SL(2,C), SU(2), and Chebyshev polynomials

Henri Bacry
Centre de Physique Théorique,[a),b)] *CNRS-Luminy-Case 907, F-13288 Marseille, Cedex 9, France*

(Received 5 March 1987; accepted for publication 6 May 1987)

When expressed in terms of the trace, the characters of SU(2) are known to be related with the Chebyshev polynomials of second kind. It is shown that those of the first kind also play a fundamental role. If $A\in SU(2)$ and $t=\mathrm{Tr}\,A$, then $f_n(t)=\mathrm{Tr}(A_n)$, $f_n(2\cos\theta)=\sin n\theta/\sin\theta$, $l_n(t)=\mathrm{Tr}(A^n)$, $l_n(2\cos\theta)=2\cos n\theta$, where A_n denotes the representative of A in the irrep of dimension n. Other polynomials related with them are of interest. They are (i) the "primordial" polynomials P_n (every f_n or l_n can be expressed in a unique way in terms of P_d, where d is a divisor of n), (ii) the "factorial" polynomials $f_n!=f_1f_2\cdots f_n$ which occur in a natural way in the representations, (iii) the g_n polynomials appearing in the generating functions of powers of f_n.

I. THE f_n POLYNOMIALS

The characteristic equation for $A\in SL(2,C)$ is

$$A^2-tA+I=0, \tag{1.1}$$

where $t=\mathrm{Tr}\,A$ and I is the unit 2×2 matrix. From (1,1) it follows that any power (positive or negative) of A is a linear combination of A and I

$$A^n=f_n(t)A-h_n(t)I, \tag{1.2}$$

where f_n and h_n only depend on t. It is a simple matter to prove that they are in fact *polynomials with integral coefficients*. By multiplying (1.2) by A, one gets in using (1.1)

$$\begin{aligned}A^{n+1}&=[f_n(t)A-h_n(t)I\,]A\\&=[tf_n(t)-h_n(t)\,]A-f_n(t)I\end{aligned}$$

and, by identification, we obtain the recurrence formula

$$\boxed{f_{n+1}(t)=tf_n(t)-f_{n-1}(t),} \tag{1.3a}$$

$$\boxed{f_0(t)=0,\quad f_1(t)=1,} \tag{1.3b}$$

together with $h_n(t)=f_{n-1}(t)$,

$$A^n=f_n(t)A-f_{n-1}(t)I. \tag{1.4}$$

Equation (1.3) could be used as a definition of the sequence $f_n(t)$. By choosing for A a diagonal matrix with entries $e^{\pm i\theta}$, we get for $n\geqslant0$,

$$f_n(2\cos\theta)=\sin n\theta/\sin\theta=U_{n-1}(\cos\theta), \tag{1.5}$$

as it can be shown by induction. Here U is the standard notation for the Chebyshev polynomial of the second kind. Our labeling is justified by the symmetry property

$$f_{-n}(t)=-f_n(t) \tag{1.6}$$

but also by the property (3.2) which will be derived later on. It is well known that the representative A_n of SU(2) in the irreducible representation of dimension n has the trace $\sin n\theta/\sin\theta$ if the eigenvalues of A are $e^{\pm i\theta}$. Therefore

$$f_n(\mathrm{tr}\,A)=\mathrm{tr}(A_n). \tag{1.7}$$

Remark 1: If we multiply (1.1) by A^{-1} we see that

$$A+A^{-1}=tI. \tag{1.8}$$

Moreover if we replace A by A^{-1} in (1.4) and subtract the equation obtained from (1.4), we get

$$(A^n-A^{-n})/(A-A^{-1})=f_n(t)I. \tag{1.9}$$

Remark 2: The choice of the letter f for denoting our set of polynomials is made to recall the well-known link between Fibonacci numbers φ_n and Chebyshev polynomials, namely,

$$\begin{aligned}&f_n(i)=i^{n-1}\varphi_n,\\&\varphi_{n+1}=\varphi_n+\varphi_{n-1},\quad \varphi_0=0,\quad \varphi_1=1.\end{aligned} \tag{1.10}$$

Remark 3: The orthogonality of the $f_n(t)$ is just the orthogonality of characters

$$\frac{1}{2\pi}\int_{-2}^{+2}(4-t^2)^{1/2}f_n(t)f_m(t)dt=\delta_{nm}. \tag{1.11}$$

Remark 4: The generating functions of the f_n's can be written

$$\Phi(t,z)=\frac{z}{\det(I-zA)}=\sum_{n=0}^{\infty}f_n(t)z^n=\frac{z}{1-tz+z^2}. \tag{1.12}$$

II. THE l_n POLYNOMIALS

The l_n's are polynomials in t defined by

$$l_n(t)=\mathrm{Tr}(A^n) \tag{2.1}$$

[compare with (1.7)]. Taking the trace of both sides of Eq. (1.4), we get

$$l_n(t)=tf_n(t)-2f_{n-1}(t)$$

and, by use of (1.3),

$$l_n(t)=f_{n+1}(t)-f_{n-1}(t). \tag{2.2}$$

It is a simple matter to deduce that the l_n's satisfy the same recurrence relation as the f_n's, namely,

$$\boxed{\begin{aligned}&l_{n+1}(t)=tl_n(t)-l_{n-1}(t),\\&\\&l_0(t)=2,\quad l_1(t)=t.\end{aligned}} \tag{2.3}$$

Instead of (1,6), we have

$$l_{-n}(t)=l_n(t). \tag{2.4}$$

Instead of (1.8), we have

$$A^n+A^{-n}=l_n(t)I, \tag{2.5}$$

[a] Université d'Aix Marseille-Faculté des Sciences de Luminy.
[b] Laboratoire Propre du CNRS-LP7061.

 0022-2488/87/102259-09$02.50 © 1987 American Institute of Physics

which can be easily proved by replacing A by A^n and t by $l_n(t)$ in Eq. (1.1).

Remark: The numbers λ_n defined by $l_n(i) = i^n\lambda_n$ are known as the Lucas numbers. They obey the same recurrence relation as the one of Fibonacci (with $\lambda_0 = 2$, $\lambda_1 = 1$).

Finally suppose that A is diagonal with entries $e^{\pm i\theta}$, we readily get from (2.5), for $n \geqslant 0$,

$$l_n(2\cos\theta) = 2\cos n\theta = 2T_n(\cos\theta) \tag{2.6}$$

or

$$l_n(t) = 2T_n(t/2), \tag{2.7}$$

which relate the l_n's to the Chebyshev polynomials of the first kind.

Remark: For a given value of t, the sequences $h_n(t)$ satisfying the recurrence relation (1.3a) form a two-dimensional vector space. The sequences $f_n(t)$ and $l_n(t)$ form a basis characterized by the fact that they are eigenvectors of the operator T transforming a sequence (h_n) into the sequence (h_{-n}). That sequence space can be given a symmetric scalar product

$$\langle g,h\rangle = \tfrac{1}{2}(2g_0h_0 - g_1h_{-1} - g_{-1}h_1). \tag{2.8}$$

We get an orthogonal (resp. pseudo-orthogonal) space if $|t| < 2$ (resp. $|t| > 2$) and an indefinite metric for $|t| = 2$. Note that $\langle g,g\rangle$ is invariant under a shift, namely,

$$\langle g,g\rangle = \det\begin{bmatrix} g_n & g_{n+1} \\ g_{n+1} & g_n \end{bmatrix} \quad \text{for any } n\in\mathbb{Z}. \tag{2.9}$$

For a diagonal matrix, the entries are $\frac{1}{2}(t \pm (t^2-4)^{1/2})$. Formulas (1.9) and (2.5) give

$$f_n(t) = (\alpha^n - \beta^n)/(\alpha - \beta), \tag{2.10}$$

$$l_n(t) = \alpha^n + \beta^n, \tag{2.11}$$

where $\alpha = \frac{1}{2}(t + (t^2-4)^{1/2})$, $\beta = \frac{1}{2}(t - (t^2-4)^{1/2}) = \alpha^{-1}$.

The sequences (α^n) and (β^n) both satisfy the recurrence relation (1.3a). They form an isotropic basis for the sequence vector space.

Other properties of the f_n and l_n: We must underline that the SL(2,C) group is a very pedagogical tool for a study of the Chebyshev polynomials. As an example the property

$$l_{mn}(t) = l_m(l_n(t)) \tag{2.12}$$

follows from

$$\mathrm{Tr}(A^{mn}) = l_m(\mathrm{Tr}(A^n)).$$

Also

$$f_n(2) = n, \quad l_n(2) = 2 \tag{2.13}$$

follow from $A = I$.

Let us mention that the expressions

$$f_n(t) = \sum_{k=0}^{n-1}\binom{2n-k-1}{k}(t-2)^{n-k-1}$$

$$= \sum_{k=0}^{n-1}\binom{2n-k-1}{k}(t+2)^{n-k-1}(-)^k$$

$$= 2^{1-2n}\sum_{k=0}^{n-1}\binom{2n}{2k+1}(t-2)^{n-k-1}(t+2)^k$$

are easily obtained by setting $A = X \pm I$ or $A = (X+I)/(X-I)$ (the same for the l_n's).

III. THE PRIMORDIAL POLYNOMIALS P_n

In the following, f_n and l_n have positive index n. We have already mentioned the relationship between the f_n and the Chebyshev polynomials,

$$U_n(t/2) = f_{n+1}(t). \tag{3.1}$$

Our labeling was clearly justified by the symmetry property (1.6), but also by the following ones:

$$f_n | f_m \quad \text{iff } n|m, \tag{3.2}$$

$$f_n \wedge f_m = f_{n\wedge m}. \tag{3.3}$$

In (3.2) the vertical bar means "is a divisor of." In (3.3) the symbol $\wedge$ means "greater common divisor" (g.c.d.). The same symbols are used both for polynomials and natural integers.

The proof of those properties is quite easy. *A matrix A is said to be of pseudo-order n* if n is the smallest integer such that A^n is a scalar matrix; it is clear that

$$A^m = \lambda I \quad \text{iff } n|m.$$

From Eq. (1.4), we get

$$f_m(t) = 0 \quad \text{iff } n|m.$$

Property (3.2) follows.

Property (3.3) can be derived from the following identity:

$$f_{ab}(t) = f_a(l_b(t))f_b(t), \tag{3.4}$$

a direct consequence of the property

$$\mathrm{Tr}(A_{ab}) = \mathrm{Tr}(A_a^b)\mathrm{Tr}(A_b) \tag{3.5}$$

obtained with the aid of identities (1.9) and (2.5). Note that A_a^b can be considered as well as $(A_a)^b$ or $(A^b)_a$ (representation property).

Let us denote by d the g.c.d. of m and n. It is clear that $f_d(t) = 0$ implies $f_m(t) = f_n(t) = 0$. Conversely suppose $f_m(t) = f_n(t) = 0$. We have $m = ad$, $n = bd$ with $a \wedge b = 1$. From (3.4) we have

$$f_a(l_d(t))f_d(t) = f_b(l_d(t))f_d(t) = 0.$$

Since $a \wedge b = 1$, f_a and f_b cannot vanish together. Therefore $f_d(t) = 0$.

As a consequence we have also the following property: *f_n and f_m are relatively prime iff n and m are relatively prime.*

An important consequence of (3.2) and (3.3) is that any polynomial f_n can be factorized in a unique way as a product of prime factors,

$$\boxed{f_n(t) = \prod_{d|n} P_d(t),} \tag{3.6}$$

where the product is taken for all divisors of n. The P_n are characterized by the properties

(i) P_n only divides the f_m such that $n|m$,

(ii) $P_n = f_n$ iff n is prime,

(iii) $P_n \wedge P_m = 1$ if $n \neq m$.

Before giving a rigorous definition of the P_n's (hereafter called primordial polynomials), let us examine some examples.

Example 1: We know that $f_2 | f_4$. Since $f_2 = P_2$ (2 is prime) we have $f_4 = P_2 P_4$.

Example 2: $f_2 | f_6$ and $f_3 | f_6$. Therefore there exists a unique polynomial P_6 such that $f_6 = P_2 P_3 P_6$.

Example 3: $f_{12} = P_2 P_3 P_4 P_6 P_{12}$,

$$f_{12}(t) = t(t^2-1)(t^2-2)(t^2-3)(t^4-4t^2+1).$$

In order to get a rigorous definition of the P_n's, we will use the Møbius function. It is an arithmetical function[1] defined as follows:

$\mu(1) = 1$,

$\mu(n) = 1$ (resp. -1) if n is a product of an even (resp. odd) number of different primes,

$\mu(n) = 0$ otherwise.

Formula (3.6) can be written

$$\log f_n(t) = \sum_{d|n} \log P_n(t).$$

By making use of the inverse Møbius formula,[1] we get

$$\log P_n(t) = \sum_{d|n} \mu\left(\frac{n}{d}\right) \log f_d(t)$$

or

$$\boxed{P_n(t) = \prod_{d|n} [f_d(t)]^{\mu(n/d)}.} \tag{3.7}$$

Consequence: A matrix A is of pseudo-order n if and only if $P_n(\operatorname{tr} A) = 0$.

Proof: $P_n(t) = 0$ implies $f_n(t) = 0$. The pseudo-order must be a divisor of n. Since $f_n(t) = \Pi_{d|n} P_d(t)$ and the P_d's are relatively prime, not other $P_d(t)$ can vanish. The pseudo-order is n.

The l_n's also can be factorized in primordial polynomials. The proof is based on the relation

$$f_{2n}(t) = f_n(t) l_n(t), \tag{3.8}$$

which is a direct consequence of (3.4) for $a = 2$ and of the value of $f_2(t) = t$. Now, any n can be written in a unique way as $n = 2^\alpha m$ where m is odd and $\alpha \geqslant 0$. We have

$$f_{2^{\alpha+1}m} = \prod_{d|m} P_d P_{2d} P_{4d} \cdots P_{2^{\alpha+1}d},$$

$$f_{2^\alpha m} = \prod_{d|m} P_d P_{2d} \cdots P_{2^\alpha d}.$$

From (3.8) it follows that

$$\boxed{l_{2^\alpha m}(t) = \prod_{d|m} P_{2^{\alpha+1}d}(t) \quad (m \text{ odd}).} \tag{3.9}$$

The inverse formula is given by

$$\boxed{P_{2^{\alpha+1}m}(t) = \prod_{d|m} [l_{2^\alpha d}(t)]^{\mu(n/d)}} \tag{3.10}$$

(the proof is easy but tedious; it is left to the reader).

As particular consequences we have

$$\text{(i) } P_{2n} | l_n, \tag{3.11}$$

$$\text{(ii) } l_p(t) = P_2(t) P_{2p}(t) \quad \text{if } p \text{ is prime}, \tag{3.12}$$

$$\text{(iii) } l_{2^\alpha}(t) = P_{2^{\alpha+1}}(t). \tag{3.13}$$

Some properties of the primordial polynomials (given without proof).

(i) They are polynomials with alternate integral coefficients.

$$\text{(ii) } P_n(t) = \prod_{p(n)} \left(t - 2\cos\frac{p\pi}{n}\right), \tag{3.14}$$

where $p(n)$ means $p < n$ and $p \wedge n = 1$.

It follows from (3.14) that

$$\deg P_n(t) = \varphi(n), \tag{3.15}$$

where $\varphi(n)$ is the Euler arithmetical function. From (3.6) we obtain the well-known formula

$$\sum_{d|n} \varphi(d) = n \tag{3.16}$$

[taking into account that degree $f_n = n - 1$ and $\varphi(1) = 1$].

(iii) We have

$$P_n(t) = t^{\varphi(n)} - [\varphi(n) + \mu(n)] t^{\varphi(n)-2} + \cdots. \tag{3.17}$$

(iv) As a consequence of (3.15) every polynomial P_n is of even degree except P_2: $P_2(t) = t$.

(v) For practical computations of the primordial polynomials it is convenient to use the following property:

$$\log P_n(t) = \log f_{p_1}(l_{n/p_1}(t)) - \sum_i \log f_{p_i}(l_{n/p_i}(t)) + \sum_{i,j} \log f_{p_i}(l_{n/p_i p_j}(t)) \cdots, \tag{3.18}$$

where $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots$ and $p_i, p_j, \ldots \neq p_1$. As examples, we have

$$P_{18}(t) = \frac{f_3(l_6(t))}{f_3(l_3(t))} = \frac{f_2(l_9(t))}{f_2(l_3(t))},$$

$$P_{30}(t) = \frac{f_2(l_{15}(t)) f_2(l_1(t))}{f_2(l_{10}(t)) f_2(l_6(t))} = \frac{f_3(l_{10}(t)) f_3(l_1(t))}{f_3(l_5(t)) f_3(l_2(t))}.$$

(vi) The divisibility property (3.2) can be proved on a physical problem: the polynomial $f_n(t)$ appears as a characteristic equation for a massless string on which $(n-1)$ identical massive points have been fixed at equal distances. Obviously any eigenfrequency for the n-interval problem is also an eigenfrequency for an nm-interval one. More precisely

$$f_n(t) = \det(tI - J_n), \tag{3.19}$$

where

$$J_n = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & 1 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \cdots & \cdots & 0 \quad 1 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots \quad 1 \quad 0 \end{bmatrix}. \tag{3.20}$$

Although the divisibility property is physically obvious, it is not evident on the structure of J_n.

(vii) For $t = 2$, $P_n(2)$ is related with an arithmetical function $\Lambda(n)$ known as the Von Mangoldt one[1]

$$\Lambda(n) = \operatorname{Log} P_n(2) = \begin{cases} \log p & \text{if } n = p^\alpha \ (p \text{ prime}, \alpha \geqslant 1), \\ 0 & \text{otherwise.} \end{cases}$$

In other words,

$$P_n(2) = p \quad \text{if } n = p^\alpha \ (p \text{ prime}),$$
$$P_n(2) = 1 \quad \text{otherwise.} \tag{3.21}$$

IV. THE CLEBSCH–GORDAN SERIES AND THE f_n's

Let us consider a Kronecker product of the type

$$B = A_{n_1} \otimes A_{n_2} \otimes \cdots \otimes A_{n_k}, \tag{4.1}$$

where B is the representative of the matrix $A \in SL(2,C)$ in the representation which is the direct product of representations of spins $(n_i - 1)/2$ $(i = 1,2,...,k)$. Suppose that B reduces as follows:

$$B = a_1 A_{m_1} \oplus a_2 A_{m_2} \oplus \cdots \oplus a_s A_{m_s}, \tag{4.2}$$

where a_i is the multiplicity of the representation of dimension m_i. One gets from (4.2)

$$\operatorname{Tr} B = a_1 f_{m_1}(t) + a_2 f_{m_2}(t) + \cdots + a_s f_{m_s}(t) \tag{4.3}$$

and from (4.1)

$$\operatorname{Tr} B = f_{n_1}(t) f_{n_2}(t) \cdots f_{n_k}(t). \tag{4.4}$$

We intend to give a formula permitting to compute the multiplicities a_i as functions of the n_i's.

Before going to the general case, let us study the case where B is of the form

$$B = A \otimes A \otimes A \cdots \otimes A \quad (k \text{ factors}). \tag{4.5}$$

Then

$$\operatorname{Tr} B = t^k. \tag{4.6}$$

From

$$A + A^{-1} = tI$$

we get

$$(A + A^{-1})^{k+1} = \sum_{r=0}^{k+1} \binom{k+1}{r} A^{k+1-2r} = t^{k+1} I$$

or

$$\frac{1}{2} \sum_{r=0}^{k+1} \binom{k+1}{r} l_{k+1-2r}(t) = t^{k+1}.$$

By taking the derivative of both sides and using the property (valid even for negative n)

$$l_n'(t) = n f_n(t) \tag{4.7}$$

we obtain

$$t^k = \frac{1}{2} \sum_{r=0}^{k+1} \frac{k!}{r!(k+1-r)!} (k+1-2r) f_{k+1-2r}(t)$$

or

$$t_k = \sum_{r=0}^{[(k+1)/2]} \frac{k!}{r!(k+1-r)!} (k+1-2r) \times f_{k+1-2r}(t), \tag{4.8}$$

where we have taken into account the property (1.6).

The coefficients appearing in (4.8) can be considered as entires of an infinite matrix K. They give the multiplicity of the representations appearing in the reduction of a tensor product of k spin-$\frac{1}{2}$ representations. The matrix K is reducible into an odd and an even part. It reads

$$K = \begin{bmatrix} 1 & & & & & & \\ 0 & 1 & & & & & \\ 1 & 0 & 1 & & & & \\ 0 & 2 & 0 & 1 & & & \\ 2 & 0 & 3 & 0 & 1 & & \\ 0 & 5 & 0 & 4 & 0 & 1 & \\ 5 & 0 & 9 & 0 & 5 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{4.9}$$

All entries[2] above the diagonal are zero. The entries of the first column are known as Catalan numbers.[3]

Let us now examine the most general problem; namely, the reduction of the Kronecker product (4.1). The number of times the representation of dimension n appears is given by the scalar product

$$a_{n_1 n_2 \cdots n_k n} = \frac{1}{2\pi} \int_{-2}^{2} f_{n_1}(t) f_{n_2}(t) \cdots f_{n_k}(t) f_n(t) (4 - t^2)^{1/2} \, dt. \tag{4.10}$$

Because of the symmetry of that formula we will write n_{k+1} instead of n.

Instead of computing the formula (4.10) we will derive a generating function for the a's. For such a purpose, we multiply (4.10) by $x_1^{n_1} x_2^{n_2} \cdots x_{k+1}^{n_{k+1}}$ and sum over all n_i's. We get, with the aid of (1.12),

$$F_k(x_1, x_2, ..., x_{k+1}) = \sum_{n_i=0}^{\infty} a_{n_1 n_2 \cdots n_{k+1}} x_1^{n_1} x_2^{n_2} \cdots x_{k+1}^{n_{k+1}}$$

$$= \frac{1}{2\pi} \int_{-2}^{+2} \frac{x_1 x_2 \cdots x_{k+1}}{(1 - t x_1 + x_1^2) \cdots (1 - t x_{k+1} + x_{k+1}^2)} (4 - t^2)^{1/2} \, dt$$

$$= x_1 x_2 \cdots x_{k+1} \sum_{i=1}^{k+1} x_i^k \prod_{j(\neq i)} \frac{1}{(x_i - x_j)(1 - x_i x_j)}. \tag{4.11}$$

The simplest functions F_k are

$$F_1(x_1,x_2) = x_1x_2/(1-x_1x_2),$$

$$F_2(x_1,x_2,x_3) = \frac{x_1x_2x_3}{(1-x_1x_2)(1-x_2x_3)(1-x_1x_3)},$$

$$F_3(x_1,x_2,x_3,x_4) = \frac{x_1x_2x_3x_4(1-x_1x_2x_3x_4)}{(1-x_1x_2)(1-x_1x_3)(1-x_1x_4)(1-x_2x_3)(1-x_2x_4)(1-x_3x_4)},$$

$$F_4(x_1,x_2,x_3,x_4,x_5) = \frac{\sigma_5}{\Pi_{i<j}(1-x_ix_j)}[1-\sigma_4+\sigma_5(1-\sigma_5)],$$

where σ_i are the elementary symmetric functions of degree i. Let us note that the F_k's for larger values of k have more and more complicated expressions.

V. FACTORIAL CHEBYSHEV POLYNOMIALS

In the present section, we study two kinds of polynomials which are involved in representations of SL(2,C) and SU(2). They are

$$f_n(t)! = f_1(t)f_2(t)\cdots f_n(t), \quad f_0(t)! = 1, \tag{5.1}$$

$$l_n(t)! = l_1(t)l_2(t)\cdots l_n(t), \quad l_0(t)! = 1, \tag{5.2}$$

and the $f_{\binom{n}{k}}(t)$ defined by

$$(A^{n-1}-z)(A^{n-3}-z)\cdots(A^{-n+1}-z) = \sum_{k=0}^{n} f_{\binom{n}{k}}(t)(-1)^k z^k I, \tag{5.3}$$

where $t = \mathrm{Tr}\,A$.

Proposition 5.1: We have the recurrence relation

$$2f_{\binom{n+1}{k+1}}(t) = l_{k+1}(t)f_{\binom{n}{k+1}}(t) + l_{n-k}(t)f_{\binom{n}{k}}(t). \tag{5.4}$$

Proof: First we note that for $A = I$, $t = 2$ and definition (5.3) gives $f_{\binom{n}{k}}(2) = \binom{n}{k}$ and (5.4) gives the well-known recurrence relation for binomial numbers [$l_n(2) = 2$]

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}.$$

It is clear that (5.3) is symmetric in the exchange $A \leftrightarrow A^{-1}$ which explains why the $f_{\binom{n}{k}}$'s only depend on t.

Multiply each factor of (5.3) by A and the left-hand side by $A^{-n} - Az$. It follows that the right-hand side must be multiplied by $I - A^{n+1}z$. We get

$$(A^n - Az)(A^{n-2} - Az)\cdots(A^{-n} - Az) = \sum_{k=0}^{n} f_{\binom{n}{k}}(t)(-1)^k(I - A^{n+1}z)z^k.$$

According to the definition (5.3), we have

$$\sum_{k=0}^{n+1} f_{\binom{n+1}{k}}(t)(-1)^k(Az)^k = \sum_{k=0}^{n} f_{\binom{n}{k}}(t)(-1)^k(z^k - A^{n+1}z^{k+1}).$$

Replace now z by $A^{-1}z$ and identify the terms of both sides. We get

$$f_{\binom{n+1}{k+1}}(t)I = f_{\binom{n}{k+1}}(t)A^{-k-1} + f_{\binom{n}{k}}(t)A^{n-k}.$$

Multiply by A^r and take the trace of both sides. We obtain

$$l_r(t)f_{\binom{n+1}{k+1}}(t) = l_{k-r+1}(t)f_{\binom{n}{k+1}}(t) + l_{n-k+r}(t)f_{\binom{n}{k}}(t). \tag{5.5}$$

The relation (5.4) is just a particular case of (5.5) for $r = 0$.

Proposition 5.2:

$$f_{\binom{n}{k}}(t) = f_n(t)!/f_k(t)!f_{n-k}(t)!. \tag{5.6}$$

Proof: This can be proved recurrently with the aid of (5.4). For $n = 1$, (5.3) gives

$$1 - z = \sum_{k=0}^{1} f_{\binom{1}{k}}(t)(-1)^k z^k,$$

which proves (5.6) for $n = 1$. Let us suppose that (5.6) is valid for n. The relation (5.4) gives with the aid of (5.6)

$$2f_{\binom{n+1}{k+1}}(t) = [l_{k+1}(t)f_{n-k}(t) + l_{n-k}(t)f_{k+1}(t)] \times [f_n(t)!/f_{k+1}(t)!f_{n-k}(t)!]$$

and Proposition 2 follows from the identity

$$f_{ab}(t) = f_a(t)l_b(t) + l_a(t)f_b(t) \tag{5.7}$$

[a direct consequence of (2.10) and (2.11)].

Remark 1: Other identities like

$$\sum_{k=0}^{2n+1} f_{\binom{2n+1}{k}}(t) = 2[l_n(t)!]^2,$$

$$\sum_{k=0}^{2n+1} f_{\binom{2n+1}{k}}(t)(-1)^k = 0,$$

can be proved.

Remark 2: By making $t = i$, we can discover nice properties of Fibonacci and Lucas numbers.

In the next two sections we give applications of the $f_{\binom{n}{k}}(t)$'s.

Proposition 5.3: The $f_{\binom{n}{k}}(t)$ with $k(n-k) = 0$ are polynomials with simple roots. More precisely, any $f_{\binom{n}{k}}(t)$ can be written in a unique way as a product of distinct primordial polynomials

$$f_{\binom{n}{k}}(t) = P_{m_1}(t)P_{m_2}(t)\cdots P_{m_r}(t), \quad m_i \leqslant n, \tag{5.8}$$

with the properties

$$\sum_{m_i} \varphi(m_i) = k(n-k), \tag{5.9}$$

$$\sum_{m_i} \mu(m_i) = -1, \tag{5.10}$$

$$\sum_{m_i} \Lambda(m_i) = \log \binom{n}{k}. \tag{5.11}$$

Proof: Equation (5.8) can be written

$$f_{\binom{n}{k}}(t) = \prod_{m=1}^{n} P_m(t)^{\alpha(n,k,m)}, \tag{5.12}$$

where $\alpha(n,k,m)$ must be proved to be 0 or 1 (never negative!). Euclidean divisions of n and k by m give

$$n = m[(n-\hat{n})/m] + \hat{n}, \quad 0 \leqslant \hat{n} < m,$$
$$k = m[(k-\hat{k})/m] + \hat{k}, \quad 0 \leqslant \hat{k} < m.$$

Then, the Euclidean division of $n-k$ by m will be

$$\begin{aligned} n-k &= m[[n-k-(\hat{n}-\hat{k})]/m] + \hat{n} - \hat{k} \quad \text{if } \hat{n} \geqslant \hat{k} \\ &= m[[n-\hat{k}-(\hat{n}-\hat{k})]/m - 1) \\ &\quad + \hat{n} - \hat{k} + \hat{m} \quad \text{if } \hat{n} < \hat{k}. \end{aligned}$$

It follows that in $f_{\binom{n}{k}}$, P_m will appear at the numerator $(n-\hat{n})/m$ times and at the denominator

$$\frac{k-\hat{k}}{m} + \begin{cases} [n-k-(\hat{n}-\hat{k})]/m \text{ times} & \text{if } \hat{n} \geqslant \hat{k}, \\ [n-k-(\hat{n}-\hat{k})]/m - 1 \text{ times} & \text{if } \hat{n} < \hat{k}. \end{cases}$$

Therefore it will appear once as a factor in $f_{\binom{n}{k}}$ if $\hat{n} < \hat{k}$ and not otherwise. That proves the beginning of our proposition, that is,

$$\begin{aligned} &\alpha(n,k,m) = 1 \quad \text{if } \hat{n} < \hat{k}, \\ &\alpha(n,k,m) = 0 \quad \text{otherwise.} \end{aligned} \tag{5.13}$$

Now, we know that

$$f_n(t) = t^{n-1} - (n-2)t^{n-3} + \cdots, \tag{5.14}$$

$$P_n(t) = t^{\varphi(n)} - [\mu(n) + \varphi(n)]t^{\varphi(n)-2} + \cdots. \tag{5.15}$$

From (5.14), we easily get

$$f_n(t)! = t^{n(n-1)/2} - [(n-2)(n-1)/2]t^{n(n-1)/2-2} + \cdots$$

and, if $k(n-k) \neq 0$,

$$f_{\binom{n}{k}}(t) = t^{k(n-k)}[k(n-k)-1]t^{k(n-k)-2} + \cdots.$$

But (5.15) and (5.12) give us

$$f_{\binom{n}{k}}(t) = t^N - \sum_{m=1}^{n} \alpha(n,k,m) \times [\varphi(m) + \mu(m)]t^{N-2} + \cdots,$$

where

$$N = \sum_{m=1}^{n} \alpha(n,k,m)\varphi(m).$$

It readily follows that, if $k(n-k) \neq 0$,

$$\begin{aligned} &\sum_{m=0}^{n} \alpha(n,k,m)\varphi(m) = k(n-k), \\ &\sum_{m=0}^{n} \alpha(n,k,m)[\varphi(m) + \mu(m)] = k(n-k) - 1, \end{aligned} \tag{5.16}$$

which becomes by use of (5.16),

$$\sum_{m=0}^{n} \alpha(n,k,m)\mu(m) = -1. \tag{5.17}$$

Equations (5.17) and (5.16) are alternatives for (5.9) and (5.10). The last property to prove, namely, (5.11) follows from (5.12) by replacing t by 2,

$$\log \binom{n}{k} = \sum_{m=1}^{n} \alpha(n,k,m)\Lambda(m). \tag{5.18}$$

Consequence: The $f_{\binom{n}{k}}(t)$ can be written in a "Pascal tableau" analogous to the one of binomial numbers. We get

$$\begin{array}{ccccccccccccc} & & & & & & 1 & & & & & & \\ & & & & & 1 & & 1 & & & & & \\ & & & & 1 & & P_2 & & 1 & & & & \\ & & & 1 & & P_3 & & P_3 & & 1 & & & \\ & & 1 & & P_2P_4 & & P_3P_4 & & P_2P_4 & & 1 & & \\ & 1 & & P_5 & & P_4P_5 & & P_4P_5 & & P_5 & & 1 & \\ 1 & & P_2P_3P_6 & & P_3P_5P_6 & & P_2P_4P_5P_6 & & P_3P_5P_6 & & P_2P_3P_6 & & 1 \end{array}.$$

The relations (5.16)–(5.18) permit one to check that table. For instance, $f_{\binom{7}{3}} = P_5P_6P_7$ is such that

$$3(7-3) = \varphi(5) + \varphi(6) + \varphi(7) = 4 + 2 + 6,$$
$$-1 = \mu(5) + \mu(6) + \mu(7) = -1 + 1 - 1,$$
$$\log(7 \cdot 6 \cdot 5/2 \cdot 3) = \log 5 + 0 + \log 7.$$

VI. THE GENERATING FUNCTIONS OF THE $f_n(t)^k$

We already know that [(Eq. 1.12)]

$$\sum_{n=0}^{\infty} f_n(t)z^n = \frac{z}{1 - tz + z^2}. \tag{6.1}$$

Proposition 6.1: The generating functions of the $f_n(t)^k$ is given by

$$\sum_{n=0}^{\infty} f_n(t)^k z^n = \frac{g_k(t,z)}{\sum_{l=0}^{k+1} f_{\binom{k+1}{l}}(t)(-)^l z^l}, \tag{6.2}$$

where g_k is a polynomial in t and z, of degree k in z with $g_k(t,0) = 0$.

Proof (by induction): The property is true for $k = 1$ [Eq. (6.1)]. It can be also written as follows:

$$\sum_{n=0}^{\infty} f_n(t)z^n I = \frac{z}{(A - z)(A^{-1} - z)}.$$

Equation (6.2) would follow from definition (5.3) and the relation we are going to prove recurrently:

$$\sum_{n=0}^{\infty} f_n(t)^k z^n I = \frac{g_k(t,z)}{(A^k - z)(A^{k-2} - z)\cdots(A^{-k} - z)}. \tag{6.3}$$

That relation can be written

$$g_k(t,z)I = \sum_{n=0}^{\infty} f_n(t)^k z^n (A^k - z)\cdots(A^{-k} - z).$$

Let us replace z by Az. A simple transformation leads to

$$g_k(t,Az) = \sum_{n=0}^{\infty} f_n(t)^k A^{n+k} z^n (A^{k-1} - z) \times (A^{k-3} - z)\cdots(A^{-k-1} - z)$$

and

$$(A - A^{-k}z)g_k(t,Az) = \sum_{n=0}^{\infty} f_n(t) A^n z^n (A^{k+1} - z) \times (A^{k-1} - z)\cdots(A^{-k-1} - z). \tag{6.4}$$

A similar relation can be obtained by replacing A by A^{-1} (same trace). By subtracting that relation from (6.4) and dividing by $A - A^{-1}$, one obtains

$$\frac{(A - A^{-k}z)g_k(t,Az) - (A^{-1} - A^k z)g_k(t,A^{-1}z)}{A - A^{-1}} = g_{k+1}(t,z)I. \tag{6.5}$$

We readily see that if g_k is of degree k in z, g_{k+1} is of degree $k+1$. Moreover $g_{k+1}(t,0) = 0$ and $g_{k+1}(t,z)$ is polynomial in t.

Proposition 6.2: Define

$$g_k(t,z) = \sum_{l=0}^{k-1} a_{k,l}(t) z^{l+1}, \tag{6.6}$$

where $a_{k,l}$ is a polynomial of degree $l(k-l-1)$ and

$$a_{k,l}(t) = a_{k,k-1-l}(t), \tag{6.7}$$

$$a_{k+1,l}(t) = f_{l+1}(t)a_{k,l}(t) + f_{k-l+1}(t)a_{k,l-1}(t). \tag{6.8}$$

Proof: From (6.5) and (6.6) we get

$$g_{k+1}(t,z)I = \sum_{l=0}^{k-1} a_{k,l}(t) z^{l+1} \times \frac{A^{l+1} - A^{l-k}z - A^{-l-1} + A^{k-l}z}{A - A^{-1}} = \sum_{l=0}^{k-1} a_{k,l}(t)\left[f_{l+1}(t)z^{l+1} + f_{k-l}(t)z^{l+2}\right]$$

from which (6.8) follows.

For $k = 1$, $a_{k,l}(t)$ is of degree $l(k-l-1) = 0$. Let us find the degree recurrently. From (6.8), we get

$$\deg\left[a_{k+1,l}(t)\right] = \sup\left[l(k-l-1) + l, (l-1)(k-l) + k - l\right] = l(k-l).$$

Q.E.D.

Finally (6.7) can also be proved by induction

$$a_{k+1,k-l}(t) = f_{k-l+1}(t)a_{k,k-l}(t) + f_{l+1}(t)a_{k,k-l-1}(t) = f_{k-l+1}(t)a_{k,l-1}(t) + f_{l+1}(t)a_{k,l}(t) = a_{k+1,l}(t).$$

Here are the first g_k polynomials,

$$g_1(t,z) = z,$$
$$g_2(t,z) = z + z^2,$$
$$g_3(t,z) = z + 2tz^2 + z^3,$$
$$g_4(t,z) = z + (3t^2 - 1)z + (3t^2 - 1)z^3 + z^4,$$
$$g_5(t,z) = z + (4t^3 - 3t)z^2 + (6t^4 - 8t^2 + 2)z^3 + (4t^3 - 3t)z^4 + z^5,$$
$$g_6(t,z) = z + (5t^4 - 6t^2 + 1)z^2 + (10t^6 - 25t^4 + 16t^2 - 2)z^3 + (10t^6 - 25t^4 + 16t^2 - 2)z^4 + (5t^4 - 6t^2 + 1)z^5 + z^6.$$

Remarks: (1) By taking $t = 2$, we get the generating functions of the k th powers of natural integers

$$\sum_{n=0}^{\infty} n^k z^n = \frac{g_k(2,z)}{(1-z)^{k+1}}.$$

(ii) By taking $t = i$, one would get the generating functions of φ_n^k's.

(iii) An analogous computation could be made for the k th powers of the l_n polynomials. The results are

$$\sum_{n=0}^{\infty} l_n(t)^k z^n = \frac{h_n(t,z)}{\sum_{l=0}^{k+1} f_{\binom{k+1}{l}}(t)(-1)^l z^l}, \tag{6.2'}$$

where h_k is a polynomial in t and z, of degree k in z. If we state

$$h_k(t,z) = \sum_{m=0}^{k} b_{k,m}(t) z^m,$$

where $b_{k,m}(t)$ is a polynomial of degree $m(k+1-m)$ satisfying

$$b_{k+1,m}(t) = l_{m+1}(t)b_{k,m}(t) - l_{k-m+1}(t)b_{k,m-1}(t).$$

Moreover $h_k(t,0) = 2^k$.

VII. CAYLEY–HAMILTON RELATION FOR A_n

Proposition 7.1: The Cayley–Hamilton equation for A_n is

$$\sum_{k=0}^{n} (-1)^k f_{\binom{n}{k}}(t) A_n^k = 0. \tag{7.1}$$

Proof: If A is diagonal and unitary, i.e.,

$$A = \begin{bmatrix} \exp(i\theta) & 0 \\ 0 & \exp(-i\theta) \end{bmatrix},$$

its representant A_n is

$$A_n = \begin{bmatrix} \exp(i(n-1)\theta) & & & 0 \\ & \exp(i(n-3)\theta) & & \\ & & \ddots & \\ 0 & & & \exp(-i(n-1)\theta) \end{bmatrix}.$$

If we replace z in (7.1) by one of the eigenvalues of A_n, one obtains an identity [see (5.3)]. That proves our proposition.

Consequence: Using the identity (3.4)

$$f_{ab}(t) = f_a(t) f_b(l_a(t))$$

and taking the trace of (7.1), we get

$$\sum_{k=0}^{n} (-1)^k f_{\binom{n}{k}}(t) \frac{f_{nk}(t)}{f_n(t)} = 0.$$

VIII. CONCLUSION

The present work has two possible continuations. One involves the SL(2,F) groups where F is a finite field. An elementary study of that kind of group proves that our variable t instead of $t/2$ is natural. To give an argument for that we state without proof the following property: If p is an odd prime and t an integer

$$f_{((p+1)/2)}(t) \equiv \frac{1}{2}\left[\left(\frac{t-2}{p}\right) + \left(\frac{t+2}{p}\right)\right],$$

$$f_{((p-1)/2)}(t) \equiv \frac{1}{2}\left[\left(\frac{t-2}{p}\right) - \left(\frac{t+2}{p}\right)\right],$$

where (a/p) is the Legendre symbol.[1]

Another study concerns the generalization of Chebyshev polynomials for SL(N,C). The main results in that domain have been presented in some conferences[4] and will be written in a forthcoming paper.

Notes added in proof

(1) Equations (3.9) and (3.10) can be conveniently rewritten as

$$l_n(t) = \prod_{d \| n} P_{2n/d}(t),$$

$$P_{2n}(t) = \prod_{d \| n} [l_{n/d}(t)]^{\mu(d)},$$

where $d \| n$ means d is an *odd* divisor of n.

(2) If we rewrite $\alpha(n,k,m)$ as a function of k and $l = n - k$, for a fixed value of m, α is a periodic function in k and l of period m.

ACKNOWLEDGMENTS

The author is grateful to Professor E. Grosswald and J. Zak for their interest in the beginning of the present work. This work started at the Physics Department of the Technion (Haifa) and it is a pleasure for the author to express his gratitude to the members of that department for their warm hospitality.

APPENDIX A: THE FIRST f_n POLYNOMIALS

$$f_0(t) = 0,$$
$$f_1(t) = 1,$$
$$f_2(t) = t,$$
$$f_3(t) = t^2 - 1,$$
$$f_4(t) = t^3 - 2t = t(t^2 - 2) = P_2P_4,$$
$$f_5(t) = t^4 - 3t^2 + 1,$$
$$f_6(t) = t^5 - 4t^3 + 3t = P_2P_3P_6,$$
$$f_7(t) = t^6 - 5t^4 + 6t^2 - 1,$$
$$f_8(t) = t^7 - 6t^5 + 10t^3 - 4t = P_2P_4P_8,$$
$$f_9(t) = t^8 - 7t^6 + 15t^4 - 10t^2 + 1 = P_3P_9,$$
$$f_{10}(t) = t^9 - 8t^7 + 21t^5 - 20t^3 + 5t = P_2P_5P_{10},$$
$$f_{11}(t) = t^{10} - 9t^8 + 28t^6 - 35t^4 + 15t^2 - 1,$$
$$f_{12}(t) = t^{11} - 10t^9 + 36t^7 - 56t^5 + 35t^3 - 6t = P_2P_3P_4P_6P_{12},$$
$$f_{13}(t) = t^{12} - 11t^{10} + 45t^8 - 84t^6 + 70t^4 - 21t^2 + 1,$$
$$f_{14}(t) = t^{13} - 12t^{11} + 55t^9 - 120t^7 + 126t^5 - 56t^3 + 7t = P_2P_7P_{14},$$
$$f_{15}(t) = t^{14} - 13t^{12} + 66t^{10} - 165t^8 + 210t^6 - 126t^4 + 28t^2 - 1 = P_3P_5P_{15},$$
$$f_{16}(t) = t^{15} - 14t^{13} + 78t^{11} - 220t^9 + 330t^7 - 252t^5 + 84t^3 - 8t = P_2P_4P_8P_{16},$$
$$f_{17}(t) = t^{16} - 15t^{14} + 91t^{12} - 286t^{10} + 495t^8 - 462t^6 + 210^4 - 36t^2 + 1,$$
$$f_{18}(t) = t^{17} - 16t^{15} + 105t^{13} - 364t^{11} + 715t^9 - 792t^7 + 462t^5 - 120t^3 + 9t = P_2P_3P_6P_9P_{18},$$
$$f_{19}(t) = t^{18} - 17t^{16} + 120t^{14} - 455t^{12} + 1001t^{10} - 1287t^8 + 924t^6 - 330t^4 + 45t^2 - 1,$$
$$f_{20}(t) = t^{19} - 18t^{17} + 136t^{15} - 560t^{13} + 1365t^{11} - 2002t^9 + 1716t^7 - 792t^5 + 165t^3 - 10t = P_2P_4P_5P_{10}P_{20}.$$

APPENDIX B: THE FIRST l_n POLYNOMIALS

$$l_0 = 2,$$
$$l_1 = t = P_2,$$

$$l_2 = t^2 - 2 = P_4,$$
$$l_3 = t^3 - 3t = P_2P_6,$$
$$l_4 = t^4 - 4t^2 + 2 = P_8,$$
$$l_5 = t^5 - 5t^3 + 5t = P_2P_{10},$$
$$l_6 = t^6 - 6t^4 + 9t^2 - 2 = P_4P_{12},$$
$$l_7 = t^7 - 7t^5 + 14t^3 - 7t = P_2P_{14},$$
$$l_8 = t^8 - 8t^6 + 20t^4 - 16t^2 + 2 = P_{16},$$
$$l_9 = t^9 - 9t^7 + 27t^5 - 30t^3 + 9t = P_2P_6P_{18},$$
$$l_{10} = t^{10} - 10t^8 + 35t^6 - 50t^4 + 25t^2 - 2 = P_4P_{20},$$
$$l_{11} = t^{11} - 11t^9 + 44t^7 - 77t^5 + 55t^3 - 11t = P_2P_{22},$$
$$l_{12} = t^{12} - 12t^{10} + 54t^8 - 112t^6 + 105t^4 - 36t^2 + 2 = P_8P_{24},$$
$$l_{13} = t^{13} - 13t^{11} + 65t^9 - 156t^7 + 182t^5 - 91t^3 + 13t = P_2P_{26},$$
$$l_{14} = t^{14} - 14t^{12} + 77t^{10} - 210t^8 + 294t^6 - 196t^4 + 49t^2 - 2 = P_4P_{28},$$
$$l_{15} = t^{15} - 15t^{13} + 90t^{11} - 275t^9 + 450t^7 - 378t^5 + 140t^3 - 15t = P_2P_6P_{10}P_{30},$$
$$l_{16} = t^{16} - 16t^{14} + 104t^{12} - 352t^{10} + 660t^8 - 672t^6 + 336t^4 - 64t^2 + 2 = P_{32},$$
$$l_{17} = t^{17} - 17t^{15} + 119t^{13} - 442t^{11} + 935t^9 - 1122t^7 + 714t^5 - 204t^3 + 17t = P_2P_{34},$$
$$l_{18} = t^{18} - 18t^6 + 135t^{14} - 546t^{12} + 1287t^{10} - 1782t^8 + 1386t^6 - 540t^4 + 81t^2 - 2 = P_4P_{12}P_{36},$$
$$l_{19} = t^{19} - 19t^{17} + 152t^{15} - 665t^{13} + 1729t^{11} - 2717t^9 + 2508t^7 - 1254t^5 + 285t^3 - 19t = P_2P_{38}.$$

APPENDIX C: THE FIRST PRIMORDIAL POLYNOMIALS

If p is prime, $P_p(t) = f_p(t)$,

$$P_4(t) = t^2 - 2,$$
$$P_6(t) = t^2 - 3,$$
$$P_8(t) = t^4 - 4t^2 + 2,$$
$$P_9(t) = t^6 - 6t^4 + 9t^2 - 1,$$
$$P_{10}(t) = t^4 - 5t^2 + 5,$$
$$P_{12}(t) = t^4 - 4t^2 + 1,$$
$$P_{14}(t) = t^6 - 7t^4 + 14t^2 - 7,$$
$$P_{15}(t) = t^8 - 9t^6 + 26t^4 - 24t^2 + 1,$$
$$P_{16}(t) = t^8 - 8t^6 + 20t^4 - 16t^2 + 2,$$
$$P_{18}(t) = t^6 - 6t^4 + 9t^2 - 3,$$
$$P_{20}(t) = t^8 - 8t^6 + 19t^4 - 12t^2 + 1,$$
$$P_{21}(t) = t^{12} - 13t^{10} + 64t^8 - 146t^6 + 148t^4 - 48t^2 + 1,$$
$$P_{22}(t) = t^{10} - 11t^8 + 44t^6 - 77t^4 + 55t^2 - 11,$$
$$P_{24}(t) = t^8 - 8t^6 + 20t^4 - 16t^2 + 1,$$
$$P_{25}(t) = t^{20} - 20t^{18} + 170t^{16} - 800t^{14} + 2275t^{12} - 4003t^{10} + 4280t^8 - 2605t^6 + 775t^4 - 75t^2 + 1,$$
$$P_{26}(t) = t^{12} - 13t^{10} + 65t^8 - 156t^6 + 182t^4 - 91t^2 + 13,$$
$$P_{27}(t) = t^{18} - 18t^{16} + 135t^{14} - 546t^{12} + 1287t^{10} - 1782t^8 + 1386t^6 - 540t^4 + 81t^2 - 1,$$
$$P_{28}(t) = t^{12} - 12t^{10} + 53t^8 - 104t^6 + 86t^4 - 24t^2 + 1,$$
$$P_{30}(t) = t^8 - 7t^6 + 14t^4 - 8t^2 + 1,$$
$$P_{32}(t) = t^{16} - 16t^{14} + 104t^{12} - 352t^{10} + 660t^8 - 672t^6 + 336t^4 - 64t^2 + 2.$$

[1] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers* (Clarendon, Oxford, 1965), Chaps. 16 and 17.
[2] R. Pauncz, *Spin Eigenfunctions, Construction and Use* (Plenum, New York, 1979), p. 21.
[3] M. Aigner, *Combinatorial Theory* (Springer, Berlin, 1979).
[4] H. Bacry, in *Lecture Notes in Physics*, Vol. 201 (Springer, Berlin, 1984), p. 483; *Lecture Notes in Mathematics*, Vol. 1171 (Springer, Berlin, 1985), p. 564; *Group Theoretical Methods in Physics* (Nauka, Moscow, 1986), p. 239.

Pedestrian approach to two-cocycles for unitary ray representations of Lie groups

J. Krause
Facultad de Física, Pontificia Universidad Católica de Chile, Casilla 6177, Santiago 22, Chile

(Received 28 August 1986; accepted for publication 17 June 1987)

Necessary and sufficient conditions for unitary ray representations of connected Lie groups are reexamined. Thus a systematic constructive method is obtained for calculating the admissible exponent factors (two-cocycles). The gauge freedom of the unitary ray representation formalism is also considered. This introduces the distinction between trivial and genuine ray representations. A special gauge is then adopted, within which the two-cocycle is almost unique. The only prerequisite of the exponent factor calculus is the knowledge of the binary combination rules for the essential parameters of the group. The attained method affords a simple, general, and explicit (i.e., coordinate-dependent) two-cocycle calculus. The aim of this paper is merely instrumental.

I. INTRODUCTION

This work concerns non-Abelian two-cocycle calculus, as required in many applications of the quantum theory of symmetries. Indeed, it is a well-known fact that quantum mechanics does not fix the phase of the vectors describing pure states, and one associates such states to rays rather than to vectors. Therefore, unitary *ray* representations should be used in quantum theory, in general.[1] The extension of the unitary formalism, from "true" (vector) to "projective" (ray) representations, faces no difficulties, as long as one is able to calculate the admissible exponent functions (i.e., the two-cocycles) of the corresponding group G. The current techniques for calculating two-cocycles, however, are conceptually difficult and complicated to handle, because they are usually presented within a highbrow mathematical formalism that goes beyond the standard curriculum of most physicists in these matters. In fact, these techniques seem to be reserved to those physicists who are specialists in cohomology theory and other sophisticated issues of Lie group theory.[2]

There are two (perhaps more) available methods; one is purely group-theoretic, the other has a more geometric flavor. Briefly stated, the group-theoretic method is as follows. It is known that the exponents of a Lie group G are related to the true unitary representations of a "larger" group G_k associated with G. This G_k is a central extension of the universal covering group $\tilde{G}$ of [the Lie algebra $L(G)$ of] G by the one-dimensional Abelian group R [that is, by the additive group of real numbers, which, by its turn, is the universal covering group of $U(1)$: $R \approx \tilde{U}(1)$]. This fact immediately affords a constructive method for obtaining the two-cocycles of G. The starting point of the process is the choice of an admissible extension[3] $L_k(G)$ of the Lie algebra $L(G)$ by the trivial algebra $L_1 = L(R)$. Then, once an allowable extended algebra $L_k(G)$ has been established, one can determine the group G_k (by means of the Campbell–Baker–Hausdorff formula, for instance). In this way, the two-cocycles of G can be read off by inspection of the group multiplication law of the parameters of G_k. Moreover, it has also been shown that for the determination of the associated two-cocycle of G one can use directly any faithful representation of the extended algebra $L_k(G)$.[4] Exponent factors for several Lie groups of physical interest have been calculated by means of this technique. Perhaps this is the approach to unitary ray representations of G preferred by people familiar with cohomology theory[5] of Lie groups and Lie algebras.[6]

The other method of construction of continuous (in fact, of class C^∞) exponents that figures in the literature[7] is analytic and uses the powerful coordinate-independent techniques of modern differential geometry.[8] This method was introduced by Houard in classical mechanics (in connection with the problem of determining the Lagrangians whose Euler–Lagrange equations are invariant under a given transformation group[9]). Clearly, the same technique serves the purposes of quantum theory as well. After introducing some differential forms, obtained from any given C^∞ exponent of G, in this approach one proceeds *reciprocally* to deduce a general formula giving the exponents in terms of the closed left-invariant two-forms of the group. In effect, Houard proves the following theorem: For a Lie group, any C^∞ local exponent explicitly determines a closed left-invariant two-form and, conversely, to any closed left-invariant two-form corresponds a family of C^∞ local exponents that can be explicitly calculated (see Ref. 7). In this manner, the trivial (local) exponents correspond to those two-forms that are differentials of left-invariant one-forms and, moreover, the classes of equivalent local exponents[3] on G correspond biunivocally to the classes of equivalent closed left-invariant two-forms (i.e., those that differ by the differential of a left-invariant one-form). (Let us here remark that, for *any* given Lie group, every conceivable method of construction of its two-cocycles, *in general*, produces only local exponents.[3])

These are beautiful and powerful methods indeed. Yet, it seems that there is still some room for a more *pedestrian* approach to this subject. Thus this paper addresses the problem of developing a general and elementary method for calculating local two-cocycles of a given Lie group, *once the group multiplication law of its essential parameters is known*. To our knowledge, an approach such as the elementary and explicit (i.e., parameter-dependent) constructive approach to exponent factors of Lie groups is lacking in the current literature, notwithstanding the many important features

 0022-2488/87/102268-12$02.50 © 1987 American Institute of Physics

concerning this issue that can be found in the excellent "classical" works of Wigner,[10] Bargmann,[3] and others.[4,7] Our main purpose here is to bring this matter into a tailor-made formalism, manageable enough for the needs of quantum-theoretic application. Of course, we do not claim complete originality for the mathematical contents of this paper, whose aim is purposely instrumental. Our approach to the issue of two-cocycles is analytic and, in many respects, comes very close to Houard's work,[7] as a simplified, coordinate-dependent version of it. (Though there are some novelties in our scope of the subject.)

Clearly, this is an endeavor interesting for its own sake; even more interesting and timely since two-cocycles are becoming fashionable in several areas of theoretical physics. For instance, two-cocycles have been investigated recently in connection with the Wess–Zumino–Witten anomaly,[11] as nontrivial extensions of current algebra,[12] and also in the cohomology of Wess–Zumino Lagrangians of gauge fields.[13] In a different context, a special two-cocycle of the Euclidean group E_2 has been used recently to obtain the kinematic quantum model[14] of the simple harmonic oscillator.[15] It should also be mentioned here that a suitable (well-known) two-cocycle of the Galilei group[16] yields the kinematic quantum model of a Newtonian free particle.[17] Furthermore, even three-cocycles are becoming fashionable (as they can be found, for example, in the quantum mechanics of a point particle moving in an external magnetic field that is not divergence-free[18]), and several physicists have pointed out their usefulness in the quantum mechanical description of magnetic monopoles.[19] (We would like to mention this fact here, although we do not touch on three-cocycles in the present article.) Hence, a simplified, self-contained formalism of two-cocycles, presenting its own calculational tools, may be of some help to a wide realm of physicists.

Let us outline the contents of this paper. We first examine the consequences of the various (well-known) functional relations for two-cocycles that occur in the theory of unitary ray representations of Lie groups (Sec. II). In Sec. III we obtain a set of necessary and sufficient conditions for the required functional relations that characterize an admissible two-cocycle. Therefore, a systematic constructive procedure arises (and this is the main point in the present approach). In Secs. IV and V we discuss the gauge transformations of unitary ray representations of Lie groups. In order to illustrate the main features of the method, some miscellaneous examples are also included in this paper (Sec. VI). For the sake of completeness, we add an Appendix where we present an elementary introduction to "non-Abelian calculus." We hope that his appendix will help the reader to grasp the discussions contained in the following context. Let us finally remark that in this paper we shall proceed in a straightforward formal manner, since our emphasis is on *method*. For all the important topological details we refer the reader to Bargmann's paper.[3]

II. EXPONENT FACTORS FOR UNITARY RAY REPRESENTATIONS REVISITED

We begin our work presenting the general features of the ray formalism conducive to the allowable exponent functions of an arbitrary, connected, r-parameter Lie group G. We denote by $q=(q^1,...,q^r)$ a generic point of the group manifold $M(G)$; the coordinates q^a, $a=1,...,r$, are real and form a set of r essential parameters of G. We shall write $\bar{q}=(\bar{q}^1,...,\bar{q}^r)$ (instead of q^{-1}) to denote that point in $M(G)$, uniquely associated with the point q, that labels that element of G and that is the inverse of the element corresponding to q; the point $e=(e^1,...,e^r)$ of $M(G)$ labels the unit element of G. Thus the r-dimensional manifold $M(G)$ carries an analytic mapping g: $M(G)\times M(G)\to M(G)$, which is endowed with the group properties. This mapping is realized by a set of r group-multiplication functions of the parameters, say, $g(q';q)=q''$.[20] (Indices are often suppressed when there is no danger of confusion.) In the Appendix we present a summary of useful formulas pertaining to the affine geometry of $M(G)$, which shall be needed in the sequel; in particular, some general properties of the analytic functions g^a can be recalled from the Appendix.

In order to simplify our discussion, and concentrate on the main problems of two-cocycle calculus, in this paper we assume that the coordinate patch $\{q\}$ covers a whole submanifold $N(e)\subset M(G)$, which is a neighborhood of the identity point e. Sometimes, however, when G is a noncompact, connected, and simply connected Lie group (as, for instance, the universal covering group of a noncompact Lie group), we also formally assume that the coordinate patch $\{q\}$, containing the identity point e, covers the whole manifold $M(G)$ and maintains everywhere the required one-to-one correspondence with the elements of G. (Certainly, if G is compact, this last assumption would be inaccurate.)

We now turn to the unitary ray representations of G. In quantum mechanics G establishes an isomorphism between rays that preserves all transition probabilities. Therefore, it is useful to define unitary (or antiunitary) operators rays, in analogy to the notion of vector rays. In this fashion, according to Wigner's theorem,[21] the operators of the isomorphism are representatives selected from the corresponding operators rays. Hence one infers (by well-known arguments) the ray representation property

$$U_k(q')U_k(q)=e^{i\phi_k(q';q)}U_k[g(q';q)]\,, \tag{2.1}$$

where ϕ_k is a real function of the points q' and q, and where $U_k(q')$ and $U_k(q)$ are suitable selected representative operators. (Right now, k is a label for the selection of representatives.) It is clear that for a different choice of representatives, a different function ϕ_k will appear in (2.1). We shall discuss this *gauge freedom* in Sec. IV. Since G is connected, it can be shown that the operators $U_k(q)$ are necessarily *unitary*.[3] However, as a consequence of (2.1) one gets

$$U_k^{+}(q)\equiv U_k^{-1}(q)=e^{-i\mu_k(q)}U_k(\bar{q})\,, \tag{2.2}$$

where $\mu_k(q)$ is defined by

$$\mu_k(q)=\phi_k(q;\bar{q})\equiv\phi_k(\bar{q};q) \tag{2.3}$$

(which identity can be proved rather easily). The relation stated in (2.2) holds for unitary ray representations in general, instead of

$$U_0^{+}(q)=U_0^{-1}(q)=U_0(\bar{q})\,, \tag{2.4}$$

which holds for unitary *vector* representations. Of course,

one recovers (2.4) within the ray formalism by choosing a gauge such that $\mu_k(q)\equiv 0$ (cf. Sec. V). Also, the "initial value" property of ϕ_k,

$$\phi_k(e;q)=\phi_k(q;e)=0\,, \tag{2.5}$$

for all $q\in N(e)$, need not be assumed since it can be proved quite directly. (That is, this property is *not* a gauge condition imposed on the exponent function.) Thus

$$\mu_k(e)=0 \tag{2.6}$$

follows, wherefrom

$$U_k(e)=I \tag{2.7}$$

is attained for *all* selections of representatives. Now, the associative property of the representation (2.1) yields the following well-known functional relation for the exponent function:

$$\begin{aligned}&\phi_k(q';q)+\phi_k[q'';g(q';q)]\\&\quad=\phi_k(q'';q')+\phi_k[g(q'',q');q]\,.\end{aligned} \tag{2.8}$$

This three-point relation entails the fundamental property for an admissible local exponent function, and thus it represents the backbone of two-cocycle analysis.[3] Finally, the operator identity for the Hermitian adjoint of the product of two operators gives

$$\begin{aligned}&\phi_k(q';q)+\phi_k(\bar{q};\bar{q}')\\&\quad=\mu_k(q')+\mu_k(q)-\mu_k[g(q';q)]\,.\end{aligned} \tag{2.9}$$

Altogether, the functional properties we have sketched above enhance the two-cocycle $\phi_k(q';q)$ with the group properties of the $U_k(q)$'s, at least on $N(e)\subset M(G)$.

Before proceeding to examine the technicalities of two-cocycle calculus, we wish to mention two important features. First, let us recall that, using an elegant construction of Iwasawa's,[22] one can show that every *local* exponent $\phi_k(q';q)$ is *equivalent* (cf. Sec. IV) to a local exponent $\phi_k'(q';q)$ that, on some neighborhood $N(e)$ of the identity, has continuous partial derivatives of all orders with respect to the parameters q' and q. Iwasawa's theorem is a direct consequence of the functional relation (2.8). Next, let us also recall that, for a connected *and* simply connected Lie group G, every *local* exponent $\phi_k(q';q)$, defined on some neighborhood $N(e)$ of the identity, can be extended to a *global* two-cocycle function defined on the whole group manifold $M(G)$ (cf. Theorem 5.1. in Bargmann's paper[3]). Moreover, the extended two-cocycle is differentiable everywhere on $M(G)\times M(g)$ if $\phi_k(q';q)$ is differentiable on $N(e)\times N(e)$.

III. TWO-COCYCLE CALCULUS ON THE GROUP MANIFOLD

This section deals with the general theory of exponent calculus corresponding to a connected Lie group G. Looking first for the necessary conditions of the formalism, let us define the functions

$$r_a^{(k)}(q)=D_a(q')\phi_k(q';q)\,, \tag{3.1}$$

$$l_a^{(k)}(q)=D_a(q')\phi_k(q;q')\,, \tag{3.2}$$

attached to a given phase function ϕ_k. The operators $D_a(q')$ stand for $\lim_{q'\to e}\partial_a'$ [cf. Appendix, Eq. (A8)]. These functions will be called *right* and left *exponent generators*, respectively. Clearly, they satisfy the initial conditions:

$$r_a^{(k)}(e)=l_a^{(k)}(e)=0\,. \tag{3.3}$$

Now, we subject the three-point functional relation of ϕ_k to the following manipulations: On both members of Eq. (2.8) we perform the operations (1) $D_a(q'')$, (2) $D_a(q)$, and (3) $D_a(q')$, separately. Thus we obtain

$$X_a(q')\phi_k(q';q)=r_a^{(k)}[g(q';q)]-r_a^{(k)}(q')\,, \tag{3.4}$$

$$Y_a(q)\phi_k(q';q)=l_a^{(k)}[g(q';q)]-l_a^{(k)}(q)\,, \tag{3.5}$$

$$[X_a(q)-Y_a(q')]\phi_k(q';q)=l_a^{(k)}(q')-r_a^{(k)}(q)\,, \tag{3.6}$$

respectively, where $X_a(q)$ and $Y_a(q)$ are Lie's infinitesimal operators in $M(G)$ [cf. Eqs. (A9) and (A10)].

However, the exponent generators are not completely arbitrary. In fact, if one performs (1) $D_b(q')$ and (2) $D_b(q)$ separately on Eq. (3.6), one gets

$$\begin{aligned}&X_a(q)r_b^{(k)}(q)-\sigma_{ba}^c r_c^{(k)}(q)=l_{a,b}^{(k)}(e)+u_{ab}^{(k)}(q)\,,\\&Y_a(q)l_b^{(k)}(q)-\sigma_{ab}^c l_c^{(k)}(q)=r_{a,b}^{(k)}(e)+v_{ab}^{(k)}(q)\,,\end{aligned} \tag{3.7}$$

where we have defined

$$\begin{aligned}u_{ab}^{(k)}(q)&=\lim_{q'\to e}\partial_a'\,\partial_b'\,\partial\phi_k(q';q)\,,\\v_{ab}^{(k)}(q)&=\lim_{q'\to e}\partial_a'\,\partial_b'\,\phi_k(q;q')\,,\end{aligned} \tag{3.8}$$

and where the σ_{ab}^c denote constants defined in Eq. (A21). Hence, taking $q=e$ in Eqs. (3.7), one obtains

$$r_{a,b}^{(k)}(e)=l_{b,a}^{(k)}(e)\,. \tag{3.9}$$

Moreover, $D_b(q)$ applied to Eq. (3.5) yields

$$X_a(q)l_b^{(k)}(q)=Y_b(q)r_a^{(k)}(q)\,. \tag{3.10}$$

In this way, it can be shown that Eqs. (3.7) and (3.10) are the only first-order differential equations for the exponent generators one can obtain from the set (3.4)–(3.6). Accordingly, besides Eqs. (3.10), taking the skew-symmetric parts of Eqs. (3.7), one concludes that the generators have to satisfy also the following inhomogeneous *non-Abelian curl* equations:

$$X_a(q)r_b^{(k)}(q)-X_b(q)r_a^{(k)}(q)-f_{ab}^c r_c^{(k)}(q)=-k_{ab}\,, \tag{3.11}$$

$$Y_a(q)l_b^{(k)}(q)-Y_b(q)l_a^{(k)}(q)+f_{ab}^c l_c^{(k)}(q)=k_{ab}\,, \tag{3.12}$$

where $f_{ab}^c=\sigma_{ba}^c-\sigma_{ab}^c$ are the structure constants of G (cf. the Appendix), and where the k_{ab} are constants such that

$$k_{ab}=r_{a,b}^{(k)}(e)-r_{b,a}^{(k)}(e)=l_{b,a}^{(k)}(e)-l_{a,b}^{(k)}(e)\,. \tag{3.13}$$

These constants will be briefly referred to as the *ray constants* of the extended representation.

Finally, let us point out that the ray constants themselves are not all independent, in general, for they have to satisfy two sets of constraints. Indeed, we first observe the following constraints:

$$k_{ab}+k_{ba}=0\,. \tag{3.14}$$

Furthermore, if we operate with $D_c(q)$ in Eq. (3.11), we get

$$r_{b,ac}^{(k)}(e)-r_{a,bc}^{(k)}(e)=f_{ab}^c r_{d,c}^{(k)}(e)-\sigma_{ac}^d r_{d,c}^{(k)}(e)+\sigma_{bc}^d r_{a,d}^{(k)}(e)\,,$$

from which the *Bargmann constraints*[3] easily follow:

$$f^d_{ab}k_{cd} + f^d_{ca}k_{bd} + f^d_{bc}k_{ad} = 0\,. \tag{3.15}$$

This completes our analysis of the *necessary* consequences of an admissible exponent function of G.[23]

In the sequel we discuss this framework from a synthetic point of view, since we are searching for necessary and sufficient conditions on which a constructive method of exponent calculus can be grounded. To this end, we organize our discussion in a series of lemmas, which may be proved rather easily once we are in possession of the typical manipulations of "non-Abelian calculus" (cf. the Appendix). For the sake of briefness, however, we here omit the proofs of these lemmas.

Hereafter, any given set $\{r_a(q), l_a(q);\ a = 1,...,r\}$ for which there exists a solution $\phi(q';q)$ of Eqs. (3.4) and (3.5), with the initial conditions (2.5), will be called an *admissible set of exponent generators*. Then, according to Lemma A.V, one has the following lemma.

Lemma 3.I: The solution ϕ associated with an admissible set of generators is unique.

Albeit trivial, this lemma is important because it means that all the gauge freedom one has for settling two-cocycle functions comes from the gauge freedom of the generators themselves (cf. Sec. IV). The following lemma is "crucial" for the issue we are studying.

Lemma 3.II: If $\phi(q';q)$ is the solution associated with an admissible set of generators, then $\phi(q';q)$ satisfies the three-point functional relation (2.8).

Thus admissible sets of generators produce admissible two-cocycle functions. The consistency of Eqs. (3.4) and (3.5) with the initial conditions (2.5) can be checked quite directly, without recourse to the relation (2.8), so that the problem tackled in the previous lemma is well posed indeed.

We now present two lemmas concerning the relations between the several differential equations of the formalism.

Lemma 3.III: For an admissible set of exponent generators such that $r_a(e) = l_a(e)$, $a = 1,...,r$, the associated solution $\phi(q';q)$ of Eqs. (3.4) and (3.5) also satisfies Eq. (3.6).

[Of course, *sensu stricto*, the condition $r_a(e) = l_a(e)$ of the previous lemma is a necessary feature of a set of admissible generators.] Hence, Eq. (3.6) plays no essential role in the constructive approach. Furthermore, we have the following lemma.

Lemma 3.IV: If for a given set of right exponent generators $\{r_a(q)\}$, the function $\phi(q';q)$ satisfies Eq. (3.4) with the initial conditions (2.5), and if there exists a set of left generators $\{l_a(q)\}$ such that Eq. (3.10) holds, then $\phi(q';q)$ also satisfies Eq. (3.5) (and vice versa).

In this way, recollecting the previous results, one has proved the following theorem.

Theorem 3.I: The necessary and sufficient condition for $\phi_k(q';q)$ to be an admissible two-cocycle function of G is that it satisfies either Eqs. (3.4) or Eqs. (3.5), with the initial conditions (2.5).

Next, we present three lemmas concerning the exponent generators.

Lemma 3.V: If $\{r_a(q), l_a(q)\}$ is a set of functions satisfying Eqs. (3.10), and such that $f^c_{ab}r_c(e) = f^c_{ab}l_c(e)$, then these functions also satisfy Eqs. (3.11) and (3.12), where (now) the k_{ab} correspond to a set of constants of integration. (Clearly, $k_{ab} + k_{ba} = 0$.)

As an immediate consequence one has the following lemma.

Lemma 3.VI: For a given set of ray constants, if one assumes the initial conditions (3.3), then Eqs. (3.10) and (3.12) imply Eq. (3.11).

Finally, our last lemma is a consequence of Lemmas A.II and A.VIII.

Lemma 3.VII: If a set of functions $\{r_a(q)\}$ satisfy Eq. (3.11) then there exists a function $\phi(q';q)$ that satisfies (3.4). In the same manner, if a set of functions $\{l_a(q)\}$ satisfies Eqs. (3.12), then there exists a function $\phi(q';q)$ that satisfies (3.5).

Thus one has the following theorem.

Theorem 3.II: For a given set of ray constants, a necessary and sufficient condition for the existence of an admissible exponent function $\phi(q';q)$ is either (1) to have a set of right generators $r^{(k)}_a(q)$ satisfying Eqs. (3.11), or else (2) to have a set of left generators $l^{(k)}_a(q)$ satisfying Eqs. (3.12), together with the initial conditions (3.3).

In summary, given a connected Lie group G, the general procedure for obtaining an admissible two-cocycle function $\phi_k(q';q)$ is clear. One first introduces a suitable set of ray constants, which have to satisfy the required constraints (3.12) and (3.15), and may be otherwise arbitrary. Then one may solve the problem completely "from the right" [i.e., solve Eqs. (3.11) and (3.4)], or completely "from the left" [i.e., solve (3.12) and (3.5)]. We present a resume of this general method in Table I, adopting the "right" framework for the sake of concreteness. (However, this method is far from producing a unique two-cocycle function, because of

TABLE I. General procedure (from the right) for obtaining an admissible nontrivial local exponent function of a connected non-Abelian Lie group.

Step	Symbol	Solve	Initial conditions
(1) Ray constants	k_{ab}	$k_{ab} + k_{ba} = 0$ $f^d_{ab}k_{cd} + f^d_{ca}k_{bd} + f^d_{bc}k_{ad} = 0$	$k_{ab} \neq f^c_{ab}k_c$
(2) Right exponent generators	$r^{(k)}_a(q)$	$X_a(q)r^{(k)}_b(q) - X_{(b)}(q)r^{(k)}_a(q) - f_{ab}r^{(k)}_c(q) = -k_{ab}$	$r^{(k)}_a(e) = 0$
(3) Two-cocycle	$\phi_k(q';q)$	$X_a(q')\phi_k(q';q) = r^{(k)}_a[g(q';q)] - r^{(k)}_a(q')$	$\phi_k(e;q) = 0$

the ample gauge freedom one has in selecting the representative operators.)

As a very simple example of the method we observe that, in particular, for an Abelian Lie group all the ray constants can be taken arbitrarily (within $k_{ab}+k_{ba}=0$). In this case, if the q's are canonical parameters, the exponent generators are simply given by

$$r_a^{(k)}(q)=l_a^{(k)}(q)=\tfrac{1}{2}k_{ab}q^b\,, \tag{3.16}$$

and the admissible two-cocycles are all of the form[3]

$$\phi_k(q';q)=\tfrac{1}{2}k_{ab}q'^a q^b\,. \tag{3.17}$$

More interesting examples are discussed in Sec. VI.

It must be borne in mind that for a connected Lie group the previous method produces only a *local* two-cocycle $\phi_k(q';q)$ defined on $N(e)\times N(e)$. In the applications one usually obtains a two-cocycle of class C^∞ quite directly (i.e., without recourse to Iwasawa's construction[22]). Moreover, if G is a noncompact, connected, and simply connected Lie group, the method usually yields a *global* two-cocycle which is C^∞ on $M(G)\times M(G)$.

IV. GAUGE TRANSFORMATIONS AND GENUINE UNITARY RAY REPRESENTATIONS

As was already mentioned, it is evident that the exponent factor one uses in a ray representation depends on the selection of representative operators. Thus if one considers a different choice of representatives, say,

$$U_{k'}(q)=e^{i\gamma_{k'k}(q)}U_k(q)\,, \tag{4.1}$$

taken from the corresponding operator rays, then a new two-cocycle function appears in Eq. (2.1); namely,

$$\phi_{k'}(q';q)=\phi_k(q';q)+\gamma_{k'k}(q')+\gamma_{k'k}(q)-\gamma_{k'k}[g(q';q)]\,. \tag{4.2}$$

Here $\gamma_{k'k}(q)$ is an arbitrary real function, provided it satisfies

$$\gamma_{k'k}(e)=0. \tag{4.3}$$

Two exponents related in this fashion are called *equivalent*. Hence one has a *local* gauge freedom inherent to the unitary ray representations formalism, since Eq. (4.2) is a gauge transformation of the second kind.

Taking $q'=\bar{q}$ in Eq. (4.2) yields

$$\mu_{k'}(q)=\mu_k(q)+\gamma_{k'k}(\bar{q})+\gamma_{k'k}(q)\,. \tag{4.4}$$

(We shall return to this equation presently.) From the definitions (3.1) and (3.2), the general gauge transformation law induced by Eq. (4.1) on the exponent generators immediately obtains; viz.,

$$r_a^{(k')}(q)=r_a^{(k)}(q)-X_a(q)\gamma_{k'k}(q)+\gamma_{k'k,a}(e)\,, \tag{4.5}$$

$$l_a^{(k')}(q)=l_a^{(k)}(q)-Y_a(q)\gamma_{k'k}(q)+\gamma_{k'k,a}(e)\,. \tag{4.6}$$

As a consequence, the general gauge transformation induced on the ray constants follows:

$$k'_{ab}=k_{ab}+f^c_{ab}\gamma_{k'k,c}(e)\,. \tag{4.7}$$

Clearly, we are using the label $k=(k_{1,2},\ldots,k_{r-1,r})$ to denote the set of ray constants used in the determination of the two-cocycle $\phi_k(q';q)$, attached to the $U_k(q)$'s according to Eq. (2.1). Hence, if $\gamma_{k'k}(e)=0$ and $f^c_{ab}\gamma_{k'k,c}(e)=0$ (i.e., $k'_{ab}=k_{ab}$), we write $\gamma_{kk}(q)=\gamma_k(q)$ and we say that the two exponent functions $\phi_k(q';q)$ and $\phi'_k(q';q)$ are equivalent within a *restricted* gauge transformation. Briefly, *restricted gauge transformations leave the ray constants invariant.*

It can be proved quite directly that the consistency of the *general* gauge transformation scheme demands

$$[X_a(q)-Y_a(q')]\gamma_{k'k}[g(q';q)]=0\,. \tag{4.8}$$

However, according to Lemma A.II, this is an identity, and therefore it sets no restriction on the gauge generating function $\gamma_{k'k}(q)$. Thus, in conclusion, one has a consistent scheme of gauge transformations (of the second kind), and every *unitary ray representation of G defines in a unique way only a class of equivalent exponent functions.*[3]

We next supply some lemmas that concern the gauge transformations of unitary ray representations.

Lemma 4.I: For a *given* set k of ray constants, the generators $\{r_a^{(k)}(q), l_a^{(k)}(q)\}$ are defined only within the restricted gauge transformation,

$$r'^{(k)}_a(q)=r_a^{(k)}(q)-X_a(q)\gamma_k(q)\,, \tag{4.9}$$

$$l'^{(k)}_a(q)=l_a^{(k)}(q)-Y_a(q)\gamma_k(q)\,, \tag{4.10}$$

provided

$$\gamma_{k,a}(e)=0\,. \tag{4.11}$$

This lemma (4.I) is an immediate consequence of the Lie algebra $L(G)$ (i.e., cf. Lemma A.III).

Lemma 4.II: A general gauge transformation of the exponent generators [viz., Eqs. (4.5) and (4.6)] induces a gauge transformation of the exponent function [i.e., Eq. (4.2)].

Furthermore, as was already remarked (cf. Lemma 3.I), all the gauge freedom of a two-cocycle function comes from the gauge freedom of the exponent generators. One also proves the following lemma as a corollary to Lemma 4.I.

Lemma 4.III: An arbitrary gauge transformation of the ray constants, i.e.,

$$k'_{ab}=k_{ab}+f^c_{ab}k_c \tag{4.12}$$

(with $k_a, a=1,\ldots,r$, arbitrary real constants), induces a general gauge transformation of the exponent generators [i.e., Eqs. (4.5) and (4.6)], with the only proviso that

$$\gamma_{k'k,a}(e)=k_a\,. \tag{4.13}$$

So we have shown that the whole formalism of gauge transformations, as deduced from Eq. (4.1), is inversible. Let us epitomize this in the following theorem.

Theorem 4.I: The general gauge transformation [i.e., Eq. (4.12), with some $f^c_{ab}k_c\neq 0$] of the ray constants is a necessary and sufficient condition for having a nonrestricted gauge transformation [i.e., Eq. (4.2)] of the exponent function.

In Table II we present a summary of general gauge transformations for unitary ray representations of non-Abelian connected Lie groups.

Of course, *genuine* unitary ray representations of G are those ray representations that may *not* be reduced to a vector representation by a mere gauge transformation, as presented above. In order to further study this matter, we better con-

TABLE II. Gauge transformation scheme for unitary ray representations. The transformations shown are necessary and sufficient conditions for having equivalent unitary ray representations.

Step	Gauge transformations	Conditions
(1) Ray constants	$k_{ab} \to k'_{ab} = k_{ab} + f^c_{ab} k_c$	...
(2) Exponent generators	$r_a(q) \to r'_a(q) = r_a(q) - X_a(q)\gamma(q) + k_a$ $l_a(q) \to l'_a(q) = l_a(q) - Y_a(q)\gamma(q) + k_a$	$\gamma_{,a}(e) = k_a$
(3) Two-cocycle	$\phi(q';q) \to \phi'(q';q) = \phi(q';q) - \gamma[g(q';q)] + \gamma(q') + \gamma(q) - \gamma(e)$	...
(4) Gauge generator	$\gamma(q) \to \gamma'(q) = \gamma(q) - \gamma(e)$	$\gamma'(e) = 0$

sider instead *trivial* unitary ray representations; namely, those whose exponent functions are of the form

$$\phi(q';q) = \gamma[g(q';q)] - \gamma(q') - \gamma(q) . \tag{4.14}$$

Two-cocycles of this kind are nothing but gauge artifacts; as such, they bear no physical meaning in quantum theory. Therefore, it is worthwhile to learn how to *avoid* these trivial solutions.

