deformation of its boundary.

As a consequence of the above-mentioned remark, we obtain, in the string isothermal gauge $[X'_\mu(\sigma)\cdot X_\mu(\sigma)=0]$, the value in Eq. (B1):

$$\lim_{\tau \to 0^+} \hat{T}^{\mu\nu}(\phi_\mu(\sigma,\zeta))\hat{T}(\phi_\mu(\bar{\sigma},\bar{\zeta})) = \left[X'_\mu(\sigma)/\sqrt{X'_\mu(\sigma)^2}\right]\left[X'_\mu(\bar{\sigma})/\sqrt{X'_\mu(\bar{\sigma})^2}\right], \quad \text{(B3)}$$

where we have taken into account that $X_\mu(\sigma) = X_\mu(\bar{\sigma})$ in Eq. (B1). By making Eq. (B3) covariant, i.e., $\sqrt{(X^1_\mu(\sigma))^2} \to e(\sigma)$, we obtain the β term in Eq. (27), which for $M=22$ $[e(\sigma)=\text{const}]$, is simply given by

$$\frac{\beta}{2}\int_{-\pi}^{\pi} d\bar{\sigma}\, X'_\mu(\sigma)(\delta^{(D)}(X_\mu(\sigma)-X_\mu(\bar{\sigma})))X'_\mu(\bar{\sigma}). \quad \text{(B4)}$$

APPENDIX C: THE MIGDAL–ELFIN STRING AS A PARTICULAR CASE

Our aim in this Appendix is to show how to obtain the proposed Migdal–Elfin string for QCD $[SU(\infty)]$[12] as a particular case of our proposed self-suppressing fermionic string when the string world sheet does not possess nontrivial self-intersections, i.e., $\phi_\mu(\sigma,\zeta) = \phi_\mu(\bar{\sigma},\bar{\zeta})$ means that $\sigma=\bar{\sigma}$, $\zeta=\bar{\zeta}$.

In order to analyze this case let us introduce orthonormal coordinates on the string surface $\{\phi_\mu(\sigma,\zeta)\}$:

$$\partial_\sigma\phi_\mu\,\partial_\zeta\phi^\mu = 0, \quad (\partial_\sigma\phi^\mu)^2 = (\partial_\zeta\phi_\mu)^2,$$
$$h(\sigma,\zeta) = \det\{h_{ab}(\sigma,\zeta)\} = \det\{\partial^a\phi^\mu\,\partial_b\phi^\mu\} = (\partial_\sigma\phi^\mu)^2 = (\partial_\zeta\phi^\mu)^2. \quad \text{(C1)}$$

Note that this is possible since we have concealed the model's conformal anomaly by choosing $M=22$.

By introducing a tangent vector along coordinates lines $\partial\phi^\mu/\partial\zeta$ and $\partial\phi^\mu/\partial\sigma$, we have the relationship (see the Appendix of Ref. 13)

$$\delta^{(D)}(\phi^\mu(\sigma,\zeta)-\phi^\mu(\bar{\sigma},\bar{\zeta})) = \delta_\epsilon^{(D-2)}(0)([1/h(\sigma,\zeta)^2]\delta^{(2)}((\sigma-\bar{\sigma}),(\zeta-\bar{\zeta}))), \quad \text{(C2)}$$

where $\delta_\epsilon^{(D-2)}(0)$ means a regularized form of the delta function singular value $\delta^{(D-2)}(0)$ (See Ref. 13).

Substituting Eq. (C2) into the string self-interaction term [Eq. (25)] we obtain the more invariant expression for the fermion action:

$$\beta^{(R)}\int_D(\bar{\psi}_{(k)}\psi_{(k)})(\delta,\zeta) \times\left(\sum_{\{\phi_\mu(\sigma,\zeta)=\phi_\mu(\bar{\sigma},\bar{\zeta})\}}\hat{T}^{\mu\nu}(\phi_\mu(\sigma,\zeta))\hat{T}^{\mu\nu}(\phi_\mu(\bar{\sigma},\bar{\zeta}))\right), \quad \text{(C3)}$$

where $\beta^{(R)} = \beta\,\delta_\epsilon^{(D-2)}(0)$ is the regularized string constant.

At this point we can see that Eq. (C3) reduces to a mass term for the intrinsic SO(22) fermion field $\psi_\zeta(\sigma,\zeta)$, which, in the case of the string world sheet has only the trivial self-intersection

$$\phi_\mu(\sigma,\zeta) = \phi_\mu(\bar{\sigma},\bar{\zeta}) \Rightarrow \sigma=\bar{\sigma},\ \zeta=\bar{\zeta},$$

since

$$\hat{T}^{\mu\nu}(\phi_\mu(\sigma,\zeta))\hat{T}_{\mu\nu}(\phi_\mu(\sigma,\zeta)) = 1.$$

We thus get

$$\sum_{k=1}^{22}\beta^{(R)}\int_D(\bar{\psi}_{(k)}\psi_{(k)})(\sigma,\zeta)d\sigma\,d\zeta. \quad \text{(C4)}$$

For the nontrivial self-intersecting case [σ multivalued $\phi^\mu(\sigma,\zeta)$ functions] we have to add to Eq. (C4) the term responsible for the theory's interaction, which is supported at the nontrivial string's surface self-intersection lines $\phi_\mu(\sigma,\zeta) = \phi_\mu(\bar{\sigma}\bar{\zeta})$ with $\sigma \neq \bar{\sigma}$ as given by our interaction action [Eq. (25C)] and previously conjectured in Ref. 14.

[1] R. P. Feynman, Phys. Rev. **80**, 440 (1950).
[2] A. M. Polyakov, Phys. Lett. B **103**, 207 (1981).
[3] A. M. Polyakov, Nucl. Phys. B **164**, 175 (1979); A. M. Migdal, Phys. Rep. **102**, 199 (1983); L. C. L. Botelho, Phys. Lett. B **169**, 428 (1986).
[4] L. Brink, P. Di Vecchia, and P. Howe, Phys. Lett. B **65**, 471 (1976).
[5] K. Fujikawa, Phys. Rev. D **25**, 2584 (1982).
[6] J. B. Hartle and S. W. Hawking, Phys. Rev. **28**, 2966 (1983).
[7] C. Rebbi, Phys. Rep. C **12** (1974); J. Scherk, Rev. Mod. Phys. **47**, 123 (1975).
[8] CC. Marshall and P. Ramond, Nucl. Phys. B **85**, 375 (1975).
[9] L. C. L. Botelho and J. C. Mello, J. Phys. A **20**, 2217 (1987).
[10] S. Deser and B. Zumino, Phys. Lett. B **65** (1976); P. S. Howe, J. Phys. A **12**, 393 (1979).
[11] L. C. L. Botelho, Phys. Lett. B **152**, 358 (1985).
[12] A. A. Migdal, Nucl. Phys. B **189**, 253 (1981).
[13] P. Olesen and J. L. Petersen, Nucl. Phys. B **181**, 157 (1981).
[14] L. C. L. Botelho, Rev. Bras. Fis. **16**, 279 (1986); Caltech preprint 68-1444, 1987.

Field theory of geometric p-branes

Choon-Lin Ho[a)]
School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455

(Received 29 December 1988; accepted for publication 3 May 1989)

In this paper, p-brane field theory is constructed in terms of functions of geometric p-surfaces. Nambu–Goto p-brane dynamics is incorporated in the Dirac form. The field equation for toroidal p-branes in $p+2$ dimensions is exactly solved for, and is shown to admit an equally spaced mass-squared spectrum containing massless states.

I. INTRODUCTION

The successes of string theory[1] have inspired interest in the theory of higher dimensional objects, generally known as p-branes. Considerable progress in understanding p-brane theory has been made in the past year or so.[2] Particularly, super-p-brane theories have been constructed.[3] It is also realized that these super-p-brane theories are closely related to the four classical superstring theories.[4] More recently, it was realized that the area-preserving membrane algebra contains the Virasoro algebra as a subalgebra.[5] It is therefore hoped that a better understanding of p-brane theory could provide new insights into string theory. As a first step in this direction, various schemes of quantizing membrane theory have been proposed.[6] Despite all these attempts, two important issues have not yet been settled, namely, the existence of massless particles, and the consistency of quantum p-brane theory.[4,7]

In a previous paper,[8] we proposed a quantum theory of geometric membranes which generalizes the line functional approach to string theory proposed by Carson and Hosotani,[9] which is a reparametrization-invariant formulation of string field theory.[10] There we formulate our membrane field theory as a theory of surface functionals (functions of geometric surfaces) which reproduces classical Nambu–Goto membrane dynamics in a certain limit. The connection of our field theory with classical membrane dynamics is done in accordance with Dirac's treatment of spin-$\frac{1}{2}$ particles. Our membrane fields therefore transform nontrivially under Lorentz transformations. As such, the basic entities of our theory are multicomponent surface functionals. This is the main difference between our approach and the standard ones, such as light-cone formulation. Reparametrization invariance is kept manifest in our approach. We solve the field equation exactly in $1+3$ dimensions for toroidal membranes. We find that the solution contains massless states and that it admits equally spaced mass-squared spectrum which is characteristic of free string theories.

Similar results were also obtained in the line functional theory of string in $1+2$ dimensions.[9] One immediately realizes that there is something in common for a string in two-space and a membrane in three-space; they are both hypersurfaces in their respective embedding spaces. A question naturally arises: Do the same results generalize to p-branes which are hypersurfaces in $p+2$ space dimensions? The purpose of this paper is to show that our theory for toroidal p-branes in $p+2$ dimensions does admit an equally spaced mass-squared spectrum containing massless states.

This paper is organized as follows. In Sec. II we give a brief review of surface functional theory of p-branes, which was presented in Ref. 8. We then discuss in Sec. III our theory in $p+2$ dimensions and in $X_0=\tau$ gauge. Connection of our theory with various differential-geometric quantities of hypersurfaces is discussed in great detail. In Sec. IV, the field equation for toroidal p-branes is solved exactly, yielding an equally spaced mass-squared spectrum with massless states. Section V summarizes the paper. In Appendix A, we outline calculations of principal and total curvatures of an n-torus T^n. Appendix B gives the explicit forms of function U, defined in Sec. IV, that are required to solve the field equation.

II. COVARIANT FIELD EQUATIONS OF CLOSED p-BRANES

In this section, we shall review briefly the derivation of covariant equations for multicomponent fields Ψ of closed p-branes in d-dimensional spacetime, as presented in Ref. 8.

We start with the classical Nambu–Goto action for p-branes given by

$$S=-\frac{1}{\kappa}\int d^{p+1}\xi\sqrt{(-1)^p\det\hat{h}_{\alpha\beta}}\,, \tag{2.1}$$

$$\hat{h}_{\alpha\beta}(\xi)=\partial_\alpha X^\mu\partial_\beta X_\mu, \tag{2.2}$$

where $X^\mu(\xi)$ $(\mu=0,...,d-1)$ and $\xi^\alpha=(\tau,\sigma_1,...,\sigma_p)$ $(\alpha=0,...,p)$ are space-time and world volume coordinates, respectively. Canonical conjugate momenta $p^\mu(\sigma)$, where σ denotes collectively the set $\{\sigma_1,...,\sigma_p\}$, are defined to be $\delta S/\delta\,\partial_\tau X_\mu(\sigma)$. It can be shown that the canonical Hamiltonian vanishes.

There are $p+1$ sets of primary constraints,

$$\begin{aligned}&\chi_0(\sigma)\equiv\tfrac{1}{2}\{p^2-[(-1)^p/\kappa^2]\hat{h}\}=0,\\&\chi_k(\sigma)\equiv p\cdot\partial_k X=0\quad(k=1,...,p),\end{aligned} \tag{2.3}$$

where $\hat{h}=\det\hat{h}_{jk}$ $(j,k=1,...,p)$ is the cofactor of $\hat{h}_{00}$. The Poisson-bracket algebra of these constraints is given by

$$\begin{aligned}&\{\chi_i(\sigma),\chi_j(\sigma')\}\\&\quad=\chi_j(\sigma)\frac{\partial}{\partial\sigma_i}\delta(\sigma-\sigma')\\&\qquad+\chi_i(\sigma')\frac{\partial}{\partial\sigma_j}\delta(\sigma-\sigma'),\\&\{\chi_0(\sigma),\chi_j(\sigma')\}\\&\quad=[\chi^0(\sigma)+\chi^0(\sigma')]\frac{\partial}{\partial\sigma_j}\delta(\sigma-\sigma'),\end{aligned} \tag{2.4a}$$

[a)] Address after September 1, 1989: Institute of Physics, Academia Sinica, Taipei, Taiwan 11529, Republic of China.

 0022-2488/89/092168-09$02.50

$$\{\chi_0(\sigma),\chi_0(\sigma')\}$$
$$= \frac{(-1)^{p+1}}{\kappa^2}\sum_{i,j}\left[\frac{\partial\hat{h}(\sigma)}{\partial h_{ij}(\sigma)}\chi_i(\sigma) + \frac{\partial\hat{h}(\sigma')}{\partial h_{ij}(\sigma')}\chi_i(\sigma')\right]\frac{\partial}{\partial\sigma_j}\delta(\sigma-\sigma'),$$

where

$$\delta(\sigma-\sigma')\equiv\prod_{i=1}^{p}\delta(\sigma_i-\sigma_i').$$

In terms of the integral transforms of χ's,

$$\chi_\alpha[f]\equiv\int d^p\sigma f(\sigma)\chi_\alpha(\sigma),$$

the algebra reads

$$\{\chi_i[f],\chi_j[g]\}=\chi_j[f\partial_i g]-\chi_i[g\partial_j f],$$
$$\{\chi_0[f],\chi_j[g]\}=\chi_0[f\partial_j g-g\partial_j f], \tag{2.4b}$$

and

$$\{\chi_0[f],\chi_0[g]\} = \frac{(-1)^{p+1}}{\kappa^2}\sum_{i,j}\chi_i\left[(f\partial_j g-g\,\partial_i f)\frac{\partial\hat{h}}{\partial h_{ij}}\right].$$

For $p\geqslant 2$, the algebra closes only weakly. This is due to the fact that the right-hand side of the Poisson bracket of $\chi_0(\sigma)$ and $\chi_0(\sigma')$ involves X^μ-dependent coefficients. This makes the equations essentially nonlinear. Furthermore, in a quantum theory the commutator algebra of the constraints would give rise to operator anomalies.

We avoid these difficulties by proposing a field theory of geometric p-branes. The fundamental objects in the theory are p-dimensional surfaces, or p-surfaces in short. Fields are functions of p-surfaces, which are invariant under σ-reparametrization (σ-RP). The theory is required to reproduce the Nambu–Goto dynamics in the $X_0=\tau$ gauge. In this gauge, dynamical variables are spatial coordinates $\mathbf{X}(\sigma)$, and their conjugate momenta $\mathbf{p}(\sigma)=\delta S/\delta\dot{\mathbf{X}}(\sigma)$. Now there are only p sets of constraints, $\chi_k(\sigma)\equiv-\mathbf{p}\cdot\partial_k\mathbf{X}=0$, and the canonical Hamiltonian is nontrivial,

$$H_0=\int d^p\sigma\sqrt{(-1)^p\hat{h}}\,\sqrt{(-1)^p(\mathbf{p}^2/\hat{h})+1/\kappa^2}. \tag{2.5}$$

It can be checked that the Poisson bracket of $\chi_k(\sigma)$ and H_0 vanishes. The Schrödinger equation and constraints for the field $\Psi_{X_0=\tau}=\Psi_{X_0=\tau}[\tau;\mathbf{X}(\sigma)]$ are

$$i\frac{\partial}{\partial\tau}\Psi_{X_0=\tau}=H_0\Psi_{X_0=\tau}, \tag{2.6}$$

$$\partial_k\mathbf{X}\cdot\mathbf{p}\,\Psi_{X_0=\tau}=0\quad(k=1,\ldots,p), \tag{2.7}$$

with $\mathbf{p}=-i[\delta/\delta\mathbf{X}(\sigma)]$. According to Eq. (2.7), $\Psi_{X_0=\tau}$ is σ-RP invariant, and is therefore a function of spatial p-surfaces. The second square root in Eq. (2.5) is replaced, according to Dirac's treatment of relativistic point particles, by

$$\mathcal{M}(\sigma)=\boldsymbol{\alpha}(\sigma)\cdot\mathbf{p}/\sqrt{(-1)^p\hat{h}}+\beta(\sigma)(1/\kappa), \tag{2.8}$$

with $\mathcal{M}^2(\sigma)=(-1)^p(\mathbf{p}^2/\hat{h})+1/\kappa^2$. Note that $\boldsymbol{\alpha}(\sigma)$ and $\beta(\sigma)$ are matrices to be determined. This then leads to the covariant equation of motion for multicomponent fields Ψ of space-time p-surfaces:

$$\int d^p\sigma\left[\Gamma^\mu(\sigma)P_\mu(\sigma)-\frac{\sqrt{(-1)^p\hat{h}}}{\kappa^2}\Lambda(\sigma)\right]\times\Psi[X^\mu(\sigma)]=0, \tag{2.9}$$

$$(\partial_k X^\mu)P_\mu\Psi=0\quad(k=1,\ldots,p). \tag{2.10}$$

Here $P_\mu=i[\delta/\delta X^\mu(\sigma)]$, and $\Gamma^\mu(\sigma)$ and $\Lambda(\sigma)$ are generalized Dirac matrices.

In order to determine the forms of $\Gamma^\mu(\sigma)$ and $\Lambda(\sigma)$, we impose the following requirements[11]:

(i) Equation (2.9) must be Lorentz covariant, translation invariant, and manifestly σ-RP invariant;

(ii) In the $X_0=\tau$ subspace, Eq. (2.9) reduces to Eqs. (2.6)–(2.8) with $\boldsymbol{\alpha}(\sigma)$ and $\beta(\sigma)$ local in σ;

(iii) $\Gamma^\mu(\sigma)$ and $\Lambda(\sigma)$ depend on σ only through $X^\mu(\sigma)$;

(iv) $\Gamma^\mu(\sigma)$ and $\Lambda(\sigma)$ commute with $[1/\sqrt{(-1)^p\hat{h}}\,]P_\mu$.

Note that in transition to the $X_0=\tau$ gauge, we make the following replacements:

$$\Psi[X^\mu(\sigma)]\to\Psi_{X_0=\tau}[\tau;\mathbf{X}(\sigma)],$$
$$\frac{i}{\sqrt{(-1)^p\hat{h}}}\frac{\delta}{\delta X^0(\sigma)}\to\frac{i}{V_p}\frac{\partial}{\partial\tau}, \tag{2.11}$$

where

$$V_p=\int d^p\sigma\sqrt{(-1)^p\hat{h}}.$$

It turns out to be convenient to define a totally antisymmetric tangent tensor by

$$t_{\mu_1\cdots\mu_p}=\frac{1}{\sqrt{(-1)^p\hat{h}}}\frac{\partial(X_{\mu_1},\ldots,X_{\mu_p})}{\partial(\sigma_1,\ldots,\sigma_p)}, \tag{2.12}$$

with $t^2=(-1)^p p!$. Then the conditions (i), (iii), and (iv) imply that $\Gamma^\mu(\sigma)$ and $\Lambda(\sigma)$ depend on σ only through $t_{\mu_1\cdots\mu_p}$ [to arrive at this, one needs to make use of the fact that $\delta/\delta X^\mu(\sigma)$ and $(\partial/\partial\sigma_k)X_\nu(\sigma)$ commute]. Derivatives of $t_{\mu_1\cdots\mu_p}$ are excluded by the condition (iv). One can therefore expand $\Gamma^\mu(\sigma)$ and $\Lambda(\sigma)$ in a Taylor series in $t_{\mu_1\cdots\mu_p}(\sigma)$ with constant matrix coefficients. To fulfill condition (ii), terms involving products of $t_{\mu_1\cdots\mu_p}$ in the expansion of $\Gamma^\mu(\sigma)$ must be discarded, since they will lead to $\boldsymbol{\alpha}(\sigma)$ and $\beta(\sigma)$ matrices which are nonlocal in σ in general. The condition (ii) places no restriction on the form of the $\Lambda(\sigma)$ matrix: $\Lambda(\sigma)$ can contain terms of an arbitrary power of $t_{\mu_1\cdots\mu_p}$. For simplicity, we set $\Lambda(\sigma)=1$. We thus arrive at the following forms:

$$\Gamma^\mu(\sigma)=a_1\Gamma^\mu+(a_2/p!)\Gamma_A^{\mu\nu_1\cdots\nu_p}t_{\nu_1\cdots\nu_p}+(a_3/p!)\Gamma_M^{\mu\nu_1\cdots\nu_p}t_{\nu_1\cdots\nu_p},$$
$$\Lambda(\sigma)=1. \tag{2.13}$$

Here a_1, a_2, and a_3 are arbitrary constants. Γ^μ are the usual Dirac matrices. $\Gamma_A^{\mu_1\cdots\mu_{p+1}}$ and $\Gamma_M^{\mu_1\cdots\mu_{p+1}}$ have the symmetry represented by the following Young tableaux:

$\Gamma_A^{\mu_1\cdots\mu_{p+1}}$: $p+1$ boxes

$\Gamma_M^{\mu_1\cdots\mu_{p+1}}$: p boxes

Algebra satisfied by the Γ_M-piece is generally very complicated, and an equation which involves the Γ_M-piece is also difficult to solve in general. Hence we shall consider only the Γ^μ and the totally antisymmetric $\Gamma_A^{\mu_1 \cdots \mu_{p+1}}$ piece in this paper, and we shall drop the subscript A from now on.