It is an immediate consequence of Eq. (4.14) that the *trivial exponent generators* are of the form

$$r_a(q) = X_a(q)\gamma(q) + k_a , \tag{4.15}$$

$$l_a(q) = Y_a(q)\gamma(q) + k_a , \tag{4.16}$$

where we set $k_a = -\gamma_{,a}(e)$, and therefore the corresponding trivial ray constants are given by

$$k_{ab} = f^c_{ab} k_c \tag{4.17}$$

(which are obviously admissible, in principle). One can easily see that Eqs. (4.14), with (4.15) and (4.16), satisfy the system (3.4)–(3.6) in a trivial fashion, that is, identically for whatever function $\gamma(q)$ one may consider. In the same way, it also can be shown that Eqs. (4.15) and (4.16), with (4.17), satisfy Eqs. (3.10)–(3.12) in a trivial manner. Hence, one has no true differential equations for the determination of $\gamma(q)$, as it should be. In other words, albeit trivial and arbitrary, a purely gauge exponent function is an admissible solution of the problem.

Conversely (and most importantly), the reader can prove the following lemmas.

Lemma 4.IV: Trivial exponent generators produce only trivial two-cocycle functions.

Lemma 4.V: Trivial ray constants produce only trivial exponent generators.

We thus have the following theorem.

Theorem 4.II: The necessary and sufficient condition for having a trivial two-cocycle function is that *all* the ray constants are trivial.

This theorem settles the issue since, as a corollary, to have a genuine unitary ray representation of G it is enough to introduce at least one ray constant that is *not* trivial (i.e., such that $0 \neq k_{ab} \neq f^c_{ab} k_c$). A glance at Eq. (4.7) shows that *it is impossible to eliminate a nontrivial ray constant by means of a gauge transformation.* On the other hand, if a given set k contains *some* trivial ray constants, these always may be eliminated simultaneously by means of a suitable change of gauge. In particular, for an Abelian Lie group, all trivial ray constants are necessarily zero, and thus yield nothing (i.e., *all ray representations of Abelian Lie groups are genuine*).

An interesting consequence of Theorem 4.II arises, for instance, in applications to the group SU(2), for which *all* the ray constants are obviously trivial. Hence, all unitary ray representations of SU(2) are gauge artifacts (as is well known indeed). An analogous result holds for the homogeneous Lorentz group. (Even so, let us recall that the double-valued representations of these groups are more properly interpreted as ray representations.)

V. A SPECIAL GAUGE

As usual when one deals with a theory that formally contains some gauge freedom, in the unitary ray representation theory one takes advantage of the allowable gauge transformation to choose a gauge in which the two-cocycle function $\phi_k(q';q)$ becomes "simpler" (or behaves in a physically "reasonable" manner). On physical grounds, one of the most reasonable properties of ϕ_k one would like to retain when the formalism is used in quantum theory corresponds to the unitarity of the representative operators, as expressed in Eq. (2.4). There are, of course, other gauges within the unitary ray formalism. However, henceforth we impose the following gauge condition:

$$\mu_k(q) = 0 . \tag{5.1}$$

It is clear that in order to transform a given unitary ray representation into a ray representation belonging to this special gauge one uses $\gamma_k(q) = -\frac{1}{2}\mu_k(q)$ as the generator of the required gauge transformation. Moreover, it is also clear that *one still has some remaining gauge freedom for selecting the representative operators within this gauge.* Indeed, this freedom entails the property

$$\gamma_k(q) + \gamma_k(\bar{q}) = 0 , \tag{5.2}$$

which must be satisfied everywhere by the gauge generating function. In agreement with Eq. (A26), this means that the gauge generators $\gamma_k(q)$ (which operate within the special gauge) must be solutions of the differential equations

$$X_a(q)\gamma_k(q) - Y_a(\bar{q})\gamma_k(\bar{q}) = 0 , \tag{5.3}$$

together with the initial datum $\gamma_k(e) = 0$.

When expressed in terms of the exponent function $\phi_k(q';q)$ itself, our special gauge is characterized by the property

$$\phi_k(q';q) + \phi_k(\bar{q};\bar{q}') = 0\,, \tag{5.4}$$

for all $q',q \in M(G)$, which is clearly equivalent to Eq. (5.1). If one now performs the operation $D_a(q')$ in this last equation, and uses Eq. (A26), one easily obtains the relation

$$l_a^{(k)}(q) = r_a^{(k)}(\bar{q}) \tag{5.5}$$

characterizing the exponent generators within this gauge. In effect, conversely, if one assumes this relation and considers Eqs. (3.4) and (3.5), written for $\phi_k(q';q)$ *and* for $\phi_k(\bar{q};\bar{q}')$, one arrives at Eq. (5.4), because of the initial conditions (2.5). Hence we arrive at the following theorem.

Theorem 5.I: The relation $l_a^{(k)}(q) = r_a^{(k)}(\bar{q})$ is a necessary and sufficient condition for the special gauge defined in Eq. (5.1).

Furthermore, if one assumes Eq. (5.5), then the differential equations for the exponent generators are Eqs. (3.11) *and* also

$$Y_a(q) r_b^{(k)}(q) + Y_b(\bar{q}) r_a^{(k)}(\bar{q}) = 0\,. \tag{5.6}$$

In the applications, however, it may be rather cumbersome to find a solution $r_a^{(k)}(q)$, $a = 1,...,r$, of Eq. (3.11), with $r_a^{(k)}(e) = 0$, which at the same time satisfies Eq. (5.6).

According to the preceding discussion, one can always bring a (previously calculated) exponent function into the special gauge (5.1) by performing a suitable gauge transformation at the *end* of the calculations.

VI. MISCELLANEOUS EXAMPLES

With the aim of exhibiting the technicalities of non-Abelian two-cocycle calculus, in this section we present some simple instances that are, at the same time, mathematically nontrivial and physically important. Many features of the results obtained in this section are well known, of course. Our emphasis is on *method*. Here we content ourselves with obtaining the formal results, and do not delve into their physical meaning.

A. The Galilei group in one-dimensional space

Let us apply the formalism of two-cocycle calculus to the Galilei group in one-dimensional space. The group-multiplication functions for this three-parameter Lie group are given by

$$\begin{aligned} q''^1 &= g^1(q';q) = q'^1 + q^1\,, \\ q''^2 &= g^2(q';q) = q'^2 + q^2 - q'^3 q^1\,, \\ q''^3 &= g^3(q';q) = q'^3 + q^3\,, \end{aligned} \tag{6.1}$$

where q^1 corresponds to Newtonian time translation, q^2 to Euclidean space translation, and q^3 to the Galilean boost in one-dimensional space. Clearly, the group manifold corresponds to $-\infty < q^a < \infty$, $a = 1,2,3$, with the identity point at the origin, $e = (0,0,0)$, and the group-inversion formulas for the parameters are

$$\bar{q}^1 = -q^1\,, \quad \bar{q}^2 = -q^2 - q^3 q^1\,, \quad \bar{q}^3 = -q^3\,. \tag{6.2}$$

Thus one easily obtains the right and left infinitesimal operators; namely,

$$X_1 = \partial_1\,, \quad X_2 = \partial_2\,, \quad X_3 = \partial_3 - q^1 \partial_2\,, \tag{6.3}$$

and

$$Y_1 = \partial_1 - q^3 \partial_2\,, \quad Y_2 = \partial_2\,, \quad Y_3 = \partial_3\,, \tag{6.4}$$

respectively. The well-known Lie algebra follows:

$$\begin{aligned} &[X_1,X_2] = 0\,, \quad [X_1,X_3] = -X_2\,, \quad [X_2,X_3] = 0\,, \\ &[Y_1,Y_2] = 0\,, \quad [Y_1,Y_3] = Y_2\,, \quad [Y_2,Y_3] = 0\,, \end{aligned} \tag{6.5}$$

and also $[X_a,Y_b] = 0$, for $a,b = 1,2,3$. Hence the only nonzeroth structure constant is $f_{13}^2 = -1$.

After settling these details, we are ready to apply two-cocycle calculus to this group. One first considers the constraints (3.15) for the ray constants. After some simple calculations, one concludes that all the ray constants $\{k_{12},k_{13},k_{23}\}$ survive. Clearly, k_{13} is a gauge artifact and may be eliminated from the beginning. However, since our purpose is merely illustrative, it seems worthwhile to feign ignorance on this fact, and manage the issue with all three ray constants present. Then Eqs. (3.11) for the right generators yield

$$r_{1,2} - r_{2,1} = k_{12}\,, \tag{6.6}$$

$$r_{2,3} - r_{3,2} - q^1 r_{2,2} = k_{23}\,, \tag{6.7}$$

$$r_{3,1} - r_{1,3} + q^1 r_{1,2} + r_2 = k_{31}\,. \tag{6.8}$$

Of course, one can tackle these equations following several integration schemes. We shall use the following scheme.

Let us assume (without loss of generality)

$$r_1(q) = \tfrac{1}{2} k_{12} q^2 + \tfrac{1}{2} k_{13} q^3 + u(q^1,q^2,q^3)\,, \tag{6.9}$$

where $u(q)$ is an undetermined real function. Then, Eqs. (6.6)–(6.8) (also without loss of generality) yield

$$\begin{aligned} r_2(q) &= \tfrac{1}{2} k_{21} q^1 + \tfrac{1}{2} k_{23} q^3 \\ &\quad + \int_0^{q^1} dq'^1\, u_{,2} + v(q^2,q^3)\,, \end{aligned} \tag{6.10}$$

$$\begin{aligned} r_3(q) &= \tfrac{1}{2} k_{31} q^1 + \tfrac{1}{2} k_{32}(q^2 + q^3 q^1) \\ &\quad + \int_0^{q^1} dq'^1\, u_{,3} - q^1 \int_0^{q^1} dq'^1\, u_{,2} \\ &\quad + \int_0^{q^2} dq'^2\, v_{,3} - q^1 v + w(q^3)\,. \end{aligned} \tag{6.11}$$

Thus Eqs. (6.9)–(6.11) represent a general solution to Eqs. (6.6)–(6.8). However, the functions $u(q^1,q^2,q^3)$, $v(q^2,q^3)$, and $w(q^3)$ remain completely arbitrary. Clearly, defining the function

$$\gamma(q) = \int_0^{q^1} dq'^1\, u + \int_0^{q^2} dq'^2\, v + \int_0^{q^3} dq'^3\, w\,, \tag{6.12}$$

one obtains

$$\begin{aligned} x_1(q)\gamma(q) &= u(q^1,q^2,q^3)\,, \\ x_2(q)\gamma(q) &= \int_0^{q^1} dq'^1\, u_{,2} + v(q^2,q^3)\,, \\ x_3(q)\gamma(q) &= \int_0^{q^1} dq'^1\, u_{,3} - q^1 \int_0^{q^1} dq'^1\, u_{,2} \\ &\quad + \int_0^{q^2} dq'^2\, v_{,3} - q^1 v + w(q^3)\,, \end{aligned} \tag{6.13}$$

and also $\gamma(e) = \gamma(0,0,0) = 0$. So one gets, after performing

the gauge transformation generated by $\gamma(q)$ in (6.9)–(6.11) the following *completely gauge-reduced* solution:

$$\begin{aligned} r_1(q) &= \tfrac{1}{2}k_{12}q^2 + \tfrac{1}{2}k_{13}q^3 , \\ r_2(q) &= \tfrac{1}{2}k_{21}q^1 + \tfrac{1}{2}k_{23}q^3 , \\ r_3(q) &= \tfrac{1}{2}k_{31}q^1 + \tfrac{1}{2}k_{32}(q^2 + q^3q^1) . \end{aligned} \tag{6.14}$$

Thereafter, for the corresponding left exponent generators, using Eq. (3.10), one obtains

$$\begin{aligned} l_1(q) &= -\tfrac{1}{2}k_{12}(q^2 + q^3q^1) - \tfrac{1}{2}k_{13}q^3 , \\ l_2(q) &= -\tfrac{1}{2}k_{21}q^1 - \tfrac{1}{2}k_{23}q^3 , \\ l_3(q) &= -\tfrac{1}{2}k_{31}q^1 - \tfrac{1}{2}k_{32}q^2 ; \end{aligned} \tag{6.15}$$

which, according to Eq. (6.2), show that these exponent generators belong to the special gauge discussed in Sec. V [i.e., one has $l_a(q) = r_a(\bar{q})$].

We now finish our work using these exponent generators. Let us consider Eq. (3.4) for the two-cocycle function produced by the right generators. In the present case these equations are

$$\begin{aligned} &\partial'_1\phi(q';q) = \tfrac{1}{2}k_{12}(q^2 - q'^3q^1) + \tfrac{1}{2}k_{13}q^3 , \\ &\partial'_2\phi(q';q) = \tfrac{1}{2}k_{21}q^1 + \tfrac{1}{2}k_{23}q^3 , \\ &(\partial'_3 - q'^1\partial'_2)\phi(q';q) \\ &\quad = \tfrac{1}{2}k_{31}q^1 + \tfrac{1}{2}k_{32}(q^2 + q^3q'^1 + q^3q^1) . \end{aligned} \tag{6.16}$$

A straightforward integration of this system yields

$$\begin{aligned} \phi(q';q) = {} & \tfrac{1}{2}k_{12}[q'^1(q^2 - q'^3q^1) - q'^2q^1] \\ & + \tfrac{1}{2}k_{23}[q'^2q^3 - q'^3(q^2 + q^1q^3)] \\ & + \tfrac{1}{2}k_{31}(q'^3q^1 - q'^1q^3) , \end{aligned} \tag{6.17}$$

which meets the condition $\phi(e;q) = \phi(q;e) = 0$. The reader can show that $\phi(q';q)$ as given in Eq. (6.17) is also a solution of (3.5). Furthermore, a rather tedious (but easy) calculation shows that this solution satisfies the three-point functional relation (2.8). Clearly, $\phi(\bar{q};q) = \phi(q;\bar{q}) = 0$ (as expected).

Finally, we observe that the gauge transformation generated by

$$\gamma(q) = -\tfrac{1}{2}k_{31}(q^1q^3 + 2q^2) , \tag{6.18}$$

when performed on the solution (6.17), eliminates the k_{31} term. Thus our solution reduces to the form

$$\begin{aligned} \phi(q'q) = {} & \tfrac{1}{2}k_{12}[q'^1(q^2 - q'^3q^1) - q'^2q^1] \\ & + \tfrac{1}{2}k_{23}[q'^2q^3 - q'^3(q^2 + q^1q^3)] \end{aligned} \tag{6.19}$$

(which, of course, could have been assured from the beginning, since k_{31} is a gauge artifact). This solution is *unique* in the sense that it is a *completely gauge-reduced* two-cocycle satisfying $\mu(q) = 0$. It is interesting to observe that $\gamma(q)$, as defined in Eq. (6.18), generates a gauge transformation that is *not* a restricted one [i.e., one has $\gamma_{,1}(e) = 0$, $\gamma_{,2}(e) = -k_{31}, \gamma_{,3}(e) = 0$], nevertheless it operates within our gauge [i.e., $\gamma(q) = -\gamma(\bar{q})$]. Once we have a unitary ray representation of the Galilei group (corresponding to one-dimensional space), with the phase function given by

$$\phi^{(12)}(q';q) = \tfrac{1}{2}k_{12}[q'^1(q^2 - q^3q^1) - q'^2q^1] , \tag{6.20}$$

or

$$\phi^{(23)}(q';q) = \tfrac{1}{2}k_{23}[q'^2q^3 - q'^3(q^2 + q^3q^1)] \tag{6.21}$$

[or else $\phi^{(12)} + \phi^{(23)}$, as in (6.19)], then it is *not* possible to change this ray representation into a vector representation by a mere gauge transformation.

The literature on unitary ray representations of the Galilei group is rather extensive.[3,24,25] The issue has been studied mainly in connection with the Galilean invariance of the Schrödinger equation of a free particle.[26,27] Indeed, the important fact to remark is that these projective unitary representations of the Galilei group are the only ones to which we can attribute a usual physical meaning.[28,29] The two-cocycle function $\phi^{(23)}$ presented in Eq. (6.21) corresponds to the one-dimensional space version of the well-known Galilei two-cocycle function for three-dimensional space. (The constant k_{23} corresponds to the mass of the particle.[3]) The two-cocycle $\phi^{(12)}$, presented in Eq. (6.20), also figures in the current literature,[30] and its physical meaning has been discussed by Lévy-Leblond.

B. The Euclidean group in the plane

The group of translation and rotations of two-dimensional Euclidean space has the following rules of binary combination of the parameters:

$$\begin{aligned} q''^1 &= g^1(q';q) = q'^1 + q^1 , \\ q''^2 &= g^2(q';q) = q'^2\cos\omega q^1 + q'^3\sin\omega q^1 + q^2 , \\ q''^3 &= g^3(q';q) = q'^2\sin\omega q^1 + q'^3\cos\omega q^1 + q^3 , \end{aligned} \tag{6.22}$$

where $\omega q^1 = \theta$ is the angle of rotation in the plane, and $\{q^2,q^3\}$ corresponds to space translations along a system of rectangular Cartesian axes. The group manifold is given by $-\pi < \omega q^1 < +\pi$, $-\infty < q^2 < +\infty$, $-\infty < q^3 < +\infty$, the identity corresponds to the origin, $e = (0,0,0)$, and the inversion formulas for the parameters are

$$\begin{aligned} \bar{q}^1 &= -q^1 , \\ \bar{q}^2 &= -q^2\cos\omega q^1 + q^3\sin\omega q^1 , \\ \bar{q}^3 &= -q^2\sin\omega q^1 - q^3\cos\omega q^1 . \end{aligned} \tag{6.23}$$

Hence, the right and left infinitesimal operators are

$$\begin{aligned} &X_1 = \partial_1 , \quad X_2 = \cos\omega q^1\,\partial_2 - \sin\omega q^1\,\partial_3 , \\ &X_3 = \sin\omega q^1\,\partial_2 + \cos\omega q^1\,\partial_3 , \end{aligned} \tag{6.24}$$

and

$$\begin{aligned} &Y_1 = \partial_1 - \omega(q^2\,\partial_3 - q^3\,\partial_2) , \\ &Y_2 = \partial_2 , \quad Y_3 = \partial_3 , \end{aligned} \tag{6.25}$$

respectively. So one gets two nonzeroth structure constants; namely, $f^3_{21} = f^2_{13} = \omega$. Therefore, k_{23} is the only admissible ray constant that is nontrivial.

We next solve the differential equations for the right exponent generators; i.e.,

$$\begin{aligned} &(\cos\omega q^1)r_{1,2} - (\sin\omega q^1)r_{1,3} = r_{2,1} + \omega r_3 , \\ &(\sin\omega q^1)r_{1,2} + (\cos\omega q^1)r_{1,3} = r_{3,1} - \omega r_2 , \\ &(\cos\omega q^1)(r_{3,2} - r_{2,3}) - (\sin\omega q^1)(r_{2,2} + r_{3,3}) = -k_{23} . \end{aligned} \tag{6.26}$$

It seems advantageous to consider, without loss of generality,

$$r_2(q) = \tfrac{1}{2} k_{23} q^3 \sec \omega q^1 + u(q^1,q^2,q^3)\,,$$
$$r_3(q) = \tfrac{1}{2} k_{32} q^2 \sec \omega q^1 + v(q^1,q^2,q^3)\,, \tag{6.27}$$

which, when substituted into the last of Eqs. (6.26), yield

$$(\cos \omega q^1)(v_{,2} - u_{,3}) - (\sin \omega q^1)(u_{,2} + v_{,3}) = 0\,. \tag{6.28}$$

Clearly, as a simple possibility for this last relation to hold, one may choose $v_{,2} - u_{,3} = 0$, and $u_{,2} + v_{,3} = 0$, wherefrom one easily obtains

$$u(q^1,q^3) = w(q^1)q^3 + f(q^1)\,,$$
$$v(q^1,q^2) = w(q^1)q^2 + g(q^1)\,. \tag{6.29}$$

(This "ansatz" is legitimate, since we are just looking for *one* set of right generators.) Thus we have

$$r_2(q) = \tfrac{1}{2} k_{23} q^3 \sec \omega q^1 + w(q^1)q^3 + f(q^1)\,,$$
$$r_3(q) = \tfrac{1}{2} k_{32} q^2 \sec \omega q^1 + w(q^1)q^2 + g(q^1)\,; \tag{6.30}$$

which, when substituted into the first two equations of (6.26), give us

$$\begin{aligned} r_1(q) = {} & -\tfrac{1}{4} k_{23}\omega[(q^2)^2 + (q^3)^2]\sec^2 \omega q^1 + \tfrac{1}{2}[(q^2)^2 - (q^3)^2][\omega w(q^1)\cos \omega q^1 + \dot{w}(q^1)\sin \omega q^1] \\ & - q^2 q^3[\omega w(q^1)\sin \omega q^1 - \dot{w}(q^1)\cos \omega q^1] + q^2\{[\dot{f}(q^1) + \omega g(q^1)]\cos \omega q^1 + [\dot{g}(q^1) - \omega f(q^1)]\sin \omega q^1\} \\ & - q^3\{[\dot{f}(q^1) + \omega g(q^1)]\sin \omega q^1 - [\dot{g}(q^1) - \omega f(q^1)]\cos \omega q^1\} + s(q^1)\,. \end{aligned} \tag{6.31}$$

The functions $f(q^1), g(q^1), w(q^1)$, and $s(q^1)$ remain arbitrary. However, it is an obvious *general rule* that all those undefined terms in an exponent generator that do not exhibit a linear dependence on the ray constants may be eliminated by means of a suitable restricted gauge transformation. Indeed, in the present example it can be proved that the function

$$\begin{aligned} \gamma(q) = {} & \tfrac{1}{2} w(q^1)[(q^2)^2 - (q^3)^2]\sin \omega q^1 + w(q^1)q^2 q^3 \cos \omega q^1 \\ & + g(q^1)(q^2 \cos \omega q^1 - q^3 \sin \omega q^1) + f(q^1)(q^2 \sin \omega q^1 + q^3 \cos \omega q^1) + \int_0^{q^1} dq'^1\, u(q'^1) \end{aligned} \tag{6.32}$$

generates a gauge transformation that reduces the right exponent generators [presented in Eqs. (6.30) and (6.31)] to the form

$$r_1(q) = \tfrac{1}{4} k\omega[(q^2)^2 + (q^3)^2]\sec^2 \omega q^1\,, \quad r_2(q) = -\tfrac{1}{2} k q^3 \sec \omega q^1\,, \quad r_3(q) = \tfrac{1}{2} k q^2 \sec \omega q^1\,, \tag{6.33}$$

where we have written $k = k_{32}$. Therefore, the associated left exponent generators are given by

$$l_1(q) = \tfrac{1}{4} k\omega[(q^2)^2 + (q^3)^2]\sec^2 \omega q^1\,, \quad l_2(q) = \tfrac{1}{2} k q^2 \tan \omega q^1 + \tfrac{1}{2} k q^3\,, \quad l_3(q) = \tfrac{1}{2} k q^3 \tan \omega q^1 - \tfrac{1}{2} k q^2\,. \tag{6.34}$$

One can easily check that these phase generators belong to the $\mu(q) = 0$ gauge [cf. Sec. V].

We now obtain the two-cocycle corresponding to these generators. If one proceeds from the left, one has to solve the following set of differential equations:

$$\begin{aligned} & [\partial_1 - \omega(q^2 \partial_3 - q^3 \partial_2)]\phi(q';q) \\ & \quad = \tfrac{1}{4} k\omega[(q'^2)^2 + (q'^3)^2 + (q^2)^2 + (q^3)^2]\sec^2 \omega(q'^1 + q^1) + \tfrac{1}{2} k\omega(q'^2 q^2 + q'^3 q^3)\cos \omega q^1 \sec^2 \omega(q'^1 + q^1) \\ & \qquad + \tfrac{1}{2} k\omega(q'^3 q^2 - q'^2 q^3)\sin \omega q^1 \sec^2 \omega(q'^1 + q^1) - \tfrac{1}{4} k\omega[(q^2)^2 + (q^3)^2]\sec^2 \omega q^1\,, \end{aligned} \tag{6.35}$$

$$\begin{aligned} \partial_2 \phi(q';q) = {} & \tfrac{1}{2} k(q'^2 \cos \omega q^1 + q'^3 \sin \omega q^1)\tan(q'^1 + q^1) \\ & - \tfrac{1}{2} k(q'^2 \sin \omega q^1 - q'^3 \cos \omega q^1) + \tfrac{1}{2} k q^2[\tan \omega(q'^1 + q^1) - \tan \omega q^1]\,, \end{aligned} \tag{6.36}$$

$$\begin{aligned} \partial_3 \phi(q';q) = {} & -\tfrac{1}{2} k(q'^2 \sin \omega q^1 - q'^3 \cos \omega q^1)\tan \omega(q'^1 + q^1) \\ & - \tfrac{1}{2} k(q'^2 \cos \omega q^1 + q'^3 \sin \omega q^1) + \tfrac{1}{2} k q^3[\tan \omega(q'^1 + q^1) - \tan \omega q^1]\,. \end{aligned} \tag{6.37}$$

A straightforward calculation then yields the final answer:

$$\begin{aligned} \phi_k(q';q) = {} & \tfrac{1}{4} k[(q'^2)^2 + (q'^3)^2][\tan \omega(q'^1 + q^1) - \tan \omega q'^1] + \tfrac{1}{4} k[(q^2)^2 + (q^3)^2][\tan \omega(q'^1 + q^1) - \tan \omega q^1] \\ & + \tfrac{1}{2} k(q'^2 q^2 + q'^3 q^3)[\cos \omega q^1 \tan \omega(q'^1 + q^1) - \sin \omega q^1] \\ & + \tfrac{1}{2} k(q'^3 q^2 - q'^2 q^3)[\sin \omega q^1 \tan \omega(q'^1 + q^1) + \cos \omega q^1]\,. \end{aligned} \tag{6.38}$$

The reader can check this solution against the fundamental three-point functional relation (2.8). This two-cocycle belongs to the special gauge $\mu(q) = 0$.

Let us recall that the Euclidean group in two dimensions corresponds precisely to the Newtonian symmetry group of the (classical) one-dimensional harmonic oscillator.[15] In another paper we have dealt with the quantum kinematic of the harmonic oscillator, using the regular ray representation of the Newtonian group E_2 of the system, with the two-cocycle presented in Eq. (6.38).[15]

C. The Poincaré group in two-dimensional space-time

The Poincaré group in two-dimensional Minkowskian space-time $\{(t,x)\}$ is a three-parameter Lie group, with the following binary combination law:

$$\begin{aligned} q''^1 &= \lambda(q'^3)(q^1 - q'^3 q^2) + q'^1\,, \\ q''^2 &= \lambda(q'^3)(q^2 - q'^3 q^1) + q'^2\,, \\ q''^3 &= (q'^3 + q^3)/(1 - q'^3 q^3)\,, \end{aligned} \tag{6.39}$$

where

$$\lambda(q^3) = |1 - (q^3)^2|^{-1/2} . \tag{6.40}$$

Clearly, q^1 corresponds to coordinate-time translation, q^2 to (one-dimensional) space translation, and q^3 is the one-dimensional Lorentz boost. (We set $c = 1$.) The group manifold is thus defined by $-\infty < q^1 < +\infty$, $-\infty < q^2 < +\infty$, $1 < q^3 < +1$; and $e = (0,0,0)$. Thus

$$\bar{q}^1 = -\lambda(q^3)(q^1 + q^3 q^2) , \quad \bar{q}^2 = -\lambda(q^3)(q^2 + q^3 q^1) ,$$
$$\bar{q}^3 = -q^3 . \tag{6.41}$$

Using $\partial_3 \lambda(q^3) = q^3[\lambda(q^3)]^3$ one obtains the following infinitesimal operators:

$$X_1 = \partial_1 , \quad X_2 = \partial_2 ,$$
$$X_3 = -q^2 \partial_1 + [1 - (q^3)^2]\partial_3 , \tag{6.42}$$

and

$$Y_1 = \lambda(q^3)(\partial_1 - q^3 \partial_2) , \quad Y_2 = \lambda(q^3)(\partial_2 - q^3 \partial_1) ,$$
$$Y_3 = [1 - (q^3)^2]\, \partial_3 . \tag{6.43}$$

Therefore, the nonzeroth structure constants are $f^2_{31} = f^1_{32} = 1$. Hence this group has just one nontrivial ray constant; namely, $k_{12} = k \neq 0$. Now, the differential equations for the right exponent generators read

$$r_{1,2} - r_{2,1} = 1 ,$$
$$r_{3,2} + q^2 r_{2,1} + q^1 r_{2,2} - [1 - (q^3)^2] r_{2,3} + r_1 = 0 , \tag{6.44}$$
$$q^2 r_{1,1} + q^1 r_{1,2} - [1 - (q^3)^2] r_{1,3} + r_{3,1} + r_2 = 0 .$$

Starting with the assumption

$$r_1(q) = \tfrac{1}{2} k q^2 + u(q^1, q^2, q^3) , \tag{6.45}$$

one easily arrives at the completely gauge-reduced solution

$$r_1(q) = \tfrac{1}{2} k q^2 , \quad r_2(q) = -\tfrac{1}{2} k q^1 , \quad r_3(q) = 0 . \tag{6.46}$$

Hence the associated left exponent generators are such that $l_a(q) = r_a(\bar{q})$ holds, and we have a solution within the gauge $\mu(q) = 0$.

In this manner, the differential equations (from the right) for the function $\phi_k(q';q)$ become [after substitution of the first two equations into the third, cf. Eq. (3.4), with $a = 1,2,3$]

$$\partial'_1 \phi_k(q';q) = \tfrac{1}{2} k \lambda(q'^3)(q^2 - q'^3 q^1) ,$$
$$\partial'_2 \phi_k(q';q) = -\tfrac{1}{2} k \lambda(q'^3)(q'^3 q^2) ,$$
$$\partial'_3 \phi_k(q';q) = \tfrac{1}{2} k\, [\lambda(q'^3)]^3 (q^2 - q'^3 q^1) q'^2 \tag{6.47}$$
$$- \tfrac{1}{2} k\, [\lambda(q'^3)]^3 (q^1 q'^3 q^2) q'^1 ,$$

which one readily integrates to read

$$\phi_k(q';q) = \tfrac{1}{2} k(\bar{q}'^2 q^1 - \bar{q}'^1 q^2) . \tag{6.48}$$

Clearly, $\phi_k(q;\bar{q}) = \phi_k(\bar{q};q) = 0$. Moreover, one easily checks that this is an admissible *nontrivial* two-cocycle indeed.

ACKNOWLEDGMENTS

Part of this work was done while visiting the Institute des Hautes Etudes Scientifiques (Bures-sur-Yvette, France). The author is grateful to Louis Michel for his kind hospitality and encouragement. He is also indepted to Ninoslav Bralić for a critical reading of the manuscript and invaluable discussions.

A travel grant from PNUD (UNESCO) and partial support by FNDCYT(1106/85), by DIUC(21/85), and by DGI-UCV (123.746/85) are hereby acknowledged.

APPENDIX: AN INTRODUCTION TO NON-ABELIAN CALCULUS

For the sake of completeness here we append some (not as well-known albeit elementary) features of non-Abelian calculus used in this paper. We shall develop this issue in a rather sketchy fashion. Here G denotes an r-parameter connected Lie group and $M(G) = \{q = (q^1,...,q^r)\}$ denotes the group manifold. Hence $M(G)$ is endowed with r group-multiplication functions,

$$q''^a = g^a(q'^1,...,q'^r;q^1,...,q^r) , \tag{A1}$$

with $a = 1,...,r$, which realize the binary composition law of the essential parameters of the group.[20] Thus one has

$$(q';q) \in M(G) \times M(G) \to g(q';q) \in M(G) , \tag{A2}$$
$$g(q;e) = g(e;q) = q , \tag{A3}$$
$$g(q;\bar{q}) = g(\bar{q};q) = e , \tag{A4}$$
$$g[q'';g(q';q)] = g[g(q'';q');q] , \tag{A5}$$

the meaning of which is clear.

Now, let us define the following functions:

$$R^a_b(q) = D_b(q') g^a(q';q) \equiv \lim_{q' \to e} \partial'_b g^a(q';q) , \tag{A6}$$

$$L^a_b(q) = D_b(q') g^a(q;q') \equiv \lim_{q' \to e} \partial'_b g^a(q;q') . \tag{A7}$$

[Bear in mind that under the action of *one* of the operators

$$D_a(q) = \lim_{q \to e} \partial_a , \tag{A8}$$

$a = 1,...,r$, *all* the corresponding variables $q = (q^1,...,q^r)$ become ignorable in the result.] Because of well-known geometric reasons, one refers to the functions $R^a_b(q)$ and $L^a_b(q)$, $a,b = 1,...,r$, as the (elements of the) *right* and *left transport matrices* in $M(G)$, respectively. One then introduces Lie's (right and left) infinitesimal operators on $M(G)$; namely,

$$X_a(q) = R^b_a(q) \partial_b , \tag{A9}$$
$$Y_a(q) = L^b_a(q) \partial_b . \tag{A10}$$

Of course, one has $R^a_b(e) = L^a_b(e) = \delta^a_b$.

Next, we present a series of interesting lemmas. However, we omit the proofs, for the sake of briefness.

Lemma A.1: For any given Lie group, one has

$$X_b(q') g^a(q';q) = R^a_b[g(q';q)] , \tag{A11}$$
$$Y_b(q) g^a(q';q) = L^a_b[g(q';q)] , \tag{A12}$$
$$X_b(q) g^a(q';q) = Y_b(q') g^a(q';q) . \tag{A13}$$

We observe that (A11) and (A12) are nontrivial generalizations of the trivial relations $X_b(q) q^a = R^a_b(q)$ and $Y_b(q) q^a = L^a_b(q)$. These are useful results, indeed. By means of these formulas one proves immediately the following lemma.

Lemma A.II: For any differentiable function $\psi(q)$ defined on $M(G)$, one has

$$X_a[g(q';q)]\psi[g(q';q)] = X_a(q')\psi[(q';q)], \quad \text{(A14)}$$

$$Y_a[g(q';q)]\psi[g(q';q)] = Y_a(q)\psi[g(q';q)], \quad \text{(A15)}$$

$$X_a(q)\psi[g(q';q)] = Y_a(q')\psi[g(q';q)]. \quad \text{(A16)}$$

Recall that $g(q';q) = q'' \in M(G)$, so these equations are well posed indeed. We like to remark that (A14)–(A16) entail the non-Abelian generalizations of the corresponding elementary results of ordinary (i.e., Abelian) calculus. We shall also present a set of converse relations to these equations.

Furthermore, Lemma A.I permits us to derive easily the Lie algebra of the set of operators $\{X_a(q), Y_b(q);\ a,b = 1,...,r\}$.

Lemma A.III: The infinitesimal operators $X_a(q)$ and $Y_a(q)$ obey the following algebra:

$$[X_a(q), X_b(q)] = f^c_{ab} X_c(q), \quad \text{(A17)}$$

$$[Y_a(q), Y_b(q)] = -f^c_{ab} Y_c(g), \quad \text{(A18)}$$

$$[X_a(q), Y_b(q)] = 0, \quad \text{(A19)}$$

where the structure constants are given by

$$f^c_{ab} = \sigma^c_{ba} - \sigma^c_{ab}, \quad \text{(A20)}$$

with

$$\sigma^a_{bc} = R^{\ a}_{b,c}(e) = L^{\ a}_{c,b}(e) = D_c(q)R^{\ a}_{\ b}(q) = D_b(q)L^{\ a}_{\ c}(q). \quad \text{(A21)}$$

One also needs to consider matrices performing *inverse* transport on $M(G)$. These matrices are defined as follows:

$$\overline{R}^{\ a}_{\ b}(q) = \lim_{q' \to q} \partial'_b g^a(q';\bar{q}), \quad \text{(A22)}$$

$$\overline{L}^{\ a}_{\ b}(q) = \lim_{q' \to q} \partial'_b g^a(\bar{q};q'); \quad \text{(A23)}$$

so that

$$\overline{R}^{\ a}_{\ c}(q)R^{\ c}_{\ b}(q) = R^{\ a}_{\ c}(q)\overline{R}^{\ c}_{\ b}(q) = \delta^a_b, \quad \text{(A24)}$$

$$\overline{L}^{\ a}_{\ c}(q)L^{\ c}_{\ b}(q) = L^{\ a}_{\ c}(q)\overline{L}^{\ c}_{\ b}(q) = \delta^a_b. \quad \text{(A25)}$$

Lemma A.IV: For any differentiable function $\psi(q)$, the following relations hold:

$$X_a(q)\psi(\bar{q}) = -Y_a(q)\psi(q), \quad \text{(A26)}$$

$$Y_a(q)\psi(\bar{q}) = -X_a(\bar{q})\psi(\bar{q}). \quad \text{(A27)}$$

[To prove this lemma, consider the middle-point cancellation relations

$$g[g(q'';q);g(\bar{q};q')] = g(q'';q'), \quad \text{(A28a)}$$

$$g[g(q'';\bar{q});g(q;q')] = g(q'';q'), \quad \text{(A28b)}$$

which are rather obvious and may be proved easily. Then, by taking the limits (1) $\lim_{q'' \to e} \lim_{q' \to e} \partial_b$ in (A28a) and (2) $\lim_{q'' \to e} \lim_{q' \to e} \partial_b$ in (A28b), after some manipulations, one gets

$$\frac{\partial \bar{q}^a}{\partial q^b} = -L^{\ a}_{\ c}(\bar{q})\overline{R}^{\ c}_{\ b}(q) = -R^{\ a}_{\ c}(\bar{q})\overline{L}^{\ c}_{\ b}(q), \quad \text{(A29)}$$

from which Eqs. (A26) and (A27) follow.]

The result stated in Eq. (A29) is useful because it entails two explicit relations between the one-to-one related variables $\bar{q}$ and q. Let us also observe that from (A29) we obtain

$$\lim_{q \to e} \frac{\partial \bar{q}^a}{\partial q^b} = \delta^a_b, \quad \text{(A30)}$$

as it should be indeed, since $\overline{e + \delta q} = e - \delta q + O(2)$.

As a remark (referred to in the context of this paper), we present a trivial feature.

Lemma A.V: The solution of the homogeneous system of first-order linear differential equations,

$$X_a(q')\phi(q';q) = 0, \quad Y_a(q)\phi(q';q) = 0, \quad \text{(A31)}$$

$a = 1,...,r$, with the initial conditions

$$\phi(e;q) = \phi(q;e) = 0, \quad \text{(A32)}$$

for all $q \in M(G)$, is

$$\phi(q';q) \equiv 0 \quad \text{(A33)}$$

(which is obvious, indeed).

Lemma A.VI: One has

$$\det\{[X_a(q') + X_a(q)]g^b(q';q)\} \neq 0, \quad \text{(A34)}$$

$$\det\{[Y_a(q') + Y_a(q)]g^b(q';q)\} \neq 0, \quad \text{(A35)}$$

for all points q' and q in $M(G)$.

(Recall that G is connected.) Formally, the proof is an immediate consequence of Lemma A.I.

Now, let us define the following auxiliary functions on the group manifold:

$$h^{\ a}(q) = g^a(q;q), \quad \text{(A36)}$$

for $a = 1,...,r$, and for all $q \in M(G)$. Clearly, $h^{\ a}(q) \in M(G)$. It can be shown that $\det[\partial_b h^{\ a}(q)] \neq 0$ holds everywhere. So we write

$$q'^a = h^{\ a}(q) \leftrightarrow q^a = \bar{h}^{\ a}(q') \quad \text{(A37)}$$

(say). With this construct one is in position to prove the following lemma, which is tantamount to the converse of Lemma A.II.

Lemma A.VII: If either

$$[X_a(q) - Y_a(q')]\phi(q';q) = 0 \quad \text{(A38)}$$

or

$$\{X_a[g(q';q)] - X_a(q')\}\phi(q';q) = 0 \quad \text{(A39)}$$

or

$$\{X_a[g(q';q)] - Y_a(q)\}\phi(q';q) = 0 \quad \text{(A40)}$$

holds for all $q', q \in M(G)$, then there exists a function $\psi(q)$ such that

$$\phi(q';q) = \psi[g(q';q)]. \quad \text{(A41)}$$

To prove this assertion, observe that

$$\left.\begin{aligned} &\det[X_a(q')g^b(q';q)] \\ &\det[Y_a(q)g^b(q';q)] \\ &\det[X_a(q)g^b(q';q)] \\ &\det[Y_a(q')g^b(q';q)] \end{aligned}\right\} \neq 0 \quad \text{(A42)}$$

hold for all q' and q. Of course, Lemma A.VII corresponds precisely to the elementary implication

$$(\partial_a - \partial'_a)\phi(q';q) = 0 \rightarrow \phi(q';q) = \psi(q' + q),$$

valid in the Abelian case. In this sense, (A41) represents the

most *general integral* of either of Eqs. (A38), (A39), or (A40).

We next prove the following important lemma.

Lemma A.VIII: The general solutions of the homogeneous *non-Abelian curl* equations,

$$X_a(q)r_b(q) - X_b(q)r_a(q) - f^c_{ab}r_c(q) = 0\,, \quad \text{(A43)}$$

$$Y_a(q)l_b(q) - Y_b(q)l_a(q) + f^c_{ab}l_c(q) = 0\,, \quad \text{(A44)}$$

are

$$r_a(q) = X_a(q)\gamma(q)\,, \quad \text{(A45)}$$

$$l_a(q) = Y_a(q)\Gamma(q)\,, \quad \text{(A46)}$$

where $\gamma(q)$ remains arbitrary.

Proof: That (A45) and (A46) imply (A43) and (A44), respectively, is trivial. Thus let us assume (A43), and define new auxiliary functions

$$u_a(q) = \overline{R}^{\,b}_{\,a}(q)r_b(q)\,.$$

Then, because of the Lie algebra (A17), (A43) becomes

$$\begin{aligned} &X_a(q)R^{\,c}_{\,b}(q)u_c(q) - X_b(q)R^{\,c}_{\,a}(q)u_c(q) \\ &\quad = u_c(q)\left[X_a(q)R^{\,c}_{\,b}(q) - X_b(q)R^{\,c}_{\,a}(q)\right] \\ &\qquad + \left[R^{\,c}_{\,b}(q)R^{\,d}_{\,a}(q) + R^{\,c}_{\,b}(q)R^{\,d}_{\,a}(q)\right] \\ &\qquad \times \left[u_{c,d}(q) - u_{d,c}(q)\right] = f^c_{ab}R^{\,c}_{\,d}(q)u_c(q)\,. \end{aligned}$$

Thus one has

$$u_{a,b}(q) - u_{b,a}(q) = 0\,,$$

which yields $u_a(q) = \gamma_{,a}(q)$. Hence (A45) follows. One obtains (A46) from (A44) in a similar way. Thus one proves the lemma.

Let us observe that if one uses the *same* scalar field $\gamma(q)$ in the solutions (A45) and (A46), then one gets

$$Y_b(q)r_a(q) = X_a(q)l_b(q)\,, \quad \text{(A47)}$$

as a trivial consequence of (A19). Conversely (and finally), we assert the following lemma.

Lemma A.IX: If the fields $r_a(q)$ satisfy (A43) and one defines the fields $l_a(q)$ by means of (A47), then these satisfy

$$Y_a(q)l_b(q) - Y_b(q)l_a(q) + f^c_{ab}l_c(q) = C_{ab}\,, \quad \text{(A48)}$$

where the C_{ab}'s are constants given by

$$C_{ab} = f^c_{ab}\left[l_c(e) - r_c(e)\right]\,. \quad \text{(A49)}$$

[1] H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1931).

[2] An interesting introduction to the subject of group extensions in quantum mechanics can be found in L. Michel, "Invariance in quantum mechanics and group extension," in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964).

[3] V. Bargmann, Ann. Math. **59**, 1 (1954).

[4] J. M. Lévy-Leblond, Riv. Nuovo Ciomento **4**, 99 (1974).

[5] Compare S. Eilenberg and S. Maclane, Ann. Math. **48**, 426 57 (1947); S. Eilenberg, Bull. Am. Math. Soc. **55**, 3 (1949).

[6] A detailed treatment of the necessary mathematical apparatus may be found in Ref. 3; cf. also L. Michel, in *Applications of Mathematics to Problems in Theoretical Physics*, edited by F. Lurcat (Gordon and Breach, New York, 1967), p. 135, and other references therein.

[7] J. C. Houard, J. Math. Phys. **18**, 502 (1977).

[8] S. Helgason, *Differential Geometry and Symmetric Spaces* (Academic, New York, 1962).

[9] J. C. Lévy-Leblond, Commun. Math. Phys. **12**, 64 (1969); Am. J. Phys. **39**, 502 (1971).

[10] E. P. Wigner, Ann. Math. **40**, 149 (1939).

[11] R. Jackiw, "Quantum mechanical symmetry breaking," in *Recent Developments in Quantum Field Theory*, edited by J. Ambjorn *et al.* (Elsevier, New York, 1985).

[12] T. Fujiwara, Phys. Lett. **152**, 103 (1985).

[13] Y. S. Wu, Phys. Lett. B **153**, 70 (1985).

[14] J. Krause, J. Phys. A **18**, 1309 (1985).

[15] J. Krause, J. Math. Phys. **27**, 2922 (1986).

[16] J. M. Lévy-Leblond, "Galilei group and Galilean invariance," in *Group Theory and its Applications*, edited by E. M. Loebl (Academic, New York, 1971), Vol. II; see, also, Ref. 3.

[17] J. Krause, "Galilean quantum kinematics," preprint.

[18] Compare this to Ref. 11.

[19] Y. S. Wu and A. Zee, Phys. Lett. **152**, 98 (1985); B. Grossman, Phys. Lett. **152**, 93 (1985). In a similar trend of ideas, cf., also, K. S. Cheng, J. Math. Phys. **18**, 746 (1977) (no explicit use of cocycle calculus is made in this reference, however).

[20] G. Racah, Ergeb. Exakten Naturwiss. **37**, 28 (1965).

[21] E. P. Wigner, *Group Theory* (Academic, New York, 1959).

[22] K. Iwasawa, Ann. Math. **50**, 507 (1949).

[23] The *ray constants* k_{ab} are precisely the real numbers defining the extension $L_k(G)$ of the Lie algebra $L(G)$ of G (i.e., *the two-cocycles of the Lie algebra*). See Ref. 16 for an explanation of the constraints stated in Eqs. (3.14) and (3.15).

[24] H. J. Bernstein, J. Math. Phys. **8**, 406 (1967).

[25] J. Voisin, J. Math. Phys. **6**, 1519 (1965).

[26] M. Hammermesh, Ann. Phys. (NY) **9**, 518 (1960).

[27] A. G. Nikitin and V. I. Fushchich, Teor. Mat. Fiz. **44**, 34 (1980).

[28] E. Inönü and E. P. Wigner, Nuovo Cimento **9** (8), 705 (1952).

[29] J. M. Lévy-Leblond, J. Math. Phys. **4**, 776 (1963); Commun. Math. Phys. **6**, 286 (1967).

[30] J. M. Lévy-Leblond, Commun. Math. Phys. **12**, 76 (1969).

Finite-dimensional representations of the special linear Lie superalgebra sl(1,n). I. Typical representations

Tchavdar D. Palev[a)]
Arnold Sommerfeld Institut für Mathematische Physik, Technische Universität Clausthal, D-3392 Clausthal-Zellerfeld, West Germany

(Received 2 December 1986; accepted for publication 1 April 1987)

In a series of two papers all finite-dimensional irreducible representations of the special linear Lie superalgebra sl(1,n) are written down in a matrix form. This paper develops a background for constructing the representations. Expressions for the transformation of the basis under the action of the generators are given for all induced and, hence, for all typical sl(1,n) modules.

I. INTRODUCTION

In this paper and the one that follows[1] we study all finite-dimensional irreducible representations of the special linear Lie superalgebra (LS) sl(1,n) for any $n = 2,3,\ldots$. We consider sl(1,n) [$= A(0,n-1)$ in the notation of Ref. 2] as a subalgebra of the general linear Lie superalgebra gl(1,n). The latter consists of all squared $(n+1)$-dimensional matrices. We label the rows and the columns of these matrices with indices $A,B,C,D,\ldots = 0,1,2,\ldots,n$. Assign to each index A a degree (A), which is zero for $A = 0$ and 1 for $A = 1,\ldots,n$. Let $e_{AB} \in \mathrm{gl}(1,n)$ be a matrix with 1 on the A th row and the B th column and zero elsewhere. The even (resp. the odd) part of gl(1,n) is defined to be the linear envelope of all matrices e_{AB}, for which $(A)+(B)$ is an even (resp. an odd) number. The multiplication (= the supercommutator) [,] on gl(1,n) is given with the linear extension of the relations

$$[e_{AB},e_{CD}] = \delta_{BC}e_{AD} - (-1)^{[(A)+(B)][(C)+(D)]}\delta_{AD}e_{CD}. \quad (1.1)$$

The LS sl(1,n) is a subalgebra of gl(1,n) consisting of all those matrices $a \in \mathrm{gl}(1,n)$, whose supertrace (= str) vanishes, i.e.,

$$\mathrm{sl}(1,n) = \left\{ a \,\middle|\, a \in \mathrm{gl}(1,n),\ \mathrm{str}(a) = \sum_{A=0}^{n} (-1)^{(A)} a_{AA} = 0 \right\}. \quad (1.2)$$

The even subalgebra

$$\mathrm{sl}(1,n)_0 = \mathrm{lin\ env}\{E_{ij} \mid E_{ij} = e_{ij} + \delta_{ij}e_{00},\ i,j = 1,\ldots,n\} \quad (1.3)$$

is isomorphic to the general linear Lie algebra gl(n). In this case the E_{ij} are the Weyl generators of gl(n),

$$[E_{ij},E_{kl}] = \delta_{jk}E_{il} - \delta_{li}E_{kj},\quad i,j,k,l = 1,\ldots,n. \quad (1.4)$$

The algebras sl(1,n), $n = 2,3,\ldots$, belong to the class of the basic Lie superalgebras (LS's) in the classification of Kac,[2] i.e., each sl(1,n) (1) is simple, (2) has a reductive even subalgebra and (3) has a nondegenerate Killing form. All simple Lie algebras are basic Lie superalgebras. The basic LS's, which are not Lie algebras, resolve into four countable classes [$A(m,n)$, $B(m,n)$, $C(n)$, and $D(m,n)$, $m,n = 1,2,\ldots$], one continuous class of 17-dimensional algebras $D(2,1;\alpha)$, and two exceptional LS's $G(3)$ and $F(4)$.

The structure of the basic LS's resembles in many respects the structure of the simple Lie algebras. Every such algebra A can be represented as a direct-space sum $A = N^- \oplus H \oplus N^+$ of its Cartan subalgebra H, which is the Cartan subalgebra of the even part, and the subalgebras N^- and N^+ spanned on the negative and the positive root vectors, respectively. The root vectors e_α are in one-to-one correspondence with their roots α, which are elements from the space of all linear functionals (the dual space) $\overset{*}{H}$ over H. The correspondence $e_\alpha \leftrightarrow \alpha$ is determined from the relation

$$[h,e_\alpha] = \alpha(h)e_\alpha,\quad \alpha \in \overset{*}{H}. \quad (1.5)$$

One can always choose a canonical system of $3r$ elements in A ($r = \dim H$)

$$e_i,h_i,f_i,\quad i = 1,\ldots,r, \quad (1.6)$$

which generate the algebra and have the following properties: (a) $h_1,\ldots,h_r$ constitute a basis in H; (b) $e_i \in N^+$ and $f_i \in N^-$ are positive and negative root vectors, respectively; and (c) the generators (1.6) satisfy the relations

$$[e_i,f_j] = \delta_{ij}h_i,\quad [h_i,e_j] = \alpha_{ij}e_j,\quad [h_i,f_j] = -\alpha_{ij}f_j, \quad (1.7)$$

where $\alpha_{ii} = 0$ or 2, $i = 1,\ldots,r$, and, if $\alpha_{ii} = 0$, then the first nonzero element among $\alpha_{i,i+k}$, $k = 1,2,\ldots$, is 1.

The matrix $\alpha = (\alpha_{ij})$ is called a Cartan matrix of A. Up to an isomorphism the algebra A is characterized by its Kac–Dynkin diagram. The latter consists of r white, gray, and black nodes, denoted as ○, ⊗, and ●, respectively. The ith node is white, if e_i is an even element, and gray or black, if e_i is an odd element and $\alpha_{ii} = 0$ or 2, respectively. The ith and the jth nodes are joined by $|\alpha_{ij}\alpha_{ji}|$ lines [except for $D(2,1;\alpha)$].

The Cartan subalgebra H of sl(1,n), which is a Cartan subalgebra of gl(n), is a linear span of the generators $E_{11},\ldots,E_{nn}$ [see (1.3)],

$$H = \mathrm{lin\ env}\{E_{ii} \mid i = 1,\ldots,n\}. \quad (1.8)$$

We choose $E_{11},\ldots,E_{nn}$ as a basis in H and denote by $E^1,\ldots,E^n$ the dual to it basis in $\overset{*}{H}$, i.e.,

$$E^i(E_{jj}) = \delta^i_j,\quad E^i \in \overset{*}{H},\quad i,j = 1,\ldots,n. \quad (1.9)$$

As usual, we accept a lexical ordering in $\overset{*}{H}$, assuming that $\lambda = \sum_{i=1}^{n} \lambda_i E^i > 0$, if the first nonzero coordinate of this

[a)] Present address: Institute of Nuclear Research and Nuclear Energy, blvd. Lenin 72, 1184 Sofia, Bulgaria.

 0022-2488/87/102280-24$02.50 © 1987 American Institute of Physics

functional is positive, i.e.,

$$\lambda > 0, \text{ if } \lambda_1 = \lambda_2 = \cdots = \lambda_{k-1} = 0, \quad \lambda_k > 0. \quad (1.10)$$

Then for any two functionals $\lambda', \lambda'' \in \overset{*}{H}$ one defines

$$\lambda' > \lambda'', \text{ if } \lambda' - \lambda'' > 0. \quad (1.11)$$

Very often we identify λ with its coordinates, i.e., we set

$$\lambda = \sum_{i=1}^{n} \lambda_i E^i \equiv [\lambda_1, \lambda_2, \ldots, \lambda_n]. \quad (1.12)$$

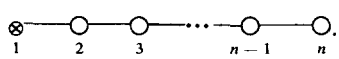
We say that $\lambda = [\lambda_1, \ldots, \lambda_n]$ is lexical if

$$\lambda_i - \lambda_{i+1} \geqslant 0 \quad \forall i = 1, \ldots, n-1. \quad (1.13)$$

From (1.5) one infers that the set of all $e_{AB}, A < B = 0,1,\ldots,n$ (resp. $A > B = 0,1,\ldots,n$) gives the positive (resp. the negative) root vectors of sl$(1,n)$. The canonical system of generators (1.6) reads

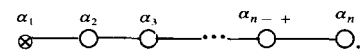
$$\begin{array}{lll} h_1 = E_{11}, & e_1 = e_{01}, & f_1 = e_{10}, \\ h_2 = E_{11} - E_{22}, & e_2 = e_{12}, & f_2 = e_{21}, \\ \vdots & \vdots & \vdots \\ h_n = E_{n-1,n-1} - E_{nn}, & e_n = e_{n-1,n}, & f_n = e_{n,n-1}, \end{array} \quad (1.14)$$

with e_1 and f_1 being the only odd generators in it. Since $[[h_1, e_1]] = 0$, i.e., $\alpha_{11} = 0$, the Kac–Dynkin diagram contains one gray and $n-1$ white nodes:



(1.15)

The structure of the finite-dimensional modules over a given basic LS is illustrated in Fig. 1. Apart from the algebras $B(0,n)$,[3] every basic LS has indecomposible (i.e., nonfully-reducible) finite-dimensional modules. Several examples of such modules are available. However, at present it is not known how to construct all indecomposible representations. In contrast to this, all finite-dimensional irreducible modules (fidirmods) are fully classified.[4] Every such fidirmod $W(\Lambda)$ is characterized by its highest weight $\Lambda \in \overset{*}{H}$. In particular, let $\overline{W}(\Lambda)$ be the module over the basic LS A, induced from the fidirmod $V_0(\Lambda)$ of the even subalgebra $A_0 \subset A$ [see Ref. 4, for sl$(1,n)$ see also Sec. II C]. Then either (1) $\overline{W}(\Lambda)$ is a fidirmod of A or (2) $\overline{W}(\Lambda)$ is an indecomposible A module, but the factor module $\overline{W}(\Lambda)/\overline{I}(\Lambda)$ with respect to the maximal (nontrivial) invariant subspace $\overline{I}(\Lambda)$ is a fidirmod. In case (1) the fidirmod $W(\Lambda) = \overline{W}(\Lambda)$ and also the representation of A in $W(\Lambda)$ is called typical. In case (2) the fidirmod $W(\Lambda) = \overline{W}(\Lambda)/\overline{I}(\Lambda)$ (and the corresponding representation) is said to be nontypical. It is remarkable that each fidirmod of the basic LS A can be constructed in this way, i.e., it is either typical or nontypical. In Ref. 4 a given fidirmod $W(\Lambda)$ is characterized by the coordinates $\alpha_i = \Lambda(h_i)$, $i = 1,\ldots,r$, of its highest weight Λ in the dual to $h_1,\ldots,h_r$ basis $h^1,\ldots,h^r$ in $\overset{*}{H}$. To visualize the A fidirmod $W(\Lambda)$ one writes above each, say the ith node $(i = 1,\ldots,r)$ of the Kac–Dynkin diagram the coordinate α_i of Λ. For instance, the sl$(1,n)$ fidirmod $W(\Lambda)$, $\Lambda = \Sigma_{i=1}^{n} \alpha_i h^i$, is denoted as



(1.16)

The method of induced representations describes all typical representations and in principle shows how one can proceed to construct the nontypical modules. In this way[5] and in other publications[6–8] several properties of the finite-dimensional irreducible representations have been established. Of interest in this respect is the generalization of the Young tableaux technique to the case of LS's.[9] Irrespective of the progress, the representation theory of the basic Lie superalgebras is still far from being complete. It is, in fact, much less worked out in comparison with the corresponding level of development of the simple Lie algebras. In particular, from a physical point of view, the important problem of computing the matrix elements of the generators within an arbitrary fidirmod has been solved so far only for the low rank Lie superalgebras osp(2,1), sl(1,2),[7] osp(3,2),[10] and sl(1,3).[11] In the present paper and in Ref. 1 we take a further step towards the solution of the general representation problem. We consider all fidirmods of the Lie superalgebra sl$(1,n)$ for any $n = 2,3,\ldots$, introduce appropriate basis, and write down expressions for the transformation of the basis under the action of the sl$(1,n)$ generators.

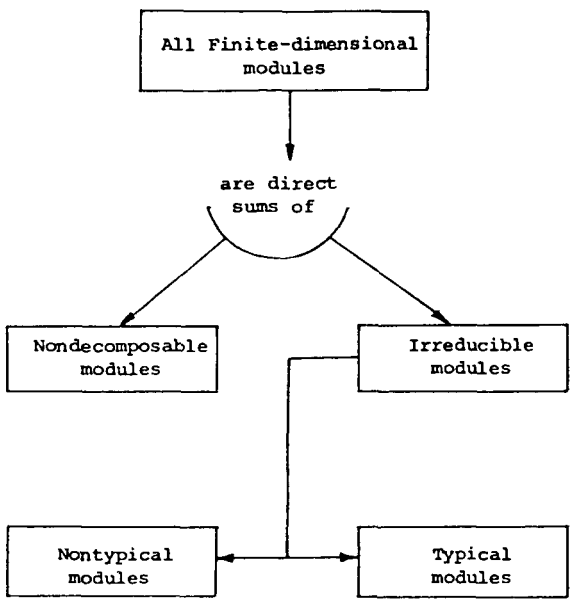


FIG. 1. The structure of the finite-dimensional modules over a given basic LS.

II. INDUCED REPRESENTATIONS OF sl(1,n)

A. Abbreviations, notation, terminology

We list here some of the abbreviations and the notation that will be used throughout the paper:

LS, LS's—Lie superalgebra, Lie superalgebras,

LA, LA's—Lie algebra, Lie algebras,

Fidirmod(s)—finite-dimensional irreducible module(s),

lin env(X)—the linear envelope of X,

$\mathbb{C}$—the complex numbers,

$\mathbb{Z}$—all integers,

$\mathbb{Z}_+$—all non-negative integers,

$\mathbb{N}$—all positive integers,

GZ basis—Gel'fand–Zetlin basis [see (2.13)],

I-basis—induced basis [see (2.54)],
$[\,,\,]$—product (supercommutator) in the LS,
$[x,y] = xy - yx$,
$\{x,y\} = xy + yx$.
Let $m_{ij} \in \mathbb{C}$. Then we set

$$[m_{1,n+1},m_{2,n+1},\ldots,m_{n,n+1}] = [m]_{n+1}, \tag{2.1}$$

$$[m_{1k},m_{2k},\ldots,m_{kk}] = [m]_k, \quad k = 1,\ldots,n, \tag{2.2}$$

$$[m_{1k}+c,m_{2k}+c,\ldots,m_{kk}+c] = [m+c]_k, \quad c\in\mathbb{C}, \tag{2.3}$$

$$[m_{1k} \pm \delta_{1i},\ldots,m_{kk} \pm \delta_{ki}] = [m]_k^{\pm i}, \tag{2.4}$$

$$[m_{1k} + \xi\delta_{1i} + \eta\delta_{2j}, m_{2k} + \xi\delta_{2i} + \eta\delta_{2j},\ldots,m_{kk} + \xi\delta_{ki} + \eta\delta_{kj}] = [m]_k^{\xi i,\eta j},$$

$$\xi,\eta = 0,\pm 1,\ i,j\in(1,2,\ldots,k), \tag{2.5}$$

$$l_{ij} = m_{ij} - i, \tag{2.6}$$

$$S(i,j) = \begin{cases} 1, & \text{for } i \leqslant j, \\ -1, & \text{for } i > j. \end{cases} \tag{2.7}$$

Definition 1: A sequence of n numbers, which are either 0 or 1, will be called a θ-tuple.

For any such θ-tuple we use three different notations,

$$\{\theta\}_n = \{\theta_1,\ldots,\theta_n\} = (i_1,\ldots,i_N), \quad \theta_1,\ldots,\theta_n = 0,1, \tag{2.8}$$

where $(i_1,\ldots,i_N)$ is the subset of $(1,2,\ldots,n)$, consisting of all those $k\in(1,2,\ldots,n)$, for which $\theta_k = 1$, i.e.,

$$\begin{aligned} &\theta_k = 1, \quad \text{if } k\in(i_1,\ldots,i_N), \\ &\theta_k = 0, \quad \text{if } k\bar{\in}(i_1,\ldots,i_N). \end{aligned} \tag{2.9}$$

Definition 2: The θ-tuple $\{\theta\}_n$ is said to be of degree N, if $\theta_1 + \cdots + \theta_n = N$, i.e.,

$$\{\theta\}_n = (i_1,\ldots,i_N). \tag{2.10}$$

B. Fidirmods of gl(n)

Throughout the paper we use the Gel'fand and Zetlin notation for the fidirmods of the Lie algebra gl(n) (see Ref. 12), accepting also some abbreviations from Ref. 13. Every fidirmod of gl(n) is labeled by its signature

$$[m]_n = [m_{1n},m_{2n},\ldots,m_{nn}], \tag{2.11}$$

where $m_{1n},\ldots,m_{nn}$ are, in general, complex numbers such that

$$m_{in} - m_{jn}\in\mathbb{Z}_+ \quad \forall i<j = 1,\ldots,n. \tag{2.12}$$

Let $V([m]_n)$ be the fidirmod of gl(n) with a signature $[m]_n$. As a basis in $V([m]_n)$ we choose the Gel'fand–Zetlin basis (GZ basis),[12]

$$\left|\begin{matrix} m_{1n},m_{2n},\ldots,m_{nn} \\ \vdots \\ m_{1i},m_{2i},\ldots,m_{ii} \\ \vdots \\ m_{11} \end{matrix}\right\rangle \equiv \left|\begin{matrix} [m]_n \\ \vdots \\ [m]_i \\ \vdots \\ m_{11} \end{matrix}\right\rangle. \tag{2.13}$$

The numbers $m_{1n},\ldots,m_{nn}\in\mathbb{C}$ are fixed and label the representation space. The others $m_{ij}\in\mathbb{C}$ distinguish between the basis vectors in $V([m]_n)$ and take all possible values, consistent with the "betweenness" condition

$$m_{i,j+1} - m_{ij}\in\mathbb{Z}_+, \quad m_{ij} - m_{i+1,j+1}\in\mathbb{Z}_+. \tag{2.14}$$

The Cartan generators $E_{11},\ldots,E_{nn}$ are diagonal in this basis, i.e., the GZ basis consists of weight vectors. Since

$$E_{kk}\left|\begin{matrix} [m]_n \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ \vdots \\ m_{11} \end{matrix}\right\rangle = \left(\sum_{i=1}^{k} m_{ik} - \sum_{i=1}^{k-1} m_{i,k-1}\right)\left|\begin{matrix} [m]_n \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ \vdots \\ m_{11} \end{matrix}\right\rangle, \tag{2.15}$$

the correspondence weight vector → weight is

$$\left|\begin{matrix} [m]_n \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ \vdots \\ m_{11} \end{matrix}\right\rangle \to \sum_{k=1}^{n}\left(\sum_{i=1}^{k} m_{ik} - \sum_{i=1}^{k-1} m_{i,k-1}\right)E^k. \tag{2.16}$$

The highest weight vector x_Λ is the one from (2.13) for which

$$m_{ii} = m_{i,i+1} = \cdots = m_{in}, \quad i = 1,\ldots,n. \tag{2.17}$$

In this case (2.15) yields

$$E_{ii}x_\Lambda = m_{in}x_\Lambda, \quad i = 1,\ldots,n, \tag{2.18}$$

and, therefore, $m_{1n},\ldots,m_{nn}$ are the coordinates of the gl(n) highest weight Λ in the dual to $E_{11},\ldots,E_{nn}$ basis $E^1,\ldots,E^n$,

$$\Lambda = \sum_{i=1}^{n} m_{in}E^i \equiv [m_{1n},\ldots,m_{nn}] = [m]_n. \tag{2.19}$$

In other words, the signature (2.11) of the gl(n) fidirmod $V([m]_n)$ consists of the coordinates of the highest weight Λ in the basis $E^1,\ldots,E^n$.

C. Induced representations

We now proceed to introduce, following Kac,[4] the sl(1,n) module $\overline{W}([m]_{n+1})$, induced from the gl(n) fidirmod $V_0([m]_{n+1})$. We recall that [see (2.1)]

$$[m]_{n+1} = [m_{1,n+1},m_{2,n+1},\ldots,m_{n,n+1}]. \tag{2.20}$$

The coordinates of the gl(n) highest weight [see (2.19)]

$$\Lambda = \sum_{i=1}^{n} m_{i,n+1}E^i \tag{2.21}$$

in $V_0([m]_{n+1})$ satisfy (2.12), which in the present notation reads

$$m_{i,n+1} - m_{j,n+1}\in\mathbb{Z}_+ \quad \forall i<j = 1,2,\ldots,n. \tag{2.22}$$

Denote by P_+ the linear envelope of all odd positive root vectors of sl(1,n),

$$P_+ = \text{lin env}\{e_{0i} \mid i = 1,\ldots,n\}. \tag{2.23}$$

Let $P = \text{gl}(n) \oplus P_+$. Extend $V_0([m]_{n+1})$ to a P module, assuming

$$P_+V_0([m]_{n+1}) = 0. \tag{2.24}$$

The sl(1,n) module, induced from the gl(n) fidirmod

$V_0([m]_{n+1})$, is defined to be the factor space

$$\overline{W}([m]_{n+1}) = U \otimes V_0([m]_{n+1})/I([m]_{n+1}), \quad (2.25)$$

of the tensor product of the sl(1,n) universal enveloping algebra U with $V_0([m]_{n+1})$ and subsequently factorized by the subspace

$$I([m]_{n+1}) = \text{lin env}\{up \otimes v - u \otimes pv | u \in U, \quad p \in P \subset U, \quad v \in V_0([m]_{n+1})\}. \quad (2.26)$$

The linear space $\overline{W}([m]_{n+1})$ is equipped with a structure of an sl(1,n) module in a natural way:

$$g(u \otimes v) = gu \otimes v, \quad g \in \text{sl}(1,n), \quad u \otimes v \in \overline{W}([m]_{n+1}). \quad (2.27)$$

From the Poincaré–Birkhoff–Witt theorem[4] follows that U is a linear span of all elements of the form

$$g = (e_{10})^{\theta_1}(e_{20})^{\theta_2}\cdots(e_{n0})^{\theta_n}p, \quad \theta_1,\ldots,\theta_n = 0,1, \quad (2.28)$$

where p is a polynomial of elements from P. The restriction $\theta_i = 0,1$ comes from the observation that $(e_{i0})^2 = 0$ in U. Since for any g, defined in (2.28), and $v \in V_0([m]_{n+1})$,

$$g \otimes v = (e_{10})^{\theta_1}(e_{20})^{\theta_2}\cdots(e_{n0})^{\theta_n}p \otimes v = (e_{10})^{\theta_1}(e_{20})^{\theta_2}\cdots(e_{n0})^{\theta_n} \otimes pv, \quad (2.29)$$

one concludes that

$$\overline{W}([m]_{n+1}) = \text{lin env}\{(e_{10})^{\theta_1}\cdots(e_{n0})^{\theta_n} \otimes v | v \in V_0([m]_{n+1}), \quad \theta_1,\ldots,\theta_n = 0,1\}. \quad (2.30)$$

Let $x_\lambda \in V_0([m]_{n+1})$ be a gl(n) weight vector with a weight $\lambda \in \overset{*}{H}$, i.e., $E_{ii}x_\lambda = \lambda(E_{ii})x_\lambda$, $i = 1,\ldots,n$. Then

$$E_{ii}\left[(e_{10})^{\theta_1}\cdots(e_{n0})^{\theta_n} \otimes x_\lambda\right] = \left[\lambda(E_{ii}) - \sum_{k \neq i = 1}^{n} \theta_k\right](e_{10})^{\theta_1}\cdots(e_{n0})^{\theta_n} \otimes x_\lambda. \quad (2.31)$$

Therefore, for any $\theta_1,\ldots,\theta_n = 0,1$

$$(e_{10})^{\theta_1}\cdots(e_{n0})^{\theta_n} \otimes x_\lambda = x_{\lambda'} \in \overline{W}([m]_{n+1}) \quad (2.32)$$

is an sl(1,n) weight vector with a weight

$$\lambda' = \sum_{i=1}^{n}\left[\lambda(E_{ii}) - \sum_{k \neq i = 1}^{n} \theta_k\right]E^i \in \overset{*}{H}, \quad (2.33)$$

which, in the lexical ordering we have accepted [see (1.11)], is less than λ, $\lambda' < \lambda$. Clearly, the highest weight Λ [see (2.21)] of the gl(n) module $V_0([m]_{n+1})$ is also a highest weight of the sl(1,n) module $\overline{W}([m]_{n+1})$. Denoting by

$$\mathbf{1} = (e_{10})^0\cdots(e_{n0})^0 \quad (2.34)$$

the unity in U, one concludes that $\mathbf{1} \otimes x_\Lambda$ is the sl(1,n) highest weight vector in $\overline{W}([m]_{n+1})$, i.e.,

$$e_{0i}(\mathbf{1} \otimes x_\Lambda) = 0, \quad E_{jk}(\mathbf{1} \otimes x_\Lambda) = 0, \quad j < k = 1,\ldots,n, \quad (2.35)$$

$$E_{ii}(\mathbf{1} \otimes x_\Lambda) = m_{i,n+1}(\mathbf{1} \otimes x_\Lambda), \quad i = 1,\ldots,n.$$

Thus, to every gl(n) fidirmod $V_0([m]_{n+1})$ there corresponds an induced sl(1,n) module $\overline{W}([m]_{n+1})$. Both of them have the same highest weight $\Lambda = \Sigma_{i=1}^{n} m_{i,n+1}E^i$. Every induced module $\overline{W}([m]_{n+1})$ is either irreducible [i.e., it is an sl(1,n) fidirmod] or indecomposible. The representations of sl(1,n), realized in the irreducible induced modules (and also the modules themselves), are said to be typical.[4] Each $\overline{W}([m]_{n+1})$, which is not irreducible, contains a maximal sl(1,n) invariant subspace $\overline{I}([m]_{n+1}) \neq 0$. The factor module $\overline{W}([m]_{n+1})/\overline{I}([m]_{n+1})$ carries an irreducible representation of sl(1,n). All such factor modules (and also the corresponding representations) are called nontypical. Since $\mathbf{1} \otimes x_\Lambda \in \overline{I}([m]_{n+1})$, the highest weight of $\overline{W}([m]_{n+1})/\overline{I}([m]_{n+1})$ is the same as the highest weight of $V_0([m]_{n+1})$, i.e., $\Lambda = \Sigma_{i=1}^{n} m_{i,n+1}E^i$. Since, moreover, the typical and the nontypical representations exhaust all finite-dimensional irreducible representations of sl(1,n), we conclude that there exists one-to-one correspondence between the fidirmods of gl(n) and the fidirmods of sl(1,n), namely,

$$V_0([m]_{n+1}) \leftrightarrow \overline{W}([m]_{n+1})/\overline{I}([m]_{n+1}), \quad (2.36)$$

where in the typical modules we assume that $\overline{I}([m]_{n+1}) = 0$.

A convenient criterion for the irreducibility of $\overline{W}([m]_{n+1})$ has been proved in Ref. 14.

Proposition 1: The induced sl(1,n) module $\overline{W}([m]_{n+1})$ is typical if and only if

$$m_{k,n+1} \neq k - 1 \quad \forall k = 1,2,\ldots,n. \quad (2.37)$$

D. Induced basis

The main difficulty to overcome in constructing the nontypical modules is the determination of the maximal (nontrivial) invariant subspace $\overline{I}([m]_{n+1})$ of each $\overline{W}([m]_{n+1})$. To simplify the problem (which will be solved in Ref. 1) we introduce a basis in the induced modules in such a way that each basis vector is either from $\overline{I}([m]_{n+1})$ or is a vector from a complement to its subspace. To this end we decompose $\overline{W}([m]_{n+1})$ into a direct sum of gl(n) fidirmods $V(\lambda_i)$, where λ_i, $i = 1,\ldots,M$, is the highest weight of $V(\lambda_i)$,

$$\overline{W}([m]_{n+1}) = \sum_{i=1}^{M} \oplus V(\lambda_i), \quad (2.38)$$

and introduce a GZ basis Γ_i within each $V(\lambda_i)$. Then as a basis in $\overline{W}([m]_{n+1})$ we take

$$\Gamma = \bigcup_{i=1}^{M} \Gamma_i. \quad (2.39)$$

Since $\overline{I}([m]_{n+1})$ is a gl(n) submodule, it is a direct sum of some of the gl(n) fidirmods $V(\lambda_i)$ $i = 1,\ldots,M$,

$$\overline{I}([m]_{n+1}) = V(\lambda_{i_1}) \oplus V(\lambda_{i_2}) \oplus \cdots \oplus V(\lambda_{i_k}), \quad i_1 \neq \cdots \neq i_k \in (1,\ldots,M). \quad (2.40)$$

Clearly, each basis vector belongs either to $\overline{I}([m]_{n+1})$ or to the complementary space $\Sigma_j \oplus V(\lambda_j)$, $j \in (i_1,\ldots,i_k)$, i.e., the basis Γ possesses the required properties.

Let $T \subset U$ be the subalgebra spanned on all polynomials of the odd negative root vectors, i.e.,

$$T = \text{lin env}\{(e_{10})^{\theta_1}(e_{20})^{\theta_2}\cdots(e_{n0})^{\theta_n} | \theta_1,\ldots,\theta_n = 0,1\}. \quad (2.41)$$

The relation (2.30) shows that the sl(1,n) module $\overline{W}([m]_{n+1})$, considered as a linear space, is a tensor prod-

uct of T and $V_0([m]_{n+1})$,

$$\overline{W}([m]_{n+1}) = T \otimes V_0([m]_{n+1}). \tag{2.42}$$

Since $[\mathrm{gl}(n), T] \subset T$, T can be viewed as a $\mathrm{gl}(n)$ module. Moreover, for any $a \in \mathrm{gl}(n)$ and $t \otimes v \in T \otimes V_0([m]_{n+1})$,

$$a(t \otimes v) = [\mathrm{ad}(a)]t \otimes v + t \otimes av, \quad \mathrm{ad}(a)t = [a,t]. \tag{2.43}$$

Therefore, $\overline{W}([m]_{n+1})$ is a tensor product of the $\mathrm{gl}(n)$ modules T and $V_0([m]_{n+1})$.

Proposition 2: For any integer $0 \leqslant N \leqslant n$ the subspace

$$T_N = \mathrm{lin\ env}\left\{ (e_{10})^{\theta_1} \cdots (e_{n0})^{\theta_n} \,\middle|\, \sum_{i=1}^{n} \theta_i = N, \quad \theta_1,\ldots,\theta_n = 0,1 \right\} \tag{2.44}$$

is a $\mathrm{gl}(n)$ fidirmod with a highest weight

$$\lambda_N = \sum_{i=1}^{N} (1-N)E^i + \sum_{i=N+1}^{n} (-N)E^i. \tag{2.45}$$

Therefore, characterizing T_N with the coordinates of its highest weight, we write

$$T_N = V([\overset{1,\ldots,N,}{1-N,\ldots,1-N,} \overset{N+1,\ldots,n}{-N,\ldots,-N}]). \tag{2.46}$$

The proof is straightforward. The subspace T_0 is one-dimensional, spanned on the unity of U. The subspace $T_1 = \mathrm{lin\ env}\{e_{i0} \mid i = 1,\ldots,n\}$ is n dimensional; the correspondence root vector $\leftrightarrow$ root in this case is

$$e_{i0} \leftrightarrow [-1+\delta_{1i}, -1+\delta_{2i}, \ldots, -1+\delta_{ni}]. \tag{2.47}$$

Since $T = \Sigma_{N=0}^{n} \oplus T_N$, we have from (2.42)

$$\overline{W}([m]_{n+1}) = \sum_{N=0}^{n} \oplus [T_N \otimes V_0([m]_{n+1})]. \tag{2.48}$$

The decomposition of $T_N \otimes V_0([m]_{n+1})$ into $\mathrm{gl}(n)$ fidirmods is easily carried out (a useful prescription for such decompositions is given in Ref. 13):

$$T_N \otimes V_0([m]_{n+1}) = \sum_{\substack{\theta_1,\ldots,\theta_n = 0,1 \\ \theta_1 + \cdots + \theta_n = N}}{}' \oplus V([m_{1,n+1} + \theta_1 - N, \ldots, m_{n,n+1} + \theta_n - N]). \tag{2.49}$$

As usual, $[m_{1,n+1} + \theta_1 - N, \ldots, m_{n,n+1} + \theta_n - N]$ are the coordinates of the highest weight in the basis $E^1,\ldots,E^n$ [see (1.9)]. The prime on the sum in (2.49) means that one has to delete all nonlexical terms, i.e., those $V([m_{1,n+1} + \theta_1 - N, \ldots, m_{n,n+1} + \theta_n - N)$ for which

$$(m_{i+1,n+1} + \theta_{i+1}) - (m_{i,n+1} + \theta_i) > 0, \tag{2.50}$$

for certain $i = 1,2,\ldots,n-1$. Combining (2.48) and (2.49) we have the following proposition.

Proposition 3: The induced $\mathrm{sl}(1,n)$ module $\overline{W}([m]_{n+1})$ decomposes into a direct sum of $\mathrm{gl}(n)$ fidirmods as follows:

$$\overline{W}([m]_{n+1}) = \sum_{\theta_1,\ldots,\theta_n = 0,1}{}' \oplus V\left(\left[m_{1,n+1} + \theta_1 - \sum_{i=1}^{n} \theta_i, \ldots, m_{n,n+1} + \theta_n - \sum_{i=1}^{n} \theta_i\right]\right). \tag{2.51}$$

Observe that the signatures of the $\mathrm{gl}(n)$ fidirmods in the direct sum decomposition (2.51) are all different, i.e., $\overline{W}([m]_{n+1})$ is a simply reducible $\mathrm{gl}(n)$ module. Let

$$m_{in} = m_{i,n+1} + \theta_i - \sum_{k=1}^{n} \theta_k. \tag{2.52}$$

As a basis $\Gamma([m]_n)$ in each

$$V\left(\left[m_{1,n+1} + \theta_1 - \sum_{k=1}^{n} \theta_k, \ldots, m_{n,n+1} + \theta_n - \sum_{k=1}^{n} \theta_k\right]\right) = V([m_{1n},\ldots,m_{nn}]) \equiv V([m]_n), \tag{2.53}$$

we choose the Gel'fand–Zetlin basis (2.13). To indicate, however, that each such vector belongs to $\overline{W}([m]_{n+1})$ we modify the notation setting

$$\left|\begin{array}{c} [m]_n \\ \vdots \\ [m]_i \\ \vdots \\ m_{11} \end{array}\right\rangle = \begin{bmatrix} m_{1,n+1}, m_{2,n+1}, \ldots, m_{n,n+1} \\ m_{1n}, m_{2n}, \ldots, m_{nn} \\ \vdots \\ m_{1i}, m_{2i}, \ldots, m_{ii} \\ \vdots \\ m_{11} \end{bmatrix} \equiv \begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \\ [m]_i \\ \vdots \\ m_{11} \end{bmatrix}. \tag{2.54}$$

Then the system

$$\Gamma = \bigcup_{\theta_1,\ldots,\theta_n} \Gamma\left(\left[m_{1,n+1} + \theta_1 - \sum_{k=1}^{n} \theta_k, \ldots, m_{n,n+1} + \theta_n - \sum_{k=1}^{n} \theta_k\right]\right) \tag{2.55}$$

gives a basis in $\overline{W}([m]_{n+1})$. The union is over all those $\theta_1,\ldots,\theta_n = 0,1$ for which the lexical condition (2.50) holds. More precisely, we have the following proposition.

Proposition 4: The basis in the induced $\mathrm{sl}(1,n)$ module $\overline{W}([m]_{n+1})$ consists of all those patterns (2.54) for which the following conditions hold:

$$(1)\quad m_{in} = m_{i,n+1} + \theta_i - \sum_{k=1}^{n} \theta_k, \quad \theta_1,\theta_2,\ldots,\theta_n = 0,1, \tag{2.56}$$

$$(2)\quad m_{i,j+1} - m_{ij} \in \mathbb{Z}_+, \quad m_{ij} - m_{i+1,j+1} \in \mathbb{Z}_+, \quad i \leqslant j = 1,\ldots,n-1. \tag{2.57}$$

The so-defined basis in $\overline{W}([m]_{n+1})$ will be called an induced basis and each vector (2.54)—an I-pattern. This basis is an analog of the Gel'fand and Zetlin basis in the fidirmods of the classical Lie algebras. Indeed, consider the chain of subalgebras

$$\mathrm{sl}(1,n) \supset \mathrm{gl}(n) \supset \mathrm{gl}(n-1) \supset \cdots \supset \mathrm{gl}(2) \supset \mathrm{gl}(1) \tag{2.58}$$

and a flag of subspaces

$$\overline{W}([m]_{n+1}) \supset V([m]_n) \supset V([m]_{n-1}) \supset \cdots \supset V([m]_2) \supset V(m_{11}), \tag{2.59}$$

where, for any $k = 1,...,n$, $V([m]_k)$ is a gl(k) fidirmod with a signature $[m]_k$. Since dim $V(m_{11}) = 1$, the flag (2.59) determines a one-dimensional subspace, spanned on the I-pattern (2.54).

From (2.56) one concludes that

$$\sum_{k=1}^{n} \theta_k = \frac{1}{n-1} \sum_{k=1}^{n} (m_{k,n+1} - m_{kn}). \tag{2.60}$$

Moreover,

$$\theta_k = \frac{1}{n-1} \sum_{i=1}^{n} (m_{i,n+1} - m_{in}) - (m_{k,n+1} - m_{kn}), \quad k = 1,...,n. \tag{2.61}$$

Consider the θ-tuple $\{\theta_1,\theta_2,...,\theta_n\} = \{\theta\}_n$ determined by (2.61).

Proposition 5: The signature $[m]_n$ of the gl(n) fidirmod $V([m]_n) \subset \overline{W}([m]_{n+1})$ is uniquely defined [see (2.56)] and defines uniquely [see (2.61)] a θ-tuple,

$$[m_{1n},m_{2n},...,m_{nn}] \leftrightarrow \{\theta_1,\theta_2,...,\theta_n\}. \tag{2.62}$$

We use this correspondence to turn $\overline{W}([m]_{n+1})$ into $\mathbb{Z}$-graded linear space.

Definition 3: We say that the gl(n) fidirmod $V([m]_n) \subset \overline{W}([m]_{n+1})$ is of degree N, deg $V([m]_n) = N$, if the degree of its θ-tuple (Definition 2) is N, i.e., $\theta_1 + \cdots + \theta_n = N$ or [see (2.10)] $\{\theta\}_n = (i_1,...,i_N)$.

If the degree of an I-pattern (2.54) is N, then (2.60) and (2.61) yield

$$\theta_k = N - (m_{k,n+1} - m_{kn}), \quad k = 1,...,n. \tag{2.63}$$

III. TRANSFORMATION OF THE I-BASIS

A. Expressions for the even generators

By construction all I-patterns (2.54) with a fixed nth row $[m_{1n},...,m_{nn}] = [m]_n$ constitute a gl(n) Gel'fand–Zetlin basis $\Gamma([m]_n)$ in $V([m]_n) \subset \overline{W}([m]_{n+1})$. The action of the gl($n$) generators, i.e., the even generators E_{ij} [see (1.3)], on this basis is known.[12] In terms of our notations (2.54) it reads

$$E_{kk} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ \vdots \\ m_{11} \end{bmatrix} = (m_{1k} + \cdots + m_{kk} - m_{1,k-1} - \cdots - m_{k-1,k-1}) \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k+1} \\ \vdots \\ m_{11} \end{bmatrix}, \tag{3.1}$$

$$E_{k,k-1} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{j=1}^{k-1} \left| \frac{\prod_{i=1}^{k} (l_{ik} - l_{j,k-1} + 1) \prod_{i=1}^{k-2} (l_{i,k-2} - l_{j,k-1})}{\prod_{i\neq j=1}^{k-1} (l_{i,k-1} - l_{j,k-1} + 1)(l_{i,k-1} - l_{j,k-1})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1}^{-j} \\ [m]_{k-2} \\ \vdots \\ m_{11} \end{bmatrix}, \tag{3.2}$$

$$E_{k-1,k} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{j=1}^{k-1} \left| \frac{\prod_{i=1}^{k} (l_{ik} - l_{j,k-1}) \prod_{i=1}^{k-2} (l_{i,k-2} - l_{j,k-1} - 1)}{\prod_{i\neq j=1}^{k-1} (l_{i,k-1} - l_{j,k-1})(l_{i,k-1} - l_{j,k-1} - 1)} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ \vdots \\ [m]_k \\ [m]_{k-1}^{j} \\ [m]_{k-2} \\ \vdots \\ m_{11} \end{bmatrix}, \tag{3.3}$$

where $l_{ij} = m_{ij} - i$.

The action of the other generators can be obtained from the commutation relations (see, for instance, Ref. 13). Therefore, it remains to determine the transformation of the I-basis under the action of the odd generators of sl(1,n).

B. Application of the Wigner–Eckart theorem

We consider first the odd negative root vectors $e_{10},...,e_{n0}$. With respect to the adjoint representation of sl(1,n), restricted to the even subalgebra gl(n), these generators transform among themselves,

$$[E_{ij},e_{k0}] = \delta_{jk} e_{i0} - \delta_{ij} e_{k0}. \tag{3.4}$$

Therefore, the linear span T_1 of $e_{10},...,e_{n0}$ is a gl(n) module, which, according to Proposition 2, is a gl(n) fidirmod with a

signature $[0,-1,\ldots,-1]_n$, i.e.,

$$T_1 = \text{lin env}\{e_{p0} | p = 1,\ldots,n\} = V([0,-1,\ldots,-1]_n). \tag{3.5}$$

The link between the generators $e_{10},\ldots,e_{n0}$ and the GZ basis in T_1 is easily established:

$$e_{p0} = \left| \begin{array}{c} [-1]_n^1 \\ \vdots \\ [-1]_p^1 \\ [-1]_{p-1} \\ \vdots \\ -1 \end{array} \right\rangle, \quad p = 1,\ldots,n, \tag{3.6}$$

where, according to the notation (2.3) and (2.4),

$$[-1]_k = \underbrace{[-1,\ldots,-1]}_{k \text{ times}} \equiv [-1,\ldots,-1]_k, \tag{3.7}$$

$$[-1]_k^1 = \underbrace{[0,-1,\ldots,-1]}_{k \text{ times}} \equiv [0,-1,\ldots,-1]_k. \tag{3.8}$$

In the representation space $\overline{W}([m]_{n+1})$ the abstract supercommutation relations (3.4) hold as operator equations

$$E_{ij}e_{k0} - e_{k0}E_{ij} \equiv [E_{ij},e_{k0}] = \delta_{jk}e_{i0} - \delta_{ij}e_{k0}. \tag{3.9}$$

For simplicity in (3.9) and everywhere in the paper we use the same notation for the elements of $\text{sl}(1,n)$ and for their representatives as endomorphisms in $\overline{W}([m]_{n+1})$. From the commutation relations (3.9) one concludes that the endomorphisms $e_{10},\ldots,e_{n0}$ are components of a $\text{gl}(n)$ irreducible tensor operator with a signature $[0,-1,\ldots,-1]_n$. Therefore, applying the Wigner–Eckart theorem,[15] taking into account that $V([-1]_n^1) \otimes V[m]_n)$ is a simply reducible $\text{gl}(n)$ module [see (2.49) for $N=1$], and using (3.6) and (2.54), we have

$$e_{p0} \begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \\ m_{11} \end{bmatrix} = \sum R([m]_{n+1},[m]_n,[m']_n) \left[\begin{array}{c|cc} [m']_n & [-1]_n^1 & [m]_n \\ \vdots & \vdots & \vdots \\ [m']_p & [-1]_p^1 & [m]_p \\ [m']_{p-1} & [-1]_{p-1} & [m]_{p-1} \\ \vdots & \vdots & \vdots \\ m'_{11} & -1 & m_{11} \end{array} \right] \begin{bmatrix} [m]_{n+1} \\ [m']_n \\ \vdots \\ m'_{11} \end{bmatrix}. \tag{3.10}$$

The sum in (3.10) is over all those I-patterns that are allowed from Proposition 4; $R([m]_{n+1},[m]_n,[m']_n)$ are the so-called reduced matrix elements for the tensor operator $(e_{10},\ldots,e_{n0})$ and

$$\left[\begin{array}{c|cc} [m']_n & [-1]_n^1 & [m]_n \\ \vdots & \vdots & \vdots \\ m'_{11} & -1 & m_{11} \end{array} \right] \tag{3.11}$$

are the $\text{gl}(n)$ Clebsch–Gordan coefficients, which relate the tensor product of two GZ bases in the decomposition

$$V([-1]_n^1) \otimes V([m]_n) = \sum_{[m']_n} \oplus V([m']_n), \tag{3.12}$$

i.e.,

$$\left| \begin{array}{c} [-1]_n^1 \\ \vdots \\ [-1]_p^1 \\ [-1]_{p-1} \\ \vdots \\ -1 \end{array} \right\rangle \otimes \left| \begin{array}{c} [m]_n \\ \vdots \\ [m]_p \\ [m]_{p-1} \\ \vdots \\ m_{11} \end{array} \right\rangle = \sum \left[\begin{array}{c|cc} [m']_n & [-1]_n^1 & [m]_n \\ \vdots & \vdots & \vdots \\ [m']_p & [-1]_p^1 & [m]_p \\ [m']_{p-1} & [-1]_{p-1} & [m]_{p-1} \\ \vdots & \vdots & \vdots \\ m'_{11} & -1 & m_{11} \end{array} \right] \left| \begin{array}{c} [m']_n \\ \vdots \\ [m']_p \\ [m']_{p-1} \\ \vdots \\ m'_{11} \end{array} \right\rangle. \tag{3.13}$$

Taking into account that

$$V([1]_n) \otimes V([m]_n) = V([m+1]_n) \tag{3.14}$$

and multiplying both sides of (3.13) with the basis vector of the one-dimensional module $V([1]_n)$, one derives that

$$\left[\begin{array}{c|cc} [m']_n & [-1]_n^1 & [m]_n \\ \vdots & \vdots & \vdots \\ [m']_p & [-1]_p^1 & [m]_p \\ [m']_{p-1} & [-1]_{p-1} & [m]_{p-1} \\ \vdots & \vdots & \vdots \\ m'_{11} & -1 & m_{11} \end{array}\right] = \left[\begin{array}{c|cc} [m'+1]_n & [0]_n^1 & [m]_n \\ \vdots & \vdots & \vdots \\ [m'+1]_p & [0]_p^1 & [m]_p \\ [m'+1]_{p-1} & [0]_{p-1} & [m]_{p-1} \\ \vdots & \vdots & \vdots \\ m'_{11}+1 & 0 & m_{11} \end{array}\right]. \tag{3.15}$$

Each such Clebsch–Gordan coefficient can be written as a product of gl(r) scalar factors ($r = 2,3,...,n$),[16]

$$\left[\begin{array}{c|cc} [m'+1]_n & [0]_n^1 & [m]_n \\ \vdots & \vdots & \vdots \\ [m'+1]_p & [0]_p^1 & [m]_p \\ [m'+1]_{p-1} & [0]_{p-1} & [m]_{p-1} \\ \vdots & \vdots & \vdots \\ m'_{11}+1 & 0 & m_{11} \end{array}\right] = \prod_{r=p+1}^{n} \left[\begin{array}{c|cc} [m'+1]_r & [0]_r^1 & [m]_r \\ [m'+1]_{r-1} & [0]_{r-1}^1 & [m]_{r-1} \end{array}\right]$$

$$\times \left[\begin{array}{cc} [m'+1]_p & [0]_p^1 \\ [m'+1]_{p-1} & [0]_{p-1} \end{array}; [m]_p \atop [m]_{p-1}\right]$$

$$\times \prod_{r=2}^{p-1} \left[\begin{array}{c|cc} [m'+1]_r & [0]_r & [m]_r \\ [m'+1]_{r-1} & [0]_{r-1} & [m]_{r-1} \end{array}\right]. \tag{3.16}$$

The expressions for the scalar factors are available.[17] The gl(n) scalar factor,

$$\left[\begin{array}{c|cc} [m'+1]_r & [0]_r^1 & [m]_r \\ [m'+1]_{r-1} & [0]_{r-1}^1 & [m]_{r-1} \end{array}\right], \tag{3.17}$$

may be nonzero only if

$$\exists i = 1,...,r \text{ such that } [m'+1]_r = [m]_r^i, \quad \exists j = 1,...,r-1 \text{ such that } [m'+1]_{r-1} = [m]_{r-1}^j. \tag{3.18}$$

If for certain $i = 1,...,r$ and $j = 1,...,r-1$ (3.18) holds, i.e.,

$$\begin{aligned} &m'_{1r} = m_{1r} - 1 + \delta_{1i},...,m'_{rr} = m_{rr} - 1 + \delta_{ri}, \\ &m'_{1,r-1} = m_{1,r-1} - 1 + \delta_{1j},...,m'_{r-1,r-1} = m_{r-1,r-1} - 1 + \delta_{r-1,j}, \end{aligned} \tag{3.19}$$

then

$$\left[\begin{array}{c|cc} [m]_r^i & [0]_r^1 & [m]_r \\ [m]_{r-1}^j & [0]_{r-1}^1 & [m]_{r-1} \end{array}\right] = S(i,j) \left| \frac{\prod_{k\neq j=1}^{r-1}(l_{k,r-1} - l_{i,r-1}) \prod_{k\neq i=1}^{r}(l_{kr} - l_{j,r-1})}{\prod_{k\neq i=1}^{r}(l_{kr} - l_{ir}) \prod_{k\neq j=1}^{r-1}(l_{k,r-1} - l_{j,r-1} - 1)} \right|^{1/2}, \tag{3.20}$$

where l_{ij} and $S(i,j)$ are defined by (2.6) and (2.7).

The gl(p) scalar factor

$$\left[\begin{array}{cc} [m'+1]_p & [0]_p^1 \\ [m'+1]_{p-1} & [0]_{p-1} \end{array}; [m]_p \atop [m]_{p-1}\right] \tag{3.21}$$

may take nonzero values only if

$$\exists j = 1,...,r-1 \text{ such that } [m'+1]_p = [m]_p^j \text{ and } [m'+1]_{p-1} = [m]_{p-1}, \tag{3.22}$$

i.e., for certain $j = 1,...,p$,

$$m'_{1p} = m_{1p} - 1 + \delta_{1j},...,m'_{pp} = m_{pp} - 1 + \delta_{pj}, \quad m'_{1,p-1} + 1 = m_{1,p-1},...,m'_{p-1,p-1} + 1 = m_{p-1,p-1}. \tag{3.23}$$

Then

$$\left[\begin{array}{cc} [m]_p^j & [0]_p^1 \\ [m]_{p-1} & [0]_{p-1} \end{array}; [m]_p \atop [m]_{p-1}\right] = \left| \frac{\prod_{k=1}^{p-1}(l_{k,p-1} - l_{jp} - 1)}{\prod_{k\neq j=1}^{p}(l_{kp} - l_{jp})} \right|^{1/2}. \tag{3.24}$$

The gl(r) scalar factor

$$\left[\begin{array}{cc} [m'+1]_r & [0]_r \\ [m'+1]_{r-1} & [0]_{r-1} \end{array}; [m]_r \atop [m]_{r-1}\right] \tag{3.25}$$

is nonzero if and only if $[m'+1]_r = [m]_r$ and $[m'+1]_{r-1} = [m]_{r-1}$. Then

$$\left[\begin{array}{cc} [m]_r & [0]_r \\ [m]_{r-1} & [0]_{r-1} \end{array}; [m]_r \atop [m]_{r-1}\right] = 1. \tag{3.26}$$

Inserting (3.15) in (3.10) and using (3.18), (3.22), and (3.26), we obtain

$$e_{p0}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\\ [m]_p\\ [m]_{p-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{j_n=1}^{n}\sum_{j_{n-1}=1}^{n-1}\cdots\sum_{j_p=1}^{p}R([m]_{n+1},[m]_n,[m-1]_n^{j_n})$$

$$\times\begin{bmatrix}[m]_p^{j_p}\\ [m]_{p-1}\end{bmatrix}\begin{bmatrix}[0]_p^1\\ [0]_{p-1}\end{bmatrix};\begin{bmatrix}[m]_p\\ [m]_{p-1}\end{bmatrix}\prod_{r=p+1}^{n}\begin{bmatrix}[m]_r^{j_r}\\ [m]_{r-1}^{j_{r-1}}\end{bmatrix}\begin{bmatrix}[0]_r^1\\ [0]_{r-1}^1\end{bmatrix};\begin{bmatrix}[m]_r\\ [m]_{r-1}\end{bmatrix}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^{j_n}\\ \vdots\\ [m-1]_p^{j_p}\\ [m-1]_{p-1}\\ \vdots\\ m_{11}-1\end{bmatrix}. \tag{3.27}$$

In a similar way as for T_1 [Eq. (3.5)] one concludes that the subalgebra P_+, defined with (2.23), is a gl(n) fidirmod with respect to the adjoint representation. Its signature is $[1]_n^{-n}$, i.e.,

$$P_+=\text{lin env}\{e_{0p}|p=1,\ldots,n\}=V([1]_n^{-n}). \tag{3.28}$$

The relation between the GZ basis in $V([1]_n^{-n})$ and the positive odd root vectors reads

$$e_{0p}=\left|\begin{matrix}[1]_n^{-n}\\ \vdots\\ [1]_p^{-p}\\ [1]_{p-1}\\ \vdots\\ 1\end{matrix}\right\rangle,\quad p=1,\ldots,n. \tag{3.29}$$

Since $V([1]_n^{-n})\otimes V_0([m]_n)$ is a simply reducible gl(n) module, from the Wigner–Eckart theorem we have

$$e_{0p}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\\ [m]_p\\ [m]_{p-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum S([m]_{n+1},[m]_n,[m']_n)\begin{bmatrix}[m']_n\\ \vdots\\ [m']_p\\ [m']_{p-1}\\ \vdots\\ m'_{11}\end{bmatrix}\begin{bmatrix}[1]_n^{-n}\\ \vdots\\ [1]_p^{-p}\\ [1]_{p-1}\\ \vdots\\ 1\end{bmatrix};\begin{bmatrix}[m]_n\\ \vdots\\ [m]_p\\ [m]_{p-1}\\ \vdots\\ m_{11}\end{bmatrix}\begin{bmatrix}[m]_{n+1}\\ [m']_n\\ \vdots\\ [m']_p\\ [m']_{p-1}\\ \vdots\\ m'_{11}\end{bmatrix}. \tag{3.30}$$

The sum is over all I-patterns (Proposition 4); $S([m]_{n+1},[m]_n,[m']_n)$ are the reduced matrix elements, corresponding to $(e_{01},\ldots,e_{0n})$. The Clebsch–Gordan coefficients can be represented as products of scalar factors

$$\begin{bmatrix}[m']_n\\ \vdots\\ [m']_p\\ [m']_{p-1}\\ \vdots\\ m'_{11}\end{bmatrix}\begin{bmatrix}[1]_n^{-n}\\ \vdots\\ [1]_p^{-p}\\ [1]_{p-1}\\ \vdots\\ 1\end{bmatrix};\begin{bmatrix}[m]_n\\ \vdots\\ [m]_p\\ [m]_{p-1}\\ \vdots\\ m_{11}\end{bmatrix}=\begin{bmatrix}[m'-1]_n\\ \vdots\\ [m'-1]_p\\ [m'-1]_{p-1}\\ \vdots\\ m'_{11}-1\end{bmatrix}\begin{bmatrix}[0]_n^{-n}\\ \vdots\\ [0]_p^{-p}\\ [0]_{p-1}\\ \vdots\\ 0\end{bmatrix};\begin{bmatrix}[m]_n\\ \vdots\\ [m]_p\\ [m]_{p-1}\\ \vdots\\ m_{11}\end{bmatrix}$$

$$=\prod_{r=p+1}^{n}\begin{bmatrix}[m'-1]_r\\ [m'-1]_{r-1}\end{bmatrix}\begin{bmatrix}[0]_r^{-r}\\ [0]_{r-1}^{-r+1}\end{bmatrix};\begin{bmatrix}[m]_r\\ [m]_{r-1}\end{bmatrix}\times\begin{bmatrix}[m'-1]_p\\ [m'-1]_{p-1}\end{bmatrix}\begin{bmatrix}[0]_p^{-p}\\ [0]_{p-1}\end{bmatrix};\begin{bmatrix}[m]_p\\ [m]_{p-1}\end{bmatrix}$$

$$\times\prod_{r=2}^{p-1}\begin{bmatrix}[m'-1]_r\\ [m'-1]_{r-1}\end{bmatrix}\begin{bmatrix}[0]_r\\ [0]_{r-1}\end{bmatrix};\begin{bmatrix}[m]_r\\ [m]_{r-1}\end{bmatrix}. \tag{3.31}$$

The scalar factors, appearing in the first multiple, may be different from zero only if

$$\exists i = 1,\ldots,r \text{ such that } [m'-1]_r = [m]_r^{-i},$$
$$\exists j = 1,\ldots,r-1 \text{ such that } [m'-1]_{r-1} = [m]_{r-1}^{-j}. \tag{3.32}$$

If (3.32) holds, then[16]

$$\begin{bmatrix} [m]_r^{-i} \\ [m]_{r-1}^{-j} \end{bmatrix} \begin{bmatrix} [0]_r^{-r} & [m]_r \\ [0]_{r-1}^{-r+1} & [m]_{r-1} \end{bmatrix} = S(i,j) \left| \frac{\Pi_{k\neq j=1}^{r-1}(l_{k,r-1}-l_{ir})\Pi_{k\neq i=1}^{r}(l_{kr}-l_{j,r-1}+1)}{\Pi_{k\neq i=1}^{r}(l_{kr}-l_{ir})\Pi_{k\neq j=1}^{r-1}(l_{k,r-1}-l_{j,r-1}+1)} \right|^{1/2}. \tag{3.33}$$

The second multiple in (3.31) may take nonzero values only if

$$\exists i = 1,\ldots,p \text{ such that } [m'-1]_p = [m]_p^{-i} \text{ and } [m'-1]_{p-1} = [m]_{p-1}. \tag{3.34}$$

Then

$$\begin{bmatrix} [m]_p^{-i} \\ [m]_{p-1} \end{bmatrix} \begin{bmatrix} [0]_p^{-p} & [m]_p \\ [0]_{p-1} & [m]_{p-1} \end{bmatrix} = \left| \frac{\Pi_{k=1}^{p-1}(l_{k,p-1}-l_{ip})}{\Pi_{k\neq i=1}^{p}(l_{kp}-l_{ip})} \right|^{1/2}. \tag{3.35}$$

The scalar factors, appearing in the last multiple of (3.31), are nonzero if and only if

$$[m'-1]_r = [m]_r \text{ and } [m'-1]_{r-1} = [m]_{r-1}. \tag{3.36}$$

In this case

$$\begin{bmatrix} [m]_r \\ [m]_{r-1} \end{bmatrix} \begin{bmatrix} [0]_r & [m]_r \\ [0]_{r-1} & [m]_{r-1} \end{bmatrix} = 1. \tag{3.37}$$

Taking into account (3.31), (3.32), (3.34), and (3.36) we write (3.30) in the following form:

$$e_{0p} \begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \\ [m]_p \\ [m]_{p-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{j_n=1}^{n} \sum_{j_{n-1}=1}^{n-1} \cdots \sum_{j_p=1}^{p} S([m]_{n+1},[m]_n,[m+1]_n^{-j_n})$$
$$\times \begin{bmatrix} [m]_p^{-j_p} \\ [m]_{p-1} \end{bmatrix} \begin{bmatrix} [0]_p^{-p} & [m]_p \\ [0]_{p-1} & [m]_{p-1} \end{bmatrix} \prod_{r=p+1}^{n} \begin{bmatrix} [m]_r^{-j_r} \\ [m]_{r-1}^{-j_{r-1}} \end{bmatrix} \begin{bmatrix} [0]_r^{-r} & [m]_r \\ [0]_{r-1}^{-r+1} & [m]_{r-1} \end{bmatrix} \begin{bmatrix} [m]_{n+1} \\ [m+1]_n^{-j_n} \\ \vdots \\ [m+1]_p^{-j_p} \\ [m+1]_{p-1} \\ \vdots \\ m_{11}+1 \end{bmatrix}. \tag{3.38}$$

C. Reduction of the problem

At this place it is convenient to change to new notation for the reduced matrix elements. According to Proposition 5 and the relation (2.8) the signature $[m]_n$ of the gl(n) submodule $V([m]_n) \subset \overline{W}([m]_{n+1})$ is in one-to-one correspondence with the set $(i_1,\ldots,i_N)$ of those indices for which $\theta_{i_1} = \theta_{i_2} = \cdots = \theta_{i_N} = 1$. Since $[m-1]_n^i$ and $[m+1]_n^{-i}$ are also determined by the same θ-tuple $(i_1,\ldots,i_N)$ and i, we set

$$R([m]_{n+1},[m]_n,[m-1]_n^i) \equiv R([m]_{n+1};i_1,\ldots,i_N;i), \tag{3.39}$$
$$S([m]_{n+1},[m]_n,[m+1]_n^{-i}) \equiv S([m]_{n+1};i_1,\ldots,i_N;i). \tag{3.40}$$

In the case $p = n$ we obtain from (3.27), and (3.38)–(3.40),

$$e_{n0} \begin{bmatrix} [m]_{n+1} \\ [m]_n \\ [m]_{n-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{i=1}^{n} R([m]_{n+1};i_1,\ldots,i_N;i) \left| \frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ [m-1]_n^i \\ [m-1]_{n-1} \\ \vdots \\ m_{11}-1 \end{bmatrix}, \tag{3.41}$$

$$e_{0n}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{i=1}^{n}S([m]_{n+1};i_1,\ldots,i_N;i)\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}\\ \vdots\\ m_{11}+1\end{bmatrix}. \tag{3.42}$$

Proposition 6: In the right-hand side of Eqs. (3.41) and (3.42) the coefficients, written below, are nonzero for any I-pattern:

$$\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|\neq 0,$$

$$\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|\neq 0,\quad i=1,\ldots,n. \tag{3.43}$$

Proof: From (2.57) one derives

$$m_{in}-m_{jn}\geqslant 0\quad \forall i<j=1,\ldots,n, \tag{3.44}$$

$$m_{in}-m_{j,n-1}\geqslant 0\quad \forall i\leqslant j=1,\ldots,n-1, \tag{3.45}$$

$$m_{i,n-1}-m_{j+1,n}\geqslant 0\quad \forall i\leqslant j=1,\ldots,n-1. \tag{3.46}$$

The inequality (3.44) together with the definition (2.6) yields

$$l_{in}-l_{jn}>0\quad \forall i<j=1,\ldots,n, \tag{3.47}$$

and, therefore,

$$\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})\neq 0. \tag{3.48}$$

Suppose that for a certain $i=1,\ldots,n$ there exists an I-pattern

$$\begin{bmatrix}[m]_{n+1}\\ [\tilde{m}]_n\\ [\tilde{m}]_{n-1}\\ \vdots\\ \tilde{m}_{11}\end{bmatrix}=\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^{i}\\ [m-1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}$$

such that

$$\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^{i}\\ [m-1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}=0. \tag{3.49}$$

Then there should exist $k=1,\ldots,n-1$ such that

$$l_{k,n-1}-l_{in}-1=m_{k,n-1}-m_{in}-k+i-1=0,$$

i.e.,

$$m_{k,n-1}-m_{in}=k-i+1. \tag{3.50}$$

(a) Suppose that $k\geqslant i$. Then $k-i+1>0$, whereas according to (3.45) $m_{k,n-1}-m_{in}\leqslant 0$. Hence (3.50) is impossible.

(b) Suppose that $k<i$. Then $k-i+1\leqslant 0$, whereas (3.46) yields $m_{k,n-1}-m_{in}\geqslant 0$. Therefore, (3.50) could be fulfilled only if $k=i-1$, i.e., if

$$m_{i-1,n-1}=m_{in}. \tag{3.51}$$

Suppose that (3.51) holds. Then [see (3.49)]

$$\tilde{m}_{i-1,n-1}=m_{i-1,n-1}-1,\quad \tilde{m}_{in}=m_{in}. \tag{3.52}$$

Combining (3.51) and (3.52) we obtain $\tilde{m}_{i-1,n-1}<\tilde{m}_{in}$, which contradicts the definition of an I-pattern [see (2.57)]. Hence, also in this case (3.50) is impossible. This proves that the coefficient from (3.41) cannot be zero. In a similar way one concludes that the coefficient from (3.42) is different from zero, i.e., (3.43) holds.

Proposition 7: If

$$\deg\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\end{bmatrix}=N,$$

then

$$\deg\begin{bmatrix}[m]_{n+1}\\ [m\mp 1]_n^{\pm i}\\ \vdots\end{bmatrix}=\tilde{N}=N\pm 1.$$

Proof: Let $[\tilde{m}]_n=[m\mp 1]_n^{\pm i}$, i.e.,

$$\tilde{m}_{kn}=m_{kn}\mp 1\pm\delta_{ki}. \tag{3.53}$$

According to (2.60) the degree $\tilde{N}$, corresponding to $[\tilde{m}]_n$, is

$$\tilde{N}=\frac{1}{n-1}\sum_{k=1}^{n}(m_{k,n+1}-\tilde{m}_{kn})$$

$$=\frac{1}{n-1}\sum_{k=1}^{n}(m_{k,n+1}-m_{kn})$$

$$\pm\frac{1}{n-1}\sum_{k=1}^{n}(1-\delta_{ki})=N\pm 1. \qquad \blacksquare$$

Proposition 8: If $i\in(i_1,\ldots,i_N)$, then

$$R([m]_{n+1};i_1,\ldots,i_N;i)=0. \tag{3.54}$$

Proof: The vector on the left-hand side of (3.41) is of degree N. For the corresponding θ-tuple we have $\theta_{i_1}=\theta_{i_2}=\cdots=\theta_{i_N}=1$. Therefore, for each $i_1,\ldots,i_p,\ldots,i_N$,

$$\theta_{i_p}=N-(m_{i_p,n+1}-m_{i_p,n})=1. \tag{3.55}$$

Suppose that for a certain $i_p\in(i_1,\ldots,i_N)$, $R([m]_{n+1};i_1,\ldots,i_N;i_p)\neq 0$. Then in the sum (3.41) there will appear a vector with $[\tilde{m}]_n=[m-1]_n^{i_p}$, i.e., with $\tilde{m}_{kn}=m_{kn}-1+\delta_{k,i_p}$. In particular, $\tilde{m}_{i_p,n}=m_{i_p,n}$ and since (Proposition 7)

$$\deg\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^{i_p}\\ \vdots\end{bmatrix}=\tilde{N}=N+1,$$

we conclude from (2.63) that

$$\tilde{\theta}_{i_p}=\tilde{N}-(m_{i_p,n+1}-\tilde{m}_{i_p,n})$$

$$=N-(m_{i_p,n+1}-m_{i_p,n})+1=\theta_{i_p}+1=2, \tag{3.56}$$

which is impossible, since it contradicts the definition of the I-basis (Proposition 4).

In a similar way one proves the next statement.

Proposition 9: If $i\bar{\in}(i_1,\ldots,i_N)$, then

$$S([m]_{n+1};i_1,\ldots,i_N;i)=0. \tag{3.57}$$

In the cases $p=n, n-1$ the relations (3.27) and (3.38) together with Propositions 8 and 9 yield

$$e_{n0}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{i\bar{\in}(i_1,\ldots,i_N)}R([m]_{n+1};i_1,\ldots,i_N;i)\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^i\\ [m-1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}, \tag{3.58}$$

$$e_{0n}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{i\in(i_1,\ldots,i_N)}S([m]_{n+1};i_1,\ldots,i_N;i)\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}\\ \vdots\\ m_{11}+1\end{bmatrix}, \tag{3.59}$$

$$\begin{aligned}e_{n-1,0}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}&=\sum_{i\bar{\in}(i_1,\ldots,i_N)}\sum_{j=1}^{n-1}R([m]_{n+1};i_1,\ldots,i_N;i)S(i,j)\\ &\times\left|\frac{\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{in}-1)\prod_{k\neq i=1}^{n}(l_{kn}-l_{j,n-1})\prod_{k=1}^{n-2}(l_{k,n-2}-l_{j,n-1}-1)}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{j,n-1})(l_{k,n-1}-l_{j,n-1}-1)}\right|^{1/2}\\ &\times\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^i\\ [m-1]_{n-1}^j\\ [m-1]_{n-2}\\ \vdots\\ m_{11}-1\end{bmatrix},\end{aligned} \tag{3.60}$$

$$\begin{aligned}e_{0,n-1}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}&=\sum_{i\in(i_1,\ldots,i_N)}\sum_{j=1}^{n-1}S([m]_{n+1};i_1,\ldots,i_N;i)S(i,j)\\ &\times\left|\frac{\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{in})\prod_{k\neq i=1}^{n}(l_{kn}-l_{j,n-1}+1)\prod_{k=1}^{n-2}(l_{k,n-2}-1_{j,n-1})}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{j,n-1})(l_{k,n-1}-l_{j,n-1}+1)}\right|^{1/2}\\ &\times\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}^{-j}\\ [m+1]_{n-2}\\ \vdots\\ m_{11}+1\end{bmatrix}.\end{aligned} \tag{3.61}$$

In (3.58), (3.60), and throughout the paper a sum over $\bar{i}\in(i_1,\ldots,i_N)$ means a sum over the compliment set $(1,\ldots,n)\setminus(i_1,\ldots,i_N)$, i.e.,

$$\sum_{\bar{i}\in(i_1,\ldots,i_N)} \equiv \sum_{\substack{i=(1,2,\ldots,n)\\ i\neq(i_1,\ldots,i_N)}} . \tag{3.62}$$

Proposition 10: The endomorphisms e_{n0} and e_{0n} of $\overline{W}([m]_{n+1})$, defined with Eqs. (3.58) and (3.59) satisfy the right commutation relations with the even generators E_{ij} [which action on $\overline{W}([m]_{n+1})$ follows from (3.1)–(3.3)], namely

$$[e_{n0},E_{ij}] = -\delta_{jn}e_{i0} + \delta_{ij}e_{n0}, \quad i,j=1,\ldots,n, \tag{3.63}$$

$$[e_{0n},E_{ij}] = \delta_{in}e_{0j} - \delta_{ij}e_{0n}, \quad i,j=1,\ldots,n. \tag{3.64}$$

A sketch of the proof is given in Appendix A.

Proposition 11: The endomorphisms e_{k0} and e_{0k}, defined in (3.27) and (3.28), satisfy the commutation relations with the even generators E_{ij}, i.e.,

$$[e_{0k},E_{ij}] = \delta_{ki}e_{0j} - \delta_{ij}e_{0k}, \quad i,j,k=1,\ldots,n, \tag{3.65}$$

$$[e_{k0},E_{ij}] = -\delta_{jk}e_{i0} + \delta_{ij}e_{k0}, \quad i,j,k=1,\ldots,n. \tag{3.66}$$

Proof: Let $i\leqslant j$ and $k<n$. Using Proposition 10 and the gl(n) commutation relations (1.4), we have

$$\begin{aligned}
&[e_{0k},E_{ij}]\\
&\quad=[e_{0n}E_{nk}-E_{nk}e_{0n},E_{ij}]\\
&\quad=[[e_{0n},E_{ij}],E_{nk}]+[e_{0n},[E_{nk},E_{ij}]]\\
&\quad=[\delta_{in}e_{0j}-\delta_{ij}e_{0n},E_{nk}]+[e_{0n},\delta_{ki}E_{nj}-\delta_{jn}E_{ik}]\\
&\quad=\delta_{in}[e_{0j},E_{nk}]-\delta_{ij}e_{0k}+\delta_{ki}(e_{0j}-\delta_{jn}e_{0n})\\
&\qquad-\delta_{jn}(\delta_{in}e_{0k}-\delta_{ik}e_{0n}).
\end{aligned} \tag{3.67}$$

Since $i\leqslant j$,

$$\delta_{in}e_{0j}=\delta_{in}e_{0n} \text{ and } \delta_{in}e_{0k}=\delta_{in}\delta_{jn}e_{0k}. \tag{3.68}$$

Inserting (3.68) and (3.67) and taking into account that $[e_{0n},E_{nk}]=e_{0k}$, we get the desired result

$$[e_{0k},E_{ij}] = \delta_{ki}e_{0j} - \delta_{ij}e_{0k} \quad \forall i\leqslant j=1,\ldots,n. \tag{3.69}$$

To complete the proof we use the relation

$$[e_{0,n-1},E_{n,n-1}]=0, \tag{3.70}$$

which is proved in Appendix B.

Let $n>i>j$. Then

$$\begin{aligned}
[e_{0,n-1},E_{ij}] &= [[e_{0n},E_{n,n-1}],E_{ij}]\\
&= [e_{0n},[E_{n,n-1},E_{ij}]]=\delta_{i,n-1}e_{0j}.
\end{aligned} \tag{3.71}$$

From (3.69)–(3.71) we conclude that the $e_{0,n-1}$ fulfill the commutation relations with gl(n):

$$[e_{0,n-1},E_{ij}] = \delta_{i,n-1}e_{0j} - \delta_{ij}e_{0,n-1}, \quad i,j=1,\ldots,n. \tag{3.72}$$

Let now $k<n-1$ and $n>i>j$. Then

$$\begin{aligned}
[e_{0k},E_{ij}] &= [e_{0n}E_{nk}-E_{nk}e_{0n},E_{ij}]\\
&= [e_{0n},[E_{nk},E_{ij}]]\\
&= \delta_{ik}[e_{0n},E_{nj}]=\delta_{ik}e_{0j}, \quad n>i>j.
\end{aligned} \tag{3.73}$$

If $k<n-1$ and $n>j$,

$$\begin{aligned}
[e_{0k},E_{nj}] &= [e_{0,n-1}E_{n-1,k}-E_{n-1,k}e_{0,n-1},E_{nj}]\\
&= -\delta_{j,n-1}[e_{0,n-1},E_{nk}]=0, \quad n>j.
\end{aligned} \tag{3.74}$$

From (3.69), (3.73), and (3.74) we conclude that

$$[e_{0k},E_{ij}] = \delta_{ki}e_{0j} - \delta_{ij}e_{0k}, \quad k<n-1,\ i,j=1,\ldots,n. \tag{3.75}$$

The last relation together with (3.72) and (3.64) gives

$$[e_{0k},E_{ij}] = \delta_{ki}e_{0j} - \delta_{ij}e_{0k} \quad \forall i,j,k=1,\ldots,n. \tag{3.76}$$

In a similar way one also proves the commutation relation (3.66).

Proposition 12: If $\{e_{0n},e_{0n}\}=0$ (i.e., $e_{0n}e_{0n}=0$) as an operator equality in $\overline{W}([m]_{n+1})$, then also

$$\{e_{0p},e_{0q}\}=0 \quad \forall p,q=1,\ldots,n. \tag{3.77}$$

Proof: Let $p<n$. From Eq. (3.76) we have

$$\begin{aligned}
\{e_{0n},e_{0p}\} &= \{e_{0n},e_{0n}E_{np}-E_{np}e_{0n}\} = -e_{0n}[e_{0n},E_{np}]\\
&\quad -[e_{0n},E_{np}]e_{0n} = -\{e_{0n},e_{0p}\}.
\end{aligned}$$

Therefore,

$$\{e_{0n},e_{0p}\}=0, \quad p=1,\ldots,n. \tag{3.78}$$

Let $p<n$, $q<n$. Then

$$\begin{aligned}
\{e_{0p},e_{0q}\} &= \{e_{0p},e_{0n}E_{nq}-E_{nq}e_{0n}\}\\
&= -\{e_{0n},[e_{0p},E_{nq}]\}=0.
\end{aligned}$$ ■

In a similar way one proves the following proposition.

Proposition 13: If $\{e_{n0},e_{n0}\}=0$ in $\overline{W}([m]_{n+1})$, then

$$\{e_{p0},e_{q0}\}=0 \quad \forall p,q=1,\ldots,n. \tag{3.79}$$

Proposition 14: If the operator equation

$$\{e_{n0},e_{0n}\}=E_{nn} \tag{3.80}$$

holds in $\overline{W}([m]_{n+1})$, then also the anticommutation relation

$$\{e_{p0},e_{0q}\}=E_{pq} \tag{3.81}$$

is fulfilled.

Proof: Let $p<n$. Using Eqs. (3.75) and (3.81) one derives

$$\begin{aligned}
\{e_{0n},e_{p0}\} &= \{e_{0n},[E_{pn},e_{n0}]\} = [E_{pn},\{e_{n0},e_{0n}\}]\\
&= [E_{pn},E_{nn}]=E_{pn}.
\end{aligned} \tag{3.82}$$

If $p<n$ and $q<n$, then we have from (1.4) and (3.82)

$$\begin{aligned}
\{e_{p0},e_{0q}\} &= \{e_{p0},[e_{0n},E_{nq}]\}\\
&= [\{e_{p0},e_{0n}\},E_{nq}]-\{e_{0n},[e_{p0},E_{nq}]\}\\
&= [E_{pn},E_{nq}]+\delta_{pq}\{e_{0n},e_{n0}\}=E_{pq}.
\end{aligned}$$ ■

From Propositions 11–14 we conclude.

Corollary: The operators e_{k0} and e_{0k}, $k=1,\ldots,n$, defined with the Eqs. (3.27) and (3.38) turn the linear space $\overline{W}([m]_{n+1})$ into an sl(1,n) module if and only if the following relations hold:

$$e_{n0}e_{n0}=0, \tag{3.83}$$

$$e_{0n}e_{0n}=0, \tag{3.84}$$

$$e_{n0}e_{0n}+e_{0n}e_{n0}=E_{nn}. \tag{3.85}$$

D. Determination of the reduced matrix elements

Considering the relations (3.83)–(3.85) as operator equations in $\bar{W}([m]_{n+1})$ and using the expressions (3.58) and (3.59) we arrive at the following equations for the unknown reduced matrix elements.

For any $i<j=1,\ldots,n$ and $i,j\bar{\in}(i_1,\ldots,i_N)$,

$$\frac{R([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N,i;j)}{|l_{i,n+1}-l_{j,n+1}+1|^{1/2}} + \frac{R([m]_{n+1};i_1,\ldots,i_N;j)R([m]_{n+1};i_1,\ldots,i_N,j;i)}{|l_{j,n+1}-l_{i,n+1}+1|^{1/2}}=0. \tag{3.86}$$

For any $i<j\bar{\in}(i_1,\ldots,i_N)$,

$$\frac{S([m]_{n+1};i_1,\ldots,i_N;i)S([m]_{n+1};i_1,\ldots,i_N\setminus i;j)}{|l_{j,n+1}-l_{i,n+1}+1|^{1/2}} + \frac{S([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N\setminus j;i)}{|l_{i,n+1}-l_{j,n+1}+1|^{1/2}}, \tag{3.87}$$

where $i_1,\ldots,i_N\setminus i$ is the set $i_1,\ldots,i_N$ from which the index i has been deleted.

For any $i\bar{\in}(i_1,\ldots,i_N)$ and $j\bar{\in}(i_1,\ldots,i_N)$,

$$S([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N\setminus i;j) + R([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N,j;i)=0, \tag{3.88}$$

$$\sum_{i\bar{\in}(i_1,\ldots,i_N)} S([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N\setminus i;i) \times\left|\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in})\right|\left|\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})(l_{kn}-l_{in}+1)\right|^{-1/2} + \sum_{j\bar{\in}(i_1,\ldots,i_N)} R([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N,j;j) \times\left|\prod_{k=1}^{n-1}(l_{k,n-1}-l_{jn}-1)\right|\left|\prod_{k\neq j=1}^{n}(l_{kn}-l_{jn})(l_{kn}-l_{jn}-1)\right|^{-1/2} = \sum_{i=1}^{n} m_{in}-\sum_{j=1}^{n-1} m_{j,n-1}. \tag{3.89}$$

A derivation of Eqs. (3.86)–(3.89) is given in Appendix C.

A hint of how to solve the equations for the reduced matrix elements give the results from Refs. 18 and 19. There we have introduced a concept of I-basis for the LS's sl(1,2) and sl(1,3). In particular, for the action of e_{n0},e_{0n} $(n=2,3)$ on the I-basis we had [see (2.32) and (2.35) in Ref. 18 and (2.32) in Ref. 19]

$$e_{n0}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix} = \sum_{i\bar{\in}(i_1,\ldots,i_N)} \operatorname{sgn}(\theta)\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^{i}\\ [m-1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}, \tag{3.90}$$

$$e_{0n}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix} = \sum_{i\in(i_1,\ldots,i_N)} \operatorname{sgn}(\theta)(l_{i,n+1}+1)\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}\\ \vdots\\ m_{11}+1\end{bmatrix}, \tag{3.91}$$

where $\operatorname{sgn}(\theta)$ is a sign function, depending on $\theta_1,\ldots,\theta_n$, i.e., $\operatorname{sgn}(\theta)=\pm 1$. Comparing (3.58) with (3.90) and (3.59) with (3.91) we conclude that in the cases $n=2,3$ the reduced matrix elements are of the form

$$R([m]_{n+1};i_1,\ldots,i_N;i) = \operatorname{sgn}(\theta)\left|\frac{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}, \tag{3.92}$$

$$S([m]_{n+1};i_1,\ldots,i_N;i) = \operatorname{sgn}(\theta)(l_{i,n+1}+1)\left|\frac{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}. \tag{3.93}$$

The solution of Eqs. (3.86)–(3.89) is not unique. One possible solution can be given with expressions of the form (3.92), (3.93) for any n. More precisely, we have the following proposition.

Proposition 15: The reduced matrix elements $R([m]_{n+1};i_1,\ldots,i_N;i)$ and $S([m]_{n+1};i_1,\ldots,i_N;i)$ of the gl(n) irreducible tensor operators $(e_{10},\ldots,e_{n0})$ and $(e_{01},\ldots,e_{0n})$ can be chosen to be

$$R([m]_{n+1};i_1,\ldots,i_N;i) = (1-\theta_i)(-1)^{\theta_1+\cdots+\theta_{i-1}} \times\left|\frac{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}, \tag{3.94}$$

$$S([m]_{n+1};i_1,\ldots,i_N;i)$$
$$=\theta_i(-1)^{\theta_1+\cdots+\theta_{i-1}}(l_{i,n+1}+1)$$
$$\times\left|\frac{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})}{\Pi_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}. \tag{3.95}$$

It is not difficult to check that the expressions (3.94) and (3.95) satisfy the first three equations (3.86)–(3.88). In the proof it is convenient to use the following representation of the reduced matrix elements:

$$R([m]_{n+1};i_1,\ldots,i_N;i)$$
$$=(1-\theta_i)(-1)^{\theta_1+\cdots+\theta_{i-1}}$$
$$\times\prod_{i\in(i_1,\ldots,i_N)}\left|\frac{l_{k,n+1}-l_{i,n+1}+1}{l_{k,n+1}-l_{i,n+1}}\right|^{1/2}, \tag{3.96}$$

$$S([m]_{n+1};i_1,\ldots,i_N;i)$$
$$=\theta_i(-1)^{\theta_1+\cdots+\theta_{i-1}}(l_{i,n+1}+1)$$
$$\times\prod_{\bar{i}\in(i_1,\ldots,i_N)}\left|\frac{l_{k,n+1}-l_{i,n+1}-1}{l_{k,n+1}-l_{i,n+1}}\right|^{1/2}. \tag{3.97}$$

To prove that the relations (3.94) and (3.95) also satisfy Eq. (3.89) we use the following identity.

Proposition 16: Let $A_1,\ldots,A_n,B_1,\ldots,B_n$ be, in general, complex numbers such that

$$A_i\neq A_j, \quad \text{if } i\neq j=1,\ldots,n, \tag{3.98}$$
$$A_i-B_j\in\mathbb{Z}\forall i,j,=1,\ldots,n. \tag{3.99}$$

Then

$$\sum_{i=1}^{n}\frac{\Pi_{k=1}^{n}(A_i-B_k)}{\Pi_{k\neq i=1}^{n}(A_i-A_k)}=\sum_{i=1}^{n}(A_i-B_i). \tag{3.100}$$

The proof is given in Appendix D.

If $B_n=0$ then (3.100) reads

$$\sum_{i=1}^{n}A_i\frac{\Pi_{k=1}^{n-1}(A_i-B_k)}{\Pi_{k\neq i=1}^{n}(A_i-A_k)}=\sum_{i=1}^{n}A_i-\sum_{j=1}^{n-1}B_j. \tag{3.101}$$

Suppose that the numbers $A_1,\ldots,A_n,B_1,\ldots,B_{n-1}$ are such that

$$\begin{aligned}&A_i-A_k\in\mathbb{N} \quad \forall i<k=1,\ldots,n,\\ &A_i-B_k\in\mathbb{Z}_+ \quad \forall i\leqslant k=1,\ldots,n-1,\\ &B_k-A_i\in\mathbb{Z}_+ \quad \forall k<i=1,\ldots,n.\end{aligned} \tag{3.102}$$

Then

$$\frac{A_i-B_k}{A_i-A_k}=\left|\frac{A_i-B_k}{A_i-A_k}\right| \quad \forall i\neq k=1,\ldots,n,$$

and, therefore, (3.101) takes the form

$$\sum_{i=1}^{n}A_i\left|\frac{\Pi_{k=1}^{n-1}(A_i-B_k)}{\Pi_{k\neq i=1}^{n}(A_i-A_k)}\right|=\sum_{i=1}^{n}A_i-\sum_{j=1}^{n-1}B_j. \tag{3.103}$$

Consider an arbitrary I-pattern (2.54) of degree N and let

$$\{\theta_1,\ldots,\theta_n\}=(i_1,\ldots,i_N) \tag{3.104}$$

be the θ-tuple corresponding to it (Proposition 5). From (3.96) and (3.97) one derives

$$R([m]_{n+1};i_1,\ldots,i_N\setminus i;i)$$
$$=(-1)^{\theta_1+\cdots+\theta_{i-1}}$$
$$\times\left|\frac{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in}+1)}{\Pi_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}, \tag{3.105}$$

$$S([m]_{n+1};i_1,\ldots,i_N,j;j)$$
$$=(-1)^{\theta_1+\cdots+\theta_{j-1}}(l_{j,n+1}+1)$$
$$\times\left|\frac{\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn}-1)}{\Pi_{k\neq j=1}^{n}(l_{k,n+1}-l_{j,n+1})}\right|^{1/2}, \tag{3.106}$$

where, we wish to underline, everything in (3.105) and (3.106) is expressed in terms of $l_{1n},\ldots,l_{nn}$ and the θ-tuple (3.104), corresponding to the initial I-pattern. Inserting the expressions (3.96), (3.97), (3.105), and (3.106) in the left-hand side of Eq. (3.89) [= lhs(3.89)], we have

lhs(3.89)
$$=\sum_{i\in(i_1,\ldots,i_N)}(l_{i,n+1}+1)\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\Pi_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|$$
$$+\sum_{\bar{j}\in(i_1,\ldots,i_N)}(l_{j,n+1}+1)$$
$$\times\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{jn}-1)}{\Pi_{k\neq j=1}^{n}(l_{k,n+1}-l_{j,n+1})}\right|. \tag{3.107}$$

The relations (2.63), (2.6), and (2.9) yield

$$\begin{aligned}&l_{in}=l_{i,n+1}+1-N, \quad i\in(i_1,\ldots,i_N),\\ &l_{jn}=l_{j,n+1}-N, \quad \bar{j\in}(i_1,\ldots,i_N).\end{aligned} \tag{3.108}$$

Combining Eqs. (3.107) and (3.108) we get

lhs(3.89)
$$=\sum_{i=1}^{n}(l_{i,n+1}+1)$$
$$\times\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{i,n+1}+N-1)}{\Pi_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|. \tag{3.109}$$

Introduce the notation

$$A_i=l_{i,n+1}+1, \quad B_k=l_{k,n-1}+N,$$
$$i=1,\ldots,n, \quad k=1,\ldots,n-1.$$

In terms of these notation (3.109) reads

$$\text{lhs}(3.89)=\sum_{i=1}^{n}A_i\left|\frac{\Pi_{k=1}^{n-1}(A_i-B_k)}{\Pi_{k\neq i=1}^{n}(A_i-A_k)}\right|. \tag{3.110}$$

For $i<k$, $m_{i,n+1}-m_{k,n+1}\in\mathbb{Z}_+$. Therefore

$$(m_{i,n+1}-i+1)-(m_{k,n+1}-k+1)\in\mathbb{N},$$

i.e.,

$$A_i-A_k\in\mathbb{N} \quad \forall i<k=1,\ldots,n. \tag{3.111}$$

From (2.63) and (2.6) we have

$$A_i=l_{in}+N+1-\theta_i=m_{in}-i+N+1-\theta_i. \tag{3.112}$$

If $i\leqslant k=1,\ldots,n-1$, then [see (2.57)]

$$A_i-B_k=m_{in}-m_{k,n-1}+1-\theta_i+k-i\in\mathbb{Z}_+. \tag{3.113}$$

If $k<i=1,\ldots,n$, then

$$B_k - A_i = m_{k,n-1} - m_{in} + \theta_i - 1 + i - k \in \mathbb{Z}_+. \tag{3.114}$$

Hence, we can apply to (3.110) the identity (3.103):

$$\text{lhs}(3.89) = \sum_{i=1}^{n} A_i - \sum_{j=1}^{n-1} B_j = \sum_{i=1}^{n} m_{in} - \sum_{j=1}^{n-1} m_{j,n-1}. \tag{3.115}$$

Thus, the expressions (3.94) and (3.95) for $R([m]_{n+1};i_1,\ldots,i_N,i)$ and $S([m]_{n+1};i_1,\ldots,i_N,i)$ satisfy also Eq. (3.89) and, therefore, these expressions can be accepted as reduced matrix elements of the gl(n) irreducible tensor operators $(e_{10},\ldots,e_{n0})$ and $(e_{01},\ldots,e_{0n})$.

E. Typical representations

Inserting (3.94), (3.20), and (3.24) in (3.27) we obtain the transformation of the I-basis (2.54) under the action of the odd positive root vectors:

$$e_{p0}\begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \\ [m]_p \\ [m]_{p-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{i_n=1}^{n} \sum_{i_{n-1}=1}^{n-1} \cdots \sum_{i_p=1}^{p} (1-\theta_{i_n})(-1)^{\theta_1+\cdots+\theta_{i_n-1}} \left| \frac{\prod_{k\neq i_n=1}^{n}(l_{kn}-l_{i_n,n})}{\prod_{k\neq i_n=1}^{n}(l_{k,n+1}-l_{i_n,n+1})} \right|^{1/2}$$

$$\times \prod_{r=p+1}^{n} S(i_r,i_{r-1}) \left| \frac{\prod_{k\neq i_{r-1}}^{r-1}(l_{k,r-1}-l_{i_r,r-1})\prod_{k\neq i_r=1}^{r}(l_{k,r}-l_{i_{r-1},r-1})}{\prod_{k\neq i_r=1}^{r}(l_{k,r}-l_{i_r,r})\prod_{k\neq i_{r-1}=1}^{r-1}(l_{k,r-1}-l_{i_{r-1},r-1}-1)} \right|^{1/2}$$

$$\times \left| \frac{\prod_{k=1}^{p-1}(l_{k,p-1}-l_{i_p,p}-1)}{\prod_{k\neq i_p=1}^{p}(l_{k,p}-l_{i_p,p})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ [m-1]_n^{i_n} \\ \vdots \\ [m-1]_p^{i_p} \\ [m-1]_{p-1} \\ \vdots \\ m_{11}-1 \end{bmatrix}, \quad p=1,\ldots,n. \tag{3.116}$$

Similarly, inserting (3.33), (3.35), and (3.95) in (3.38), we obtain

$$e_{0p}\begin{bmatrix} [m]_{n+1} \\ [m]_n \\ \vdots \\ [m]_p \\ [m]_{p-1} \\ \vdots \\ m_{11} \end{bmatrix} = \sum_{i_n=1}^{n} \sum_{i_{n-1}=1}^{n-1} \cdots \sum_{i_p=1}^{p} \theta_{i_n}(-1)^{\theta_1+\cdots+\theta_{i_n-1}}(l_{i_n,n+1}+1)$$

$$\times \prod_{r=p+1}^{n} S(i_r,i_{r-1}) \left| \frac{\prod_{k\neq i_{r-1}=1}^{r-1}(l_{k,r-1}-l_{i_r,r})\prod_{k\neq i_r=1}^{r}(l_{k,r}-l_{i_{r-1},r-1}+1)}{\prod_{k\neq i_r=1}^{r}(l_{kr}-l_{i_r,r})\prod_{k\neq i_{r-1}=1}^{r-1}(l_{k,r-1}-l_{i_{r-1},r-1}+1)} \right|^{1/2}$$

$$\times \left| \frac{\prod_{k\neq i_n=1}^{n}(l_{kn}-l_{i_n,n})}{\prod_{k\neq i_n=1}^{n}(l_{k,n+1}-l_{i_n,n+1})} \right|^{1/2} \times \left| \frac{\prod_{k=1}^{p-1}(l_{k,p-1}-l_{i_p,p})}{\prod_{k\neq i_p=1}^{p}(l_{kp}-l_{i_p,p})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ [m+1]_n^{-i_n} \\ \vdots \\ [m+1]_p^{-i_p} \\ [m+1]_{p-1} \\ \vdots \\ m_{11}+1 \end{bmatrix}. \tag{3.117}$$

The transformations (3.116) and (3.117) [see also (3.58) and (3.59)] take a particularly simple form in the case $p = n$:

$$e_{n0}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix} = \sum_{i=1}^{n}(1-\theta_i)(-1)^{\theta_1+\cdots+\theta_{i-1}}\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^i\\ [m-1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}, \tag{3.118}$$

$$e_{0n}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix} = \sum_{i=1}^{n}\theta_i(-1)^{\theta_1+\cdots+\theta_{i-1}}(l_{i,n+1}+1)\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\prod_{k\neq i=1}^{n}(l_{k,n+1}-l_{i,n+1})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}\\ \vdots\\ m_{11}+1\end{bmatrix}. \tag{3.119}$$

The transformations (3.118) and (3.119) together with the expressions (3.1)–(3.3) for the even generators determine uniquely all other generators. In this sense the relations (3.118) and (3.119) determine the representation of the LS sl(1,n) in $\overline{W}([m]_{n+1})$.

For any n-tuple $[m_{1,n+1},m_{2,n+1},\ldots,m_{n,n+1}]$, $m_{i,n+1}-m_{i+1,n+1}\in\mathbb{Z}_+$, $i=1,\ldots,n-1$, the formulas (3.116), (3.117) define completely the induced representation of sl(1,n) in $\overline{W}([m]_{n+1})$ [the transformations (3.1)–(3.3) for the even generators follow from (3.116) and (3.117), since $E_{ij}=\{e_{i0},e_{0j}\}$]. This representation is irreducible and, hence, typical if and only if (Proposition 1)

$$m_{k,n+1}\neq k-1,\quad k=1,\ldots,n. \tag{3.120}$$

If for certain $k=1,\ldots,n$, $m_{k,n+1}=k-1$, the representation is indecomposible. In this case $\overline{W}([m]_{n+1})$ contains a maximal invariant submodule $\overline{I}([m]_{n+1})$ and at the same time there exists no compliment to $\overline{I}([m]_{n+1})$ subspace, which is invariant with respect to sl(1,n). The factor module $\overline{W}([m]_{n+1})/\overline{I}([m]_{n+1})$ carries an irreducible nontypical representation. The maximal invariant subspaces, the factor modules and their transformation under the action of the sl(1,n) generators will be given in Ref. 1.

ACKNOWLEDGMENT

The author is thankful to Professor H. D. Doebner for the kind hospitality in the Arnold Sommerfeld Institute for Mathematical Physics in Clausthal and for the valuable discussions on different points of the present investigation.

APPENDIX A: A PROOF OF PROPOSITION 10

Consider first the commutator

$$[e_{n0},E_{n,n-1}] = e_{n0}E_{n,n-1} - E_{n,n-1}e_{n0}. \tag{A1}$$

Acting with the right-hand side of (A1) on an arbitrary I-pattern and using Eqs. (3.2) and (3.58) we have

$$(e_{n0}E_{n,n-1}-E_{n,n-1}e_{n0})\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}$$

$$=\sum_{j=1}^{n-1}\left|\frac{\prod_{i=1}^{n}(l_{in}-l_{j,n-1}+1)\prod_{i=1}^{n-2}(l_{i,n-2}-l_{j,n-1})}{\prod_{i\neq j=1}^{n-1}(l_{i,n-1}-l_{jn}+1)(l_{i,n-1}-l_{j,n-1})}\right|^{1/2}e_{n0}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}^{-j}\\ [m]_{n-2}\\ \vdots\\ m_{11}\end{bmatrix}$$

$$-\sum_{\bar{i}\in(i_1,\ldots,i_N)}R([m]_{n+1};i_1,\ldots,i_N;i)\left|\frac{\prod_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}E_{n,n-1}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^i\\ [m-1]_{n-1}\\ \vdots\\ m_{11}-1\end{bmatrix}$$

$$=\sum_{j=1}^{n-1}\left|\frac{(l_{in}-l_{j,n-1}+1)\prod_{k\neq i=1}^{n}(l_{kn}-l_{j,n-1}+1)\prod_{k=1}^{n-2}(l_{k,n-2}-l_{j,n-2})}{\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{j,n-1}+1)(l_{k,n-1}-l_{j,n-1})}\right|^{1/2}$$

$$\times\sum_{\bar{i}\in(i_1,\ldots,i_N)}R([m]_{n+1};i_1,\ldots,i_N;i)\left|\frac{(l_{j,n-1}-l_{in}-2)\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^i\\ [m-1]_{n-1}^{-j}\\ [m-1]_{n-2}\\ \vdots\\ m_{11}-1\end{bmatrix}$$

$$-\sum_{\bar{i}\in(i_1,\ldots,i_N)}R([m]_{n+1};i_1,\ldots,i_N;i)\left|\frac{(l_{j,n-1}-l_{in}-1)\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{in}-1)}{\prod_{k\neq i=1}^{n}(l_{kn}-l_{in})}\right|^{1/2}$$

$$\times\sum_{j=1}^{n-1}\left|\frac{(l_{in}-l_{j,n-1}+2)\prod_{k\neq i=1}^{n}(l_{kn}-l_{j,n-1}+1)\prod_{k=1}^{n-2}(l_{k,n-2}-l_{j,n-2})}{\prod_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{j,n-1}+1)(l_{k,n-1}-l_{j,n-1})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m-1]_n^i\\ [m-1]_{n-1}^{-j}\\ [m-1]_{n-2}\\ \vdots\\ m_{11}-1\end{bmatrix}=0, \tag{A2}$$

which is in agreement with (3.63).

By a straightforward computation one proves also that the following commutation relations hold:

$$[e_{n0},E_{k,k-1}]=0,\quad k=1,\ldots,n-1, \tag{A3}$$

$$[e_{n0},E_{k-1,k}]=\delta_{kn}e_{n-1,0},\quad k=1,\ldots,n, \tag{A4}$$

$$[e_{n0},E_{kk}]=(1-\delta_{kn})e_{n0},\quad k=1,\ldots,n. \tag{A5}$$

The rest of the relations (3.63) follow from (A3)–(A5) and the gl(n) commutation relations. The proof of the commutation relations (3.64) is similar.

APPENDIX B: DERIVATION OF EQ. (3.70)

Applying twice Eqs. (3.2) and (3.61) we arrive at the expression

$$(B)\equiv[e_{0,n-1},E_{n,n-1}]\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{i\in(i_1,\ldots,i_N)}\sum_{j=1}^{n-1}\sum_{s=1}^{n-1}S([m]_{n+1};i_1,\ldots,i_N;i)$$

$$\times S(i,s)\left\{\left|\frac{\Pi_{k=1}^{n}(l_{kn}-l_{j,n-1}+1)\Pi_{k=1}^{n-2}(l_{k,n-2}-l_{j,n-1})}{\Pi_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{j,n-1})(l_{k,n-1}-l_{j,n-1}+1)}\right|^{1/2}\right.$$

$$\times\left|\frac{\Pi_{k\neq s=1}^{n-1}(l_{k,n-1}-l_{in}-\delta_{kj})\Pi_{k\neq i=1}^{n}(l_{kn}-l_{s,n-1}+\delta_{sj}+1)\Pi_{k=1}^{n-2}(l_{k,n-2}-l_{s,n-1}+\delta_{sj})}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})\Pi_{k\neq s=1}^{n-1}(l_{k,n-1}-l_{s,n-1}-\delta_{kj}+\delta_{sj})(l_{k,n-1}-l_{s,n-1}-\delta_{kj}+\delta_{sj}+1)}\right|^{1/2}$$

$$-\left|\frac{\Pi_{k\neq s=1}^{n-1}(l_{k,n-1}-l_{in})\Pi_{k\neq i=1}^{n}(l_{kn}-l_{s,n-1}+1)\Pi_{k=1}^{n-2}(l_{k,n-2}-l_{s,n-1})}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{i,n+1})\Pi_{k\neq s=1}^{n-1}(l_{k,n-1}-l_{s,n-1})(l_{k,n-1}-l_{s,n-1}+1)}\right|^{1/2}$$

$$\times\left.\left|\frac{\Pi_{k=1}^{n}(l_{kn}-l_{j,n-1}-\delta_{ki}+\delta_{js}+1)\Pi_{k=1}^{n-2}(l_{k,n-2}-l_{j,n-1}+\delta_{js})}{\Pi_{k\neq j=1}^{n-1}(l_{k,n-1}-l_{j,n-1}-\delta_{ks}+\delta_{js})(l_{k,n-1}-l_{j,n-1}-\delta_{ks}+\delta_{js}+1)}\right|^{1/2}\right\}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}^{-j,-s}\\ [m+1]_{n-2}\\ \vdots\\ m_{11}+1\end{bmatrix}. \tag{B1}$$

Denote the expression in the curled brackets in (B1) as

$$F(j,s;[m]_n,[m]_{n-1},[m]_{n-2};i) \tag{B2}$$

and represent the sum as

$$(B)=\sum_{i\in(i_1,\ldots,i_N)}S([m]_{n+1};i_1,\ldots,i_N;i)$$

$$\times\left\{\sum_{j=1}^{n-1}S(i,s)F(j,j;[m]_n,[m]_{n-1},[m]_{n-2};i)+\sum_{s<j=1}^{n-1}[S(i,s)F(j,s;[m]_n,[m]_{n-1},[m]_{n-2};i)\right.$$

$$\left.+S(i,j)F(s,j;[m]_n,[m]_{n-1},[m]_{n-2};i)]\right\}\begin{bmatrix}[m]_{n+1}\\ [m+1]_n^{-i}\\ [m+1]_{n-1}^{-j,-s}\\ [m+1]_{n-2}\\ \vdots\\ m_{11}+1\end{bmatrix}. \tag{B3}$$

After some calculations one obtains

$$F(j,j;[m]_n,[m]_{n-1},[m]_{n-2};i)=0,$$

$$S(i,j)F(j,s;\ldots)+S(i,j)F(s,j;\ldots)$$

$$=\frac{G(j,s;[m]_n,[m]_{n-1};i)}{\Pi_{k=1}^{3}(l_{s,n-1}-l_{j,n-1}+k-2)}$$

$$\times\left|\prod_{\substack{k\neq s=1\\ k\neq j}}^{n-1}\frac{(l_{k,n-1}-l_{in})}{(l_{k,n-1}-l_{j,n-1}+1)(l_{k,n-1}-l_{j,n-1})(l_{k,n-1}-l_{s,n-1}+1)(l_{k,n-1}-l_{s,n-1})}\right.$$

$$\times\left.\prod_{k\neq i=1}^{n}\frac{(l_{kn}-l_{j,n-1}+1)(l_{kn}-l_{s,n-1}+1)}{(l_{kn}-l_{in})}\prod_{k=1}^{n-2}(l_{k,n-2}-l_{j,n-1})(l_{k,n-2}-l_{s,n-1})\right|^{1/2}, \tag{B4}$$

where

$$G(j,s;[m]_n,[m]_{n-1};i)$$

$$=S(i,s)|l_{in}-l_{j,n-1}|(l_{s,n-1}-l_{j,n-1}+1)-S(i,s)|l_{in}-l_{j,n-1}+1|(l_{s,n-1}-l_{j,n-1}-1)$$

$$+S(i,j)|l_{in}-l_{s,n-1}|(l_{s,n-1}-l_{j,n-1}-1)-S(i,j)|l_{in}-l_{s,n-1}+1|(l_{s,n-1}-l_{j,n-1}+1). \tag{B5}$$

To get rid of the modulus in (B5) consider the following cases.

(a) $i \leqslant s$. Since $s<j$, also $i<j$ and, therefore, $S(i,j)=S(i,s)=1$. From (2.57) and (2.6) we conclude that $l_{in}-l_{j,n-1} \geqslant 0$, $l_{in}-l_{s,n-1} \geqslant 0$. Therefore,

$$
\begin{aligned}
&G(j,s;[m]_n,[m]_{n-1};i) \\
&\quad = (l_{in}-l_{j,n-1})(l_{s,n-1}-l_{j,n-1}+1)-(l_{in}-l_{j,n-1}+1)(l_{s,n-1}-l_{j,n-1}-1) \\
&\qquad + (l_{in}-l_{s,n-1})(l_{s,n-1}-l_{j,n-1}-1)-(l_{in}-l_{s,n-1}+1)(l_{s,n-1}-l_{j,n-1}+1)=0. \qquad \text{(B6)}
\end{aligned}
$$

In a similar way one shows that

$$
G(j,s;[m]_n,[m]_{n-1};i)=0. \qquad \text{(B7)}
$$

In the other two cases,

(b) $j<i$,

(c) $s<i \leqslant j$.

Hence, (B7) holds for arbitrary values of the indices i,j,s. From (B3), (B4), and (B7) we get the desired result

$$
[e_{0,n-1},E_{n,n-1}]=0. \qquad \text{(B8)}
$$

APPENDIX C: EQUATIONS FOR THE REDUCED MATRIX ELEMENTS

Acting with $e_{n0}e_{n0}$ on an arbitrary I-pattern and using (3.58) one derives the expression

$$
\begin{aligned}
e_{n0}e_{n0}\begin{bmatrix} [m]_{n+1} \\ [m]_n \\ [m]_{n-1} \\ \vdots \\ m_{11} \end{bmatrix} &= \sum_{i \bar{\in} (i_1,\ldots,i_N)} \sum_{j \bar{\in} (i_1,\ldots,i_N)} R([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N,i;j) \\
&\quad \times \left| \frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)(l_{k,n-1}-l_{jn}-1)}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn}+\delta_{ik})} \right|^{1/2} \begin{bmatrix} [m]_{n+1} \\ [m-2]_n^{i,j} \\ [m-2]_{n-1} \\ \vdots \\ m_{11}-2 \end{bmatrix}, \qquad \text{(C1)}
\end{aligned}
$$

which can be written also as

$$
\begin{aligned}
e_{n0}e_{n0}\begin{bmatrix} [m]_{n+1} \\ [m]_n \\ [m]_{n-1} \\ \vdots \\ m_{11} \end{bmatrix} &= \sum_{i<\bar{j} \bar{\in} (i_1,\ldots,i_N)} \left| \frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)(l_{k,n-1}-l_{jn}-1)}{(l_{in}-l_{jn})\Pi_{k=1,k\neq i,j}^{n}(l_{kn}-l_{in})(l_{kn}-l_{jn})} \right|^{1/2} \\
&\quad \times \left\{ \frac{R([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N,i;j)}{|l_{i,n+1}-l_{j,n+1}+1|^{1/2}} \right. \\
&\quad \left. + \frac{R([m]_{n+1};i_1,\ldots,i_N;j)R([m]_{n+1};i_1,\ldots,i_N,j;i)}{|l_{j,n+1}-l_{i,n+1}+1|^{1/2}} \right\} \begin{bmatrix} [m]_{n+1} \\ [m-2]_n^{i,j} \\ [m-2]_{n-1} \\ \vdots \\ m_{11}-2 \end{bmatrix}. \qquad \text{(C2)}
\end{aligned}
$$

In obtaining (C2) we have used the circumstance that $\theta_i=\theta_j=0$ and, hence,

$$
l_{i,n+1}-l_{j,n+1}=l_{in}-l_{jn}. \qquad \text{(C3)}
$$

Since, as a consequence of Proposition 6,

$$
\left| \frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in}-1)(l_{k,n-1}-l_{jn}-1)}{(l_{in}-l_{jn})\Pi_{k=1,k\neq i,j}^{n}(l_{kn}-l_{in})(l_{kn}-l_{jn})} \right|^{1/2} \neq 0, \quad i<j,
$$

the operator relation (3.83) holds if and only if the expression in the curled brackets of (C2) vanishes, i.e., if $R([m]_{n+1},i_1,\ldots,i_N;i)$ satisfies Eq. (3.86).

The relation (C3) holds also if $i,j\in(i_1,\ldots,i_N)$. Using this one derives from (3.84) and (3.59):

$$e_{0n}e_{0n}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=\sum_{\substack{i,j\in(i_1,\ldots,i_N)\\ i<j}}\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})(l_{k,n-1}-l_{jn})}{(l_{in}-l_{jn})\Pi_{k\neq i=1,k\neq j}^{n}(l_{kn}-l_{in})(l_{kn}-l_{jn})}\right|^{1/2}$$

$$\times\left\{\frac{S([m]_{n+1};i_1,\ldots,i_N;i)S([m]_{n+1};i_1,\ldots,i_N\setminus i;j)}{|l_{j,n+1}-l_{i,n+1}+1|^{1/2}}\right.$$

$$\left.+\frac{S([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N\setminus j;i)}{|l_{i,n+1}-l_{j,n+1}+1|^{1/2}}\right\}\begin{bmatrix}[m]_{n+1}\\ [m+2]_n^{-i,-j}\\ [m+2]_{n-1}\\ \vdots\\ m_{11}+2\end{bmatrix}. \tag{C4}$$

Therefore, Eq. (C4) holds for any I-pattern iff the Eq. (3.87) is fulfilled.

Consider Eq. (3.85). Acting with the right-hand side of it on an arbitrary I-pattern and applying (3.1) for $k=n$ we have

$$E_{nn}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=\left(\sum_{i=1}^{n}m_{in}-\sum_{j=1}^{n-1}m_{j,n-1}\right)\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}. \tag{C5}$$

Acting with the left-hand side of (3.85) on an arbitrary I-pattern and taking into account (3.58) and (3.59) we obtain

$$C\equiv(e_{n0}e_{0n}+e_{0n}e_{n0})\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}$$

$$=\sum_{i\in(i_1,\ldots,i_N)}\sum_{\bar{j}\in(i_1,\ldots,i_N\setminus i)}S([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N\setminus i;j)$$

$$\times\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})(l_{k,n-1}-l_{jn}+\delta_{ij}-1)}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn}-\delta_{ki}+\delta_{ij})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m]_n^{-i,j}\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}$$

$$+\sum_{\bar{j}\in(i_1,\ldots,i_N)}\sum_{i\in(i_1,\ldots,i_N,j)}R([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N,j;i)$$

$$\times\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{jn}-1)(l_{k,n-1}-l_{in}-\delta_{ij})}{\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn})\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in}+\delta_{kj}-\delta_{ij})}\right|^{1/2}\begin{bmatrix}[m]_{n+1}\\ [m]_n^{-i,j}\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}=C_1+C_2, \tag{C6}$$

$$C_1=\sum_{i\in(i_1,\ldots,i_N)}\sum_{j\bar{\in}(i_1,\ldots,i_N)}\Bigg\{S([m]_{n+1};i_1,\ldots,i_N;i)$$
$$\times\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})(l_{k,n-1}-l_{jn}-1)}{\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn}-\delta_{ki})}\right|^{1/2}R([m]_{n+1};i_1,\ldots,i_N\backslash i;j)$$
$$+R([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N,j;i)$$
$$\times\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{jn}-1)(l_{k,n-1}-l_{in})}{\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn})\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in}+\delta_{ki})}\right|^{1/2}\Bigg\}\begin{bmatrix}[m]_{n+1}\\ [m]_n^{-i,j}\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix},\tag{C7}$$

$$C_2=\Bigg\{\sum_{i\in(i_1,\ldots,i_N)}S([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N\backslash i;i)\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})}{\left|\Pi_{k\neq i=1}^{n}(l_{kn}-l_{in})(l_{kn}-l_{in}+1)\right|^{1/2}}$$
$$+\sum_{j\bar{\in}(i_1,\ldots,i_N)}R([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N,j;j)\frac{\left|\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{jn}-1)\right|}{\left|\Pi_{k\neq j=1}^{n}(l_{kn}-l_{jn})(l_{kn}-l_{jn}-1)\right|^{1/2}}\Bigg\}\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}.\tag{C8}$$

The term C_1 may be written as

$$C_1=\sum_{i\in(i_1,\ldots,i_N)}\sum_{j\bar{\in}(i_1,\ldots,i_N)}\left|\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})(l_{k,n-1}-l_{jn}-1)}{(l_{in}-l_{jn})(l_{in}-l_{jn}-1)\Pi_{k\neq i=1,k\neq j}^{n}(l_{kn}-l_{in})(l_{kn}-l_{jn})}\right|^{1/2}$$
$$\times\{S([m]_{n+1};i_1,\ldots,i_N;i)R([m]_{n+1};i_1,\ldots,i_N\backslash i;j)$$
$$+R([m]_{n+1};i_1,\ldots,i_N;j)S([m]_{n+1};i_1,\ldots,i_N,j;i)\}\begin{bmatrix}[m]_{n+1}\\ [m]_n^{-i,j}\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}.\tag{C9}$$

From (C5), (C6), (C8), and (C9) we conclude that the Eq. (3.85) is fulfilled if and only if

$$C_1=0,\tag{C10}$$

$$C_2=\left(\sum_{i=1}^{n}m_{in}-\sum_{j=1}^{n-1}m_{j,n-1}\right)\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ \vdots\\ m_{11}\end{bmatrix}.\tag{C11}$$

One easily derives that if

$$\begin{bmatrix}[m]_{n+1}\\ [m]_n\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix}\quad\text{and}\quad\begin{bmatrix}[m]_{n+1}\\ [m]_n^{-i,j}\\ [m]_{n-1}\\ \vdots\\ m_{11}\end{bmatrix},\quad i\neq j,$$

are I-patterns, then

$$\frac{\Pi_{k=1}^{n-1}(l_{k,n-1}-l_{in})(l_{k,n-1}-l_{jn}-1)}{(l_{in}-l_{jn})(l_{in}-l_{jn}-1)\Pi_{k\neq i=1,k\neq j}^{n}(l_{kn}-l_{in})(l_{kn}-l_{jn})}\neq 0. \tag{C12}$$

Since, moreover, the right-hand side of (C9) is a linear combination of linearly independent I-patterns, Eq. (C10) holds iff the reduced matrix elements satisfy Eq. (3.88). Equation (3.89) is an immediate consequence of Eqs. (C8) and (C11).