We can now write Eq. (2.9) as

$$\left[a_1 \Gamma^\mu P_\mu + (a_2/p!)\Gamma^{\mu_1 \cdots \mu_{p+1}} P_{\mu_1 \cdots \mu_{p+1}} - V_p/\kappa\right]\Psi = 0, \tag{2.14}$$

where

$$P_\mu = \int d^p\sigma\, p_\mu$$

and

$$P_{\mu_1 \cdots \mu_{p+1}} = \frac{1}{p+1}\int d^p\sigma\left[\sum_\rho \text{sign}(\rho) t_{\mu_1 \cdots \mu_p} p_{\mu_{p+1}}\right].$$

Here ρ represents the cyclic permutation of the indices $\mu_1,\ldots,\mu_{p+1}$. The two constants a_1 and a_2 satisfy

$$a_1^2 + (-1)^p a_2^2 = 1. \tag{2.15}$$

The Γ's matrices obey

$$\{\Gamma^\mu,\Gamma^\nu\} = 2\eta^{\mu\nu}, \tag{2.16}$$

$$\{\Gamma^{\mu_1 \cdots \mu_{p+1}},\Gamma^{\nu_1 \cdots \nu_{p+1}}\} = 2\eta^{\mu_1 \cdots \mu_{p+1},\nu_1 \cdots \nu_{p+1}}, \tag{2.17}$$

where the generalized η-tensor is defined by

$$\eta^{\mu_1 \cdots \mu_{p+1},\nu_1 \cdots \nu_{p+1}} = \begin{vmatrix} \eta^{\mu_1\nu_1} & \eta^{\mu_1\nu_2} & \cdots & \eta^{\mu_1\nu_{p+1}} \\ \vdots & \vdots & \vdots & \vdots \\ \eta^{\mu_{p+1}\nu_1} & \eta^{\mu_{p+1}\nu_2} & \cdots & \eta^{\mu_{p+1}\nu_{p+1}} \end{vmatrix}. \tag{2.18}$$

The algebras (2.16) and (2.17) are Clifford algebras of dimensions d and $d(d-1)\cdots(d-p)/(p+1)!$, respectively.

Finally, we note that the field functions Ψ transforms nontrivially under Lorentz transformations, as a result of the generalized Dirac algebras (2.16) and (2.17). It can be checked that the generators of d-dimensional Lorentz transformations are given by

$$J^{\mu\nu} = \int d^p\sigma(X^\mu p^\nu - X^\nu p^\mu) + \frac{i}{4}[\Gamma^\mu,\Gamma^\nu] + \frac{i}{4p!}\left[\Gamma^{\mu\lambda_1 \cdots \lambda_p},\Gamma^\nu{}_{\lambda_1 \cdots \lambda_p}\right]. \tag{2.19}$$

III. p-BRANE FIELD EQUATIONS IN $d=p+2$ AND $X_0=\tau$ GAUGE

The covariant field equation (2.14) derived in Sec. II is in general difficult to solve. However, it can be greatly simplified if one considers closed p-brane theory in $(p+2)$ dimensions, and in the $X_0=\tau$ gauge. In this case, p-surfaces become hypersurfaces in E^{p+1}, and only one component of the operator $P_{\mu_1 \cdots \mu_{p+1}}$ is relevant. This operator is closely related to some differential-geometric quantities of p-dimensional hypersurfaces, such as total mean curvatures.

In this section, we shall first write down the field equation in $(p+2)$ dimensions and in the $X_0=\tau$ gauge. We then relate the relevant component of $P_{\mu_1 \cdots \mu_{p+1}}$ which appears in the field equations to various geometric quantities of p-surfaces. Exact solutions of the field equation will be given in the next section.

A. Field equation

In $d=p+2$ dimensions, one can replace the matrices $\Gamma^{\mu_1 \cdots \mu_{p+1}}$ by the dual matrices defined as

$$\Lambda^\mu \equiv \frac{1}{(p+1)!}\epsilon^{\mu\mu_1 \cdots \mu_{p+1}}\Gamma_{\mu_1 \cdots \mu_{p+1}}. \tag{3.1}$$

Using Eqs. (2.17) and (2.18), one obtains the algebra satisfied by Λ^μ,

$$\{\Lambda^\mu,\Lambda^\nu\} = (-1)^{p+1}2\eta^{\mu\nu}, \tag{3.2}$$

which is, up to a sign, just the usual Dirac algebra. Replacing the $\Gamma^{\mu_1 \cdots \mu_{p+1}}$ matrices in Eq. (2.14) by their duals, we arrive at the following simpler equation:

$$\left[a_1 \Gamma^\mu P_\mu + a_2 \Lambda^\mu F_\mu - V_p/\kappa\right]\Psi = 0, \tag{3.3}$$

where

$$F^\mu = \frac{(-1)^{p+1}}{p!}\epsilon^{\mu\mu_1 \cdots \mu_{p+1}}\int d^p\sigma t_{\mu_1 \cdots \mu_p} p_{\mu_{p+1}}. \tag{3.4}$$

It follows from Eq. (2.19) that the fields Ψ describe bosonic p-branes.

Let us now restrict our theory to the $X_0=\tau$ subspace. In this subspace, the only nonvanishing component of the tangent tensor defined in Eq. (2.12) is

$$t_{i_1 \cdots i_p} = \frac{1}{\sqrt{h}}\frac{\partial(X_{i_1},\ldots,X_{i_p})}{\partial(\sigma_1,\ldots,\sigma_p)} \quad (i_k \neq 0,\ k=1,\ldots,p), \tag{3.5}$$

where $h = \det h_{ij} = \det(\partial_i \mathbf{X}\cdot\partial_j \mathbf{X})$. It then follows from Eq. (3.4) that only the zero component of the operator F^μ survives in the field equation. Equation (3.3) now reads

$$\left[a_1 \Gamma^\mu P_\mu + i a_2 \Lambda^0 Q - V_p/\kappa\right]\Psi = 0, \tag{3.6}$$

where

$$Q \equiv -iF^0 = \int d^p\sigma\, \mathbf{N}(\sigma)\cdot\frac{\delta}{\delta \mathbf{X}(\sigma)}. \tag{3.7}$$

$\mathbf{N}(\sigma)$ is the unit normal vector at each point of the p-surface in E^{p+1}, and is related to the tangent tensor by

$$N_j = (1/p!)\epsilon_{i_1 \cdots i_p j} t_{i_1 \cdots i_p} \quad (j,i_k = 1,2,\ldots,p+1). \tag{3.8}$$

Equations (3.6)–(3.8) give the field equation in $d=p+2$ dimensions and in the $X_0=\tau$ gauge. The operator Q is in fact the generator of a small deformation of p-surface along the normal direction at each of its points. It is therefore not surprising that Q is related to the differential-geometric properties of p-dimensional hypersurfaces. This will be discussed in the next subsection.

B. Differential-geometric quantities

In this subsection, we shall derive useful identities relating the operator Q with various differential-geometric quantities of p-surfaces.

As mentioned before, in the $X_0=\tau$ gauge and in $(p+2)$ dimensions, a closed p-brane, specified by points $\mathbf{X}(\sigma)$, becomes a closed hypersurface in E^{p+1}. According to the theory of surfaces, the geometry of hypersurfaces de-

pends only on two quadratic differential forms. These are called the first and second fundamental forms, denoted, respectively, by I and II. They are defined by

$$\mathrm{I} \equiv d\mathbf{X} \cdot d\mathbf{X} = h_{ij}\, d\sigma^i d\sigma^j \quad (i,j = 1,...,p) \tag{3.9}$$

and

$$\mathrm{II} \equiv -d\mathbf{N} \cdot d\mathbf{X} = b_{ij}\, d\sigma^i\, d\sigma^j \quad (i,j = 1,...,p). \tag{3.10}$$

The coefficients h_{ij}'s and b_{ij}'s are given by

$$h_{ij} = \partial_i \mathbf{X} \cdot \partial_j \mathbf{X} \tag{3.11}$$

and

$$b_{ij} = -\partial_i \mathbf{N} \cdot \partial_j \mathbf{X} = \mathbf{N} \cdot \partial_i \partial_j \mathbf{X} = b_{ji}. \tag{3.12}$$

The second equality in Eq. (3.12) follows from the relation

$$\mathbf{N} \cdot \partial_j \mathbf{X} = 0 \quad (j = 1,...,p), \tag{3.13}$$

which can be easily proved from Eqs. (3.5) and (3.8). Also, since $\mathbf{N}^2 = 1$, we have

$$\mathbf{N} \cdot \partial_i \mathbf{N} = 0 \quad (i = 1,...,p). \tag{3.14}$$

Combining Eqs. (3.13) and (3.14), we obtain

$$\partial_i \mathbf{N} = -B_i^{\ j} \partial_j \mathbf{X}, \tag{3.15}$$

where $B_i^{\ j}$ is a $p \times p$ matrix.

Equation (3.15) is called the Weingarten formula. Eigenvalues of the matrix B give the principal curvatures, denoted by $\kappa_1,...,\kappa_p$. We define the rth mean curvature H_r as

$$H_r \equiv \frac{1}{\binom{p}{r}} a_r \quad (r = 1,...,p), \tag{3.16}$$

where a_r is the rth elementary symmetric polynomial in κ's,

$$a_r \equiv \sum_{i_1 < i_2 \cdots < i_r} \kappa_{i_1} \kappa_{i_2} \cdots \kappa_{i_r}. \tag{3.17}$$

Gauss curvature is given by $H_p = \det B$. For later convenience we define $H_0 \equiv 1$.

We shall now derive some useful identities which relate the operator Q and the differential-geometric quantities defined so far. First, we have

$$Q\mathbf{X}(\sigma) = \mathbf{N}(\sigma), \tag{3.18}$$

$$Q\partial_j \mathbf{X}(\sigma) = \partial_j \mathbf{N}(\sigma). \tag{3.19}$$

From Eqs. (3.11) and (3.19), we obtain

$$Qh_{ij} = -2b_{ij}. \tag{3.20}$$

Using Eqs. (3.11), (3.12), and (3.15), we can show that

$$b_{ij} = B_i^{\ l} h_{lj} \tag{3.21}$$

and

$$B_i^{\ j} = b_{il} h^{lj}, \tag{3.22}$$

where h^{ij} is the inverse of h_{ij}.

With the use of Eq. (3.21), we have

$$Qh = -2hh^{ik} b_{ik} = -2hB_i^{\ i} = -2h \operatorname{Tr} B \tag{3.23}$$

and

$$Q\sqrt{h} = -\sqrt{h} \operatorname{Tr} B. \tag{3.24}$$

Also, one can easily check that

$$Q\mathbf{N} = 0, \tag{3.25}$$

$$Qb_{jk} = -\partial_j \mathbf{N} \cdot \partial_k \mathbf{N} = -b_{jk} h^{kl} b_{lk}, \tag{3.26}$$

$$Qh^{jk} = 2h^{jl} b_{lm} h^{mk}, \tag{3.27}$$

$$QB_i^{\ j} = B_i^{\ k} B_k^{\ j}. \tag{3.28}$$

Repeated use of Eq. (3.28) gives the important identity

$$QS_n = nS_{n+1}, \tag{3.29}$$

where

$$S_n \equiv \operatorname{Tr} B^n = \sum_{i=1}^{p} \kappa_i^n. \tag{3.30}$$

We now use Eqs. (3.24) and (3.29) to prove an important identity which relates H_r to H_{r+1}:

$$Q(\sqrt{h} H_r) = -(p-r)(\sqrt{h} H_{r+1}) \quad (r = 0,1,...,p), \tag{3.31a}$$

or equivalently,

$$Q(\sqrt{h} a_r) = -(r+1)(\sqrt{h} a_{r+1}) \quad (a_0 \equiv 1,\ a_{p+1} \equiv 0), \tag{3.31b}$$

in view of Eq. (3.16). This identity reduces to Eq. (3.24) when $r = 0$. The proof of Eq. (3.31b) is very simple by induction using Newton's formula. The Newton formula relates the a_r's and the S_r's as follows:

$$\begin{aligned} &S_1 - a_1 = 0, \quad S_2 - S_1 a_1 + 2a_2 = 0, \\ &S_3 - S_2 a_1 + S_1 a_2 - 3a_3 = 0, \\ &S_r - S_{r-1} a_1 + S_{r-2} a_2 - \cdots + (-1)^{r-1} S_1 a_{r-1} \\ &\quad + (-1)^r r a_r = 0 \quad (r = 1,...,p). \end{aligned} \tag{3.32}$$

Let us assume that Eq. (3.31b) is true for r equals 1,..., up to $(r-1)$. Multiplying Eq. (3.32) by $\sqrt{h}$ and operating on the resulted expression by Q, we obtain, after some cancellations of terms,

$$\begin{aligned} &(-1)^{r+1} Q(\sqrt{h} a_r) \\ &\quad = \sqrt{h}\,[S_{r+1} - S_r a_1 + S_{r-1} a_2 \cdots + (-1)^r S_1 a_r] \\ &\quad = \begin{cases} \sqrt{h}\,[(-1)^{r+2}(r+1) a_{r+1}], & r \neq p, \\ 0, & r = p. \end{cases} \end{aligned} \tag{3.33}$$

In obtaining the second equality of Eq. (3.33), we have used Eq. (3.32) for the case $r \neq p$ and the following identity for the case $r = p$:

$$S_{p+1} - S_p a_1 + \cdots + (-1)^p S_1 a_p = 0. \tag{3.34}$$

Equation (3.31b) is easily proved to be true for $r = 1$. Hence by induction it is true for $r = 1,...,p-1$, and also for $r = 0$ and $r = p$, if we define $a_0 \equiv 1$, and $a_{p+1} \equiv 0$.

In order to solve the field equation (3.6), one still needs to express the operators $\mathbf{P}$ and Q in terms of some convenient coordinates. Let us define the following σ-RP invariant coordinates,

$$h_l = \int d^p\sigma \sqrt{h} H_l \quad (l = 0,1,..,p),$$

$$\mathbf{y} = \frac{1}{h_0} \int d^p\sigma \sqrt{h}\, \mathbf{X},$$

$$\boldsymbol{\lambda}_r = \binom{p}{r} \int d^p\sigma \sqrt{h}\, (\mathbf{X} - \mathbf{y}) H_r,$$

and

$$\zeta_r = \binom{p}{r} \int d^p\sigma \sqrt{h}\, \mathbf{N} H_r \quad (r = 1,...,p). \tag{3.35}$$

Note that $h_0 = V_p$, h_p is the total Gauss curvature and $\mathbf{y}$ is the center-of-mass coordinates of the p-surface.

It follows from Eq. (3.31) that

$$Q h_l = -(p-l) h_{l+1} \quad (l = 0,...,p). \tag{3.36}$$

Using the identities derived in the last subsection, it is easy to show that

$$Q\mathbf{y} = (1/h_0)\boldsymbol{\lambda}_1, \tag{3.37}$$

$$Q\boldsymbol{\lambda}_r = -(r+1)\boldsymbol{\lambda}_{r+1} + \boldsymbol{\zeta}_r + \binom{p}{r}\frac{\boldsymbol{\lambda}_1}{h_0} h_r \quad (r = 1,...,p-1),$$

$$Q\boldsymbol{\lambda}_p = \boldsymbol{\zeta}_p + \frac{\boldsymbol{\lambda}_1}{h_0} h_p, \tag{3.38}$$

and

$$Q\boldsymbol{\zeta}_r = -(r+1)\boldsymbol{\zeta}_{r+1} \quad (r = 1,...,p-1),$$

$$Q\boldsymbol{\zeta}_p = 0. \tag{3.39}$$

Finally, in terms of the coordinates $\{\tau, h_l, \mathbf{y}, \boldsymbol{\lambda}_r, \boldsymbol{\zeta}_r\}$, we have

$$P^\mu = \left(i\frac{\partial}{\partial\tau}, -i\frac{\partial}{\partial \mathbf{y}}\right) \tag{3.40}$$

and

$$\begin{aligned} Q = &-\left[p h_1 \frac{\partial}{\partial h_0} + (p-1) h_2 \frac{\partial}{\partial h_1} + \cdots + h_p \frac{\partial}{\partial h_{p-1}}\right] + \frac{\boldsymbol{\lambda}_1}{h_0}\cdot\frac{\partial}{\partial \mathbf{y}} + \left[-2\boldsymbol{\lambda}_2 + \boldsymbol{\zeta}_1 + p\frac{\boldsymbol{\lambda}_1}{h_0} h_1\right]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_1} \\ &+ \left[-3\boldsymbol{\lambda}_3 + \boldsymbol{\zeta}_2 + \binom{p}{2}\frac{\boldsymbol{\lambda}_1}{h_0} h_2\right]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_2} - 2\boldsymbol{\zeta}_2\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_1} + \cdots \\ &+ \left[-(r+1)\boldsymbol{\lambda}_{r+1} + \boldsymbol{\zeta}_r + \binom{p}{r}\frac{\boldsymbol{\lambda}_1}{h_0} h_r\right]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_r} - r\boldsymbol{\zeta}_r\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_{r-1}} + \cdots \\ &+ \left[\boldsymbol{\zeta}_p + \frac{\boldsymbol{\lambda}_1}{h_0} h_p\right]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_p} - p\boldsymbol{\zeta}_p\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_{p-1}}. \end{aligned} \tag{3.41}$$

We note that the total Gauss curvature h_p and the coordinate $\boldsymbol{\zeta}_p$ are constants of motion, as can be easily checked.

IV. EXACT SOLUTIONS

We proceed to solve the field equation (3.6) in this section. In general, the field equation (3.6) with Eqs. (3.40) and (3.41) is still difficult to solve. It can, however, be exactly solved for in a special case for which $h_2 = h_3 = \cdots = h_p = 0$, and $h_1 \neq 0$. It follows from Eq. (3.41) that these values of h's are constants of motion. We show in Appendix A that the p-dimensional torus T^p possesses this set of values of total mean curvatures.

We now look for positive energy solutions which have the following form:

$$\Psi = e^{-ip\cdot y} U(\mathbf{p}, \mathbf{y}, h_1, \boldsymbol{\lambda}_r, \boldsymbol{\zeta}_r)\Phi(h_0, h_1), \tag{4.1}$$

with $y_0 \equiv \tau$. We require that U satisfies

$$Q(e^{-ip\cdot y} U) = 0. \tag{4.2}$$

If this is so, then one can factor out the factor $(e^{-ip\cdot y}U)$ from the field equation. It is not too difficult to find the required forms of U that solve Eq. (4.2). Remember that in the case we are interested in, the operator Q has the following simple form:

$$\begin{aligned} Q = &-p h_1 \frac{\partial}{\partial h_0} + \frac{\boldsymbol{\lambda}_1}{h_0}\cdot\frac{\partial}{\partial \mathbf{y}} + \left[-2\boldsymbol{\lambda}_2 + \boldsymbol{\zeta}_1 + p\frac{\boldsymbol{\lambda}_1}{h_0} h_1\right]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_1} + [-3\boldsymbol{\lambda}_3 + \boldsymbol{\zeta}_2]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_1} - 2\boldsymbol{\zeta}_2\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_1} + \cdots \\ &+ [-(r+1)\boldsymbol{\lambda}_{r+1} + \boldsymbol{\zeta}_r]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_r} - r\boldsymbol{\zeta}_r\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_{r-1}} + \cdots + \boldsymbol{\zeta}_p\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_p} - p\boldsymbol{\zeta}_p\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_{p-1}}. \end{aligned} \tag{4.3}$$

Now we observe the following identities:

$$\begin{aligned} &\left\{[-(r+1)\boldsymbol{\lambda}_{r+1} + \boldsymbol{\zeta}_r]\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_r} - r\boldsymbol{\zeta}_r\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_{r-1}}\right\}\mathbf{L}_{r-1} = -r\mathbf{L}_r, \\ &\left[\boldsymbol{\zeta}_p\cdot\frac{\partial}{\partial \boldsymbol{\lambda}_p} - p\boldsymbol{\zeta}_p\cdot\frac{\partial}{\partial \boldsymbol{\zeta}_{p-1}}\right]\mathbf{L}_{p-1} = -p^2\boldsymbol{\zeta}_p, \end{aligned} \tag{4.4}$$

where

$$\mathbf{L}_r \equiv -(r+1)\lambda_{r+1} + r\,\xi_r \quad (r = 1,\ldots,p-1). \tag{4.5}$$

These observations enable us to solve Eq. (4.2) for U in terms of $\mathbf{p}$, $\mathbf{y}$, h_1, λ_1, ξ_p, and $\mathbf{L}_r$. We present the explicit forms of U in Appendix B.