APPENDIX D: A PROOF OF THE IDENTITY (3.100)

This identity can be derived in different ways. Here we follow a proof, which was suggested to us by D. Pfeil (University of Clausthal). In the cases $n=1,2$, (3.100) is evident. We prove by induction on n. Suppose (3.100) holds for a given n. Consider $2n+2$ numbers A_i, B_i, $i=1,\ldots,n+1$, which satisfy the conditions (3.98) and (3.99). Using the identity

$$(A_i-B_{n+1})(A_n-A_{n+1})=(A_i-A_{n+1})(A_n-B_{n+1})-(A_i-A_n)(A_{n+1}-B_{n+1}),$$

we have

$$\begin{aligned}
D\equiv\sum_{i=1}^{n+1}\frac{\Pi_{k=1}^{n+1}(A_i-B_k)}{\Pi_{k\neq i=1}^{n+1}(A_i-A_k)}&=\sum_{i=1}^{n-1}\frac{(A_i-B_n)(A_i-B_{n+1})\Pi_{k=1}^{n-1}(A_i-B_k)}{(A_i-A_n)(A_i-A_{n+1})\Pi_{k\neq i=1}^{n-1}(A_i-A_k)}\cdot\frac{A_n-A_{n+1}}{A_n-A_{n+1}}\\
&+\frac{(A_n-B_{n+1})\Pi_{k=1}^{n}(A_n-B_k)}{(A_n-A_{n+1})\Pi_{k=1}^{n-1}(A_n-A_k)}+\frac{(A_{n+1}-B_n)\Pi_{k\neq n=1}^{n+1}(A_{n+1}-B_k)}{(A_{n+1}-A_n)\Pi_{k=1}^{n-1}(A_{n+1}-A_k)}\\
&=\sum_{i=1}^{n-1}\frac{\Pi_{k=1}^{n-1}(A_i-B_k)}{\Pi_{k\neq i=1}^{n-1}(A_i-A_k)}\cdot\frac{A_i-B_n}{A_n-A_{n+1}}\left[\frac{A_n-B_{n+1}}{A_i-A_n}-\frac{A_{n+1}-B_{n+1}}{A_i-A_{n+1}}\right]\\
&+\frac{(A_n-B_{n+1})\Pi_{k=1}^{n}(A_n-B_k)}{(A_n-A_{n+1})\Pi_{k=1}^{n-1}(A_n-A_k)}+\frac{(A_{n+1}-B_n)\Pi_{k\neq n=1}^{n+1}(A_{n+1}-B_k)}{(A_{n+1}-A_n)\Pi_{k=1}^{n-1}(A_{n+1}-A_k)}=D_1-D_2,
\end{aligned}\tag{D1}$$

where

$$D_1=\frac{A_n-B_{n+1}}{A_n-B_{n+1}}\left[\sum_{i=1}^{n-1}\frac{\Pi_{k=1}^{n}(A_i-B_k)}{\Pi_{k\neq i=1}^{n}(A_i-A_k)}+\frac{\Pi_{k=1}^{n}(A_n-B_k)}{\Pi_{k\neq n=1}^{n}(A_n-A_k)}\right]. \tag{D2}$$

The expression in the brackets of (D2) reduces to (3.100) in the case n, which holds by assumption. Therefore,

$$D_1=\frac{A_n-B_{n+1}}{A_n-A_{n+1}}\sum_{i=1}^{n}(A_i-B_i), \tag{D3}$$

$$D_2=\frac{A_{n+1}-B_{n+1}}{A_n-A_{n+1}}\left[\sum_{i=1}^{n-1}\frac{\Pi_{k=1}^{n}(A_i-B_k)}{(A_i-A_{n+1})\Pi_{k\neq i=1}^{n-1}(A_i-A_k)}+\frac{(A_{n+1}-B_n)\Pi_{k=1}^{n-1}(A_{n+1}-B_k)}{\Pi_{k=1}^{n-1}(A_{n+1}-A_k)}\right]. \tag{D4}$$

Introduce a new notation

$$\bar{A}_k=A_k,\quad \bar{A}_n=A_{n+1},\quad k=1,\ldots,n-1. \tag{D5}$$

Then D_2 reads

$$D_2=\frac{A_{n+1}-B_{n+1}}{A_n-A_{n+1}}\sum_{i=1}^{n}\frac{\Pi_{k=1}^{n}(\bar{A}_i-B_k)}{\Pi_{k\neq i=1}^{n}(\bar{A}_i-\bar{A}_k)},$$

and since $\bar{A}_1,\ldots,\bar{A}_n,B_1,\ldots,B_n$ satisfy the conditions (3.98) and (3.99) and since, moreover, in this case (3.100) holds by assumption, we obtain

$$D_2=\frac{A_{n+1}-B_{n+1}}{A_n-A_{n+1}}\sum_{i=1}^{n}(\bar{A}_i-B_i)=\frac{A_{n+1}-B_{n+1}}{A_n-A_{n+1}}\left[\sum_{i=1}^{n-1}(A_i-B_i)+A_{n+1}-B_{n+1}\right]. \tag{D6}$$

Inserting (D3) and (D6) in (D1) we obtain the desired result

$$\sum_{i=1}^{n+1}\frac{\Pi_{k=1}^{n+1}(A_i-B_k)}{\Pi_{k\neq i=1}^{n+1}(A_i-A_k)}=\sum_{i=1}^{n+1}(A_i-B_i). \tag{D7}$$

Hence, if (3.100) holds for certain n, it holds also for $n+1$. Since (3.100) is an identity for $n=1$, it is an identity for any integer $n\in\mathbb{N}$.

[1]T. D. Palev, "Finite-dimensional representations of the special linear Lie superalgebra sl(1,n). II. Nontypical representations" (in preparation).
[2]V. G. Kac, Adv. Math. **26**, 8 (1977).
[3]D. Ž. Djoković and G. Hochschild, Illinois J. Math. **20**, 134 (1976).
[4]V. G. Kac, Lect. Notes Math. **626**, 597 (1978).
[5]V. G. Kac, Comm. Algebra **5**, 889 (1977).
[6]V. Rittenberg, *Lecture Notes in Physics*, Vol. 79 (Springer, Berlin, 1978); M. Scheunert, *Lecture Notes in Mathematics*, Vol. 716 (Springer, Berlin,

1979); D. A. Leites, Sov. Prob. Mat. **25**, 3 (1984) (in Russian), and references therein.
[7]M. Scheunert, W. Nahm, and V. Rittenberg, J. Math. Phys. **18**, 155 (1977); M. Marcu, *ibid.* **21**, 1277 (1980).
[8]L. Ross, Trans. Am. Math. Soc. **120**, 17 (1965); A. Pais and V. Rittenberg, J. Math. Phys. **16**, 2062 (1975); G. Hochschild, Illinois J. Math. **20**, 107 (1976); D. Ž. Djoković, J. Pure. Appl. Algebra **9**, 25 (1976); F. A. Berezin, Sov. J. Nucl. Phys. **29**, 857 (1979); **30**, 605 (1979); T. D. Palev, J. Math. Phys. **21**, 1293 (1980); T. D. Palev and O. Ts. Stoytchev, C. R. Acad. Bulg. Sci. **35**, 733 (1982); J.-P. Hurni and B. Morel, J. Math. Phys. **23**, 2236 (1982); **24**, 157 (1983); B. Gruber, T. S. Santhanam, and R. Wilson, J. Math Phys. **25**, 1253 (1984).
[9]P. H. Dondi and P. D. Jarvis, Z. Phys. C **4**, 201 (1980); J. Phys. A **14**, 547 (1981); A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1149, 1810 (1981); *ibid.* **23**, 1239 (1982); A. B. Balantekin, *ibid.* **23**, 486 (1982); I. Bars, B. Morel, and H. Ruegg, *ibid.* **24**, 2253 (1983); F. Delduc and M. Gourdin, *ibid.* **25**, 1651 (1984); **26**, 1865 (1985); I. Bars, Physica D **15**, 42 (1985).
[10]J. Van der Jeugt, J. Math. Phys. **25**, 3334 (1984).
[11]T. D. Palev, J. Math. Phys. **26**, 1640 (1985); **28**, 272 (1987).
[12]I. M. Gel'fand and M. L. Zetlin, Dokl. Akad. Nauk SSSR **71**, 825 (1950) (in Russian); see also G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963).
[13]J. D. Louck, Am. J. Phys. **38**, 1 (1970).
[14]T. D. Palev and O. Ts. Stoychev, Preprint JINR E5-82-54, Dubna 1982.
[15]E. P. Wigner, *Gruppentheorie und ihre Anwendungen auf die Quantenmechanik der Atomspektren* (Friedr. Vieweg, Braunschweig, 1931); C. Eckart, Rev. Mod. Phys. **2**, 305 (1930); see also A. O. Barut and R. Raçzka, *Theory of Group Representations and Applications* (PWN-Polish Scientific, Warsaw, 1980), pp. 247.
[16]See, for instance, A. U. Klymik, *Matrix Elements of the Clebsch–Gordan Coefficients of Representations of Groups* (Naukova Dumka, Kiev, 1979) (in Russian).
[17]G. Baird and L. Biedenharn, J. Math. Phys. **4**, 1449 (1963).
[18]T. D. Palev, J. Math. Phys. **27**, 1994 (1986).
[19]A. H. Kamupingene and T. D. Palev, Preprint ICTP, Trieste, IC/85/146, 1985.

Parastatistics and the Clifford algebra unitary group approach to the many-electron correlation problem

M. D. Gould
School of Chemistry, University of Western Australia, Nedlands, Western Australia, 6009
J. Paldus[a)]
Institute for Advanced Study Berlin, Wallotstrasse 19, D-1000 Berlin 33, West Germany

(Received 12 February 1987; accepted for publication 17 June 1987)

It is shown that the Clifford algebra unitary group approach, which is based on the subgroup chain $U(2^n) \supset SO(2n+1) \supset SO(2n) \supset U(n)$, may be described in terms of the para-Fermi algebra. Applications to the development of efficient algorithms for the evaluation of matrix elements of $U(n)$ generators and of their products are also briefly discussed.

I. INTRODUCTION

The unitary group approach (UGA) to the many-electron correlation problem[1–3] provides a versatile formalism enabling an efficient exploitation of the invariance properties of a nonrelativistic, clamped-nuclei, electronic Hamiltonian in quantum-chemical calculations of the molecular electronic structure. It represents an outgrowth of the formalism initially laid down by Moshinsky[4] in the context of the nuclear shell model and is based on the fact that the spin-independent many-electron Hamiltonian may be expressed as a bilinear form in the spin-free (orbital) $U(n)$ generators. It enabled the development of efficient methods for the evaluation of Hamiltonian matrix elements using the algorithms for matrix elements of $U(n)$ generators[1–3] and of products of generators[5] and proved to be particularly useful in large scale quantum-chemical configuration interaction (shell model) calculations.[6–11] A detailed account of these developments can be found in numerous reviews[12–15] and monographs.[16,17]

Thus, from the viewpoint of the many-electron (and, generally, many-fermion) problem, it is essential to develop efficient and versatile algorithms for the evaluation of matrix elements of the $U(n)$ generators and of their products. In the UGA formalism it is traditional to adopt the Gel'fand–Tsetlin basis although it is sometimes more convenient to adopt other bases.[18–20] (This is particularly true for the "group" function type approaches[19,20] when a molecular wave function is built from the wave functions of subsystems.) This problem dates back to the original work of Gel'fand and Tsetlin[21] and Baird and Biedenharn[22] who developed explicit formulas for matrix elements of all $U(n)$ generators in the Gel'fand–Tsetlin basis. More recently an alternative algebraic approach to this problem was developed.[23] The general formalism of Refs. 21–23 considerably simplifies for the many-electron problem since at most two-column irreps of $U(n)$ need be considered. This simplification[1,2] together with its graphical representation[3] has led to the development of numerous computational implementations.[6–11]

Recently a new approach to the evaluation of the $U(n)$ generator matrix elements was undertaken,[24,25] following an earlier work of Nikam and Sarma,[26] which exploits the imbedding of $U(n)$ in the much larger group $U(2^n)$ via the subgroups $SO(2n+1)$ and $SO(2n)$. The main advantage of this approach,[24,25] which is referred to as a Clifford algebra UGA (CAUGA) in view of the role played by spin representations, is that the $U(n)$ generator matrix elements may be efficiently evaluated by exploiting the simple action of the $U(2^n)$ generators on the basis states of the totally symmetric tensor representations. This approach also enables the treatment of particle-number-nonconserving operators and coupling schemes other than the canonical Gel'fand–Tsetlin scheme,[20] thus providing a greater flexibility in the construction of many-electron bases and in the development of pertinent computational algorithms in general.

It is our aim here to investigate this problem from the viewpoint of parastatistics, first introduced by Green[27] as a generalized method of field quantization that includes normal Fermi and Bose statistics as a special case. (For a detailed account of parastatistics and its applications in quantum field theory see Ref. 28 and references cited therein.) We shall demonstrate that, from the viewpoint of the $U(n)$ generator matrix element evaluation, it suffices to consider the pth spinor representation of $SO(2n+1)$, which always occurs exactly once in the symmetric pth-rank tensor representation of $U(2^n)$. This in turn enables the CAUGA formalism to be described in terms of the para-Fermi algebra. We have previously demonstrated[29] that parafermions of order 2 occur naturally in the spin-independent many-electron correlation problem, where this approach provides additional flexibility and convenience. However, it is felt that the same should hold for a general many-fermion problem since the CAUGA imbedding enables one to reduce the evaluation of $U(n)$ generator matrix elements for any p-column irrep to that for the totally symmetric p-box irrep $[p,0]$ of $U(2^n)$.

II. THE CAUGA FORMALISM

In the Clifford algebra unitary group approach[20,24,25] (CAUGA) we exploit a realization of the spinor algebra of the rotation group $SO(2n+1)$ in the covering algebra of $U(2n)$ to obtain explicit representation matrices for the

[a)] Permanent address: Department of Applied Mathematics, Department of Chemistry and (GWC)^2, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

 0022-2488/87/102304-06$02.50 © 1987 American Institute of Physics

$SO(2n+1)$ [or $SO(2n)$ or $U(n)$] generators in the basis symmetry adapted to the chain

$$U(2^n) \supset SO(2n+1) \supset SO(2n) \supset U(n),$$

supplemented, when desired, by the canonical Gel'fand–Tsetlin chain.

The above chain has been first employed by Moshinsky and Quesne[30] in connection with a study of dynamical or noninvariance groups for a general n-level many-fermion system and of the concept of complementary subgroups within a given irrep of a larger group. The $SO(2n+1)$ part of the chain was introduced even earlier by Helmers[31] and Judd.[32]

To achieve a desired realization of the $SO(2n+1)$ generators that also establishes the relationship with parastatistics formulation we employ the second quantization formalism and introduce p sets of fermion annihilation operators a_i^α ($\alpha = 1,...,p$, $i = 1,...,n$), which satisfy the commutation and anticommutation relations

$$\begin{aligned} &[a_i^\alpha, a_j^\beta] = [a_i^\alpha, a^{\dagger\beta}_j] = 0, \quad \alpha \neq \beta, \\ &\{a_i^\alpha, a_j^\alpha\} = 0, \quad \{a_i^\alpha, a^{\dagger\alpha}_j\} = \delta_{ij}, \end{aligned} \tag{1}$$

together with relations conjugate to these. Throughout, we assume the existence of a unique vacuum state $|0\rangle$ on which all the fermion annihilation operators a_i^α vanish, i.e.,

$$a_i^\alpha |0\rangle = 0, \quad \alpha = 1,...,p, \quad i = 1,...,n.$$

We shall find it convenient to define fermion operators a_ρ^α ($\rho = 1,...,2n$), according to the convention

$$a_{\bar{i}}^\alpha = a^{\dagger\alpha}_i, \quad \bar{i} = i + n \quad (i = 1,...,n).$$

With this notation the relations of Eq. (1) may be conveniently expressed as

$$[a_\rho^\alpha, a_\sigma^\beta] = 0, \quad \alpha \neq \beta, \quad \{a_\rho^\alpha, a_\sigma^\alpha\} = g_{\rho\sigma}, \tag{2}$$

where $g_{\rho\sigma}$ is the (symmetric) $SO(2n)$ metric defined by

$$g_{\rho\sigma} = \begin{cases} 1, & \text{if } |\rho - \sigma| = n, \\ 0, & \text{otherwise.} \end{cases}$$

If $g^{\rho\sigma}$ ($= g_{\rho\sigma}$) denotes the inverse metric, we may raise and lower the indices according to

$$(a^\alpha)^\rho = g^{\rho\sigma} a_\sigma^\alpha, \quad a_\rho^\alpha = g_{\rho\sigma}(a^\alpha)^\sigma, \quad \text{etc.} \tag{3}$$

For each $\alpha = 1,...,p$ the operators $\frac{1}{2}[a_\mu^\alpha, a_\nu^\alpha]$ and $a_\mu^\alpha/\sqrt{2}$ ($\mu,\nu = 1,...,2n$) form the generators of an $SO(2n+1)$ group,[23,33] herein denoted as $SO_\alpha(2n+1)$. We let $SO(2n+1)$ denote the diagonal subgroup of

$$\bigotimes_{\alpha=1}^{p} SO_\alpha(2n+1).$$

This is, in fact, the $SO(2n+1)$ group considered earlier.[24–26] The generators of the $U(n)$ subgroup of $SO(2n+1)$ are then expressible as[29]

$$E_j^i = \sum_{\alpha=1}^{p} a^{\dagger\alpha}_i a_j^\alpha. \tag{4}$$

We denote the Hilbert space of all polynomials in the fermion creation operators $a^{\dagger\alpha}_i$ ($\alpha = 1,...,p$) acting on the vacuum state $|0\rangle$ by $\mathscr{H}$. The set of all polynomials in the fermion creation operators $a^{\dagger\alpha}_i$ (of a fixed type α) acting on the vacuum state $|0\rangle$, herein denoted $\mathscr{H}_\alpha$, constitutes the 2^n-dimensional fundamental spinor representation of $SO_\alpha(2n+1)$. At the same time, the space $\mathscr{H}_\alpha$ constitutes the vector representation of a $U(2^n)$ group,[24–26] herein denoted $U_\alpha(2^n)$. The full space of states $\mathscr{H}$ may then be identified as the tensor product space

$$\mathscr{H} = \mathscr{H}_1 \otimes \mathscr{H}_2 \otimes \cdots \otimes \mathscr{H}_p, \tag{5}$$

with

$$\dim \mathscr{H} = 2^{np}.$$

The space of states $\mathscr{H}$ clearly carries a reducible representation (by tensors of rank p) of the diagonal subgroup $U(2^n)$ of

$$\bigotimes_{\alpha=1}^{p} U_\alpha(2^n).$$

The CAUGA formalism exploits the fact that all irreps of $U(n)$ with at most p columns in the Young tableau occur (at *least* once) in the pth fundamental symmetric tensor rep $[p,\dot{0}]$ of $U(2^n)$. Thus the fully symmetric component $\mathscr{H}_s$ of the tensor product space $\mathscr{H}$, Eq. (5), plays a fundamental role in CAUGA. Clearly, the space $\mathscr{H}_s$ is spanned by all polynomials in the fermion creation operators $a^{\dagger\alpha}_i$, acting on the vacuum state $|0\rangle$, which are totally symmetric in the superscripts α, and carries the symmetric pth-rank tensor rep $[p,\dot{0}]$ of $U(2^n)$. It should be remarked that we could also employ other components of the tensor product space $\mathscr{H}$ such as, for example, the antisymmetric rep $[1^p,\dot{0}]$ of $U(2^n)$. However, we are only guaranteed to get all the $U(n)$ irreps with at most p columns in the Young tableau if we employ the fully symmetric component $\mathscr{H}_s$ of $\mathscr{H}$ (cf. Refs. 20, 24, and 25).

The space $\mathscr{H}_s$ constitutes a reducible rep of the subgroup $SO(2n+1)$ of $U(2^n)$. The branching rules for the reduction of $\mathscr{H}_s$ into the irreps of $SO(2n+1)$ are clearly given by the reduction of the symmetrized pth power of the fundamental spinor rep $(\frac{1}{2},\frac{1}{2},...,\frac{1}{2})$ of $SO(2n+1)$. In general, the $U(2^n)\downarrow SO(2n+1)$ branching rules present a formidable problem, which dates back to the pioneering work of Brauer and Weyl,[34] Murnagham,[35] and Littlewood.[36] These earlier results have been recently extended by Butler and Wybourne[37] and King *et al.*[38,39]

The $U(2^n)\downarrow SO(2n+1)$ branching rules for the $[2,\dot{0}]$ and $[1^2,\dot{0}]$ irreps of $U(2^n)$, which are relevant to CAUGA, were shown by Brauer and Weyl[34] (see also King *et al.*[39]) to be

$$\begin{aligned} [2,\dot{0}]\downarrow SO(2n+1) &= (1^n) \oplus \Big\{ \bigoplus_r [(1^{n-3-4r},\dot{0}) \oplus (1^{n-4-4r},\dot{0})] \Big\}, \\ [1^2,\dot{0}]\downarrow SO(2n+1) &= \bigoplus_r [(1^{n-1-4r},\dot{0}) \oplus (1^{n-2-4r},\dot{0})]. \end{aligned} \tag{6}$$

These results are sufficient for the many-electron problem where it suffices to consider[20,24] the $[2,\dot{0}]$ irrep of $U(2^n)$. The process, however, quickly becomes complicated for higher tensor reps of $U(2^n)$ although a general prescription for obtaining the $U(2^n)\downarrow SO(2n+1)$ branching rules for the symmetric reps of $U(2^n)$ has been outlined by King *et al.*[39]

Although the general $U(2^n)\downarrow SO(2n+1)$ rules are complicated it can be shown that the symmetric irrep $[p,\dot{0}]$ of $U(2^n)$ contains precisely one representation of $SO(2n+1)$ with the highest weight $(\dot{p}/2) \equiv (p/2,p/2,\ldots,p/2)$. This rep of $SO(2n+1)$ will be explicitly constructed in the subsequent section using the methods of parastatistics. Using the representation theory for the para-Fermi algebras we shall find that, from the viewpoint of the evaluation of $U(n)$ matrix elements, it suffices to restrict ourselves to the irreducible $SO(2n+1)$ subrepresentations $(\dot{p}/2)$ of the (symmetrized) space $\mathscr{H}_s$.

In this context we should note that the dimension of the $SO(2n+1)$ irrep $(\dot{p}/2)$ given by Bracken and Green[33] (second equation on p. 353 of Ref. 33) is, in fact, the dimension of the totally symmetric p-column irrep $[p,\dot{0}]$ of $U(2^n)$,

$$\dim[p,\dot{0}]_{U(2^n)} = \binom{2^n+p-1}{p}. \tag{7}$$

The dimension of $(\dot{p}/2)$ of $SO(2n+1)$ cannot be easily expressed by a simple formula. Designating this dimension by D_p we can find, however, the recursion formulas

$$D_p = 2^n\left[\binom{2k}{k}\binom{n+2k}{k}^{-1}\right]D_{p-1},$$
$$\text{when } p = 2k+1 \text{ (odd)}, \tag{8a}$$

and

$$D_p = 2^{-n}\left[\binom{2(k+n)-1}{n}\binom{n+k-1}{n}^{-1}\right]D_{p-1},$$
$$\text{when } p = 2k \text{ (even)}. \tag{8b}$$

Since

$$D_0 = \mathrm{Dim}(\dot{0}) = 1, \tag{9a}$$

we find easily that

$$D_1 = 2^n, \tag{9b}$$

$$D_2 = \binom{2n+1}{n}, \tag{9c}$$

$$D_3 = \frac{2^{n+1}}{n+2}\binom{2n+1}{n}, \tag{9d}$$

$$D_4 = \binom{2n+3}{n}\binom{2n+1}{n}\binom{n+2}{n}^{-1}, \text{ etc.} \tag{9e}$$

It is also instructive to examine the actual subduction for the irreps $[2,\dot{0}]$ and $[1^2,\dot{0}]$ of $U(2^n)$ to both $SO(2n+1)$ and $U(n)$ subgroups. [The latter can be easily obtained by considering the symmetric and antisymmetric components of the second tensor power $([0]\oplus[1]\oplus[1^2]\oplus\cdots\oplus[1^n])^{\otimes 2}$ using the Littlewood–Richardson rules.] We note that only in the lowest-dimensional case, i.e., $U(2)\supset SO(3)\supset SO(2)\supset U(1)$, are both subductions multiplicity-free since $[2,0]\downarrow SO(3) = (1)$ and $[1^2]\downarrow SO(3) = (0)$ with corresponding dimensions 3 and 1, respectively. Subducing to the Abelian group $U(1)$ we have $[2,0]\downarrow U(1) = [0] \dot{+} [1] \dot{+} [2]$ and $[1^2]\downarrow U(1) = [1]$. The case $n = 2$ yields already two irreps for the $[1^2,\dot{0}]\downarrow SO(5)$ subduction, namely $(1) \dot{+} (0)$, while $[2,\dot{0}]\downarrow SO(5) = (1^2)$. Thus $[2,\dot{0}]\downarrow U(2) = [0] \dot{+} [1] \dot{+} [1^2] \dot{+} [2] \dot{+} [2,1] \dot{+} [2^2]$ and $[1^2,\dot{0}] = [1] + 2[1^2] + [2,1]$. However, starting with $n = 3$ both subductions yield multiple irreps. We present below cases $n = 3$ and 5 as examples. Note that while the symmetric irrep $[2,\dot{0}]$ of $U(2^n)$ contains all possible two-columned irreps of $U(n)$, this is not the case for the $[1^2,\dot{0}]$ irrep of $U(2^n)$, which has the irreps of the type $[2^r]$ (and the scalar irrep $[\dot{0}]$) missing. For simplicity we drop zeros in the $SO(2n+1)$ and $U(n)$ irrep labels. The dimensions of the respective irreps are indicated in parentheses.

<u>$n=3$:</u> $U(8)\supset SO(7)\supset SO(6)\supset U(3)$:

$$[2,\dot{0}]\downarrow SO(7) = \underset{(36)}{(1^3)} \dot{+} \underset{(35)}{(0)}\ \underset{(1)}{},$$

$$[1^2,\dot{0}]\downarrow SO(7) = \underset{(28)}{(1^2)} \dot{+} \underset{(21)}{(1)}\ \underset{(7)}{},$$

$$[2,\dot{0}]\downarrow U(3) = \underset{(36)}{[0]} \dot{+} \underset{(1)}{[1]} + \underset{(3)}{[1^2]} + \underset{(3)}{2[1^3]} \dot{+} \underset{2\times(1)}{[2]} \dot{+} \underset{(6)}{[2,1]} \dot{+} \underset{(8)}{[2,1^2]} \dot{+} \underset{(3)}{[2^2]} \dot{+} \underset{(6)}{[2^2,1]} \dot{+} \underset{(3)}{[2^3]} \dot{+} \underset{(1)}{},$$

$$[1^2,\dot{0}]\downarrow U(3) = \underset{(28)}{[1]} \dot{+} \underset{(3)}{2[1^2]} \dot{+} \underset{2\times(3)}{2[1^3]} \dot{+} \underset{2\times(1)}{[2,1]} \dot{+} \underset{(8)}{2[2,1^2]} \dot{+} \underset{2\times(3)}{[2^2,1]}.$$

<u>$n=5$:</u> $U(32)\supset SO(11)\supset SO(10)\supset U(5)$:

$$[2,\dot{0}]\downarrow SO(11) = \underset{(528)}{(1^5)} \dot{+} \underset{(462)}{(1^2)} \dot{+} \underset{(55)}{(1)}\ \underset{(11)}{},$$

$$[1^2,\dot{0}]\downarrow SO(11) = \underset{(496)}{(1^4)} \dot{+} \underset{(330)}{(1^3)} \dot{+} \underset{(165)}{(0)}\ \underset{(1)}{},$$

$$[2,\dot{0}]\downarrow U(5) = [0] \dot{+} [1] \dot{+} [1^2] \dot{+} 2[1^3] \dot{+} 3[1^4] \dot{+} 3[1^5] \dot{+} [2] \dot{+} [2,1] \dot{+} [2,1^2] \dot{+} 2[2,1^3] \dot{+} 3[2,1^4]$$
$$(528) \quad (1) \quad (5) \quad (10) \quad 2\times(10) \quad 3\times(5) \quad 3\times(1) \quad (15) \quad (40) \quad (45) \quad 2\times(24) \quad 3\times(5)$$
$$\dot{+} [2^2] \dot{+} [2^2,1] \dot{+} [2^2,1^2] \dot{+} 2[2^2,1^3] \dot{+} [2^3] \dot{+} [2^3,1] \dot{+} [2^3,1^2] \dot{+} [2^4] \dot{+} [2^4,1] \dot{+} [2^5],$$
$$(50) \quad (75) \quad (45) \quad 2\times(10) \quad (50) \quad (40) \quad (10) \quad (15) \quad (5) \quad (1)$$

$$[1^2,\dot{0}]\downarrow U(5) = [1] \dot{+} 2[1^2] \dot{+} 2[1^3] \dot{+} 2[1^4] \dot{+} 3[1^5] \dot{+} [2,1] \dot{+} 2[2,1^2] \dot{+} 2[2,1^3] \dot{+} 2[2,1^4] \dot{+} [2^2,1]$$
$$(496) \quad (5) \quad 2\times(10) \quad 2\times(10) \quad 2\times(5) \quad 3\times(1) \quad (40) \quad 2\times(45) \quad 2\times(24) \quad 2\times(5) \quad (75)$$
$$\dot{+} 2[2^2,1^2] \dot{+} 2[2^2,1^3] \dot{+} [2^3,1] \dot{+} 2[2^3,1^2] \dot{+} [2^4,1].$$
$$2\times(45) \quad 2\times(10) \quad (40) \quad 2\times(10) \quad (5)$$

Note that the dimension of the irrep $(\dot{p}/2)$ of $SO(2n+1)$ for the case $p=2$, i.e., of the irrep (1^n) of $SO(2n+1)$, which is always contained in the irrep $[2,\dot{0}]$ of $U(2^n)$ [cf. Eq. (6)] and whose dimension is given by Eq. (9c), i.e.,

$$n=3, \quad \dim(1^3) = \binom{7}{3} = 35,$$

$$n=5, \quad \dim(1^5) = \binom{11}{5} = 462, \text{ etc.},$$

equals the sum of the dimensions of all the two-column irreps of $U(n)$, as may be easily verified in the case of the examples given above.

III. PARA-FERMI ALGEBRAS

Following the ansatz prescribed by Green[27] we define the operators a_ρ $(\rho = 1,...,2n)$,

$$a_\rho = \sum_{\alpha=1}^{p} a_\rho^\alpha, \tag{10}$$

which satisfy the relations

$$[a_\rho,[a_\mu,a_\nu]] = 2(g_{\rho\mu}a_\nu - g_{\rho\nu}a_\mu), \tag{11a}$$

$$a_i a^j|0\rangle = p\delta_i^j|0\rangle,$$

$$i,j = 1,...,n, \quad \mu,\nu,\rho = 1,...,2n, \tag{11b}$$

where we raise indices in accordance with Eq. (3); i.e.,

$$a^\rho = g^{\rho\sigma}a_\sigma, \text{ etc.} \tag{12}$$

Equations (11a) and (11b) are the defining relations for parafermions of order p.[27,33] We note that parafermions of order 1 correspond to normal fermions.

In the following we denote the space of all polynomials in the para-Fermi creation operators $a^i = a_{\bar{i}}$ $(i = 1,...,n)$ acting on the vacuum state $|0\rangle$ (i.e., the para-Fermi Fock space) by $\mathscr{F}_p$. Since the para-Fermi creation operators, as defined by Eq. (10), are symmetric in superscripts α, it follows that the para-Fermi Fock space $\mathscr{F}_p$ is contained in the fully symmetric component $\mathscr{H}_s$ of the full space of $\mathscr{H}$.

Using commutation relations of Eq. (11a) it may be easily verified that the operators

$$\alpha_{\mu\nu} = \tfrac{1}{2}[a_\mu,a_\nu] \tag{13}$$

satisfy the following commutation relations:

$$[\alpha_{\mu\nu},\alpha_{\rho\sigma}] = g_{\nu\rho}\alpha_{\mu\sigma} + g_{\mu\sigma}\alpha_{\nu\rho} - g_{\mu\rho}\alpha_{\nu\sigma} - g_{\nu\sigma}\alpha_{\mu\rho}, \tag{14a}$$

$$[\alpha_{\mu\nu},a_\rho] = g_{\rho\nu}a_\mu - g_{\rho\mu}a_\nu. \tag{14b}$$

The commutation relations of Eq. (14a) show that the operators (13) constitute the generators of the group $SO(2n)$ while the operators

$$\alpha_{\mu\nu}, \quad a_\rho/\sqrt{2} \tag{15}$$

constitute the generators of the group $SO(2n+1)$. It thus follows that the unirreps (unitary irreps) of the para-Fermi algebra are to comprise finite-dimensional irreps of $SO(2n+1)$. The representation carried by the Fock space $\mathscr{F}_p$, for parastatistics of order p is uniquely characterized as that representation which admits a unique vacuum state satisfying the conditions of Eq. (11b). We remark that it is also possible to introduce the pseudo-orthogonal group $SO(2n+1,1)$ into the parafermion algebra,[29] although this will not be done in the present treatment.

The $SO(2n+1)$ group with infinitesimal generators (15) constitutes the $SO(2n+1)$ subgroup of $U(2^n)$ employed by Sarma *et al.*[24–26] As a Cartan subalgebra of $SO(2n+1)$ [and of $SO(2n)$] we choose the operators

$$h_i = \alpha_i^i, \quad i = 1,...,n, \tag{16}$$

which serve to uniquely label the weights of $SO(2n+1)$. It is easily seen[33] that the vacuum state $|0\rangle$ constitutes a minimal weight state of $SO(2n+1)$ weight

$$(-p/2,-p/2,...,-p/2).$$

It thus follows that the space of para-Fermi states $\mathscr{F}_p \subseteq \mathscr{H}_s$ carries the $SO(2n+1)$ irrep with the highest weight

$$(p/2,p/2,...,p/2) \equiv (\dot{p}/2). \tag{17}$$

This is precisely the $SO(2n+1)$ irrep referred to in Sec. II.

The para-Fermi number-preserving operators

$$b_j^i = \tfrac{1}{2}[a^i,a_j] \tag{18}$$

constitute the generators of the unitary subgroup $U(n)$ of $SO(2n)$. Following Refs. 29 and 33 we work instead with the shifted $U(n)$ generators (cf., also, Refs. 20 and 24)

$$E_j^i = \tfrac{1}{2}[a^i,a_j] + \tfrac{1}{2}p\delta_j^i, \tag{19}$$

where p is the order of parastatistics. It is easily demonstrated[29] that the $U(n)$ generators (19) agree with the prescription of Eq. (4) and hence constitute the generators of the $U(n)$ subgroup of $U(2^n)$ considered in CAUGA,[24–26] as required. The para-Fermi Fock space $\mathscr{F}_p$ possesses the remarkable property that it decomposes into a direct sum of $U(n)$ irreps of the form $\lambda = (\lambda_1,\lambda_2,...,\lambda_n)$ with

$$p \geqslant \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0, \tag{20}$$

and that all such irreps occur exactly once.[28,33] In other words, all irreps of $U(n)$ with no more than p columns in the

Young tableau occur in $\mathcal{F}_p$ with unit multiplicity.

Since $\mathcal{F}_p \subseteq \mathcal{H}_s$, the above results imply that all p-columned representations of $U(n)$ occur in the $[p,\dot{0}]$ irrep of $U(2^n)$ at least once, as noted earlier in Refs. 24 and 25. In fact, our approach demonstrates an even stronger result, namely that from the viewpoint of $U(n)$ matrix element evaluation it suffices to work in the subspace $\mathcal{F}_p \subseteq \mathcal{H}_s$ corresponding to parastatistics of order p.

IV. CONCLUSIONS

We have shown that the CAUGA formalism (or its generalization to fermions with p internal degrees of freedom) may be described in terms of the para-Fermi algebra of order p. This adds a new insight into the CAUGA and opens up the possibility of exploiting the existing extensive work on para-Fermi algebras (cf., e.g., Ref. 28). It is believed that the CAUGA approach is of particular relevance for the spin-independent many-electron correlation problem where the second-order para-Fermi creation and annihilation operators, corresponding to the creation and annihilation of spin-averaged paraparticles, occur naturally.[29]

Finally, we remark that from the viewpoint of Clifford algebras, an alternative (but equivalent) representation of the para-Fermi algebra (11) may be given in terms of the elements of a generalized Clifford algebra (cf. Ramakrishnan[40])

$$\beta_j = a_j + a^j, \quad \beta_{\bar{j}} = i(a_j - a^j), \quad \bar{j} = j + n, \quad j = 1,...,n. \tag{21}$$

This defines Clifford algebra elements β_ρ for $\rho = 1,...,2n$. It is easily verified, using the para-Fermi relations (11), that the Clifford algebra elements of Eq. (21) satisfy the generalized Clifford algebra relations

$$[\beta_\rho,[\beta_\mu,\beta_\nu]] = 4(\delta_{\rho\mu}\beta_\nu - \delta_{\rho\nu}\beta_\mu). \tag{22}$$

The relations of Eq. (14) show that the operators

$$\beta_{\mu\nu} = \tfrac{1}{4}[\beta_\mu,\beta_\nu] \tag{23}$$

satisfy the relations

$$\begin{aligned} [\beta_{\mu\nu},\beta_\rho] &= \delta_{\rho\nu}\beta_\mu - \delta_{\rho\mu}\beta_\nu, \\ [\beta_{\mu\nu},\beta_{\rho\sigma}] &= \delta_{\rho\nu}\beta_{\mu\sigma} - \delta_{\mu\sigma}\beta_{\rho\nu} - \delta_{\rho\mu}\beta_{\nu\sigma} + \delta_{\nu\sigma}\beta_{\rho\mu}, \end{aligned} \tag{24}$$

which we recognize as the $SO(2n+1)$ commutation relations, with choice of metric $g_{RS} = \delta_{RS}$ $(R,S = 1,...,2n+1)$. The elements of a normal Clifford algebra,[40] which corresponds to a generalized Clifford algebra of order 1, satisfy the relations of Eqs. (22) and (24), as required.

ACKNOWLEDGMENTS

The work presented in this paper has been kindly supported by an ARGS research grant (M.D.G.) and an NSERC grant-in-aid of research (J.P.), which are hereby gratefully acknowledged.

[1] J. Paldus, J. Chem. Phys. **61**, 5321 (1974); Int. J. Quantum Chem., Symp. **9**, 165 (1975).
[2] J. Paldus, in *Theoretical Chemistry: Advances and Perspectives*, edited by H. Eyring and D. J. Henderson (Academic, New York, 1976), Vol. 2, pp. 131–290.
[3] I. Shavitt, Int. J. Quantum Chem., Symp. **11**, 131 (1977); **12**, 5 (1978).
[4] M. Moshinsky, in *Many-Body Problems and Other Selected Topics in Theoretical Physics*, edited by M. Moshinsky, T. A. Brody, and G. Jacob (Gordon and Breach, New York, 1966), p. 289 [also published as a separate book: *Group Theory and the Many-Body Problem* (Gordon and Breach, New York, 1968)].
[5] J. Paldus and M. J. Boyle, Phys. Scr. **21**, 295 (1980).
[6] M. J. Downward and M. A. Robb, Theor. Chim. Acta **46**, 129 (1977); D. Hegarty and M. A. Robb, Mol. Phys. **38**, 1795 (1979).
[7] B. R. Brooks and H. F. Schaefer, J. Chem. Phys. **70**, 5092 (1979).
[8] P. E. M. Siegbahn, J. Chem. Phys. **72**, 1647 (1980).
[9] H. Lischka, R. Shepard, F. Brown, and I. Shavitt, Int. J. Quantum Chem., Symp. **15**, 91 (1981).
[10] P. Saxe, D. J. Fox, H. F. Schaefer, and N. C. Handy, J. Chem. Phys. **77**, 5584 (1982).
[11] V. R. Saunders and J. H. van Lenthe, Mol. Phys. **48**, 923 (1983).
[12] *The Unitary Group for the Evaluation of Electronic Energy Matrix Elements, Lecture Notes in Chemistry*, Vol. 22, edited by J. Hinze (Springer, Berlin, 1981).
[13] M. A. Robb and U. Niazi, Comp. Phys. Rep. **1**, 127 (1984).
[14] W. Duch and J. Karwowski, Comp. Phys. Rep. **2**, 93 (1985).
[15] J. Paldus, in *Symmetries in Science II*, edited by B. Gruber and R. Lenczewski (Plenum, New York, 1986), pp. 429–446.
[16] R. Pauncz, *Spin Eigenfunctions: Construction and Use* (Plenum, New York, 1979).
[17] S. Wilson, *Electron Correlation in Molecules* (Clarendon, Oxford, 1984), Chap. 5.
[18] P. E. S. Wormer, in *Electron Correlation: Proceedings of the Daresbury Study Weekend*, edited by M. F. Guest and S. Wilson (Science Research Council, Daresbury Laboratory, Warrington, U. K., 1980), p. 49.
[19] M. D. Gould and J. Paldus, Int. J. Quantum Chem. **30**, 327 (1986); M. D. Gould, *ibid.* **30**, 365 (1986).
[20] J. Paldus, M. J. Gao, and J. Q. Chen, Phys. Rev. A **35**, 3197 (1987).
[21] I. M. Gel'fand and M. L. Tsetlin, Dokl. Akad. Nauk SSSR **71**, 825, 1070 (1950).
[22] G. E. Baird and L. C. Biedenharn, J. Math. Phys. **4**, 1449 (1963).
[23] M. D. Gould, J. Math. Phys. **21**, 444 (1980); **22**, 15, 2376 (1981).
[24] J. Paldus and C. R. Sarma, J. Chem. Phys. **83**, 5135 (1985).
[25] C. R. Sarma and J. Paldus, J. Math. Phys. **26**, 1140 (1985).
[26] R. S. Nikam and C. R. Sarma, J. Math. Phys. **25**, 1199 (1984).
[27] H. S. Green, Phys. Rev. **90**, 270 (1953); Prog. Theor. Phys. (Kyoto) **47**, 1400 (1972).
[28] Y. Ohnuki and S. Kamefuchi, *Quantum Field Theory and Parastatistics* (Springer, New York, 1982).
[29] M. D. Gould and J. Paldus, Phys. Rev. A **34**, 804 (1986).
[30] M. Moshinsky and C. Quesne, J. Math. Phys. **11**, 1631 (1970).
[31] K. Helmers, Nucl. Phys. **23**, 594 (1961).
[32] B. R. Judd, in *Group Theory and its Applications*, Vol. I, edited by E. M. Loebl (Academic, New York, 1968), pp. 183–220.
[33] A. J. Bracken and H. S. Green, Nuovo Cimento A **9**, 349 (1972).
[34] R. Brauer and H. Weyl, Am. J. Math. **57**, 425 (1935).
[35] F. D. Murnaghan, *The Theory of Group Representations* (Johns Hopkins, Baltimore, 1938).
[36] D. E. Littlewood, Proc. London Math. Soc. **49**, 307 (1947); **50**, 349

(1948); *The Theory of Group Characters* (Clarendon, Oxford, 1950).
[37]P. H. Butler and B. G. Wybourne, J. Phys. (Paris) **30**, 655 (1969).
[38]R. C. King, *Lecture Notes in Physics*, Vol. 50 (Springer, New York, 1975), p. 481; J. Phys. A **8**, 429 (1975).
[39]R. C. King, L. Dehuai, and B. G. Wybourne, J. Phys. A (Math. Gen.) **14**, 2509 (1981).
[40]A. Ramakrishnan, *L-Matrix Theory or the Grammar of Dirac Matrices* (McGraw-Hill, Bombay, 1972).

On the initial value problem for a class of nonlinear integral evolution equations including the sine–Hilbert equation

P. M. Santini,[a] M. J. Ablowitz, and A. S. Fokas
Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13676

(Received 14 October 1986; accepted for publication 27 May 1987)

A method for solving a class of nonlinear singular integral evolution equations for decaying initial values on the line is presented. The underlying scattering problem is a matrix Riemann–Hilbert problem. Scattering analysis shows that the spectrum is purely discrete. An application is to the so-called sine–Hilbert equation $H\theta_t = -c \sin\theta$, where c is a constant and H denotes the Hilbert transform.

I. INTRODUCTION

The inverse scattering (or spectral) transform (IST) method has been shown to be a powerful tool for solving suitable initial value problems for certain nonlinear evolution equations (see, for example, Refs. 1 and 2). All of the physically interesting equations solvable by the inverse scattering transform take on a very simple form. Indeed often the work of the asymptologist is to derive special, simple equations in a suitable asymptotic zone. The governing equation here [see (5) below] is a particularly simple one. Although Eq. (5) has not yet appeared in a concrete physical situation, it provides a simple solvable model for a nonlinear evolution equation with a dispersion relation $w(k) = \alpha\,\mathrm{sgn}(k)$. Perhaps publication of this work might motivate asymptologists to look for such a system. It should be pointed out that the mathematical structure of the direct and inverse scattering problem associated with this nonlinear equation is quite different. Given the fact that the inverse scattering transform is associated with many physically relevant problems, we feel that researchers would want to be knowledgeable about such novel features of related problems. Indeed it can be expected that such situations would arise in other problems as well.

II. THEORY

In this paper we present the IST associated with the following class of matrix nonlinear evolution equations[3,4]:

$$Q_t = \sigma_3 p(L)Q, \quad Q = Q(x,t), \tag{1}$$

where

$$LF \equiv i\sigma_3\{\sqrt{1+Q^2}HF - \tfrac{1}{2}QH([Q,F]/\sqrt{1+Q^2})\},$$
$$\sqrt{1+Q^2} = \sqrt{1+q_{12}q_{21}}. \tag{2}$$

Here F and Q are off-diagonal 2×2 matrices, $\sigma_3 = \mathrm{diag}(1,-1)$ is the usual Pauli spin matrix, $p(y)$ is an arbitrary polynomial in y, [,] is the usual commutator, and

$$(Hf)(x) \equiv \frac{1}{\pi} \int\!\!\!\!\!\!- _{-\infty}^{\infty} dy (y-x)^{-1} f(y) \tag{3}$$

is the Hilbert transform.

Equation (1) is the first known example of a class of nonlinear evolution equations which are purely integral in space. It is solvable via a purely local Riemann–Hilbert spectral problem. Its introduction[3] was originally motivated by the discovery that physically relevant integrodifferential equations such as the intermediate long wave equation[5–9] are solvable via differential Riemann–Hilbert IST schemes.[10–12]

It is known that (1) possesses an infinite family of conservation laws.[3] The derivation of this result requires only the use of the elementary properties of the associated scattering problem.

As alluded to above, this scattering problem is an example of a pure (nondifferential) Riemann–Hilbert (RH) problem[13] in configurational space x. The solution exhibits, as we will see below, a new type (for problems on the infinite line) of singularity structure in the spectral variable z, consisting only of polar singularities clustering at finite points of the z plane.

Before describing these properties in detail, we briefly discuss the first element of class (1), obtained by choosing $p(y) = -icy$,

$$Q_{12_t} = c\sqrt{1+Q_{12}Q_{21}}HQ_{12}, \tag{4a}$$

$$Q_{21_t} = c\sqrt{1+Q_{12}Q_{21}}HQ_{21}, \tag{4b}$$

which, in the obvious reduction $Q_{12} = Q_{21} = v$, can be written in the following suggestive form (taking account of the property $H^2 = -1$):

$$H\theta_t = -c \sin\theta, \quad v(x,t) = i \sin\theta(x,t). \tag{5}$$

The analogy between Eq. (5) and the sine–Gordon equation $\theta_{xt} = \sin\theta$ led us to refer to it as the sine–Hilbert equation.[4] We also note in pasing the compelling analog with the Korteweg–de Vries equation and the Benjamin–Ono equation whereby going from the KdV equation $u_t + uu_x + u_{xxx} = 0$ to the Benjamin–Ono equation $u_t + uu_x + Hu_{xx} = 0$ we simply replace one of the x derivatives by the Hilbert transform H.

The solution of the initial value problem associated with the linearized version

$$H\theta_t = -c\theta \tag{6}$$

of Eq. (5) is given by

$$\theta(x,t) = e^{ict}A^+(x) + e^{-ict}A^-(x), \tag{7}$$

where $A^{\pm}(x)$, defined by

[a] Permanent address: Dipartimento di Fisica, Universitá di Roma, La Sapienza, 00185 Roma, Italy.

 0022-2488/87/102310-07$02.50 © 1987 American Institute of Physics

$$A^{\pm}(x) \equiv \pm \int_0^{\pm\infty} \frac{dk}{2\pi} \hat{\theta}(k,0) e^{ikx}, \tag{8}$$

are the boundary values on the line $\mathrm{Im}\, x = 0$ of functions holomorphic in the upper and, respectively, lower half x plane, and $\hat{\theta}(k,0)$ is the Fourier transform of the (localized) initial condition $u(x,0)$.

The peculiar time dependence of the general solution (7) corresponds to the dispersion relation

$$\omega(k) = -c\,\mathrm{sgn}(k), \tag{9}$$

in which the frequency ω depends only on the sign of the wave number k and not on its magnitude. The independence of ω from $|k|$ is, for instance, a property of the wave propagation in some fluid dynamical systems[14] and seems to allow the possibility of a physical interpretation of Eq. (9) and, perhaps, of the full nonlinear equation (5); these two questions are still open.

The Lax pair associated with Eq. (1) is given by the following 2×2 matrix equations[3,4]:

$$\psi^-(x,t,z) = G(x,t,z)\psi^+(x,t,z),$$
$$G(x,t,z) \equiv I + z\sigma_3 + U(x,t), \quad x\in\mathbb{R}, \tag{10}$$

$$\psi_t^{\pm}(x,t,z) = \frac{a_n}{2}\left(z^n\sigma_3 + \sum_{j=0}^{n-1} z^j (P^{\pm} V_{n,j}^{\pm})(x,t)\right) \times \psi^{\pm}(x,t,z). \tag{11}$$

In formula (10) I is the identity matrix, z plays the role of a spectral parameter, and $U(x,t)$ is a z-independent 2×2 matrix given in the form

$$U(x,t) = \sqrt{I + Q^2(x,t)} - I + Q(x,t), \tag{12}$$

where Q, introduced in (1), is the off-diagonal part of U. In what follows we take the $p(y)$, introduced in (1), to be $p(y) = a_n y^n$, $n\in\mathbb{N}$; this is for convenience only. We have

$$V_{n,j}^{\pm} \equiv 2L^{n-j-1}Q \mp (1/\sqrt{1+Q^2})[Q, L^{n-j-1}Q], \tag{13}$$

LQ is given by (2), and the $P^{\pm}$ are the usual projection operators

$$(P^{\pm}f)(x) \equiv \pm \frac{1}{2\pi i}\int_{-\infty}^{\infty} dy (y - (x \pm i0))^{-1} f(y). \tag{14}$$

Given a Holder matrix function $U(x)$, Eq. (10) defines a homogeneous matrix RH problem on the line $\mathrm{Im}\, x = 0$ of the complex x plane, and ψ^+ (ψ^-) is the boundary value of a function holomorphic in the upper (lower) half x plane. Equations (11) describe the corresponding time evolution of $\psi^{\pm}$.

Equation (12) implies that

$$\det[I + z\sigma_3 + U(x,t)] = 1 - z^2, \tag{15}$$

with the following important consequences[4]: (i) the matrix $I + z\sigma_3 + U(x,t)$ is invertible for every $x\in\mathbb{R}$ [this is a necessary condition for the solvability of (10)]; and (ii) the total index κ of the matrix RH problem (10) is zero, since

$$\kappa = (1/2\pi)[\arg(\det(I + z\sigma_3 + U(x,t)))]_{-\infty}^{\infty}, \tag{16}$$

where $[f(x)]_{-\infty}^{\infty} = f(\infty) - f(-\infty)$. Then an important theorem due to Gohberg and Krein[15] shows that "generically" the two partial indices κ_1,κ_2 ($\kappa = \kappa_1 + \kappa_2$) are both zero.

This fact guarantees the existence and uniqueness of the fundamental matrix solutions $\psi^{\pm}(x,z)$ of (10) (here and in the following we omit the time dependence when not needed), satisfying the following boundary conditions:

$$\psi^+(x,z) \underset{|x|\to\infty}{\to} I, \quad \psi^-(x,z) \underset{|x|\to\infty}{\to} I + z\sigma_3, \tag{17}$$

which are a consequence of the requirement that $Q(x)$ vanishes as $|x|\to\infty$ sufficiently rapidly. Equations (17) suggest the introduction of functions $\mu^{\pm}(x,z)$ defined by $\mu^+(x,z) \equiv \psi^+(x,z)$ and $\mu^-(x,z) \equiv \psi^-(x,z)(I + z\sigma_3)^{-1}$. They obviously satisfy the RH boundary value problem

$$\mu^+(x,z) - \mu^-(x,z) + z(\sigma_3\mu^+(x,z) - \mu^-(x,z)\sigma_3) + U(x)\mu^+(x,z) = 0, \tag{18a}$$

$$\mu^{\pm}(x,z) \underset{|x|\to\infty}{\to} I, \tag{18b}$$

and have the following interesting property:

$$\det \mu^{\pm}(x,z) = 1, \tag{19}$$

which is a direct consequence of Eq. (15) (see Proposition IV of the Appendix).

If (18) has a unique solution then it can be given in terms of the following matrix integral equations:

$$\mu^+(x,z) + (I + z\sigma_3)^{-1}P^+(U\mu^+)(x,z) = I, \tag{20a}$$

$$\mu^-(x,z) - P^-(U\mu^+)(x,z)(I + z\sigma_3)^{-1} = I. \tag{20b}$$

These equations can have homogeneous matrix solutions $\phi^{\pm(j)}(x)$, $j = 1,2,\ldots$ (corresponding to the eigenvalues $z = z_j$) that satisfy the equations

$$\phi^{+(j)}(x) + (I + z_j\sigma_3)^{-1}P^+(U\phi^{+(j)})(x) = 0, \tag{21a}$$

$$\phi^{-(j)}(x) - P^-(U\phi^{+(j)})(x)(I + z_j\sigma_3)^{-1} = 0, \tag{21b}$$

with the boundary conditions

$$\phi^{\pm(j)}(x) = 0(x^{-1}), \quad |x| \gg 1. \tag{22}$$

Because of (22), the eigenvalues z_j are bound states of (18a); they correspond to the nongeneric case in which the partial indices κ_1 and κ_2 are different from zero ($\kappa_1 = -\kappa_2 \neq 0$), when there does not exist a unique solution of (18).

For suitable potentials $Q(x)$ the structure of (20) and Fredholm theory[16] imply that the matrices $\mu^{\pm}(x,z)$ are holomorphic in the complex z plane except for possible poles that generically cluster at ± 1. A sketch of the proof is given in the Appendix (Proposition I and its consequences). The poles of $\mu^{\pm}(x,z)$ correspond to the homogeneous solutions of Eqs. (20). Here we assume that they are simple (with a genericity argument) and that the following representation formula holds:

$$\mu^{\pm}(x,z) = I + \sum_{j=-\infty}^{\infty} \frac{\phi^{\pm(j)}(x)}{z - z_j}, \quad z_j \underset{j\to+\infty}{\to} \pm 1, \quad j\in\mathbb{Z}. \tag{23}$$

The first consequence of Eqs. (21) is that the 2×2 matrices $\phi^{\pm(l)}(x)$ are singular; precisely

$$\begin{pmatrix} \phi_{12}^{\pm(l)}(x) \\ \phi_{22}^{\pm(l)}(x) \end{pmatrix} = \alpha_l^{\pm} \mathbf{\Pi}^{\pm(l)}(x), \quad l\in\mathbb{Z}, \tag{24}$$

where $\Pi^{\pm(l)}$ indicate the first column vectors of the matrices $\phi^{\pm(l)}$:

$$\Pi^{\pm(l)}(x)=\begin{pmatrix}\Pi_1^{\pm(l)}(x)\\ \Pi_2^{\pm(l)}(x)\end{pmatrix}\equiv\begin{pmatrix}\phi_{11}^{\pm(l)}(x)\\ \phi_{21}^{\pm(l)}(x)\end{pmatrix}, \tag{25}$$

and

$$\alpha_l^- = [(1+z_l)/(1-z_l)]\alpha_l^+. \tag{26}$$

Equations (21) also imply that

$$\lim_{|x|\to\infty} x\Pi_1^{+(l)}(x)=\lim_{|x|\to\infty} x\Pi_1^{-(l)}(x)\equiv c_l, \tag{27}$$

$$(1-z_l)\lim_{|x|\to\infty} x\Pi_2^{+(l)}(x)=(1+z_l)\lim_{|x|\to\infty} x\Pi_2^{-(l)}(x), \tag{28}$$

while Eqs. (19), (23), (24), (25), (26), and (27) imply that

$$\lim_{|x|\to\infty} x\Pi_2^{\pm(l)}(x)=-c_l/\alpha_l^{\pm}. \tag{29}$$

Finally, useful information about $U(x)$ is obtained by expanding Eq. (18a) for large z and by using Eq. (23);

$$U(x)=\sum_{j=-\infty}^{\infty}[\phi^{-(j)}(x)\sigma_3-\sigma_3\phi^{+(j)}(x)]. \tag{30}$$

Significantly it turns out that the vector solutions $\Pi^{\pm(l)}(x)$, $l\in\mathbb{Z}$, of Eqs. (21) satisfy the following infinite-dimensional algebraic system:

$$(x+\gamma_l^{\pm})\Pi^{\pm(l)}(x)=c_l\left[\begin{pmatrix}1\\ -1/\alpha_l^{\pm}\end{pmatrix}+\sum_{\substack{j=-\infty\\ j\neq l}}^{\infty}\frac{\alpha_l^{\pm}-\alpha_j^{\pm}}{\alpha_l^{\pm}(z_l-z_j)}\Pi^{\pm(j)}(x)\right], \quad l\in\mathbb{Z}. \tag{31}$$

The proof amounts to showing that the right- and left-hand sides of Eq. (31) satisfy the same integral equation. Here we give a sketch of the proof for Eq. $(31)^+$ [the analogous proof for Eq. $(31)^-$ is omitted].

Using Eq. $(21)^+$ and the asymptotic properties $(27)^+$ and $(29)^+$, one can show that $x\Pi^{+(l)}(x)$ satisfies the following nonhomogeneous vector equations:

$$x\Pi^{+(l)}(x)+(I+z_l\sigma_3)^{-1}P^+(U(x\Pi^{+(l)}))(x)=c_l\begin{pmatrix}1\\ -1/\alpha_l^+\end{pmatrix}, \quad l\in\mathbb{Z}. \tag{32}$$

On the other hand, manipulating Eq. $(21)^+$, one obtains

$$\nu^{+(l)}(x)+(I+z_l\sigma_3)^{-1}P^+(U\nu^{+(l)})(x)=c_l\begin{pmatrix}1\\ -1/\alpha_l^+\end{pmatrix}+\nu_0^{+(l)}(x), \quad l\in\mathbb{Z}, \tag{33}$$

where

$$\nu^{+(l)}(x)\equiv c_l\left[\begin{pmatrix}1\\ -1/\alpha_l^+\end{pmatrix}+\sum_{\substack{j=-\infty\\ j\neq l}}^{\infty}\frac{\alpha_l^+-\alpha_j^+}{\alpha_l^+(z_l-z_j)}\Pi^{+(j)}(x)\right], \tag{34a}$$

$$\nu_0^{+(l)}(x)\equiv c_l(I+z_l\sigma_3)^{-1}\left(\sigma_3\sum_{j=-\infty}^{\infty}\left(1-\frac{\alpha_j^+}{\alpha_l^+}\right)\times\Pi^{+(j)}(x)+P^+\left(U\begin{pmatrix}1\\ -1/\alpha_l^+\end{pmatrix}\right)(x)\right). \tag{34b}$$

Since the $(+)$ part of Eq. (30) implies that $\nu_0^{+(l)}(x)=\mathbf{0}$ the vector functions $x\Pi^{+(l)}(x)$ and $\nu^{+(l)}(x)$ satisfy the same nonhomogeneous integral equation (32), then their difference is proportional to the corresponding homogeneous solution $\Pi^{+(l)}(x)$; $\nu^{+(l)}(x)-x\Pi^{+(l)}(x)=\gamma_l^+\Pi^{+(l)}(x)$, which is Eq. $(31)^+$.

The constant $\gamma_l^{\pm}$ may be evaluated from Eqs. (31) for large x and making use of Eqs. (27),

$$\gamma_l^{\pm}=\sum_{\substack{j=-\infty\\ j\neq l}}^{\infty}\frac{\alpha_l^{\pm}-\alpha_j^{\pm}}{\alpha_l^{\pm}(z_l-z_j)}c_j+\lim_{|x|\to\infty}\left[x\left(1-\frac{x\Pi_1^{\pm(l)}(x)}{c_l}\right)\right], \quad l\in\mathbb{Z}. \tag{35}$$

Equations $(35)^{\pm}$, (27), (18a), and (30) finally imply that γ_l^+ and γ_l^- are connected through the following simple expression:

$$\gamma_l^+-\gamma_l^-=2c_l/(1-z_l^2), \quad l\in\mathbb{Z}. \tag{36}$$

Conversely, one could prove that if the $\Pi^{\pm(l)}(x)$ are solutions of the ∞-dimensional algebraic system (31), and (26) and (36) hold, then $\phi^{\pm(l)}(x)$ and $U(x)$, defined by Eqs. (24), (25), and (30), respectively, satisfy Eq. (18a) [with $\mu^{\pm}(x,z)$ replaced by $\phi^{\pm(l)}(x)$ when z is replaced by z_l]. Moreover the reconstructed potential $U(x)$ satisfies Eq. (15). A sketch of the proof is given in the Appendix (Propositions II–IV).

The direct problem is a linear mapping from the potential $Q(x)$ to the scattering data

$$S\equiv\{z_l,c_l,\alpha_l^+,\gamma_l^+; \quad l\in\mathbb{Z}\}. \tag{37}$$

More precisely, given $Q(x)$ [and then $U(x)$ through Eq. (12)], $\mu^{\pm}(x,z)$ and the bound states z_l are in principle given by solving Eqs. (20), and then the $\phi^{\pm(l)}(x)$ are obtained taking the limit $\phi^{\pm(l)}(x)=\lim_{z\to z_l}(z-z_l)\mu^{\pm}(x,z)$. All the scattering data S can be constructed through the following sequence of steps: c_l via Eq. (27); α_l^+ via $\alpha_l^+=-c_l/\lim_{|x|\to\infty} x\Pi_2^{+(l)}(x)$ [consequence of Eq. $(29^+]$; γ_l^+ via Eq. $(35)^+$ and the α^- and γ_l^-, which are not independent data, can be obtained using Eqs. (26) and (36), respectively.

The inverse scattering problem, a linear mapping from the scattering data S to the potential Q, if formally performed by solving the infinite-dimensional algebraic system (31) and then by making use of formulas (30), (24), and (25).

The time evolution of the scattering data is obtained by observing first that the solutions $\mu^{\pm}$ of (18) evolve in time according to the following equations:

$$\mu_l^{\pm}(x,t,z) = \frac{a_n}{2}\left(z^n\left[\sigma_3,\mu^{\pm}(x,t,z)\right] + \sum_{j=0}^{n-1} z^j (P \pm V_{n,j}^{\pm})(x,t)\right)\mu^{\pm}(x,t,z). \tag{38}$$

The data evolve as follows:

$$z_l(t) = z_l(0) = z_l, \quad c_l(t) = c_l(0) = c_l,$$
$$\alpha_l^{\pm}(t) = \alpha_l^{\pm}(0)e^{a_n z_l^n t}, \tag{39}$$
$$\gamma_l^{\pm}(t) = \gamma_l^{\pm}(0) - n a_n c_l z_l^{n-1} t.$$

In order to obtain the above equations for the time evolution of the scattering data we first substitute Eqs. (23) and (30) into (38). This yields $\dot{z}_l = 0$. The other relationships can be found by taking the time derivative of (31) using (38) at $z = z_l$, with $\mu^+(x,t,z_l)$ replaced by $\phi^{+(l)}$ when z is replaced by z_l and evaluating the results as $x \to \infty$.

Equations (39) complete the IST scheme

$$Q(x,0) \xrightarrow[\text{direct problem}]{} S(0) \xrightarrow[\text{evolution of scattering data}]{} S(t) \xrightarrow[\text{inverse problem}]{} Q(x,t), \tag{40}$$

which reduces the solution of the Cauchy problem for Eqs. (1) to a sequence of linear steps.

The pure soliton solutions associated with the class of Eqs. (1) correspond to a finite number N of poles in formula (23). If $N = 1$ ($c_j = 0$, $|j| \geqslant 1$), Eqs. (31), (30), and (39) yield the following one-soliton solution:

$$Q_{12}(x,t) = -\frac{2c_0\alpha_0^+(t)}{1-z_0}\left(x + \gamma_0^+(t) - \frac{c_0}{1+z_0}\right) \times [(x+\gamma_0^+(t))(x+\gamma_0^-(t))]^{-1}, \tag{41a}$$

$$Q_{21}(x,t) = -\frac{2c_0}{(1+z_0)\alpha_0^+(t)}\left(x + \gamma_0^+(t) - \frac{c_0}{1-z_0}\right) \times [(x+\gamma_0^+(t))(x+\gamma_0^-(t))]^{-1}. \tag{41b}$$

It should be mentioned that in the reduction case $Q_{12} = Q_{21}$ [which is allowed for all the evolution equations (1) corresponding to n odd] the singularities of $\mu^{\pm}(x,t,z)$ come in pairs $z_{-l} = -z_l$ and, correspondingly, $\alpha_l^+(0)\alpha_{-l}^+(0) = 1$, $c_{-l} = c_l$, and $\gamma_{-l}^{\pm} = \gamma_l^+$; if $l = 0$, $z_0 = 0$, and $\alpha_0^{+2}(0) = 1$.

III. LINEAR LIMIT

We conclude this paper by noting that the one-soliton solution (41) (when $n = 1$) provides, in an appropriate limit, a particular solution of the linearized sine–Hilbert equation. The choices $c_0 = i\epsilon$, $z_0 = 1 - \epsilon$, $\epsilon \ll 1$ imply that

$$Q_{12} \simeq -2ie^{a_1 t}(x + \gamma_0^-(0))^{-1}$$

and

$$Q_{21} \simeq -i\epsilon e^{-a_1 t}(x + \gamma_0^+(0))^{-1}$$

are particular solutions of the linearized version of (4). Then the linear combination

$$Q(x,t) = b^+ e^{-a_1 t}(x + \gamma_0^+(0))^{-1} + b^- e^{a_1 t}(x + \gamma_0^-(0))^{-1},$$

$$b^{\pm} \text{ arbitrary constants} \tag{42}$$

is the particular solution of Eq. (6) coming from (41); it coincides with (7) and (8) through the identifications $A^{\pm}(x) = b^{\pm}(x + \gamma_0^{\pm}(0))^{-1}$ and $c = ia_1$.

ACKNOWLEDGMENTS

We wish to thank D. Bar Yaacov for suggesting to us the content of Proposition I. We also thank A. Degasperis, H. Segur, S. Pierini, and M. D. Kruskal for useful discussions and comments on this work. We also thank the referee for useful suggestions.

This research was supported in part by the National Science Foundation under Grant No. MCS-8202117, the Office of Naval Research under Grant No. N00014-76-C09867, and the Air Force Office of Scientific Research under Grant No. 84-0005.

APPENDIX: RELEVANT PROPOSITIONS

In order to prove that the matrices $\mu^{\pm}(x,z)$ are holomorphic in the complex z plane $\mathbb{C}\backslash\{\pm 1\}$ except for possible poles that generically cluster at ± 1, we first convert the RH problem (10), (17) into the linear integral equation

$$\eta(x,z) + (\hat{K}\eta)(x,z) = h(x,z), \tag{A1}$$

where

$$\eta(x,z) \equiv G(x,z)(\psi^+(x,z) - I) = G(x,z)(\mu^+(x,z) - I), \tag{A2a}$$

$$(\hat{K}f)(x,z) \equiv \int K_1(x,y,z)G^{-1}(y,z)f(y)dy, \tag{A2b}$$

$$K_1(x,y,z) \equiv (2\pi i(y-x))^{-1}(G(y,z) - G(x,z)), \tag{A2c}$$

$$h(x,z) \equiv -(2\pi i)^{-1}\int_{-\infty}^{\infty}(y - (x+i0))^{-1}U(y)dy. \tag{A2d}$$

The procedure is standard.[11] We define

$$\tilde{\psi}^+(x,z) \equiv \psi^+(x,z) - I,$$
$$\tilde{\psi}^-(x,z) \equiv \psi^-(x,z) - (I + z\sigma_3); \tag{A3}$$

then

$$\tilde{\psi}^-(x,z) = G(x,z)\tilde{\psi}^+(x,z) + U(x), \tag{A4a}$$

$$\tilde{\psi}^{\pm}(x,z) \underset{|x|\to\infty}{\to} 0, \tag{A4b}$$

where $\tilde{\psi}^+(x,z)$ and $\tilde{\psi}^-(x,z)$ are $(+)$ and $(-)$ functions, respectively, then

$$\int_{-\infty}^{\infty}(y - (x \mp i0))^{-1}\tilde{\psi}^{\pm}(y,z)dy = \int_{-\infty}^{\infty}(y-x)^{-1}\tilde{\psi}^{\pm}(y,z)dy \mp \pi i\tilde{\psi}^{\pm}(x,z) = 0. \tag{A5}_{\pm}$$

Multiplying from the left Eq. $(A5)_+$ by $G(x,z)$ and subtracting it from Eq. $(A5)_-$ [in which $\tilde{\psi}^-$ is replaced by formula (A4a)], we finally obtain Eq. (A1).

Proposition I: The hypotheses on $U(x)$,

$$U(x), U_x(x) \in L^{\infty}, \tag{A6a}$$

$$\mathrm{tr}(\sigma_3 U(x)) = \mathrm{tr}\, U(x) + \det U(x) = 0, \tag{A6b}$$

$$U(x) = 0, \quad |x| > M, \quad M > 0, \tag{A6c}$$

imply that (a) the operator $\hat{K}_2$, defined by

$$(\hat{K}_2 f)(x,z) \equiv G^{-1}(x,z) f(x), \tag{A7}$$

is bounded in L^2; (b) the operator $\hat{K}_1$, defined by

$$(\hat{K}_1 f)(x,z) \equiv \int_{-\infty}^{\infty} K_1(x,y,z) f(y)\, dy, \tag{A8}$$

is compact in L^2. In the above z is fixed. Furthermore, it can be shown that (A6) are preserved under the time flow.

Remarks: Hypotheses (A6a) and (A6b) imply that there exist three positive constants b_1, b_2, b_3 such that

$$\sup |G^{-1}(x,z)| < b_1, \quad x \in \mathbb{R}, \tag{A9a}$$

$$|K_1(x,y,z)| < b_2, \quad (x,y) \in \mathbb{R}, \tag{A9b}$$

$$|G(y,z) - G(x,z)| < b_3, \quad (x,y) \in \mathbb{R}^2. \tag{A9c}$$

Finally, the compact support potential $U(x)$ in (A6c) yields a compact support kernel $K_1(x,y,z)$:

$$K_1(x,y) = 0, \quad (x,y) \in A_0 \equiv \{(x,y)/|x| > M, \quad |y| > M\}. \tag{A9d}$$

Proof of Proposition I: (a) is a direct consequence of (A9a):

$$\|\hat{K}_2 f\|_2 = \left(\int_{\mathbb{R}} dx |G^{-1}(x) f(x)|^2 \right)^{1/2}$$

$$\leqslant b_1 \left(\int_{\mathbb{R}} dx |f(x)|^2 \right)^{1/2}$$

$$= b_1 \|f\|_2, \quad f \in L^2.$$

(b) follows from the fact that $K_1(x,y) \in L^2(\mathbb{R}^2)$. Indeed (A9d) implies that

$$\int\int_{\mathbb{R}^2} |K_1(x,y)|^2\, dx\, dy = \int\int_{A} |K_1(x,y)|^2\, dx\, dy, \tag{A10}$$

where $A = \cup_{j=1}^{3} A_j$, and $A_2 \equiv \{(x,y)/|x| > M+1, |y| < M\}$,

$$A_3 \equiv \{(x,y)/|x| < M, \ |y| > M+1\},$$

$$A_1 \equiv \mathbb{R}^2 \setminus A_3 \cup A_0.$$

Then (A9) implies $\int\int_{A_1} |K_1|^2\, dx\, dy < b_2^2 (M+1)^2 < \infty$, and (A9c) implies

$$\int\int_{A_3} |K_1|^2\, dx\, dy < \frac{M b_3^2}{2\pi^2} \left(\int_{M+1}^{\infty} |y - M|^{-2}\, dy + \int_{-\infty}^{-(M+1)} |y + M|^{-2}\, dy \right) < \infty.$$

Since an analogous formula holds for the region A_3, then $\int\int_{\mathbb{R}^2} |K_1|^2\, dx\, dy < \infty$.

Consequences: (i) The compactness of $\hat{K}_1$ and the boundedness of $\hat{K}_2$ in L^2 imply that $\hat{K} = \hat{K}_1 \hat{K}_2$ is compact in L^2 (Ref. 16); (ii) the compactness of the operator $\hat{K}$ and the particular z dependence of its kernel

$$K(x,y,z) = K_1(x,y,z) G^{-1}(y,z)$$

$$= (1 - z^2)^{-1} [\beta_1(x,y) + z\beta_2(x,y)],$$

$$\beta_1(x,y) \equiv (2\pi i (y-x))^{-1} (U(x) + \sigma_3 U(y) \sigma_3 + U(x) \sigma_3 U(y) \sigma_3),$$

$$\beta_2(x,y) \equiv (2\pi i (y-x))^{-1} (U(y) - U(x)) \sigma_3,$$

imply that the solution $\eta(x,z)$ of Eq. (A1) is holomorphic in the complex z plane $\mathbb{C} \setminus \{\pm 1\}$, except for possible poles clustering at $z = \pm 1$, which are the singularities of the kernel in the complex z plane.[14]

Proposition II: If the $\mathbf{\Pi}^{\pm(l)}(x)$ are solutions of Eqs. (31) within conditions (26) and (36), then the $\mathbf{\Pi}^{\pm(l)}(x)$ also satisfy the equations

$$\mathbf{E}^{(l)} \equiv (I + z_l \sigma_3 + U) \mathbf{\Pi}^{+(l)} - (I + z_l \sigma_3) \times \begin{pmatrix} 1 & 0 \\ 0 & (1+z_l)/(1-z_l) \end{pmatrix} \mathbf{\Pi}^{-(l)} = 0, \tag{A11}$$

with $U(x)$ given by

$$U(x) = \sum_j \begin{pmatrix} -\pi_1^{+(j)} + \pi_1^{-(j)} & -\alpha_j^{+} \pi_1^{+(j)} - \alpha_j^{-} \pi_1^{-(j)} \\ \pi_2^{+(j)} + \pi_2^{-(j)} & \alpha_j^{+} \pi_2^{+(j)} - \alpha_j^{-} \pi_2^{-(j)} \end{pmatrix}. \tag{A12}$$

It should be observed that since

$$\phi^{\pm(l)} = \begin{pmatrix} \pi_1^{\pm(l)} & \alpha_l^{\pm} \pi_1^{\pm(l)} \\ \pi_2^{\pm(l)} & \alpha_l^{\pm} \pi_2^{\pm(l)} \end{pmatrix}, \tag{A13}$$

(33) is equivalent to

$$\phi^{+(l)} - \phi^{-(l)} + z_l (\sigma_3 \phi^{+(l)} - \phi^{-(l)} \sigma_3) + U \phi^{+(l)} = 0, \quad l \in \mathbb{Z}. \tag{A14}$$

The above assertion is demonstrated as follows. From (31) we have

$$(x + \gamma_l^{-})(I + z_l \sigma_3) \begin{pmatrix} 1 & 0 \\ 0 & (1+z_l)/(1-z_l) \end{pmatrix} \mathbf{\Pi}^{-(l)}$$

$$= (I + z_l \sigma_3) \begin{pmatrix} 1 & 0 \\ 0 & (1+z_l)/(1-z_l) \end{pmatrix} \begin{pmatrix} c_l \\ -c_l/\alpha_l^{-} \end{pmatrix} + \sum_{j \neq l} \frac{c_l (\alpha_l^{-} - \alpha_j^{-})}{\alpha_l^{-} (z_l - z_j)} (I + z_l \sigma_3) \begin{pmatrix} 1 & 0 \\ 0 & (1+z_l)/(1-z_l) \end{pmatrix} \mathbf{\Pi}^{-(j)}, \tag{A15}$$

$$(x+\gamma_l^+)(I+z_l\sigma_3)\mathbf{\Pi}^{+(l)} = (I+z_l\sigma_3)\begin{pmatrix} c_l \\ -c_l/\alpha_l^+ \end{pmatrix} + \sum_{j\neq l}\frac{c_l(\alpha_l^+ - \alpha_j^+)}{\alpha_l^+(z_l-z_j)}(I+z_l\sigma_3)\mathbf{\Pi}^{+(j)}, \tag{A16}$$

$$(x+\gamma_l^+)U\mathbf{\Pi}^{+(l)} = U\begin{pmatrix} c_l \\ -c_l/\alpha_l^+ \end{pmatrix} + \sum_{j\neq l}\frac{c_l(\alpha_l^+ - \alpha_j^-)}{\alpha_l^+(z_l-z_j)}U\mathbf{\Pi}^{+(j)}. \tag{A17}$$

Adding (37a)–(37c) yields

$$(x+\gamma_l^+)\mathbf{E}^{(l)} - \sum_{j\neq l}\frac{c_l(\alpha_l^+ - \alpha_j^+)}{\alpha_l^+(z_l-z_j)}\mathbf{E}^{(j)} = \begin{pmatrix} d_1^{(l)} \\ d_2^{(l)} \end{pmatrix}, \tag{A18}$$

where

$$\begin{aligned}\mathbf{d}^{(l)} = \begin{pmatrix} d_1^{(l)} \\ d_2^{(l)} \end{pmatrix} &= -(\gamma_l^+ - \gamma_l^-)(I+z_l\sigma_3)\begin{pmatrix} 1 & 0 \\ 0 & (1+z_l)/(1-z_l) \end{pmatrix}\mathbf{\Pi}^{-(l)} + c_l U\begin{pmatrix} 1 \\ -1/\alpha_l^+ \end{pmatrix} \\ &+ \sum_{j\neq l}\frac{c_l(\alpha_l^+ - \alpha_j^+)}{\alpha_l^+}\sigma_3\mathbf{\Pi}^{+(j)} + \sum_{j\neq l}\left[\frac{c_l(\alpha_l^+ - \alpha_j^+)}{\alpha_l^+(z_l-z_j)}(I+z_j\sigma_3)\begin{pmatrix} 1 & 0 \\ 0 & (1+z_j)/(1-z_j) \end{pmatrix}\right. \\ &\left. - \frac{c_l(\alpha_l^- - \alpha_j^-)}{\alpha_l^+(z_l-z_j)}(I+z_l\sigma_3)\begin{pmatrix} 1 & 0 \\ 0 & (1+z_l)/(1-z_l) \end{pmatrix}\right]\mathbf{\Pi}^{-(j)}.\end{aligned} \tag{A19}$$

Thus if $\alpha^{\pm(l)}$ and $\gamma^{\pm(l)}$ satisfy (26) and (36) then $\mathbf{d}^{(l)} = 0$ and $\mathbf{E}^{(l)} = 0$ for $l\in\mathbb{Z}$ since it satisfies

$$(x+\gamma_l^+)\mathbf{E}^{(l)} - \sum_{j\neq l}\frac{c_l(\alpha_l^+ - \alpha_j^+)}{\alpha_l^+(z_l-z_j)}\mathbf{E}^{(j)} = 0, \tag{A20}$$

i.e., the homogeneous version of (31).Then (A11) follows from the assumption that the infinite-dimensional system (31) has a unique solution. Proposition II immediately implies that $\mu^{\pm}(x,z)$, defined by Eqs. (23), satisfy the RH boundary value problem (18).

Proposition III: If the $\mathbf{\Pi}^{\pm(l)}(x)$ are solutions of Eqs. (31), then they satisfy the following equations:

$$\begin{aligned}&\Pi_1^{\pm(l)}(x) + \alpha_l^{\pm}\Pi_2^{\pm(l)}(x) \\ &+ \sum_{\substack{j=-\infty \\ j\neq l}}^{\infty}\frac{\alpha_l^{\pm} - \alpha_j^{\pm}}{z_l - z_j}(\Pi_2^{\pm(l)}(x)\Pi_1^{\pm(j)}(x) \\ &- \Pi_2^{\pm(j)}(x)\Pi_1^{\pm(l)}(x)) = 0, \quad l\in\mathbb{Z}.\end{aligned} \tag{A21}$$

It should be observed that Eqs. (A21) are equivalent to the equation $\det\mu^{\pm}(x,z) = 1$.

In order to demonstrate Proposition III we manipulate Eqs. (31):

$$\begin{aligned}&(x+\gamma_l^{\pm})(\Pi_1^{\pm(l)}(x) + \alpha_l^{\pm}\Pi_2^{\pm(l)}(x)) \\ &= \sum_{j\neq l}\frac{c_l(\alpha_l^{\pm} - \alpha_j^{\pm})}{\alpha_l^{\pm}(z_l-z_j)} \\ &\times(\Pi_1^{\pm(j)}(x) + \alpha_l^{\pm}\Pi_2^{\pm(j)}(x)), \quad l\in\mathbb{Z},\end{aligned} \tag{A22}$$

$$\begin{aligned}&(x+\gamma_l^{\pm})\sum_{j\neq l}\frac{\alpha_l^{\pm} - \alpha_j^{\pm}}{z_l - z_j}(\Pi_2^{\pm(j)}(x)\Pi_1^{\pm(l)}(x) \\ &- \Pi_1^{\pm(j)}(x)\Pi_2^{\pm(l)}(x)) \\ &= \sum_{j\neq l}\frac{c_l(\alpha_l^{\pm} - \alpha_j^{\pm})}{\alpha_l^{\pm}(z_l-z_j)}(\Pi_1^{\pm(j)}(x) + \alpha_l^{\pm}\Pi_2^{\pm(j)}(x)) \\ &+ \sum_{\substack{j\neq l \\ \nu\neq l}}A_{j\nu}^{\pm(l)}(x), \quad l\in\mathbb{Z},\end{aligned} \tag{A23a}$$

where

$$\begin{aligned}A_{j\nu}^{\pm(l)}(x) \equiv &\frac{c_l}{\alpha_l^{\pm}}\frac{(\alpha_l^{\pm} - \alpha_j^{\pm})(\alpha_l^{\pm} - \alpha_\nu^{\pm})}{(z_l - z_j)(z_l - z_\nu)} \\ &\times(\Pi_1^{\pm(\nu)}(x)\Pi_2^{\pm(j)}(x) \\ &- \Pi_2^{\pm(\nu)}(x)\Pi_1^{\pm(j)}(x)), \quad l\in\mathbb{Z}.\end{aligned} \tag{A23b}$$

Since $A_{j\nu}^{\pm(l)}(x) = -A_{\nu j}^{\pm(l)}(x)$, then

$$\sum_{\substack{j\neq l \\ \nu\neq l}}A_{j\nu}^{\pm(l)}(x) = 0,$$

and the difference between Eqs. (A22) and (A23a) gives just Eq. (A21).

Proposition IV: The soutions $\mu^{\pm}(x,z)$ of Eq. (18) satisfy the equation $\det\mu^{\pm}(x,z) = 1$ if and only if the potential matrix $U(x)$ satisfies the following basic constraints:

$$\operatorname{tr}(\sigma_3 U(x)) = 0, \tag{A24a}$$

$$\operatorname{tr}(U(x)) + \det(U(x)) = 0. \tag{A24b}$$

The above equations are equivalent to Eq. (15), or to Eq. (12).

In order to prove Proposition IV it is convenient to introduce the functions $F^+(x,z) \equiv \mu^+(x,z)$ and $F^-(x,z) \equiv (I + z\sigma_3)^{-1}\mu^-(x,z)(I + z\sigma_3)$; then the $F^\pm(x,z)$ satisfy the RH boundary value problem

$$F^-(x,z) = [I + (I + z\sigma_3)^{-1}U(x)]F^+(x,z), \qquad \text{(A25a)}$$

$$F^\pm(x,z) \underset{|x|\to\infty}{\to} I, \qquad \text{(A25b)}$$

and, obviously, $\det F^\pm(x,z) = \det \mu^\pm(x,z)$. If Eqs. (A24) hold, then Eq. (A25a) implies $\det F^-(x,z) = \det F^+(x,z)$ which, together with Eq. (A25b) implies $\det F^\pm(x,z) = 1$. Conversely, if $\det F^\pm(x,z) = 1$, then $\det[I + (I + z\sigma_3)^{-1}U(x)] = 1$, which is equivalent to Eq. (15), or Eq. (12), or Eqs. (A24).

Recapitulating, we have shown that if the $\mathbf{\Pi}^{\pm(l)}(x)$ are solutions of Eqs. (31) and the matrix functions $\mu^\pm(x,z)$ and $U(x)$ are defined in terms of $\mathbf{\Pi}^{\pm(l)}(x)$ through Eqs. (23), (A13), and (30), respectively, then (i) $\mu^\pm(x,z)$ and $U(x)$ satisfy the RH boundary value problem (18) (Proposition II), (ii) $\det \mu^\pm(x,z) = 1$ (Proposition III), and then (iii) $\mathrm{tr}(\sigma_3 U(x)) = \mathrm{tr}(U(x)) + \det(U(x)) = 0$ (Proposition IV.)

[1] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[2] F. Calogero and A. Degasperis, *Spectral Transform and Solitons I* (North-Holland, Amsterdam, 1982).
[3] A. Degasperis and P. M. Santini, Phys. Lett. A **98**, 240 (1983).
[4] A. Degasperis, P. M. Santini, and M. J. Ablowitz, J. Math. Phys. **26**, 10 (1986).
[5] R. I. Joseph, Phys. A **10**, L225 (1977).
[6] R. I. Joseph and R. Egri, J. Phys. A **11**, L97 (1978).
[7] T. Kubota, D. R. S. Ko, and D. Dodds, J. Hydronaut. **12**, 157 (1978).
[8] T. B. Benjamin, J. Fluid Mech. **29**, 559 (1967).
[9] H. Ono, J. Phys. Soc. Jpn. **39**, 1082 (1975).
[10] Y. Kodama, M. J. Ablowitz, and J. Satsuma, J. Math. Phys. **23**, 564 (1982).
[11] A. S. Fokas and M. J. Ablowitz, Stud. Appl. Math. **68**, 1 (1983).
[12] P. M. Santini, M. J. Ablowitz, and A. S. Fokas, J. Math. Phys. **25**, 892 (1984).
[13] N. I. Muskhelishvili, *Singular Integral Equations* (Gordon and Breach, New York, 1967).
[14] H. Segur and S. Pierini (private communication); P. H. Leblond and L. A. Mysak, *Waves in the Ocean* (Elsevier, Amsterdam, 1978).
[15] I. Gohberg and M. G. Krein, Usp. Mat. Nauk. **13** (2), 2 (1958).
[16] M. Reed and B. Simon, *Functional Analysis I* (Academic, New York, 1975).

Exact solutions for the nonlinear Klein–Gordon and Liouville equations in four-dimensional Euclidean space

Y. Matsuno
Department of Physics, Faculty of Liberal Arts, Yamaguchi University, Yamaguchi, 753, Japan

(Received 24 February 1987; accepted for publication 10 June 1987)

A systematic method for constructing particular solutions of the nonlinear Klein–Gordon and Liouville equations in four-spatial dimensions is developed. The method of solution presented here first consists of reducing nonlinear partial differential equations to ordinary differential equations (ODE's) by introducing symmetry variables and then seeking exact solutions for more tractable ODE's. Various exact solutions are presented, in which new solutions with nonspherical symmetries are included. Furthermore, the exact method is applied to the above equations in general n-spatial dimensions. Among them, a conformally invariant nonlinear Klein–Gordon equation is particularly interesting from the viewpoint of field theories. The exact solutions for these equations are generalizations of those for the corresponding equations in four-spatial dimensions.

I. INTRODUCTION

In this paper, we shall construct exact solutions for the following nonlinear Klein–Gordon equation:

$$\Box_4\phi + \lambda\phi^p = 0, \quad p \neq 0,1, \tag{1.1}$$

and the Liouville equation

$$\Box_4\phi + e^{\lambda\phi} = 0, \tag{1.2}$$

in four-dimensional Euclidean space. Here, $\phi = \phi(x_\mu)$ is a scalar function of the Euclidean coordinates x_μ ($\mu = 1 \sim 4$), the symbol $\Box_4$ denotes the four-dimensional Laplace operator defined by

$$\Box_4 = \sum_{\mu=1}^{4} \frac{\partial^2}{\partial x_\mu^2}, \tag{1.3}$$

and λ and p are real parameters.

These equations play an important role in various fields in physics.[1–4] In particular, Eq. (1.1) with $p = 3$ is closely related to scalar ϕ^4 theory[2] as well as the Euclidean Yang–Mills equation with the t'Hooft ansatz[3] while Eq. (1.2) represents a four-dimensional version of the Poisson equation. Therefore the investigation of exact solutions for these important equations may lead us to a deeper understanding of underlying physical phenomena.

Although various exact methods such as the inverse scattering method,[5–7] Bäcklund transformation,[8,9] and bilinear transformation method[10,11] etc. have been developed for analyzing nonlinear wave equations, the applicabilities of exact methods available nowadays are mainly restricted to lower-dimensional equations, namely those with one- or two-spatial variables in addition to one time variable. As for the higher-dimensional cases, however, there exists an exact method called "symmetry reduction."[12–15] This method consists of reducing partial differential equations (PDE's) to ordinary differential equations (ODE's) by considering symmetry groups acting on the space of independent and dependent variables. The solutions constructed by this method are a generalization of so-called similarity solutions.

Recently, the method of symmetry reduction has been applied to Eq. (1.1) with $p = 5$ and various special solutions have been presented.[16] On the other hand, to our knowledge, exact solutions for Eq. (1.2) have not been found in the literature except for a few works[17] while the Liouville equation in three-spatial dimensions has received much attention in connection with the soliton theory.[18,19]

The purpose of the present paper is to construct exact solutions of Eqs. (1.1) and (1.2) by reducing them to nonlinear ODE's. The basic idea is to introduce the following elementary symmetric functions of four Euclidean coordinates x_μ as symmetry variables:

$$s_1 = x_1 + x_2 + x_3 + x_4, \tag{1.4a}$$

$$s_2 = x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4, \tag{1.4b}$$

$$s_3 = x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4, \tag{1.4c}$$

$$s_4 = x_1x_2x_3x_4, \tag{1.4d}$$

and to rewrite Eqs. (1.1) and (1.2) in terms of these independent variables. It should be noted that the symmetry variables (1.4) are different from those due to Grundland *et al.*[16] The equations thus obtained are then reduced to nonlinear ODE's by introducing a new independent variable $y = s_2/s_1^2$. By solving these ODE's, various exact solutions are constructed for Eqs. (1.1) and (1.2).

In Sec. II, we present exact solutions of the nonlinear Klein–Gordon equation in four-spatial dimensions by reducing it to a second-order nonlinear ODE and study the properties of solutions. In Sec. III, the same procedure is applied to the Liouville equation in four-spatial dimensions to construct exact solutions. In Sec. IV, we discuss the nonlinear Klein–Gordon and Liouville equations in general n-spatial dimensions and present some exact solutions for these equations. In particular, a conformally invariant nonlinear Klein–Gordon equation considered here will be quite interesting from the viewpoint of field theories. Section V is devoted to concluding remarks.

II. NONLINEAR KLEIN–GORDON EQUATION

A. Reduction to ODE

In this section, we shall consider the nonlinear Klein–Gordon equation (1.1). Since the operator $\Box_4$ is invariant

 0022-2488/87/102317-06$02.50 © 1987 American Institute of Physics

under all possible permutations of the four coordinates x_μ, the solutions of Eq. (1.1) may be expressed as functions of the elementary symmetric functions defined by relations (1.4). In terms of s_μ, the operator $\Box_4$ is written in the form

$$\Box_4 = 4\frac{\partial^2}{\partial s_1^2} + (3s_1^2 - 2s_2)\frac{\partial^2}{\partial s_2^2} + (2s_2^2 - 2s_1s_3 - 4s_4)\frac{\partial^2}{\partial s_3^2} + (s_3^2 - 2s_2s_4)\frac{\partial^2}{\partial s_4^2} + 6s_1\frac{\partial^2}{\partial s_1\partial s_2} + 4s_2\frac{\partial^2}{\partial s_1\partial s_3} + 2s_3\frac{\partial^2}{\partial s_1\partial s_4} + 2(2s_1s_2 - 3s_3)\frac{\partial^2}{\partial s_2\partial s_3} + 2(s_1s_3 - 4s_4)\frac{\partial^2}{\partial s_2\partial s_4} + 2(s_2s_3 - 3s_1s_4)\frac{\partial^2}{\partial s_3\partial s_4}. \tag{2.1}$$

Introducing (2.1) into Eq. (1.1), however, yields a quite complicated nonlinear PDE and a reduction to an ODE seems to be very difficult although the possibility cannot be denied. Therefore, in this paper, we shall confine ourselves to seeking solutions that are functions of only two independent variables s_1 and s_2 and the more general cases will be treated elsewhere. Then, Eq. (1.1) takes the following form:

$$4\frac{\partial^2\phi}{\partial s_1^2} + 6s_1\frac{\partial^2\phi}{\partial s_1\partial s_2} + (3s_1^2 - 2s_2)\frac{\partial^2\phi}{\partial s_2^2} + \lambda\phi^p = 0, \quad p\neq 0,1. \tag{2.2}$$

At this point, one should remark that Eq. (2.2) is invariant under the scale transformations, $s_1\to\gamma s_1$, $s_2\to\gamma^2 s_2$, and $\phi\to\gamma^{-2/(p-1)}\phi$ (γ: constant). Keeping this property of Eq. (2.2) in mind, we introduce the *ansatz* function f of one variable y as follows:

$$\phi = [\sqrt{\lambda}\, s_1 f(y)]^{-2/(p-1)}, \quad y = s_2/s_1^2. \tag{2.3}$$

It then turns out that Eq. (2.2) is transformed into the following nonlinear ODE for f:

$$(2y-1)(8y-3)\left(ff'' - \frac{p+1}{p-1}f'^2\right) + \left[\frac{8(3p+1)}{p-1}y - \frac{12p}{p-1}\right]ff' - \frac{4(p+1)}{p-1}f^2 - \frac{p-1}{2} = 0, \quad p\neq 0,1, \tag{2.4}$$

where the prime appended to f means the differentiation with respect to the independent variable y. This abbreviation will be used throughout the paper. Equation (2.4) is a basic equation that we consider in Sec. II B hereafter.

B. Construction of exact solutions

Now, let us seek exact solutions of Eq. (2.4). First of all, one readily notices that Eq. (2.4) possesses a constant solution, namely,

$$f = \pm i(p-1)/\sqrt{8(p+1)}, \tag{2.5}$$

which, when substituted in (2.3), yields an exact solution of Eq. (1.1) in the form

$$\phi = [\pm i(p-1)[\lambda/8(p+1)]^{1/2}s_1]^{-2/(p-1)}. \tag{2.6}$$

We shall discuss the property of solution (2.6) in a special case of $p = 3$ later.

Next, we shall seek solutions of Eq. (2.4) that behave like $f\sim y^\alpha$ when $y\to-\infty$. One finds that α is determined by the following algebraic equation:

$$(\alpha - \tfrac{1}{2})(\alpha - (p+1)/4) = 0. \tag{2.7}$$

Various possibilities arise according to the values of α and p. We first investigate the case of $\alpha = \frac{1}{2}$. For this case, we assume the solution in the form

$$f = \pm(ay+b)^{1/2}, \tag{2.8}$$

and substitute (2.8) into (2.4), where a and b are unknown constants. Then, it readily follows that a and b are determined by the following system of algebraic equations:

$$a = [(p-1)/p](-4b + (p-1)/2), \tag{2.9a}$$

$$(p-2)(p-3)b^2 - \tfrac{1}{8}(p-1)^2(5p-12)b + \tfrac{3}{32}(p-1)^4 = 0. \tag{2.9b}$$

There exist three possible solutions of Eq. (2.9) according to the values of p, namely $p\neq 2,3$, $p = 3$, and $p = 2$, respectively.

1. $\alpha = \frac{1}{2}$, $p\neq 2,3$

In the case of $p\neq 2,3$, we have two pairs of solutions for Eq. (2.9). They are given by

$$a = -(p-1)^2/(p-3), \tag{2.10a}$$

$$b = \tfrac{3}{8}[(p-1)^2/(p-3)], \tag{2.10b}$$

and

$$a = -(p-1)^2/2(p-2), \tag{2.11a}$$

$$b = \tfrac{1}{4}[(p-1)^2/(p-2)], \tag{2.11b}$$

respectively. Substitution of (2.8) with (2.10) and (2.11) into (2.3) yields exact solutions of Eq. (1.1) in the forms

$$\phi = [\pm[\sqrt{\lambda}(p-1)/\sqrt{8(p-3)}] \times (3s_1^2 - 8s_2)^{1/2}]^{-2/(p-1)}, \tag{2.12}$$

$$\phi = [\pm[\sqrt{\lambda}(p-1)/\sqrt{4(p-2)}] \times (s_1^2 - 2s_2)^{1/2}]^{-2/(p-1)}. \tag{2.13}$$

It follows by the definition of s_1 and s_2 that

$$s_1^2 - 2s_2 = \sum_{\mu=1}^{4} x_\mu^2 \equiv x^2, \tag{2.14}$$

and hence, expression (2.13) is the spherically symmetric solution of Eq. (1.1). This solution has already been obtained by Burt.[4] On the other hand, we obtain

$$3s_1^2 - 8s_2 = \sum_{\substack{\mu,\nu=1\\(\mu<\nu)}}^{4} (x_\mu - x_\nu)^2, \tag{2.15}$$

and accordingly, expression (2.15) cannot be reduced to the form (2.14) by means of any transformation that leaves Eq. (1.1) invariant under four-dimensional Euclidean groups. This fact implies that (2.12) represents a new exact solution. For $p>1$, solution (2.12) is singular at $x_\mu = x_\nu$ ($\mu,\nu = 1\sim 4$) while for $p<1$, it is regular for all values of x_μ. As already mentioned in the Introduction, Eq. (1.1) with

$p = 5$ has been discussed by Grundland *et al.*[16] However, it should be remarked here that the solution (2.12) with $p = 5$ does not belong to the category of solutions that they have obtained.

2. $\alpha = \frac{1}{2}$, $p = 3$

This special case is of considerable physical interest, since Eq. (1.1) with $p = 3$ becomes the field equation of the familiar scalar ϕ^4 theory.[2] It is a well-known fact that Eq. (1.1) with $p = 3$ is invariant under the following specific conformal transformation[2]:

$$x_\mu \to \tilde{x}_\mu = (x_\mu + c_\mu x^2)/\sigma(x), \tag{2.16a}$$

$$\phi(x) \to \tilde{\phi}(x) = \phi(\tilde{x})/\sigma(x), \tag{2.16b}$$

with

$$\sigma(x) = 1 + 2\sum_{\mu=1}^{4} c_\mu x_\mu + c^2 x^2 \quad \left(c^2 = \sum_{\mu=1}^{4} c_\mu^2\right), \tag{2.16c}$$

where c_μ is a constant four-vector.

Now, we shall begin to discuss the solutions. As already shown in (2.6), the equation exhibits a solution of the form

$$\phi = \pm i\sqrt{\frac{8}{\lambda}}\left(\sum_{\mu=1}^{4} x_\mu\right)^{-1}. \tag{2.17}$$

At first sight, this solution seems to be a trivial plane-wave type one. However, if we apply both the conformal transformation (2.16) and the translation of the coordinates $x_\mu \to x_\mu + i\beta/2$ ($\mu = 1 \sim 4$) to (2.17), the solution can be transformed into the form

$$\phi = \pm (8\beta^2/\lambda)^{1/2}[1/(x^2 + \beta^2)], \tag{2.18a}$$

where β is a real constant related to a four-vector c_μ by the relation

$$\beta = i\left(\sum_{\mu=1}^{4} c_\mu\right)^{-1}. \tag{2.18b}$$

The solution (2.18) is nothing but the well-known instanton solution first derived by Belavin *et al.*[20]

Another solution of Eq. (1.1) with $p = 3$ is obtained from (2.8) and (2.9). The result is expressed in the form

$$\phi = \pm [\lambda(s_1^2 - 2s_2)]^{-1/2} = \pm (\lambda x^2)^{-1/2}. \tag{2.19}$$

This solution is spherically symmetric and is called the meron solution in gauge theory.[3]

3. $\alpha = \frac{1}{2}$, $p = 2$

In this case, an exact solution is given by (2.6) with $p = 2$, namely,

$$\phi = -\frac{24}{\lambda}\left(\sum_{\mu=1}^{4} x_\mu\right)^{-2}. \tag{2.20}$$

The solution is real, but singular on a four-dimensional plane, $\Sigma_{\mu=1}^{4} x_\mu = 0$. On the other hand, another solution follows from (2.3), (2.8), and (2.9) with $p = 2$ in the form

$$\phi = -\frac{8}{\lambda}\left[\sum_{\substack{\mu,\nu=1 \\ (\mu<\nu)}}^{4} (x_\mu - x_\nu)^2\right]^{-1}, \tag{2.21}$$

which seems to be a new exact solution. It is interesting to observe that spherically symmetric solutions of the form (2.13) do not exist for $p = 2$ in a striking contrast to the other cases.

4. $\alpha = (p+1)/4$, α: integer

In the preceding subsections, we have investigated the solutions for the case of $\alpha = \frac{1}{2}$. Here we shall discuss another possibility, namely, $\alpha = (p + 1)/4$ [see Eq. (2.7)]. First, consider the case where α is an integer. In this situation, the value of the parameter p can be taken to be $p = 4m + 3$ (m: integer) without loss of generality, so that $\alpha = m + 1$. Now, we assume solutions of Eq. (2.4) with $p = 4m + 3$ in the polynomial form

$$f = \sum_{j=0}^{m+1} a_j y^{m+1-j}, \quad a_0 \neq 0, \tag{2.22}$$

which is consistent with the asymptotic behavior of solutions $f \sim y^{m+1}$, and substitute (2.22) into Eq. (2.4). The resulting equation is the algebraic equation in y of order $2m + 2$. Equating the coefficients of y^j ($j = 0 \sim 2m + 2$) to zero, respectively, results in $2m + 3$ algebraic equations for unknowns a_j ($j = 0 \sim m + 1$). However, the equation that stems from the coefficient of y^{2m+2} is satisfied identically because of Eq. (2.7). Therefore the number of independent equations for a_j is at most $2m + 2$ while the number of unknowns is obviously equal to $m + 2$. Hence, if the inequality $m + 2 \geqslant 2m + 2$ holds, solutions of the form (2.22) would exist. The only possible value of m is $m = 0$, namely, $p = 3$, or equivalently $\alpha = 1$. Indeed, we have found the following exact solution of Eq. (2.4) with $p = 3$:

$$f = \pm(-3y + \tfrac{5}{4}). \tag{2.23}$$

Substituting (2.23) into (2.3) with $p = 3$, we obtain a new exact solution of Eq. (1.1) with $p = 3$ in the form[21]

$$\phi = \pm (1/\sqrt{\lambda})[4s_1/(5s_1^2 - 12s_2)] = \pm \frac{2}{\sqrt{\lambda}} \frac{\Sigma_{\mu=1}^{4} x_\mu}{x^2 + \frac{1}{4}\Sigma_{\mu,\nu=1}^{4} (x_\mu - x_\nu)^2}. \tag{2.24}$$

This solution is regular except for the origin $x_\mu = 0$ ($\mu = 1 \sim 4$) and decays asymptotically like $\phi \sim 4/(5\sqrt{\lambda}\, x_\mu)$ when $x_\mu \to \pm\infty$.

Furthermore, if we notice the invariance of Eq. (1.1) with $p = 3$ under the conformal transformation (2.16), another new exact solution can be generated starting from (2.24). The explicit form of the solution is written as follows:

$$\phi = \pm \frac{4}{\sqrt{\lambda}} \frac{\tilde{c}_1 x^2 + \Sigma_{\mu=1}^{4} x_\mu}{(5\tilde{c}_1^2 - 12\tilde{c}_2)x^4 + 2(6\Sigma_{\mu=1}^{4} c_\mu x_\mu - \tilde{c}_1 \Sigma_{\mu=1}^{4} x_\mu)x^2 + 6x^2 - \Sigma_{\mu,\nu=1}^{4} x_\mu x_\nu}, \tag{2.25a}$$

where

$$\tilde{c}_1 = \sum_{\mu=1}^{4} c_\mu, \tag{2.25b}$$

$$\tilde{c}_2 = \sum_{\substack{\mu,\nu=1 \\ (\mu<\nu)}}^{4} c_\mu c_\nu. \tag{2.25c}$$

At the present time, solutions of the form (2.22) have not been found except for $m = 0$ although we generally cannot deny the possibility of the existence of solutions.

5. $\alpha=(p+1)/4$, α: noninteger

Finally, we shall briefly discuss solutions of Eq. (2.4) that correspond to nonintegral values of α. The simplest assumption for possible solutions may be of the form

$$f = \pm (ay+b)^{(p+1)/4}, \quad p \neq 0,1. \tag{2.26}$$

Substitution of (2.26) into Eq. (2.4) yields

$$\begin{aligned}&[\tfrac{1}{4}(p+1)(9p-7)a^2 + 6(p^2-1)ab\,]y - [\tfrac{3}{8}(p+1)\\ &\quad\times(3p-1)a^2 + 3p(p+1)ab + 4(p+1)b^2]\\ &\quad-\tfrac{1}{2}(p-1)^2(ay+b)^{-(p-3)/2} = 0.\end{aligned} \tag{2.27}$$

There exists only one possibility in order for this equation to be satisfied for arbitrary values of y; one must choose $p = 3$. However, for $p = 3$, α becomes 1 and it must be excluded by the assumption of nonintegral values of α. Thus we can conclude that solutions of the form (2.26) with nonintegral values of $(p+1)/4$ do not exist for Eq. (2.4).

III. LIOUVILLE EQUATION

A. Reduction to ODE

First, we introduce the dependent variable transformation

$$\phi = (1/\lambda)\ln g, \tag{3.1}$$

to recast the Liouville equation (1.2) into the following form:

$$g\Box_4 g - \sum_{\mu=1}^{4}\left(\frac{\partial g}{\partial x_\mu}\right)^2 + \lambda g^3 = 0. \tag{3.2}$$

Assuming that g is a function of the two variables s_1 and s_2, Eq. (3.2) takes the form

$$\begin{aligned}&g\left[4\frac{\partial^2 g}{\partial s_1^2} + 6s_1\frac{\partial^2 g}{\partial s_1\partial s_2} + (3s_1^2 - 2s_2)\frac{\partial^2 g}{\partial s_2^2}\right]\\ &\quad-4\left(\frac{\partial g}{\partial s_1}\right)^2 - 6s_1\frac{\partial g}{\partial s_1}\frac{\partial g}{\partial s_2}\\ &\quad-(3s_1^2 - 2s_2)\left(\frac{\partial g}{\partial s_2}\right)^2 + \lambda g^3 = 0.\end{aligned} \tag{3.3}$$

Moreover, we employ, by noting that Eq. (3.3) is invariant under the scale transformations, $s_1 \to \gamma s_1$, $s_2 \to \gamma^2 s_2$, and $g \to \gamma^{-2} g$, the following *ansatz*:

$$g = [\lambda s_1^2 f(y)\,]^{-1}, \quad y = s_2/s_1^2. \tag{3.4}$$

With the aid of this ansatz, Eq. (3.3) is transformed into the nonlinear ODE for f as follows:

$$\begin{aligned}&(2y-1)(8y-3)(ff'' - f'^2)\\ &\quad+12(2y-1)ff' - 8f^2 - f = 0.\end{aligned} \tag{3.5}$$

One may also observe by introducing a new dependent variable h through the relation

$$f = h^2, \tag{3.6}$$

that Eq. (3.5) is reduced to the form

$$\begin{aligned}&(2y-1)(8y-3)(hh'' - h'^2)\\ &\quad+12(2y-1)hh' - 4h^2 - \tfrac{1}{2} = 0.\end{aligned} \tag{3.7}$$

The resemblance of Eq. (3.7) to Eq. (2.4) should be remarked.

B. Construction of exact solutions

The simplest solution of Eq. (3.5) is a constant solution given by

$$f = -\tfrac{1}{8}, \tag{3.8}$$

which yields an exact solution of Eq. (1.2) in the form

$$\phi = -\frac{1}{\lambda}\ln\left[-\frac{\lambda}{8}\left(\sum_{\mu=1}^{4} x_\mu\right)^2\right]. \tag{3.9}$$

Next, we shall seek solutions of Eq. (3.5) with an asymptotic form $f \to y^\alpha$ $(y \to -\infty)$. As easily confirmed, the only possible value of α is given by $\alpha = 1$. This situation is different from that of the nonlinear Klein–Gordon equation where there exist two possible values of α [see Eq. (2.7)]. Keeping the asymptotic form of the solution in mind, we take the solution in the form

$$f = ay + b, \quad a \neq 0, \tag{3.10}$$

and substitute (3.10) into Eq. (3.5). It turns out that the unknown constants a and b are determined by the following system of algebraic equations:

$$b = -a/4 + \tfrac{1}{8}, \tag{3.11a}$$

$$(2a+1)(a+1) = 0. \tag{3.11b}$$

Equations (3.11) have two pairs of solutions, namely $a = -1$, $b = \frac{3}{8}$ and $a = -\frac{1}{2}$, $b = \frac{1}{4}$, respectively. We shall treat the two cases separately.

1. $a=-1$, $b=\frac{3}{8}$

In this case, it follows from (3.1), (3.4), and (3.10) that

$$\phi = -\frac{1}{\lambda}\ln\left[\frac{\lambda}{8}\sum_{\substack{\mu,\nu=1\\(\mu<\nu)}}^{4}(x_\mu - x_\nu)^2\right]. \tag{3.12}$$

The solution (3.12) is regular except for $x_\mu = x_\nu$ $(\mu,\nu = 1\sim 4)$.

2. $a=-\frac{1}{2}$, $b=\frac{1}{4}$

In this case, the solution is expressed in the form

$$\phi = -(1/\lambda)\ln[(\lambda/4)x^2], \tag{3.13}$$

which is spherically symmetric and is regular except for the origin $x_\mu = 0$ $(\mu = 1\sim 4)$.

IV. GENERALIZATION TO n-DIMENSIONAL EUCLIDEAN SPACE

In this section, we shall generalize the solutions presented in the previous sections to those for the nonlinear Klein–Gordon and Liouville equations in n-dimensional Euclidean space. Since the procedure for constructing exact solutions for these equations is almost the same as that for corresponding four-dimensional equations, we shall not discuss the details and present only the main results.

A. Nonlinear Klein–Gordon equation in n-spatial dimensions

The equation that we consider reads

$$\Box_n \phi + \lambda \phi^p = 0, \quad p \neq 0,1, \tag{4.1a}$$

with

$$\Box_n = \sum_{\mu=1}^{n} \frac{\partial^2}{\partial x_\mu^2}, \tag{4.1b}$$

where $\Box_n$ is the n-dimensional Laplace operator. If we introduce the dependent variable transformation

$$\phi = [\sqrt{\lambda}\, s_1 f(y)]^{-2/(p-1)}, \quad y = s_2/s_1^2, \tag{4.2}$$

Eq. (4.1) is transformed into the following nonlinear ODE for f:

$$\begin{aligned}&[4ny^2 - 2(2n-1)y + n - 1] \\ &\quad \times (ff'' - [(p+1)/(p-1)]f'^2) \\ &\quad + [2/(p-1)][n(3p+1)y - 2(n-1)p]ff' \\ &\quad - [n(p+1)/(p-1)]f^2 - (p-1)/2 = 0.\end{aligned} \tag{4.3}$$

An exact solution of Eq. (4.1) that corresponds to a constant solution of Eq. (4.3), namely

$$f = \pm i(p-1)/\sqrt{2n(p+1)}, \tag{4.4}$$

is given by the form

$$\phi = [\pm i(p-1)[\lambda/2n(p+1)]^{1/2} s_1]^{-2/(p-1)}, \tag{4.5a}$$

with

$$s_1 = \sum_{\mu=1}^{n} x_\mu. \tag{4.5b}$$

Expression (4.5) is a generalization of (2.6) in n-spatial dimensions.

If we seek solutions of Eq. (4.3) with an asymptotic form $f \sim y^\alpha$ $(y \to -\infty)$, we find two exact solutions. The first solution is given by the expression

$$\begin{aligned}f = &\frac{p-1}{(2n)^{1/2}[(p-1)n - 3p + 1]^{1/2}} \\ &\times (-2ny + n - 1)^{1/2}, \quad (p-1)n \neq 3p - 1,\end{aligned} \tag{4.6}$$

which, substituted in (4.2), yields an exact solution of Eq. (4.1),

$$\begin{aligned}\phi = &\left\{ \pm \frac{\sqrt{\lambda}\,(p-1)}{(2n)^{1/2}[(p-1)n - 3p + 1]^{1/2}} \right. \\ &\left. \times [(n-1)s_1^2 - 2ns_2]^{1/2} \right\}^{-2/(p-1)}, \\ &(p-1)n \neq 3p - 1,\end{aligned} \tag{4.7a}$$

with

$$s_2 = \sum_{\substack{\mu,\nu=1 \\ (\mu<\nu)}}^{n} x_\mu x_\nu, \tag{4.7b}$$

and s_1 given by (4.5b). It is worthwhile to see that

$$(n-1)s_1^2 - 2ns_2 = \sum_{\substack{\mu,\nu=1 \\ (\mu<\nu)}}^{n} (x_\mu - x_\nu)^2, \tag{4.8}$$

which is an analog of (2.15) in n-spatial dimensions. One readily notices that the solution (4.7) is a generalization of (2.12).

On the other hand, the second solution of Eq. (4.3) is written in the form

$$\begin{aligned}f = &\frac{p-1}{2^{1/2}[(p-1)n - 2p]^{1/2}} \\ &\times (-2y + 1)^{1/2}, \quad (p-1)n \neq 2p.\end{aligned} \tag{4.9}$$

Substitution of (4.9) into (4.2) yields an exact solution of Eq. (4.1),

$$\begin{aligned}\phi = &\left[\pm \frac{\sqrt{\lambda}\,(p-1)}{2^{1/2}[(p-1)n - 2p]^{1/2}} (s_1^2 - 2s_2)^{1/2} \right]^{-2/(p-1)}, \\ &(p-1)n \neq 2p,\end{aligned} \tag{4.10}$$

which is seen to be the n-dimensional generalization of solution (2.13). Expression (4.10) represents the spherically symmetric solution due to the relation

$$s_1^2 - 2s_2 = \sum_{\mu=1}^{n} x_\mu^2 \equiv x^2. \tag{4.11}$$

Among various exact solutions presented here, a particularly interesting case arises for a special value of $p = (n+2)/(n-2)$. In this case, Eq. (4.1) becomes

$$\Box_n \phi + \lambda \phi^{(n+2)/(n-2)} = 0, \tag{4.12}$$

which is known as a conformally invariant nonlinear scalar field equation in n-spatial dimensions.[22] Equation (1.1) with $p = 3$ is a special case of Eq. (4.12) with $n = 4$. With this choice of the parameter p, the solutions (4.5), (4.7), and (4.10) are reduced to the expressions

$$\phi = \left[\pm i \frac{2}{n} \sqrt{\frac{\lambda}{n-2}} \left(\sum_{\mu=1}^{n} x_\mu \right) \right]^{-(n-2)/2}, \quad n \neq 2, \tag{4.13}$$

$$\begin{aligned}\phi = &\left\{ \pm \frac{2\sqrt{\lambda}}{[n(n-2)(n-4)]^{1/2}} \right. \\ &\left. \times \left[\sum_{\substack{\mu,\nu=1 \\ (\mu<\nu)}}^{n} (x_\mu - x_\nu)^2 \right]^{1/2} \right\}^{-(n-2)/2}, \quad n \neq 2,4,\end{aligned} \tag{4.14}$$

$$\phi = (\pm [2\sqrt{\lambda}/(n-2)]\sqrt{x^2})^{-(n-2)/2}, \quad n \neq 2, \tag{4.15}$$

respectively. Of course, one can use conformal transformations in n-spatial dimensions to generate new exact solutions of Eq. (4.12) starting from (4.13)–(4.15). But the details will be omitted here. Finally, it should be pointed out that Eq. (4.12) never possesses solutions of the form (2.24) except for $n = 4$. This fact may indicate a peculiar aspect of four-dimensional Euclidean space.

B. Liouville equation in n-spatial dimensions

The equation that we will consider is written

$$\Box_n \phi + e^{\lambda\phi} = 0. \tag{4.16}$$

Introducing a new dependent variable f through the relation

$$\phi = -(1/\lambda)\ln[\lambda s_1^2 f(y)], \quad y = s_2/s_1^2, \tag{4.17}$$

transforms Eq. (4.16) into the nonlinear ODE

$$[4ny^2 - 2(2n-1)y + n - 1](ff'' - f'^2) + [6ny - 4(n-1)]ff' - 2nf^2 - f = 0. \quad (4.18)$$

We have found three exact solutions of Eq. (4.18) as follows:

$$f = -1/2n, \quad (4.19)$$

$$f = [1/(n-3)](-y + (n-1)/2n), \quad n \neq 3, \quad (4.20)$$

$$f = [1/2(n-2)](-2y + 1), \quad n \neq 2, \quad (4.21)$$

which, substituted in (4.7), yield exact solutions of Eq. (4.16) as follows:

$$\phi = -\frac{1}{\lambda}\ln\left[-\frac{\lambda}{2n}\left(\sum_{\mu=1}^{n} x_\mu\right)^2\right], \quad (4.22)$$

$$\phi = -\frac{1}{\lambda}\ln\left[\frac{\lambda}{2n(n-3)}\sum_{\substack{\mu,\nu=1 \\ (\mu<\nu)}}^{n}(x_\mu - x_\nu)^2\right], \quad n \neq 3, \quad (4.23)$$

$$\phi = -(1/\lambda)\ln[[\lambda/2(n-2)]x^2], \quad n \neq 2. \quad (4.24)$$

The solutions (4.22)–(4.24) are generalizations of the solutions (3.9), (3.12), and (3.13) in n-spatial dimensions, respectively.

V. CONCLUDING REMARKS

In this paper, we have developed a systematic method for constructing exact solutions of the nonlinear Klein–Gordon and Liouville equations in four and general n-spatial dimensions. These equations are of course physically important and various particular solutions obtained here may be employed to elucidate the physical phenomena governed by the equations. The method of analysis presented here may also be applied to other types of nonlinear PDE's of physical interest such as the Yang–Mills equation, the Einstein equation of general relativity and other gauge field equations.

From the mathematical point of view, the broader classes of solutions may exist if we introduce other symmetry variables in addition to s_1 and s_2. In this respect, it should be remarked that the number of independent variables in Eqs. (1.1) and (1.2), for example, can be reduced to three by introducing the new independent variables $y_1 = s_2/s_1^2$, $y_2 = s_3/s_1^3$, and $y_3 = s_4/s_1^4$. Furthermore, one may pursue the possibility of generalizations of our solutions to elliptic functions. For this purpose, it will be useful to refer to several works concerning elliptic solutions of Eq. (1.1) with $p = 3$, 5.[3,16,23,24]

Various problems proposed here will be dealt with in the near future.

ACKNOWLEDGMENT

The author wishes to thank Professor M. Nishioka for useful discussions about gauge theories.

[1] S.-k. Ma, *Modern Theory of Critical Phenomena* (Benjamin, Reading, MA, 1976).
[2] R. Jackiw, Rev. Mod. Phys. **49**, 681 (1977).
[3] A. Actor, Rev. Mod. Phys. **51**, 461 (1979).
[4] P. B. Burt, *Quantum Mechanics and Nonlinear Waves* (Harwood Academic, New York, 1981).
[5] M. J. Ablowitz and H. Segur, *Solitons and Inverse Scattering Transform* (SIAM, Philadelphia, 1981).
[6] F. Calogero and A. Degasperis, *Spectral Transform and Solitons I* (North-Holland, Amsterdam, 1982).
[7] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and Nonlinear Wave Equations* (Academic, London, 1983).
[8] *Bäcklund Transformations, Lecture Notes in Mathematics*, Vol. 515, edited by R. M. Miura (Springer, Berlin, 1976).
[9] C. Rogers and W. F. Shadwick, *Bäcklund Transformations and Their Applications* (Academic, New York, 1982).
[10] "Solitons," *Topics in Current Physics*, Vol. 17, edited by R. K. Bullough and P. J. Caudrey (Springer, Berlin, 1980).
[11] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).
[12] G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer, New York, 1974).
[13] P. Winternitz, "Lie groups and solutions of nonlinear differential equations," in *Lecture Notes in Physics*, Vol. 189, edited by K. B. Wolf (Springer, Berlin, 1983).
[14] A. M. Grundland, J. Harnad, and P. Winternitz, "Symmetry reduction for nonlinear wave equations in Riemanian and pseude-Riemanian spaces," in *Wave Phenomena: Modern Theory and Applications*, edited by C. Rogers and T. B. Moodie (North-Holland, Amsterdam, 1984); J. Math. Phys. **25**, 791 (1984).
[15] P. J. Oliver, *Applications of Lie Groups to Differential Equations* (Springer, Berlin, 1986).
[16] A. M. Grundland, J. A. Tuszynski, and P. Winternitz, Phys. Lett. A **119**, 340 (1987).
[17] D. T. Stoyanov, Lett. Math. Phys. **12**, 93 (1986).
[18] G. Leibbrant, S.-s. Wang, and N. Zamani, J. Math. Phys. **23**, 1566 (1982).
[19] G. Leibbrandt, "Do Liouville's solutions possess topological charge?," in *Wave Phenomena: Modern Theory and Applications*, edited by C. Roger and T. B. Moodie (North-Holland, Amsterdam, 1984).
[20] A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Yu. S. Tyupkin, Phys. Lett. B **59**, 85 (1975).
[21] Y. Matsuno, to appear in Lett. Math. Phys.
[22] S. Fubini, Nuovo Cimento A **34**, 521 (1976).
[23] J. Cervero, L. Jacobs, and C. R. Nohl, Phys. Lett. B **69**, 351 (1977).
[24] A. Actor, Lett. Math. Phys. **2**, 275 (1978); Ann. Phys. (NY) **121**, 181 (1979); **131**, 269 (1981).

Tensors with icosahedral symmetry that are invariant under a certain wreath product

Adalbert Kerber and Thomas Scharf
Lehrstuhl II für Mathematik, Universität Bayreuth, Postfach 101251, 8580 Bayreuth, West Germany

(Received 12 March 1987; accepted for publication 27 May 1987)

Faithful icosahedral symmetry exists only for tensors of rank higher than 5. The most relevant tensor of this type is the one for third-order elastic constants $C^{(3)}_{ijklmn}$ defined by the series expansion $T_{ij} = C^{(2)}_{ijkl}\epsilon_{kl} + \frac{1}{2}C^{(3)}_{ijklmn}\epsilon_{kl}\epsilon_{mn} + \cdots$ of the stress tensor T in terms of the deformation tensor ϵ. A basis for those tensors in $(\mathbb{R}^3)^{\otimes 6}$ that are invariant under a certain action of both the icosahedral group $S_2\times A_5$ and the wreath product $S_2 \wr S_3$ of the symmetric groups S_2 and S_3 are evaluated.

I. INTRODUCTION

H.-R. Trebin has drawn our attention to the following problem concerning the elasticity of quasicrystalline structures: Faithful icosahedral symmetry exists only for tensors of rank higher than 5. The most relevant tensor of this type is the one for third-order elastic constants $C^{(3)}_{ijklmn}$ defined by the series expansion

$$T_{ij} = C^{(2)}_{ijkl}\epsilon_{kl} + \tfrac{1}{2}C^{(3)}_{ijklmn}\epsilon_{kl}\epsilon_{mn} + \cdots$$

of the stress tensor T in terms of the deformation tensor ϵ. Hence the question arises of how one can get a basis for those tensors in $(\mathbb{R}^3)^{\otimes 6}$ that are invariant under both the icosahedral group $S_2\times A_5$ and the wreath product $S_2 \wr S_3$ of the symmetric groups S_2 and S_3, which acts on $(\mathbb{R}^3)^{\otimes 6}$ via permuting in $C^{(3)}_{ijklmn}$ the elements in the pairs (i,j), (k,l), *and* (m,n) as well as the pairs themselves among each other (i.e., $S_2 \wr S_3$ is canonically imbedded into the S_6).

This problem can be attacked and solved in the three steps we shall describe in this paper. At first we formulate the situation in terms of representation theory, so that the space of invariant tensors is just the subspace of tensors that afford the identity representation.

II. THE REPRESENTATION THEORETICAL FORMULATION

$S_2\times A_5$ affords on $\mathbb{R}^3$ the faithful representations[1]

$$[1^2]\#[3,1^2]^+ \quad \text{and} \quad [1^2]\#[3,1^2]^-.$$

The wreath product[1] $S_2 \wr S_3$ affords on $(\mathbb{R}^3)^{\otimes 6}$ the representation[1]

$$\widetilde{\#^3[2]\otimes[3]},$$

which induces in S_6 the plethysm[1] $[2]\odot[3]$.

Thus in terms of representation theory the desired space is the subspace affording the identity representation in the representation

$$([1^2]\#[3,1^2]^{\pm})\boxdot([2]\odot[3]). \tag{2.1}$$

But we can simplify the considerations as follows.

The restriction of this representation to the subgroup S_2 is the identity representation since S_2 acts diagonally on $\mathbb{R}^3$ and the power 6 of $\mathbb{R}^3$ is taken, which is even. Therefore the factor $[1^2]$ can be neglected. Furthermore it does not matter whether we take $[3,1^2]^+$ or $[3,1^2]^-$ since the corresponding matrix groups are the same—these representations are conjugate[1] with respect to S_5—and the construction below does not need further information. Thus we have obtained the following corollary.

Corollary 2.1: We need only to consider the representation $[3,1^2]^+\boxdot([2]\odot[3])$.

III. THE SUBSPACE OF INVARIANT TENSORS

Corollary 2.1 has shown that we have to find a basis of the subspace of $(\mathbb{R}^3)^{\otimes 6}$ whose elements are 0 or can afford the identity representation. In order to get it we evaluate a generating system (in three steps) and finally pick a basis. The steps to be peformed use the following lemma.

Lemma 3.1: If a group G acts on a vector space V and if N is a normal subgroup of G, then we have a natural action of G/N on the space V_N of vectors invariant under N. Therefore the corresponding spaces of invariants satisfy

$$(V_N)_{G/N} = V_G.$$

In the present situation we can use the following chain of normal subgroups:

$$S_2\times S_2\times S_2 \lhd S_2 \wr S_3 \lhd A_5\times S_2 \wr S_3,$$

where

$$(A_5\times S_2 \wr S_3)/S_2 \wr S_3 \cong A_5, \quad S_2 \wr S_3/(S_2\times S_2\times S_2) \cong S_3.$$

IV. PICKING A BASIS

Let us first describe $(\mathbb{R}^3)^{\otimes 6}_{S_2\wr S_3}$. As one of its bases $\mathbb{R}^3$ has the set of standard unit vectors $\{e_i \mid 1\leqslant i\leqslant 3\}$. Thus

$$(\mathbb{R}^3)^{\otimes 2} = \langle\langle e_i\otimes e_j \mid 1\leqslant i,j\leqslant 3\rangle\rangle_{\mathbf{R}}.$$

Hence the subspace of tensors invariant under S_2 is

$$((\mathbb{R}^3)^{\otimes 2})_{S_2} = \langle\langle \tfrac{1}{2}(e_i\otimes e_j + e_j\otimes e_i) \mid 1\leqslant i\leqslant j\leqslant 3\rangle\rangle_{\mathbf{R}}. \tag{4.1}$$

This shows that we may identify the elements of this space with the linear combinations of indeterminates y_k, $1\leqslant k\leqslant 6$, where $k = 1,\ldots,6$ corresponds to the pairs $(i,j) = (1,1),(1,2),(1,3),(2,2),(2,3),(3,3)$.

Now we form the tensor cube of this space obtaining $((\mathbb{R}^3)^{\otimes 6})_{S_2\times S_2\times S_2}$ and symmetrize it with S_3 getting the desired space

$$((\mathbb{R}^3)^{\otimes 6})_{S_2\wr S_3}. \tag{4.2}$$

Thus we have obtained the following corollary.

Corollary 4.1: This space (4.2) can be considered as be-

 0022-2488/87/102323-02$02.50 © 1987 American Institute of Physics

ing the space of homogeneous polynomials (in the indeterminates y_k) of degree 3.

The representation $[3,1^2]^+$ of A_5 gives in an obvious manner a representation in the space V of polynomials in the y_k homogeneous of degree 3. By Corollary 4.1 it is sufficient to evaluate a basis of the subspace of this representation space affording the identity representation.

Consider a five-cycle $\tau\in A_5$ and a $\sigma\in A_5$ such that $\rho := \sigma\tau\sigma^{-1}\neq\tau^i$ for all i. This means that ρ, τ are rotations corresponding to different axes of symmetry. Hence ρ and τ generate A_5: $A_5 = \langle\rho,\tau\rangle$. By restricting the representation of A_5 to the subgroups $\langle\tau\rangle$ and $\langle\rho\rangle$, it is easy to evaluate a basis of the subspaces V_1, V_2 that afford the identity representation of $\langle\tau\rangle$ and $\langle\rho\rangle$, respectively. The intersection of these spaces is just the desired space.

A generating set of V_1 will be obtained by applying the so-called Reynolds operator

$$R:\ V\to V:\ v\mapsto \frac{1}{|\langle\tau\rangle|}\sum_{0<i\leq 4}\tau^i v$$

corresponding to $\langle\tau\rangle$ and V to the standard basis of monomials. A generating set of V_2 will result from a base change.

We chose a realization of the representation $[3,1^2]^+$ such that the corresponding matrix representation θ—relative to the e_i—has a nice structure[2]:

$$\theta(\tau)=\begin{pmatrix}1&0&0\\0&a&-b\\0&b&a\end{pmatrix},$$

$$\theta(\sigma)=\begin{pmatrix}d&-(1+d)/2&-1/(2b)\\2d&(1+d)/4&1/(4b)\\0&-2ab&(1+1/d)/4\end{pmatrix},$$

$$\text{with}\quad a:=\cos(2\pi/5),\quad b:=\sin(2\pi/5),\quad d:=1/\sqrt{5}. \tag{4.3}$$

A computation with the help of the MACSYMA programming system gave a basis of V_1 and V_2 and finally we obtained the following corollary.

Corollary 4.2: A basis of the desired space $V_1\cap V_2$ is the set $\{b_1,b_2,b_3,b_4\}$, where

$$\begin{aligned}
b_1 :=& \tfrac{1}{16}(y_6^3+7y_4y_6^2+7y_1y_6^2-4y_5^2y_6+7y_4^2y_6+42y_1y_4y_6-4y_3^2y_6-28y_2^2y_6+7y_1^2y_6-4y_4y_5^2\\
&-28y_1y_5^2+48y_2y_3y_5+y_4^3+7y_1y_4^2-28y_3^2y_4-4y_2^2y_4+7y_1^2y_4-4y_1y_3^2-4y_1y_2^2+y_1^3),\\
b_2 :=& \tfrac{1}{16}(5y_6^3+3y_4y_6^2+3y_1y_6^2+12y_5^2y_6+3y_4^2y_6+18y_1y_4y_6+12y_3^2y_6-12y_2^2y_6+3y_1^2y_6+12y_4y_5^2-12y_1y_5^2\\
&+48y_2y_3y_5+5y_4^3+3y_1y_4^2-12y_3^2y_4+12y_2^2y_4+3y_1^2y_4+12y_1y_3^2+12y_1y_2^2+5y_1^3),\\
b_3 :=& -2y_1y_4y_6+2y_2^2y_6+2y_1y_5^2-4y_2y_3y_5+2y_3^2y_4,\\
b_4 :=& \tfrac{1}{32}(2y_2y_6^2+y_1y_6^2+8y_3y_5y_6-4y_2y_4y_6+6y_1y_4y_6+4y_3^2y_6-4y_2^2y_6-y_1^2y_6-8y_2y_5^2-4y_1y_5^2\\
&-8y_3y_4y_5+16y_2y_3y_5+2y_2y_4^2+y_1y_4^2-4y_3^2y_4+4y_2^2y_4-y_1^2y_4-4y_1y_3^2-4y_1y_2^2+y_1^3).
\end{aligned}$$

In Sec. VI we will be able to give a nicer basis.

V. A GENERALIZATION

This method can be applied analogously for any wreath product $S_2\wr S_j$ and icosahedral symmetry. To cover these cases we need only to consider the algebra of polynomials in the y_i, $1\leqslant i\leqslant 6$.

In this case the Molien series[3]

$$P_{A_5}(T) := \sum_{j>0}\dim(I_j)T^j\in\mathbb{R}[T], \tag{5.1}$$

where I_j is the space of invariant polynomials of degree j, can be evaluated in order to get the dimension of the spaces invariant under $S_2\wr S_j$ and the icosahedral group. We will apply Molien's formula[3]

$$P_{A_5}(T)=\frac{1}{|A_5|}\sum_{g\in A_5}\frac{1}{\det(I-T\mu(g))}, \tag{5.2}$$

where I denotes the identity matrix and μ a matrix representation of A_5 on the span $\langle\langle y_i\,|\,1\leqslant i\leqslant 6\rangle\rangle_{\mathbb{R}}$. We may clearly pass to $\mathbb{C}$. Once for each $g\in A_5$ the eigenvalues of $\mu(g)$ are known, we are able to compute (5.2). But μ has been constructed from $[3,1^2]^+$ and the eigenvalues of the matrices corresponding to this representation can be easily reconstructed from the well-known character table[1] of A_5. The first of these dimensions (starting with $j=0$) are

$$1,1,2,4,6,10,17,24,36,53,74,102,141,$$
$$186,246,322,412,523,661,820. \tag{5.3}$$

VI. A NOTE ON ORTHOGONAL INVARIANTS

Consider the following polynomials:

$$\begin{aligned}
c_1 :=& y_1+y_4+y_6,\\
c_2 :=& y_1^2+2y_2^2+2y_3^2+y_4^2+2y_5^2+y_6^2,\\
c_3 :=& y_1^3+3y_1y_2^2+3y_1y_3^2+3y_2^2y_4+6y_2y_3y_5\\
&+3y_3^2y_6+y_4^3+3y_4y_5^2+3y_5^2y_6+y_6^3.
\end{aligned} \tag{6.1}$$

It should be noted that the c_i are also invariant, if the full orthogonal group $O(3)$ instead of $S_2\times A_5$ is taken, but b_4 is not. Furthermore we have

$$\begin{aligned}
&I_1=\langle\langle c_1\rangle\rangle_{\mathbb{R}},\quad I_2=\langle\langle c_1^2,c_2\rangle\rangle_{\mathbb{R}},\\
&I_3=\langle\langle c_1^3,c_1c_2,c_3,b_4\rangle\rangle_{\mathbb{R}}.
\end{aligned} \tag{6.2}$$

Hence we obtain the following corollary.

Corollary 6.1: Three is the smallest positive integer, where the icosahedral case differs from the $O(3)$ case.

Thus Corollary 6.1 shows the importance of the third-order elastic constants $C^{(3)}_{ijklmn}$ for icosahedral symmetry.

[1] G. D. James and A. Kerber, *The Representation Theory of the Symmetric Group* (Addison-Wesley, Reading, MA, 1981).
[2] P. Kramer, Z. Naturforsch. **40a**, 775 (1985).
[3] R. P. Stanley, Bull. Am. Math. Soc. (New Series) **1**, 475 (1979).

Local Green's function. I

Renchuan Wang
Department of Physics, Texas A&M University, College Station, Texas 77843 and Center for Astrophysics, USTC, Hefei, Anhui, China

(Received 25 September 1986; accepted for publication 20 May 1987)

The local Green's function is used in many physical problems. In this paper, the properties of the local Green's function are studied, and it is proved that the $N\times N$ local Green's function can represent the results of the full $N_1\times N_1$ Green's function, where N is small (or at least finite) and N_1 is large (or infinite). The accuracy of cutting the general Green's function into the local Green's function is also discussed.

I. INTRODUCTION

The Green's function G is an infinite matrix in the representation of a complete set. Usually we know G^{-1} of a physical system from the requirement of satisfying physics. The eigenvalues and eigenvectors of the system correspond to the condition Det $G^{-1}=0$. Generally speaking, it is very difficult to find exact solutions, because the system is infinite. In many physical problems, however, we can use the approximation of cutting the infinite matrix into a finite matrix, i.e., let $H_{mn}=0$ when $n,m>N$, and this is the main idea of the subspace Hamiltonian technique.[1] Here we will prove that under a certain restriction, the inverse matrix of the local Green's function—i.e., the finite submatrix on the diagonal of the whole Green's function matrix—can represent all the results of the original Green's function (i.e., it gives all the eigenvalues of G^{-1}). (For simplicity, we will say that G_{II} represents G in the rest part of this paper.) Therefore the solution of the local Green's function is equivalent to the solution of the full Green's function. We will also discuss the accuracy of cutting the full Green's function into the local Green's functions (later on we will refer to this as the cutting approximation).

II. LOCAL GREEN'S FUNCTION

The local Green's function G_{II} is a submatrix on the diagonal of the whole matrix. Suppose that in the representation of a complete set of orthonormal functions ($|n\rangle$, $n=1,2,...$), the Green's function can be expressed as[1-3]

$$G=\begin{pmatrix} G_{II} & G_{IF} \\ G_{FI} & G_{FF} \end{pmatrix}, \tag{1}$$

where $I,I^c\subset\mathcal{N}$, $\mathcal{N}=\{1,2,3,...\}$, $I\cup I^c=\mathcal{N}$, and $I\cap I^c=\varnothing$. Here I is a finite subset of $\mathcal{N}$. For clarity let us suppose that $I=\{1,2,3,...,N\}$.

Normally, in physical problems, G is the matrix to be solved for and G^{-1} is given by

$$G^{-1}=\begin{pmatrix} G^{-1}_{II} & G^{-1}_{IF} \\ G^{-1}_{FI} & G^{-1}_{FF} \end{pmatrix}=EI-\begin{pmatrix} H_{II} & H_{IF} \\ H_{FI} & H_{FF} \end{pmatrix}, \tag{2}$$

where I is the unit matrix

$$I=\begin{pmatrix} I_{II} & 0 \\ 0 & I_{FF} \end{pmatrix}$$

and I_{II} is the unit matrix of order N,

$$I_{II}=\begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots \\ & & & 1 \end{pmatrix}\begin{matrix} \\ \\ \\ \\ \end{matrix}\begin{matrix} 1\to N \\ \\ \\ N+1\to\infty \end{matrix},$$

$$I_{I^cI^c}=\begin{pmatrix} 1 & & & \\ & 1 & & 0 \\ & & 1 & \\ & 0 & & \ddots \end{pmatrix}. \tag{3}$$

The inverse of the local Green's function is G^{-1}_{II}, which satisfies

$$G^{-1}_{II}G_{II}=I_{II}. \tag{4}$$

Theorem 1:

$$\det G^{-1}(E)=\det G^{-1}_{II}(E)\det G^{-1}_{I^cI^c}(E). \tag{5}$$

Proof: First we prove that when $E\notin J=J_1\cup J_2$, where $J_1=\{E,\det G^{-1}_{FF}(E)=0\}$ and $J_2=\{E,\det G_{II}(E)=0\}$, Eq. (5) is true.

Notice that $G(E)G^{-1}(E)=I$, we can obtain

$$\begin{aligned} &G_{II}(E)G^{-1}_{IF}+G_{IF}(E)G^{-1}_{FF}(E)=0, \\ &G_{FI}(E)G^{-1}_{IF}+G_{FF}(E)G^{-1}_{FF}(E)=I_{FF}. \end{aligned} \tag{6}$$

Owing to $E\notin J$, Eq. (6) can be written as follows:

$$\begin{aligned} &G^{-1}_{II}(E)G_{IF}(E)=-G^{-1}_{IF}G^{-1}_{FF}{}^{-1}(E), \\ &G_{FI}(E)G^{-1}_{IF}G^{-1}_{FF}{}^{-1}(E)+G_{FF}(E) \\ &\quad=G^{-1}_{FF}{}^{-1}(E). \end{aligned} \tag{7}$$

Using Eq. (7), we have

$$\begin{aligned} \det G^{-1}(E)&=\left[\det\begin{pmatrix} G_{II}(E) & 0 \\ 0 & I_{FF} \end{pmatrix}\det\begin{pmatrix} I_{II} & G^{-1}_{II}(E)G_{IF}(E) \\ G_{I^cI}(E) & G_{FF}(E) \end{pmatrix}\right]^{-1} \\ &=\det G^{-1}_{II}(E)\left[\det\begin{pmatrix} I_{II} & -G^{-1}_{IF}G^{-1}_{FF}{}^{-1}(E) \\ G_{FI}(E) & G_{FF}(E) \end{pmatrix}\right]^{-1} \\ &=\det G^{-1}_{II}(E)\left[\det\begin{pmatrix} I_{II} & 0 \\ G_{FI}(E) & G_{FI}(E)G^{-1}_{IF}G^{-1}_{FF}{}^{-1}(E)+G_{FF}(E) \end{pmatrix}\right]^{-1} \\ &=\det G^{-1}_{II}(E)\det G^{-1}_{FF}(E). \end{aligned}$$

 0022-2488/87/102325-04$02.50 © 1987 American Institute of Physics

Second, when $E \in J$, it is not difficult to prove that Eq. (5) is still true. In fact, since $\det G_{I^cI^c}^{-1}(E)$ and $\det G_{II}(E)$ are a polynomial and a rational function of E, respectively, the J would be a numerable set. In the physical sense, it is obvious that J is not dense in any subregion of complex plane of E. For any $E_m \in J$, there is always a small circle $U_{\delta_0}(E_m)$, which is centered at E_m and with radius $\delta_0(>0)$, and $U_{\delta_0}(E_m) \cap J = E_m$. Since $\det G^{-1}(E)$ is a polynomial function of E, $\det G^{-1}(E)$ is an analytical function in the E complex plan. For any $\epsilon > 0$, there is a small positive number $\delta(\leqslant \delta_0)$, when $|E - E_m| < \delta$, we have

$$|\det| G^{-1}(E_m) - \det G^{-1}(E)| < \epsilon .$$

If $E \neq E_m$, then $E \notin J$, according to the above proof,

$$\det G^{-1}(E) = \det G_{II}^{-1}(E) \det G_{I^cI^c}^{-1}(E) .$$

So, when $|E - E_m| < \delta$, we have

$$|\det G^{-1}(E_m) - \det G_{II}^{-1}(E) \det G_{I^cI^c}^{-1}(E)| < \epsilon ,$$

then

$$\lim_{E \to E_m} \det G_{II}^{-1}(E) \det G_{I^cI^c}^{-1}(E) = \det G^{-1}(E_m) .$$

That is,

$$\det G^{-1}(E_m) = \det G_{II}^{-1}(E_m) \det G_{I^cI^c}^{-1}(E_m) .$$

This is the complete proof of Theorem 1.

From Theorem 1 we can obtain three important properties of the local Green's function as the following.

Remark 1: Since $\det G_{I^cI^c}^{-1}(E)$ is a polynomial function of E it is always limited by any $|E| < \infty$. When $\det G_{II}^{-1}(E) = 0$, from Theorem 1, $\det G^{-1}(E)$ must be zero. The roots of $\det G^{-1}(E)$ are real, so the roots of the inverse determinant of the local Green's function must be real and a part of the roots of the inverse determinant of the Green's function. That is,

$$\{E_K, \det G_{II}^{-1}(E_K) = 0\} \subset \{E_J, \det G^{-1}(E_J) = 0\} . \tag{8}$$

Remark 2: Since

$$\{E_L, \det G_{II}(E_L) = 0\} = \{E_L, \det G_{II}^{-1}(E_L) = \infty\}$$

and $\det G^{-1}(E)$ is limited, by Theorem 1, we have $J_2 \subset J_1$, i.e.,

$$\{E_L, \det G_{II}(E_L) = 0\} \subset \{E_J, \det G_{I^cI^c}^{-1}(E_J) = 0\}$$

so the roots of the determinant of the local Green's function are real.

Remark 3: Since the set of the roots of the inverse determinant of the Green's function is determined, the total number N_1 of the roots of the $\det G_{II}^{-1}(E)$ is equal to the dimension N of the local Green's function $G_{II}(E)$ plus the roots number m of the determinant of the local Green's function $G_{II}(E)$, that is,

$$N_1 = N + m . \tag{9}$$

Here, saying that the local Green's function G_{II} represents all the solutions of the original general Green's function just means that all the solutions of the equation $\det G_{II}^{-1}(E) = 0$ will be exactly the same as all the solutions of the equation $\det G^{-1}(E) = 0$, i.e.,

$$\{E_k, \det G_{II}^{-1}(E_K) = 0\} = \{E_J, \det G^{-1}(E_J) = 0\} .$$

From the preceding remarks, we can get the sufficient and necessary condition to represent G^{-1} by G_{II}^{-1} as follows:

$$\{E_L, \det G_{II}(E_L) = 0\} = \{E_J, \det G_{I^cI^c}^{-1}(E_J) = 0\} . \tag{10}$$

On the other hand, we must find G_{II}^{-1} from G^{-1}, since we only know G^{-1}. So we have Theorem 2,

$$G_{II}^{-1} = G_{II}^{-1} - G_{II^c}^{-1} G_{I^cI^c}^{-1} G_{I^cI}^{-1} . \tag{11}$$

Proof: From the equation $G \cdot G^{-1} = I$ we have

$$G_{II} G_{II}^{-1} + G_{II^c} G_{I^cI}^{-1} = I_{II} .$$

Multiplying by G_{II}^{-1} from the left, we have

$$G_{II}^{-1} = G_{II}^{-1} + G_{II}^{-1} G_{II^c} G_{I^cI}^{-1} . \tag{12}$$

By using Eq. (7), and substituting $G_{II}^{-1} G_{II^c}$ for $-G_{II^c}^{-1} G_{I^cI^c}^{-1}$, we obtain Eq. (11).

Equation (11) is an identical equation, with only the requirement that $\mathrm{Det}(G_{I^cI^c}^{-1}) \neq 0$. It looks as if we have found G_{II}^{-1} from G^{-1} by using Eq. (11), because I is finite. We certainly can find G_{II} as a part of G. But the really difficult problem has not been solved yet, because I^c is infinite, and there is a sum over infinite terms in Eq. (11). So the real task is to find the exact form or an approximate form of G_{II}^{-1} by using Eq. (11). By using the projection operators, we can write Eq. (11) in another equivalent form.

Let us use $\{|n\rangle, n \in \mathcal{N}\}$ as a set of complete orthonormal functions

$$\langle m|n\rangle = \delta_{mn} ,$$

$$\sum_{n \in \mathcal{N}} |n\rangle\langle n| = I .$$

Also define

$$P = \sum_{n \in I} |n\rangle\langle n| ,$$

$$\Theta = \sum_{n \in I^c} |n\rangle\langle n| . \tag{13}$$

Thus P and Θ are projection operators, and they satisfy

$$PP = P, \quad P\Theta = 0, \quad \Theta\Theta = \Theta, \quad P + \Theta = I . \tag{14}$$

The submatrices of G and G^{-1} can be expressed as

$$\begin{aligned} G_{II}^{-1} &= (PGP)^{-1} , \\ G_{II}^{-1} &= PG^{-1}P , \\ G_{II^c}^{-1} &= PG^{-1}\Theta , \\ G_{I^cI}^{-1} &= \Theta G^{-1} P , \\ G_{I^cI^c}^{-1} &= (\Theta G^{-1}\Theta)^{-1} . \end{aligned} \tag{15}$$

Equation (11) can be expressed in the form

$$(PGP)^{-1} = PG^{-1}P - PG^{-1}\Theta(\Theta G^{-1}\Theta)^{-1}\Theta G^{-1}P . \tag{16}$$

In a more general matrix representation, suppose there exists the inverse matrix A^{-1} of A, and $\det(\Theta A^{-1}\Theta) \neq 0$. There then is a relation as follows:

$$(PAP)^{-1} = PA^{-1}P - PA^{-1}\Theta(\Theta A^{-1}\Theta)^{-1}\Theta A^{-1}P . \tag{17}$$

III. THE APPROXIMATION METHOD OF THE LOCAL GREEN'S FUNCTION (THE CUTTING APPROXIMATION)

We have

$$\hat{H}_0 = -(\hbar^2/2m)\nabla^2 + U(\mathbf{x}),$$
$$\hat{H} = \hat{H}_0 + V(\mathbf{x}), \quad \hat{H}_0|n\rangle = E_n|n\rangle,$$
$$\langle m|n\rangle = \delta_{mn}, \quad \sum_{n\in\mathcal{N}} |n\rangle\langle n| = I,$$

and

$$V_{mn} = \langle m|V|n\rangle . \tag{18}$$

Suppose that

$$|V_{mn}| \ll |E_n| \quad (n = 1,2,...,N) .$$

Then the "cutting approximation" of the previous section is given by the matrix representation

$$\begin{pmatrix} & m & n \\ & 1 & 1\to\infty \\ V_{mn} & \downarrow & \\ & \infty & \end{pmatrix} \cong \begin{pmatrix} V_{11},...,V_{1N} & & 0 \\ \cdots & & \\ V_{N1},...,V_{NN} & & \\ 0 & & 0 \end{pmatrix} = \begin{pmatrix} V_{II} & 0 \\ 0 & 0 \end{pmatrix},$$

$$G_0^{-1} = E - \hat{H}_0, \tag{19}$$
$$G^{-1} = E - \hat{H}.$$

Using Eq. (18), we obtain

$$PG^{-1}\Theta = -PV\Theta, \quad \Theta G^{-1}P = -\Theta VP. \tag{20}$$

Substituting into Eq. (16), we have

$$(PGP)^{-1} = PG^{-1}P - PV\Theta(\Theta G^{-1}\Theta)^{-1}\Theta VP.$$

By using the above cutting approximation, we get

$$PV\Theta \cong 0, \quad \Theta VP \cong 0$$

so

$$(PGP)^{-1} \cong PG^{-1}P. \tag{21}$$

This is the method used most frequently. When

$$\det(PGP)^{-1} \cong \det(PG^{-1}P)$$

is fulfilled, we can obtain N first-order energy eigenvalues E of accuracy $V_{m,n}$ $(m,n\leqslant N)$. Equation (21) is called the first-order approximate equation. We can obtain the second-order approximate equation by using Eq. (16).

Lemma 1: We have

$$|V_{m,n}| \to 0 \quad (m \text{ or } n \to \infty) .$$

Proof: Clearly, we have

$$\langle k|V^2|k\rangle = \sum_{s=1}^{\infty} \langle k|V|s\rangle\langle s|V|k\rangle$$
$$= \sum_{s=1}^{\infty} V_{sk}^* V_{sk} = \sum_{s=1}^{\infty} V_{ks} V_{ks}^* . \tag{22}$$

Since $\langle k|V^2|k\rangle$ is infinite, when $s\to\infty$, $|V_{sk}|$, or $|V_{ks}| \to 0$.

Roughly speaking, we can always divide V_{mn} into powers of a small quantity λ. Suppose

$$N < N_1 < N_2 < \cdots ,$$

$$V_{11} \sim \lambda ,$$
$$V_{mn}(m,n\leqslant N_1) \sim \lambda , \tag{23}$$
$$V_{mn}(N_1 < m \text{ or } n\leqslant N_2) \sim \lambda^2 ,$$
$$\vdots .$$

Using the projection operators,

$$P_1 = \sum_{n=N+1}^{N_1} |n\rangle\langle n|, \quad \Theta_1 = \sum_{n=N_1+1}^{\infty} |n\rangle\langle n| , \tag{24}$$

where we have

$$P_1P_1 = P_1, \quad P_1P = 0, \quad P_1\Theta_1 = 0,$$
$$\Theta_1\Theta_1 = \Theta_1, \quad P_1 + \Theta_1 = \Theta .$$

The second-order equation is

$$(PGP)^{-1} = PG^{-1}P - PVP_1(\Theta G^{-1}\Theta)^{-1}P_1VP . \tag{25}$$

Since PVP_1 and P_1VP are of the order of λ, we need only take the zero order of $P_1(\Theta G^{-1}\Theta)^{-1}P_1$.

Let $(\Theta G^{-1}\Theta)^{-1} = A$. By using Eq. (17), we obtain

$$[P_1(\Theta G^{-1}\Theta)^{-1}P_1]^{-1}$$
$$= P_1G^{-1}P_1 - P_1V\Theta_1(\Theta_1G^{-1}\Theta_1)^{-1}\Theta_1VP_1$$
$$= P_1G_0^{-1}P_1 - P_1VP_1$$
$$- P_1V\Theta_1(\Theta_1G^{-1}\Theta_1)^{-1}\Theta_1VP_1 . \tag{26}$$

The zero order of $[P_1(\Theta G^{-1}\Theta)^{-1}P_1]^{-1}$ is $P_1G_0^{-1}P_1$.

Lemma 2: If the zero order of matrix B is B_0, then the zero order of matrix B^{-1} is B_0^{-1}.

Proof: We have

$$B = B_0 + \lambda B_1 + \lambda^2 B_2 + \cdots ,$$
$$B^{-1} = C_0 + \lambda C_1 + \lambda^2 C_2 + \cdots ,$$

where C_0 is the zero-order quantity of B^{-1}. By using $BB^{-1} = I$ we obtain

$$C_0 = B_0^{-1},$$
$$C_1 = B_0^{-1}B_1B_0^{-1} \tag{27}$$
$$\vdots .$$

Using Lemma 2, in $P_1(\Theta G^{-1}\Theta)^{-1}P_1$ the zero-order quantity is $(P_1G_0^{-1}P_1)^{-1}$, so the final form of the second-order equation is

$$(PGP)^{-1} = PG^{-1}P - PVP_1(P_1G_0^{-1}P_1)^{-1}P_1VP . \tag{28}$$

Since $\mathrm{Det}[PGP]^{-1} = 0$, from the second-order equation we can obtain N eigenvalues with an accuracy of the order of V_{11}^2. The advantage of using the Green's function is that the number of eigenvalues that we can obtain from the determinant is bigger than the order of the determinant.

Usually the variation of $V_{m,n}$ is not very obvious, so the number of N_1 is not very accurate. If we write the second-order equation as

$$(PGP)^{-1} = PG^{-1}P - PVP_1(P_1G^{-1}P_1)^{-1}P_1VP , \tag{29}$$

without distinguishing the quantities of different order in λ, from $\mathrm{Det}[PGP]^{-1} = 0$, we can obtain N_1 eigenvalues of E, which actualy are N_1 eigenvalues of the first-order equation $(P'GP')^{-1} \cong P'G^{-1}P'$, when N is taken as N_1, where $P' = P + P_1$.

Since

$$V \cong \begin{pmatrix} V_{11},\ldots,V_{1N_1} & & \\ \cdots & & 0 \\ V_{N_1 1},\ldots,V_{N_1 N_1} & & \\ 0 & & 0 \end{pmatrix}, \tag{30}$$

larger values of N_1 obviously give both more eigenvalues E and higher accuracy in the lowest-order eigenvalue, at the expense of a more complicated calculation.

IV. CONCLUSION

The above cutting approximation is a very simple and general method, previously used in many physical applications to find the approximate solutions. Its biggest weakness is that the accuracy cannot be estimated exactly, and the number of eigenvalues is the same as the order of the determinant. But there is one important point: We can obtain Eq. (29) by using the local Green's function, in which case we will not lose eigenvalues when we solve the high-order determinant by using the low-order determinants.

ACKNOWLEDGMENTS

I would like to thank Professor R. E. Allen, Professor C. M. Ko, and Professor L. C. Tsu for their encouragement and helpful suggestions. I am also grateful to Dr. T. X. He, Dr. S. F. Ren, and Mr. R. Yuan for their beneficial discussions and assistance.

This work was supported by the U. S. Office of Naval Research under Contract No. N00014-82-K-0447.

[1] R. E. Allen and M. Menon, Phys. Rev. B **33**, 5611 (1986).
[2] K. Kumar, *Perturbation Theory and Nuclear Many Body Problem* (North-Holland, Amsterdam, 1962).
[3] K. F. Freed, Chem. Phys. Lett. **13**, 181 (1972); **15**, 331 (1972).

Local Green's function. II

Renchuan Wang
Department of Physics, Texas A&M University, College Station, Texas 77843 and Center for Astrophysics, USTC, Hefei, Anhui, China

(Received 25 September 1986; accepted for publication 20 May 1987)

As shown in the preceding paper [J. Math. Phys. **28**, 2325 (1987)], the local Green's function can represent the results of the original general Green's function. However, it is difficult to find the exact local Green's function in the general case. In this paper, a special case—"the chain matrix"—is studied, which is a generalization of the tridiagonal matrice.

I. INTRODUCTION

In solid state physics, when studying electrons or phonons, one often assumes that each atom interacts with only a finite number of neighboring atoms. In this case, one obtains a "chain matrix" of the form

$$\begin{pmatrix} V(1) & & & 0 \\ & V(2) & & \\ & & V(3) & \\ 0 & & & \ddots \end{pmatrix}, \tag{1}$$

where

$$V(K)=\begin{pmatrix} V(K)_{11},\ldots,V(K)_{1N} \\ \cdots \\ V(K)_{N1},\ldots,V(K)_{NN} \end{pmatrix}.$$

The elements in the overlap of $V(k)$ with $V(k-1)$ or $V(k+1)$, are the same, even though their indices are not the same. (The total number for overlap matrix elements of each of the two matrices are not larger than $\frac{1}{4}N^2$.) In a surface problem, the deeper the layer from the surface, the bigger the value of k and the smaller the difference between $V(k)$'s with different values of k. So in a certain approximation, we have $V(k_1)=V(k_2)$ when $k_1,k_2 \geqslant M$, where M is a positive integer. We will then obtain an equation, which we call the chain equation, and by solving the chain equation, we will obtain the exact local Green's function. Although we cannot prove in general that it will represent the whole general Green's function, in principle, we can discuss the exact solution of the eigenvalues of the original general Green's function.[1-3] For convenience, we will treat an $N\times N$ matrix as a vector of $N\times N$ dimensions in the linear space, and the product of matrices can be treated by defining the product of basis vectors. This approach makes our discussion much simpler and clearer.

II. MATHEMATICAL PREPARATION

Then $N\times N$ matrix is treated as an $(N\times N)$-dimensional vector. The basis vectors of the linear space, $e_{i,j}$ $(i,j=1,\ldots N)$, have components

$$(e_{ij})_{lm}=\delta_{il}\delta_{jm}. \tag{2}$$

The multiplication of the basis vectors is defined as the multiplication of the basis matrices, so we have

$$e_{ij}e_{kl}=\delta_{jk}e_{il}. \tag{3}$$

Generally, an $N\times N$ matrix can be written as

$$A=\begin{pmatrix} a_{11},\ldots,a_{1N} \\ \cdots \\ a_{N1},\ldots,a_{NN} \end{pmatrix}=\sum_{i,j=1}^{N} a_{ij}e_{ij}. \tag{4}$$

Let $\mathcal{N}=\{1,2,\ldots,N\}$ be a set of N integers, and I be subsets of $\mathcal{N}$, $I=\{i_1,i_2,\ldots,i_r\}$, where $i_1,i_2,\ldots,i_r$ are different numbers between 1 and N.

Definition 1: The dimension of the subset $I=\{i_1, i_2,\ldots,i_r\}$ is defined as $n(I)=r$.

If $I,J\subset\mathcal{N}$, the submatrix of A is

$$A_{IJ}=\sum_{i\in I,\,j\in J} a_{ij}e_{ij}. \tag{5}$$

Definition 2:

$$A_{\varnothing I}=A_{J\varnothing}=0, \tag{6}$$

where $\varnothing$ is the empty set.