With Eqs. (4.1), (4.2), (4.3), and (3.40), the field equation (3.6) reads

$$\left[a_1\Gamma^\mu p_\mu - ipa_2\Lambda^0 h_1 \frac{\partial}{\partial h_0} - \frac{h_0}{\kappa}\right]\Phi(h_0,h_1) = 0, \tag{4.6}$$

where p_μ is the energy-momentum vector. To solve Eq. (4.6), we need explicit representations of the Γ^μ and Λ^μ matrices. It is convenient to make the following choices according to d, the dimensionality of space-time (direct product of Pauli matrices σ's are implied):

(i) For $d = 2n$ (a_1,a_2 real)

$$\begin{aligned}
&\Gamma^0 = \sigma_1, \quad \Gamma^1 = i\sigma_2, \quad \Gamma^2 = i\sigma_3\sigma_1, \quad \Gamma^3 = i\sigma_3\sigma_2,\ldots,\\
&\Gamma^{2n-1} = \underbrace{i\sigma_3\cdots\sigma_2}_{n \text{ terms}}\\
&\overline{\Gamma} = \overbrace{\sigma_3\cdots\sigma_3}^{},\\
&\Lambda^\mu = i\overline{\Gamma}\otimes\Gamma^\mu.
\end{aligned} \tag{4.7}$$

(ii) For $d = 2n+1$ (a_1 real, a_2 pure imaginary),

$$\begin{aligned}
&\Gamma^0 = \sigma_2, \quad \Gamma^1 = i\sigma_1, \quad \Gamma^2 = i\sigma_3\sigma_2, \quad \Gamma^3 = i\sigma_3\sigma_1, \ldots,\\
&\Gamma^{2n} = \underbrace{i\sigma_3\cdots\sigma_3\sigma_2}_{n+1 \text{ terms}},\\
&\Lambda^0 = \overbrace{\sigma_3\cdots\sigma_3\sigma_1}^{},\\
&\Lambda^1 = i\sigma_3\cdots\sigma_3\sigma_3\sigma_2,\ldots,\\
&\Lambda^{2n} = \underbrace{i\sigma_3\cdots\cdots\cdots\sigma_3\sigma_1}_{2n+1 \text{ terms}}.
\end{aligned} \tag{4.8}$$

The choice of the constants a_1 and a_2 in each case is to ensure that the mass squared, $p^2 = m^2$, is bounded from below. The wave function Φ has 2^d components, which can be decomposed into four 2^{2n-2} (2^{2n-1}) component fields Φ_{ab} for $d = 2n(2n+1)$. The indices a and b ($=$ $+$ or $-$) denote eigenvalues of $\overline{\Gamma}$ and $I\otimes\overline{\Gamma}$ for $d = 2n$ ($I\otimes\overline{\Gamma}\otimes\sigma_3$ for $d = 2n+1$), respectively, where I is a $2^n\times 2^n$ matrix of unity. After taking the square of Eq. (4.6) and rearranging the components, we obtain

$$\begin{aligned}
&a_1^2 m^2(\Phi_{++} \pm \Phi_{+-})\\
&\quad = \left[-p^2 h_1^2 |a_2|^2 \frac{\partial^2}{\partial h_0^2} + \frac{h_0^2}{\kappa^2} \mp \frac{ph_1}{\kappa}|a_2|\right]\\
&\qquad \times(\Phi_{++} \pm \Phi_{+-}),\\
&a_1^2 m^2(\Phi_{-+} \pm \Phi_{--})\\
&\quad = \left[-p^2 h_1^2 |a_2|^2 \frac{\partial^2}{\partial h_0^2} + \frac{h_0^2}{\kappa^2} \pm \frac{ph_1}{\kappa}|a_2|\right]\\
&\qquad \times(\Phi_{-+} \pm \Phi_{--}).
\end{aligned} \tag{4.9}$$

Here $|a_2|$ represents the modulus of a_2. Now let us define the operators

$$\begin{pmatrix} C \\ C^+ \end{pmatrix} \equiv \frac{1}{\sqrt{2}}\left[\pm S\frac{\partial}{\partial h_0} + \frac{h_0}{S}\right] \tag{4.10}$$

with

$$S \equiv \sqrt{p|a_2|h_1\kappa},$$

and the mass

$$m_1 \equiv \sqrt{(2ph_1|a_2|/a_1^2\kappa)}. \tag{4.11}$$

It can be easily checked that C and C^+ satisfy

$$[C, C^+] = 1. \tag{4.12}$$

Also, if we define ϕ_0 such that $C\phi_0 = 0$, then the eigenvalues of the operator C^+C are non-negative integers, $n = 0,1,\ldots$, with eigenfunctions $\phi_n \equiv (C^+)^n\phi_0$. In terms of the operators C, C^+ and the mass m_1, we can write Eq. (4.9) as

$$\begin{aligned}
&m^2(\Phi_{++} \pm \Phi_{+-}) = m_1^2 \begin{bmatrix} C^+C \\ CC^+ \end{bmatrix}(\Phi_{++} \pm \Phi_{+-}),\\
&m^2(\Phi_{-+} \pm \Phi_{--}) = m_1^2 \begin{bmatrix} CC^+ \\ C^+C \end{bmatrix}(\Phi_{-+} \pm \Phi_{--}).
\end{aligned} \tag{4.13}$$

It follows that the eigenvalues of $p^2 = m^2$ are quantized with values

$$m^2 = m_1^2 n, \quad n = 0,1,2,\ldots. \tag{4.14}$$

Hence we conclude that our field equation admits an equally spaced mass-squared spectrum, and massless states.

V. CONCLUSIONS

In this paper, p-brane field theory has been defined as a surface functional theory which incorporates p-brane dynamics in Dirac form, as given by the covariant field equation (2.14). Connection of this field equation to the Nambu–Goto p-brane dynamics is made in the $X_0 = \tau$ gauge. Such correspondence necessarily requires introduction of multicomponent surface functionals, and new generalization of the Dirac algebra, Eqs. (2.16)–(2.18). As such, the p-brane fields transform nontrivially under Lorentz transformations, and describe either bosons or fermions depending on the parameters of the equations and the dimensionality of the space-time. This is very different from the conventional approaches, where the p-brane fields are always described by scalar functionals. Our theory thus represents a new type of p-brane theory.

In Sec. III and IV, we solve the p-brane field equation, Eq. (2.14), in $p+2$ dimensions in the $X_0 = \tau$ gauge. The field equation in this case naturally involves various differential-geometric quantities of hypersurfaces. It turns out that the field equation can be exactly solved for toroidal p-brane, of which all but the first total mean curvature vanish. The solutions yield an equally spaced mass-squared spectrum that contains massless states.

While we obtain the stringlike spectrum for toroidal p-brane in $p+2$ dimensions, there exists a "dual" approach to ours by which Morris[12] constructs string field theory using a set of $d-2$ functions to describe a string world-sheet embedded in d dimensions implicitly. It would be interesting to see the connection between these two approaches, which are apparently very different in their philosophy. We should also mention that Fujikawa[13] has recently suggested that a p-brane in $p+1$ dimensions essentially corresponds to a point

particle. He has examined this statement for $p = 1$, 2, and 3 in the light-cone gauge.

At present, the connection between our surface functional approach and the conventional ones, such as the light-cone and BRST schemes, is not clear to us. We hope to clarify it in the near future. Our field equations are constructed by incorporating the p-brane dynamics in the Dirac form in the $X_0 = \tau$ gauge. As it is well known, the Dirac equation of the point particle can also be obtained from quantization of a spinning particle (a particle whose action possesses local world-line supersymmetry).[14] It is therefore very tempting to see if our p-brane field equation follows from quantization of some "spinning p-brane," whose actions possess local world-volume supersymmetry.[15] To answer this question, one would require a Neveu–Schwarz–Ramond type of formulation of the spinning p-brane. Such possibility for membranes is considered by Castro,[16] using the new bosonic p-brane Lagrangians proposed by Dolan and Tchrakian,[17] and by Lindström and Roček.[18] However, Bergshoeff *et al.*[15] have shown that such models for spinning membranes cannot be constructed within the framework of the three-dimensional super-Poincaré tensor calculus.

ACKNOWLEDGMENTS

I would like to thank Professor Y. Hosotani for his many suggestions and continual support in this work. I am also grateful to Professor M. Hamermesh for helpful discussions on symmetric functions, and to the late Professor W. F. Pohl for enlightening me on differential geometry.

This research was supported in part by the U.S. Department of Energy under Contract No. DE-AC02-83ER 40105.

APPENDIX A: PRINCIPAL AND TOTAL MEAN CURVATURES OF T^n

We give a brief outline of the calculations of principal and total mean curvatures of an n-dimensional torus T^n.

First we give some formulas from the theory of surfaces in the language of differential forms.

Let $\mathbf{X}$ denote a hypersurface in E^{n+1}. At each point we can set up an orthonormal frame $\mathbf{e}_i$ $(i = 1,...,n+1)$. In terms of this basis, we have

$$\begin{aligned} &d\mathbf{X} = \sum_{i=1}^{n} \omega_i \mathbf{e}_i, \quad \omega_{n+1} = 0, \\ &d\mathbf{e}_i = \sum_{j=1}^{n} \omega_{ij}\mathbf{e}_j + \omega_{i,n+1}\mathbf{e}_{n+1}, \\ &\omega_{ij} = -\omega_{ji} \quad (i = 1,...,n), \\ &d\mathbf{e}_{n+1} = \sum_{i=1}^{n} \omega_{n+1,i}\mathbf{e}_i, \end{aligned} \tag{A1}$$

where ω_i's and ω_{ij}'s are one-forms. We also have the following structure equations:

$$\begin{aligned} &d\omega_i = \sum_{j=1}^{n} \omega_{ij} \wedge \omega_j, \\ &d\omega_{ij} = \sum_{k=1}^{n+1} \omega_{ik} \wedge \omega_{kj} \quad (i,j = 1,...,n+1). \end{aligned} \tag{A2}$$

Here $\wedge$ represents the exterior (wedge) product of the one-forms. From (A2), we have

$$d\omega_{n+1} = 0 = \sum_{k=1}^{n} \omega_{n+1,k} \wedge \omega_k. \tag{A3}$$

According to Cartan's lemma, Eq. (A3) implies

$$\begin{aligned} &\omega_{n+1,k} = -\sum_{j=1}^{n} h_{jk}\omega_j \quad (k = 1,...,n), \\ &h_{ij} = h_{ji}. \end{aligned} \tag{A4}$$

The principal curvatures κ_i $(i = 1,...,n)$ are eigenvalues of the matrix (h_{ij}). In terms of the one-forms ω_i's and ω_{ij}'s, the "volume" element of the surface and the two fundamental forms are given by

$$dV = \omega_1 \wedge \omega_2 \wedge \cdots \wedge \omega_n, \tag{A5}$$

$$\mathrm{I} = d\mathbf{X} \cdot d\mathbf{X} = \sum_{i=1}^{n} (\omega_i)^2, \tag{A6}$$

$$\mathrm{II} = -d\mathbf{X} \cdot d\mathbf{e}_{n+1} = \sum_{i,j=1}^{n} h_{ij}\omega_i\omega_j. \tag{A7}$$

Here the dot $\cdot$ represents the ordinary inner product of vectors.

Next, we shall derive a formula giving the principal curvatures of a hypersurface in E^{n+1}, obtained from the rotation of a hypersurface in E^n, whose principal curvatures are known.

Let $\mathbf{X}$ be a hypersurface in E^n with basis $\hat{i}_1,...,\hat{i}_n$. Let $\mathbf{Y}$ be a hypersurface in E^{n+1} (basis $\hat{i}_1,...,\hat{i}_{n+1}$) obtained from $\mathbf{X}$ when the axis $\hat{i}_n$ is rotated through an angle 2π in the $\hat{i}_n - \hat{i}_{n+1}$ plane (all other axes being kept fixed). The angle of rotation, measured from $\hat{i}_n$, is denoted by θ_n. The orthonormal basis at each point of $\mathbf{X}$ and $\mathbf{Y}$ are denoted by $(\mathbf{e}'_i, \mathbf{N}^{(n)} \equiv \mathbf{e}'_n)$ and $(\mathbf{e}_j, \mathbf{N}^{(n+1)} \equiv \mathbf{e}_{n+1})$ $(i = 1,...,n-1, j = 1,...,n)$, respectively. We obviously have

$$\begin{aligned} &\mathbf{Y} = \mathbf{X} - (\mathbf{X}\cdot\hat{i}_n)\hat{i}_n + R_n(\cos\theta_n\hat{i}_n + \sin\theta_n\hat{i}_{n+1}), \\ &R_n = r_n + \mathbf{X}\cdot\hat{i}_n, \end{aligned} \tag{A8}$$

if the surface $\mathbf{X}$ is translated along the axis $\hat{i}_n$ by an amount r_n before it is rotated. Suppose

$$d\mathbf{X} = \sum_{k=1}^{n-1} \omega'_k \mathbf{e}'_k. \tag{A9}$$

Then we have

$$d\mathbf{Y} = \sum_{k=1}^{n-1} \omega_k \mathbf{e}_k + \omega_n \mathbf{e}_n, \tag{A10}$$

where

$$\begin{aligned} \mathbf{e}_k \equiv{}& \mathbf{e}'_k - (\mathbf{e}'_k\cdot\hat{i}_n)\hat{i}_n + (\mathbf{e}'_k\cdot\hat{i}_n)(\cos\theta_n\hat{i}_n \\ &+ \sin\theta_n\hat{i}_{n+1}) \quad (k = 1,...,n-1), \\ \mathbf{e}_n \equiv{}& -\sin\theta_n\hat{i}_n + \cos\theta_n\hat{i}_{n+1} \end{aligned} \tag{A11}$$

and

$$\begin{aligned} &\omega_k = \omega'_k \quad k = 1,...,n-1, \\ &\omega_n = R_n\, d\theta_n. \end{aligned} \tag{A12}$$

Also, for the normal vectors $\mathbf{N}^{(n)}$ and $\mathbf{N}^{(n+1)}$, we have

$$\begin{aligned} \mathbf{N}^{(n+1)} = {}& \mathbf{N}^{(n)} - (\mathbf{N}^{(n)}\cdot\hat{i}_n)\hat{i}_n \\ &+ (\mathbf{N}^{(n)}\cdot\hat{i}_n)(\cos\theta_n\hat{i}_n + \sin\theta_n\hat{i}_{n+1}), \end{aligned}$$

$$d\mathbf{N}^{(n)} = \sum_{k=1}^{n-1} \omega'_{n,k}\, \mathbf{e}'_k, \quad \omega'_{n,k} = -\sum_{j=1}^{n-1} h'_{jk}\omega'_j,$$
$$d\mathbf{N}^{(n+1)} = -\sum_{k=1}^{n-1}\left(\sum_{j=1}^{n-1} h'_{jk}\omega_j\right)\mathbf{e}_k + (\mathbf{N}^{(n)}\cdot\hat{i}_n)d\theta_n\, \mathbf{e}_n, \tag{A13}$$

where in the last expression we have used the fact that $\omega'_k = \omega_k$ $(k = 1,...,n-1)$ from Eq. (A12). Now, using

$$\omega_{i,n+1} = -\mathbf{e}_i \cdot d\mathbf{N}^{(n+1)} \tag{A14}$$

and Eqs. (A4) and (A12), we finally obtain

$$\begin{bmatrix} \omega_{i,n+1} \\ \hline \omega_{n,n+1} \end{bmatrix} = \left[\begin{array}{c|c} h'_{ij} & \\ \hline & -\mathbf{N}^{(n)}\cdot\hat{i}_n/R_n \end{array}\right]\begin{bmatrix} \omega_j \\ \hline \omega_n \end{bmatrix}. \tag{A15}$$

The first square bracket on the right-hand side of Eq. (A15) gives the new h_{ij}'s. Hence knowing the principal curvatures and the normal vector of $\mathbf{X}$, we can calculate the principal curvatures of $\mathbf{Y}$ according to Eq. (A15).

Now we take the hypersurface to be an n-dimensional torus T^n. Let the radius of circle S^1 on the $\hat{i}_k - \hat{i}_{k+1}$ plane be r_k. We also define $\theta_0 \equiv 0$, and $\mathbf{N}^{(1)} \equiv -\hat{i}_1$. Then it is easy, from the above formulas, to obtain the following results:

$$R_1 \equiv r_1, \quad R_j = r_j + R_{j-1}\sin\theta_{j-1} \quad (\theta_0 \equiv 0),$$
$$-\mathbf{N}^{(1)}\cdot\hat{i}_1 = +1, \quad -N^{(j)}\cdot\hat{i}_n = \prod_{k=1}^{j-1}\sin\theta_k, \tag{A16}$$
$$\omega_j = R_j\, d\theta_j \quad (j = 1,...,n).$$

According to Eq. (A15), the principal curvatures of T^n are obtained to be

$$\kappa_1 = 1/R_1, \quad \kappa_2 = \sin\theta_1/R_2,...$$
$$\kappa_n = \prod_{k=1}^{n-1}\sin\theta_k/R_n. \tag{A17}$$

Our final task is to calculate the total mean curvatures $h_l \equiv \int H_l\, dV$ $(l = 1,...,n)$. It is easy to check, with Eqs. (A16) and (A17), that the only nonvanishing integral of the forms $\int \kappa_i \cdots \kappa_j\, dV$ is $\int \kappa_1\, dV$. Since H_l is the symmetric polynomials of l κ's, we conclude that

$$h_1 \neq 0,$$
$$h_k = 0 \quad (k = 2,...,n). \tag{A18}$$

APPENDIX B: EXPLICIT FORMS OF THE FUNCTIONS U

Let $\mathbf{L}_r \equiv -(r+1)\lambda_{r+1} + r\zeta_r$.

(i) For $p = 2m$,

$$m = 1, \quad U = \exp i\left\{-\frac{\mathbf{p}\cdot\boldsymbol{\lambda}_1}{2h_1} - \frac{(\mathbf{p}\cdot\mathbf{L}_1)^2}{2^4 h_1(\mathbf{p}\cdot\boldsymbol{\zeta}_2)}\right\},$$

$$m \geqslant 2, \quad U = \exp i\left\{-\frac{\mathbf{p}\cdot\boldsymbol{\lambda}_1}{ph_1} + \frac{1}{(p^3 h_1)(\mathbf{p}\cdot\boldsymbol{\zeta}_p)}\left[-(\mathbf{p}\cdot\mathbf{L}_1)(\mathbf{p}\cdot\mathbf{L}_{2m-1})\right.\right.$$
$$+\frac{2!}{(2m-1)}(\mathbf{p}\cdot\mathbf{L}_2)(\mathbf{p}\cdot\mathbf{L}_{2m-2}) - \frac{3!}{(2m-1)(2m-2)}(\mathbf{p}\cdot\mathbf{L}_3)(\mathbf{p}\cdot\mathbf{L}_{2m-3}) + \cdots$$
$$\left.\left.+\frac{(-1)^{m-1}(m-1)!}{(2m-1)\cdots(m+2)}(\mathbf{p}\cdot\mathbf{L}_{m-1})(\mathbf{p}\cdot\mathbf{L}_{m+1}) + \frac{1}{2}\frac{(-1)^m m!}{(2m-1)\cdots(m+1)}(\mathbf{p}\cdot\mathbf{L}_m)^2\right]\right\}.$$

(ii) For $p = 2m+1$,

$$m = 0, \quad U = 1,$$

$$m = 1, \quad U = \exp i\left\{-\frac{\mathbf{p}\cdot\boldsymbol{\lambda}_1}{3h_1} - \frac{(\mathbf{p}\cdot\mathbf{L}_1)(\mathbf{p}\cdot\mathbf{L}_2)}{3^3 h_1(\mathbf{p}\cdot\boldsymbol{\zeta}_3)} + \frac{2(\mathbf{p}\cdot\mathbf{L}_2)^3}{3^6 h_1(\mathbf{p}\cdot\boldsymbol{\xi}_3)^2}\right\},$$

$$m \geqslant 2, \quad U = \exp i\left\{-\frac{\mathbf{p}\cdot\boldsymbol{\lambda}_1}{ph_1} + \frac{1}{(p^4 h_1)(\mathbf{p}\cdot\boldsymbol{\zeta}_p)}\left[-p(\mathbf{p}\cdot\mathbf{L}_1)(\mathbf{p}\cdot\mathbf{L}_{2m}) + \frac{2!}{2m}[1+2(m-2)](\mathbf{p}\cdot\mathbf{L}_2)(\mathbf{p}\cdot\mathbf{L}_{2m-1})\right.\right.$$
$$-\frac{3!}{2m(2m-1)}[1+2(m-3)](\mathbf{p}\cdot\mathbf{L}_3)(\mathbf{p}\cdot\mathbf{L}_{2m-2}) + \cdots$$
$$+\frac{(-1)^m m!}{2m(2m-1)\cdots(m+2)}(\mathbf{p}\cdot\mathbf{L}_m)(\mathbf{p}\cdot\mathbf{L}_{m+1})\Big] + \frac{2(\mathbf{p}\cdot\mathbf{L}_{2m})}{p^6 h_1(p\cdot\boldsymbol{\zeta}_p)^2}\left[2!(\mathbf{p}\cdot\mathbf{L}_2)(\mathbf{p}\cdot\mathbf{L}_{2m})\right.$$
$$-\frac{3!}{2m}(\mathbf{p}\cdot\mathbf{L}_3)(\mathbf{p}\cdot\mathbf{L}_{2m-1}) + \cdots$$
$$\left.\left.+\frac{(-1)^m m!}{2m(2m-1)\cdots(m+3)}(\mathbf{p}\cdot\mathbf{L}_m)(\mathbf{p}\cdot\mathbf{L}_{m+2}) + \frac{1}{2}\frac{(-1)^{m+1}(m+1)!}{2m(2m-1)\cdots(m+2)}(\mathbf{p}\cdot\mathbf{L}_{m+1})^2\right]\right\}.$$