Definition 3: The right inverse matrix of A_{IJ}, if it exists, is written as A_{JI}^{R-1}, which satisfies

$$A_{IJ}A_{JI}^{R-1}=\sum_{i\in I}e_{ii}. \tag{7}$$

The left matrix of A_{IJ}, if it exists, is written as A_{JI}^{L-1}, which satisfies

$$A_{JI}^{L-1}A_{IJ}=\sum_{j\in J}e_{jj}. \tag{8}$$

It is not difficult to prove that the following properties hold for the submatrices.

(1) An arbitrary matrix can always be written as the sum of the submatrices

$$A=\sum A_{I_pJ_q},$$

where

$$\begin{aligned} &\cup I_p=\mathcal{N}, \quad \forall p,p', \quad I_p\cap I_{p'}=\varnothing, \\ &\cup J_q=\mathcal{N}, \quad \forall q,q', \quad J_q\cap J_{q'}=\varnothing. \end{aligned} \tag{9}$$

(2) The product of the submatrices is given by

$$A_{IJ}B_{LK}=A_{IJ\cap L}B_{J\cap LK}=\delta_{J\cap L}C_{IK},$$

$$\delta_{J\cap L}=\begin{cases} 1, & J\cap L\neq\varnothing, \\ 0, & J\cap L=\varnothing. \end{cases} \tag{10}$$

(3) The inverse of the submatrix A_{IJ}, supposing the order of the submatrix is $\min\{n(I),n(J)\}$, satisfies the following: if $n(I)<n(J)$, there is no left inverse, and the right inverse is not unique; if $n(I)>n(J)$, there is no right inverse, and the left inverse is not unique; if $n(I)=n(J)$, $I\neq J$, there

 0022-2488/87/102329-04$02.50 © 1987 American Institute of Physics

exist a unique left inverse and right inverse, which are not equal to each other; and if $I = J$, there exists a unique and equal left and right inverse.

(4)

$$A_{II}^{-1} = A^{-1}_{II} - A^{-1}_{IIc} A^{-1}_{IcIc}{}^{-1} A^{-1}_{IcI}. \tag{11}$$

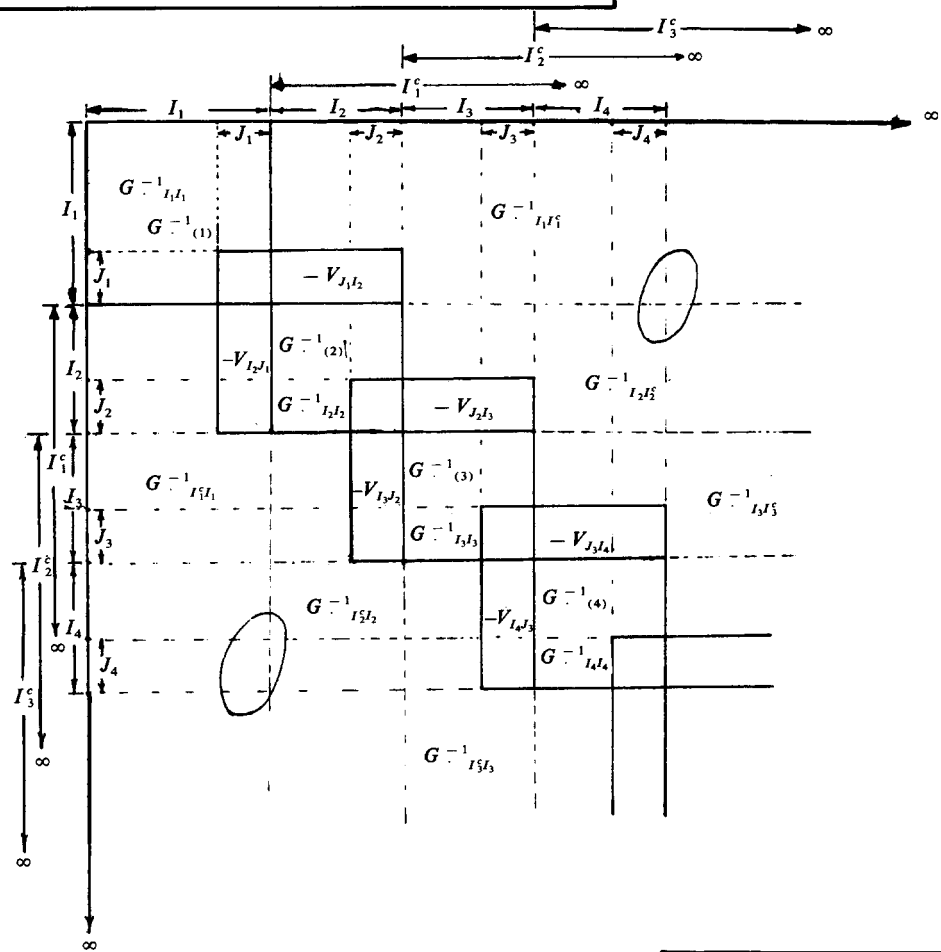
The proof of Eq. (11) is given in Ref. 1, where it is supposed that determinant of the matrix A is not zero. The inverse of the matrix A is A^{-1}, the inverse of the submatrix of A is A_{II}^{-1} and $A^{-1}_{IIc}, A^{-1}_{IcI}, A^{-1}_{IcIc}$ all are submatrices of the inverse of A. The only condition for Eq. (11) to hold is that $\det A^{-1}_{IcIc} \neq 0$, where $I \cup I^c = \mathcal{N}$, $I^c \cap I = \emptyset$, i.e., I^c is called the supplemental set of I, $\mathcal{N} = \{1,2,3,...\}$ is the set of all natural numbers.

III. LOCAL GREEN'S FUNCTIONS AND CHAIN EQUATION

The Green's function satisfies $G^{-1} = E - \hat{H} = E - \hat{T} - V$. In the representation where the kinetic energy is diagonal, the submatrix of the matrix T is

$$T(K) = \begin{pmatrix} T(K)_{11} & & 0 \\ & \ddots & \\ 0 & & T(K)_{NN} \end{pmatrix}, \tag{12}$$

which is the same as $V(k)$. The overlaps of $T(k)$ with $T(k+1)$ and $T(k)$ with $T(k-1)$ are the same even though the indices are not the same, and the bigger the value of k, the smaller the difference between $T(k)$. For $k_1, k_2 \geqslant M$ (an integer), $T(k_1) = T(k_2)$. So G^{-1} also has the chain form shown below:



(13)

Here all the areas are zeros except the area with shadow. The local Green's function can be written as

$$G_{I_1I_1}^{-1} = G^{-1}_{I_1I_1} - G^{-1}_{I_1I_1^c} G^{-1}_{I_1^cI_1^c}{}^{-1} G^{-1}_{I_1^cI_1},$$

where

$$G^{-1}_{I_1I_1^c} = -V_{J_1I_2}, \quad G^{-1}_{I_1^cI_1} = -V_{I_2J_1},$$

so we obtain

$$G_{I_1I_1}^{-1} = G^{-1}_{I_1I_1} - V_{J_1I_2} G^{-1}_{I_1^cI_1^c}{}^{-1} V_{I_2J_1}. \tag{14}$$

From Eq. (14) we see that if we substitute $G_{I_1I_1}^{-1}$ for $G^{-1}_{I_1I_1}$, the only modification is the lower right corner of $G^{-1}_{I_1I_1}$, the part in the J_1 row and J_1 column. From the property of the product of the submatrices, i.e., Eq. (10), for $G^{-1}_{I_1^cI_1^c}{}^{-1}$ we only need its submatrices $G^{-1}_{I_1^cI_1^c}{}^{-1}_{I_2I_2}$, because

$$V_{J_1I_2} G^{-1}_{I_1^cI_1^c}{}^{-1} V_{I_2J_1} = V_{J_1I_2} G^{-1}_{I_1^cI_1^c}{}^{-1}_{I_2I_2} V_{I_2J_1}.$$

The inverse of $G^{-1}_{I_1^cI_1^c}{}^{-1}_{I_2I_2}$ can be obtained by using Eq. (11).

Let

$$I_2^c = I_1^c - I_2 = \mathcal{N} - I_1 - I_2, \quad A = G^{-1}_{I_1^cI_1^c}{}^{-1}, \tag{15}$$

$$G^{-1}_{I_1^cI_1^c}{}^{-1}_{I_2I_2} = G^{-1}_{I_2I_2} - G^{-1}_{I_2I_2^c} G^{-1}_{I_2^cI_2^c}{}^{-1} G^{-1}_{I_2^cI_2},$$

$$G^{-1}_{I_2I_2^c} = -V_{J_2I_3}, \quad G^{-1}_{I_2^cI_2} = -V_{I_3J_2}, \tag{16}$$

$$G^{-1}_{I_1^cI_1^c}{}^{-1}_{I_2I_2} = G^{-1}_{I_2I_2} - V_{J_2I_3} G^{-1}_{I_2^cI_2^c}{}^{-1} V_{I_3J_2}.$$

For the same reason if we substitute $G^{-1}_{I_1^c I_1^c}{}^{-1}_{I_2 I_2}{}^{-1}$ for $G^{-1}_{I_2 I_2}$, we need only to modify the J_2 column and the J_2 row, and for $G^{-1}_{I_2^c I_2^c}{}^{-1}$, we only need its submatrix $G^{-1}_{I_2^c I_2^c}{}^{-1}_{I_3 I_3}$. Similarly its inverse can be obtained by using Eq. (11).

Let $I_3^c = I_2^c - I_3$, $A = G^{-1}_{I_2^c I_2^c}{}^{-1}$.

We can obtain

$$G^{-1}_{I_2^c I_2^c}{}^{-1}_{I_3 I_3}{}^{-1} = G^{-1}_{I_3 I_3} - V_{J_3 I_4} G^{-1}_{I_3^c I_3^c}{}^{-1} V_{I_4 J_3} \tag{17}$$

and we can continue with this procedure.

We have

$$G^{-1}_{I_k^c I_k^c}{}^{-1}_{I_{k+1} I_{k+1}}{}^{-1} = G^{-1}_{I_{k+1} I_{k+1}} - V_{J_{k+1} I_{k+2}} G^{-1}_{I_{k+1}^c I_{k+1}^c}{}^{-1} V_{I_{k+2} J_{k+1}},$$

where

$$I_k^c = I_{k-1}^c - I_k = \mathcal{N} - \sum_{L=1}^{k} I_L,$$
$$I_{k+1}^c = I_k^c - I_{k+1}. \tag{18}$$

Since $\mathcal{N}$ is an infinite set, $n(I_k) = n(I_2)$, $n(J_k) = n(J_1)$ $(k = 2,3,...)$. From the assumption, when k is large enough, $G^{-1}_{I_k^c I_k^c}$, $G^{-1}_{I_k I_k}$, $V_{J_k I_{k+1}}$, and $V_{I_{k+1} J_k}$ are not related to k. If this is satisfied when $k \geqslant M$, then

$$G^{-1}_{I_M^c I_M^c}{}^{-1}_{I_{M+1} I_{M+1}} \to X_{II},$$
$$G^{-1}_{I_{M+1}^c I_{M+1}^c}{}^{-1}_{I_{M+2} I_{M+2}} \to X_{II},$$
$$G^{-1}_{I_{M+1} I_{M+1}} \to B_{II}, \tag{19}$$
$$V_{J_{M+1} I_{M+2}} \to V_{JI},$$
$$V_{I_{M+2} J_{M+1}} \to V_{IJ}.$$

Finally we can obtain the equation of X_{II}, i.e., the chain equation

$$X_{II}{}^{-1} = B_{II} - V_{JI} X_{II} V_{IJ}.$$

This can be treated as the natural cutting equation. It can also be written as

$$V_{JI} X_{II} V_{IJ} X_{II} - B_{II} X_{II} + I_{II} = 0. \tag{20}$$

We also have $n(I) = N - n(J)$, where $N = n(I_1)$, I_{II} is a unit matrix, and B_{II}, V_{JI} are known matrices. To solve the chain equation is actually to solve the set of $n(J) \times n(I)$ second-order equations and the set of $[n(I) - n(J)] \times n(I)$ first-order equations.

After we obtain X_{II}, we substitute it into the equation as follows:

$$G^{-1}_{I_{M-1}^c I_{M-1}^c}{}^{-1}_{I_M I_M}{}^{-1} = G^{-1}_{I_M I_M} - V_{J_M I_{M+1}} X_{II} V_{I_{M+1} J_M} = X_{II}{}^{-1},$$

from which we can obtain the inverse of

$$G^{-1}_{I_{M-2}^c I_{M-2}^c}{}^{-1}_{I_{M-1} I_{M-1}},$$
$$G^{-1}_{I_{M-2}^c I_{M-2}^c}{}^{-1}_{I_{M-1} I_{M-1}}{}^{-1} = G^{-1}_{I_{M-1} I_{M-1}} - V_{J_{M-1} I} X_{II} V_{I J_{M-1}}.$$

Then we obtain

$$G^{-1}_{I_{M-2}^c I_{M-2}^c}{}^{-1}_{I_{M-1} I_{M-1}} = (G^{-1}_{I_{M-1} I_{M-1}} - V_{J_{M-1} I} X_{II} V_{I J_{M-1}})^{-1}.$$

Continuing with this procedure, we have

$$G^{-1}_{I_{M-2}^c I_{M-2}^c}{}^{-1}_{I_{M-1} I_{M-1}} \Rightarrow G^{-1}_{I_{M-3}^c I_{M-3}^c}{}^{-1}_{I_{M-2} I_{M-2}} \Rightarrow \cdots \Rightarrow \cdots \Rightarrow G^{-1}_{I_2^c I_2^c}{}^{-1}_{I_3 I_3} \Rightarrow G^{-1}_{I_1^c I_1^c}{}^{-1}_{I_2 I_2}.$$

Substituting into Eq. (14), we can obtain the exact form of the local Green's function.

It is difficult to prove mathematically that

$$\{E_k, \det G^{-1}(E_k) = 0\} \cap \{E'_L, \det G^{-1}_{I_1^c I_1^c}(E'_L) = 0\} = \varnothing,$$

because we do not know the actual form of the elements of the matrix [even though after the I_2th columns and the rows of G^{-1} and of $G^{-1}_{I_1^c I_1^c}$ the matrix elements are exactly the same, i.e., $G^{-1}_{I_2^c I_2^c} = (G^{-1}_{I_1^c I_1^c})_{I_2^c I_2^c}$]. Roughly speaking, G^{-1} and $G^{-1}_{I_1^c I_1^c}$ may represent different surface costructions of the same crystal in physics, so they will not have the same eigenvalues. If this can be proved exactly, then from Ref. 1, we can know that the local Green's function reserved all the information of the original Green's function.

IV. CONCLUSION

In the general case, we can only obtain an approximate form of the local Green's function, and the chain potential we discussed in this paper is the only case of which we can find the exact solution. In addition to requiring the chain form of G^{-1}, we also require the following.

(a) The ring submatrix on the chain of G^{-1}

$$G^{-1}(k) = \begin{pmatrix} G^{-1}(k)_{11},...,G^{-1}(k)_{1N} \\ \cdots \\ G^{-1}(k)_{N1},...,G^{-1}(k)_{NN} \end{pmatrix}$$

has an overlap with its neighbors $G^{-1}(k-1)$ or $G^{-1}(k+1)$, which is less than one-fourth of the matrix elements.

(b) When k is large enough, all the $G^{-1}(k)$ are the same. If (a) is not satisfied, then Eq. (14) is not true, and all the later derivations are not true. If (b) is not satisfied, then we cannot obtain the natural cutting equation, and we cannot obtain the exact form of the local Green's function.

Suppose $\hat{H} = \hat{H}_0 + V$, and in the representation of the eigenstates of $\hat{H}_0$, V is a chain potential. Then if (a) is satisfied and (b) is not, but we can treat V as a perturbation term and can cut it artifically we can still obtain a good approximate form of the local Green's function by using a similar method.

ACKNOWLEDGMENTS

I would like to thank Professor R. E. Allen and Professor C. M. Ko for their encouragement and helpful suggestions. I am also grateful to Dr. D. R. Zhou, Dr. S. F. Ren,

and Mr. R. Yuan for their beneficial discussions and assistance.

This work was supported by the U.S. Office of Naval Research under Contract No. N00014-82-K-0447.

[1]S. Graffi and V. Grecchi, Lett. Nuovo Cimento **12**, 425 (1975).
[2]M. Znojil, J. Math. Phys. **25**, 2979 (1984).
[3]M. Znojil, Phys. Rev. D **24**, 903 (1981).

Local Green's function. III

Renchuan Wang
Department of Physics, Texas A&M University, College Station, Texas 77843 and Center for Astrophysics, USTC, Hefei, Anhui, China

(Received 25 September 1986; accepted for publication 20 May 1987)

In this paper a way of solving for the local Green's function by means of a projection operator method is presented. Both a procedure for attacking the problem and a general formula for calculation are obtained. It is also proved that the calculation can be done with a finite number of basis functions. A short discussion of the accuracy is also presented.

I. INTRODUCTION

Under certain conditions, the local Green's function can reproduce the results of the complete Green's function, therefore, to solve the local Green's function precisely has long been a goal. However, it is very hard to do this except in some very special cases (Ref. 1). The most used method is the "cutting approximation" (Ref. 2). This method neglects the higher-order terms in V_{nm} $(m,n>N)$. It has the advantage of simplicity, but it is not quite accurate. It also cannot reproduce the results of the complete Green's function. The method we use here has certain advantages. It avoids the summation of an infinite series, and the results can be obtained by an analytical method. Another difference from the cutting approximation is that in the approximation method, the number of the eigenvalues equals the number of the states participated the calculation, but we do not have such a restriction in our method. In principle, we can obtain all the eigenvalues if we continue to do the calculation. This is also very tempting.

II. SPACE RESOLVE

Definition 1: Define R to be a Hilbert space spanned by a complete orthonormal state function set $|1\rangle,|2\rangle,\ldots|N\rangle,\ldots$. Set

$$\langle m|n\rangle=\delta_{m,n},\quad \sum_{n=1}^{\infty}|n\rangle\langle n|=I. \tag{1}$$

Definition 2: Define R^{P_0} to the N-dimensional subspace spanned by the basis $\{|0,k\rangle,|0,k\rangle=|k\rangle,\ k=1,\ldots,N\}$.

We have projection operators

$$\begin{aligned}&P_0=\sum_{n=1}^{N}|n\rangle\langle n|=\sum_{n=1}^{N}|0,n\rangle\langle 0,n|,\\&\Theta_0=\sum_{n=N+1}^{\alpha}|n\rangle\langle n|,\quad P_0+\Theta_0=1.\end{aligned} \tag{2}$$

According to the properties of the projection operators, R^{P_0} can be rewritten as

$$R^{P_0}=P_0R,\quad R^{\Theta_0}=\Theta_0R. \tag{3}$$

It is easy to verify that these two subspaces are orthogonal and

$$R^{\Theta_0}+R^{P_0}=R. \tag{4}$$

Definition 3: Define R^{P_1} to be the N-dimensional subspace of R spanned by basis $\{|1,K\rangle;|1,K\rangle=\Theta_0\hat{H}|K\rangle,\ K=1,2,\ldots,N\}$, where H is the Hamiltonian representing a certain physical system.

It is obvious that R^{P_1} is also a subspace of R^{Θ_0} because

$$\forall|1,K\rangle\in R^{P_1},\quad \hat{H}|0,K\rangle\in R,$$
$$\Theta_0\hat{H}|0,K\rangle=|1,K\rangle\in\Theta_0R=R^{\Theta_0},\quad R^{P_1}\subset R^{\Theta_1}.$$

It is very easy to find a reciprocal vector in R^{P_1} and find the projection operator from R to R^{P_1} through the reciprocal vector.

Definition 4: Define the reciprocal basis in R^{P_1} to be

$$|1,\bar{K}\rangle=\sum_{J=1}^{N}(\langle 1,m|1,n\rangle)^{-1}_{JK}|1,J\rangle.$$

The dual reciprocal basis is

$$\langle 1,\bar{K}|=\sum_{J=1}^{N}(\langle 1,m|1,n\rangle)^{-1}_{KJ}\langle 1,J|. \tag{5}$$

According to Definition 4, $|1,\bar{K}\rangle\in R^{P_1}$ and $\langle 1,\bar{K}|$ belongs to the dual space of R^{P_1}. Evidently we have

$$\langle 1,L|1,\bar{K}\rangle=\delta_{LK},\quad \langle 1,\bar{K}|1,L\rangle=\delta_{KL}. \tag{6}$$

The projection operator in R^{P_1} is

$$P_1=\sum_{K=1}^{N}|1,K\rangle\langle 1,\bar{K}|=\sum_{K=1}^{N}|1,\bar{K}\rangle\langle 1,K|. \tag{7}$$

It is easy to verify the properties of the projection operator P_1:

$$\begin{aligned}&P_1P_1=P_1,\quad P_1\Theta_0=\Theta_0P_1=P_1,\quad P_1P_0=P_0P_1=0,\\&\forall|1,\alpha\rangle\in R^{P_1},\quad P_1|1,\alpha\rangle=|1,\alpha\rangle.\end{aligned} \tag{8}$$

Then R^{P_1} can also be written as $R^{P_1}=P_1R$.

Let $\Theta_1=\Theta_0-P_1=1-P_0-P_1$. One can verify that Θ_1 is also a projection operator, and it satisfies the following relationships:

$$\begin{aligned}&\Theta_1\Theta_1=\Theta_1,\quad \Theta_1\Theta_0=\Theta_0\Theta_1=\Theta_1,\\&\Theta_1P_1=P_1\Theta_1=0.\end{aligned} \tag{9}$$

Here $\Theta_1R=R^{\Theta_1}\subset R^{\Theta_0}$ and R^{Θ_1} is orthogonal with both R^{P_1} and R^{P_0},

$$R=R^{P_0}+R^{P_1}+R^{\Theta_1}. \tag{10}$$

In the same way we define the N-dimensional subspace R^{P_2} in R^{Θ_1}, and it is spanned by the basis

$$\{|2,K\rangle;\ |2,K\rangle=\Theta_1\hat{H}|1,K\rangle,\ K=1,2,\ldots,N\},$$

and we can define the reciprocal vector in the same manner,

$$|2,\bar{K}\rangle=\sum_{J=1}^{N}(\langle 2,m|2,n\rangle)^{-1}_{JK}|2,J\rangle$$

and its dual form

 0022-2488/87/102333-05$02.50 © 1987 American Institute of Physics

$$\langle 2,\bar{K}| = \sum_{J=1}^{N} (\langle 2,m|2,n\rangle)^{-1}_{KJ} \langle 2,J|. \tag{11}$$

The projection operator from R to R^{P_2} is

$$P_2 = \sum_{k=1}^{N} |2,K\rangle\langle 2,\bar{K}| = \sum_{k=1}^{N} |2,\bar{K}\rangle\langle 2,K| \tag{12}$$

satisfying

$$\begin{aligned} &P_2P_2 = P_2, \quad P_2\Theta_1 = \Theta_1P_2 = P_2, \\ &P_2P_1 = P_1P_2 = 0, \quad P_2P_0 = P_0P_2 = 0. \end{aligned} \tag{13}$$

Here R^{P_2} can also be written as $R^{P_2} = P_2R$. We now again define $\Theta_2 = \Theta_1 - P_2 = 1 - P_0 - P_1 - P_2$ and it is easy to verify that

$$\Theta_2\Theta_2 = \Theta_2, \quad \Theta_2P_2 = P_2\Theta_2 = 0,$$
$$\Theta_2\Theta_1 = \Theta_1\Theta_2 = \Theta_2.$$

Also Θ_2 is the projection operator

$$\Theta_2R = R^{\Theta_2} \subset R^{\Theta_1} \subset R^{\Theta_0}.$$

One can see from the preceding that we can define all these in a more general way.

Definition 5: R^{P_m} is the N-dimensional subspace constructed by the basis

$$\{|m,K\rangle;\ |m,K\rangle = \Theta_{m-1}\hat{H}|m-1,K\rangle,\ K = 1,2,...,N\},$$

with projection operator

$$P_m = \sum_{K=1}^{N} |m,K\rangle\langle m,\bar{K}| = \sum_{K=1}^{N} |m,\bar{K}\rangle\langle m,K|, \tag{14}$$

reciprocal vector

$$|m,\bar{K}\rangle = \sum_{J=1}^{N} (\langle m,L|m,S\rangle)^{-1}_{JK} |m,J\rangle,$$

dual reciprocal vector

$$\langle m,\bar{K}| = \sum_{J=1}^{N} (\langle m,L|m,S\rangle)^{-1}_{KJ} \langle m,J|, \tag{15}$$

and

$$\Theta_m = \Theta_{m-1} - P_m.$$

From Definition 5, if Θ_{m-1} is a projection operator, then both P_m and Θ_m will satisfy the properties of a projection operator, and they are orthogonal with each other $P_m\Theta_m = \Theta_mP_m = 0$. Since we took Θ_0 to be a projection operator. Using Definition (5), we can get the bases

	$m=1$	$m=2$	$m=3$	$\cdots$
$\{\|0,K\rangle\}$	$\{\|1,K\rangle\}$	$\{\|2,K\rangle\}$	$\{\|3,K\rangle\}$	$\cdots$

and the projection operators

P_0	P_1	P_2	P_3	$\cdots$
Θ_0	Θ_1	Θ_2	Θ_3	$\cdots$

with resolving subspaces

$$R^{P_0} = P_0R, \quad R^{P_1} = P_1R, \quad R^{P_2} = P_2R, \ldots . \tag{16}$$

Theorem 1: The resolved subspaces R^{P_0}, R^{P_1}, R^{P_2},... are orthogonal with each other and R can be expressed as the direct summation of these subspaces:

$$R = R^{P_0} + R^{P_1} + R^{P_2} + R^{P_3} + \cdots . \tag{17}$$

Proof: Using the inductive method, R can be expressed following the direct summation of these subspaces that are orthogonal with each other,

$$R = R^{P_0} + R^{P_1} + R^{P_2} + \cdots + R^{P_m} + R^{\Theta_m}. \tag{18}$$

If we prove this for any m, and then set $m \to \infty$, the theorem is true.

We know that for $m = 0,1$ the statement is true. Now we suppose for $m - 1$

$$\begin{aligned} R &= R^{P_0} + R^{P_1} + R^{P_2} + \cdots + R^{P_{m-1}} + R^{\Theta_{m-1}} \\ &= (P_0 + P_1 + P_2 + \cdots + P_{m-1} + \Theta_{m-1})R \end{aligned} \tag{19}$$

is true, where $P_0, P_1, P_2, \ldots, P_{m-1}, \Theta_{m-1}$ are projection operators independent of each other. So,

$$\begin{aligned} &P_iP_j = P_jP_i = \delta_{ij}P_j, \\ &P_j\Theta_{m-1} = \Theta_{m-1}P_j = 0, \quad 0 \leqslant i,j \leqslant m-1, \\ &\Theta_{m-1}\Theta_{m-1} = \Theta_{m-1} = I - \sum_{j=0}^{m-1} P_j. \end{aligned} \tag{20}$$

From Definition 5,

$$P_m = \sum_{k=1}^{N} |m,K\rangle\langle m,\bar{K}| = \sum_{K=1}^{N} |m,\bar{K}\rangle\langle m,K|,$$

using the definition of the reciprocal vectors, we have

$$\langle m,\bar{K}|m,L\rangle = \delta_{KL}, \quad \langle m,L|m,\bar{K}\rangle = \delta_{LK}.$$

Obviously P_m is a projection operator,

$$P_mP_m = P_m$$

and

$$\Theta_{m-1}P_m = \sum_{K=1}^{N} \Theta_{m-1}|m,K\rangle\langle m,\bar{K}| = P_m. \tag{21a}$$

For the same reason

$$P_m\Theta_{m-1} = P_m. \tag{21b}$$

Also from Definition 5,

$$\Theta_m = \Theta_{m-1} - P_m,$$

so Θ_m is also a projection operator,

$$\Theta_m\Theta_m = \Theta_m, \tag{22a}$$

$$\Theta_mP_m = P_m\Theta_m = 0, \tag{22b}$$

and

$$\Theta_m\Theta_{m-1} = \Theta_{m-1}\Theta_m = \Theta_m. \tag{22c}$$

Now we prove that

$$P_jP_m = P_mP_j = \delta_{mj}P_m.$$

When $j = m$, it is obvious. When $j < m$ we have

$$P_jP_m = P_j\Theta_{m-1}P_m = 0,$$
$$P_mP_j = P_m\Theta_{m-1}P_j = 0.$$

Formulas (21a), (21b), and (20) have been used in the preceding proof. We can also prove that

$$P_j\Theta_m = \Theta_mP_j = 0.$$

When $j = m$, that is true, when $j < m$, using (22c) we get

$$P_j\Theta_m = P_j\Theta_{m-1}\Theta_m = 0,$$
$$\Theta_mP_j = \Theta_m\Theta_{m-1}P_j = 0.$$

Combining the preceding together, we have

$$P_iP_j = P_jP_i = \delta_{ij}P_j,$$

$$P_j\Theta_m = \Theta_m P_j = 0, \quad 0 \leqslant i,j \leqslant m,$$

$$\Theta_m\Theta_m = \Theta_m = I - \sum_{j=0}^{m} P_j.$$

So

$$P_0 + P_1 + \cdots + P_m + \Theta_m = I. \tag{23}$$

Therefore

$$R = R^{P_0} + R^{P_1} + R^{P_2} + \cdots + R^{P_m} + R^{\Theta_m}$$

is true.

III. APPLICATIONS OF THE FORMALISM

In the R space constructed by the complete orthonormal function set, the system's Hamiltonian can be expressed as

$$H = \sum_{m,n=1}^{N} H_{mn}|m\rangle\langle n|, \tag{24}$$

where $H_{mn} = \langle m|\hat{H}|n\rangle$. When m or $n > N$, $|H_{mn}|$ can be viewed as small compared to $|H_{11}|$.

The inverse of the system's Green's function can be written as

$$G^{-1} = E - \hat{H}$$

and the inverse of the local Green's function is

$$(PGP)^{-1} = PG^{-1}P - PG^{-1}\Theta(\Theta G^{-1}\Theta)^{-1}\Theta G^{-1}P, \tag{25}$$

where

$$P = \sum_{n=1}^{N} |n\rangle\langle n| = \sum_{n=1}^{N} |0,n\rangle\langle 0,n| = P_0,$$

$$\Theta = 1 - P = 1 - P_0 = \Theta_0,$$

and

$$\begin{aligned} PG^{-1}\Theta &= -P\hat{H}\Theta_0 \\ &= -\sum_{K=1}^{N} |0,K\rangle\langle 0,K|\hat{H}\Theta_0 \\ &= -\sum_{K=1}^{N} |0,K\rangle\langle 1,K| \\ &= -\sum_{K=1}^{N} |0,K\rangle\langle 1,K|P_1 = -P\hat{H}P_1. \end{aligned}$$

In the same way, we have

$$\Theta G^{-1}P = -P_1\hat{H}P. \tag{26}$$

Formula (25) can be rewritten as

$$(PG^{-1}P)^{-1} = PG^{-1}P - P\hat{H}P_1(\Theta G^{-1}\Theta)^{-1}P_1\hat{H}P, \tag{27}$$

where $P_1(\Theta G^{-1}\Theta)^{-1}P_1$ can be obtained from its inverse operator

$$[P_1(\Theta G^{-1}\Theta)^{-1}P_1]^{-1} = P_1G^{-1}P_1 - P_1G^{-1}\Theta_1(\Theta_1G^{-1}\Theta_1)^{-1}\Theta_1G^{-1}P_1.$$

Again we have

$$P_1G^{-1}\Theta_1 = -P_1\hat{H}P_2, \quad \Theta_1G^{-1}P_1 = -P_2\hat{H}P_1.$$

The preceding formula can be expressed as

$$[P_1(\Theta G^{-1}\Theta)^{-1}P_1]^{-1} = P_1G^{-1}P_1 - P_1\hat{H}P_2(\Theta_1G^{-1}\Theta_1)^{-1}P_2\hat{H}P_1. \tag{28}$$

Again we have

$$[P_2(\Theta_1G^{-1}\Theta_1)^{-1}P_2]^{-1} = P_2G^{-1}P_2 - P_2\hat{H}P_3(\Theta_2G^{-1}\Theta_2)^{-1}P_3\hat{H}P_2, \tag{29}$$

$$\vdots$$

$$[P_k(\Theta_{k-1}G^{-1}\Theta_{k-1})^{-1}P_k]^{-1} = P_kG^{-1}P_k - P_k\hat{H}P_{k+1}(\Theta_kG^{-1}\Theta_k)^{-1}P_{k+1}\hat{H}P_k, \tag{30}$$

$$\vdots.$$

If we regard $P\hat{H}P_1$, $P_1\hat{H}P$, $P_1\hat{H}P_2$, $P_2\hat{H}P_1$,... as of the same order, then the following is true.

The biggest contribution of $P_1G^{-1}P_1$ to $(PGP)^{-1}$ is second order.

The biggest contribution of $\Theta_1G^{-1}\Theta_1$ to $(PGP)^{-1}$ is fourth order.

The biggest contribution of $P_2G^{-1}P_2$ to $(PGP)^{-1}$ is fourth order.

The biggest contribution of $\Theta_2G^{-1}\Theta_2$ to $(PGP)^{-1}$ is sixth order.

$$\vdots$$

The biggest contribution of $P_kG^{-1}P_k$ to $(PGP)^{-1}$ is $(2k)$th order.

The biggest contribution of $\Theta_kG^{-1}\Theta_k$ to $(PGP)^{-1}$ is $(2k+2)$th order.

If we terminate our accuracy of calculation at $(2k)$th order, then

$$\begin{aligned} &[P_k(\Theta_{k-1}G^{-1}\Theta_{k-1})^{-1}P_k]^{-1} \cong P_kG^{-1}P_k, \\ &[P_{k-1}(\Theta_{k-2}G^{-1}\Theta_{k-2})^{-1}P_{k-1}]^{-1} \\ &\quad \cong P_{k-1}G^{-1}P_{k-1} - P_{k-1}\hat{H}P_k[P_kG^{-1}P_k]^{-1}P_kHP_{k-1}, \\ &[P_{k-2}(\Theta_{k-3}G^{-1}\Theta_{k-3})^{-1}P_{k-2}]^{-1} \\ &\quad \cong P_{k-2}G^{-1}P_{k-2} - P_{k-2}\hat{H}P_{k-1}[P_{k-1}G^{-1}P_{k-1} \\ &\qquad - P_{k-1}\hat{H}P_k(P_kG^{-1}P_k)^{-1}P_kHP_{k-1}]^{-1} \\ &\qquad \times P_{k-1}\hat{H}P_{k-2}, \end{aligned} \tag{31}$$

$$\vdots.$$

All the calculations are $N \times N$ matrix calculations: addition, subtraction, multiplication, and inversion. In principle we continue to perform these operations until we get the exact expression of $(PGP)^{-1}$.

Theorem 2: In the expression of $(PGP)^{-1}$, all the calculations can be done in the space $R^{P_0} = PR$.

Proof: The construction factors in the expression of $(PGP)^{-1}$ are P_k, $P_k\hat{H}P_{k+1}$, $P_{k+1}\hat{H}P_k$, $P_k\hat{H}P_k$,... $(k = 0,1,2,3,...)$.

If the calculation of P_k can be done in R^P, then it is obvious that the other calculations can be done in R^{P_0}. Now we use inductive method to prove this.

Suppose the calculation of P_m $(m \leqslant k)$ is only related to the states in space R^{P_0}, then the calculation of the P_{k+1}'s is only related to them, too. Obviously this is true for $k=0$.

The calculation of $P_0, P_1, \ldots, P_k$ is only related to the states in space R^{P_0}, which means that the calculation of the space basis of $R^{P_1}, R^{P_2}, \ldots, R^{P_k}$ is only related to the basis of R^{P_0}, also, and the calculation of $P_L \hat{H} |I,J\rangle$ $(I \leqslant K,\ L \leqslant I)$ is only related to the states in R^{P_0},

$$|K+1,J\rangle = \Theta_k \hat{H} |K,J\rangle = \hat{H} |K,J\rangle - \sum_{L=0}^{K} P_L \hat{H} |K,J\rangle. \tag{32}$$

It is natural that the calculation of the $|K+1,J\rangle$'s is only related to the states in R^{P_0}. According to the definition

$$P_{k+1} = \sum_{J=1}^{N} |K+1,J\rangle\langle K+1,\bar{J}| = \sum_{J,L=1}^{N} (\langle K+1,m|K+1,n\rangle)^{-1}_{JL} |K+1,J\rangle \times \langle K+1,L|.$$

Obviously the calculation of P_{k+1} can be done in R^{P_0}.

If the initial states $|0,J\rangle$ are not orthonormal, then the expression of the P_0 projection operator will be

$$P_0 = \sum_{J=1}^{N} |0,J\rangle\langle 0,\bar{J}| = \sum_{J,K=1}^{N} (\langle 0,m|0,n\rangle)^{-1}_{kJ} |0,K\rangle\langle 0,J|. \tag{33}$$

According to Theorem 2, it is not required that we know all the initial states, but it is required that $\langle 0,J|\hat{H}^n|0,K\rangle$ $(n=0,1,2,\ldots)$ be finite. Otherwise the $(PGP)^{-1}$ factor will diverge (which has no physical meaning), so we can get the following deduction.

Deduction: Given N independent state functions $\{|0,J\rangle,\ J=1,2,\ldots,N\}$, if only the $\langle 0,K|\hat{H}^n|0,J\rangle$ $(n=0,1,\ldots)$ are all finite, then, in principle, we can get the matrix expression of the local Green's function $(PGP)^{-1}$.

IV. DETAILED CALCULATION

The second-order accuracy local Green's function is

$$(PGP)^{-1} = PG^{-1}P - P\hat{H}P_1(P_1G^{-1}P_1)^{-1}P_1\hat{H}P. \tag{34}$$

From the definitions (2) and (7) we have

$$P\hat{H}P_1 = \sum_{J=1}^{N} |0,J\rangle\langle 1,J|,$$

$$P_1\hat{H}P = \sum_{J=1}^{N} |1,J\rangle\langle 0,J|, \tag{35}$$

$$(P_1G^{-1}P_1)^{-1} = \left\{\sum_{K,J}^{N} G^{-1}(1)_{KJ} |1,\bar{K}\rangle\langle 1,\bar{J}|\right\}^{-1} = \sum_{K,J}^{N} G(1)_{KJ} |1,K\rangle\langle 1,J|,$$

where

$$G^{-1}(1)_{KJ} = \langle 1,K|G^{-1}|1,J\rangle = EB(1)_{KJ} - H(1)_{KJ},$$

$$B(1)_{KJ} = \langle 1,K|1,J\rangle = H^2_{KJ} - \sum_{L=1}^{N} H_{KL}H_{LJ},$$

$$H(1)_{KJ} = \langle 1,K|H|1,J\rangle = H^3_{KJ} - \sum_{L=1}^{N} (H_{KL}H^2_{LJ} + H^2_{KL}H_{LJ}) + \sum_{I,L=1}^{N} H_{KI}H_{IL}H_{LJ}, \tag{36}$$

$$\sum_{L=1}^{N} G(1)_{KL}G^{-1}(1)_{LJ} = \delta_{KJ}.$$

We have used the following relationship in the proof of (36):

$$(P_1G^{-1}P_1)^{-1}(P_1G^{-1}P_1) = P_1.$$

This is the general expression of the local Green's function to second order. We have

$$(PGP)^{-1} = \sum_{K,J=1}^{N} \left\{E\delta_{KJ} - H_{KJ} - \sum_{L,I=1}^{N} B(1)_{KL}G(1)_{LI}B(1)_{IJ}\right\} |0,K\rangle\langle 0,J|. \tag{37}$$

If $\hat{H} = \hat{H}_0 + V$, and $|0,K\rangle$ is the eigenstate of $\hat{H}_0$ $(\hat{H}_0|0,K\rangle = \epsilon_k |0,K\rangle)$, then

$$B(1)_{KJ} = \left(V^2_{KJ} - \sum_{L=1}^{N} V_{KL}V_{LJ}\right),$$

$$H(1)_{KJ} = (V\hat{H}_0V)_{KJ} - \sum_{L=1}^{N} \epsilon_L V_{KL}V_{LJ} + V^3_{KJ} - \sum_{L=1}^{N} (V_{KL}V^2_{LJ} + V^2_{KL}V_{LJ}) + \sum_{L,I=1}^{N} V_{KL}V_{LI}V_{IJ}. \tag{38}$$

The local Green's function under the accuracy of fourth order is

$$(PGP)^{-1} = PG^{-1}P - P\hat{H}P_1(\Theta G^{-1}\Theta)^{-1}P_1\hat{H}P,$$
$$[P_1(\Theta G^{-1}\Theta)^{-1}P_1]^{-1} = P_1G^{-1}P_1 - P_1\hat{H}P_2(P_2G^{-1}P_2)^{-1}P_2\hat{H}P_1. \tag{39}$$

From the definitions (12) and (7) we have

$$P_2HP_1 = \sum_{J=1}^{N} |2,J\rangle\langle 1,\bar{J}|, \quad P_1HP_2 = \sum_{J=1}^{N} |1,\bar{J}\rangle\langle 2,J|,$$

$$(P_2G^{-1}P_2)^{-1} = \left(\sum_{K,J=1}^{N} G^{-1}(2)_{KJ} |2,\bar{K}\rangle\langle 2,\bar{J}|\right)^{-1} = \sum_{K,J=1}^{N} G(2)_{KJ} |2,K\rangle\langle 2,J|. \tag{40}$$

Using

$$(P_2G^{-1}P_2)^{-1}(P_2G^{-1}P_2) = P_2$$

we deduce that

$$\sum_{L=1}^{N} G(2)_{KL}G(2)^{-1}_{LJ} = \delta_{KJ},$$

where

$$G^{-1}(2)_{KJ} = \langle 2,K|G^{-1}|2,J\rangle = EB(2)_{KJ} - H(2)_{KJ},$$

$$B(2)_{KJ} = \langle 2,K|2,J\rangle$$
$$= H^2(1)_{KJ} - \sum_{L=1}^{N} B(1)_{KL}B(1)_{LJ}$$
$$- \sum_{L,I=1}^{N} H(1)_{KL}B^{-1}(1)_{LI}H(1)_{IJ},$$
$$H^2(1)_{KJ} = \langle 1,K|\hat{H}^2|1,J\rangle$$
$$= H^4_{KJ} - \sum_{L=1}^{N} (H^3_{KL}H_{LJ} + H_{KL}H^3_{LJ})$$
$$+ \sum_{L,I=1}^{N} H_{KL}H^2{}_{LI}H_{IJ},$$
$$H(2)_{KJ} = \langle 2,K|H|2,J\rangle$$
$$= \langle 0,K|\hat{H}(1-P)\hat{H}(1-P-P_1)$$
$$\times\hat{H}(1-P-P_1)\hat{H}(1-P)\hat{H}|0,J\rangle.$$

By following the previous method we can continue the calculation. Only at P_1 does the factor of $B^{-1}(1)$ appear. It is similar to the last term in the expression of $B(2)_{KJ}$,

$$[P_1(\Theta G^{-1}\Theta)^{-1}P_1]^{-1}$$
$$= \sum_{K,I=1}^{N} \Big\{EB(1)_{KJ} - H(1)_{KJ}$$
$$- \sum_{L,I=1}^{N} B(2)_{KL}G(2)_{LI}B(2)_{IJ}\Big\}|1,\bar{K}\rangle\langle 1,\bar{J}|$$
$$= \sum_{K,J=1}^{N} C^{-1}(2)_{KJ}|1,\bar{K}\rangle\langle 1,\bar{J}|, \quad (41)$$
$$C^{-1}(2)_{KJ} = EB(1)_{KJ} - H(1)_{KJ}$$
$$- \sum_{L,I=1}^{N} B(2)_{KL}G(2)_{LI}B(2)_{IJ}.$$

Then

$$P_1(\Theta G^{-1}\Theta)^{-1}P_1 = \sum_{k,J=1}^{N} C(2)_{KJ}|1,K\rangle\langle 1,J|, \quad (42)$$

where

$$\sum_{L=1}^{N} C(2)_{KL}C^{-1}(2)_{LJ} = \delta_{KJ}.$$

Finally we get the formula for the local Green's function under fourth order:

$$(PGP)^{-1} = \sum_{K,J=1}^{N} \Big\{E\delta_{KJ} - H_{KJ}$$
$$- \sum_{L,I=1}^{N} B(1)_{KL}C(2)_{LI}B_{IJ}(1)\Big\}|0,K\rangle\langle 0,J|. \quad (43)$$

V. CONCLUSION

The method presented above is able to give a solution for the local Green's function, though it is a bit complicated. In principle, besides the requirement of the finiteness of $\langle 0,K|\hat{H}^n|0,J\rangle$ $(K,J=1,...,N,\ n=0,1,...)$ for the initial states, there is no other restriction imposed. In actual calculations, we always want higher efficiency, so it is better, of course, if we can choose a set of initial states close to the eigenstates of the Hamiltonian.

ACKNOWLEDGMENTS

I thank Professor R. E. Allen and Professor C. M. Ko for their encouragements and helpful suggestions. I am also grateful to Dr. D. R. Zhou, Dr. S. F. Ren, and Mr. R. Yuan for their beneficial discussions and assistance.

This work was supported by the U. S. Office of Naval Research No. N00014-82-K-0447.

[1]M. Znojil, J. Math. Phys. **21**, 1629 (1980); **25**, 2979 (1984).
[2]R. C. Wang, J. Math. Phys. **28**, 2325, 2329 (1987).

The "spectral Wronskian" tool and the $\bar{\partial}$ investigation of the KdV hierarchy

M. Jaulent and M. Manna[a)]
Laboratoire de Physique Mathématique,[b)] U.S.T.L., 34060, Montpellier Cedex, France

(Received 18 December 1986; accepted for publication 29 April 1987)

The recently introduced "spatial transform" (ST) method for providing solutions to nonlinear evolution equations is developed when the basic $\bar{\partial}$ equation is the "spectral Schrödinger" equation (S). A fundamental tool is the "spectral Wronskian," which allows one to take advantage of the structure of a two-dimensional module for some set of solutions of (S). This leads easily to the KdV hierarchy. Contrary to the usual spectral transform (or inverse scattering transform) method there is no *a priori* assumption on the long distance behavior of the solutions. A recursion operator is exhibited. Local conservation laws and Bäcklund relations are also derived.

I. INTRODUCTION

In a previous paper[1] we introduced the "spatial transform" (ST) method for investigating nonlinear evolution equation (NE's). We can summarize the method as follows. Let $\mathscr{U}$ and $\mathscr{R}$ be sets of complex vector- (or matrix-) valued functions whose elements are denoted, respectively, by $U(x)$ and $R(k)$; x is the real "space" variable and k is the complex "spectral" variable. In the following, S refers to spatial and $\hat{S}$ to spectral and we sometimes point out that a letter occuring in an expression plays the role of a parameter by underlining it. A "$\bar{\partial}$" (DBAR) problem means any problem that consists of finding a function $F(k)$ in the complex plane from a relation called the "$\bar{\partial}$ equation" involving $F(k)$ and its "defect of holomorphy"

$$\frac{\partial F}{\partial \bar{k}} \doteq \frac{1}{2}\left(\frac{\partial F}{\partial k_1} + i\frac{\partial F}{\partial k_2}\right), \quad k = k_1 + ik_2, \quad (k_1,k_2)\in\mathbb{R}^2. \tag{1}$$

Other conditions such as "normalization at ∞" [i.e., information on the behavior of $F(K)$ as $|k| \to \infty$] must be added to this $\bar{\partial}$ equation in order to obtain the uniqueness of the solution. Note that a Riemann–Hilbert problem can be viewed formally as a particular $\bar{\partial}$ problem, where $F(k)$ is analytic everywhere in $\mathbb{C}$ except on a contour Γ where it has a jump.

Then, generally speaking, a ST is a map

$$\mathscr{S}\colon\ R(k)\in\mathscr{R} \to U(x)\in\mathscr{U}$$

defined through a $\mathbb{C}$-linear $\bar{\partial}$ equation $\hat{\mathrm{E}}\mathrm{qu}\{\underline{x},k,R(k)\}$ with a normalization at ∞ (N), where k is the variable, x is a parameter, and $R(k)$ $(k\in\mathbb{C})$, the "spectral potential," fixes the equation. The image of $R(k)$ by $\mathscr{S}$, $U(x)$, is defined from the asymptotic behavior as $|k| \to \infty$ of the solution $F(k,\underline{x})$ of the $\bar{\partial}$ problem $[\hat{\mathrm{E}}\mathrm{qu}\{\underline{x},k,R(k)\},(N)]$. Then $F(\underline{k},x)$ should satisfy a linear differential equation $\mathrm{Equ}\{x,\underline{k},U(x)\}$, where x is the variable, k is a parameter, and $U(x)$, the "spatial potential," fixes the equation. Then, introducing the "time" t, we assume a linear evolution $(\hat{L})$ for $R(k,t)$. (Note that as a general rule we only write the parameter t when we think it is necessary.) It is expected that $\mathscr{S}$ transforms $(\hat{L})$ into a NE in $U(x,t)$ so that we will then conclude that $\mathscr{S}$ provides solutions to this NE.

We also expect that for some subclass $\mathscr{R}_0$ of spectral potentials, $\mathscr{S}$ admits a bijective restriction

$$\mathscr{S}_0\colon\ R(k)\in\mathscr{R}_0\subset\mathscr{R} \mapsto U(x)\in\mathscr{U}_0\subset\mathscr{U},$$

which allows us not only to provide solutions to NE's but also to solve initial value problems according to

$$U(x,0) \overset{\mathscr{S}_0}{\mapsto} R(k,0) \overset{(\hat{L})}{\mapsto} R(k,t) \overset{\mathscr{S}_0}{\mapsto} U(x,t),$$

where $\hat{\mathscr{S}}_0$, the "spectral transform" $(\hat{\mathrm{ST}})$, is the inverse map of $\mathscr{S}_0$. In some way $\hat{\mathrm{E}}\mathrm{qu}\{\underline{x},k,R(k)\}$ plays the same role for the ST $\mathscr{S}$ that $\mathrm{Equ}\{\underline{x},\underline{k},U(x)\}$ plays for the $\hat{\mathrm{ST}}$ $\hat{\mathscr{S}}_0$. We shall refer to such a fact as an "$(\underline{x},k)$–$(\underline{x},\underline{k})$ analogy." This imitates what happens for the direct and inverse Fourier transform formulas.

The ST method was suggested in analogy with the well-known $\hat{\mathrm{ST}}$ (or inverse scattering transform) method [see the monographs (Refs. 2–4) and, in order to take into account the ever greater importance of the $\bar{\partial}$ problem in the field of NE's, see the review (Ref. 5)]. There is a connection (see Ref. 1) with the direct linearization method (Fokas and Ablowitz[6] and Ablowitz, Fokas, and Anderson[7]). Contrary to the $\hat{\mathrm{ST}}$ method there is no *a priori* assumption on the long distance behavior of the obtained solutions. Concerning the technique of the proofs to be used in the ST method, a "$\bar{\partial}$ analysis" has to be (constructed and) used instead of the (classical) "differential" analysis of the $\hat{\mathrm{ST}}$ method. Putting this aside, the general strategy of the ST method is in our opinion much easier than that of the $\hat{\mathrm{ST}}$ method. This should appear clearly in the following example.

In Ref. 1 we began the application of the ST method when $\hat{\mathrm{E}}\mathrm{qu}\{\underline{x},k,R(k)\}$, (N), and $(\hat{\mathrm{L}})$ are, respectively,

$$(\hat{\mathrm{S}})\colon\ \frac{\partial}{\partial\bar{k}}F(k,x) = R(k)F(-k,x), \tag{2}$$

$$(\mathrm{N})\colon\ F(k,x) = e^{-ikx}(1+O(1/k)) \quad \text{as } |k|\to\infty, \tag{3}$$

$$(\hat{\mathrm{L}})\colon\ \frac{\partial}{\partial t}R(k,t) = p(k)R(k,t), \tag{4}$$

where $x\in\mathbb{R}$, $k\in\mathbb{C}$, $F(k,x)$, $R(k)$, and $p(k)$ are complex val-

[a)] On leave of absence from Pontificia Universidade Católica de São Paulo, São Paulo, Brazil.

[b)] Unité Associée au Centre National de la Recherche Scientifique n° 040768, Recherche Coopérative sur programme n° 080264.

 0022-2488/87/102338-05$02.50 © 1987 American Institute of Physics

ued, and $p(k)$ is a given odd polynormial in k, which we write in the form

$$p(k) = -2ik\,\Omega(k^2) \quad \text{with } \Omega(K) = \sum_{n=0}^{N} \alpha_n K^n. \tag{5}$$

Recall that if a function of k, $\psi(k,x)$, can be written in the form

$$\psi(k,x) = e^{-ikx}[P(k,x) + O(1/k)] \quad \text{as } |k| \to \infty \quad (x \in \mathbb{R}), \tag{6}$$

where $P(k,x)$ is a polynomial in k with x-dependent coefficients, we say that $\psi(k,x)$ admits a "polynomial normalization" and we write $\text{Nor}\,\psi(k,x) = e^{-ikx}P(k,x)$. Using an assumption on $R(k)$ ($k \in \mathbb{C}$), which needs a further investigation and is equivalent to demanding that the $\bar{\partial}$ problem [$(\hat{S})$, (N)] admits a unique solution $F(k,x)$, we have proved in Ref. 1 that, for any P the $\bar{\partial}$ problem [$(\hat{S})$, "$\text{Nor}\,\psi(k,x) = e^{ikx}P(k,x)$"] has a unique solution $\psi(k,x)$ and that $\psi(k,x)$ admits an asymptotic expansion (AE)

$$\psi^{\infty}(k,x) = e^{-ikx} \sum_{n=-N}^{\infty} \frac{\psi_n(x)}{k^n} \quad (N \in \mathbb{Z}).$$

In particular $F(k,x)$ admits the AE

$$F^{\infty}(k,x) = e^{-ikx} \sum_{n=0}^{\infty} \frac{F_n(x)}{k^n}, \quad \text{with } F_0(x) = 1.$$

We suppose that $F(k,x)$ and $F_n(x)$ ($n \geqslant 0$) admit x-derivatives of any order.

Furthermore, $\mathcal{P}$ being the commutative ring of even polynomials in k and $\mathcal{F}$ being (for fixed x) the set of solutions of $(\hat{S})$ that admit a polynomial normalization, we have shown that $\mathcal{F}$ forms a $\mathcal{P}$-module of dimension 2 and of basis $\{F(k,x), (\partial/\partial x)F(k,x)\}$, i.e., for any $\psi(k,x) \in \mathcal{F}$ there exist $\mathcal{P}$-scalars $A(k,x)$ and $B(k,x)$ such that

$$\psi(k,x) = A(k,x)F(k,x) + B(k,x)\frac{\partial}{\partial x}F(k,x). \tag{7}$$

The motivation for introducing $\mathcal{F}$ is that "$(\partial/\partial t - \frac{1}{2}p(k)) \times F(k,x,t)$" belongs to $\mathcal{F}$, so that there exist $\mathcal{P}$-scalars $A(k,x,t)$ and $B(k,x,t)$ such that

$$(\text{L}): \frac{\partial}{\partial t}F(k,x,t) = \left(\frac{1}{2}p(k) + A(k,x,t)\right)F(k,x,t) + B(k,x,t)\frac{\partial}{\partial x}F(k,x,t). \tag{8}$$

Then, noticing that $\partial/\partial x$ acts $\mathcal{P}$-linearly in $\mathcal{F}$, we have concluded that it can be characterized by a 2×2 $\mathcal{P}$-matrix. This yields that $F(\underline{k},x)$ satisfies the (spatial) Schrödinger equation

$$(\text{S}): \left[\frac{\partial^2}{\partial x^2} + k^2 - U(x)\right]F(k,x) = 0, \quad x \in \mathbb{R}, \quad k \in \mathbb{C}, \tag{9}$$

with

$$U(x) \doteqdot -2iF_1'(x) = \frac{1}{\pi}\frac{d}{dx}\left(\iint_{\mathbb{R}^2} R(l)e^{ilx}F(-l,x)\,dl \wedge d\bar{l}\right), \tag{10}$$

where $l = l_1 + il_2$, $(l_1,l_2) \in \mathbb{R}^2$, and a prime means the derivative with respect to x. This result suggests we name $(\hat{S})$ "spectral Schrödinger" equation. Note the $(\underline{x},k)$–$(x,\underline{k})$ analogy: the set $\mathcal{F}$ of solutions of $(\hat{S})$ forms a $\mathcal{P}$-module of dimension 2; the set of solutions of (S) forms a $\mathbb{C}$-vector space of dimension 2.

It is well known that a "higher" KdV in $U(x,t)$, corresponding to the given odd polynomial $p(k)$, can then be obtained as the "compatibility condition" of the Schrödinger equation (S) and the "auxiliary Lax equation" (L). Note that no assumption on the behavior of $U(x)$ as $|x| \to \infty$ is needed (at least surely if we do not demand the existence of a "recursion" operator). In the logic of our approach another kind of derivation of the KdV hierarchy is needed using the AE's. We have outlined it in Ref. 1 and we have detailed the case $p(k) = 8ik^3$, which corresponds to KdV.

In this paper we complete our investigation of the ST method when the basic $\bar{\partial}$ equation is the spectral Schrödinger equation $(\hat{S})$. In Secs. III–V we prove that the ST method provides solutions to a hierarchy of NE's, that this hierarchy is purely differential, and that it is the KdV hierarchy. A recursion operator (defined in an appropriate space) is exhibited and local conservation laws are derived. In Sec. VI Bäcklund relations are investigated. Finally in Sec. VII we consider the case where the solutions go to zero for large distances, which corresponds to the range of application of the usual $\hat{S}$T method.

A fundamental tool in our $\bar{\partial}$ analysis is the "spectral Wronskian" we introduce in Sec. II. This develops further the already mentioned $(\underline{x},k)$–$(x,\underline{k})$ analogy. We recall the importance of "Wronskian relations" in the $\hat{S}$T method.[3] In some way we have introduced a "$\bar{\partial}$ differential version" of the "spectral integral relations" mentioned in Ref. 3.

II. THE SPECTRAL WRONKIAN TOOL; THE "BILINEAR" RELATION (B)

As $g(k)$ and $h(k)$ ($k \in \mathbb{C}$) are complex functions we define their "spectral Wronskian"

$$\hat{W}[g(k),h(k)] \doteqdot \begin{vmatrix} g(k) & h(k) \\ g(-k) & h(-k) \end{vmatrix} (2ik)^{-1} = [g(k)h(-k) - g(-k)h(k)](2ik)^{-1}. \tag{11}$$

This $\hat{W}$ is an alternating $\mathcal{P}$-bilinear form. We now prove a "spectral Wronskian property" for the solutions of $(\hat{S})$: if $G(k,\underline{x})$ and $H(k,\underline{x})$ belong to $\mathcal{F}$, then $\hat{W}[G(k,\underline{x}),H(k,\underline{x})]$ is a $\mathcal{P}$-scalar and

$$\hat{W}[G(k,\underline{x}),H(k,\underline{x})] = \hat{W}[G^{\infty}(k,\underline{x}),H^{\infty}(k,\underline{x})].$$

Note the $(\underline{x},k)$–$(x,\underline{k})$ analogy: if $g(\underline{k},x)$ and $h(\underline{k},x)$ are solutions of (S), then their (usual) Wronskian $W(g(\underline{k},x),h(\underline{k},x))$ is a $\mathbb{C}$-scalar and

$$W(g(\underline{k},x),h(\underline{k},x)) = W(g(\underline{k},\infty),h(\underline{k},\infty)).$$

Using $(\hat{S})$, it is easy to verify that

$$\frac{\partial}{\partial \bar{k}}\hat{W}[G(k,\underline{x}),H(k,\underline{x}] = 0$$

so that $\hat{W}[G(k,\underline{x}),H(k,\underline{x})]$ is an (even) entire function in k. The proof is completed by using the AE of $G(k,\underline{x})$ and $H(\underline{k},x)$ and the Liouville theorem.

A direct application of the spectral Wronskian property yields

$$(\mathrm{B}):\ \hat{W}\left[F(k,\underline{x}),\frac{\partial}{\partial x}F(k,\underline{x})\right]=1,$$

i.e.,

$$W[F(\underline{k},x),F(-\underline{k},x)]=2ik. \tag{12}$$

Since (S) implies $(\partial/\partial x)W[F(\underline{k},x),F(-\underline{k},x)]=0$, the bilinear and first-order (in the x-derivatives) formula (B) can be viewed as completing the linear and second-order formula (S). Note the $(\underline{x},k)$–$(x,\underline{k})$ analogy: $\{F(k,\underline{x}),(\partial/\partial x)F(k,\underline{x})\}$ is a basis of the $\mathscr{P}$-module $\mathscr{F}$; $\{F(\underline{k},x),F(-\underline{k},x)\}$ ($k\in\mathbb{C}^*$) is a basis of the $\mathbb{C}$-vector space of solutions of (S).

The spectral Wronskian tool allows the calculation of the $\mathscr{P}$-scalars $A(k,x)$ and $B(k,x)$ in Eq. (7):

$$\begin{aligned}B(k,x)&=\hat{W}[F(k,x),\psi(k,x)],\\ A(k,x)&=-\hat{W}\left[\frac{\partial}{\partial x}F(k,x),\psi(k,x)\right].\end{aligned} \tag{13}$$

III. THE FUNCTIONS $\varphi(k,x),\varphi^{\infty}(k,x),\varphi_n(x)$; DERIVATION OF THE NE [EQ. (17)]

As an application of formulas (13) we find that the $\mathscr{P}$-scalar $B(k,x,t)$ in Eq. (8) satisfies the relation

$$B(k,x,t)=\hat{W}\left[F(k,x,t),\frac{\partial}{\partial t}F(k,x,t)\right]+\frac{p(k)}{2ik}\varphi(k,x,t), \tag{14}$$

where $\varphi(k,x)\doteq F(k,x)F(-k,x)$. A similar relation can be written for $A(k,x,t)$. Since $\varphi(k,x)$ admits the AE

$$\varphi^{\infty}(k,x)=\sum_{n=0}^{\infty}\frac{\varphi_n(x)}{k^{2n}},$$

with

$$\varphi_n(x)=\sum_{m=0}^{2n}(-1)^m F_m(x)F_{2n-m}(x), \tag{15}$$

we find that $B(k,x,t)$ and $A(k,x,t)$ can be calculated with the formulas

$$\begin{aligned}B(k,x,t)&=\mathrm{Nor}\left[\frac{p(k)}{2ik}\varphi^{\infty}(k,x,t)\right],\\ A(k,x,t)&=-\frac{1}{2}\frac{\partial}{\partial x}B(k,x,t).\end{aligned} \tag{16}$$

The equalization of the terms in $1/k^2$ in Eq. (14) gives the equation

$$U_t=\left(-2\sum_{n=0}^{N}\alpha_n\varphi_{n+1}\right)_x \tag{17}$$

[recall the notation (5)], where now lower indices are used for indicating the derivatives with respect to x and t. Since the α_n's can be varied freely we have thus obtained a "hierarchy" of equations. Now $\varphi^{\infty}(k,x)$ can be viewed as the "generating function" for this hierarchy. It is not yet clear that (17) is a NE in $U(x,t)$. To show this we need to investigate the structure of the φ_n's. This will be done in Sec. IV. As a consequence we will obtain in Sec. V that (17) is in fact a "higher" KdV in $U(x,t)$ and that the operators L and L^+ (defined in appropriate spaces) exist that "generate" the hierarchy according to $\varphi_{n+1}=L\varphi_n$, $(\varphi_{n+1})_x=L^+(\varphi_n)_x$. We will also derive local conservation laws.

IV. PROPERTIES OF THE φ_n's; THE SPACES $\mathscr{E}$, $\mathscr{E}'$, $\mathscr{G}$

We call $\mathscr{E}$ the $\mathbb{C}$-vector space of functions $f:\mathbb{R}\to\mathbb{C}$ that are polynomials in the x-derivatives of U—U, U_x, U_{xx},...—without constant term (this last point is of importance). Here $\mathscr{E}'=(d/dx)(\mathscr{E})$ is the image of the space $\mathscr{E}$ by the map $d/dx:\mathscr{E}\to\mathscr{E}$. We also introduce $\mathscr{G}=\{f\in\mathscr{E},\ U_x f\in\mathscr{E}'\}$.

We will prove the following properties for the φ_n's:

$$(\mathrm{P}_1):\ \forall n\geqslant 1,\quad \varphi_n\in\mathscr{E};$$

$$\begin{aligned}(\mathrm{P}_2):\ (\varphi_{n+1})_x&=-\tfrac{1}{4}(\varphi_n)_{xxx}+U(\varphi_n)_x+\tfrac{1}{2}U_x\varphi_n\\ &=-\tfrac{1}{4}(\varphi_n)_{xxx}+(U\varphi_n)_x-\tfrac{1}{2}U_x\varphi_n\\ &\qquad(n\geqslant 0);\end{aligned} \tag{18}$$

$$(\mathrm{P}_3):\ \forall n\geqslant 0,\quad \forall m\geqslant 0,\quad g_{m,n}\doteq\varphi_m(\varphi_n)_x\in\mathscr{E}';$$

$$(\mathrm{P}_3'):\ \forall n\geqslant 1,\quad \varphi_n\in\mathscr{G}.$$

Proof of (P_1): We use (15) and the recurrence formula for the F_n's,

$$\begin{aligned}F'_{n+1}(x)&=-(i/2)F''_n(x)+(i/2)U(x)F_n(x),\\ F_0(x)&=1,\end{aligned} \tag{19}$$

obtained from (S) and the AE of $F(k,x)$. As a kind of substitute to the knowledge of constants for integrating (19) we use the formula

$$i\varphi_{n+1}(x)=\sum_{m=0}^{2n}(-1)^m F_m(x)F'_{2n+1-m}(x),\quad n\geqslant 0, \tag{20}$$

obtained from (B) [Eqs. (12)] and the AE of $F(k,x)$. [Note that in the standard case where $U(x)$ goes to zero as $|x|\to\infty$, Eq. (19) can be completed with the boundary condition $F_n(\infty)=0$.] Because of the structure of (19) we are led to an "inflationist" proof in the order of the derivatives of the F_n's, i.e., in order to prove that $\varphi_n\in\mathscr{E}$ $(n\geqslant 1)$ we prove that

$$\psi_{N,q,r}\doteq\sum_{m=0}^{N}(-1)^m F_m^{(q)}F_{N-m}^{(r)}\in\mathscr{E},$$

$$\text{for } N\geqslant 0,\quad q\geqslant 0,\quad r\geqslant 0,\quad (N,q,r)\neq(0,0,0).$$

This can be done by induction: introduce the recurrence assumption

$$(\mathrm{H}_N)\ (N\geqslant 0):\quad \forall q\geqslant 0,\quad \forall r\geqslant 0,\quad \psi_{N,q,r}\in\mathscr{E}\,((N,q,r)\neq(0,0,0));$$

then use the identities $\psi_{2n,0,0}=\varphi_n$ [Eq. (15)], $\psi_{2N+1,0,1}=i\varphi_{n+1}$ [Eq. (20)], and the recurrence formula

$$\psi_{N+1,q,r+1}=-\frac{i}{2}\psi_{N,q,r+2}+\sum_{k=0}^{r}C_r^k(F'_1)^{(k)}\psi_{N,q,r-k},\quad N\geqslant 0,\quad q\geqslant 0,\quad r\geqslant 0, \tag{21}$$

derived from (19) and the Leibnitz formula.

Proof of (P_2): Use the AE of $\varphi(k,x)$ in the following equation derived from (S):

$$\varphi_{xxx}+4(k^2-U)\varphi_x-2U_x\varphi=0,\quad x\in\mathbb{R},\quad k\in\mathbb{C}. \tag{22}$$

Proof of (P_3): Use $g_{N,0}\in\mathscr{E}'$ and $(g_{n,m+1}-g_{n+1,m})\in\mathscr{E}'$

$(n \geqslant 0,\ m \geqslant 0)$ which is a consequence of $\varphi_n \in \mathscr{E}$ $(n \geqslant 1)$ and (P_2). Property (P_3') follows from (P_3):

$$U_x \varphi_n = 2\varphi_n (\varphi_1)_x = 2g_{n,1} \in \mathscr{E}'.$$

V. CONSEQUENCES: EQUATION (17) IS A HIGHER KdV; THE OPERATORS $\int dx$, L, AND L^+; LOCAL CONSERVATION LAWS

Property (P_1) implies that Eq. (17) is a nonlinear partial differential equation in $U(x,t)$ and is therefore a higher KdV [the behavior of $U(x,t)$ as $|x| \to \infty$ does not play any role in the form of Eq. (17)].

Now we use (P_2) for expliciting the dependence of φ_{n+1} [resp. $(\varphi_{n+1})_x$] with respect to φ_n [resp. $(\varphi_n)_x$]. We have to find some canonical way of choosing primitives. [In the standard case where $U(x)$ goes to zero as $|x| \to \infty$ this is easily done by introducing $\int_{\pm\infty}^{x} dy$.] We remark that the operator d/dx: $\mathscr{E} \to \mathscr{E}'$ is a $\mathbb{C}$-linear isomorphism, so that we can define the inverse operator $\int dx \doteq (d/dx)^{-1}$: $\mathscr{E}' \to \mathscr{E}$ and the $\mathbb{C}$-linear operators L and L^+, L: $\mathscr{S} \to \mathscr{E}$, L^+: $\mathscr{E}' \to \mathscr{E}$:

$$L = -\frac{1}{4}\frac{d^2}{dx^2} + U - \frac{1}{2}\int dx\, U_x,$$
$$L^+ = -\frac{1}{4}\frac{d^2}{dx^2} + U + \frac{1}{2} U_x \int dx. \tag{23}$$

Here L and L^+ are connected by $L^+ \circ (d/dx) = (d/dx) \circ L$ (identity on $\mathscr{S}$). Using (P_2) and (P_3') it is easy to verify that

$$\varphi_{n+1} = L\varphi_n, \quad (\varphi_{n+1})_x = L^+ (\varphi_n)_x \quad (n \geqslant 1). \tag{24}$$

Therefore using $\varphi_1 = \frac{1}{2}U$ [set $n = 0$ in (20)] and recalling the notation (5) we find that the higher KdV (17) can be put into the two equivalent forms:

$$U_t + (\Omega(L)U)_x = 0, \quad U_t + \Omega(L^+)U_x = 0. \tag{25}$$

Here L^+ is the "recursion" operator. Note that the generating function φ^∞ satisfies the property $L\varphi^\infty = k^2(\varphi^\infty - 1)$, where we have used $(\varphi^\infty - 1) \in \mathscr{S}$ and we have set $L\,1 = \varphi_1$.

From (L) [Eq. (8)] we derive the following evolution law for $\varphi(k,x)$: $\varphi_t = B\varphi_x - B_x\varphi$, which yields $\varphi_t^\infty = B\varphi_x^\infty - B_x\varphi^\infty$. Inserting (16) we find that $(\varphi_n)_t$ is a linear combination of terms $g_{p,q}$. Hence, using (P_3): $(\varphi_n)_t \in \mathscr{E}'$ $(n \geqslant 0)$, i.e., there exists $\gamma_n \in \mathscr{E}$ such that $(\varphi_n)_t = (\gamma_n)_x$. We have obtained for any higher KdV an infinite set of local conservation laws. The densities—the φ_n's—are common to all equations of the hierarchy. This is not the case for the currents—the γ_n's.

VI. BÄCKLUND RELATIONS

Now we start from two spectral Schrödinger equations $(\hat{S}_1)$ and $(\hat{S}_2)$ with spectral potentials $R_1(k,t)$ and $R_2(k,t)$ connected by the formula

$$R_2(k,t) = R_1(k,t)G(k)/G(-k)$$

with

$$G(k) = g(k^2) + 2ikh(k^2),$$
$$g(k^2) = \sum_{n=0}^{N_1} \beta_n k^{2n},$$
$$h(k^2) = \sum_{n=0}^{N_2} \gamma_n k^{2n},$$

the β_n's and the γ_n's being time-independent coefficients. Here $R_1(k,t)$ and $R_2(k,t)$ satisfy the same time evolution $(\hat{L})$. The ST method provides spatial potentials $U_1(x,t)$ and $U_2(x,t)$, which are solutions of the same higher KdV.

Since $F_1(k,x)$ and $F_2(k,x)$ are the solutions of $(\hat{S}_1,1)$ and $(\hat{S}_2,1)$, it is easy to see that both $F_2(k,x)$ and $G(k)F_1(k,x)$ are solutions of $(\hat{S}_2)$. From Eq. (7) there exist $\mathscr{P}$-scalars $C(k,x)$ and $D(k,x)$ such that

$$G(k)F_1(k,x) = C(k,x)F_2(k,x) + D(k,x)\frac{\partial}{\partial x}F_2(k,x). \tag{26}$$

Then using the spectral Wronskian tool and a procedure similar to Sec. III one can find the implicit "Bäcklund" relation between $U_1(x,t)$ and $U_2(x,t)$:

$$i\sum_{n=0}^{N_1} \beta_n \tilde{\varphi}_{2n+1}(x,t) - 2\sum_{n=0}^{N_2} \gamma_n \tilde{\varphi}_{2n+2}(x,t) = 0, \tag{27}$$

where the $\tilde{\varphi}_n$'s occur in the AE $\tilde{\varphi}^\infty(k,x) = \Sigma_{n=0}^{\infty} \tilde{\varphi}_n(x)/k^n$ of $\tilde{\varphi}(k,x) \doteq F_1(k,x)F_2(-k,x)$.

We consider the particular case $g(k^2) = 2p$, $h(k^2) = 1$. Then Eq. (27) reduces to $\tilde{\varphi}_2 = ip\tilde{\varphi}_1$, which can be explicited in the form

$$U_1 + U_2 = -\frac{1}{2}\left[\int_a^x (U_1 - U_2)(y)dy - M\right] \times \left[4p - \int_a^x (U_1 - U_2)(y)dy + M\right], \tag{28}$$

where a is a chosen real number and M is a constant determined from a, $R_1(k)$, and $R_2(k)$.

VII. THE SUBCLASS $\mathscr{R}_0$ AND THE USUAL SCHRÖDINGER $\hat{S}T$

We consider the subclass $\mathscr{R}_0$ of spectral potentials $R(k)$ $(k \in \mathbb{C})$ in the form

$$R(k) = \frac{i}{2}\delta(k_2)r(k) + \sum_{n=1}^{N} \pi\delta(k - k_n)C_n, \tag{29}$$

with $r(k) = O(1/k)$ as $|k| \to \infty$ $(k \in \mathbb{R})$, $\operatorname{Im} k_n > 0$, which corresponds to a subclass $\mathscr{U}_0$ of spatial potentials $U(x)$ $(x \in \mathbb{R})$, which go to zero as $|x| \to \infty$. Then the previously defined ST $\mathscr{S}$ admits a bijective restriction: $\mathscr{S}_0$: $\mathscr{R}_0 \to \mathscr{U}_0$. Then $\mathscr{S}_0 \doteq (\mathscr{S}_0)^{-1}$ is exactly the usual Schrödinger $\hat{S}T$ and $r(k)$ $(k \in R)$ is the reflection coefficient, the k_n's and the C_n's correspond to the bound states, and $F(k,x)$ coincides with $t(k)\,\varphi(k,x)$ for $\operatorname{Im} k > 0$ and $\psi(-k,x)$ for $\operatorname{Im} k \leqslant 0$, where $t(k)$ is the transmission coefficient and $\varphi(k,x)$ and $\psi(k,x)$ are the Jost solutions of (S): $\varphi(k,x) \sim e^{-ikx}$ $(x \to -\infty)$, $\psi(k,x) \sim e^{ikx}$ $(x \to \infty)$.

In this case it is easy to find that

$$\left(\int dx\right) f = \int_{\infty}^{x} f(y)dy = \int_{-\infty}^{x} f(y)dy, \quad \text{for } f \in \mathscr{E}'. \tag{30}$$

Hence we find the usual expressions for the operators L and L^+. The local conservation laws can be integrated (the system is "closed"), which yields an infinite set of constants of motion $C_n \doteq \int_{-\infty}^{\infty} \varphi_n(x,t)dx$. Since $\mathscr{S}_0$ is bijective the

Bäcklund relations now provide Bäcklund transformations. Note that for $a = \infty$ in Eq. (28) we have $M = 0$.

ACKNOWLEDGMENTS

The authors would like to thank Professor P. C. Sabatier for stimulating discussions. M. J. is grateful to Professor G. Soliani, Professor P. Rotelli, and the Department of Physics of the University of Lecce (Italy) for kind hospitality.

M. M. acknowleges the Instituto de Física Teórica de São Paulo for scientific support and the Coordenação Aperfeiçoamento de Pessoal de Nivel Superior for financial support within the agreement with the Comité Français d'Evaluation de la Coopération avec le Brésil. This work was supported in part by M.P.I., Italy.

[1] M. Jaulent and M. Manna, "Construction of "spatial" transforms from $\bar{\partial}$ equations: The 2 module of solutions of a $\bar{\partial}$ equation and the KdV hierarchy," in *Inverse Problems with Interdisciplinary Applications*, edited by P. C. Sabatier (Academic, New York, to appear).

[2] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transforms* (SIAM, Philadelphia, 1981).

[3] F. Calogero and A. Degasperis, *The Spectral Transform and Solitons: Tools to Solve and Investigate Nonlinear Evolution Equations* (North-Holland, Amsterdam, 1982), Vol. 1.

[4] V. E. Zakharov, S. V. Manakov, S. P. Novikov, and L. P. Pitaevskii, *Theory of Solitons: The Inverse Scattering Method* (Plenum, New York, 1984).

[5] A. S. Fokas and M. J. Ablowitz, *Lectures on the Inverse Scattering Transform for Multidimensional (2+1) Problems, in Nonlinear Phenomena, Lecture Notes in Physics*, edited by K. B. Wolf (Springer, Berlin, 1983).

[6] A. S. Fokas and M. J. Ablowitz, Phys. Rev. Lett. **47**, 1096 (1981).

[7] M. J. Ablowitz, A. S. Fokas, and R. L. Anderson, Phys. Lett. A **93**, 375 (1983).

Yang–Mills–Higgs theory on a compact Riemann surface

Mitsunori Noguchi
Department of Mathematics, Duke University, Durham, North Carolina 27706

(Received 18 March 1987; accepted for publication 3 June 1987)

Jaffe and Taubes [*Vortices and Monopoles* (Birkhauser, Boston, 1980)] have shown the existence and uniqueness of n-vortex solutions on the complex plane. In this paper, their results are generalized to an arbitrary $U(1)$ bundle over a compact Riemann surface with a Hermitian metric. Berger's "nonlinear analysis" [*Nonlinearity and Functional Analysis* (Academic, New York, 1977)] has provided an effective method to prove the existence part of the main theorem of this paper.

I. INTRODUCTION

Throughout this paper (M,g) will denote a compact Riemann surface equipped with a Hermitian metric g and (L,h) will denote a complex line bundle with a fiber metric h. The Abelian Higgs functional is defined to be

$$A(\nabla,\varphi) = \|\nabla\varphi\|_2^2 + \|\Xi_\nabla\|_2^2 + \int_M V(\varphi)*1$$

for each connection-section pair (∇,φ) of (L,h), where $\|\cdot\|_2$ is a suitable L_2 norm, Ξ_∇ the curvature of ∇, and $V(\varphi) = (\lambda/4)[1 - h(\varphi,\varphi)]^2$ the Higgs potential with a constant λ, and where $*$ is the Hodge operator of g. Let $c_1(L)$ be the first Chern class of (L,h) and let $\nabla^{(0,1)}\varphi$ be the $(0,1)$ component of $\nabla\varphi$. If one assumes that $c_1(L)\geqslant 0$ and $\lambda = 1$, the Euler–Lagrange equations of $A(\nabla,\varphi)$ reduce to the following first-order system (the vortex equations):

$$\nabla^{(0,1)}\varphi = 0, \quad i*\Xi_\nabla = \tfrac{1}{2}[1 - h(\varphi,\varphi)], \tag{1.1}$$

and $A(\nabla,\varphi)$ achieves a topological minimum $2\pi c_1(L)$ at a solution (∇,φ) to (1.1).[1–3]

The solutions to (1.1) are called n-vortex solutions when $n = c_1(L)\geqslant 0$, and the group $\mathrm{Aut}(L)$ of $U(1)$ automorphisms of L defines an equivalence relation, called gauge equivalence, on the set of n-vortex solutions. Jaffe and Taubes have shown that on $M = \mathbb{C}$, for each effective divisor D, there exists a finite action smooth solution (∇,φ) to (1,1), unique up to gauge equivalence, such that $A(\nabla,\varphi) = 2\pi \deg D$ and φ determines D.[2] The purpose of this paper is to prove a compact analog of their result.

In what follows, $\mathrm{Div}(M)$ will denote the set of divisors on (M,g), $[D]$ the invertible sheaf of $D\in\mathrm{Div}(M)$, $\mathcal{O}(L)$ the sheaf of sections whenever (L,h) is given a holomorphic structure, $\mathcal{O}(U)$ the sheaf of holomorphic functions on the open set $U\subseteq(M,g)$, $H^0(M,\mathcal{O}(L))$ the global holomorphic sections, (φ) the divisor of $\varphi\in H^0(M,\mathcal{O}(L))$, and $\mathcal{O}(D) = \mathcal{O}([D])$ as commonly written.

Main Theorem: Let (M,g) be a compact Riemann surface equipped with a Hermitian metric g. Let (L,h) be a complex line bundle over (M,g) such that $c_1(L) = n\geqslant 0$. Under these assumptions, the following statements hold.

(i) An n-vortex solution (∇,φ) exists if and only if $n < (4\pi)^{-1}\,\mathrm{Vol}(M)$, where $\mathrm{Vol}(M)$ is the volume of (M,g). Note that if the Gaussian curvature k_g of g is a nonzero constant, the above condition is equivalent to $n < (2k_g)^{-1}\chi(M)$, where $\chi(M)$ is the Euler characteristic of (M,g).

(ii) Let $D\in\mathrm{Div}(M)$ be an effective divisor of degree n, where n satisfies the condition in (i). There exists an n-vortex solution (∇,φ) such that $\varphi\in H^0(M,\mathcal{O}(L))$ satisfies $(\varphi) = D$ when L is given the holomorphic structure defined by ∇.

(iii) The solution (∇,φ) described in (ii) in unique up to gauge equivalence.

II. PROOF OF MAIN THEOREM

Let $D\in\mathrm{Div}(M)$ be an effective divisor of the form $D = \Sigma_{i=1}^N a_i\cdot p_i$, $a_i\geqslant 0$, $p_i\in M$, for $i = 1,2,\ldots,N$. Let $\{f_\alpha\}$ be a set of local defining functions of D w.r.t. some covering $\{U_\alpha\}$ of M.[4] We may define a globally defined distribution $\delta(D)\equiv(2\pi)^{-1}\Delta\ln|f_\alpha|$ on (M,g), where Δ is the Laplacian of g. The entire proof of the Main Theorem depends upon solving the following problem.

Given an effective divisor $D\in\mathrm{Div}(M)$, find a function $u\in C^\infty(M - D)$ satisfying

$$\Delta u = \tfrac{1}{2}(e^{2u} - 1) + 2\pi\,\delta(D). \tag{2.1}$$

We have the following crucial lemmas.

Lemma 1: A solution to (2.1) is unique.

Lemma 2: A necessary and sufficient condition for the existence of a solution to (2.1) is that

$$\deg D < (4\pi)^{-1}\,\mathrm{Vol}(M),$$

where $\deg D = \Sigma_{i=1}^N a_i$.

We will prove these lemmas later in the paper.

Proof of (i) and (ii): Suppose we have a smooth solution (∇,φ) to

$$\nabla^{(0,1)}\varphi = 0, \tag{2.2a}$$

$$i*\Xi_\nabla = \tfrac{1}{2}[1 - h(\varphi,\varphi)]. \tag{2.2b}$$

Local solutions of $\nabla^{(0,1)}s = 0$ define a holomorphic structure on L. Consequently, (2.2a) simply says $\varphi\in H^0(M,\mathcal{O}(L))$, and ∇ will be the unique metric connection compatible with the holomorphic structure. Locally, $\Xi_\nabla = \bar\partial\partial\ln h^2$ when we write $h(s,s) = h^2$. Writing $\varphi = f\cdot s$ for some $f\in\mathcal{O}(U)$, (2.2b) becomes

$$\Delta\ln h = \tfrac{1}{2}(h^2|f|^2 - 1). \tag{2.3}$$

Adding $\Delta\ln|f|$ to both sides, (2.3) becomes

$$\Delta\ln h|f| = \tfrac{1}{2}(h^2|f|^2 - 1) + \Delta\ln|f|,$$

which is simply

$$\Delta\ln[h(\varphi,\varphi)]^{1/2} = \tfrac{1}{2}[h(\varphi,\varphi) - 1] + 2\pi\,\delta((\varphi)). \tag{2.4}$$

 0022-2488/87/102343-04$02.50 © 1987 American Institute of Physics

Thus $u=\ln[h(\varphi,\varphi)]^{1/2}$ solves (2.1) for $D=(\varphi)$, and Lemma 2 implies $\deg D<(4\pi)^{-1}\,\mathrm{Vol}(M)$. But $\deg D=\deg(\varphi)=c_1(L)=n$, and $n<(4\pi)^{-1}\,\mathrm{Vol}(M)$ as desired.

Next, suppose $n<(4\pi)^{-1}\,\mathrm{Vol}(M)$. Given an effective divisor D of degree n, solve problem (2.1). Since D is effective, we may choose a global section $\varphi'\in H^0(M,\mathcal{O},(D))$ such that $(\varphi')=D$. Let s be a local holomorphic frame of $[D]$ and write $\varphi'=fs$. Define a Hermitian metric $h'=e^u/|f|$ locally. Let ∇' be the canonical connection of h' on $[D]$. It follows that

$$\nabla'^{(0,1)}\varphi'=0,\quad i*\Xi_{\nabla'}=\tfrac{1}{2}[1-h'(\varphi',\varphi')] \tag{2.5}$$

holds on $([D],h')$. Choose a $U(1)$ isomorphism Φ: $(L,h)\to([D],h')$ such that $\Phi^*h'=h$. Define $\varphi=\Phi^*\varphi'$, $\nabla=\Phi^*\nabla'$, where Φ^* indicates an obvious pullback map. Equations (2.5) give

$$\nabla^{(0,1)}\varphi=0,\quad i*\Xi_{\nabla}=\tfrac{1}{2}[1-h(\varphi,\varphi)],$$

and the (∇,φ) above defines a gauge equivalence class of n-vortex solutions. □

Proof of (iii): Suppose there are two distinct solutions, (∇_i,φ_i), $i=1,2$, to (2.2a) and (2.2b). Let s_i be a local holomorphic frame defined by ∇_i. Write $\varphi_i=f_ie_i$. By our assumption, we have $(\varphi_1)=(\varphi_2)=D$ which implies $\Delta\ln|f_1|=\Delta\ln|f_2|=2\pi\delta(D)$. Equation (2.2b) becomes

$$\Delta\ln h_i|f_i|=\tfrac{1}{2}(h_i^2|f_i|^2-1)+2\pi\delta(D),$$

and Lemma 1 says $h_1^2|f_1|^2=h_2^2|f_2|^2$, or $h(\varphi_1,\varphi_1)=h(\varphi_2,\varphi_2)$. Thus we find a smooth function $e^{i\chi}$: $M\to U(1)$ such that $\varphi_2=e^{i\chi}\varphi_1$. Locally, we may define $g_{12}=e^{i\chi}(f_1/f_2)$ so that $s_2=g_{12}s_1$. The connection form of ∇_i w.r.t. s_i is given by $A_i=\partial\ln h_i^2$. As we change our local frame from s_1 to s_2, $A_1=\partial\ln h_1^2$ transforms into $A_1'=\partial\ln h_1^2+g_{12}^{-1}dg_{12}$. Compute

$$A_2-A_1'=\partial\ln(h_2^2/h_1^2)-\partial\ln(f_1/f_2)-i\,d\chi.$$

But $h_1^2|f_1|^2=h_2^2|f_2|^2$ implies

$$\partial\ln(h_2^2/h_1^2)=\partial\ln(f_1/f_2),$$

and we obtain $\nabla_2=\nabla_1-i\,d\chi$ as claimed. □

III. PROOF OF LEMMA 1

Here we prove the uniqueness of a solutionn to the problem (2.1). In what follows, $\|\cdot\|_p$ will denote the norm in $L_p(M,g)$ and $\|\cdot\|_{\mathscr{H}}$ will denote the norm in $\mathscr{H}$ whenever $\mathscr{H}$ is a Hilbert space. First, consider the Poisson equation

$$\Delta u_0=-2\pi c^{-1}(\deg D)F+2\pi\delta(D),$$

where $F\in C^\infty(M)$ such that $\int_M F*1=c>0$. There is a solution $u_0\in C^\infty(M-D)$ uniquely determined up to the addition of a constant.[5,6] Fix a pair (F,u_0) on (M,g). Let $K\equiv-1+4\pi c^{-1}(\deg D)F$. Equation (2.1) is equivalent to

$$\Delta v=\tfrac{1}{2}(e^{2u_0}\cdot e^{2v}+K), \tag{3.1}$$

where $v=u-u_0\in C^\infty(M)$. Let H be the Sovolev space $W_{1,2}(M,g)$ of functions on (M,g). Define a functional a: $H\to\mathbb{R}$ by $a(v)=\frac{1}{2}\|\nabla v\|_2^2+\frac{1}{4}(p,e^{2v})+\frac{1}{2}(K,v)$, where $p=e^{2u_0}\geqslant 0$ is smooth, $(\cdot,\cdot)$ is the obvious bilinear pairing, and ∇ is the gradient operator of g. We will prove Lemma 2 in the following steps.

Step 1: $a(v)$ is defined for $v\in H$.

Step 2: $a(v)\in C^1(H,\mathbb{R})$.

Step 3: $a(v)$ is strictly convex.

As a result, a weak solution to (3.1) will be proved unique.

Proof of step 1: We have the following inequalities.

(i) $\|\nabla v\|_2\leqslant\|v\|_H$.

(ii) $$|(p,e^{2v})|\leqslant\sup_M p\cdot\int_M e^{2|v|}*1$$
$$\leqslant\sup_M p\{\gamma\exp[2|\bar v|+(2\|\nabla v\|_2)^2/4\beta]\}$$

for some γ,β, where

$$\bar v=\int_M v*\left(\int_M *1\right)^{-1}$$

(see Ref. 7). Note that

$$\|v\|_p\leqslant cp^{1/2}(\|v\|_2+\|\nabla v\|_2),\quad \text{for } p\geqslant 1$$

(see Ref. 7).

(iii) $|(K,v)|\leqslant\|K\|_2\|v\|_2\leqslant\|K\|_2\|v\|_H$. □

Proof of step 2: Write $a(v)=a_1(v)+a_2(v)+a_3(v)$, where $a_1(v)=\frac{1}{2}\|\nabla v\|_2^2$, $a_2(v)=\frac{1}{4}(p,e^{2v})$, and $a_3(v)=\frac{1}{2}(K,v)$. Obviously, $a_1,a_3\in C^1(H,\mathbb{R})$, and it suffices to show $a_2(v)\in C^1(H,\mathbb{R})$. To this end compute the Gauteaux derivative at $v\in H$ for $h\in H$ as

$$da_2(v,h)=\frac{d}{dt}\bigg|_{t=0}\frac{1}{4}(p,e^{2(v+th)})=\frac{1}{2}\int_M p\,e^{2v}h*1,$$

where the last step can be justified by a standard theorem of calculus.[8]

Next, we show $da_2(v,\cdot)\in H^*$ is continuous in order to guarantee the Fréchet derivative $a_2'(v)$ exists and continuous.[9] Let $v_n\to v$ in H. We have

$$\|da_2(v_n,\cdot)-da_2(v,\cdot)\|_{H^*}$$
$$\leqslant\tfrac{1}{2}\sup_M p\cdot\|e^{2v_n}-e^{2v}\|_2\cdot\sup_{\|h\|_H=1}\|h\|_2,$$

and $\|e^{2v_n}-e^{2v}\|_2\to 0$ can be found in Ref. 7. □

Proof of step 3: Since $a_1(v)$ is quadratic, it is strictly convex. The linearity of $a_3(v)$ implies that $a_1(v)+a_3(v)$ is strictly convex. $a_2(v)$ is clearly convex because $p\geqslant 0$. □

IV. PROOF OF LEMMA 2

Proof of necessity: Recall Eq. (3.1) given by

$$\Delta v=\tfrac{1}{2}(e^{2u_0}\cdot e^{2v}+K), \tag{4.1}$$

where $K\equiv-1+4\pi c^{-1}(\deg D)F$. Upon integrating (4.1) over M, we get

$$\int_M K*1=-\int e^{2u_0}\cdot e^{2v}*1<0, \tag{4.2}$$

since $e^{2u_0}\geqslant 0$. However, (4.2) implies

$$\deg D<(4\pi)^{-1}\,\mathrm{Vol}(M).$$ □

Proof of sufficiency: For the sufficiency we must show the existence of a smooth solution to Eq. (4.1) assuming condition (4.2).

Letting $p=e^{2u_0}\geqslant 0$ as before, define operators L,P:

$H \to H$ by

$$(Lu,v) = \int_M \nabla u \cdot \nabla v * 1, \quad (Pu,v) = \frac{1}{2}\int_M p e^{2u} v * 1.$$

Note that L is a bounded self-adjoint operator and P makes sense because $e^{2u}, v \in L_2(M,g)$ due to the fact that $x \in W_{1,2}(M,g)$ implies

(i) $x \in L_p(M,g)$, for $1 \leqslant p < \infty$ (Sobolev inequality),

(ii) $e^x \in L_p(M,g)$, for $1 \leqslant p < \infty$ (see Ref. 7).

We may consider the following operator equation on H:

$$Lv + Pv = f, \tag{4.3}$$

where $(f,u) = -\frac{1}{2}\int_M Ku * 1$ for all $u \in H$.

Now we verify the following claim.

Claim: If $\int_M K * 1 < 0$, then the operator equation $Lv + Pv = f$ can be solved for v.

Note that the claim together with the combined use of the L_p and Schauder regularity theorems[7] will complete our sufficiency proof.

Proof of the claim: Decomposing H as $H = \operatorname{Ker} L \oplus H_1$, we obtain projection operators, π_1 and π_2, as in the following diagram:

Writing $v \in H$ as $v = c + w$, $c \in \operatorname{Ker} L$, and $w \in H_1$, (4.3) becomes

$$\pi_1 P(c + w) = \pi_1 f, \tag{4.4a}$$

$$Lw + \pi_2 P(c + w) = \pi_2 f. \tag{4.4b}$$

Note that (4.4a) simply says

$$e^{2c}\int_M p e^{2w} * 1 = -\int_M K * 1, \tag{4.5}$$

since $\operatorname{Ker} L$ consists of constant functions.

Let $A \equiv -\int_M K * 1$ which is positive by our hypothesis. Regarding c in (4.5) as a function of w, we have

$$c(w) = \frac{1}{2}\left[\ln A - \ln\left(\int_M p e^{2w} * 1\right)\right].$$

We will complete the proof in the following steps.

Step 1: $P \in C^1(H,H)$.

Step 2: $c: H_1 \to \operatorname{Ker} L$ is of class C^1.

Consider the map $T: H_1 \to H_1$ defined by

$$T(w) = Lw + \pi_2 P[c(w) + w].$$

Steps 1 and 2 together with the chain rule implies $T \in C^1(H_1,H_1)$.

Step 3: $(T'(w)v,v) \geqslant \|v\|^2_{H_1}$.

By the Lax–Milgram lemma, step 3 implies that $T'(w) \in L(H_1,H_1)$ is an invertible linear operator and $\|[T'(w)]^{-1}\|_{L(H_1,H_1)} \leqslant 1$. Hadamard's theorem applied to T concludes T is a homeomorphism of H_1 onto H_1.[9] This completes the proof of our claim.

Proof of step 1: Compute the Gauteaux derivative

$$dP(x_0,h) = \frac{d}{dt}\bigg|_{t=0} P(x_0 + th)$$

at $x_0 \in H$, for $h \in H$. For $v \in H$, we have

$$(dP(x_0,h),v) = \frac{d}{dt}\bigg|_{t=0}\frac{1}{2}\int_M p e^{2(x_0+th)} v * 1$$

$$= \int_M p e^{2x_0} h v * 1,$$

and the last step can be justified using a standard theorem of calculus.[8] We must show $dP(x,\cdot) \in L(H,H)$ is continuous in order to guarantee the Fréchet derivative $P'(x)$ exists and continuous.[9] To this end let $x_n \to x$ in H. We have

$$\|dP(x_n,\cdot) - dP(x,\cdot)\|_{L(H,H)}$$

$$\leqslant \sup_M p \|e^{2x_n} - e^{2x}\|_2 \sup_{\|h\|_H = 1} \sup_{\|v\|_H = 1} \|h\|_4 \|v\|_4$$

and $\|e^{2x_n} - e^{2x}\|_2 \to 0$ can be found in Ref. 7. □

Proof of step 2: As in step 1 we compute

$$dc(w,v) = -\left(\int_M p e^{2w} * 1\right)^{-1}\int p e^{2w} v * 1.$$

One only needs to show $dc(w,\cdot)$ is continuous in w. To this end, let $w_n \to w$ in $H_1 \subset H$. Compute

$$\|dc(w_n,\cdot) - dc(w,\cdot)\|_{H_1^*}$$

$$= \sup_{\|v\|_{H_1}=1}\left|\int_M\left[\left(\int_M p e^{2w_n} * 1\right)^{-1} p e^{2w_n}\right.\right.$$

$$\left.\left. - \left(\int_M p e^{2w} * 1\right)^{-1} p e^{2w}\right] v * 1\right|$$

$$\leqslant \sup_M p \left\|\left(\int_M p e^{2w_n} * 1\right)^{-1} e^{2w_n}\right.$$

$$\left. - \left(\int_M p e^{2w} * 1\right)^{-1} e^{2w}\right\|_2 \sup_{\|v\|_{H_1}=1} \|v\|_2.$$

Note that $\|e^{2w_n} - e^{2w}\|_2 \to 0$, and

$$\left|\int_M p e^{2w_n} * 1 - \int_M p e^{2w} * 1\right| \leqslant \sup_M p \left|\int_M e^{2w_n} - \int_M e^{2w}\right| \to 0$$

since $x \to \int_M e^x$ is continuous w.r.t. the weak convergence in H.[10] □

Proof of step 3: We compute

$$(T'(w)v,v)$$

$$= \int_M |\nabla v|^2 * 1 + (\pi_2 P'[c(w) + w][c'(w)v + v], v)$$

$$= \int_M |\nabla v|^2 * 1 + e^{2c(w)}\left(\int_M p e^{2w} * 1\right)^{-1}$$

$$\times\left[\left(\int_M p^{2w} v^2 * 1\right)\left(\int_M e^{2w} * 1\right)\right.$$

$$\left. - \left(\int_M p e^{2w} v * 1\right)^2\right] \geqslant \int_M |\nabla v|^2 * 1.$$

Moreover,

$$\int_M |\nabla v|^2 * 1 = \|v\|^2_{H_1}$$

(see Ref. 11).

ACKNOWLEDGMENTS

I wish to thank Professor W. K. Allard for helpful discussions and Ms. B. Farrell for typing this paper.

[1]E. B. Bogomol'nyi, Sov. J. Nucl. Phys. **24**, 449 (1976).
[2]A. Jaffe and C. Taubes, *Vortices and Monopoles* (Birkhauser, Boston, 1980).
[3]M. Noguchi, Ph.D. thesis, Duke University, 1985.
[4]P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).
[5]S. Chern, Am. J. Math. **82**, 323 (1960).
[6]M. Noguchi, "On the existence of certain pseudo-metrics on a compact Riemann surface," submitted to J. Differ. Geom.
[7]J. Kazdan and F. Warner, Ann. Math. **99**, 14 (1974).
[8]S. Lang, *Analysis II* (Addison-Wesley, Reading, MA, 1969).
[9]M. Berger, *Nonlinearity and Functional Analysis* (Academic, New York, 1977).
[10]N. Trudinger, J. Math. Mech. **17**, 473 (1967).
[11]M. Berger, J. Differ. Geom. **5**, 325 (1971).

Differential geometric aspects of the Cartan form: Symmetry theory

David Betounes
Mathematics Department, University of Southern Mississippi, Hattiesburg, Mississippi 39406-5045

(Received 16 September 1985; accepted for publication 6 May 1987)

This paper demonstrates that the classical Cartan form θ_L^1 is not adequate for the determination of all the natural symmetries and conservation laws for a Lagrangian L. It is shown that the various extensions $\theta_L^2,\ldots,\theta_L^r$ of the classical Cartan form, introduced in recent papers, give larger symmetry groups: $G_1 \subset G_2 \subset \cdots \subset G_r$. This paper also introduces the notion of contact equivalent Lagrangians, which serves to clarify the idea that different Lagrangians can give rise to the same variational and symmetry theories.

I. INTRODUCTION

The standard geometric formulation of the variational theory for first-order Lagrangians $L: J^1E \to \mathbb{R}$ uses the classical Cartan form θ_L^1, a certain differential p-form on the jet bundle J^1E associated with a fiber bundle $E \to N$ [$\dim(N) = p$, $\dim(E) = p+q$].[1] Recent papers[2,3] have discussed various improved versions $\theta_L^2,\ldots,\theta_L^r$, $r = \min(p,q)$, of the classical Cartan form and have argued that θ_L^r is the most suitable of these forms for use in the variational theory. In this paper we examine the symmetry theory connected with each of these forms and show that the Cartan form θ_L^r determines a larger group of symmetries of L than the classical one.

Specifically, we first show (Sec. III) that the Cartan forms θ_L^k, $k = 1,\ldots,r$ are globally defined forms on J^1E and have the mapping properties

$$f^{1*}(\theta_L^k) = \theta_{f(L)}^k \quad \text{and} \quad \mathscr{L}_{X^1}(\theta_L^k) = \theta_{X(L)}^k. \tag{1.1}$$

These properties prove useful in the ensuing symmetry theory, and curiously enough seem indigenous to first-order field theories: the analogs of the forms θ_L^k for higher-order Lagrangians $L: J^mE \to \mathbb{R}$ are *local* p-forms on J^mE but fail to extend to *global* forms.

The Cartan forms θ_L^k differ from one another by contact element terms and so each gives the same global differential geometric formulation of the Euler–Lagrange equations: $\sigma^{1*}(X^1 \lrcorner d\theta_L^k) = 0$. However, it is shown (Sec. IV) that the natural symmetry groups $G_k = \{f \mid f^{1*}(d\theta_L^k) = d\theta_L^k\}$ associated with each Cartan form are in general distinct: $G_1 \subset G_2 \subset \cdots \subset G_r$. This demonstrates the extent to which the symmetry group G_1 determined by the classical Cartan form fails to encompass all the natural symmetries of the Lagrangian L. The mapping properties (1.1) serve to simplify the computation of the symmetry groups (and algebras) as well as the corresponding conservation laws. Finally we introduce (Sec. V) the concept of contact equivalence, two Lagrangian K,L being contact equivalent if their Cartan forms are related by $d\theta_K^r = h^{1*}(d\theta_L)$. It is shown that the extremals, symmetry groups, and conservation laws for K and L are the same (up to isomorphism) and thus the physics connected with either Lagrangian is the same.

II. PRELIMINARY DEFINITIONS AND NOTATION

The natural setting for Lagrangian field theories involves a (smooth) fiber bundle $\pi: E \to N$ over a base manifold N with standard fiber $F \cong \pi^{-1}\{x\}$ [$\dim(N) = p$ and $\dim(F) = q$]. The sections $\sigma: N \to E$ are the classical fields of interest [$\pi \circ \sigma(x) = x$] and we denote the collection of all such sections by $\Gamma(E)$. Each pair of charts (U,x_i) and (V,y^a) on N and F gives rise to a fibered chart (C,x_i,y^a) on E ($i = 1,\ldots,p$ and $a = 1,\ldots,q$). The first-order variational theory is based on the geometry of the bundle of one-jets J^1E, This bundle is the collection of all equivalence classes $[\sigma]_x$ of sections of E: two sections $\sigma, \bar{\sigma}$ being equivalent at x, N if $\sigma(x) = \bar{\sigma}(x)$ and $\partial(y^a \circ \sigma)(x)/\partial x_i = \partial(y^a \circ \bar{\sigma})(x)/\partial x_i$ in some (hence every) fibered chart. Then J^1E is endowed with a natural differentiable structure and surjections $\pi_0^1([\sigma]_x) = \sigma(x)$ and $\pi^1([\sigma]_x) = x$ on E and N which make it into a fiber bundle over E (with standard fiber $\mathbb{R}^{pq}$) and a fiber bundle over N (in the extended sense that π^1 is a submersion). In the sequel we adopt the standard practice of not distinguishing notationally between the differential forms and functions ϕ on E and ψ on N and their pullbacks $\pi_0^{1*}(\phi)$, $\pi^{1*}(\psi)$ to J^1E. The fibered chart (C,x_i,y^a) on E extends to a fibered chart (W,x_i,y^a,y_i^a) on J^1E, with the coordinate functions given by $x_i([\sigma]_x) = x_i(x)$, $y^a([\sigma]_x) = y^a \circ \sigma(x)$, and $y_i^a([\sigma]_x) = \partial(y^a \circ \sigma)(x)/\partial x_i$.

We let $B(E)$ denote the group of bundle maps of E. These are the fiber preserving diffeomorphisms f: $E \to E$. Each bundle map induces a diffeomorphism $\hat{f}$: $N \to N$ on the base space ($\pi \circ f = \hat{f} \circ \pi$). The corresponding algebra of infinitesimal bundle maps is denoted by $\mathrm{IB}(E)$. These are the projectible vector fields X: $E \to TE$ on E ($d\pi|_z X_z = \hat{X}_{\pi(z)}$ for some unique vector field $\hat{X}$ on N). There is a natural action of the group $\mathrm{B}(E)$ on the set of sections $\Gamma(E)$. This is given by $f(\sigma) = f \circ \sigma \circ \hat{f}^{-1}$ and is basic to the entire symmetry theory.

The geometric objects associated with the bundle E prolong in a functorial way to corresponding ones on J^1E: each section $\sigma: N \to E$ prolongs to a section $\sigma^1: N \to J^1E$ defined by $\sigma^1(x) = [\sigma]_x$; each bundle map f: $E \to E$ prolongs to a bundle map f^1: $J^1E \to J^1E$ defined by $f^1([\sigma]_x) = [f(\sigma)]_{\hat{f}(x)}$; and each infinitesimal bundle map X: $E \to TE$ prolongs to a vector field X^1 on J^1E defined by $X_m^1(\phi) = d[\phi \circ f_t^1(m)]/dt|_{t=0}$, where $m \in J^1E$, $\phi \in C^\infty(J^1E)$, and f_t is the flow generated by X (thus f_t^1 is the flow generated by

 0022-2488/87/102347-07$02.50 © 1987 American Institute of Physics

X^1). The functorial properties of these prolongations are expressed by the identities

$$(f\circ g)^1 = f^1\circ g^1, \tag{2.1}$$

$$[X,Y]^1 = [X^1, Y^1], \tag{2.2}$$

$$f(\sigma)^1 = f^1\circ\sigma^1\circ\hat{f}^{-1}. \tag{2.3}$$

For the variational theory we assume that N is a semi-Riemannian, orientable manifold with metric g and volume form Δ. A (first-order) Lagrangian is a (smooth) map L: $J^1E\to\mathbb{R}$, and the associated Lagrangian form on J^1E is $L\Delta$. Each bundle map $f\in B(E)$ induces a map (transformation on the set of Lagrangians according to

$$f(L) = (L\circ f^1)J(\hat{f}), \tag{2.4}$$

where $J(\hat{f})$ denotes the Jacobian of f. From this definition it follows that $f^{1*}(L\Delta) = f(L)\Delta$ (here and in the sequel the $*$ denotes the pullback operation induced on the differential forms by a map between manifolds). The corresponding infinitesimal transformation on the set of Lagrangians [induced by an infinitesimal bundle map $X\in IB(E)$] is defined by

$$X(L) = \mathcal{L}_{X^1}(L) + \operatorname{div}(\hat{X})L, \tag{2.5}$$

where $\mathcal{L}_{X^1}$ denotes the Lie derivative. From this it follows that $\mathcal{L}_{X^1}(L\Delta) = X(L)\Delta$.

III. THE CARTAN FORMS

We review here the definitions of the various Cartan forms $\theta^1_L, \theta^2_L,\ldots,\theta^r_L$, $r = \min(p,q)$, and prove several mapping properties of these forms which will be of importance in the symmetry theory. We take a simple direct approach of defining these forms in local coordinates on J^1E and show in Theorem 1 that these coordinate expressions agree on the overlap of any two fibered charts, thus giving rise to *global* Cartan forms. In general one can use the Tulczyjew bicomplex[4] to naturally construct *local* Cartan forms $\theta^1_L,\ldots,\theta^r_L$ for any mth order Lagrangian $L: J^mE\to\mathbb{R}$.[5] For $m = 1$ this construction gives the Cartan forms introduced below, and thus gives a different perspective on the origin and naturality of these forms. Regrettably, for $m > 1$ the construction does not give rise to *global* Cartan forms on J^mE, and therefore one must proceed along different avenues.[6]

Definition: Suppose that $1\leqslant n\leqslant r = \min(p,q)$, $i_1,\ldots,i_n \in\{1,\ldots,p\}$ and that $a_1,\ldots,a_n\in\{1,\ldots,q\}$. For a fibered chart (W,x_i,y^a,y^a_i) on J^1E define the following differential forms on W:

$$\omega^a = dy^a - y^a_i\,dx_i, \tag{3.1}$$

$$\Delta_{i_1\cdots i_n} = \frac{\partial}{\partial x_{i_n}} \lrcorner\cdots\lrcorner \frac{\partial}{\partial x_{i_1}} \lrcorner\,\Delta, \tag{3.2}$$

$$M^n_L(W) = \frac{1}{(n!)^2}\frac{\partial^n L}{\partial y^{a_1}_{i_1}\cdots\partial y^{a_n}_{i_n}}(\omega^{a_1}\cdots\omega^{a_n})\,\Delta_{i_1\cdots i_n}. \tag{3.3}$$

Here and in the sequel there is implied summation over repeated indices and the wedge symbol $\wedge$ has been omitted from the exterior products. Also $\lrcorner$ denotes contraction of a vector field with a differential form. We denote the basic Lagrangian form $L\Delta$ by $\theta^0_L = M^0_L = L\Delta$.

Theorem 1: There is a globally defined form M^n_L on J^1E whose coordinate expression on each fibered chart W is $M^n_L(W)$. This form has the mapping property

$$f^{1*}(M^n_L) = M^n_{f(L)}. \tag{3.4}$$

Consequently the Cartan forms, defined by

$$\theta^k_L = M^0_L + M^1_L + \cdots M^k_L \tag{3.5}$$

($k = 1,\ldots,r$), are global forms on J^1E with the mapping property

$$f^{1*}(\theta^k_L) = \theta^k_{f(L)}. \tag{3.6}$$

Proof: Suppose that f: $E\to E$ is a bundle map and that (W,x_i,y^a,y^a_i), $(\overline{W},\bar{x}_i,\bar{y}^a,\bar{y}^a_i)$ are two fibered charts on J^1E with $f^1(W)\cap\overline{W}$ nonempty. Restricting to these charts f^1: $W\cap(f^1)^{-1}(\overline{W})\to f^1(W)\cap\overline{W}$, and to prove the theorem it suffices to prove that $f^{1*}(M^n_L(\overline{W})) = M^n_{f(L)}(W)$. To prove this introduce the notation $C_{ij} = \partial(\bar{x}_i\circ\hat{f})/\partial x_j$, $C^{-1}_{ij} = [\partial(x_i\circ\hat{f}^{-1})/\partial\bar{x}_j]\circ\hat{f}$, $f^a = \bar{y}^a\circ f$, and $J = J(\hat{f})$. Then the coordinate representation of f^1 on the above charts is $\bar{x}_i\circ f^1 = \bar{x}_i\circ\hat{f}$, $\bar{y}^a\circ f^1 = f^a$, and

$$\bar{y}^a_i\circ f^1 = \left[\frac{\partial f^a}{\partial x_j} + y^b_j\frac{\partial f^a}{\partial y^b}\right]C^{-1}_{ji}. \tag{3.7}$$

From this it follows that $f^{1*}(d\bar{x}_i) = C_{ij}\,dx_j$, $f^{1*}(d\bar{y}^a) = (\partial f^a/\partial x_j)dx_j + (\partial f^a/\partial y^b)dy^b$, $f^{1*}(\bar{y}^a_i) = \bar{y}^a_i\circ f^1$, and consequently

$$f^{1*}(\bar{\omega}^a) = f^{1*}(d\bar{y}^a - \bar{y}^a_i d\bar{x}_i) = \frac{\partial f^a}{\partial y^b}\omega^b, \tag{3.8}$$

$$f^{1*}(\bar{\Delta}_{i_1\cdots i_n}) = (C^{-1}_{j_1i_1}\cdots C^{-1}_{j_ni_n})J\Delta_{j_1\cdots j_n}. \tag{3.9}$$

Thus

$$f^{1*}(M^n_L(\overline{W})) = \frac{1}{(n!)^2}\left(\frac{\partial^n L}{\partial\bar{y}^{a_1}_{i_1}\cdots\partial\bar{y}^{a_n}_{i_n}}\circ f^1\right)\left(\frac{\partial f^{a_1}}{\partial y^{b_1}}\cdots\frac{\partial f^{a_n}}{\partial y^{b_n}}\right) \times(\omega^{b_1}\cdots\omega^{b_n})(C^{-1}_{j_1i_1}\cdots C^{-1}_{j_ni_n})J\Delta_{j_1\cdots j_n}. \tag{3.10}$$

However, using the chain rule on $\partial^n f(L)/(\partial y^{b_1}_{j_1}\cdots\partial y^{b_n}_{j_n})$ together with Eq. (3.7) one finds that Eq. (3.10) is the same as $M^n_{f(L)}(W)$, and this completes the proof.

Corollary 1: Suppose that $X\in IB(E)$ is an infinitesimal bundle map. Then

$$\mathcal{L}_{X^1}(M^n_L) = M^n_{X(L)}, \tag{3.11}$$

and consequently

$$\mathcal{L}_{X^1}(\theta^n_L) = \theta^n_{X(L)} \tag{3.12}$$

for each $n = 1,\ldots,r$.

Proof: Let f_t be the flow generated by X, so that f^1_t is the flow generated by X^1. Then since $f^{1*}_t(L\Delta) = f_t(L)\Delta$, and $\mathcal{L}_{X^1}(L\Delta) = X(L)\Delta$, it follows that

$$X(L) = \frac{d}{dt}f_t(L)\Big|_{t=0}. \tag{3.13}$$

Thus using Eq. (3.4), one finds that

$$\mathscr{L}_{X^1}(M_L^n) = \frac{d}{dt}(f_t^{1*}(M_L^n))\Big|_{t=0}$$
$$= \frac{d}{dt} M^n_{f_t(L)}\Big|_{t=0} = M^n_{X(L)}.$$

Comments: The differential ideal $\mathscr{C}$ of contact forms on J^1E is the set of all differential forms ϕ such that $\sigma^{1*}(\phi) = 0$ for every section $\sigma \in \Gamma(E)$. The forms ω^a in Eq. (3.1) are the basic contact one-forms and the forms $(\omega^{a_1} \cdots \omega^{a_n})\ \Delta_{i_1 \cdots i_n}$ constitute a basis for the contact ideal. Thus the p-forms M_L^n $(n = 1,...,r)$ are contact forms, and from this one can show, using basic properties of contractions and pullbacks, that

$$\sigma^{1*}(\theta_L^k) = (L \circ \sigma^1)\Delta, \tag{3.14}$$

$$\sigma^{1*}(X^1 \rfloor \theta_L^k) = \sigma^{1*}(X^1 \rfloor \theta_L^1), \tag{3.15}$$

$$\sigma^{1*}(X^1 \rfloor d\theta_L^k) = \sigma^{1*}(X^1 \rfloor d\theta_L^1), \tag{3.16}$$

for every σ, X, and $k = 1,...,r$. The identity (3.16) shows that each of the Cartan forms is equally suitable for formulating the global version of the Euler–Lagrange equations on J^1E (rather than on J^2E).

Definition: [cf. Ref. 1(e)] A section $\sigma \in \Gamma(E)$ is an extremal of L if

$$\forall X \in \mathrm{IB}(E), \quad \sigma^{1*}(X^1 \rfloor d\theta_L^1) = 0. \tag{EL1}$$

We let $\mathrm{Ext}(L)$ denote the set of all extremals of L. Several useful alternative versions of the Euler–Lagrange equations (EL1) arise from the following observations. First since

$$X^1 \rfloor d\theta_L^1 = \mathscr{L}_{X^1}(\theta_L^1) - d(X^1 \rfloor \theta_L^1)$$
$$= \theta^1_{X(L)} - d(X^1 \rfloor \theta_L^1),$$

one sees that (EL1) is equivalent to

$$\forall X \in \mathrm{IB}(E), \quad d\left[\sigma^{1*}(X^1 \rfloor \theta_L^1)\right] = [X(L) \circ \sigma^1]\Delta. \tag{EL2}$$

Next observe that there exists a vector field $W = W(\sigma, X, L)$ on N such that $W \rfloor \Delta = \sigma^{1*}(X^1 \rfloor \theta_L^1)$, and thus (EL2) is equivalent to

$$\forall X \in \mathrm{IB}(E), \quad \mathrm{div}(W(\sigma, X, L)) = X(L) \circ \sigma^1. \tag{EL3}$$

These global equations give the classical Euler–Lagrange equations locally on each chart. Namely suppose that locally $\hat{X} = \xi_i(x)(\partial/\partial x_i)$ and $X = \xi_i(x)(\partial/\partial x_i) + \eta^a(x,y)(\partial/\partial y^a)$. Then the local expression for X^1 is

$$X^1 = \xi_i \frac{\partial}{\partial x_i} + \eta^a \frac{\partial}{\partial y^a} + \zeta^a_i \frac{\partial}{\partial y^a_i}, \tag{3.17}$$

where

$$\zeta^a_i = \frac{\partial \eta^a}{\partial x_i} + \frac{\partial \eta^a}{\partial y^b} y^b_i - \frac{\partial \xi_j}{\partial x_i} y^a_j. \tag{3.18}$$

Then since

$$X^1 \rfloor \theta_L^1 = L\xi_i \Delta_i + \frac{\partial L}{\partial y^a_i}(\eta^a - y^a_j \xi_j)\,\Delta_i - \frac{\partial L}{\partial y^a_i} \omega^a X^1 \rfloor \Delta_i,$$

one finds for the components of W (suppressing the σ on the right-hand side)

$$W^i(\sigma, X, L) = L\xi_i + \frac{\partial L_a}{\partial y_i}(\eta^a - y^a_j \xi_j). \tag{3.19}$$

Thus, in particular, for $X = \partial/\partial y^a$, the local version of (EL3) is

$$|g|^{-1/2} \frac{\partial}{\partial x_i}\left(|g|^{1/2} \frac{\partial L}{\partial y^a_i}\right) = \frac{\partial L}{\partial y^a}. \tag{3.20}$$

IV. THE LAGRANGIAN SYMMETRY THEORY

Various treatments of the Lagrangian symmetry theory have been developed using either the classical Cartan form θ_L^1 or, more generally, Lepagian equivalents of it.[7] Indeed by abstracting the essential structure of the Euler–Lagrange equations to $\sigma^{1*}(X^1 \rfloor d\phi) = 0$, one can develop a general symmetry theory based on any differential form ϕ on J^1E. However, the real substance of the theory comes from the Cartan forms θ_L^k, since the association $L \to \theta_L^k$ together with the mapping properties (1.1) allow one to reduce the analysis from working with differential forms to working with Lagrangians.

Definition: For a Lagrangian L and for $k \in \{0,1,..., r\}$ let G_k and $\mathscr{G}_k$ be the subsets of $\mathrm{B}(E)$ and $\mathrm{IB}(E)$ defined by

$$G_k = G_k(L) = \{f \mid f^{1*}(d\theta_L^k) = d\theta_L^k\}, \tag{4.1}$$

$$\mathscr{G}_k = \mathscr{G}_k(L) = \{X \mid \mathscr{L}_{X^1}(d\theta_L^k) = 0\}. \tag{4.2}$$

Using the functorial properties (2.1) and (2.2) of the prolongation operation together with properties of $*$ and $\mathscr{L}$ one can easily show that G_k is a group and that $\mathscr{G}_k$ is a Lie algebra. Now on the most general level the *symmetries of* L are those bundle maps $f: E \to E$ which permute the extremals of L around $\sigma \in \mathrm{Ext}(L) \Leftrightarrow f(\sigma) \in \mathrm{Ext}(L)$; that is, σ is a solution of the field equations if and only if $f(\sigma) = f \circ \sigma \circ \hat{f}^{-1}$ is also a solution. Thus the *complete symmetry group of* L is defined by

$$S = \mathrm{S}(L) = \{f \mid f(\mathrm{Ext}(L)) = \mathrm{Ext}(L)\}. \tag{4.3}$$

The following proposition shows that each G_k is a group of symmetries of L (with $\mathscr{G}_k$ the corresponding algebra of infinitesimal symmetries).

Proposition 1: For any f, σ, and X the following identity holds:

$$f(\sigma)^{1*}(X^1 \rfloor d\theta_L^k) = (\hat{f}^{-1})^* \circ \sigma^{1*}\left[(f^*X)^1 \rfloor f^{1*} d\theta_L^k\right]. \tag{4.4}$$

Consequently, $G_k(L)$ is a subgroup of the complete symmetry group $\mathrm{S}(L)$.

Proof: Identity (4.4) follows from Eq. (2.3) since

$$f(\sigma)^{1*} = (f^1 \circ \sigma^1 \circ \hat{f}^{-1})^* = (\hat{f}^{-1})^* \circ \sigma^{1*} \circ f^{1*},$$

$$f^{1*}(X^1 \rfloor \phi) = f^{1*} X^1 \rfloor f^{1*}\phi = (f^*X)^1 \rfloor f^{1*}\phi.$$

One can now use identity (4.4) to easily prove that $G_k(L) \subset \mathrm{S}(L)$.

Comments: The above definition describes the symmetry subgroup G_k as the group of isometries of $d\theta_L^k$ and the Lie algebra $\mathscr{G}_k$ as the corresponding algebra of Killing vectors of $d\theta_L^k$. However, the mapping properties of the Cartan forms reduce these descriptions to ones involving trivial Lagrangians. By a *trivial Lagrangian* we mean a Lagrangian L whose Euler–Lagrange equations vanish identically; more precisely, every section is an extremal of L. We denote the set of trivial Lagran-

gians by Z. In a previous paper[2] it was shown that Z is determined by the Cartan form θ_L^r—this being one of the reasons for wanting an improved version of the classical Cartan form. The other Cartan forms determine subsets of Z. Since this is central to our presentation of the symmetry theory we summarize the results on trivial Lagrangians.

(T1) $d\theta_L^k = 0$ if and only if L is trivial and has nullity k (i.e., $\partial^{k+1}L/\partial y_{i_1}^{a_1}\cdots\partial y_{i_{k+1}}^{a_{k+1}} = 0$ on each chart). Consequently $d\theta_L^k = 0$ implies that $d\theta_L^{k+1} = 0$.

(T2) Every trivial Lagrangian has nullity $r = \min(p,q)$. Consequently $Z = \{L \mid d\theta_L^r = 0\}$.

(T3) Letting $Z_k = \{L \mid d\theta_L^k = 0\}$, one has that

$$Z_0 \subset Z_1 \cdots \subset Z_r = Z. \tag{4.5}$$

In general the containments in (4.5) are proper since if $L \in Z$, then the nullity condition forces L to locally have the form

$$L = F(x,y) + F_i^a(x,y)y_i^a + \cdots (1/r!)F_{i_1\cdots i_r}^{a_1\cdots a_r}(x,y)y_{i_1}^{a_1}\cdots y_{i_r}^{a_r}. \tag{4.6}$$

Furthermore, the coefficients $F_{i_1\cdots i_k}^{a_1\cdots a_k}$ $(k \geqslant 2)$ must be antisymmetric in the upper and lower indices separately. Because of this

$$\theta_L^r = F\Delta + F_i^a\, dy^a\Delta_i + \cdots + [1/(r!)^2]\, F_{i_1\cdots i_r}^{a_1\cdots a_r}\, dy^{a_1}\cdots dy^{a_r}\,\Delta_{i_1\cdots i_r}. \tag{4.7}$$

The remaining conditions for the triviality of L are just the partial differential equations that arise from $d\theta_L^r = 0$. From this one sees that the subset Z_k of Z is characterized locally by $F_{i_1\cdots i_n}^{a_1\cdots a_n} = 0$ $(n = k+1,\ldots,r)$.

Combining the results (T1)–(T3) with the mapping properties of the Cartan form, one obtains the following theorem.

Theorem 2: Alternative characterizations of the symmetry groups and algebras of L are

$$S(L) = \{f \mid \mathrm{Ext}(f(L)) = \mathrm{Ext}(L)\}, \tag{4.8}$$

$$G_k(L) = \{f \mid f(L) - L \in Z_k\}, \tag{4.9}$$

$$\mathcal{G}_k(L) = \{X \mid X(L) \in Z_k\}. \tag{4.10}$$

Consequently because of the containments in (4.6) it follows that

$$G_{00} \subset G_0 \subset G_1 \cdots \subset G_r \subset S,$$
$$\mathcal{G}_{00} \subset \mathcal{G}_0 \subset \mathcal{G}_1 \cdots \subset \mathcal{G}_r. \tag{4.11}$$

Here G_{00} and $\mathcal{G}_{00}$ are defined by

$$G_{00} = \{f \mid f(L) - L = 0\} \quad \text{and} \quad \mathcal{G}_{00} = \{X \mid X(L) = 0\}.$$

The theorem exhibits the distinctions among the symmetry theories determined by the various Cartan forms and shows the extent to which the improved version θ_L^r of the Cartan form is more suitable than the classical version. In general, the containments in (4.11) are proper, although for particular Lagrangians there is always the possibility that some of the groups and algebras in these chains coincide. The subgroup G_{00} consists of those symmetries which leave L invariant, $f(L) = L$, and is the one most commonly encountered in the literature, primarily because these symmetries are the easiest to determine. Previous works, based on the classical Cartan form θ_L^1, led to the symmetry subgroup G_1, which is now seen to be unnecessarily restrictive. The *natural* symmetry group is G_r, which consists of those bundle maps f which leave L invariant modulo the addition of trivial Lagrangians: $f(L) = L + L'$. To illustrate the new aspects of the theory we offer the following examples.

Example 1: For the sake of simplicity assume that $N = \mathbb{R}^2$ and let $E = \mathbb{R}^2 \times \mathbb{R}^2$ be the trivial bundle over N. Then one can identify J^1E with $\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^4$. Consider the Lagrangian L defined by

$$L = L(x,y,y) = y_1^1 y_1^2 - y_2^1 y_2^2 + cx_1(y_1^1 y_2^2 - y_2^1 y_1^2), \tag{4.12}$$

where c is a constant. To see that the containments $G_0 \subset G_1 \subset G_2 \subset S$ are proper, let f,g,h be the following bundle maps of E:

$$f(x,y) = (x_1, x_2, y^1 + mx_1, y^2),$$
$$g(x,y) = (x_1 + m, x_2, y^1, y^2),$$
$$h(x,y) = (x_1, x_2, my^1, my^2),$$

where $m \neq 0$ is a constant. One finds that (1) $f(L) = L + my_1^2 + mcx_1y_2^2 = L + L'$ and L' is trivial with nullity 1; so $f \in G_1 \setminus G_0$; (2) $g(L) = L + mc(y_1^1 y_2^2 - y_2^1 y_1^2) = L + L'$ and L' is trivial with nullity 2; so $g \in G_2 \setminus G_1$; (3) $h(L) = m^2 L$, so that $h \in S \setminus G_2$. Note also that the Euler–Lagrange equations for the extremals $\sigma = (\sigma^1, \sigma^2)$ are $\sigma_{x_1x_1}^a - \sigma_{x_2x_2}^a + (-1)^a c\sigma_{x_2}^a = 0$, $a = 1,2$. Thus one sees that $f(\sigma)$, $g(\sigma)$, and $h(\sigma)$ are also extremals, which gives an alternative verification that f, g, and h are symmetries of L. The example demonstrates the necessity of using the extended Cartan form θ_L^2 to determine the natural symmetries of L: in the previous theory which uses θ_L^1 the bundle map g does not classify as a symmetry, $g^{1*}(d\theta_L^1) \neq d\theta_L^1$, when in fact it should be so classified.

Example 2: The Lagrangians of interest in elementary particle theory are quadratic Lagrangians. To keep things simple we just consider a trivial bundle $E = \mathbb{R}^p \times \mathbb{R}^q$ over $N = \mathbb{R}^p$. A quadratic Lagrangian $L: \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^{pq} \to \mathbb{R}$ then has the form

$$L(x,y,y') = A(x,y) + A_i^a(x,y)\, y_i^a + A_{ij}^{ab}(x,y)\, y_i^a y_j^b, \tag{4.13}$$

where we assume, without loss of generality, that $A_{ij}^{ab} = A_{ji}^{ba}$. The determination of the infinitesimal symmetries X of L proceeds as follows. Using the notation in Eqs. (3.17) and (3.18) one finds that

$$X(L) = \mathcal{L}_{X^1}(L) + \mathrm{div}(\xi)L = \mathcal{L}_X(L) + \zeta_k^c \frac{\partial L}{\partial y_k^c}$$
$$+ \mathrm{div}(\xi)L = F + F_i^a y_i^a + F_{ij}^{ab} y_i^a y_j^b,$$

where

$$F = \mathscr{L}_X(A) + \operatorname{div}(\xi)A + A^c_k \frac{\partial \eta^c}{\partial x_k},$$

$$F^a_i = \mathscr{L}_X(A^a_i) + \operatorname{div}(\xi)A^a_i + 2A^{ca}_{ki}\frac{\partial \eta^c}{\partial x_k} + A^c_i \frac{\partial \eta^a}{\partial y^c} - A^a_k \frac{\partial \xi_i}{\partial x_k},$$

$$F^{ab}_{ij} = \mathscr{L}_X(A^{ab}_{ij}) + \operatorname{div}(\xi)A^{ab}_{ij} + 2\left[A^{cb}_{ij}\frac{\partial \eta^c}{\partial y^a} - A^{ab}_{kj}\frac{\partial \xi_i}{\partial x_k}\right].$$

Now X is an infinitesimal symmetry of L if and only if $X(L)$ is a trivial Lagrangian, and so substituting the above expression for $X(L)$ into the Euler–Lagrange equations gives a very explicit (yet complex) system of partial differential equations which the components ξ_i, η^a of X must satisfy. In the simplest case, for $X \in \mathscr{G}_{00}$ [i.e., $X(L) = 0$], this system reduces to the first-order system, $F = 0, F^a_i = 0, F^{ab}_{ij} = 0$. For the cases $X \in \mathscr{G}_0$ or $X \in \mathscr{G}_1$ the system is simplified by the auxiliary equations $F^a_i = 0$, $F^{ab}_{ij} = 0$, or $F^{ab}_{ij} = 0$, respectively. Also note that since $X(L)$ has nullity 2 for any X, it follows that $\mathscr{G}_2 = \mathscr{G}_3 = \cdots = \mathscr{G}_r$. The example illustrates the distinctions among the various types of infinitesimal symmetries for quadratic Lagrangians.

Conservation laws: The conservation laws associated with the infinitesimal symmetries of L (Noether's theorem and its generalizations) are easily derived from the Euler–Lagrange equations as we have formulated them in (EL2) or (EL3).

Theorem 3: For $X \in \mathscr{G}_{00}(L)$ the associated conservation law is

$$d\left[\sigma^{1*}(X^1 \lrcorner \theta^1_L)\right] = 0.$$

Equivalently,

$$\operatorname{div}(W(\sigma,X,L)) = 0. \tag{4.14}$$

More generally, for $X \in \mathscr{G}_k(L)$ the associated (local) conservation law is

$$d\left[\sigma^{1*}(X^1 \lrcorner \theta^1_L - \omega)\right] = 0.$$

Equivalently,

$$\operatorname{div}(W(\sigma,X,L) - \Omega) = 0. \tag{4.15}$$

These equations [(4.14) and (4.15)] hold for every extremal σ of L. In Eq. (4.15) ω is a (local) $(p-1)$-form on J^1E such that $d\omega = \theta^k_{X(L)}$ and $\Omega = \Omega(\sigma,X,L)$ is a (local) vector field on N such that $\Omega \lrcorner \Delta = \sigma^{1*}\omega$ [and consequently $\operatorname{div}(\Omega) = X(L) \circ \sigma^1$].

Proof: If $X(L) = 0$ then Eqs. (4.14) clearly follow from the Euler–Lagrange equations (EL2) and (EL3). More generally if $X(L)$ is any trivial Lagrangian of nullity k, then Eqs. (4.15) follow from the fact that any trivial Lagrangian can be expressed as a "divergence." More precisely, since $d\theta^k_{X(L)} = 0$ Poincaré's lemma gives the local existence of a form ω such that $d\omega = \theta^k_{X(L)}$. Taking pullbacks gives $[X(L) \circ \sigma^1]\Delta = d\sigma^{1*}\omega = d(\Omega \lrcorner \Delta) = \operatorname{div}(\Omega)\Delta$. Using this to rewrite (EL2) and (EL3), one obtains Eqs. (4.15).

Versions of Noether's theorem based on the classical Cartan form θ^1_L have been used extensively in the literature [cf. Refs. 1(e) and 1(g)] but Theorem 3 shows that for more general symmetries $X \in \mathscr{G}_k(L) \setminus \mathscr{G}_1(L)$, the corresponding conserved current $\lambda = (X^1 \lrcorner \theta^1_L) - \omega$ depends on the extended Cartan form θ^k_L via the solution of $d\omega = \theta^k_{X(L)}$ for ω. Also note that $d\lambda = -(X^1 \lrcorner d\theta^k_L)$ and so the division of λ into the parts $X^1 \lrcorner \theta^1_L$ and ω, while serving to illustrate the connection with the classical Noether theorem [Eqs. (4.14)], is in some respects rather artificial.

V. CONTACT EQUIVALENCE

A simplification of the symmetry theory is afforded by the notion of contact equivalent Lagrangians. The idea is that different Lagrangians can lead to isomorphic extremal sets and symmetry theories, and thus either Lagrangian (preferably the simpler one) can be used to model the physics. Various notions of equivalence have been studied in the literature, but since the Cartan forms play a prominent role in the variational theory, it is natural to base the equivalence on these forms.

Definition: Two Lagrangians K and L are called *contact equivalent* if there exists a bundle map $h: E \to E$ such that

$$d\theta^r_K = h^{1*}(d\theta^r_L). \tag{5.1}$$

Using the mapping property in Theorem 1 together with the characterization of trivial Lagrangians, one easily sees that condition (5.1) is equivalent to

$$K = h(L) + L_0 \tag{5.2}$$

for some trivial Lagrangian L_0.

Theorem 4: Suppose that K and L are contact equivalent Lagrangians with h as in Eq. (5.1). Then h induces an isomorphism between the respective extremal sets, symmetry groups, and conservation laws for K and L. Specifically,

(i) $h^{-1}(\operatorname{Ext}(L)) = \operatorname{Ext}(K)$;

(ii) $h^{-1}S(L)h = S(K)$,
$h^{-1}G_r(L)h = G_r(K)$,
$h^*(\mathscr{G}_r(L)) = \mathscr{G}_r(K)$;

and (iii) the identity $\sigma^{1*}(\lambda) = \hat{h}^{-1*} \circ [h^{-1}(\sigma)]^{1*} \times [h^{1*}(\lambda)]$, which holds for all σ and λ, establishes the relationship between the conservation laws for K and L, i.e.,

$$h^{1*}(\operatorname{Cons}(L)) = \operatorname{Cons}(K).$$

Proof: The proof of (i) follows from identity (4.4) with $f = h$ and $k = r$. Next, by (i) $(h^{-1} \circ f \circ h)(\operatorname{Ext}(K)) = h^{-1}[f(\operatorname{Ext}(L))]$; hence $f \in S(L)$ iff $h^{-1} \circ f \circ h \in S(K)$. Similarly, by (5.1) $(h^{-1} \circ f \circ h)^{1*}(d\theta^r_K) = h^{1*} \circ f^{1*} \circ (h^{1*})^{-1}(d\theta^r_K) = h^{1*} \circ f^{1*}(d\theta^r_L)$; hence $f \in G_r(L)$ iff $h^{-1} \circ f \circ h \in G_r(K)$. The corresponding result at the infinitesimal level follows from looking at flows; namely, if f_t is the flow for X, then it is well known that $h^{-1} \circ f_t \circ h$ is the flow for $h^*(X)$. Hence $h^*(X) \in \mathscr{G}_r(K)$ iff $h^{-1} \circ f_t \circ h \in G_r(K)$ for every t iff $f_t \in G_r(L)$ for every t iff $X \in \mathscr{G}_r(L)$. This proves (ii). Finally the proof of (iii) follows from the various functorial properties,

$$\sigma^{1*} = [h(h^{-1}(\sigma))]^{1*} = [h^{1}\circ(h^{-1}(\sigma))^{1}\circ\hat{h}^{-1}]^{*}$$
$$= \hat{h}^{-1*}\circ[h^{-1}(\sigma)]^{1*}\circ h^{1*}.$$

Comments: The theorem shows that the complete symmetry groups $S(K)$, $S(L)$, and the natural symmetry groups $G_r(K), G_r(L)$ are conjugate subgroups of $B(E)$. The same is not necessarily true for $G_k(K)$, $G_k(L)$, $k = 00,0,1,...,r-1$. Indeed suppose that $L_0 = K - h(L)$ has nullity s (take $s = 00$ if $L_0 = 0$). Then it is easy to show that $h^{-1}G_k(L)h = G_k(K - L_0)$. Furthermore $G_k(K - L_0) = G_k(K)$ for $k = s+1,...,r$. These results again illustrate that θ^r is the best choice for the Cartan form, the other Cartan forms giving unnatural and incomplete symmetry theories.

Ideally one would wish to classify the set of Lagrangians on J^1E by exhibiting a set of canonical (representative) Lagrangians, one from each contact equivalence class, and perhaps also to have a procedure for reducing a given Lagrangian to its canonical form by a sequence of operations consisting of bundle transformations and subtraction of trivial Lagrangians. At present it seems unlikely that such ambitious goals can be achieved in general. We offer the following example to illustrate the difficulties involved for the important restricted class of quadratic Lagrangians.

Example 3: We return to the trivial bundle case in example 2, and for notational convenience consider $y' = \{y_i^a\}$ as a point in $\mathbb{R}^{pq}$ and let $\langle\ ,\ \rangle$ denote the standard inner product on $\mathbb{R}^{pq}$. Then suppressing the x, y dependence, we rewrite the quadratic Lagrangian L in (4.13) as

$$L = A + \langle A_1, y'\rangle + \langle A_2 y', y'\rangle,$$

where $A_1 = \{A_i^a\} \in \mathbb{R}^{pq}$ and $A_2 = \{A_{ij}^{ab}\}$ is a $pq \times pq$ matrix. Using the notation from the proof of Theorem 1, the prolongation of a bundle map $h(x,y) = (\hat{h}(x),\ \phi(x,y))$ is $h^1(x,y,y') = (\hat{h}(x),\ \phi(x,y),\ M + Hy')$, where $M_i^a = (\partial\phi^a/\partial x_j)C_{ji}^{-1}$ and $H_{ij}^{ab} = (\partial\phi^a/\partial y^b)C_{ji}^{-1}$. With this notation the transformation $h(L) = (L\circ h^1)J$ is given by

$$h(L) = (A + \langle (A_1 + A_2 M), M\rangle + \langle H^t[A_1 + (A_2 + A_2^t)M], y'\rangle + \langle H^t A_2 H y', y'\rangle)J. \tag{5.3}$$

This identity illustrates an avenue for constructing a procedure to reduce L to canonical form. At present this procedure is incomplete, and so we limit the discussion to the following remarks.

Restricting attention to the case where A_2 is constant (independent of x and y), one can further assume without loss of generality that A_2 is symmetric [otherwise, by subtracting the trivial Lagrangian $\langle \frac{1}{2}(A_2 - A_2^t)y', y'\rangle$ from L, one obtains an equivalent Lagrangian with this property]. Now A_2 can always be diagonalized by an orthogonal matrix, $Q^t A_2 Q = D$, and if it is possible to do this by means of a bundle map h [say, $h(x,y) = (Cx, Sy)$ with the matrices C and S chosen so that $H^t A_2 H = D$] then L reduces to an equivalent Lagrangian of the form $h(L) = B + B_i^a y_i^a + D_i^a (y_i^a)^2$. Further reductions can then proceed from here. One can characterize the Lagrangians for which such a diagonalization is possible. For brevity here we just present the following examples of this.

(1) Classical dynamical systems ($p=1$): $L = A + A^{ab}\dot{y}^a\dot{y}^b$. The choice of $C = 1$ and $S = Q$ diagonalizes A_2 and gives an equivalent Lagrangian of the form $h(L) = B + D^a(\dot{y}^a)^2$.

(2) Single particle field ($q=1$): $L = A + A_{ij}\psi_{x_i}\psi_{x_j}$. This also diagonalizes to an equivalent Lagrangian of the form $h(L) = B + D_i(\psi_{x_i})^2$.

(3) The scalar meson field: $L = \psi_t\psi_t^* - \psi_x\psi_x^* - \psi_y\psi_y^* - \psi_z\psi_z^* + \mu^2\psi\psi^*$ involving two scalar fields ψ, ψ^*. This diagonalizes with C equal to the 4×4 identity matrix and

$$S = 2^{-1/2}\begin{bmatrix} 1 & & 1 \\ 1 & & -1 \end{bmatrix}.$$

VI. CONCLUSION

The purpose of the paper was to exhibit the utility of the extended Cartan form θ^r in the Lagrangian symmetry theory. Since there are alternative approaches to this, and indeed since it is a special case of the general geometric theory of partial differential equations (PDE's), we should perhaps include here a few additional comments for the sake of a broader perspective.

The basic geometric object for formulating an mth-order system of PDE's with q functions and p variables is the contact element manifold C^mE, where E is an appropriate manifold with $\dim(E) = p + q$ [cf. Ref. 1(e)]. Here C^mE is constructed as a fiber bundle over E, a point (contact element) in the fiber above $z \in E$ being an equivalence class $[(N,\sigma,x)]_z$ of submanifolds of E, $\sigma: N \to E$, $\sigma(x) = z$, all of which have the same mth-order contact at z. As in the jet bundle theory, each submanifold $\sigma: N \to E$ prolongs to a smooth map $\sigma^m: N \to C^mE$, and each diffeomorphism $f: E \to E$ prolongs to a diffeomorphism $f^m: C^mE \to C^mE$. A system of PDE's is modeled by a collection $H = \{H_\alpha\}_{\alpha=1}^s$ of smooth maps $H_\alpha: C^mE \to \mathbb{R}$. The solutions $\mathrm{Ext}(H)$ of this system are thus submanifolds $\sigma: N \to E$ for which $H_\alpha(\sigma^m(x)) = 0$, that is, for which $\sigma(N)$ is contained in the variety $\Omega_H = \{w \in C^mE \mid H_\alpha(w) = 0\ \forall\alpha\}$. The complete symmetry group $S(H)$ of H consists of those diffeomorphisms f for which $f\circ\sigma \in \mathrm{Ext}(H)$ for every $\sigma \in \mathrm{Ext}(H)$. Because of the identity $(f\circ\sigma)^m = f^m\circ\sigma^m$, $S(H)$ is alternatively given by the geometrically preferable characterization $S(H) = \{f \mid f^m(\Omega_H) \subset \Omega_H\}$. By the famous Lie–Bäcklund theorem[8] each contact transformation $h: C^mE \to C^mE$ is actually (when $q > 1$) the prolongation $h = f^m$ of some $f: E \to E$. Thus $S(H)$ is represented as precisely those contact transformations that leave Ω_H invariant. One can specialize to the case where $E \to N$ is a fiber bundle and H is comprised of smooth functions on the subbundle $J^mE \subset C^mE$. We recommend the new text[9] by Olver as an excellent reference for a wealth of details and history on this subject. In particular Olver presents a method for the explicit computation of $S(H)$ (at the infinitesimal level and for $E = N \times Q$, $N \subset \mathbb{R}^p$, $Q \subset \mathbb{R}^q$).

For the Lagrangian theory, one notes that each Lagrangian $L: J^mE \to \mathbb{R}$ has its Euler–Lagrange equations expressed as

a system H of PDE's on $J^{2m}E$. Then $S(H)$ is the complete symmetry group for L. Olver shows (at least the infinitesimal level) that $G(L) = \{f \mid f(L) - L \in Z\}$ is a subgroup of the complete symmetry group. Others (cf. Refs. 8 and 10) have discussed $G(L)$ as well. However, one should note that historically there has been some confusion in the literature as to the nature of Z. One (correct, but local) characterization of Z is as a set of divergences (cf. Ref. 9). The characterization here (and in Refs. 2 and 3) is preferable since it is global. [For comparison, note that the generalization of θ^r to a form on J^mE mentioned in Sec. III gives $Z = \{L \mid d\theta_L^r = 0\}$. Thus $L \in Z$ implies that $\theta_L^r = d\omega$ for some locally defined $(p-1)$-form ω on J^mE (constructed say, using the Poincaré homotopy operator). Taking pullbacks gives $L \circ \sigma^m = d\sigma^{m*}\omega$ for every $\sigma \in \Gamma(E)$. Thus locally $L = D_{x_i} W_i$ (on $J^\infty E$) where the W_i's are certain functions on J^mE and the D_{x_i}'s are the total derivative operators.]

In summary, while the variational theory can be formulated without the use of the Cartan forms, we have stressed here the naturality of the Cartan form approach, and in particular the benefit of the extended Cartan form θ^r.

[1]The Cartan form for classical dynamical systems ($p = 1$) can be traced back to (a) H. Poincaré, *Les Methods Nouvelles de la Mechanique céleste* (Dover, New York, 1957), original edition 1892, Vol. II; (b) E. Cartan, *Lecons Sur Les Invariants Integraux* (Hermann, Paris, 1922); Modern treatments of classical dynamical systems in terms of the Cartan form are numerous, e.g., (c) P. Griffiths, *Exterior Differential Systems and the Calculus of Variations* (Birkhauser, Boston, 1983); (d) W. Sarlet and F. Cantrijn, SIAM Rev. **23**, 467 (1981). Generalizations of the Cartan form to cover field theories ($p > 1$) as well can be found in (e) R. Hermann, *Geometry, Physics, and Systems* (Dekker, New York, 1973); (f) R. Hermann, *Differential Geometry and the Calculus of Variations* (Math Sci. Press, Brookline, MA, 1977); (g) H. Goldschmidt and S. Sternberg, Ann. Inst. Fourier (Grenoble) **23**, 203 (1973); (h) D. Krupka, *Some Geometric Aspects of Variational Problems in Fibered Manifolds* (J. E. Purkyne U. P., Brno, Czechoslovakia, 1973).

[2]D. Betounes, Phys. Rev. D **29**, 599 (1984).

[3]H. Rund, Lec. Notes Pure Appl. Math. **100**, 455 (1985).

[4]W. M. Tulczyjew, Lect. Notes Math. **836**, 22 (1980).

[5]I. M. Anderson (private communication).

[6]Many recent works have solved the global existence problem for higher-order Cartan forms by using connections on the base space N. See, for example, M. Ferraris, "Fibered connections and global Poincaré–Cartan forms in higher-order calculus of variations," *Proceedings of the Conference on Differential Geometry and its Applications*, Nove Mesto na Morave (Univ. Karlova, Praga, Czech., 1984),Vol. II; I. Kolár, J. Geom. Phys. **1**, 127 (1984).

[7]See Refs. 1(e)–1(h).

[8]R. L. Anderson and N. H. Ibragimov, *Lie–Bäcklund Transformations in Applications*, SIAM Studies in Applied Mathematics (SIAM, Philadelphia, 1979).

[9]P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer, New York, 1986).

[10]E. L. Hill, Rev. Mod. Phys. **23**, 253 (1951).

On the dimensional reduction of invariant fields and differential operators

P. A. Nikolov
Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Sciences, Boul. Lenin 72, Sofia 1784, Bulgaria

(Received 9 December 1986; accepted for publication 13 May 1987)

In the present paper the most general type of group action introducing some relevant exact sequences for the dimensional reduction of invariant fields and differential operators is studied.

I. INTRODUCTION

An action of a group G on a vector bundle ξ by bundle morphisms induces naturally an action of G on the sets of sections, differential operators, etc. When the G action is "good," there is a natural one-to-one correspondence between all G-invariant sections of ξ and all sections of another "reduced" vector bundle ξ_G. In principle, the G-invariant structures on ξ (like G-invariant sections of some tensor power of ξ, differential operators, the action of another group, etc.) define the corresponding reduced structures on ξ_G. In the present paper we study the reduction procedure for the most general type of group action introducing some relevant exact sequences. In Sec. II the reduced bundle ξ_G is described by means of a set of coordinate realizations and transition bundle isomorphisms. In Sec. III the bundle ξ is specified to be the tangent bundle or its tensor powers. The symmetric second tensor power of the exact sequence (3.5) reproduces the "reduction theorem" in Ref. 1. The relevant basic notions and notations about differential operators on vector bundles in terms of the jet bundles are summarized in Sec. IV. Section V deals with a crucial detail—a restriction of a differential operator on a submanifold. This is not a natural operation and the problem reduces to a splitting of the exact sequence (5.1) of jet bundles. The major problem—the dimensional reduction of a differential operator D—is considered in Sec. VI. In each coordinate bundle of the reduced bundle ξ_G we have the situation of Sec. V. Here the group action and the G invariance of D provide a splitting of the corresponding exact sequences and Propositions 6.1–6.4 assure that the restricted differential operators are compatible with the cocycle of ξ_G. The application of the general construction of ξ_G to $\mathrm{Hom}(J^k(\xi),\eta)$ leads to a description of the G-invariant linear differential operators (or intertwining differential operators) as a section of a bundle with a typical fiber—all intertwining operators between two finite-dimensional representations of the isotropy group. As a consequence we have a description in this language of all G-invariant (linear) connections on ξ. A dimensional reduction of a group action is considered in Sec. VII. As an example illustrating the discussed constructions we reexamine[2] in Sec. VIII the dimensional reduction of the SU(2) Yang–Mills equation by means of a reduction group $SL(2,\mathbb{C})$. In Sec. IX we show that the dimensional reduction of the six-dimensional Maxwell equation, followed by a restriction on the projected six-dimensional light cone by means of a $SO_0(2,4)$-invariant splitting of the corresponding exact sequence of the type (5.1) leads to the discussed in the literature "conformal electrodynamics."

II. DIMENSIONAL REDUCTION OF A G-VECTOR BUNDLE. REDUCED VECTOR BUNDLE

Consider a connected Lie group G (not necessarily compact) acting from the left on a (real or complex) vector bundle $\xi=(E,\pi,B)$ by bundle morphisms. We shall assume that all manifolds, bundles, and maps are (C^∞) smooth. Denote this action by (T,t) or by (T_g,t_g), $g\in G$, where T: $G\times E\to E$ is the action of G on E and t: $G\times B\to B$ is the projected action on B. By definition, $\pi\circ T_g=t_g\circ\pi$ and T_g: $\xi_b\to\xi_{t_g(b)}$, $\xi_b=\pi^{-1}(b)$, $b\in B$, is a linear isomorphism. The action of G on ξ induces naturally an action on $C^\infty(\xi)$—the space of all sections of ξ ("the linear matter fields") by the equation

$$g(\psi)(b)=T_g\psi t_{g^{-1}}(b)\,, \tag{2.1}$$

$\psi\in C^\infty(\xi)$, $b\in B$. A section $\psi\in C^\infty(\xi)$ is G invariant (in other terminology "G equivariant") if it is a stable point for the action (2.1),

$$g(\psi)=\psi\equiv T_g\psi(b)=\psi(t_g(b))\,,\quad g\in G\,. \tag{2.2}$$

Denote by $C^\infty(\xi)_G\subset C^\infty(\xi)$ the subspace of all G-invariant sections of ξ. Our first goal is to describe $C^\infty(\xi)_G$. In the general case this is a complicated problem. One may impose here some simplifying conditions, assuming that all orbits of the action t are of the same type (say G/H_0) and that they form a locally trivial bundle (B,p,M)

$$p\colon B\to B/G\equiv M\,, \tag{2.3}$$

where B/G is naturally a manifold and p is the natural projection. Let $G_b\subset G$ be the isotropy group of $b\in B$ for the G action t on B. The restriction T: $G_b\times\xi_b\to\xi_b$ is a linear representation. Denote by $\mathrm{st}\,\xi_b\subset\xi_b$ the subspace of all stable vectors. We shall assume that the family of linear subspaces ξ_b is a (smooth) vector subbundle $\mathrm{st}\,\xi\subset\xi$. In this case $C^\infty(\xi)_G$ has the structure of the set of all sections of another vector bundle. That is to say, one can construct a *reduced vector bundle* ξ_G over the base M ($=B/G$), and a natural one-to-one correspondence between *all sections* of ξ_G and the *G-invariant sections* of ξ. Let θ: $C^\infty(\xi_G)\to C^\infty(\xi)_G$ give this correspondence. We shall only work in this case and shall briefly say that ξ is a *reducible G-vector bundle*. The most convenient construction of ξ_G for our study is the following: Let $\{U_\alpha\}$, $\alpha\in A$, be a sufficiently fine open covering of M and for each U_α we fix a section transversal to the fibers of the bundle (2.3), σ_α: $U_\alpha\to B$. Denote by $\tilde U_\alpha=\sigma_\alpha(U_\alpha)$ the graph of σ_α, $\xi_\alpha=\mathrm{st}\,\xi_{\tilde U_\alpha}$ the restriction of $\mathrm{st}\,\xi$ on $\tilde U_\alpha$, $\tilde U_{\alpha,\beta}=\sigma_\alpha(U_\alpha\cap U_\beta)$ if $U_\alpha\cap U_\beta\neq 0$, $\xi_{\alpha,\beta}=\xi_{\alpha\tilde U_{\alpha,\beta}}$ ($=\mathrm{st}\,\xi_{U_{\alpha,\beta}}$). Let $\varphi_{\alpha,\beta}$: $U_\alpha\cap U_\beta\to G$ define a local action of

 0022-2488/87/102354-09$02.50 © 1987 American Institute of Physics

G (over $U_\alpha \cap U_\beta \neq 0$), mapping $\tilde{U}_{\beta,\alpha}$ on $\tilde{U}_{\alpha,\beta}$:

$$t_{\varphi_{\alpha\beta}(x)}(\sigma_\beta(x)) = \sigma_\alpha(x), \quad x \in U_\alpha \cap U_\beta . \tag{2.4}$$

The pairs $(T_{\varphi_{\alpha\beta}(x)}, t_{\varphi_{\alpha\beta}(x)})$, $x \in U_\alpha \cap U_\beta$, define an isomorphism $\tilde{\varphi}_{\alpha\beta}$: $\xi_{\beta,\alpha} \to \xi_{\alpha,\beta}$. This isomorphism does not depend on the freedom of choice of $\varphi_{\alpha\beta}$ from (2.4) and so it is uniquely defined for a pair $(\alpha,\beta) \in A \times A$ if $U_\alpha \cap U_\beta \neq 0$. It is crucial here that ξ_β is a restriction of the stable subbundle st ξ. The set of bundle isomorphisms $\tilde{\varphi}_{\alpha\beta}$ forms a cocycle

$$\tilde{\varphi}_{\alpha\beta} = \tilde{\varphi}_{\beta\alpha}^{-1}, \tag{2.5}$$

$$\tilde{\varphi}_{\alpha\beta} \circ \tilde{\varphi}_{\beta\gamma} = \tilde{\varphi}_{\alpha\gamma} \tag{2.6}$$

over $U_\alpha \cap U_\beta \cap U_\gamma \neq 0$. [We shall not emphasize hereafter that equations like (2.6) are considered when they are correctly defined.] The cocycle $\tilde{\varphi}_{\alpha\beta}$ defines the reduced bundle ξ_G gluing the "coordinate bundles" ξ_α. (For a general treatment of the gluing procedure in category language see Ref. 3.) A section $S \in C^\infty(\xi_G)$ corresponds to a set of sections $S_\alpha \in C^\infty(\xi_\alpha)$ compatible with the cocycle

$$S_\alpha = \tilde{\varphi}_{\alpha\beta}(S_\beta) . \tag{2.7}$$

Now the correspondence θ: $C^\infty(\xi_G) \to C^\infty(\xi)_G$ is evident. Here $S_\alpha \in C^\infty(\text{st}\,\xi_{\bar{U}_\alpha})$ and there is only one G-invariant section $\theta(S) \equiv \psi \in C^\infty(\xi)_G$ such that $\psi(\sigma_\alpha(x)) = S_\alpha(\sigma_\alpha(x))$, $x \in U_\alpha$. For $b \in B$ we can take $\psi(b) = T_g S_\alpha(\sigma_\alpha(x))$, where $x = p(b) \in U_\alpha$ for some $\alpha \in A$ and $g \in G$ satisfies $t_g(\sigma_\alpha(x)) = b$. Due to (2.7) the definition of ψ is correct. If ψ is G invariant, from (2.2) $\psi \in C^\infty(\text{st}\,\xi)$ and the restrictions $S_\alpha = \psi_{\bar{U}_\alpha}$ satisfy (2.7).

In coordinates (b,u) of the bundle ξ, the group action is

$$T_g(b,u) = (t_g(b), T(g,b)u), \tag{2.8}$$

where $T(g,b) \in \mathrm{GL}(n)$ ($n = \dim \xi$) and satisfies

$$T(g_1 g_2, b) = T(g_1, t_{g_2}(b)) \circ T(g_2, b) . \tag{2.9}$$

Sometimes the G action t on B is given and it is a problem to lift it to a bundle morphism action (T,t) on ξ. Throughout this paper we shall assume that this problem is solved and is a part of the initial condition.

When the bundle morphism action on ξ is of a special kind, some structures arise in the reduced bundle.

III. DIMENSIONAL REDUCTION OF THE TANGENT AND COTANGENT BUNDLES AND THEIR TENSOR POWERS

We specify here that ξ is the tangent bundle $T(B)$ or the cotangent bundle $T^*(B)$ or $\otimes^k T(B)$, $S^k T(B)$, $\Lambda^k T(B)$ [the k th tensor, symmetric, and antisymmetric tensor power of $T(B)$]. The manifold B is equipped with a G action t satisfying the usual assumptions for the orbits plus a more specific one, namely, we want that for each point $b \in B$ there is a local cross section $\tilde{U}$ of the bundle (2.3) such that $b \in \tilde{U}$ and for each $b' \in U$, $G_{b'} = G_b$. When G is compact, this is always the case because the bundle (2.3) has a Lie structure group.[4] The lifted bundle morphism action (t_*, t) of G on $T(B)$ makes $T(B)$ a reducible G-vector bundle. The major feature of this case is that $T(B)_G$ and $T(M)$ are involved in an exact sequence.

Lemma 3.1[5]: There is a natural exact sequence

$$0 \to \tau \xrightarrow{i} T(B)_G \xrightarrow{j} T(M) \to 0 , \tag{3.1}$$

where $\tau = (T^v(B))_G$, $T^v(B) \subset T(B)$ is the subbundle of the vertical vectors.

We have the following theorem.

Theorem 3.2[5]: There is a natural one-to-one correspondence between all splittings of (3.1) and the *G-invariant connections* on the bundle (2.3).

The space of all splittings of (3.1) is an affine space with a linear group $\mathrm{Hom}(T(M),\tau) = C^\infty(T^*(M) \otimes \tau)$. The G-invariant connections of the bundle (2.3) are the equipments of $T(B)$ with a G-invariant "horizontal" subbundle $T^h(B) \subset T(B)$, complementary to the vertical one, $T(B) = T^v(B) \oplus T^h(B)$. Because (B,p,M) is a locally trivial bundle with a given structure on the fibers (the structure of a homogeneous space G/H_0), it can be considered as a bundle associated with a principal bundle $P(K,M)$ over M with a structure group $K = \mathrm{Aut}(G/H_0) = N(H_0)/H_0$, where $N(H_0)$ is the normalizer of H_0 in G. The parallel transport of the G-invariant connections preserves the structure in the fibers and so they correspond to connections on the principal bundle $P(K,M)$ (see, for example, Ref. 1).

If we choose a splitting of (3.1),

$$T(B)_G = \tau \oplus T(M) , \tag{3.2}$$

then one can say that the G-invariant vector fields on B [the sections of $T(B)_G$] correspond to pairs of a scalar field (a section of τ) and a vector field on M.