[1]See, e. g., M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge U.P., New York, 1987), Vols. I and II.
[2]See the following reviews, and the references therein: M. J. Duff, Class. Quantum Grav. **5**, 189 (1988); C. N. Pope, University of Southern California preprint USC-87/HEP 07; K. S. Stelle and P. K. Townsend, Imperial/TP/87-88/5; E. Bergshoeff, E. Sezgin, and P. K. Townsend, Ann. Phys. (NY) **185**, 330 (1988).
[3]J. Hughes, J. Liu, and J. Polchinski, Phys. Lett. B **180**, 370 (1986); E. Bergshoeff, E. Sezgin, and P. K. Townsend, *ibid.* B **189**, 75 (1987); A. Achucarro, J. Evans, P. K. Townsend, and D. L. Wiltshire, *ibid.* **198**, 441 (1987).
[4]See A. Achucarro *et al.* in Ref. 3 and the review by M. J. Duff in Ref. 2.
[5]E. G. Floratos and J. Illiopoulos, Phys. Lett. B **201**, 237 (1988); I. Antoniadis, P. Ditsas, E. Floratos, and J. Illiopoulos, Nucl. Phys. B **300**, 549 (1988); I. Bars, C. N. Pope, and E. Sezgin, Phys. Lett. B **210**, 85 (1988).
[6]M. J. Duff, T. Inami, C. N. Pope, E. Sezgin, and K. S. Stelle, Nucl. Phys. B **297**, 515 (1988); E. Bergshoeff, E. Sezgin, and Y. Tanii, *ibid.* **298**, 187 (1988); K. Fujikawa and J. Kubo, Phys. Lett. B **199**, 75 (1987); K. Fujikawa, *ibid.* **206**, 18 (1988); J. Gamboa and M. Ruiz-Altaba, *ibid.* **205**, 245 (1988).
[7]K. Kikkawa and M. Yamasaki, Prog. Theor. Phys. **76**, 1379 (1986); M. J. Duff *et al.* in Ref. 6; I. Bars, C. N. Pope, and E. Sezgin, Phys. Lett. B **198**, 455 (1987); L. Mezincescu, I. Nepomechie, and P. Van Nieuwenhuizen, University of Miami preprint No. UMTG-139, 1987 and SUNY Stony Brook preprint No. ITP-SB-87-43, 1987.
[8]C. L. Ho and Y. Hosotani, Phys. Rev. Lett. **60**, 885 (1988).
[9]L. Carson and Y. Hosotani, Phys. Rev. Lett. **56**, 2144 (1986); Phys. Rev. D **37**, 1492 (1988).
[10]For other reparametrization-invariant approaches to string field theory, see Ref. 7 quoted in the second paper in Ref. 9, and also the following recent papers: G. Kleppe, P. Ramond, and R. R. Viswanathan, Phys. Lett. B **206**, 466 (1988); University of Florida preprint UFTP-88-9, 1988.
[11]The treatment outlined below is a straightforward generalization of that given in the second paper in Ref. 9. We shall refer the reader to that paper, particularly Sec. IV, for a detailed discussion.
[12]T. R. Morris, Phys. Lett. B **202**, 222 (1988).
[13]K. Fujikawa, Phys. Lett. B **213**, 425 (1988).
[14]F. A. Berezin and M. S. Marinov, Ann. Phys. NY **104**, 336 (1977); L. Brink, S. Deser, B. Zumino, P. di Vecchi, and P. Howe, Phys. Lett. B **64**, 435 (1976); P. G. O. Freund in A. Ferber, Nucl. Phys. B **132**, 55 (1978); L. Brink and J. H. Schwarz, Phys. Lett. B **100**, 310 (1981); P. G. O. Freund, *Introduction to Supersymmetry* (Cambridge U.P. Cambridge, 1986).
[15]E. Bergshoeff, E. Sezgin, and P. K. Townsend, Phys. Lett. B **209**, 451 (1988).
[16]C. Castro, "A supersymmetric Lagrangian for the spinning membrane," University of Texas (Austin) preprint (1988).
[17]B. P. Dolan and D. H. Tchrakian, Phys. Lett. B **202**, 211 (1988).
[18]U. Lindström and M. Roček, SUNY Stony Brook preprint ITP-SB-88-61 (1988).

Wigner's little group and decomposition of Lorentz transformations

Dimitris V. Vassiliadis
Department of Physics, University of Maryland, College Park, Maryland 20742

(Received 5 January 1989; accepted for publication 3 May 1989)

It is shown how an arbitrary Lorentz transformation can be expressed in terms of elements of Wigner's little group and its cosets. This yields a natural parametrization for the little group, while its coset members turn out to be helicity-preserving transformations. The associated Wigner angle and its relation to the actual change in helicity are discussed. Finally, the extension to zero-mass particles shows how the little group becomes a gauge transformation in that limit.

I. INTRODUCTION

In a 1957 paper,[1] Wigner discusses the role of the Lorentz transformations in understanding the internal symmetries of space-time, as they refer to a particle's four-momentum and spin state.

In particular, there are certain transformations that leave such quantities invariant after their application. By definition, those that leave a four-momentum invariant form the so-called Wigner's little group for that momentum. Other transformations change the momentum but leave the helicity invariant, as, for instance, any rotation would do. A generic transformation will then affect both the four-momentum and the helicity. In fact, the resulting state of these two quantities will depend on the particles's mass, as well as on the "path" it follows, as it is boosted and rotated in the momentum space.

The theorem to be shown in Sec. II states that an arbitrary Lorentz transformation may be written as the product of a member of the little group times a member of its (left or right) coset. That coset member is a transformation that leaves the helicity invariant. Expressions for the parameters of the little group and the coset members can then be calculated. Working in the O(3) formalism (Sec. III) we find an expression for the Wigner angle[1-4] related to the little-group member; subsequently we calculate the actual amount of spin rotation relative to the momentum direction now using the spinor representation. It is interesting to see how the little-group transformation changes form and meaning when it is applied to massless particles.[5] In Sec. IV we find that it becomes a gauge transformation matrix; its different contents in the two representations are then discussed.

The above theorem has been discussed previously in the literature,[2,4] but this was only done for special cases of the parameters involved. This paper aims to generalize those results in an attempt to unify them.

II. COSET DECOMPOSITION OF THE LORENTZ GROUP

Lorentz transformations can be visualized in the environment of the three-dimensional momentum space. A momentum state is specified by the triad (p_x,p_y,p_z) in that space. For the sake of simplicity the axial directions will be referred to as x,y,z instead of p_x,p_y,p_z, respectively. Boosts are performed by vector additions and rotations refer to the origin. Spin and its rotations, helicity, and other features related to spin are not shown in such a picture.

Considering an arbitrary boost $B_\phi(\beta)$, of boost parameter β, it is interesting to examine if it can be decomposed in parts, each of which keeps either the helicity or the four-momentum constant. The boost changes the momentum state of a particle in a well-defined way, bringing it from $\mathbf{p}$ to $\mathbf{p}'$ (Fig. 1). It is not obvious, though, how the helicity changes in the process. For simplicity we choose the initial state ($\mathbf{p}$) to lie along the z axis, and then that axis together with B_ϕ define the x-z plane (no loss of generality). The angle ϕ can have any value between 0° and 180°, measured from the z axis. Using the four-vector representation, we first construct the state $\mathbf{p}$ by applying a boost $A_z(\alpha)$ on the unit-mass particle, initially defined to be at rest and in the positive-helicity state:

$$A_z(\alpha)\begin{bmatrix}0\\0\\0\\1\end{bmatrix}=\begin{bmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1/a&\alpha/a&0&\alpha/a\\0&0&\alpha/a&1/a&1&1/a\end{bmatrix}=\begin{matrix}0\\0\\\alpha/a\\1/a\end{matrix}\equiv\mathbf{p} \qquad (1)$$

[where α is the boost velocity of A_z and $a=(1-\alpha^2)^{1/2}$]; no change of helicity has occurred.

After that, we can apply $B_\phi(\beta)$ on state $\mathbf{p}$. Here

$$B_\phi(\beta)=\begin{bmatrix}1+(\gamma-1)\beta_x^2/\beta^2 & 0 & (\gamma-1)\beta_x\beta_z/\beta^2 & \beta_x\gamma\\ 0 & 1 & 0 & 0\\ (\gamma-1)\beta_x\beta_z/\beta^2 & 0 & 1+(\gamma-1)\beta_z^2\beta^2 & \beta_z\gamma\\ \beta_x\gamma & 0 & \beta_z\gamma & \gamma\end{bmatrix}, \qquad (2)$$

where $\beta_x=\beta\sin(\phi)$, $\beta_z=\beta\cos(\phi)$, and $\gamma=(1-\beta^2)^{-1/2}$.

The resulting momentum state (point $\mathbf{p}'$) can be reached by other paths on the x-z plane. All of them will have the same effect on four-momentum, but they will affect helicity differently. A transformation that preserves helicity is, for example, a rotation $R(\vartheta)$ applied in succession to a boost in the z direction $B_z(\epsilon)$, in such a way that we again reach $\mathbf{p}'$

 0022-2488/89/092177-04$02.50

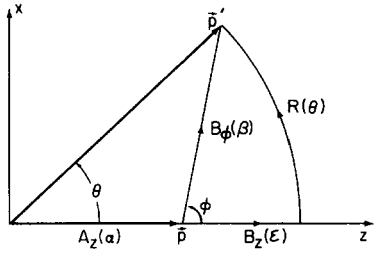


FIG. 1. The two-dimensional momentum space for Sec. II. A state **p** is first created by boost A_z and then boosted by B_ϕ to **p**'. Here B_ϕ can be decomposed in a helicity-preserving part $R(\vartheta)B_z(\epsilon)$ and a little-group member. The latter is momentum preserving by definition (therefore absent from momentum space), but gives an additional rotation to the particle spin.

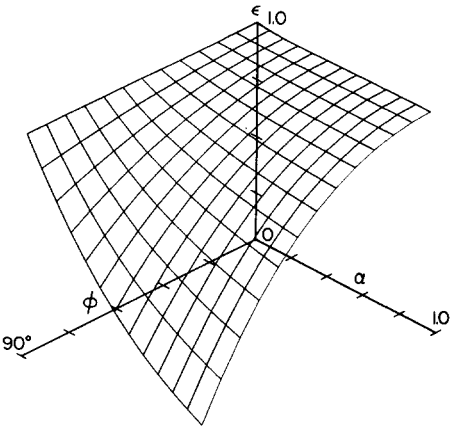


FIG. 2. Boost parameter ϵ of zB_z depends on all α,β,ϕ parameters (cf. Fig. 1). Here ϵ is plotted versus ϕ (from 0°–90°) and α. The resulting surface corresponds to a β of 0.6. Note that ϵ becomes 1 for $\alpha = 0, \phi = 0$ (irrespective of β).

eventually,

$$B_z(\epsilon) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/e & \epsilon/e \\ 0 & 0 & \epsilon/e & 1/e \end{bmatrix},$$

$$R(\vartheta) = \begin{bmatrix} \cos(\vartheta) & 0 & \sin(\vartheta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\vartheta) & 0 & \cos(\vartheta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

with $\epsilon =$ (boost velocity of B_z) and $e = (1-\epsilon^2)^{1/2}$.

Both transformations $R(\vartheta)$ and $B_z(\epsilon)$ are helicity preserving, so if the particle follows the second path, its spin will not change its orientation relative to the momentum. A direct way to show that the two paths B_ϕ and RB_z are not totally equivalent, is to compute the "closed-loop" matrix product $B_\phi^{-1}RB_z$. Thus

$$D = B_\phi^{-1}R(\vartheta)B_z$$

$$= \begin{bmatrix} T-1 & 0 & -u & \alpha u \\ 0 & 1 & 0 & 0 \\ u & 0 & 1-u^2/T & \alpha u^2/T \\ \alpha u & 0 & -\alpha u^2/T & 1+\alpha^2 u^2/T \end{bmatrix}, \tag{4}$$

where

$$F = \sqrt{(\alpha\beta_z+1)^2 - a^2b^2},\ u = -\beta_x/F,$$
$$T = 1 + (\alpha+\beta_z)/F.$$

In general, matrix (4) is not the identity matrix. We can say that D is "equivalent to B_ϕ with respect to helicity" (both result in the same change of helicity when applied to **p**). Since D maps a point of the momentum space in itself, or keeps its four-momentum invariant, it is a member of the little group for **p**, by definition.

We see that Wigner's little group can be parametrized using α (boost parameter of **p**), β, ϕ (parameters of the helicity-equivalent boost B_ϕ) as the relevant quantities [ϵ,ϑ are functions of those three variables as follows

$$\epsilon = \frac{-\alpha+\gamma^2(\alpha\beta_z+1)}{\alpha^2+\gamma^2(\alpha\beta_z+1)^2},$$
$$\sin(\vartheta) = \beta_x \frac{\alpha\beta_z+1+1/\gamma}{(\alpha\beta_z+1)(\gamma+1)}; \tag{5}$$

also see Fig. 2 for the variation of ϵ with α, ϕ].

The above choice of parameters is justified if we consider the case where $\epsilon = 0$ (no difference between initial and final boost parameters). In this case the parametrization of D becomes the so-called Eulerian parametrization[4] of Wigner's little group, which is now locally isomorphic to O(3). The D is expressed in terms of the z boost it acts on, and an angle around the y direction.

Now we can easily prove the theorem. We solve the above matrix equation for B_ϕ:

$$B_\phi(\beta) = R(\vartheta)B_z(\epsilon)D^{-1}; \tag{6}$$

from this we see that arbitrary B_ϕ has been broken down in two pieces: RB_z is helicity invariant and only changes the four-momentum, while D^{-1} does not affect the momentum state and only transforms the helicity. In the above equation the helicity-invariant part is a member of the left coset of the little group, but the expression can be rewritten in terms of a right coset member [$B_\phi = (RB_zD^{-1}R^{-1}B_z^{-1})(B_zR)$].

III. WIGNER ANGLE AND CHANGE OF HELICITY

In this section we go on deriving some results using the formalism of Sec. II. Then we see how this approach fits to the spinor representation.

A. O(3) formalism

In Eq. (6) the only terms that bring about a change in the momentum **p** are the helicity-preserving R,B_z; the action of D^{-1} does not apply on the kinematics of the particle. Therefore, in order to change helicity, D^{-1} can only rotate the spin. A measure of this rotation can be obtained if we consider an alternative momentum-preserving transformation: $A_zR_wA_z^{-1}$ and set it equal to D. Essentially we move the particle to the rest frame, rotate its spin there, and then restitute its momentum. Then

$$R_W(\Phi_W) = A_z^{-1}DA_z. \tag{7}$$

The angle Φ_W is the so-called Wigner angle corresponding to the little-group element D. However, $R_W(\Phi_W)$ is *not* the actual rotation of the spin as the particle is boosted from

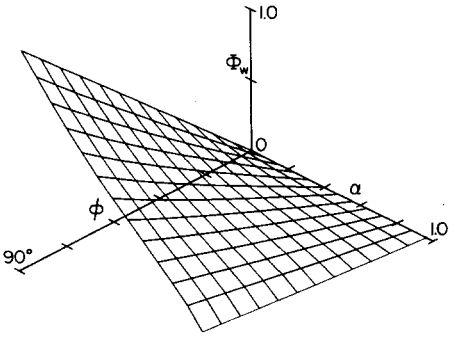


FIG. 3. When a spinor **p** is boosted, there is a change in the angle between the spin and momentum directions (helicity). This change is expressed by means of the conventional (not actual) "Wigner angle" Φ_W. Its dependence (in radians) on angle ϕ and boost parameter α is shown. The boost parameter of B_ϕ is $\beta = 0.6$. Notice how Φ_W goes to zero for $\alpha \to 1$ (then D becomes a gauge transformation of massless particles).

p to **p**', but a mere convention. This is so because, in the sequence $A_z R_W A_z^{-1}$, after R_W has rotated the spin, the boost A_z is not parallel to the spin direction anymore: its application now produces an additional spin rotation. The actual value(s) of the angle between the spin and the momentum after the boost will be given more elegantly, in terms of the spinor formalism below.

A direct comparison of the rhs of (7) with the rotation matrix by angle Φ_W around the y axis yields

$$\tan(\Phi_W) = \beta_x a/(\alpha + \beta_z) \tag{8}$$

(see Fig. 3). So Φ_W gives a measure of the angle by which the spin is rotated when a boost with components $(\beta_x, 0, \beta_z)$ is applied on a particle at velocity α [again taking into account Eq. (5)].

B. SL(2,c) formalism

For these particles one can use the spinor formalism. The spinors obey the Lie algebra:

$$[S_i, S_j] = i\epsilon_{ijk} S_k,$$
$$[S_i, K_j] = i\epsilon_{ijk} K_k,$$
$$[K_i, K_j] = -i\epsilon_{ijk} S_k,$$
$$S_i = \tfrac{1}{2}\sigma_i, \quad K_i = \pm (1/2)\sigma_i.$$

We note that if a set of K_i's satisfies the above equations, then they hold for the set of opposite elements ($-K_i$'s) as well; however, a similar statement does not hold for the S_i's.

This Lie group acts on the normalized Pauli spinors. We distinguish between spinors that obey a Lie algebra where K_i's have a positive and a negative sign. So

$$\chi_+, \chi_-, \quad \text{for} \quad K_i = 1/2i\sigma_i,$$
$$\dot{\chi}_+, \dot{\chi}_-, \quad \text{for} \quad K_i = -1/2i\sigma_i,$$

where $+$ and $-$ stand for "up" and "down," respectively.

The Dirac equation relates the dotted spinors to the undotted ones, thus leaving only two out of four. However, it seems that supersymmetric theories have a richer structure if all spinors are taken to be independent, and therefore we will treat them as such. In the following considerations we will deal with undotted spinors only. Results for the dotted ones will be simply quoted in the end.

The matrices of Sec. II will take on the form

$$B_\phi(\beta) = \begin{bmatrix} C + Sn_z & Sn_x \\ Sn_x & C - Sn_z \end{bmatrix}, \quad R(\vartheta) = \begin{bmatrix} c & -s \\ s & c \end{bmatrix},$$
$$A_z = \begin{bmatrix} N & 0 \\ 0 & 1/N \end{bmatrix},$$

with $C = (\frac{1}{2}(\gamma - 1))^{1/2}$, $S = (\frac{1}{2}(\gamma + 1))^{1/2}$, $n_x = \sin(\phi)$, $n_z = \cos(\phi)$, $c = \cos(\vartheta/2)$, $s = \sin(\vartheta/2)$, and $N = [(1+\alpha)/(1-\alpha)]^{1/4}$. Then the transformation D equals

$$D^{(+)} = \begin{bmatrix} \sqrt{T/2} & auN^2 \\ -au/N^2 & \sqrt{T/2} \end{bmatrix} \tag{9}$$

(where u, T are the expressions defined in Sec. II).

We could now go on and repeat the calculations of Sec. III A to compute the Wigner angle. Instead, though, we can easily find the angle formed by the spin and momentum directions. When the particle is at the origin, it can be represented by a spinor of the form $\chi_\pm$ (assume a polarization along the z axis). After $A_z(\alpha)$ is applied, the resulting matrix, $\chi_\pm(p) = A_z(\alpha)\chi_\pm = N^{\pm 1}\chi_\pm$, is boosted by B_ϕ to become

$$\chi_\pm(\mathbf{p}') = B_\phi(\beta)\chi_\pm(\mathbf{p}) = N^{\pm 1}\chi'_\pm,$$

where $\chi'_\pm$ is

$$\chi'_+ = \begin{bmatrix} \left(\frac{1+b}{2b}\right)^{1/2} - \cos(\phi)\left(\frac{1+b}{2b}\right)^{1/2} \\ -\sin(\phi)\left(\frac{1-b}{2b}\right)^{1/2} \end{bmatrix},$$
$$\chi'_- = \begin{bmatrix} -\sin(\phi)\left(\frac{1-b}{2b}\right)^{1/2} \\ \left(\frac{1+b}{2b}\right)^{1/2} + \cos(\phi)\left(\frac{1-b}{2b}\right)^{1/2} \end{bmatrix},$$

with $b = (1-\beta^2)^{1/2} = 1/\gamma$.

The above spinors can also be obtained (in normalized form) using a pure rotation by an angle $\omega_\pm$ around the y axis. The angles of rotation for dotted and undotted spinors will equal

$$\tan\left(\frac{\omega_+}{2}\right) = \frac{1 + b \pm \beta\cos(\phi)}{\pm\beta\sin(\phi)},$$
$$\tan\left(\frac{\omega_-}{2}\right) = \frac{\pm\beta\sin(\phi)}{1 + b \mp \beta\cos(\phi)}, \tag{10}$$

where the upper sign refers to undotted spinors and the lower to dotted ones.

So the angle δ between the spin and momentum at the point **p**' (in other words, the change in helicity resulting from B_ϕ) will be

$$\delta_\pm = \vartheta - \omega_\pm, \tag{11}$$

ϑ still having the same value as the one given by (5).

These angles are plotted versus α and ϕ in Fig. 4 for the case of undotted spinors. In Fig. 5 their relative magnitudes are shown in comparison to Φ_W.

IV. THE MASSLESS-PARTICLE LIMIT

So far our approach has been mass independent. It is interesting to check what happens if, keeping the momen-

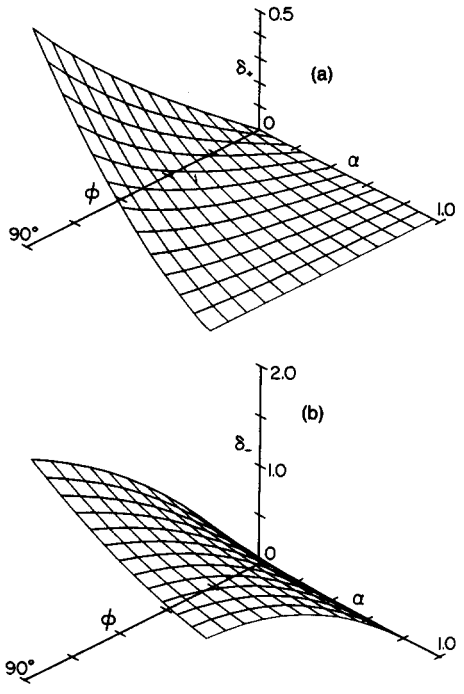


FIG. 4. The *actual* change in helicity, $\delta_{\pm}$ (sign referring to the initial z polarization of the spinor). These angles are plotted against ϕ and α (for undotted spinors). (a) δ_+ and (b) δ_-; $\beta = 0.6$. Notice that for $\alpha \to 1$, δ_+ goes to zero together with Φ_W, unlike δ_- (that spinor remains nonaligned).

tum finite, we take the mass to be zero, in other words, if the particle moves at the velocity of light. That "infinite momentum/zero-mass" limit, as it is often called, can be taken in the end of our calculations, by setting $\alpha = 1$. In this case there is no notion of a rest frame. So the approach taken to derive (7) cannot be repeated here. This is expressed in the limiting value of the Wigner angle Φ_W: it becomes identically zero [cf. (8)]. However, because the elements of A_z take on infinite values, D remains finite. It has the form

$$D(\alpha = 1,\beta,\phi) = \begin{bmatrix} 1 & 0 & -u & u \\ 0 & 1 & 0 & 0 \\ u & 0 & 1-u^2/2 & u^2/2 \\ u & 0 & -u^2/2 & 1+u^2/2 \end{bmatrix},$$

with $u = \beta_x/(\beta_z + 1)$, which is identical to that of a gauge transformation that can be applied on a photon four-potential $A^\mu = (A_x,0,A_z,\omega)$. The above expression for $D(\alpha = 1)$ remains a gauge transformation seen from any frame of reference: $D'(u) = B(\beta)D(u)B^{-1}(\beta) = D(u')$; where $u' = [(1+\beta)/(1-\beta)]^{1/2}u$. For the special case $\varepsilon = 0$, this transition from the Wigner rotation to a gauge transformation matrix shows how the little group becomes locally isomorphic to the E(2) group. The nature of the contraction and the singularity has already been discussed in the literature.[4,6]

Similar to Sec. III, we can extend these conclusions for spinors. If we apply $D^{(\pm)}$ to $\chi_\pm, \dot{\chi}_\pm$, we see that, in the $\alpha = 1$ limit, two of them remain invariant, while each of the

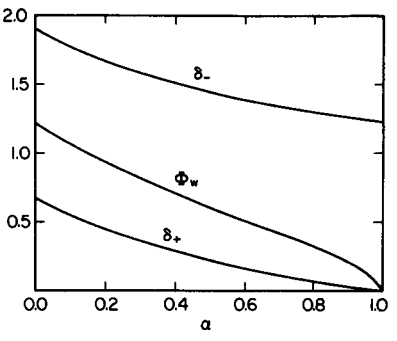


FIG. 5. The Wigner and $\delta_\pm$ angles (in radians) against α; parameters are $\beta = 0.6$ and $\phi = 70°$.

two others mixes with the opposite-polarization spinor:

$$D^{(+)}\chi_+ = \chi_+, \quad D^{(+)}\chi_- = \chi_- + u\chi_+,$$
$$D^{(-)}\dot{\chi}_- = \dot{\chi}_- - u\chi_+, \quad D^{(-)}\dot{\chi}_- = \dot{\chi}_-. \tag{12}$$

Physically the invariant set can be polarized neutrinos. For the two others, helicity is not preserved under the action of the little group as we go to the zero-mass limit; they are not forced to align. These two extra degrees of freedom correspond to the gauge degrees of freedom in the case of photons.

V. CONCLUDING REMARKS

In this paper, the general form of a decomposition theorem for Lorentz transformations was shown. It turns out that a boost can be decomposed in two factors: one keeps the four-momentum invariant (is a member of the little group for that momentum), the other one preserves helicity. In this way we are led to a natural parametrization of Wigner's little group. A member of that group, D, is expressed in terms of the boost it acts on and the parameters (boost β, relative angle ϕ) of a boost that is equivalent to D with respect to helicity. Little-group members also have a Wigner's angle associated with them. This gives a measure of the amount of spin rotation when D acts on the particle (momentum and spin) state. Here we calculated this angle as well as the actual change in helicity (using the spinor representation). Finally, it was shown that, at the zero-mass limit (case of photons or neutrinos), the little group turns into a gauge transformation matrix. Only one out of two spinors in each pair aligned with the momentum in this limit. The nonalignment of the remaining spinors gives rise to the particle's gauge degrees of freedom.

ACKNOWLEDGMENTS

We wish to thank Professor Y. S. Kim for his helpful comments during the preparation of this paper.

[1]E. P. Wigner, Rev. Mod. Phys. **29**, 255 (1957).
[2]D. Han, Y. S. Kim, and D. Son, J. Math. Phys. **28**, 2373 (1987).
[3]V. I. Ritus, Soviet Phys. JETP **13**, 240 (1961).
[4]D. Han, Y. S. Kim, and D. Son, J. Math. Phys. **27**, 2228 (1986).
[5]J. Kupersztych, Nuovo Cimento B **31**, 1 (1976); Phys. Rev. D **17**, 629 (1978).
[6]Y. S. Kim and E. P. Wigner, J. Math. Phys. **28**, 1175 (1987).

Gauge anomalies on S^2 and group extensions

Alan Carey
Department of Pure Mathematics, University of Adelaide, Adelaide, South Australia

John Palmer
Department of Mathematics, University of Arizona, Tucson, Arizona 85721

(Received 28 February 1989; accepted for publication 3 May 1989)

A geometric connection is established between the group cocycle used by Mickelsson [*Topological and Geometrical Methods in Field Theory* (World Scientific, Singapore, 1986), pp. 117–131] to realize the Kac–Moody extension of the group of loops in SU(n) as the quotient of a topologically trivial extension of maps from the disk into SU(n) and the gauge anomaly for Quillen's determinant bundle over the Cauchy–Riemann operators on the spin bundle over S^2.

I. INTRODUCTION

Let G denote the group SU(n). In Ref. 1, Mickelsson observed that the gauge anomaly for chiral fermions on the two sphere[2] could be used to define a central extension, $\widehat{DG}$, of the maps from the disk, D, into G, which has the Kac–Moody extension of the loop group as a natural quotient. Because the central extension of the maps from the disk into G is topologically trivial and this is not the case for the Kac–Moody extension, one has achieved something in this realization (one may find some further discussion of Mickelsson's construction in Frenkel's paper[3]).

In this paper we are concerned with examining the relation between the group extension defined by Mickelsson and the gauge anomaly for chiral fermions. In Sec. II we use some ideas from Pressley and Segal[4] to give an alternative picture of the group extension for maps from the disk into G defined by Mickelsson in Ref. 1. Some ideas from geometric quantization figure here.[5]

In Sec. III we introduce the spin bundle over the Riemann sphere and examine Quillen's determinant line bundle over the space of Cauchy–Riemann operators on this bundle. We split the sphere into the upper and lower hemispheres, identifying the lower hemisphere with the disk and the equator with the circle S^1. We consider the space of Cauchy–Riemann operators which "live" on the disk and using some results of G. Segal and Wilson[6] show that there is a natural map from this space of Cauchy–Riemann operators into the Grassmannian of subspaces of $L^2(S^1)$ that are "close" to the Hardy space H_+ in a precise sense. The map essentially assigns to each Cauchy–Riemann operator the subspace of boundary values of sections over the disk, which are holomorphic with respect to the associated complex structure. We show that det* bundle over the Grassmannian pulls back to Quillen's determinant bundle under this map. This is an expression of the correspondence between "path integrals" and an operator formalism on the boundary that has attracted much attention in recent work on conformal field theories.[7]

In Sec. IV we use some results of Quillen[8] to calculate the gauge anomaly for the action of a subgroup of the gauge group acting on the spin bundle in the holomorphic trivialization introduced by Quillen. We arrive at the Wess–Zumino term, whose appearance is understood from more sophisticated topological considerations.[2] This calculation is rather similar to the calculation of gauge anomalies on the sphere that can be found in Kupianen and Mickelsson,[9] although they do not explicitly consider Quillen's holomorphic trivialization.

Finally, in Sec. V we show that Mickelsson's group extension $\widehat{DG}$ acts on a line bundle over a base, which can be identified with the contractible space DG/G. This line bundle intersects a piece of Quillen's determinant bundle over the base D_0G of maps from D to G, which are the identity on the boundary S^1. Over this intersection the gauge action of D_0G in Quillen's determinant bundle is shown to agree with the action of a subgroup in $\widehat{DG}$ (which acts naturally in the det* bundle). We use the gauge anomaly result for Quillen's trivialization to show that there is a natural trivialization of the pull back bundle $\det^* \to DG/G$, which agrees with Quillen's trivialization over the cosets $D_0G \cdot G$. Relative to this common trivialization, the gauge anomaly for Quillen's trivialization of det can be identified with the group cocycle for $\widehat{D_0G}$. This is the principal result of this paper and constitutes our account of the surprising relation between these two notions.

Not all of the work in the first three sections is essential for the final result. However, we believe that most of the material in the first three sections is of independent interest as a simple concrete example of the developments in Refs. 4, 6, and 8.

II. TWO GROUP EXTENSIONS

Let D denote the closed unit disk $\{z \in \mathbf{C}: |z| \leqslant 1\}$ in $\mathbf{C}$ and write $S^1 = \partial D$ for the boundary of D. Let G denote the group SU(n), of unitary maps on $\mathbf{C}^n$ with determinant 1. Let DG denote the group of C^∞ maps from D into G with the group operation being pointwise multiplication (each element in DG is the restriction of a smooth map from a neighborhood of D in $\mathbf{C}$ into G). Let LG denote the group of C^∞ maps from S^1 into G under pointwise multiplication. There is a natural homomorphism b: $DG \to LG$, which sends an element $\phi \in DG$ to $b\phi = \phi|_{S^1}$, the boundary value of the map ϕ.

In this section we will construct central extensions $\widehat{DG}$

 0022-2488/89/092181-11$02.50

and $\widehat{LG}$, which covers the homomorphism b. The central extension $\widehat{LG}$ will be the Kac–Moody extension of LG[4] and the group $\widehat{DG}$ is related to the extension defined by Mickelsson in Ref. 1 (the reader should be aware that the group referred to as "DG" in Ref. 1 has a base point condition that we have dropped here). An alternative construction of this extension can be found in Ref. 10. The cocycle in Ref. 10 for $\widehat{DG}$ matches the one we find here. The homomorphism $\hat{b}$ has a kernel that is naturally homomorphic to D_0G, the kernel of the homomorphism b. Thus as was done in Ref. 1, we may realize the Kac–Moody extension $\widehat{LG}$ as a quotient $\widehat{DG}/D_0G$.

The construction of $\widehat{LG}$ and $\widehat{DG}$ and the homomorphism $\hat{b}$ all follow directly from proposition (4.42) in Pressley and Segal.[4] Before we explain this we will review some facts about line bundles with connections that may serve as motivation and that will provide a link with later developments concerning determinant bundles. A detailed account of the material we review here may be found in Kostant[5] and some infinite dimensional generalizations in Ref. 4.

To keep the exposition free of extra explanations, we will review the situation in the finite-dimensional case, even though we have in mind an application to a special infinite-dimensional problem. Let X denote a smooth finite-dimensional, connected, simply connected manifold and suppose that π: $L\to X$ is a smooth complex line bundle over X. A section of L is a smooth map s: $X\to L$ whose value $s(x)$ at any point $x\in X$ is in the fiber of L over x. A connection ∇ on L is a way of differentiating sections of L along tangent directions in X. To each local section s of L, and each local smooth vector field v, both defined on some open set $U\subseteq X$, the connection determines a new local section $\nabla_v s$ on U (the derivative of s with respect to the vector field v). The section $\nabla_v s$ is a linear function of v and for any C^∞ function f on U we have $\nabla_{fv}s = f\nabla_v s$ and $\nabla_v(fs) = v(f)s + f\nabla_v s$. It makes sense to differentiate a section s along a curve $\gamma(t)$ in the base X, and this makes possible the notion of parallel translation in L. If the curve $[a,b]\ni t\to\gamma(t)\in X$ joins $x_a = \gamma(a)$ to $x_b = \gamma(b)$, then we can transport an element $l\in\pi^{-1}(x_a)$ to $\pi^{-1}(x_b)$ by lifting $\gamma(t)$ to a section $\hat{\gamma}(t)$ over $\gamma(t)$, which starts at l and is flat with respect to the connection along γ. That is $\nabla_{\gamma(t)}\hat{\gamma}(t) = 0$. The end point $\hat{\gamma}(b)$ is the parallel translation of l to $\pi^{-1}(x_b)$ along γ, for which we write $P_\gamma^\nabla(l)$ or just $P_\gamma(l)$ when the connection is understood.

The curvature of a connection is the infinitesimal parallel translation about a parallelogram in the base (determined by two tangent vectors u and v). For line bundles this may be identified with a closed two form ω on the base X. The global version of this is a fundamental result for line bundles. Suppose γ is a closed oriented curve in X and let σ denote an oriented surface in X whose boundary is γ. Parallel translation about γ is multiplication by a complex number $Q(\gamma)$ (the holonomy of γ) and we have

$$Q(\gamma) = e^{i\int_\sigma\omega}.$$

Since parallel translation along γ does not depend on the choice of surface element σ it follows that $\omega/2\pi$ must be an integral two form.

A smooth map ϕ: $X\to X$ of the base is said to have a lift, $\hat{\phi}$, into L if there is a smooth map $\hat{\phi}$: $L\to L$ that is linear on the fibers of L and covers the action of ϕ on the base. If $\hat{\phi}$ is an invertible lift of a map ϕ on X, then $\hat{\phi}$ acts on sections, s, of L by $\hat{\phi}^*s = \hat{\phi}^{-1}s(\phi)$ (the pullback of s by ϕ). A lift $\hat{\phi}$ is said to be flat, relative to a connection ∇ if it preserves the connection in the following sense:

$$\nabla_v(\hat{\phi}^*s) = \hat{\phi}^*\nabla_{d\phi(v)}s.$$

A smooth map ϕ of the base X will have a flat lift into L if and only if $\phi^*\omega = \omega$, that is, it preserves the curvature two-form on X. This condition may also be expressed as the invariance $Q(\phi\gamma) = Q(\gamma)$ of the holonomy function on loops under the action of ϕ on loops. Any two flat lifts of the same map on the base differ by a connection preserving cover of the identity map on X. Such a map is necessarily constant when X is connected.[5]

Suppose that Γ is a group of diffeomorphisms of X that preserves the curvature ω of a connection on L. The group $\hat{\Gamma}$ of flat lifts of elements in Γ is a central extension of Γ by $\mathbf{C}^*$: $= \mathbf{C} - \{0\}$. We can be more explicit about the group $\hat{\Gamma}$. Fix a choice of base point $x_0\in X$. Suppose that $\phi\in\Gamma$, let p denote a path in X that joins x_0 to $\phi(x_0)$, and let $u\in\mathbf{C}^*$. The triple (ϕ,p,u) determines a flat lift of ϕ in the following manner. The action of (ϕ,p,u) on the fiber L_{x_0}, of L over x_0 is given by parallel translation P_p along p from L_{x_0} to $L_{\phi(x_0)}$ followed by multiplication by u. To obtain the action of (ϕ,p,u) on the fiber L_x choose a path γ from x_0 to x and define

$$(\phi,p,u)\cdot l = uP_{\phi\cdot\gamma}P_p(P_\gamma)^{-1}\cdot l$$

for $l\in L_x$. This map does not depend on the choice of the curve γ, since a change in this curve produces holonomy changes in P_γ and $P_{\phi\cdot\gamma}$ that exactly cancel, because the holonomy function is invariant under the action of ϕ. This action on the fibers associates a flat lift of ϕ to each triple (ϕ,p,u) and every flat lift can be so realized. It is easy to check that the composition law for the triples (ϕ,p,u) is

$$(\phi_1,p_1,u_1)\cdot(\phi_2,p_2,u_2) = (\phi_1\phi_2,p_1*\phi_1p_2,u_1u_2), \tag{2.1}$$

where $p_1*\phi_1p_2$ denotes the path obtained by first following the path p_1 and then following the path ϕ_1p_2 ($*$ is the homotopy product). Two triples (ϕ_1,p_1,u_1) and (ϕ_2,p_2,u_2) determine the same map on L provided that

$$\phi_1 = \phi_2 \quad\text{and}\quad u_2 = u_1e^{i\int_{\sigma(p_1*p_2^{-1})}\omega}, \tag{2.2}$$

where $\sigma(p_1*p_2^{-1})$ is an oriented surface whose boundary is $p_1*p_2^{-1}$. In proposition (4.4.2) in Ref. 4 it is observed that if Γ is a group acting on a manifold X (possibly infinite dimensional), which leaves invariant a closed integral two-form $\omega/2\pi$ on X, then the set of triples (ϕ,p,u), as above, subject to the equivalence relation (2.2), forms a group $\hat{\Gamma}$ under the composition law (2.1). This group is a central extension of Γ by $\mathbf{C}^*$ and the map $(\phi,p,u)\to\phi$ is a homomorphism from $\hat{\Gamma}$ to Γ. If the two-form ω is real, then it is clear that we can reduce the extension to the one torus $\mathbf{T}$ by restricting u to be of absolute value 1.

In the situation of interest for us X is an infinite-dimensional Grassmannian. Let H_+ denote the Hardy space of analytic functions on the disk with boundary values in $L^2(S^1,\mathbf{C}^n)$, thought of as a subspace of $L^2(S^1,\mathbf{C}^n)$. The

group LG acts on $L^2(S^1,\mathbf{C}^n)$ by left multiplication. The orbit of H_+ under this action can be identified with the group ΩG of based loops in G (loops that start and finish at the identity in G).[4] This is because any unitary loop with a holomorphic invertible extension into the interior of the disk is necessarily constant (Schwartz' reflection produces an invertible homomorphic map on P^1). With the identification of the Grassmannian with ΩG, the action of $\phi\in LG$ on $g\in\Omega G$ is given by $g(e^{i\theta})\to\phi(e^{i\theta})g(e^{i\theta})\phi(1)^{-1}$.

The space ΩG has a closed integral two-form $\omega/2\pi$, which is invariant under the action of LG on ΩG. To write this form down suppose that $g\in\Omega G$ and that $\delta_i g$ for $i=1,2$, are tangent vectors to ΩG at g identified in the natural way with $n\times n$ matrix valued functions on S^1. Then

$$\omega_g(\delta_1 g,\delta_2 g)=\frac{1}{2\pi}\int_0^{2\pi}\mathrm{Tr}(\partial_\theta(g^{-1}\delta_1 g)g^{-1}\delta_2 g)d\theta.$$

That this form is invariant is easily checked. The integrality is proved in propositions (4.44) and (4.45) in Ref. 4. If one prefers, this integrality also follows from the independent construction of the group extension for LG (with the appropriate Lie algebra cocycle) acting on the determinant bundle over ΩG (see Chap. 7 in Ref. 4). The form ω is easily seen to be real. The set of triples (ϕ,p,u) with $\phi\in LG$, p a path in the Grassmannian connecting H_+ with ϕH_+, and $u\in C$ with $|u|=1$ is thus seen to be a group under the composition law (2.1) subject to the equivalence relation (2.2). We denote this group by $\widehat{LG}$ and note that it is clearly a central extension of LG by the one torus T.

The definition of $\widehat{DG}$ is very much the same. However, for reasons that will not be apparent until the last section we will modify the sort of "path" that appears in the definition. The boundary value map $DG\ni\phi\to b\phi=\phi|_{S^1}\in LG$ is a homomorphism and $b\phi$ acts on ΩG preserving the form ω. Thus we define DG as the set of triples (ϕ,p,u) with $\phi\in DG$, p a path in DG connecting some constant map, $g\in G$, to ϕ, and $u\in\mathbf{C}$ with $|u|=1$. If p is a path in DG we write $\bar{p}:=b\phi H_+$ for the image of this path in the Grassmannian. Note that since G is connected we can move around the initial constant map $g\in G$ without effecting the induced map $\bar{p}$. The composition law for $\widehat{DG}$ is (2.1) with the understanding that ϕ acts on the Grassmannian via $b\phi$, the paths that appear in the definition are $\bar{p}_j$ rather than p_j, and the equivalence relation is (2.2). Observe that the composition law and the equivalence relation that define $\widehat{DG}$ depend only on the induced maps $\bar{p}$. The extra information in p will only be used to simplify matters in Sec. V and can be ignored for the present. It is clear that $\widehat{DG}$ is a central extension of DG by T.

We define a homomorphism $\hat{b}:\widehat{DG}\to\widehat{LG}$ by

$$\hat{b}(\phi,p,u)=(b\phi,\bar{p},u).$$

It is trivial to see that this is well defined and a homomorphism. To connect this more explicitly with the results in Ref. 10, note that the bundle $\widehat{DG}\to DG$ has an obvious section. For $\phi\in DG$, let ϕ^r denote the "radial homotopy" given by

$$\phi^r(z)=\phi(rz),\quad 0\leqslant r\leqslant 1.$$

Associated with any $\phi\in DG$ there is a natural path p in DG joining the constant map $\phi(0)$ to ϕ. Namely, $[0,1]\ni r\to\phi^r$. The section we have in mind is $\phi\to(\phi,\phi^r,1):=\hat{\phi}$. We now multiply $\hat{\phi}_1\cdot\hat{\phi}_2$ and reexpress the result as a multiple of $\widehat{\phi_1\phi_2}$ to find

$$\hat{\phi}_1\cdot\hat{\phi}_2=c(\phi_1,\phi_2)\widehat{\phi_1\phi_2},$$

where

$$c(\phi_1,\phi_2)=e^{i\int_\sigma\omega},$$

and σ is the surface obtained by mapping the triangle $\Delta:=\{(r,s)\,|\,0\leqslant s\leqslant r\leqslant 1\}$ to the Grassmannian by

$$\Delta\ni(r,s)\to b\phi_1^r\phi_2^s H_+.$$

We identify the surface $\sigma(r,s)=b\phi_1^r\phi_2^s H_+$ in the Grassmannian with

$$\sigma(r,s)=\phi_1(re^{i\theta})\phi_2(se^{i\theta})\phi_2(s)^{-1}\phi_1(r)^{-1}$$

in ΩG and we write ϕ_1 for $\phi_1(re^{i\theta})$ and ϕ_2 for $\phi_2(se^{i\theta})$. Then

$$\int_\Delta\sigma^*\omega=-\frac{1}{2\pi}\int_\Delta dr\,ds\int_0^{2\pi}d\theta\,\mathrm{Tr}[\phi_1^{-1}\partial_r\phi_1\,\phi_2\times\partial_\theta(\phi_2^{-1}\partial_s\phi_2)\phi_2^{-1}].$$

This result is precisely the cocycle found in Ref. 10. One may also calculate the kernel of the homomorphism $\hat{b}$. It is the set of triples $(\phi,p,e^{-i\int_{\sigma(\bar{p})}\omega})$ with $b\phi=$ identity, $\bar{p}$ a closed path starting and finishing at H_+, and $\sigma(\bar{p})$ is an oriented surface with boundary $\bar{p}$. Let $\sigma_r=\sigma(\hat{\phi}^r)$. It is not difficult to check that the radial homotopy $\phi^r(z)$ gives us a homomorphism:

$$\phi\to(\phi,\bar{\phi}^r,e^{-i\int_{\sigma_r}\omega})$$

from D_0G onto the kernel of $\hat{b}$.

Mickelsson made the interesting observation that the term $e^{-i\int_{\sigma_r}\omega}$ is related to the Wess–Zumino term in the gauge anomaly for chiral Fermions on S^2 (though it might be fairer to say that the cocycle property for the Wess–Zumino term inspired his definition of the group extension for DG, see also Ref. 11). He accounted for this by noting a homotopy equivalence between ΩG and the space of potentials modulo gauge equivalence. The rest of this paper will be devoted to an alternative explanation of this observation by making a connection with Quillen's determinant bundle for the space of Cauchy–Riemann operators on the spin bundle on S^2.[8]

III. CAUCHY–RIEMANN OPERATORS ON S^2

In this section we will introduce a C^∞ vector bundle on S^2 and a family of Cauchy–Riemann operators on this vector bundle. The vector bundle is chosen so that the index of the Cauchy–Riemann operators is 0 (it is the direct sum of n copies of the spin bundle on the sphere). The family of Cauchy–Riemann operators we are interested in lives in just one hemisphere (which we identify with the unit disk D). The boundary of this hemisphere we identify with the unit circle S^1. For each fixed Cauchy–Riemann operator the boundary values of holomorphic sections in the disk determine a subspace of $L^2(S^1,\mathbf{C}^n)$, which is in the Grassmannian of subspaces introduced by Segal and Wilson in Ref. 6. The principal result of this section is that under the map, which

takes a Cauchy–Riemann operator to a subspace in the Grassmannian, the det* bundle over the Grassmannian pulls back to Quillen's determinant bundle over the Cauchy-Riemann operators. This observation will allow us to directly relate the cocycle that appears in the lift of D_0G, into $\widehat{DG}$ with the gauge anomaly for Quillen's determinant bundle. We will do this in Sec. V.

The vector bundle on S^2 that we are interested in is the direct sum of n copies of a distinguished line bundle on S^2 that arises naturally from the identification of S^2 with $\mathbf{P}^1$. If we think of $\mathbf{P}^1$ as $\mathbf{C}\cup\{\infty\}$, then the maps $\mathbf{P}^1-\{\infty\}\ni z\to\{(l,lz)\colon l\in\mathbf{C}\}$ and $\mathbf{P}^1-\{0\}\ni z\to\{(l/z,l)\colon l\in\mathbf{C}\}$ give the usual identification of $\mathbf{C}\cup\{\infty\}$ with projective one space. Consider the line bundle over $\mathbf{P}^1$, which consists of pairs of points (m,v) where $m\in\mathbf{P}^1$ and v is an element of the line in $\mathbf{C}^2$ to which m maps. There are natural trivializations of this line bundle given by

$$\mathbf{P}^1-\{\infty\}\ni z\to e_0(z):=(z,(1,z))$$

and

$$\mathbf{P}^1-\{0\}\ni z\to e_\infty(z):=(z,(1/z,1)),$$

where $1/\infty=0$ in the second definition. The transition function between the two trivializations is

$$e_\infty(z)=z^{-1}e_0(z),\quad \text{for } z\in\mathbf{C}-\{0\}.$$

We denote the line bundle over S^2 defined by this transition function as E (it is the spin bundle over S^2). The bundle E has a natural Hermitian structure that arises from the fact that its fibers can be regarded as subspaces of $\mathbf{C}^2$. Thus we define

$$\mu e_0(z)\cdot\nu e_0(z):=(\mu,\mu z)\cdot(\nu,\nu z)=\bar\mu\nu(1+|z|^2)$$

and

$$\mu e_\infty(z)\cdot\nu e_\infty(z):=\bar\mu\nu(1+|z|^{-2}).$$

The vector bundle we wish to consider is the direct sum of n copies of E, which we denote by E^n. Let $e_{k,j}$ denote the jth copy of the k trivialization for E^n $(k=0,\infty)$ and write

$$\mathbf{e}_k(z):=[e_{k,1}(z),\ldots,e_{k,n}(z)].$$

Define $D=\{z\colon |z|\leqslant 1\}$ and $D_\infty=\{z\colon |z|>1\}\cup\{\infty\}$ (note that D is closed and D_∞ is open). Now choose $\epsilon>0$ and define $D_\epsilon=\{z:|z|<1+\epsilon\}$ and $D_{\infty,\epsilon}=\{z\colon |z|>1-\epsilon\}\cup\{\infty\}$. We may clearly regard $\mathbf{e}_0(z)$ and $\mathbf{e}_\infty(z)$ as trivializations of E^n over D_ϵ and $D_{\infty,\epsilon}$, respectively. If $f_k(z)$ is a $\mathbf{C}^n$-valued function then we write

$$\mathbf{e}_k(z)f_k(z):=\sum_{j=1}^n e_{k,j}(z)f_{k,j}(z)$$

for the corresponding local section of E^n.

Next we wish to introduce a family of Cauchy–Riemann operators on the bundle E^n. Let $\mathscr{A}_0$ denote the set of C^∞ $n\times n$ matrix-valued functions on the interior of the disk D with support in the closed disk D. Each "potential" $A\in\mathscr{A}_0$ determines a Cauchy–Riemann operator, $\bar\partial_A$, on the bundle E^n in the following manner:

$$\bar\partial_A\mathbf{e}_0(z)f_0(z):=d\bar z\,\mathbf{e}_0(z)(\bar\partial_z+A(z))f_0(z),$$

$$\bar\partial_A\mathbf{e}_\infty(w)f_\infty(w):=d\bar w\,\mathbf{e}_\infty(w)\bar\partial_w f_\infty(w),$$

where $w=z^{-1}$ is the natural local parameter in $D_{\infty,\epsilon}$ and in the first equation the potential A has been smoothly extended to D_ϵ by setting it equal to zero outside of D.

A Cauchy–Riemann operator maps sections of E^n to (0,1) form valued sections of E^n. A local section, s, of E^n is holomorphic with respect to $\bar\partial_A$, provided $\bar\partial_A s=0$. It is not difficult to show that it is always possible to find local trivializations holomorphic with respect to $\bar\partial_A$. Any two such trivializations differ by an $n\times n$ matrix-valued function that is holomorphic in the usual sense. Thus a Cauchy–Riemann operator determines a holomorphic structure for the bundle E^n (for more details about Cauchy–Riemann operators see Ref. 12).

For $A\in\mathscr{A}_0$, let W_A denote the subspace of $L^2(S^1,C^n)$ obtained by taking L^2 boundary values of solutions, f, to $(\bar\partial_z+A(z))f(z)=0$ for z in the interior of D. The subspace W_A is in the Grassmannian of subspaces of $L^2(S^1,C^n)$, which are close to the usual Hardy space

$$H_+:=\{f\in L^2(S^1)\colon f \text{ has an analytic continuation into the interior of } D\},$$

in a sense that we now explain. Let Gr denote the collection of subspaces, W, of $L^2(S^1,C^n)$ with the property that the orthogonal projection on W differs from the orthogonal projection on H_+ by a Schmidt class map. Let Gr_0 denote the connected component of Gr containing the subspace H_+. Then proposition (8.11.10) in Ref. 4 implies that $W_A\in\mathrm{Gr}_0$. The trivialization used in Ref. 4 is the "exterior" one rather than the "interior" one we use. Confusion will result if this difference is overlooked. We will not pause at this point to adapt proposition (8.11.10) in more detail to our circumstances, since the adaptation required will emerge naturally in the course of explaining the main result of this section, to which we now turn.

In Ref. 6, Segal and Wilson construct a holomorphic line bundle, det, over Gr_0, which is a natural extension of the notion of a determinant bundle for finite-dimensional Grassmannians (the line bundles whose fibers over a subspace is the highest exterior power of that subspace). The dual bundle, det*, is most important for us. In Ref. 8, Quillen defines a determinant line bundle over the space of Cauchy–Riemann operators on a fixed C^∞ vector bundle with a compact Riemann surface for its base. The principal result of this section is that under the map $\mathscr{A}_0\ni A\to W_A\in\mathrm{Gr}_0$ the det* bundle over Gr_0 pulls back to Quillen's determinant bundle over $\mathscr{A}_0$. This is one version of the transition between a path integral formalism in the disk D and an operator formalism on the boundary, which is much studied in the physics literature.[7]

The fiber of the determinant bundle over $A\in\mathscr{A}_0$ is $\lambda(\ker(\bar\partial_A))^*\otimes\lambda(\mathrm{coker}(\bar\partial_A))$, where $\lambda(\cdot)$ is the highest exterior power of a finite-dimensional vector space (see Ref. 8). It is useful to identify the subspaces $\ker(\bar\partial_A)$ and $\mathrm{coker}(\bar\partial_A)$ in cohomological terms. Let $\mathscr{E}_A$ denote the sheaf of germs of sections of E^n holomorphic with respect to $\bar\partial_A$. Let $\mathscr{E}_A(U)$ denote the space of holomorphic sections of E^n over the open set U. Define a map

$$\mathscr{E}_A(D_\epsilon)\oplus\mathscr{E}_A(D_\infty)\to\mathscr{E}_A(D_\epsilon\cap D_\infty)$$

by

$$h_0 \oplus h_\infty \to h_0 - h_\infty . \tag{3.1}$$

The kernel of this map is $H^0(S^2,\mathscr{E}_A)$ and may clearly be identified with $\ker(\bar\partial_A)$. The cokernel of this map is $H^1(S^2,\mathscr{E}_A)$ and may be identified with the cokernel of $\bar\partial_A$, as we now indicate. To solve $\bar\partial_A f = g$ for f, given g an L^2 section of $E^n \otimes T^{(0,1)}$ [here $T^{(p,q)}$ is the bundle of (p,q) forms on $\mathbf{P}^1$], one might try to solve the two problems:

$$\bar\partial_A f_0 = g|_{D_\epsilon} \quad \text{and} \quad \bar\partial_A F_\infty = g|_{D_\infty} \tag{3.2}$$

then piece the results together. It is always possible to solve the local problems (see Ref. 13, Theorem 13.2). If one has a pair of solutions f_0 and f_∞ to (3.2) then the difference $f_0 - f_\infty$ is defined and holomorphic in the usual sense on $D_\epsilon \cap D_\infty$ [when this difference is identified with a function via the trivialization $\mathbf{e}_0(z)$ on D_ϵ]. If $f_0 - f_\infty$ is in the image of the map (3.1) we have $f_0 - f_\infty = h_0 - h_\infty$ or

$$f_0 - h_0 = f_\infty - h_\infty \quad \text{on} \quad D_\epsilon \cap D_\infty .$$

But $\bar\partial_A(f_0 - h_0) = g - 0$ and $\bar\partial_A(f_\infty - h_\infty) = g - 0$, so that the global section defined by $f_0 - h_0$ on D_ϵ and $f_\infty - h_\infty$ on D_∞ is a solution to $\bar\partial_A f = g$. Let $R(X)$ denote the range of a map X and let R temporarily denote the range of the map (3.1). Consider the map that sends $g + R(\bar\partial_A)$ into the coset $f_0 - f_\infty + R$, where f_0 and f_∞ are any two solutions to (3.2). It is easy to check that this map is well defined and the calculation above shows that the map is injective. To see that it is surjective suppose that H is a holomorphic section of E_m over $D_\epsilon \cap D_\infty$, which has L^2 boundary values at $|z| = 1$. By subtracting from H a function that is holomorphic in the exterior of the unit disk, we can ensure that the difference has a smooth extension into the interior of D. Thus, in considering the coset $H + R$, we can suppose that H has a smooth extension into the interior of D. But $\bar\partial_A H = 0$ outside of D and so $\bar\partial_A H$ has a smooth extension to a global (0,1) form valued section g of E^n. Now choose $f_0 = H$ and $f_\infty = 0$ to see that the coset for H is in the image of the map from $\operatorname{coker}(\bar\partial_A)$ to $H^1(S^2,\mathscr{E}_A)$ defined above. The isomorphism between $\operatorname{coker}(\bar\partial_A)$ and $H^1(S^2,\mathscr{E}_A)$ is known as Dolbeault's theorem (see Theorem 15.14 in Ref. 13 for a more precise account).

Next we sketch the connection with the Grassmannian in $L^2(S^1)$. It is useful, at this point, to think in terms of the covering of S^2 given by D_ϵ and $D_{\infty,\epsilon}$ (we want to let $\epsilon \to 0$ to get $D_\epsilon \cap D_{\infty,\epsilon} \to S^1$). The map that takes $h_0 \oplus h_\infty \in \mathscr{E}_A(D_\epsilon) \oplus \mathscr{E}_A(D_{\infty,\epsilon})$ to $h_0 - h_\infty$ formally becomes the map

$$W_A \oplus H_\infty \ni h_0 \oplus h_\infty \to h_0 - h_\infty \in L^2(S^1,\mathbf{C}^n),$$

where H_∞ is the space of sections of E^n holomorphic in D_∞ with L^2 boundary values on the unit circle. To be more precise we must specify the trivialization with respect to which we are to understand $h_0 - h_\infty$ as a function on S^1. The trivialization we choose is the interior trivialization $\mathbf{e}_0(z)$ over D_ϵ [this is not the trivialization considered in proposition (8.11.10) of Ref. 4, which we are otherwise following at this point]. We now identify H_∞ in this trivialization. A section f in E^n is holomorphic over D_∞, provided that the vector-valued function f_∞ defined by $f = \mathbf{e}_\infty(z) f_\infty(z)$ is holomorphic in D_∞. To have boundary values in L^2, we must have

$$f_\infty(z) = \sum_{n=0}^{\infty} f_{\infty,n} z^{-n}$$

where $f_{\infty,n}$ is a square summable function of n. The coordinates of this same section in the $\mathbf{e}_0(z)$ trivialization are given by $f_0(z) = z^{-1} f_\infty(z)$. Thus we see that the L^2 boundary values of H_∞ are identified in the $\mathbf{e}_0(z)$ trivialization with the Hardy space $H_-(S^1,\mathbf{C}^n) =$ the set of L^2 functions on the circle with analytic continuations into the exterior of the disk that vanish at ∞.

We have identified the kernel and cokernel of the map:

$$W_A \oplus H_-(S^1,\mathbf{C}^n) \ni h_0 \oplus h_\infty \to h_0 - h_\infty \in L^2(S^1,\mathbf{C}^n), \tag{3.3}$$

with the kernel and cokernel of $\bar\partial_A$. The kernel of (3.3) is $W_A \cap H_-$, which is also the kernel of the orthogonal projection of W_A on H_+. The cokernel of (3.3) may be identified with the cokernel of the very same projection $P_+ : W_A \to H_+$, as we now indicate. Let R denote the range of (3.3) and suppose that $g \in L^2(S^1)$. Consider the map that takes $g + R$ to $P_+ g + P_+ W_A$. This map is well defined since two representatives g_1 and g_2 of the same class differ by $w - v$ ($w \in W_A$ and $v \in H_-$), which is zero in the coset class on the right [$P_+(w - v) = P_+ w \in P_+ W_A$]. The map is injective since $P_+ h + P_+ W_A = 0$ implies that $P_+ h \in P_+ W_A$, which in turn implies that for some $w \in W_A$ we have $P_+ h = P_+ w$. But then

$$h = P_+ h + P_- h$$
$$= w - P_- w + P_- h \in W_A + H_- = R.$$

The map is surjective since if h is any element in H_+ we have

$$h + R \to h + P_+ W_A .$$

Thus we have identified the kernel and cokernel of $\bar\partial_A$ with the kernel and cokernel of the orthogonal projection $P_+ : W_A \to H_+$.

The kernel and cokernel of the projection $P_+ : W_A \to H_+$ are the data for the determinant bundle over Gr_0 as we now explain (following Segal and Wilson[6]). The first thing to note is that because we are working on E^n the index of $\bar\partial_A$ is 0 [Riemann–Roch gives the index as $n(1 - g)$ + degree, with $g = 0$ and the degree $= n \times (-1)$ for n copies of the spin bundle]. This implies that the index of the orthogonal projection of W_A on H_+ is 0 which in turn means that W_A is in the connected component of the Grassmannian containing H_+ (that is, Gr_0).[4] Recall from Ref. 6 that the information needed to define an element in the fiber of the det* bundle over $W \in \mathrm{Gr}_0$ is an isomorphism $w : H_+ \to W$, such that $P_+ w$ is a trace class perturbation of the identity on H_+. Such an isomorphism is called an admissible basis for W. If one knows $\ker(P_+|_W)$ and $\operatorname{coker}(P_+|_W)$ it is easy to construct such maps. Let F denote the orthogonal complement of $P_+ W$ in H_+ and let $i : \ker(P_+|_W) \to \operatorname{coker}(P_+|_W)$ denote any isomorphism of these two finite-dimensional vector spaces. There is a natural isomorphism of F with the cokernel of $P_+|_W$ given by $F \ni x \to x + P_+ W$. Compose i with this isomorphism to get a map (which we still call i) $i : \ker(P_+|_W) \to F$. Now extend i to the rest of W by making it 0 on the orthogonal complement of $\ker(P_+|_W)$. Keep the notation i for this extension and define

$$w_i := (P_+ + i)^{-1}.$$

The map $P_+ + i$ is invertible since it is Fredholm with index 0 and has no kernel. One may easily check that $P_+ w_i$ is a finite rank perturbation of the identity and so w_i is suitable as an admissible basis for W. The fiber in the det* bundle over $W \in \mathrm{Gr}_0$ is a set of equivalence classes of pairs (w,μ) where w is an admissible basis for W and μ is an element of $\mathbf{C}$. A pair (w_1,μ_1) is equivalent to (w_2,μ_2) if and only if

$$\mu_2 = \mu_1 \det(w_1^{-1} w_2).$$

What we will show next is that the fiber over W in the det* bundle may be naturally identified with the vector space

$$\lambda(\ker(P_+|_W))^* \otimes \lambda(\mathrm{coker}(P_+|_W)),$$

where $\lambda(\cdot)$ denotes the highest exterior power. Let w_i and w_j denote two admissible bases arising from isomorphisms i and j of the kernel and cokernel of $P_+|_W$. Then

$$\begin{aligned}\det(w_i^{-1} w_j) &= \det((P_+ + i)(P_+ + j)^{-1}) \\ &= \det(I + (i-j)(P_+ + j)^{-1}).\end{aligned}$$

Since the range of i and the range of j are both contained in F, the operator whose determinant appears above has the matrix $\begin{pmatrix} I & * \\ 0 & A \end{pmatrix}$, relative to the decomposition $F^{\perp} \oplus F$. Thus $\det(w_i^{-1} w_j) = \det P_F(I + (i-j)(P_+ + j)^{-1}) P_F$, where P_F denotes the orthogonal projection on F. But $(P_+ + j)^{-1} P_F = j^{-1} P_F$, so that finally

$$\det(w_i^{-1} w_j) = \det(ij^{-1}).$$

Now we define a map from the fiber of det* at W to

$$\lambda(\ker(P_+|_W))^* \otimes \lambda(\mathrm{coker}(P_+|_W)),$$

which is a natural isomorphism. Choose an isomorphism i of $\ker(P_+|_W)$ with $\mathrm{coker}(P_+|_W)$ as above. By taking highest exterior powers this defines a map $\lambda(i)$ from $\lambda(\ker(P_+|_W))$ to $\lambda(\mathrm{coker}(P_+|_W))$. Thus we may regard $\lambda(i)$ as an element of

$$\lambda(\ker(P_+|_W))^* \otimes \lambda(\mathrm{coker}(P_+|_W)).$$

With this understood define a map

$$(w_i,\mu) \to \mu\lambda(i).$$

The transition function (2.3) above shows that this is a well-defined identification of the fiber in det* over $W \in \mathrm{Gr}_0$ with $\lambda(\ker(P_+|_W))^* \otimes \lambda(\mathrm{coker}(P_+|_W))$. We have concentrated on the description of the fibers over the noninvertible Cauchy–Riemann operators in $\mathscr{A}_0$. The fiber over the invertible elements in $\mathscr{A}$ is canonically identified with $\mathbf{C}$ (via the canonical section of the determinant bundle[8]). This is also true in the det* bundle. The fibers in the det* bundle over the subspaces in the Grassmannian that are transverse to H_- are canonically identified with $\mathbf{C}$ via the canonical section of this bundle.[4]

Return now to the map $\mathscr{A}_0 \ni A \to W_A$. The fiber in the det* bundle over $W_A \in \mathrm{Gr}_0$ is naturally identified with $\lambda(\ker(P_+|W_A))^* \otimes \lambda(\mathrm{coker}(P_+|W_A))$ (or $\mathbf{C}$ when W_A is transverse to H_-); however, this is the same space as $\lambda(\ker(\bar\partial_A))^* \otimes \lambda(\mathrm{coker}(\bar\partial_A))$ (or $\mathbf{C}$ when the Cauchy–Riemann operator associated with A is invertible), which is the fiber in Quillen's determinant bundle over the Cauchy–Riemann operators [since $\ker(\bar\partial_A)$ is isomorphic to $\ker(P_+|W_A)$ and $\mathrm{coker}(\bar\partial_A)$ is isomorphic to $\mathrm{coker}(P_+|W_A)$ as shown above]. Observe that it is clear from our description that the canonical section on det* pulls back to the canonical section on det under the map: $A \to W_A$ (the canonical section in det* is $1 \in \mathbf{C}$ over the subspaces transverse to H_- and 0 over the nontransverse subspaces; the canonical section in det is $1 \in \mathbf{C}$ over the invertible Cauchy–Riemann operators and 0 over the noninvertible ones). This will be of use to us in Sec. V.

The space $\mathscr{A}_0$ and the Grassmannian Gr_0 have natural holomorphic structures[4] and it is not hard to show that the map $\mathscr{A}_0 \ni A \to W_A \in \mathrm{Gr}_0$ is a holomorphic map. The simple calculation we did above shows that the pullback of the det* bundle over Gr_0 may be algebraically identified with Quillen's determinant bundle over $\mathscr{A}_0$. We did not show that the holomorphic structure of the pullback bundle agrees with the holomorphic structure of Quillen's determinant bundle. We do not need this result and for simplicity we will not take up this matter here.

IV. THE GAUGE ANOMALY IN QUILLEN'S DETERMINANT BUNDLE

In this section we will examine the lift of the group $D_0 G$ of gauge transformations acting on $\mathscr{A}_0$ (thought of as Cauchy–Riemann operators) into the determinant bundle over $\mathscr{A}_0$. To be explicit we will consider the lift relative to the holomorphic trivialization introduced by Quillen.[8]

We begin by explaining the results from Ref. 8 that we need to calculate the gauge anomaly in the holomorphic trivialization. Suppose that E is a fixed C^∞ vector bundle (E^n in our case) over a compact Riemann surface M ($\mathbf{P}^1$ in our case). A Cauchy–Riemann operator $\bar\partial_A$ on E is a first-order differential operator that maps smooth sections of E into one form valued sections of E and, which, relative to some local parameter z and local frame on E, has the form

$$\bar\partial_A = d\bar z(\bar\partial_z + A(z)),$$

where $A(z)$ is a smooth matrix function. Let $\mathscr{A}$ denote the space of such operators. Suppose now that we are given an inner product on E (described in Sec. II for E^n) and a metric on M compatible with its complex structure (we take the induced metric on S^2 regarded as a submanifold of $\mathbf{R}^3$). The spaces $\Omega^{0,q}(E)$, of $(0,q)$ form valued sections of E then have inner products, which allows one to define the adjoint $\bar\partial_A^*$ and the Laplacian $\bar\partial_A^* \bar\partial_A$. Quillen shows that the zeta function determinant for $\bar\partial_A^* \bar\partial_A$ can be interpreted as a Hermitian structure on the holomorphic determinant line bundle over the space of Cauchy–Riemann operators $\mathscr{A}$. A holomorphic line bundle with Hermitian structure has a unique connection compatible with these two structures. Quillen's main result is that the curvature of the connection in the determinant line bundle associated with the Hermitian structure derived from the zeta determinant of $\bar\partial_A^* \bar\partial_A$ is the Kahler form on $\mathscr{A}$, which it has as an affine space relative to the inner product space: $\mathscr{B} := \Omega^{0,1}[\mathrm{End}(E)]$. Because of the special form of the curvature one may scale the inner product in the determinant bundle by the exponential of a quadratic form in $\mathscr{B}$ to get a metric whose associated connection has zero

curvature. A flat section of the determinant bundle relative to this flat connection then gives Quillen the desired holomorphic trivialization of the determinant bundle. The inner product on $\mathscr{B}$ may be described as follows. Given B in $\mathscr{B}$, say $B = A(z)d\bar{z}$, relative to an orthonormal framing of E, let $B^* = A(z)^* dz$ in $\Omega^{1,0}[\mathrm{End}(E)]$. Then $\mathrm{Tr}_E(B^*B)$ is a (1,1) form, which can be integrated, and we define

$$\|B\|^2 := \frac{i}{2\pi}\int_M \mathrm{Tr}_E(B^*B).$$

Let $\bar{\partial}_0$ denote a fixed Cauchy–Riemann operator in E and define

$$q(\bar{\partial}_A) = \|\bar{\partial}_A - \bar{\partial}_0\|^2.$$

The Kahler form on $\mathscr{A}$ is $\partial\bar{\partial}q$. This form does not depend on the choice of base point $\bar{\partial}_0$ but the function e^q by which one scales the metric on the det bundle does depend on $\bar{\partial}_0$.

Now suppose that the index of the Cauchy–Riemann operators on E is zero. Let $\delta(\bar{\partial}_A)$ denote the trivialization of the determinant bundle described above and let $\sigma(\bar{\partial}_A)$ denote the canonical section of det. Then there is a holomorphic function of $\bar{\partial}_A$, $\det(\bar{\partial}_A;\bar{\partial}_0)$, such that $(\bar{\partial}_A) = \det(\bar{\partial}_A;\bar{\partial}_0)\delta(\bar{\partial}_A)$. Furthermore one has[4]

$$|\det(\bar{\partial}_A;\bar{\partial}_0)|^2 = e^{\|\bar{\partial}_A - \bar{\partial}_0\|^2}\det{}_\zeta(\bar{\partial}_A^*\,\bar{\partial}_A) \tag{4.1}$$

where $\det_\zeta(\bar{\partial}_A^*\,\bar{\partial}_A) = e^{-\zeta'(0)}$ is the zeta function determinant for $\bar{\partial}_A^*\,\bar{\partial}_A$ [for $\mathrm{Re}(s) > 1$, $\zeta(s) = \Sigma\lambda^{-s}$ and the sum runs over the nonzero eigenvalues λ of $\bar{\partial}_A^*\,\bar{\partial}_A$].

We will use (4.1) and some further results in Ref. 8 to calculate the gauge anomaly for Quillen's holomorphic trivialization over $\mathscr{A}_0$. Before we do this we discuss the action of the gauge group $\Omega[\mathrm{Hom}(E)]$ on the determinant bundle. Suppose that the index of the Cauchy–Riemann operators on E is zero. There is a way of constructing the determinant bundle over $\mathscr{A}$, which makes it clear that the gauge group $\Omega[\mathrm{Hom}(E)]$ acts on det. Each Cauchy–Riemann operator $\bar{\partial}_A \in \mathscr{A}$ determines a Fredholm map with index 0 from the Sobolev space of sections of E that are square integrable together with their first derivatives to the space of square integrable (0,1) form valued sections of E. Because each Cauchy–Riemann operator $\bar{\partial}_A$ has index 0, it is possible to find a finite rank map F so that $\bar{\partial}_A + F$ is invertible. Let U_F denote the open set of $\bar{\partial}_A \in \mathscr{A}$ such that $\bar{\partial}_A + F$ is invertible. Over each such open set consider the trivial bundle $U_F \times \mathbf{C}$ with the section $\sigma_F(\bar{\partial}_A) = (\bar{\partial}_A, 1)$. Define a line bundle over $\mathscr{A}$ by relating two such sections, σ_F and σ_G, by the transition function

$$\sigma_F(\bar{\partial}_A) = \det((\bar{\partial}_A + F)^{-1}(\bar{\partial}_A + G))\sigma_G(\bar{\partial}_A).$$

The determinant in this definition makes sense since $\bar{\partial}_A + G$ is a finite rank perturbation of $\bar{\partial}_A + F$ and the cocycle condition that must be satisfied for this to define a line bundle follows from the multiplicative property of determinants. The canonical section of det is $\det(\bar{\partial}_A(\bar{\partial}_A + F)^{-1})\,\sigma_F$ over U_F. Suppose now that $g \in \Omega[\mathrm{Hom}(E)]$ is a gauge transformation. We can define a lift of the action $\bar{\partial}_A \to g\,\bar{\partial}_A g^{-1}$ into det by

$$\hat{g}\sigma_F(\bar{\partial}_A) := \sigma_{gFg^{-1}}(g\,\bar{\partial}_A g^{-1}).$$

It is not hard to check that because of the similarity invariance of the determinant this lift is well defined. It obviously covers the action of g on $\mathscr{A}$. Thus we may lift the action of the gauge group into the determinant bundle. Of special interest to us is the way in which the holomorphic trivialization $\delta(\bar{\partial}_A)$ transforms under the gauge action. In a special case we will calculate $\hat{g}\delta(g^{-1}\bar{\partial}_A g)/\delta(\bar{\partial}_A)$, the gauge anomaly for the holomorphic trivialization. The special case we consider has $\bar{\partial}_A$ in the space of Cauchy–Riemann operators associated with $\mathscr{A}_0$ on E^n and $g \in D_0 G$, thought of as a gauge transformation on all of E^n (one extends it smoothly outside of the disk, D, as the constant map to the identity).

To do this explicitly we require some further information about the Green's function for invertible Cauchy–Riemann operators on E^n. We will explain slightly more than we need for our final calculation, since the relation between the asymptotics of the Green's function and matrix factorizations is of independent interest in other contexts.[1]

Recall from Sec. III that $\mathbf{e}_0(z)$ is the trivialization of E^n over the epsilon collar of the unit disk, D_ϵ, and $\mathbf{e}_\infty(z)$ is the trivialization of E^n over D_∞ related by $\mathbf{e}_\infty(z) = z^{-1}\mathbf{e}_0(z)$ on the overlap. Suppose that $A \in \mathscr{A}_0$ and let $\bar{\partial}_A$ denote the associated Cauchy–Riemann operator on E^n, defined in Sec. III. To make use of a formula for the derivative of the zeta function given in Ref. 8 we need to know the behavior of the Green's function near the diagonal for invertible $\bar{\partial}_A$. It will be enough for our purposes to analyze the solution of $\bar{\partial}_A f = g$ when g has support in D_ϵ (it is not hard to use a partition of unity to remove this restriction in any case). Let f_0 and g_0 denote the coordinates of f and g relative to the $\mathbf{e}_0$ trivialization of E^n over D_ϵ. Then we have

$$(\bar{\partial}_z + A(z))f_0(z) = g_0(z), \quad \text{for } z \in D_\epsilon. \tag{4.2}$$

Since D_ϵ is a noncompact Riemann surface the holomorphic bundle $(E^n, \bar{\partial}_A)$ is holomorphically trivial[13] over D_ϵ. This is the same as saying that there exists a C^∞ matrix-valued function $M(z)$, defined for $z \in D_\epsilon$, such that $M(z)^{-1}(\bar{\partial}_z + A(z))\,M(z) = \bar{\partial}_z$ [the columns of M are the coordinates of a basis of holomorphic sections of $(E^n, \bar{\partial}_A)$ over D_ϵ]. The Birkhoff factorization theorem (Theorem 8.1.2 in Ref. 4) implies that the restriction of M to the unit circle S^1 has a factorization

$$M|_{S^1} = M_-\lambda M_+,$$

where $M_+(z)$ has an invertible holomorphic extension into the interior of the unit disk, $M_-(z)$ has an invertible holomorphic extension into the exterior of the unit disk [which we normalize so that $M_-(\infty) = \text{identity}$], and λ is a diagonal matrix with jth diagonal entry z^{k_j} for some integer k_j. Now define $N(z) = M(z)M_+(z)^{-1}$ for $z \in D$. Then since $M_+(z)^{-1}$ is holomorphic inside D, it is clear that $N^{-1}(\bar{\partial}_z + A)N = \bar{\partial}_z$ inside D. However, $N|_{S^1} = M_-\lambda$ has a holomorphic extension to the exterior of the disk with perhaps a pole at ∞. Thus if we extend N by $M_-\lambda$ into D_ϵ and note that $A = 0$ outside D, then it follows that

$$N(z)^{-1}(\bar{\partial}_z + A(z))N(z) = \bar{\partial}_z, \quad \text{for } z \in D_\epsilon.$$

Now we introduce the trivialization $\hat{\mathbf{e}}_0 := \mathbf{e}_0 N$ on D_ϵ and $\hat{\mathbf{e}}_\infty := \mathbf{e}_\infty M_-$ on D_∞. The Cauchy–Riemann operator $\bar{\partial}_A$ acts on coordinates in the $\hat{\mathbf{e}}_0$ frame by $d\bar{z}\,\bar{\partial}_z$ and in the $\hat{\mathbf{e}}_\infty$

frame by $d\bar{w}\,\bar{\partial}_w$, where $w = z^{-1}$. The two frames are related by

$$\hat{\mathbf{e}}_0(z) = \hat{\mathbf{e}}_\infty(z) z\lambda(z). \tag{4.3}$$

Suppose now that for some j we have $k_j < 0$. Let $\hat{e}_{pj}$ denote the jth section in the row vector $\hat{\mathbf{e}}_p$. Then the section that is $\hat{e}_{0j}$ on D_ϵ and $z^{k_j+1}\hat{e}_{\infty j}$ on D_∞ is an element of the kernel of $\bar{\partial}_A$. Thus if $\bar{\partial}_A$ is invertible no k_j can be negative. The degree of the bundle E^n is $-n$, but by (4.3) this must be equal to $-n-\Sigma k_j$. Thus since no k_j is negative $\Sigma k_j = 0$ implies that all the integers k_j are 0 and we have $\lambda(z) =$ identity.

We now return to the solution of (4.2). We find

$$N^{-1}(\bar{\partial}_z + A) f_0 = N^{-1} g_0,$$

$$\bar{\partial}_z (N^{-1} f_0) = N^{-1} g_0,$$

$$N(z)^{-1} f_0(z) = \frac{i}{2\pi}\int_{D_\epsilon} \frac{N(u)^{-1} g_0(u)}{u-z}\, d\bar{u} \wedge du,$$

or

$$f_0(z) = \frac{i}{2\pi}\int_{D_\epsilon} \frac{N(z)N(u)^{-1}}{u-z} g_0(u) d\bar{u} \wedge du, \tag{4.4}$$

where we have used a standard Green's function for $\bar{\partial}_z$ (see Ref. 13). What is interesting is that this formula is the correct formula for $\bar{\partial}_A^{-1} g$ (or more properly the coordinates of this section over D_ϵ). To see this it is enough to check that $zf_0(z)$ extends to a holomorphic function in the exterior of the disk. Since $N(z)$ and the integral in (4.4) are holomorphic for z outside D_ϵ and $N(z) = I + O(z^{-1})$ near ∞, it follows that $f_0(z) = O(z^{-1})$, so that $zf_0(z)$ is indeed holomorphic in the complement of D_ϵ.

We are now in a position to calculate the asymptotics of the Green's function for $\bar{\partial}_A$ near the diagonal. To make use of Quillen's results we need the Green's function in an orthonormal frame for E^n. The frame $\mathbf{e}'(z) := \mathbf{e}(z)/\sqrt{1+|z|^2}$ is an orthonormal frame for E^n with respect to the Hermitian structure defined in Sec. III. Define

$$n(z) = \sqrt{1+|z|^2} N(z), \quad \text{for } z \in D_\epsilon.$$

Then the Green's function for $\bar{\partial}_A$, relative to the orthonormal frame $\mathbf{e}'(z)$, is

$$G(z,u) := \frac{i}{2\pi}\frac{du}{z-u} n(z) n(u)^{-1}$$

in the normalization used by Quillen.[8] In the notation used by Quillen, the relevant asymptotics near the diagonal are

$$G(z,u) = \frac{i}{2\pi}\frac{du}{z-u}\{I + (z-u)\beta(u) - (\bar{z}-\bar{u})\alpha(u) + O(|z-u|^2)\},$$

where

$$\beta(u) = (\partial_u n(u)) n(u)^{-1}$$

and

$$\alpha(u) = -(\bar{\partial}_u n(u)) n(u)^{-1}.$$

Note that $\alpha(u)$ represents the potential A in the orthonormal frame $\mathbf{e}'$. That is,

$$\alpha(u) = A(u) + [\bar{u}/(1+|u|^2)]I.$$

Define $\rho(z) := \frac{1}{2}(1+|z|^2)$.[2] Then the metric, ds^2, on S^2, which it has as a submanifold of $\mathbf{R}^3$ (in the z coordinate of stereographic projection), is given by

$$ds^2 = \rho(z)|dz|^2.$$

Finally we make one last definition before we state the result from Ref. 8 that we will use. Define $J(z)$ by

$$J(z) = \frac{i\,dz}{2\pi}\left(\beta(z) - \alpha(z)^* - \frac{1}{2}\partial_z \log\rho(z)\right).$$

In Ref. 8, J naturally arises as the difference "on the diagonal" between the Green's function for $\bar{\partial}_A$ and a geometrically defined parametrix for $\bar{\partial}_A$. Suppose now that $A \in \mathscr{A}_0$ is a family of potentials that depends holomorphically on the parameter w and is such that each associated Cauchy–Riemann operator is invertible. Then in Ref. 8 it is shown that

$$\partial_w \log(\det(\bar{\partial}_A^* \bar{\partial}_A)) = \int_D \mathrm{Tr}(J(z)\partial_w A(z) d\bar{z}). \tag{4.5}$$

We will now use (4.1) and (4.5) to calculate the gauge anomaly in Quillen's holomorphic trivialization. Suppose that $A \in \mathscr{A}_0$ and $g \in D_0 G$. Define

$$A\cdot g := g^{-1} A g + g^{-1}\bar{\partial}_z g,$$

so that $g^{-1}\bar{\partial}_A g = \bar{\partial}_{A\cdot g}$. We wish to calculate

$$\frac{\hat{g}\delta(\bar{\partial}_{A\cdot g})}{\delta(\bar{\partial}_A)}.$$

Write $0 \in \mathscr{A}_0$ and choose the base point $\bar{\partial}_0$ in the space of Cauchy–Riemann operators. It is not hard to check that the canonical section σ is equivariant under the action of the gauge group on the determinant bundle. Using this equivariance, one finds that when $\bar{\partial}_A$ is invertible,

$$\frac{\hat{g}\delta(\bar{\partial}_{A\cdot g})}{\delta(\bar{\partial}_A)} = \frac{\det(\bar{\partial}_A;\bar{\partial}_0)}{\det(\bar{\partial}_{A\cdot g};\bar{\partial}_0)}. \tag{4.6}$$

We will make use of this result by connecting $g \in D_0 G$ to the identity map by a "piecewise holomorphic" path in $D_0(G_C)$, the smooth maps from D into the complexification of G, which are equal to the identity on S^1. The determinant, $\det(\bar{\partial}_A,\bar{\partial}_0)$, depends holomorphically on A, so that if $g \in D_0(G_C)$ depends holomorphically on a parameter w then $\bar{\partial}_{A\cdot g}$ and consequently also $\det(\bar{\partial}_{A\cdot g};\bar{\partial}_0)$ are holomorphic functions of the parameter w. Suppose now that $g \in D_0(G_C)$ depends holomorphically on the parameter w. Then it follows from (4.1) that

$$\partial_w \log(\det(\bar{\partial}_{A\cdot g};\bar{\partial}_0)) = \partial_w \|\bar{\partial}_{A\cdot g} - \bar{\partial}_0\|^2 + \partial_w \log(\det_\zeta(\bar{\partial}_{A\cdot g}^* \bar{\partial}_{A\cdot g})). \tag{4.7}$$

The right-hand side of this last equation is rendered more explicit by Quillen's formula for the logarithmic derivative of the zeta determinant of the Laplacian (4.5).

Suppose now that $g \in D_0 G$. Then g can be connected to the constant map from D to the identity in G by a continuous path in $D_0 G$. To see this observe that a map $g \in D_0 G$ can be thought of as a map from the two sphere S^2 into G, obtained by identifying the boundary of the disk with the north pole ∞. The map so obtained is the identity at the north pole. The second homotopy group $\pi_2(G)$ is known to be trivial, so that any map from S^2 to G can be continuously deformed to the

map that is constant and equal to the identity at every point of S^2. It is easy to see that the homotopy can be chosen to respect the base point condition $g(\infty) = I$. Let $[0,1] \ni t \to g_t \in D_0 G$ denote a continuous path in $D_0 G$ with $g^0 =$ identity and $g^1 = g$. We will now deform this homotopy to a piecewise holomorphic homotopy in $D_0(G_C)$. Choose $\delta > 0$ and define $U_t(\delta)$, an open neighborhood of $t \in [0,1]$ by

$$U_t(\delta) := \{r \in [0,1] : \|g_r - g_t\| < \delta\},$$

where $\|\cdot\|$ is the supremum of the operator norm over the disk D. Suppose now that t_1 and t_2 are two elements of $U_t(\frac{1}{4})$. Then it is easy to see that g_{t_1} and g_{t_2} are close enough in norm so that the line segment

$$[t_1,t_2] \ni u \to \frac{u - t_1}{t_2 - t_1} g_{t_2} + \frac{t_2 - u}{t_2 - t_1} g_{t_1},$$

joining g_{t_1} to g_{t_2}, stays inside the space of maps into the complexification of G (the $n \times n$ complex invertible matrices). It is also clear that the elements in this line segment are equal to the identity on the boundary of D. As t ranges over $[0,1]$ the sets $U_t(\frac{1}{4})$ cover the unit interval. Let $\delta > 0$ denote a Lebesgue number for this covering (any set with diameter less than δ is completely contained in some element of this covering). Choose a partition $\{t_1,\ldots,t_n\}$ of $[0,1]$ whose norm (the maximum of the adjacent differences $|t_{j+1} - t_j|$) is less than δ. The path γ which consists of the line segments joining g_{t_j} to $g_{t_{j+1}}$ is then a piecewise holomorphic path in $D_0(G_C)$ connecting g to the identity, in the sense that along each such line segment one may extend the path from a function of u to a holomorphic function of $w := u + iv$ in a neighborhood of $[t_i,t_{i+1}]$ in $\mathbf{C}$.

Suppose now that $A \in \mathscr{A}_0$ and that the associated Cauchy–Riemann operator is invertible. Let $g \in D_0 G$ and suppose that γ is a piecewise holomorphic path in G_C joining the identity to g as above. Define

$$f(\gamma) = \frac{\det(\bar\partial_{A\cdot\gamma};\bar\partial_0)}{\det(\bar\partial_A;\bar\partial_0)}.$$

We wish to differentiate $f(\gamma)$ with respect to w along the path γ and then integrate over γ to find $f(g)$. There is, of course, more than one parameter w along γ and one ought to do the calculation along each line segment separately then add the results up in the end. To avoid burdensome notation we will not distinguish the different parameters w on γ with the understanding that the calculation of derivatives is done in the interior of the line segments, which the path γ comprises. Using (4.5) and (4.7) one finds that

$$\frac{\partial_w f(\gamma)}{f(\gamma)} = \partial_w \|\alpha\cdot\gamma\|^2 + \frac{i}{2\pi}\int_D \mathrm{Tr}((\beta_\gamma - \alpha\cdot\gamma^* - c)\partial_w A\cdot\gamma) dz\wedge d\bar z, \tag{4.8}$$

where $\beta_\gamma := \partial_z(\gamma^{-1}n)(\gamma^{-1}n)^{-1}$ and $c = \frac{1}{2}\partial_z \log\rho$ and we used the fact that $\partial_w \alpha\cdot\gamma = \partial_w A\cdot\gamma$ One immediate simplification in (4.8) is that the term $\partial_w\|\alpha\cdot\gamma\|^2$ is exactly cancelled by the term in the integral, which involves $\alpha\cdot\gamma^*$ (these are the only two terms that do not depend holomorphically on A). Now define

$$\varphi = \gamma^{-1}N$$

and recall that $A = -\bar\partial_z N N^{-1}$ so that $A\cdot\gamma = -\bar\partial_z\varphi\varphi^{-1}$. One also finds that since $n = (2\rho)^{1/2}N$ one has $\beta_\gamma = \partial_z\varphi\varphi^{-1} + \partial_z \log\rho$. Substituting these expressions in (4.8) and integrating one finds that

$$f(g) = e^F,$$

where

$$F = -\frac{i}{2\pi}\int_\gamma dw \int_D \mathrm{Tr}(\partial_z\varphi\varphi^{-1}\partial_w(\bar\partial_z\varphi\varphi^{-1})) \times dz\wedge d\bar z.$$

The term involving ρ in this calculation is zero since $\mathrm{Tr}(A\cdot g - A) = 0$. This is a consequence of the similarity invariance of the trace and the fact that $\bar\partial_z g g^{-1}$ has zero trace when g takes values in $G = \mathrm{SU}(n)$.

Suppose that as above γ is a piecewise linear approximation to a smooth homotopy $t \to g_t$. Then in the limit in which the norm of the partition $\{t_1,\ldots,t_n\}$ tends to 0, one finds that the path $\gamma(t)$ tends to g_t and the derivative $\partial_w\gamma$ tends to $\partial_t g_t$ in a fashion that is regular enough to permit the substitution of these limits in the integral for F given above. Make these substitutions and write

$$\varphi = g_t^{-1}N.$$

One finds

$$F = -\frac{i}{2\pi}\int_0^1 dt \int_D \mathrm{Tr}(\partial_z\varphi\varphi^{-1}\partial_t(\bar\partial_z\varphi\varphi^{-1})) \times dz\wedge d\bar z. \tag{4.9}$$

We have one further observation to make regarding this calculation. Suppose that U is an open connected set in $\mathbf{R}^n$ with coordinates $\{u_1,\ldots,u_n\}$ and that $U \ni u \to g(u) \in D_0 G$ is a smooth map. Suppose that $A \in \mathscr{A}_0$ and that the associated Cauchy–Riemann operator is invertible with $N(z)$ the solution to

$$N(z)^{-1}(\bar\partial_z + A(z))N(z) = \bar\partial_z,$$

which is holomorphic in the exterior of the unit disk and equal to the identity at ∞. As above write

$$\varphi = g(u)^{-1}N.$$

Now write d_u for exterior differentiation with respect to the parameters u and define the one-form $\mathbf{F}$ by

$$\mathbf{F} := -\frac{1}{2\pi}\int_D \mathrm{Tr}(\partial_z\varphi\varphi^{-1} d_u(\bar\partial_z\varphi\varphi^{-1}))dz\wedge d\bar z.$$

Suppose that u_0 and u_1 are two points in U and that γ is a smooth path in U with initial point u_0 and final point u_1. Define $A_k = A\cdot g(u_k)$ for $k = 0,1$. Then a calculation along the lines given above shows that

$$\frac{\det(\bar\partial_{A_1};\bar\partial_0)}{\det(\bar\partial_{A_0};\bar\partial_0)} = e^{i\int_\gamma \mathbf{F}}.$$

For this to make sense it is clearly necessary that the integral of $\mathbf{F}$ is independent of the path joining u_0 to u_1 (modulo 2π). We will show that $\mathbf{F}$ is locally exact by calculating the exterior derivative $d_u\mathbf{F}$. Because we wish to make use of some of the calculations in the next section we will work in a more general setting than is required for the anomaly calcu-

lation. Suppose now that ϕ: $D \times U \to \mathrm{GL}(n,C)$ is a smooth map and write $d = dz\,\partial_z + d\bar{z}\bar{\partial}_z$ for the exterior derivative with respect to $z, \bar{z}$. Making use of the identity

$$d_u \operatorname{Tr}(\partial_z\phi\,\phi^{-1}\bar{\partial}_z\phi\,\phi^{-1}) = \operatorname{Tr}(d_u(\partial_z\phi\,\phi^{-1})\bar{\partial}_z\phi\,\phi^{-1}) + \operatorname{Tr}(\partial_z\phi\,\phi^{-1}\,d_u(\bar{\partial}_z\phi\,\phi^{-1})),$$

one sees that

$$\operatorname{Tr}(\partial_z\phi\,\phi^{-1}d_u(\bar{\partial}_z\phi\,\phi^{-1}))dz\wedge d\bar{z} = -\tfrac{1}{2}\operatorname{Tr}(d\phi\,\phi^{-1}\,d_u(d\phi\,\phi^{-1})) + d_u\tfrac{1}{2}\operatorname{Tr}(\partial_z\phi\,\phi^{-1}\,\bar{\partial}_z\phi\,\phi^{-1})dz\wedge d\bar{z}, \tag{4.10}$$

where the wedge product of forms is understood in each of the terms. From this it is clear that to calculate the exterior derivative $d_u\mathbf{F}$, it suffices to compute

$$d_u\int_D \operatorname{Tr}(d\phi\,\phi^{-1}\,d_u(d\phi\,\phi^{-1})) = \int_D \operatorname{Tr}(d_u(d\phi\,\phi^{-1})d_u(d\phi\,\phi^{-1})). \tag{4.11}$$

Now observe that

$$d_u(d\phi\,\phi^{-1}) = -\phi d(\phi^{-1}\,d_u\phi)\phi^{-1}.$$

Substituting this onto the right-hand side of (4.10) and using the similarity invariance of the trace, one finds that

$$d_u\mathbf{F} = \frac{1}{4\pi}\int_D \operatorname{Tr}(d(\phi^{-1}\,d_u\phi)d(\phi^{-1}\,d_u\phi)).$$

Using Stoke's theorem we obtain the identity we desire:

$$d_u\mathbf{F} = \frac{1}{4\pi}\int_{S^1} \operatorname{Tr}(\phi^{-1}\,d_u\phi\,d(\phi^{-1}\,d_u\phi)), \tag{4.12}$$

where

$$\mathbf{F} := -\frac{1}{2\pi}\int_D \operatorname{Tr}(\partial_z\phi\,\phi^{-1}\,d_u(\bar{\partial}_z\phi\,\phi^{-1}))dz\wedge d\bar{z}.$$

For $\phi = \varphi$, given above in the anomaly calculation, the right-hand side of (4.11) vanishes, since the integrand on the right-hand side has a holomorphic extension to a neighborhood of the exterior of the unit disk in $\mathbf{P}^1$.

In the next section (4.12) will allow us to define a natural trivialization of a pullback of the det* bundle over the Grassmannian.

V. COMPATIBLE TRIVIALIZATIONS

In this section we will connect the group extensions of the first section with the gauge action analyzed in Sec. IV. Let $\mathscr{A}_D$ denote the family of C^∞ maps from D into the $n\times n$ complex matrices. We identify $A\in\mathscr{A}_D$ with the "Cauchy–Riemann" operator $\bar{\partial}_z + A$ on the disk D. If $A\in\mathscr{A}_0$ then we may extend $\bar{\partial}_z + A$ to a Cauchy–Riemann operator on the bundle E^n, as was done in Sec. III. If $A\in\mathscr{A}_D$ but A is not in $\mathscr{A}_0$, then no such natural identification is possible. An element $\phi\in DG$ acts on $\mathscr{A}_D$ by

$$\bar{\partial}_z + A \to \phi(\bar{\partial}_z + A)\phi^{-1}.$$

We are especially interested in the orbit of $\bar{\partial}_z$ under the action of DG. Since the only invertible holomorphic maps on D with boundary values in G are constants, it follows that this orbit space can be identified with DG/G. There is a natural map from DG/G into Gr_0, given by

$$DG/G \ni \phi G \to b(\phi)H_+,$$

where the boundary value map b is the same as in Sec. II. We may pull back the det* bundle from Gr_0 to DG/G via this map. To avoid introducing extra notation we will refer to this pullback as the det* bundle over DG/G. Note that the results of Sec. III show that det* is an extension of Quillen's determinant bundle over $D_0G\cdot G$ (which can be identified with the orbit $\bar{\partial}_0$ under D_0G in the space of Cauchy–Riemann operators on E^n).

Next we will show that the group $\widehat{DG}$ acts on the det* bundle over DG/G. Recall that an element (ϕ,p,u) of $\widehat{DG}$ consists of an element ϕ of DG, a path p connecting the identity e in DG to ϕ in DG, and a complex number u of absolute value 1. As before, let $\bar{p}$ denote the path $b(p)H_+$ in the Grassmannian joining H_+ to $b(\phi)H_+$. Let P denote parallel translation in the det* bundle over Gr_0, relative to the connection whose curvature is the form, ω, of Sec. II (see Ref. 4), the action of (ϕ,p,u) on an element v in the fiber of det* over $gG\in DG/G$ is given by

$$(\phi,p,u)\cdot v = uP_{b(\phi)\gamma}P_{\bar{p}}(P_\gamma)^{-1}v, \tag{5.1}$$

where γ is a path joining H_+ to $b(g)H_+$ in Gr_0 and we have freely identified the fiber of det* over gG with the fiber of det* over $b(g)H_+$. As in Sec. II the fact that ω is invariant under the action of LG implies that this action in the det* bundle does not depend on the choice of the path γ.

The space DG/G is contractible and this means that the pullback of the closed curvature form ω on Gr_0 to DG/G under the map given above is necessarily exact. In fact we may reinterpret Eq. (4.12) as an explicit expression of this fact. Suppose that, as at the end of Sec. IV the set U is an open subset of $\mathbf{R}^n$ and that ϕ: $D\times U\to G$ is a smooth map. Then (4.12) is

$$\omega = d_u\mathbf{F}, \tag{5.2}$$

where we use the notation ω for the pullback of the curvature from the det* bundle over DG/G.

A line bundle with a connection over a simply connected base that has a curvature form which is an exact differential on the base may be trivialized in a canonical fashion. Let π: $L\to X$ denote a line bundle with connection one-form $\mathbf{a}$ on L. Suppose that the curvature of this connection ω is equal to $d\alpha$ for a one-form α on the base X. The difference $\mathbf{a} - \pi^*(\alpha)$ satisfies the conditions necessary for it to be a connection one-form on the bundle L and the curvature of this connection is clearly zero. We can choose a flat section to trivialize the bundle L. It is possible to describe parallel translation with respect to the flat connection in terms of the parallel translation P with respect to the connection $\mathbf{a}$. If γ is a curve in the base X then parallel translation along γ relative to the flat connection described above is given by

$$e^{-i\int_\gamma\alpha}P_\gamma.$$

We may now apply this observation to the det* bundle over DG/G. The pullback of the connection on det* $\to \mathrm{Gr}_0$ has curvature $\omega = d\,\mathbf{F}$. Thus one may trivialize this bundle in a natural fashion, choosing $\alpha = \mathbf{F}$ [this is not perhaps as

"natural" a choice as say, the left-hand side of (4.11), but it is the choice that will make the connection with Quillen's trivialization particularly transparent]. We are especially interested in this trivialization over the coset space $D_0G\cdot G$ in DG/G. Suppose that $\phi\in D_0G$ and γ is a path in D_0G which joins the identity e to ϕ. The path $b(\gamma)H_+ = H_+$ is constant so that parallel translation in the pullback bundle along γ is trivial. If $\sigma^*(H_+)$ denotes the canonical section over H_+, then the trivialization discussed above is

$$e^{-i\int_\gamma \mathbf{F}}\sigma^*(H_+)$$

in the fiber over $\phi\cdot G$. The factor $e^{-i\int_\gamma \mathbf{F}}$ is the determinant of $\overline{\partial}_{0\cdot\phi^{-1}}$ which in turn is the same as the ratio of the canonical section of Quillen's det bundle to the holomorphic trivialization. Since the canonical section in the det bundle is identified with the canonical section in the det* bundle under the isomorphism discussed in Sec. III, it follows that we have identified a natural trivialization on the pullback bundle $\det^* \to DG/G$ which agrees with Quillen's holomorphic trivialization over the cosets $D_0G\cdot G$ where the two bundles can be identified.

Next observe that since the canonical section in det is equivariant with respect to the gauge action of D_0G and the lift $\widehat{D_0 G}$ given by (5.1) also leaves invariant the canonical section of the det* bundle over $D_0G\cdot G$, it follows that these two actions may be identified. Relative to the trivialization discussed above it follows that the cocycle for the action of $\widehat{D_0 G}$ agrees with the gauge anomaly in Quillen's trivialization. This is, finally, the promised connection between the gauge anomaly and the group cocycle described in the Introduction.

To conclude we would like to point out that the relation $d\mathbf{F} = \omega$ does not give a relationship between the curvature form and Quillen's holomorphic anomaly outside of $D_0G\cdot G$. The reason is that the form $\mathbf{F}$ represents a holomorphic anomaly only when ϕ has a holomorphic continuation into the exterior of the unit disk and ω gives the curvature form only when ϕ takes values in $G = \mathrm{SU}(n)$. The intersection of these two classes consists of functions that are constant on the boundary S^1.

ACKNOWLEDGMENTS

The second named author (JP) would like to acknowledge numerous helpful conversations with Doug Pickerell (who explained Mickelsson's construction to him) and the hospitality of the University of Adelaide where much of the work on this paper was done.

JP was supported by National Science Foundation Grant DMS-8703169.

[1] J. Mickelsson, "Kac–Moody groups and the Dirac determinant line bundle," in *Topological and Geometrical Methods in Field Theory* (World Scientific, Singapore, 1986), pp. 117–131.
[2] M. Atiyah and I. Singer, "Dirac operators coupled to vector potentials," Proc. Natl. Acad. Sci. (USA) **81**, 2597 (1984).
[3] I. Frenkel, "Beyond affine Lie algebras," *Proceedings of the International Congress of Mathematicians*, 1986 (American Math. Soc., Providence, RI, 1986), Vol. 1, pp. 821–839.
[4] A. Pressley and G. Segal, *Loop Groups* (Clarendon, Oxford, 1986).
[5] B. Kostant, "Quantization and unitary representations," in *Lecture Notes in Mathematics*, Vol. 170 (Springer, New York, 1970), pp. 87–207.
[6] G. Segal and G. Wilson, "Loop groups and equations of KdV type," Pub. Math. IHES **61**, 5 (1985).
[7] E. Witten, "Quantum field theory, Grassmannians and algebraic curves," Commun. Math. Phys. **113**, 529 (1988).
[8] D. Quillen, "Determinants of Cauchy–Riemann operators on a Riemann surface," Funct. Anal. Appl. **19**, 37 (1985).
[9] A. Kupianen and J. Mickelsson, "What is the effective action in two dimensions?," Phys. Lett. B **185**, 107 (1987).
[10] M. Murray, "Another construction of the central extension of the loop group," Commun. Math. Phys. **116**, 73 (1988).
[11] A. Carey and M. Murray, "Holonomy and the Weiss–Zumino term," Lett. Math. Phys. **12**, 323 (1986).
[12] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[13] O. Forster, *Lectures on Riemann surfaces* (Springer, New York, 1981).