For a free G action t the calculation of $(\otimes^k T(B))_G$, $(S^k T(B))_G$, $(\Lambda^k T(B))_G$ is purely algebraic. For example, after a splitting of (3.1) we have

$$(S^k T(B))_G = S^k(T(B)_G) = S^k(\tau \oplus T(M)) = \bigoplus_{l=0}^{k} S^l\tau \otimes S^{k-l}T(M) . \tag{3.3}$$

Equation (3.3) can be interpreted as a correspondence between all G-invariant k-fold symmetric tensor fields and the $(k+1)$-ples consisting of a scalar field (a section of $S^k\tau$), a vector field on M with coefficients in $S^{k-1}\tau$,..., a k-fold symmetric tensor field on M.

If the G action t is not free, in principle, we only have an inclusion $S^k(T(B)_G) \to (S^k T(B))_G$, but then Eq. (3.3) reads

$$(S^k T(B))_G = \bigoplus_{l=0}^{k} \widetilde{S^l\tau} \otimes S^{k-l}T(M) , \tag{3.4}$$

where $\widetilde{S^l\tau} = (S^l T^v(B))_G$ ($\neq S^l\tau$).

For the cotangent bundle $T^*(B)$ we have the dual exact sequence

$$0 \leftarrow \tau^* \xleftarrow{i^*} T^*(B)_G \xleftarrow{j^*} T^*(M) \leftarrow 0 , \tag{3.5}$$

where $\tau^* = (T^v(B))^*_G$, and, respectively, the dual theorem (Theorem 3.2) states that there is a natural one-to-one correspondence between all splittings of (3.5) and the G-invariant connections on the bundle (2.3).

Finally we consider in this language the dimensional reduction of $S^2T^*(B)$ because this case contains the dimensional reduction of G-invariant metrics discussed in the literature[1] and is in some sense degenerate. After splitting of (3.5) we have

$$(S^2T^*(B))_G = \widetilde{S^2\tau^*} \oplus \tau^* \otimes T^*(M) \oplus S^2T^*(M) , \tag{3.6}$$

where $\widetilde{S^2\tau^*} = (S^2T^v(B))^*_G$. This is a correspondence between all G-invariant symmetric bilinear forms ψ on $T(B)$ and the triples $(\psi_{2,0}, \psi_{1,1}, \psi_{0,2})$, where $\psi_{2,0}$ is a scalar field [a section of $\widetilde{S^2\tau^*}$; $\psi_{2,0}(x)$ is a G-invariant bilinear symmetric form on $T(p^{-1}(x))$, $x \in M$], $\psi_{1,1}$ is a one-form on M with coefficients in τ^*, and $\psi_{0,2}$ is a symmetric bilinear form on $T(M)$. We need a condition on $(\psi_{2,0}, \psi_{1,1}, \psi_{0,2})$ assuring that ψ is nondegenerate. In the Euclidean case this condition is simple. Here $\psi_{2,0}$ and $\psi_{0,2}$ are always nondegenerate if ψ is nondegenerate. Here $\psi_{2,0}$ provides an isomorphism $\tau^* = \tau$ and hence an isomorphism $C^\infty(\tau^* \otimes T^*(M)) = C^\infty(\tau \otimes T^*(M)) = \mathrm{Hom}(\tau^*, T^*(M))$. But $C^\infty(\tau \otimes T^*(M))$ parametrizes all the splittings of (3.5) and according to the dual Theorem 3.2, all the G-invariant connections on the bundle (B,p,M). We obtain the reduction theorem[1]; in the Euclidean case there is a one-to-one correspondence between all G-invariant metrics on B and the triples $(\psi_{2,0}, \psi_{1,1}, \psi_{0,2})$, where $\psi_{2,0}$ is a scalar field (non degenerate section of $\widetilde{S^2\tau^*}$), $\psi_{1,1} \in C^\infty(\tau \otimes T^*(M))$ is a G-invariant connection on (B,p,M) and $\psi_{0,2}$ is a metric on M.

In the pseudo-Euclidean case the same conditions on $(\psi_{2,0}, \psi_{1,1}, \psi_{0,2})$ do not describe all the G-invariant metrics on B. For example, let us consider $B = R^2 \setminus \{0\}$, $G = R^* = R \setminus \{0\}$ acting by multiplications. Then $\psi = [1/(x^2+y^2)]\,(dx \otimes dx - dy \otimes dy)$ is R^* invariant but the corresponding field $\psi_{2,0}$ on PR^2 is degenerate.

IV. DIFFERENTIAL OPERATORS ON VECTOR BUNDLES

We summarize here, following Refs. 6 and 7, the basic notions and notations about the differential operators on vector bundles, which we shall need.

Let ξ, η be vector bundles over B. Denote by $J^k(\xi) = (E^k, \pi^k, B)$ the k-jet bundle of ξ. The fiber of $J^k(\xi)$ over a point $b \in B$ is the quotient of the space of germs of sections of ξ at b by the subspace of germs vanishing to order $k+1$ at b. So the elements of $J^k(\xi)_b$ are the coordinate free notion of "the field ψ and its derivatives up to order k at the point b." Denote $E^0 = E$, $E^{-1} = B$ and let $\pi^{k,l}$: $J^k(\xi) \to J^l(\xi)$, $k > l \geqslant 0$, be the natural projections, and J^k: $C^\infty(\xi) \to C^\infty(J^k(\xi))$ be the k-jet lifting of the sections of ξ. Coordinates (x^μ, z^a), $\mu = 1,2,\ldots,\dim B$, $a = 1,2,\ldots,\dim \xi$, of ξ induce coordinates $(x, z^a, z^a_{\mu_1}, \ldots, z^a_{\mu_1 \cdots \mu_k})$ of $J^k(\xi)$, $1 \leqslant \mu_1 \leqslant \cdots \leqslant \mu_i \leqslant \dim B$, $i = 1,2,\ldots,k$, where

$$z^a_{\mu_i \mu_2 \cdots \mu_i}(J^k(\psi)_b) = \frac{\partial^i}{\partial x^{\mu_1}\, \partial x^{\mu_2} \cdots \partial x^{\mu_i}} \psi(b)\,. \tag{4.1}$$

So $\dim J^k(\xi) = \dim \xi \binom{\dim B + k}{k}$.

There is a natural morphism

$$i: S^k T^*(B) \otimes \xi \to J^k(\xi)\,, \tag{4.2}$$

and the sequences

$$0 \to S^k T^*(B) \otimes \xi \xrightarrow{i} J^k(\xi) \xrightarrow{\pi^{k,k-1}} J^{k-1}(\xi) \to 0\,, \tag{4.3}$$

$k = 1,2,\ldots$, are exact.

A linear differential operator D: $C^\infty(\xi) \to C^\infty(\eta)$ of order k may be identified with a vector bundle morphism $\widetilde{D}$: $J^k(\xi) \to \eta$ or, equivalently, with a section of $L(J^k(\xi), \eta)$. So we have the isomorphisms $\mathrm{LDiff}_k(\xi,\eta) = \mathrm{Hom}(J^k(\xi), \eta) = C^\infty(L(J^k(\xi)\eta))$ [$\mathrm{LDiff}_k(\xi,\eta)$ denotes the space of all linear differential operators D: $C^\infty(\xi) - C^\infty(\eta)$ of order up to k]. The main symbol $\sigma(D)$ is the composition

$$\sigma(D) = \widetilde{D} \circ i: S^k T^*(B) \otimes \xi \to \eta\,. \tag{4.4}$$

Because $\mathrm{Hom}(S^k T^*(B) \otimes \xi, \eta) = C^\infty(P^k(T^*(B), L(\xi,\eta)))$, we can consider $\sigma(D)$ at $b \in B$ as a homogeneous polynomial of degree k from $T^*(B)_b$ to $L(\xi_b, \eta_b)$. To calculate $\sigma(D)(b,p)(e)$, $(b,p) \in T^*(B)_b$, $e \in \xi_b$, one must take $\psi \in C^\infty(\xi)$, $\psi(b) = e$, $f: B \to R$ with $df_b = p$ and then

$$\sigma(D)(b,p)(e) = D((1/k!)(f - f(b))^k \cdot \psi)(b)\,. \tag{4.5}$$

For $k = 1$ from (4.3) we have

$$0 \to T^*(B) \otimes \xi \xrightarrow{i} J^1(\xi) \xrightarrow{\pi^{1,0}} \xi \to 0\,. \tag{4.6}$$

Any splitting of (4.6) is equivalent to a linear connection on ξ. A splitting morphism $\widetilde{\nabla}: J^1(\xi) \to T^*(B) \otimes \xi$, $\widetilde{\nabla} \circ i = \mathrm{id}$ corresponds to a linear differential operator (a covariant derivative) ∇: $C^\infty(\xi) \to C^\infty(T^*(B) \otimes \xi)$ satisfying

$$\nabla(f \cdot \psi) = df \otimes \psi + f \cdot \nabla(\psi), \quad f \in C^\infty(B)\,. \tag{4.7}$$

A splitting of (4.6) can also be given by a morphism S: $\xi \to J^1(\xi)$ satisfying $\pi^{0,1} \circ S = \mathrm{id}$. So the set of all linear connections on ξ coincides with the set of all linear differential operators in $\mathrm{LDiff}_1(\xi, T^*(B) \otimes \xi)$ with a main symbol $\sigma(\nabla)(b,p)(e) = p \otimes e$, or with the set of all sections of $\xi^* \otimes J^1(\xi)$ satisfying $\pi^{0,1} \circ S = \mathrm{id}$ and is an affine space with a linear group $\mathrm{Hom}(\xi, T^*(B) \otimes \xi) = C^\infty(T^*(B) \otimes L(\xi,\xi))$.

We can differentiate simultaneously both sides of the equation $D(\psi) = \varphi$. In the coordinate free language this is a prolongation of D. The l th prolongation $p^l(D)$ of a differential operator D: $C^\infty(\xi) \to C^\infty(\eta)$ is the unique morphism $p^l(D)$: $J^{k+l}(\xi) \to J^l(\eta)$ such that the following diagram is commutative:

$$\begin{array}{ccc} C^\infty(J^{k+l}(\xi)) & \xrightarrow{p^l(D)} & C^\infty(J^l(\eta)) \\ \uparrow J^{k+l} & & \uparrow J^l \\ C^\infty(\xi) & \xrightarrow{D} & C^\infty(\eta) \end{array}\,. \tag{4.8}$$

We set $R^{k,l} = \ker p^l(D)$. In the general case $R^{k,l}$ is a family of linear subspaces of the bundle $J^{k+l}(\xi)$. One says that a linear differential operator D: $C^\infty(\xi) - C^\infty(\eta)$ is *formally integrable* if for $l \geqslant 0$, $R^{k,l} \subset J^{k+l}(\xi)$ is a vector subbundle and $\pi^{k+l+1,k+l}$: $R^{k,l+1} \to R^{k,l}$ is an epimorphism. For formally integrable operators the subbundle $R^{k,0} = \ker \widetilde{D}$ is called its *equation*.

A nonlinear differential operator of order k [$D \in \mathrm{Diff}_k(\xi,\eta)$], will be identified with a fiber preserving map D: $J^k(\xi) \to \eta$.

Comment 4.1: A differential operator may be used as an equation or as a field. The operators D and $f \cdot D$, $f \in C^\infty(B)$, $f \neq 0$, have the same equations. A typical example for the second role is the gauge field (the linear connections). It will be convenient for the dimensional reduction of the gauge field (in Sec. VIII) to consider the linear connections on as sections $S \in C^\infty(\xi^* \otimes J^1(\xi))$ satisfying $\pi^{1,0} \circ S = \mathrm{id}$.

V. RESTRICTION OF A DIFFERENTIAL OPERATOR

Let $\xi,\eta,\ldots$ be vector bundles over B, $D: C^\infty(\xi)\to C^\infty(\eta)$ be a differential operator (not necessarily linear), and $N\subset B$ be a submanifold. Denote by $i: N\to M$ the natural embedding and by ξ_N, η_N ($=i^*\xi$, $i^*\eta$), ... the restrictions on N. The operator D does not define naturally an operator $C^\infty(\xi_N)\to C^\infty(\eta_N)$. One needs additional information. Let $I^k_N\subset J^k(\xi)_N$ be the subbundle of all jets of sections of ξ vanishing on N and $i: I^k_N\to J^k(\xi)_N$ be the natural embedding. There is a natural bundle morphism j: $J^k(\xi)_N\to J^k(\xi_N)$ determined by the restriction on N $j(J^k(\psi)(b))=J^k(\psi\circ i)(b)$, $b\in N$. The following sequence is exact:

$$0\to I^k_N \xrightarrow{i} J^k(\xi)_N \xrightarrow{j} J^k(\xi_N)\to 0\,. \tag{5.1}$$

A differential operator D is *internal for* N if the map D: $J^k(\xi)_N\to\eta_N$ may go through j, i.e., there is a fiber preserving map $\tilde\varphi$: $J^k(\xi_N)\to\eta_N$ such that

$$\tilde D=\tilde\varphi\circ j\,. \tag{5.2}$$

In this case the value $D(\psi)(b)$, $b\in N$, depends only on the restriction of ψ on N and we have a correctly defined differential operator $C^\infty(\xi_N)\to C^\infty(\eta_N)$ with a total symbol $\tilde\varphi$: $J^k(\xi_N)\to\eta_N$. If the operator is not internal for N, we do not have a natural restriction. Actually we need a bundle preserving map S: $J^k(\xi_N)\to J^k(\xi)_N$ satisfying $S\circ j=\mathrm{id}$. As a restricted operator we can take

$$\tilde D_N=\tilde D\circ S: J^k(\xi_N)\to\eta_N\,. \tag{5.3}$$

The case when S is a bundle morphism is equivalent to a splitting of (5.1) and so any splitting of (5.1) defines a restricted operator D_N: $C^\infty(\xi_N)\to C^\infty(\eta_N)$ by means of (5.3).

VI. DIMENSIONAL REDUCTION OF INVARIANT DIFFERENTIAL OPERATORS

Let $\xi,\eta,\ldots$ be reducible G-vector bundles over B and have the same G action t on B. A differential operator D: $C^\infty(\xi)\to C^\infty(\eta)$ is G invariant if

$$D(g(\psi))=g(D(\psi))\,, \tag{6.1}$$

$\psi\in C^\infty(\xi)$, $g\in G$ [g denotes the action on $C^\infty(\xi)$ and on $C^\infty(\eta)$]. The G invariance of D provides, by means of the correspondence θ: $C^\infty(\xi_G)\to C^\infty(\xi)_G$, an operator D_G: $C^\infty(\xi_G)\to C^\infty(\eta_G)$, the *reduced operator*. We want an explicit form of D_G on the coordinate bundles ξ_α of ξ_G or, more exactly, for any two coordinate bundles ξ_α, η_α a differential operator $D_\alpha: C^\infty(\xi_\alpha)\to C^\infty(\eta_\alpha)$ compatible with the cocycles $\tilde\varphi_{\alpha\beta}$ (of ξ_G and η_G). Because $\xi_\alpha=\mathrm{st}\,\xi_{\tilde U_\alpha}$, we have the situation described in Sec. V and must consider the corresponding exact sequence

$$0\to I^k_\alpha \xrightarrow{i} J^k(\xi)_\alpha \xrightarrow{j} J^k(\xi_\alpha)\to 0\,, \tag{6.2}$$

where $J^k(\xi)_\alpha=J^k(\mathrm{st}\,\xi)_{\tilde U_\alpha}$. Now the G action on ξ provides a natural splitting $S_\alpha: J^k(\xi_\alpha)\to J^k(\xi)_\alpha$ of (6.2) corresponding to the unique extension of a section of ξ_α to a G-invariant section of ξ. Then $\tilde D_\alpha=\tilde D\circ S_\alpha: J^k(\xi_\alpha)\to\eta_\alpha$. The following propositions rule the calculation and the properties of D_α.

Let L: $C^\infty(\xi)\to\mathfrak{g}^*\otimes_R C^\infty(\xi)$ be the Lie derivative of the G action on ξ. Here $\mathfrak{g}^*$ is the dual Lie algebra of G. For $a\in\mathfrak{g}$ and $\psi\in C^\infty(\xi)$ we have

$$L(\psi)(a)=L_a(\psi)\,, \tag{6.3}$$

where

$$L_a(\psi)=\frac{d}{dt}\exp(ta)(\psi)_{t=0}\,. \tag{6.4}$$

Proposition 6.1: The first-order differential operator L is formally integrable for st ξ. [$L: C^\infty(\mathrm{st}\,\xi)\to\mathfrak{g}^*\otimes_R C^\infty(\mathrm{st}\,\xi)$ is correctly defined because st ξ is a G-invariant subbundle.]

Let $\{a_i\}$ be a basis in $\mathfrak{g}$ and Z_i be the corresponding fundamental vector fields on the total space of st ξ. In coordinates (x^μ,z^a) of st ξ they have the form

$$Z_i(x,z)=X^\mu_{i,}(x)\frac{\partial}{\partial x^\mu}+Y^a_{i,b}(x)z^b\frac{\partial}{\partial z^a}\,. \tag{6.5}$$

The equations

$$L_i(\psi)^a=0\equiv -X^\mu_{i,}(x)z^a_\mu+Y^a_{i,b}(x)z^b=0\,, \tag{6.6}$$

$i=1,2,\ldots,\dim G$, $a=1,2,\ldots,\dim\mathrm{st}\,\xi$, have a constant rank with respect to the variables z^a in a neighborhood of each point $b\in B$ due to the assumptions for the G action on ξ, and this assures the formal integrability of L on st ξ.

Let $R^{1,k-1}=\ker p^{k-1}(L)$, [$\subset J^k(\mathrm{st}\,\xi)$] and $R^k_\alpha=R^{1,k-1}{}_{\tilde U_\alpha}$.

Proposition 6.2: R^k_α is a transversal to I^k_α [and hence defines a splitting of (6.2) or, equivalently, a bundle morphism $S_\alpha: J^k(\xi_\alpha)\to J^k(\xi)_\alpha$ satisfying $S_\alpha\circ j_\alpha=\mathrm{id}$].

In a neighborhood of each point $b\in\tilde U_\alpha$ there are coordinates (x^μ,z^a) of st ξ adapted to the G action t on B; $x^\mu=(x^\nu,x^\rho)$, $\nu=1,2,\ldots,\dim M$, $\rho=\dim M+1,\ldots,\dim B$, $x^\rho(b')=0$ for $b'\in\tilde U_\alpha$, $x^\nu(t_g(b'))=x^\nu(b')$, $g\in G$. Due to the transversality of $\sigma_\alpha: U_\alpha\to B$, Eq. (6.6) can be solved with respect to z^a_ρ:

$$z^a_\rho=z^a_\rho(x^\nu,z^b,z^c_\nu)\,,\quad x\in\tilde U_\alpha\,. \tag{6.7}$$

The $(k-1)$-jet lifting of (6.7) defines a bundle morphism

$$S_\alpha: J^k(\xi_\alpha)\to J^k(\xi)_\alpha\,, \tag{6.8}$$

giving a splitting of (6.2). Here $S_\alpha(J^k(\xi_\alpha))=R^k_\alpha$ is transversal to I^k_α.

Proposition 6.3: For a G-invariant section ψ of ξ, $(D(\psi))_\alpha=\tilde D(S_\alpha(J^k(\psi_\alpha)))$ [$\equiv D_\alpha(\psi_\alpha)$].

We must show that $J^k(\psi)(b)=S_\alpha(J^k(\psi_\alpha)(b))$, $b\in\tilde U_\alpha$. But from the G invariance of ψ, $p^{k-1}(L)(\psi)=0$, $J^k(\psi)(b)\in(R^k_\alpha)_b$, $R^k_\alpha=S_\alpha(J^k(\xi_\alpha))$ and so $J^k(\psi)(b)=S_\alpha(J^k(\psi_\alpha)(b))$.

Proposition 6.4: Let ξ,η be reducible G-vector bundles over B with the same G action t on B and let D: $C^\infty(\xi)\to C^\infty(\eta)$ be a G-invariant differential operator. Then

$$D_\alpha(\psi_\alpha)=\tilde\varphi_{\alpha\beta}D_\beta\tilde\varphi^{-1}_{\alpha\beta}(\psi_\alpha)\,. \tag{6.9}$$

Indeed for each $b\in\tilde{U}_\alpha$, $D_\alpha(\psi_\alpha)(b)=D(\psi)(b)$,

$$(\tilde{\varphi}_{\alpha\beta}D_\beta\tilde{\varphi}_{\alpha\beta}^{-1})(\psi_\alpha)(b)$$
$$=(\tilde{\varphi}_{\alpha\beta}D_\beta(\psi_\beta))(b)=T_g(D_\beta(\psi_\beta))(t_{g^{-1}}(b))$$
$$=T_gD(\psi)(t_{g^{-1}}(b))=D(\psi)(b),$$
$$g=\varphi_{\alpha\beta}(x)\in G,\ x=p(b).$$

That is, the operators D_α, $\alpha\in A$, are compatible with the cocycles of ξ_G and η_G and define the reduced operator D_G: $C^\infty(\xi_G)\to C^\infty(\eta_G)$.

Comment 6.1: Description of all G-invariant linear differential operators.

It may happen that two different G-invariant operators D_1 and D_2 have the same reduced operator, $(D_1)_G=(D_2)_G$. The set $\mathrm{LDiff}_k(\xi_G,\xi_G)$ does not describe all G-invariant linear differential operators. To do this we must consider $\mathrm{LDiff}_k(\xi,\eta)=C^\infty((J^k(\xi)^*\otimes\eta)$ as a reducible G-vector bundle. There is a one-to-one correspondence between all G-invariant linear differential operators of order k, $C^\infty(\xi)\to C^\infty(\eta)$, and the sections of the bundle $((J^k(\xi))^*\otimes\eta)_G$ with a fiber isomorphic to $\mathrm{st}((J^k(\xi))^*\otimes\eta)_b=$ all intertwining linear operators $J^k(\xi)_b\to\eta_b$ between the two finite-dimensional representations of the isotropy group G_b, $b\in B$. [The representation of G_b on $J^k(\xi)_b$ comes from the jet-lifted action of G on ξ.]

Comment 6.2: Description of the G-invariant linear connections on a reducible G-vector bundle.

Comments 4.1 and 6.1 lead to a description (in this language) of all G-invariant linear connections on ξ. They are in one-to-one correspondence with the sections $\varphi\in C^\infty(\xi^*\otimes J^1(\xi))_G$, satisfying $\pi^{1,0}\circ\theta(\varphi)=\mathrm{id}$. In other words, the reduced bundle for the G-invariant linear connections has as a typical fiber all the linear maps $A:\xi_b\to J^1(\xi)_b$ intertwining the two finite-dimensional representations of G_b on ξ_b and $J^1(\xi)_b$, right inverse to $\pi^{1,0}$; $\pi^{1,0}\circ A=\mathrm{id}$.

VII. DIMENSIONAL REDUCTION OF A GROUP ACTION

Let ξ be a reducible G-vector bundle. We call G a reduction group because we shall consider another group O, also acting on ξ by bundle morphisms (F,f). When the two actions commute; $F_o\circ T_g=T_g\circ F_o$, $g\in G$, $o\in O$, the group O has a natural bundle morphism action on ξ_G. In terms of the sections of ξ_G we have

$$o(\varphi)=\theta^{-1}(o(\theta(\varphi))),\quad o\in O,\quad \varphi\in C^\infty(\xi_G). \tag{7.1}$$

The induced representations are a special case of a reduced action. Let G be a group and $H\subset G$ a close subgroup, L a finite-dimensional vector space and V_h a linear representation of H on L. One takes for ξ the trivial bundle $G\times L\to G$ with a G action

$$F_g(g_1,u)=(g_1g^{-1},u),\quad g,g_1\in G,\quad u\in L. \tag{7.2}$$

The reduction group is H with an action on ξ

$$T_h(g_1,u)=(hg_1,V_hu),\quad h\in H. \tag{7.3}$$

The both actions commute and the reduced action of G on $C^\infty(\xi_G)$ is just the representation of G induced by H and V_h.

When ξ is the cotangent bundle, as in Sec. III, the reduced action of O on $C^\infty(T^*(B)_G)$ has an invariant subspace. This is $C^\infty(T^*(M))$ because the reduced action of O preserves the exact sequence (3.5). Such a reduced action of the group $\mathrm{SO}_0(2,4)$ will appear in Sec. IX and it will be nondecomposable.

The dimensional reduction preserves the "symmetry properties" of the differential operators. Let, as usual, ξ, η be two reducible G-vector bundles and D: $C^\infty(\xi)\to C^\infty(\eta)$ a G-invariant differential operator. Let O be another group acting on ξ, η by bundle morphisms. If the actions of O on ξ, η commute with the corresponding actions of G and D is invariant with respect to the actions of O,

$$o(D(\psi))=D(o(\psi)),\quad o\in O,\quad \psi\in C^\infty(\xi), \tag{7.4}$$

then the reduced operator $D_G:C^\infty(\xi_G)\to C^\infty(\eta_G)$ is invariant with respect to the reduced action of O on ξ_G and η_G.

VIII. AN EXAMPLE OF DIMENSIONAL REDUCTION OF A GAUGE FIELD AND YANG–MILLS EQUATION

Here we rederive, in the developed language, the results of Ref. 2. The goal is to make a dimensional reduction of the SU(2) Yang–Mills equation on Minkowski space M^4 by means of a reduction group $\mathrm{SL}(2,\mathbb{C})$ with projected action on M^4—the natural action of the Lorentz group.

We consider a complex two-dimensional Hermitian vector bundle ξ over M^4. Here (x^μ,z^a), $\mu=0,1,2,3$, $a=1,2$, are global canonical coordinates, the metric tensor on M^4 is $g_{\mu\nu}=\mathrm{diag}(-1,1,1,1)$ and (x^μ,z^a,z^a_μ) are the canonical coordinates of $J^1(\xi)$. A linear connection on ξ will be considered as a splitting of (4.6) given by the covariant derivative $\tilde{\nabla}$: $J^1(\xi)\to T^*(M^4)\otimes\xi$, $\tilde{\nabla}\circ i=\mathrm{id}$ or by S: $\xi\to J^1(\xi)$, $\pi^{1,0}\circ S=\mathrm{id}$. We have

$$\pi^{1,0}:(x,z^a,z^a_\mu)\to(x,z^a), \tag{8.1}$$

and

$$S(A):(x^\mu,z^a)\to(x^\mu,z^a,-A^a_{\mu b}z^b), \tag{8.2}$$

where the potentials A_μ are the same as in the covariant derivatives

$$\nabla_\mu=\partial_\mu+A_\mu \tag{8.3}$$

because $\ker\tilde{\nabla}=\mathrm{im}\,S$. The SU(2) connections correspond to $A^+_\mu=-A_\mu$, $\mathrm{tr}\,A_\mu=0$ and the Yang–Mills equation for the potentials A_μ is

$$F_{\mu\nu}=\partial_\mu A_\nu-\partial_\nu A_\mu+[A_\mu,A_\nu],$$
$$\partial^\mu F_{\mu\nu}+[A^\mu,F_{\mu\nu}]=0, \tag{8.4}$$

$\partial_\mu=\partial/\partial x^\mu$, $\partial^\mu=g^{\mu\nu}\partial_\nu$. We need a bundle morphism action of $\mathrm{SL}(2,\mathbb{C})$ on ξ. Let $g\to\Lambda_g$ be the double covering $\mathrm{SL}(2,\mathbb{C})\to\mathrm{SO}_0(1,3)$. The projected action of $\mathrm{SL}(2,C)$ on M^4 is taken to be

$$g(x)=\Lambda_g(x),\quad g\in\mathrm{SL}(2,\mathbb{C}),\quad x\in M^4. \tag{8.5}$$

This action has different types of orbits and we shall only work on $V_+=\{x\in M^4|x^2<0,\ x^0>0\}$ $(x^2=g_{\mu\nu}x^\mu x^\nu)$, where we have orbits of one type and the dimensional reduction is possible. On V_+ the bundle morphism action of $\mathrm{SL}(2,\mathbb{C})$ is taken to be

$$g(x^\mu,z^a)=(\Lambda^\mu_{g\nu}x^\nu,U(g,x)^a_bz^b), \tag{8.6}$$

where

$$U(g,x)=B^{-1}(\Lambda_{g^{-1}}(x))\circ T_g\circ B(x), \tag{8.7}$$

$$B(x)=a^\mu(x)\sigma_\mu, \tag{8.8}$$

$\sigma_0 = \mathbf{1}$, $\sigma_i = 1,2,3$, are the Pauli matrices and

$$a^0(x) = [(x^0 + \sqrt{x^2})/2\sqrt{x^2}]^{1/2}, \tag{8.9}$$

$$a^i(x) = x^i/[2\sqrt{x^2}(x^0 + \sqrt{x^2})]^{1/2}, \quad i = 1,2,3, \tag{8.10}$$

and T_g is the natural representation of $SL(e,\mathbb{C})$ on $\mathbb{C}^2$. Here $U(g,x)$ is the Wigner rotation between the two timelike vectors x and $\Lambda_g(x)$. It satisfies (2.9) and defines a bundle morphism action. Furthermore $U(g,x) \in SU(2)$, this bundle morphism action preserves the Hermitian structure on ξ and the Yang–Mills equation (8.4) is invariant. The one-jet lifting of the action (8.6) on $J^1(\xi)$ is

$$g(x^\mu, z^a, z^a_\mu) = (\Lambda^\mu_{g\nu} x^\nu,\ U(g,x)^a_b z^b,\ \partial_\mu U(g,x)^a_b z^b + U(g,x)^a_b \Lambda_{g^{-1}\mu}{}^\nu z^b_n). \tag{8.11}$$

The corresponding action of $SL(2,\mathbb{C})$ on A_μ [from $S(A) \in \mathrm{Hom}(\xi, J^1(\xi))$] is

$$g(A)_\mu(x) = \Lambda_{g^{-1}\mu}{}^\nu U(g,x') \circ A_\nu(x') \circ U^{-1}(g,x') + U(g,x')\partial_\mu U^{-1}(g,x'), \quad x' = \Lambda_{g^{-1}}(x). \tag{8.12}$$

The first step is to describe the reduced bundle for the $SL(2,\mathbb{C})$-invariant linear connections on ξ, according to Comment 6.2. In this case we have a global coordinate bundle. We take a cross section of all orbits $\tilde{U}_\alpha = \{x \in V_+ | x^i = 0, i = 1,2,3\}$ and this is the base of the reduced bundle. For $x_0 = (1,0,0,0) \in \tilde{U}_\alpha$ the isotropy group is $SL(2,\mathbb{C})_{x_0} = SU(2)$. The representations of $SU(2)$ on the fibers ξ_{x_0} and $J^1(\xi)_{x_0}$ are

$$g(z^a) = T^a_{gb} z^b, \tag{8.13}$$

$$g(z^a, z^a_\mu) = (T^a_{gb} z^b,\ T^a_{gb} \Lambda_{g^{-1}\mu}{}^\nu z^b_\nu), \tag{8.14}$$

$g \in SU(2) \subset SL(2,\mathbb{C})$. The representation on $J^1(\xi)_{x_0}$ is reducible; $(0, z^a_\mu)$ is an invariant subspace. The typical fiber of the reduced bundle for the $SL(2,C)$-invariant linear connections on ξ is the set of all intertwining linear operators

$$(z^a) \to (A(z)^a_\mu) = (A^a{}_{\mu b} z^b) \tag{8.15}$$

between the two representations of $SU(2)$: $(\frac{1}{2})$ on ξ_{x_0} and $(\frac{1}{2}) \otimes ((0) \oplus (1))$ on $(0, a^a_\mu)$. [When $g \in SU(2)$, Λ_g has the form

$$\Lambda_g = \begin{bmatrix} 1 & 0 \\ 0 & R_g \end{bmatrix}, \tag{8.16}$$

$R_g \in SO(3)$.] But $(\frac{1}{2}) \otimes ((0) \oplus (1)) = (\frac{1}{2}) \oplus (\frac{1}{2}) \oplus (\frac{3}{2})$ and it is clear that the space of all intertwining operators has a complex dimension 2. The result is

$$A_0 = c_1 \sigma_0, \quad A_i = c_2 \sigma_i, \quad i = 1,2,3. \tag{8.17}$$

Hence, the restriction of a $SL(2,\mathbb{C})$-invariant gauge field A_μ on $\tilde{U}_\alpha$ is

$$A_0(x^0) = A_0(x^0,0,0,0) = c_1(x^0)\sigma_0, \quad A_i(x^0) = A_i(x^0,0,0,0) = c_2(x^0)\sigma_i, \quad i = 1,2,3. \tag{8.18}$$

The differential operator (8.4) is not internal for $\tilde{U}_\alpha$. The symmetry condition $g(A) = A$ in an infinitesimal form together with its first prolongation provides an expression for the transversal derivatives $\partial_i A\ (x^0)$, $\partial_i \partial_j A_\mu(x^0)$ by means of $A_\mu(x^0)$, $\partial_0 A_\mu(x^0)$, $\partial_0 \partial_0 A_\mu(x^0)$. The result is

$$\partial_i A_0(x^0) = -(1/x^0) A_i(x^0), \tag{8.19}$$

$$\partial_i A_j(x^0) = -[i/4(x^0)^2]\epsilon_{ijk}\sigma_k - (1/x^0)\delta_{ij} A_0(x^0),$$

$$\partial_i \partial_j A_0(x^0) = \delta_{ij}([1/(x^0)^3] - 1/x^0) A_0(x^0),$$

$$\partial_i \partial_j A_k(x^0) = -(1/x^0)\delta_{ij}\partial_0 A_k(x^0) + [1/(x^0)^2]\delta_{jk} A_i(x^0) - [i/4(x^0)^2]\epsilon_{ijn}[\sigma_n, A_k(x^0)], \quad i,j,k = 1,2,3, \tag{8.20}$$

where ϵ_{ijk} is a fully antisymmetric tensor, $\epsilon_{123} = 1$.

The set of equations (8.19) and (8.20) is an explicit form of the splitting morphism S_α in Proposition 6.2. The reduced operator (the analog of D_α in Proposition 6.3) will be obtained if we consider the operator (8.4) on $\tilde{U}_\alpha$ and replace the derivatives $\partial_i A_\mu(x^0)$, $\partial_i \partial_j A_\mu(x^0)$ by the expressions of (8.19) and (8.20). This is a global coordinate realization D_α of the reduced operator. The invariant $SU(2)$ connections correspond to the choice

$$A_0(x^0) = 0, \quad A_i(x^0) = if(x^0)\sigma_i, \quad i = 1,2,3, \tag{8.21}$$

where f is a real-valued function. In this case, setting $x^0 = t^2$, the reduced operator (8.2) is

$$4t^2 f'' + 8tf' + 3f + 8tf^3 = 0. \tag{8.22}$$

An explicit global form of an $SL(2,\mathbb{C})$-invariant connection on ξ is known.[2] But we used here the connections in an infinitesimal neighborhood of $\tilde{U}_\alpha$. This technique is used for dimensional reduction in different papers.[8,9] We stress here that the infinitesimal symmetry condition and its jet prolongations provide a splitting of a relevant exact sequence of the type (5.1).

The splittings of (5.1) [or (6.2)] coming from some symmetry group are not all splittings. There are situations where a crucial role is played by another type of splitting. The next section gives an example of this.

IX. DIMENSIONAL REDUCTION OF THE SIX-DIMENSIONAL MAXWELL EQUATION. CONFORMAL ELECTRODYNAMICS AND NONDECOMPOSABLE REPRESENTATIONS

Here we show that the conformally extended Maxwell equation, the additional scalar fields, and the used nondecomposable representations of the conformal group $C(1,3)$ (see Refs. 10–15) can be obtained by a dimensional reduction of the six-dimensional Maxwell equation, followed by a restriction on the projected light cone.

We start from $\mathbb{R}^6$, $v = (v^\mu, v^5, v^6) = (v^a) \in \mathbb{R}^6$, $\mu = 0,1,2,3$, $g_{ab} = \mathrm{diag}(-1,1,1,1,1,-1)$ is the metric tensor, the Maxwell equation for the electromagnetic potential A_a is

$$F_{ab} = \partial_a A_b - \partial_b A_a, \quad \partial^a F_{ab} = J_b, \tag{9.1}$$

where J_b plays the role of an external current. The group $O(2,4)$ acts naturally on the one-forms A_a, J_a and Eq. (9.1) is invariant. We want to reduce simultaneously the differential operator and the group action.

The reduction group is $R^* = \mathbb{R}\setminus\{0\}$. It acts on $\mathbb{R}^6$ by multiplications,

$$\rho(v) = \rho v, \quad \rho \in R^* . \tag{9.2}$$

For any real λ, the action (9.2) has an extension to a bundle morphism action $T^{(\lambda)}$ on $T(\mathbb{R}^6)$ and $T^*(\mathbb{R}^6)$ by the formulas

$$T_\rho^{(\lambda)}(v,u) = (\rho v, \rho^\lambda u) , \tag{9.3}$$

$(v,u) \in T(\mathbb{R}^6)_v$ and

$$T_\rho^{(\lambda)}(v,p) = (\rho v, \rho^{-\lambda} p) , \tag{9.4}$$

$(v,p) \in T^*(\mathbb{R}^6)_v$. The conditions (2.2) and (6.6) for a one-form A_a read

$$A_a(\rho v) = \rho^{-\lambda} A_a(v) , \tag{9.5}$$

$$v^b \partial_b A_a(v) = -\lambda A_a(v) . \tag{9.6}$$

The action $T^{(\lambda)}$ commutes with the action of $O(2,4)$ and the latter may be reduced. For the pair $T^{(1)}$, $T^{(3)}$ the Maxwell equation (9.1) is invariant and also may be reduced. We shall restrict our attention to the identity-connected component $SO_0(2,4)$ of the group $O(2,4)$. The reduced cotangent bundle $T^*(\mathbb{R}^6)_{R^*}$ [we simplify the notation $T^*(\mathbb{R}^6\setminus\{0\})_{R^*}$] does not depend on λ and the corresponding exact sequence

$$0 \leftarrow \tau^* \xleftarrow{i^*} T^*(\mathbb{R}^6)_{R^*} \xleftarrow{j^*} T^*(P\mathbb{R}^6) \leftarrow 0 , \tag{9.7}$$

$P\mathbb{R}^6 = \mathbb{R}^6/R^*$, is invariant with respect to the reduced action of $SO_0(2,4)$. Hence the subspace $C^\infty(T^*(P\mathbb{R}^6))$ is invariant with respect to the realized representation of $SO_0(2,4)$ on $C^\infty(T^*(\mathbb{R}^6)_{R^*})$. The dim $\tau^* = 1$ and according to the interpretation given in Sec. III, the reduced Maxwell operator acts on the pairs consisting of a scalar field and a one-form on $P\mathbb{R}^6$.

We shall do all calculations in the adapted (nonglobal) coordinates

$$\begin{aligned} &k = v^5 + v^6 , \quad x^\mu = v^\mu/(v^5 + v^6) , \\ &\varphi = g_{ab} v^a v^b / 2(v^5 + v^6)^2 . \end{aligned} \tag{9.8}$$

We set " $-$ " indices for k and " $+$ " for φ. In these coordinates we have

$$\begin{aligned} A_-(k,x,\varphi) &= A\left(\frac{\partial}{\partial k}\right) = x^\nu A_\nu(v) + \left(\frac{1-x^2}{2} + \varphi\right) A_5(v) \\ &\quad + \left(\frac{1+x^2}{2} - \varphi\right) A_6(v) , \\ A_\mu(k,x,\varphi) &= A\left(\frac{\partial}{\partial x^\mu}\right) \\ &= k A_\mu(v) - k x_\mu A_5(v) + k x_\mu A_6(v) , \\ A_+(k,x,\varphi) &= A\left(\frac{\partial}{\partial \varphi}\right) = k A_5(v) - k A_6(v) , \end{aligned} \tag{9.9}$$

where $x_\mu = \eta_{\mu\nu} x^\nu$, $\eta_{\mu\nu} = \mathrm{diag}(-1,1,1,1)$, $\mu,\nu = 0,1,2,3$.

The Maxwell equation for $F(k,x_\mu,\varphi)$ is

$$\begin{aligned} &\frac{1}{k^2}\partial^\mu F_{\mu -} + \frac{1}{k}\partial_- F_{+-} + \frac{3}{k} F_{+-} \\ &\quad - \frac{2\varphi}{k^2}\partial_+ F_{+-} = J_- , \\ &\frac{1}{k^2}\partial^\mu F_{\mu\nu} + \frac{1}{k}\partial_+ F_{-\nu} + \frac{1}{k}\partial_- F_{+\nu} \\ &\quad - \frac{2\varphi}{k^2}\partial_+ F_{+\nu} = J_\nu , \\ &\frac{1}{k^2}\partial^\mu F_{\mu +} + \frac{1}{k}\partial_+ F_{-+} = J_+ . \end{aligned} \tag{9.10}$$

Let $X_{ab} = v_a \partial_b - v_b \partial_a$ be the generators of $SO_0(2,4)$. The physical generators[16] are $X_{\mu\nu}$—the Lorentz transformations, X_{56}—the dilatations, $T_\mu = X_{\mu 6} - X_{\mu 5}$—the translations, and $C_\mu = X_{\mu 5} + X_{\mu 6}$—the special conformal transformations. The action of the special conformal transformations $K = \exp(c^\mu C_\mu)$ in terms of the adapted coordinates is

$$\begin{aligned} &K(k,x^\mu,\varphi) = (k',x'^\mu,\varphi') , \\ &k' = k\cdot(1 + 2c_\nu x^\nu + c^2(x^2 - 2\varphi)) , \\ &x'^\mu = \frac{x^\mu + c^\mu(x^2 - 2\varphi)}{1 + 2c_\nu x^\nu + c^2(x^2 - 2\varphi)} , \\ &\varphi' = \frac{\varphi}{(1 + 2c_\nu x^\nu + c^2(x^2 - 2\varphi))^2} , \end{aligned} \tag{9.11}$$

where $x^2 = \eta_{\mu\nu} x^\mu x^\nu$. The action (9.2) of the reduction group takes the form

$$\rho(k,x^\mu,\varphi) = (\rho k, x^\mu, \varphi) . \tag{9.12}$$

We shall consider a reduction of the six-dimensional Maxwell equation for the pair $T^{(1)}$ and $T^{(3)}$. For a coordinate realization of the reduced bundle $T^*(\mathbb{R}^6)_{R^*}$ we take $\tilde{U}_\alpha = \{v \in \mathbb{R}^6 | k = 1\}$. Then the section of $T^*(\mathbb{R}^6)_\alpha$ has the form $A_{-,\mu,+}(x,\varphi) = A_{-,\mu,+}(1,x,\varphi)$, $J_{-,\mu,+}(x,\varphi) = J_{-,\mu,+}(1,x,\varphi)$. Equation (9.6) on $\tilde{U}_\alpha$ reads (for $\lambda = 1$)

$$\begin{aligned} &\partial_- A_\mu(x,\varphi) = 0, \quad \partial_- A_+(x,\varphi) = 0 , \\ &\partial_- A_-(x,\varphi) = -A_-(x,\varphi) . \end{aligned} \tag{9.13}$$

The reduced Maxwell equation [for the fields $A_{-,\mu,+}(x,\varphi)$] is

$$\begin{aligned} &\partial^\mu \partial_\mu A_- + 2\partial_+ A_- - 2\varphi \partial_+ \partial_- A_- = J_- , \\ &\partial^\mu \partial_\mu A_\nu - \partial_\nu \partial^\mu A_\mu - \partial_\nu \partial_+ A_- - 2\varphi(\partial_+ \partial_+ A_\nu - \partial_+ \partial_\nu A_+) = J_\nu , \\ &\partial^\mu \partial_\mu A_+ - \partial_+ \partial^\mu A_\mu - \partial_+ \partial_+ A_- = J_+ . \end{aligned} \tag{9.14}$$

[Equations (9.14) are obtained from (9.10) by setting $k=1$ and replacing $\partial_- A_{-,\mu,+}(x,\varphi)$ by (9.13). This is $\tilde{D}_\alpha = \tilde{D}\circ S_\alpha$ in Proposition 6.3.]

The adapted coordinates are canonical for the exact sequence (9.7),

$$\begin{aligned} &i^*(A_-,A_\mu,A_+) = A_-\,, \\ &j^*(A_\mu,A_+) = (0,A_\mu,A_+)\,. \end{aligned} \tag{9.15}$$

The reduced action of the special conformal transformations $K = \exp(c^\nu C_\nu)$ on the base of $(T^*(\mathbb{R}^6))$ is

$$\begin{aligned} &K(x^\mu,\varphi) = (x'^\mu,\varphi')\,, \\ &x'^\mu = \frac{x^\mu + c^\mu(x^2 - 2\varphi)}{1 + 2c_\nu x^\nu + c^2(x^2-2\varphi)}\,, \\ &\varphi' = \frac{\varphi}{(1+2c_\nu x^\nu + c^2(x^2-2\varphi))^2}\,, \end{aligned} \tag{9.16}$$

and their action on the fields $A_{-,\mu,+}(x,\varphi)$ for arbitrary λ is

$$K^{-1}(A)_-(x,\varphi) = \frac{1}{p^{\lambda-1}} A_-(x',\varphi')\,,$$

$$\begin{aligned} K^{-1}(A)_\mu(x,\varphi) &= \frac{2c_\mu + 2c^2 x_\mu}{p^\lambda} A_-(x',\varphi') + \frac{1}{p^{\lambda-1}} \frac{\partial x'^\nu}{\partial x^\mu} A_\nu(x',\varphi') \\ &\quad - \frac{4\varphi(c_\mu + c^2 x_\mu)}{p^{\lambda+2}} A_+(x',\varphi')\,, \end{aligned} \tag{9.17}$$

$$\begin{aligned} K^{-1}(A)_+(x,\varphi) = &-\frac{2c^2}{p^\lambda} A_-(x',\varphi') \\ &+ \frac{-2c^\nu p + 2c^2(x^\nu + c^\nu(x^2-2\varphi))}{p^{\lambda+1}} \\ &\times A_\nu(x',\varphi') + \frac{p + 4\varphi c^2}{p^{\lambda+2}} A_+(x',\varphi')\,, \end{aligned}$$

where $p = 1 + 2c_\nu x^\nu + c^2(x^2 - 2\varphi)$.

The submanifold $M^4 = \{(x,\varphi)\in\tilde{U}_\alpha \,|\, \varphi = 0\}$ is invariant and is identified with the Minkowski space. This is the Dirac embedding $M^4 \to Q_{2,4}/R^*$, $Q_{2,4} = \{v\in\mathbb{R}^6\setminus\{0\} \,|\, g_{ab}v^a v^b = 0\}$. On M^4 the special conformal transformations are, as expected,

$$K(x^\mu) = (x^\mu + c^\mu x^2)/(1 + 2c_\nu x^\nu + c^2 x^2)\,. \tag{9.18}$$

The action of $SO_0(2,4)$ on the restriction $(T^*(\mathbb{R}^6)_\alpha)_{M^4}$ is natural and was, in fact, calculated in Ref. 15.

The embedding[17] $M^4 \to \tilde{U}_\alpha$ is fixed in our considerations and leads to the following exact sequences:

$$0 \to T(M^4) \xrightarrow{k} T(\tilde{U}_\alpha)_{M^4} \xrightarrow{l} N(M^4) \to 0\,, \tag{9.19}$$

$$0 \leftarrow T^*(M^4) \xleftarrow{k^*} T^*(\tilde{U}_\alpha)_{M^4} \xleftarrow{l^*} N(M^4)^* \leftarrow 0\,, \tag{9.20}$$

where $N(M^4)$ is the normal bundle, $N(M^4) = T(\tilde{U}_\alpha)_{M^4}/T(M^4)$. In the adapted coordinates we have

$$\begin{aligned} &k^*(A_\mu(x),A_+(x)) = A_\mu(x)\,, \\ &l^*(A_+(x)) = (0,0,0,0,A_+(x))\,, \end{aligned} \tag{9.21}$$

where $A_{-,\mu,+}(x) = A_{-,\mu,+}(x,0)$.

The operator (9.14) is not internal for M^4 and we do not have a natural restriction on M^4. We want to define a restriction on M^4 in such a way that the restricted operator would be invariant with respect to the reduced action of $SO_0(2,4)$ on $(T^*(\mathbb{R}^6))_{M^4}$ [for the special conformal transformations this is (9.17) with $\varphi = 0$]. This may be done if the expressions for the transversal to M^4 derivatives ∂_+, $\partial_+\partial_+$ are invariant.

Let us consider the six-dimensional Lorentz condition

$$\partial^a A_a(v) = 0\,. \tag{9.22}$$

The reduction of (9.22) on $T^*(\mathbb{R}^6)_\alpha$ is ($\lambda = 1$),

$$\begin{aligned} &\partial^\mu A_\mu(x,\varphi) + \partial_+ A_-(x,\varphi) + 2A_+(x,\varphi) \\ &\quad - 2\varphi\,\partial_+ A_+(x,\varphi) = 0\,. \end{aligned} \tag{9.23}$$

Equation (9.23) is automatically invariant with respect to the reduced bundle morphism action of $SO_0(2,4)$ on $T^*(\mathbb{R}^6)_\alpha$ and, together with its first prolongation, gives on $M^4(\varphi = 0)$

$$\begin{aligned} &\partial^\mu A_\mu(x) + \partial_+ A_-(x) + 2A_+(x) = 0\,, \\ &\partial_+\partial^\mu A_\mu(x) + \partial_+\partial_+ A_-(x) = 0\,. \end{aligned} \tag{9.24}$$

Combining (9.14) (considered on M^4, $\varphi = 0$) and (9.24), we have

$$\begin{aligned} &\Box A_-(x) - 2\,\partial^\mu A_\mu(x) - 4A_+(x) = J_-(x), \\ &\Box A_\nu(x) + 2\,\partial_\nu A_+(x) \qquad\quad = J_\nu(x), \\ &\Box A_+(x) \qquad\qquad\qquad\quad = J_+(x), \end{aligned} \tag{9.25}$$

where $\Box = \partial^\mu\partial_\mu$. Equations (9.25) are also automatically invariant with respect to the reduced ($T^{(1)}$ for A_a and $T^{(3)}$ for J_a) bundle morphism action of $SO_0(2,4)$ on $(T^*(\mathbb{R}^6)_\alpha)_{M^4}$, i.e., it is conformally invariant.

One can impose some invariant conditions. Let $i^*(A) = 0$, $i^*(J) = 0$. Then $A_-(x) = 0 = J_-(x)$ and we have

$$\begin{aligned} &-2\,\partial^\mu A_\mu - 4A_+ = 0\,, \\ &\Box A_\nu + 2\,\partial_\nu A_+ \quad = J_\nu\,, \\ &\Box A_+ \qquad\quad = J_+\,. \end{aligned} \tag{9.26}$$

Excluding A_+ we have

$$\partial^\mu F_{\mu\nu} = J_\nu, \quad -\tfrac{1}{2}\Box\,\partial^\nu A_\nu = J_+\,, \tag{9.27}$$

$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\nu$. Equations (9.26) were derived in Refs. 11 and 14, and Eqs. (9.27) were used in Ref. 15.

If only $J_- = 0$ and excluding A_+, we have from (9.25)

$$\begin{aligned} &\partial^\mu F_{\mu\nu} + \tfrac{1}{2}\Box\,\partial_\nu A_- = J_\nu\,, \\ &\tfrac{1}{4}\Box^2 A_- - \tfrac{1}{2}\Box\,\partial^\mu A_\mu = J_+\,. \end{aligned} \tag{9.28}$$

The equation $\partial^a J_a = 0$ on $\mathbb{R}^6$, reduced by means of $T^{(3)}$ on $T^*(\mathbb{R}^6)_\alpha$, for $J_- = 0$ is an internal differential operator for the submanifold $M^4\subset\tilde{U}_\alpha$ and on M^4 gives

$$\partial^\mu J_\mu(x) = 0\,. \tag{9.29}$$

Hence it is conformally invariant and together with (9.28) leads to the conformally invariant equation

$$\Box^2 A_-(x) = 0\,. \tag{9.30}$$

The differential operators of (9.28) and (9.30) lead to a one-parameter family of conformally invariant differential

equations

$$\partial^\mu F_{\mu\nu} + \tfrac{1}{2}\Box\,\partial_\nu A_- = J_\nu\,,$$
$$\beta\,\Box^2 A_- - \tfrac{1}{2}\Box\partial^\mu A_\mu = J_+\,, \tag{9.31}$$

$\beta \in R$. Equations (9.31) were introduced in Refs. 10 and 12 by studying the conformally invariant two-point functions, and discussed in Ref. 13.

The new result of this section is the observation that the set of equations (9.27) and (9.28) can be obtained in two steps. The first is the standard dimensional reduction of the six-dimensional Maxwell equation by means of a reduction group R^* and actions $T^{(1)}$, $T^{(3)}$. The second is a conformally invariant restriction of the reduced Maxwell equation by means of the reduced six-dimensional Lorentz condition. The additional fields come from the exact sequence (3.5) and have the same nature as the scalar fields and the gauge field in the dimensional reduction of G-invariant metrics (see Ref. 1 and Sec. III).

ACKNOWLEDGMENTS

The author would like to thank Dr. L. Hadjiivanov, Dr. V. Tsanov, and Dr. V. Molotkov for a critical reading of the manuscript and useful discussions.

[1]R. Coquereaux and A. Jadczyk, "Geometry of multidimensional universes," Commun. Math. Phys. **90**, 79 (1983).
[2]L. K. Hadjiivanov and D. T. Stoyanov, "Gauge representations of the Lorentz group and classical solutions of the Yang–Mills equation," JINR preprint E2-81-705, Dubna, 1981.
[3]V. Molotkov, *Glutoi*, to be published.
[4]E. Bierstone, "The equivalent covering homotopy property for differentiable G-fibre bundles," J. Differ. Geom. **8**, 615 (1973).
[5]P. A. Nikolov, "Cohomology of groups and Kaluza–Klein theories," preprint IC/85/274, Trieste, 1985.
[6]R. S. Palais, *Seminar on the Atiyah–Singer Index Theorem* (Princeton U.P., Princeton, NJ, 1965), Chap. IV.
[7]D. C. Spencer, "Overdetermined systems of linear partial differential equations," Bull. Am. Math. Soc. **75**, 179 (1969); "Curvatures associated with differential operators," in *Lecture Notes in Mathematics*, Vol. 392 (Springer, Berlin, 1974).
[8]P. Forgacs and N. S. Manton, "Space-time symmetry in gauge theories," Commun. Math. Phys. **72**, 15 (1980).
[9]V. N. Romanov, A. S. Schwarz, and Yu. S. Tyupkin, "On spherically symmetric fields in gauge theories," Nucl. Phys. B **130**, 209 (1978).
[10]R. P. Zaikov, "On conformal invariance in gauge theories; electrodynamics," JINR preprint E2-83-28, Dubna, 1983.
[11]F. Bayen and M. Flato, "Remarks on conformal space," J. Math. Phys. **17**, 1112 (1976).
[12]B. Binegar, C. Fronsdal, and W. Heidenrech, "Conformal QED," J. Math. Phys. **24**, 2828 (1983).
[13]P. Furlan, V. B. Petkova, G. M. Sotkov, and I. T. Todorov, "Conformal quantum electrodynamics with a 5-potential," ISAS preprint 52/83/EP, Trieste; "Conformal quantum electrodynamics and non-decomposable representations," Riv. Nuovo Cimento **8**, 3 (1985); V. B. Petkova, G. M. Sotkov, and I. T. Todorov, "Conformal gauges and renormalized equation of motion in massless quantum electrodynamics," Commun. Math. Phys. **97**, 227 (1985).
[14]D. H. Mayer, "Vector and tensor fields on conformal space," J. Math. Phys. **16**, 884 (1975).
[15]G. M. Sotkov and D. T. Stoyanov, "Conformal quantization of electrodynamics," J. Phys. A **16**, 2817 (1983).
[16]A. Salam and J. Strathdee, "Nonlinear realizations II. Conformal symmetry," Phys. Rev. **184**, 1760 (1969).
[17]P. A. M. Dirac, "Wave equation in conformal space," Ann. Math. **37**, 429 (1936).

Cohomology of supermanifolds

Claudio Bartocci and Ugo Bruzzo
Dipartimento di Matematica, Università di Genova, Via L. B. Alberti 4, 16132 Genova, Italy

(Received 17 March 1987; accepted for publication 27 May 1987)

The cohomological properties of supermanifolds (intended in the sense of De Witt [*Supermanifolds* (Cambridge U. P., London, 1984)] and Rogers [J. Math. Phys. **21**, 1352 (1980)]) are investigated, paying particular attention to the de Rham cohomology of supersmooth differential forms (SDR cohomology). The SDR cohomology of De Witt supermanifolds is shown to be equivalent to the de Rham cohomology of their body. The SDR cohomology is explicitly computed for some topologically nontrivial supermanifolds and some general conclusions concerning the geometric structure of supermanifolds and the properties of the SDR cohomology are drawn. In particular, it is shown that the SDR cohomology is neither a topological nor a real differentiable invariant, but rather a "superdifferentiable" invariant.

I. INTRODUCTION

In this paper we undertake a systematic investigation of the cohomology of supermanifolds. Supermanifolds are intended in the sense of De Witt and Rogers, namely, they are manifolds whose coordinates take values in an exterior algebra B_L. We consider here the case where B_L is finitely generated, i.e., $L < \infty$. Apart from its possible applications in theoretical physics, mainly related to the study of quantum anomalies in supergauge and superstring theories, this investigation has an interest of its own, since it can help to unravel the not yet well-understood geometrical structure of supermanifolds.

Our main object of interest is the cohomology of the differential complex of "supersmooth" differential forms on a supermanifold M, with the differential operator represented by the exterior differential d. The crucial point is that this "supersmooth de Rham cohomology" (SDR cohomology) is different from the Čech cohomology of the locally constant sheaf $\mathscr{B}_{L^\cdot}$ on M with stalk $B_{L^\cdot}$, thus breaking the analogy with real manifolds, where the de Rham cohomology of smooth forms and the Čech cohomology of the locally constant sheaf with stalk $\mathbb{R}$ (the real field) do actually coincide. This implies that, in general, the supersmooth and the ordinary de Rham cohomologies of a supermanifold are different. This behavior is basically due to the fact that supersmooth forms have in some sense holomorphic properties, and therefore the sheaves of supersmooth forms have nonvanishing Čech cohomology, contrary to the sheaves of smooth forms on a real manifold.

This state of affairs has an interesting consequence: SDR cohomology is not a topological invariant. Indeed we shall discuss an example where two supermanifolds, isomorphic as real manifolds but carrying nonisomorphic superdifferentiable structures, have different SDR cohomologies. Thus this cohomology in some sense carries information about the superdifferentiable structure, and therefore could be a useful tool to study supermanifolds.

This analysis shows that supermanifolds have a richer cohomological structure than graded manifolds. Basically, graded manifolds are sheaves of Z_2-graded commutative algebras on a real manifold; their de Rham cohomology is equivalent to the de Rham cohomology of the base manifold.[1] In this respect, graded manifolds behave like De Witt supermanifolds (see Sec. V).

Let us now describe the contents of this paper. In Sec. II we recall some basic definitions in sheaf theory and a few results in sheaf cohomology. Section III contains a brief introduction to supermanifolds, with an emphasis on the definition of the function sheaf. In Sec. IV we introduce the supermanifold cohomologies we are interested in (SDR cohomology, ordinary de Rham cohomology, Čech cohomology of the locally constant sheaf $\mathscr{B}_{L^\cdot}$ and of the sheaf of supersmooth functions) and show the most obvious relationships among them. In Sec. V we deal with a "degenerate" case, i.e., the case of the so-called De Witt supermanifolds (supermanifolds that are locally trivial bundles on a real manifold), while in Sec. VI we discuss some examples of topologically nontrivial supermanifolds.

Finally, let us recall that SDR cohomology has already been considered by Rabin in Ref. 2, where some of our results can already be found.

II. SHEAF THEORY

Since in this work we shall mainly use techniques related to sheaf cohomology, we shall start with some definitions and results in sheaf theory and sheaf cohomology.

Sheaves[3]: Let X be a topological space. A *sheaf* $\mathscr{F}$ of Abelian groups on X is a correspondence that to each open set U in X assigns an Abelian group $\mathscr{F}(U)$, called the group of *sections* of $\mathscr{F}$ over U, so as to verify the following properties.

(i) For any inclusion of open sets $V \subset U$ there exists a group morphism ρ_V^U: $\mathscr{F}(U) \to \mathscr{F}(V)$, called restriction morphism.

(ii) For all open sets U, $\rho_U^U = \mathrm{id}$.

(iii) If $W \subset V \subset U$ are inclusions of open sets, then $\rho_V^U \circ \rho_W^V = \rho_W^U$.

(iv) If $\{U_i,\ i \in I\}$ is a collection of open sets in X, $U = \cup_{i \in I} U_i$, and $s, t \in \mathscr{F}(U)$ are such that $\rho_{U_i}^U(s) = \rho_{U_i}^U(t)\ \forall i \in I$, then $s = t$.

(v) If $\{U_i, i \in I\}$ and U are as above, and a collection

 0022-2488/87/102363-06$02.50 © 1987 American Institute of Physics

$\{s_i \in \mathcal{F}(U_i), i \in I\}$ is given such that $\rho^{U_i}_{U_i \cap U_j}(s_i) = \rho^{U_j}_{U_i \cap U_j}(s_j)$, then there exists a section $s \in \mathcal{F}(U)$ such that $\rho^{U}_{U_i}(s) = s_i$.

For all $x \in X$ one defines the *stalk* $\mathcal{F}_x$ of $\mathcal{F}$ at x as the direct limit of the $\mathcal{F}(U)$'s over all open neighborhoods U of x. Here $\mathcal{F}_x$ is an Abelian group whose elements are called the *germs* of sections of $\mathcal{F}$ at x.

Given two sheaves $\mathcal{F}$ and $\mathcal{G}$ on X a sheaf morphism $\lambda: \mathcal{F} \to \mathcal{G}$ is a collection $\{\lambda_U, U \text{ open in } X\}$ of group morphisms $\lambda_U: \mathcal{F}(U) \to \mathcal{G}(U)$ such that for all inclusions of open sets $V \subset U$ one has $\rho^U_V \circ \lambda_U = \lambda_U \circ \rho^U_V$.

Soft and fine sheaves: In the following, $\mathcal{F}$ will always denote a sheaf on a topological space X. A sheaf $\mathcal{F}$ is said to be *soft* if any section of $\mathcal{F}$ on a closed subset of X can be extended to all of X.[3] An example of a soft sheaf is provided by the sheaf on continuous real-valued functions on a normal space.

A sheaf $\mathcal{F}$ is said to be *fine*[4] if, given any locally finite open cover $\mathcal{U} = \{U_i, i \in I\}$ of X, there exists a collection $\{\phi^i, i \in I\}$ of endomorphisms of $\mathcal{F}$ such that (i) if $s \in \mathcal{F}(V)$, then $\rho^V_{V - U_i} \circ \phi^i_V(s) = 0$ for all $i \in I$; (ii) if V intersects only a finite number of $U_i \in \mathcal{U}$, and $s \in \mathcal{F}(V)$, then

$$s = \sum_{i \in I} \phi^i_V(s).$$

Examples of fine sheaves are the sheaves of differential p-forms on a real C^∞ differentiable manifold, $p \geqslant 0$.

If the base space is paracompact, it is easily verified that any fine sheaf is soft.[3] Under the same hypothesis, any sheaf can be canonically imbedded into a fine sheaf (by paracompact we mean Hausdorff such that any open cover has a locally finite refinement).

Čech cohomology[3]: Let $\mathcal{F}$ be a sheaf of Abelian groups on a topological space X, and $\mathcal{U} = \{U_\alpha, \alpha \in J\}$ an open cover of X, with J an ordered set; for all $\alpha_0 \cdots \alpha_p \in J$ define $U_{\alpha_0 \cdots \alpha_p} = U_{\alpha_0} \cap \cdots \cap U_{\alpha_p}$. Then define the complex of Abelian groups $C^*(\mathcal{U}, \mathcal{F})$ whose pth term is

$$C^p(\mathcal{U}, \mathcal{F}) = \prod_{\alpha_0 < \cdots < \alpha_p} \mathcal{F}(U_{\alpha_0 \cdots \alpha_p})$$

and a differential operator $\delta: C^p(\mathcal{U}, \mathcal{F}) \to C^{p+1}(\mathcal{U}, \mathcal{F})$ as follows: if $f = \{f_{\alpha_0 \cdots \alpha_p}\} \in C^p(\mathcal{U}, \mathcal{F})$, then

$$(\delta f)_{\alpha_0 \cdots \alpha_{p+1}} = \sum_{k=0}^{p+1} (-1)^k f_{\alpha_0 \cdots \hat{\alpha}_k \cdots \alpha_{p+1}},$$

where the caret denotes that the index has been omitted. The *Čech cohomology of X with values in $\mathcal{F}$ with respect to the cover $\mathcal{U}$* is defined as the cohomology of the differential complex $(C^*(\mathcal{U}, \mathcal{F}), \delta)$, i.e.,

$$\check{H}^p(\mathcal{U}, \mathcal{F}) = \frac{\ker\{\delta: C^p(\mathcal{U}, \mathcal{F}) \to C^{p+1}(\mathcal{U}, \mathcal{F})\}}{\mathrm{Im}\{\delta: C^{p-1}(\mathcal{U}, \mathcal{F}) \to C^p(\mathcal{U}, \mathcal{F})\}}.$$

The *Čech cohomology of X with values in $\mathcal{F}$* is defined as the direct limit of the Čech cohomologies with values in $\mathcal{F}$ over all the open covers of X (this involves some set-theoretical difficulties, see Ref. 3); the cohomology groups so obtained are denoted by $\check{H}^p(X, \mathcal{F})$. One has naturally a group morphism

$$\check{H}^p(\mathcal{U}, \mathcal{F}) \to \check{H}^p(X, \mathcal{F}). \tag{2.1}$$

The following results, taken from Ref. 3, will be useful in the following.

Theorem 2.1: If X is paracompact, a sufficient condition for the morphism (2.1) to be one-to-one is that for all nonvoid intersections $U_{\alpha_0 \cdots \alpha_q}$, one has

$$\check{H}^p(U_{\alpha_0 \cdots \alpha_q}, \mathcal{F}) = 0 \quad \text{for } p \geqslant 1. \qquad \square$$

Theorem 2.2: If X is paracompact, and

$$0 \to \mathcal{F} \to \mathcal{G} \to \mathcal{A} \to 0$$

is an exact sequence of sheaves on X, there is a long exact sequence in Čech cohomology

$$0 \to \check{H}^0(X, \mathcal{F}) \to \check{H}^0(X, \mathcal{G}) \to \check{H}^0(X, \mathcal{A}) \xrightarrow{\partial} \check{H}^1(X, \mathcal{F})$$
$$\to \cdots \to \check{H}^p(X, \mathcal{F}) \to \check{H}^p(X, \mathcal{G}) \to \check{H}^p(X, \mathcal{A})$$
$$\xrightarrow{\partial} \check{H}^{p+1}(X, \mathcal{F}) \to \cdots,$$

where the ∂'s are the so-called connecting morphisms.[3] $\square$

Theorem 2.3: If the sheaf $\mathcal{F}$ on a paracompact space X is soft (and, *a fortiori*, if it is fine), its Čech cohomology vanishes, i.e.,

$$\check{H}^p(X, \mathcal{F}) = 0 \quad \text{for all } p \geqslant 1. \qquad \square$$

De Rham cohomology[5]: Let M be a differentiable manifold, and Ω^p the sheaf of the real valued p-forms on M, $p \geqslant 0$. The cohomology of the differential complex

$$\Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \cdots, \tag{2.2}$$

where d is Cartan's exterior differential, is called the *de Rham cohomology* of M; its cohomology groups will be denoted by $H^*_{\mathrm{DR}}(M)$. The classical de Rham theorem can be stated as follows.

Theorem 2.4: For all $p \geqslant 0$,

$$H^p_{\mathrm{DR}}(M) \simeq \check{H}^p(M, \mathcal{R}), \tag{2.3}$$

where $\mathcal{R}$ denotes the locally constant sheaf on M with stalk $\mathbb{R}$ (the real field). $\square$

III. SUPERMANIFOLDS

In this section we describe the fundamentals of supermanifold theory, mainly following Rogers.[6,7] Let B_L be the exterior algebra over $\mathbb{R}^L$, $L < \infty$, with its natural Z_2 gradation $B_L = (B_L)_0 \oplus (B_L)_1$ (in the following, "graded" will always mean Z_2 graded). A basis for B_L (as a graded vector space) is conveniently indexed by M_L, the set of strictly increasing sequences of integers $\mu = \{0 \leqslant \mu_1 < \cdots < \mu_s \leqslant L\}$, as follows: if $\{e_1 \cdots e_L\}$ are generators of B_L, then $\beta_\mu = e_{\mu_1} \wedge \cdots \wedge e_{\mu_s}$; moverover, we set $\beta_0 = 1$.

With the wedge product, B_L is a graded commutative algebra, in the sense that $(B_L)_i (B_L)_j \subset (B_L)_{i+j \bmod 2}$ and

$$a \wedge b = (-1)^{ij} b \wedge a \quad \text{if } a \in (B_L)_i, \quad b \in (B_L)_j$$

(in the following, the wedge product symbol will be understood). Let N denote the ideal of nilpotents of B_L; then $B_L = \mathbb{R} \oplus N$. We denote by $\sigma: B_L \to \mathbb{R}$ (body map) and $s: B_L \to N$ (soul map) the projections. The Cartesian product $(B_L)^{m+n}$ has a natural structure of graded B_L module, with the gradation

$$(B_L)^{m+n} = B_L^{m,n} \oplus B_L^{\bar{m},\bar{n}}$$

given by $B_L^{m,n} = (B_L)_0^m \times (B_L)_1^n$, $B_L^{\bar{m},\bar{n}} = (B_L)_1^m \times (B_L)_0^n$. A body map $\sigma^{m,n}$: $B_L^{m,n} \to \mathbb{R}^m$ is defined by setting $\sigma^{m,n}(x^1 \cdots x^m, y^1 \cdots y^n) = (\sigma(x^1) \cdots \sigma(x^m))$.

Here $B_L^{m,n}$ can be naturally endowed with two distinct topologies: its topology as a $2^{L-1}(m+n)$-dimensional real vector space (that we shall call *fine topology*) and a *coarse topology* whose open sets are the counterimages through $\sigma^{m,n}$ of open sets in $\mathbb{R}^m$. If not otherwise stated, in the following we shall consider in $B_L^{m,n}$ the fine topology.

Next we come to the definition of a sheaf of B_L-valued functions on $B_L^{m,n}$, in terms of which the concept of supermanifold is introduced. We denote by $\mathscr{C}[V;Q]$ the sections over $V \subset X$ of the sheaf of Q-valued C^∞ functions on a manifold X. Let U be an open set in $\mathbb{R}^m$, L and L' two positive integers with $L' \leqslant L$, and denote by[7]

$$Z_{L',L}: \mathscr{C}[U;B_{L'}] \to \mathscr{C}[(\sigma^{m,0})^{-1}(U);B_L]$$

the mapping explicitly given by

$$Z_{L',L}(f)(x^1 \cdots x^m) = \sum_{i_1 \cdots i_m = 0}^{L} \frac{1}{i_1! \cdots i_m!} (\partial^{i_1}_1 \cdots \partial^{i_m}_m f)\Big|_{(\sigma(x^1) \cdots \sigma(x^m))} \times s(x^1)^{i_1} \cdots s(x^m)^{i_m}. \tag{3.1}$$

Here $Z_{L',L}$ is injective; we denote by $\hat{\mathscr{G}}[(\sigma^{m,0})^{-1}(U)]$ its image. Thus $\hat{\mathscr{G}}[(\sigma^{m,0})^{-1}(U)]$ is the ring of GH^∞ functions of even variables on $(\sigma^{m,0})^{-1}(U)$. The GH^∞ functions of even and odd variables are naturally defined on open sets $(\sigma^{m,n})^{-1}(U)$, where U is an open set in $\mathbb{R}^m$. The ring of GH^∞ functions on a set of this type is denoted by $\mathscr{G}[(\sigma^{m,n})^{-1}(U)]$ and its elements have the form

$$F(x^1 \cdots x^m, y^1 \cdots y^n) = \sum_{\mu \in M^L} F_\mu(x^1 \cdots x^m) y^\mu, \tag{3.2}$$

where $y^\mu = y^{\mu_1} \cdots y^{\mu_s}$ and $F_\mu \in \hat{\mathscr{G}}[(\sigma^{m,0})^{-1}(U)]$. The derivatives of F are *uniquely* determined by the development

$$F(x+h, y+k) = F(x,y) + \sum_{i=1}^{m} h^i \frac{\partial F}{\partial x^i}(x,y) + \sum_{\alpha=1}^{n} k^\alpha \frac{\partial F}{\partial y^\alpha}(x,y) + O(h,k)^2$$

provided that

$$L - L' \geqslant n. \tag{3.3}$$

Remarks: (i) Here $\mathscr{G}[(\sigma^{m,n})^{-1}(U)]$ is also endowed with a structure of graded $B_{L'}$ module. For all open sets V in $B_L^{m,n}$ we let $\mathscr{G}(V) = \mathscr{G}[(\sigma^{m,n})^{-1}\sigma^{m,n}(V)]$; this defines a sheaf $\mathscr{G}$ of graded $B_{L'}$ modules on $B_L^{m,n}$.

(ii) For all open U in $\mathbb{R}^m$, Eq. (3.2) defines an epimorphism of graded $B_{L'}$ modules

$$\hat{\mathscr{G}}[(\sigma^{m,0})^{-1}(U)] \otimes_{B_{L'}} \Lambda[n] \to \mathscr{G}[(\sigma^{m,n})^{-1}(U)], \tag{3.4}$$

where $\Lambda[n]$ is the exterior algebra generated by $B_L^{0,n}$ over $B_{L'}$. This morphism is one-to-one if Eq. (3.3) holds.

(iii) The sheaf $\mathscr{G}$ is apparently not soft, and therefore, since $B_L^{m,n}$ is paracompact, not even fine. This precludes the existence of GH^∞ partitions of unity on $B_L^{m,n}$, and *a fortiori* on any supermanifold.

The concept of GH^∞ function, due to Rogers,[7] is a refinement of that of G^∞ function [which is recovered by setting $L = L'$ in Eq. (3.1)], motivated by the fact that G^∞ functions are not well behaved in many regards [partial derivatives with respect to odd variables are not defined, i.e, the map (3.4) is not injective, and, as a consequence, the sheaf of derivations of G^∞ functions is not locally free[8]].

In the following, in order to avoid the above-mentioned drawbacks, we shall always assume that condition (3.3) holds.

Definition 3.1: An (m,n)-dimensional GH^∞ supermanifold is a Hausdorff, second countable topological space M together with an atlas $\mathscr{A} = \{(U_\alpha, \psi_\alpha) | \psi_\alpha: U_\alpha \to B_L^{m,n}\}$ such that the transition functions are GH^∞ mappings. □

Supermanifolds defined in this way are quite general as far as topology is concerned, as explicit examples show.[9] One can strongly constrain the topological structure by requiring that the images $\psi_\alpha(U_\alpha)$ are open in $B_L^{m,n}$ also in the coarse topology, thus obtaining the so-called De Witt supermanifolds.[10] The structural result stated in the following theorem will be useful later on.

Theorem 3.1: Any (m,n)-dimensional De Witt supermanifold M is a locally trivial C^∞ bundle $\Phi: M \to M_0$ over an m-dimensional real differentiable manifold M_0, with typical fiber $P^{m,n} = P^m \times (B_L)_1^n$, P being the ideal of nilpotents in $(B_L)_0$. □

The real manifold M_0 is usually called the *body* of M. Theorem 3.1 establishes the existence of C^∞ isomorphisms $\rho_\alpha: \Phi^{-1}(U_\alpha) \to U_\alpha \times P^{m,n}$; however, in general M is not a vector bundle since the mappings $\rho_\alpha \circ \rho_\beta^{-1}(x,\cdot): P^{m,n} \to P^{m,n}$, with $x \in U_\beta$ fixed, may fail to be vector space morphisms.

IV. SUPERMANIFOLD COHOMOLOGY THEORIES

We wish now to describe some cohomology theories that are natural to be considered on a GH^∞ supermanifold M. According to Definition 3.1, the topological space underlying M has a structure of $2^{L-1}(m+n)$ dimensional C^∞ real manifold, where (m,n) is the dimension of M as a supermanifold. So we can consider on M the sheaves $\mathscr{C}^p$ of $B_{L'}$-valued C^∞ p-forms, $p \geqslant 0$. We define a $B_{L'}$-valued de Rham cohomology, $H^*_{DR}(M, B_{L'})$, as the cohomology of the complex $\mathscr{C}^*(M)$,

$$\mathscr{C}^0(M) \xrightarrow{d} \mathscr{C}^1(M) \xrightarrow{d} \cdots, \tag{4.1}$$

where d is the obvious extension of Cartan's exterior differential. There is a relationship between $H^*_{DR}(M, B_{L'})$ and $H^*_{DR}(M)$, which is a straightforward consequence of the identity $B_{L'} = \mathbb{R} \otimes_{\mathbb{R}} B_{L'}$ (in the following, all tensor products will be taken over $\mathbb{R}$). In order to show this relationship, let us proceed as follows. The cochain complex (4.1) can be written as $\mathscr{C}^*(M) = \Omega^*(M) \otimes B_{L'}$, where $\Omega^*(M)$ is the de Rham complex (2.2). The cochain complexes $\mathscr{C}^*(M)$ and $\Omega^*(M)$ can be regarded as chain complexes, $\mathscr{C}_*(M)$ and $\Omega_*(M)$, by defining $\mathscr{C}_{-p}(M) = \mathscr{C}^p(M)$, $\Omega_{-p}(M) = \Omega^p(M)$ for $p \geqslant 0$, $\mathscr{C}_{-p}(M) = \Omega_{-p}(M) = 0$ otherwise, and their cohomologies regarded as homologies. This trick permits us to apply the *universal coefficient theorem*,[3] which yields

$$H^p_{DR}(M,B_{L'}) = H^p_{DR}(M) \otimes B_{L'} \oplus \mathrm{Tor}_{\mathbb{R}}(H^{p+1}_{DR}(M),B_{L'}).$$

The torsion functor Tor_Σ is defined as follows: if A, B are Σ modules, and $0 \to F' \to F \to A \to 0$ is any exact sequence with F free, then the sequence $0 \to \mathrm{Tor}_\Sigma(A,B) \to F' \otimes B \to F \otimes B \to A \otimes B \to 0$ is exact. Since $B_{L'}$ is free as an $\mathbb{R}$ module, $\mathrm{Tor}_{\mathbb{R}}(H^{p+1}_{DR}(M),B_{L'}) = 0$, so that

$$H^*_{DR}(M,B_{L'}) = H^*_{DR}(M) \otimes B_{L'}. \tag{4.2}$$

Denoting by $\mathcal{G}^p$ the sheaf of GH^∞ p-forms on M, we consider the complex $\mathcal{G}^*$

$$\mathcal{G}^0(M) \overset{d}{\to} \mathcal{G}^1(M) \overset{d}{\to} \cdots.$$

The GH^∞ (or supersmooth) de Rham cohomology of M, denoted by $H^*_{SDR}(M)$, is defined as the cohomology of this complex. Apart from the obvious isomorphism $H^0_{SDR}(M) \simeq H^0_{DR}(M,B_{L'})$, $H^*_{SDR}(M)$ is *a priori* different from the $B_{L'}$-valued de Rham cohomology of M, and we shall indeed give in Sec. VI examples where $H^*_{SDR}(M) \neq H^*_{DR}(M,B_{L'})$. It is obvious that the H^*_{SDR} are functors from the category of GH^∞ supermanifolds to the category of Abelian groups. Indeed, GH^∞ supermanifold maps f_1: $M_1 \to M_2$ and f_2: $M_2 \to M_3$ induce group morphisms f_i^*: $H^*_{SDR}(M_{i+1}) \to H^*_{SDR}(M_i)$ such that $(f_2 \circ f_1)^* = f_1^* \circ f_2^*$ (actually, the f_i^* are morphisms of graded $B_{L'}$ modules).

Following Rabin,[2] it is possible to state a theorem of GH^∞ homotopic invariance for SDR cohomology.

Theorem 4.1: Let $f,g: M \to N$ be GH^∞ maps. If there is a GH^∞ map $F: M \times (B_L)_0 \to N$ such that $F(x,y) = f(x)$ for $\sigma(y) \geqslant 1$ and $F(x,y) = g(x)$ for $\sigma(y) \leqslant 0$, then $f^* = g^*$. □

Other cohomologies we can consider on a supermanifold M are the Čech cohomologies $\check{H}^*(M,\mathcal{R})$ and $\check{H}^*(M,\mathcal{B}_{L'})$ of the locally constant sheaves $\mathcal{R}$ and $\mathcal{B}_{L'}$ on M, whose stalks are $\mathbb{R}$ and $B_{L'}$, respectively. Since M is paracompact, these cohomologies fulfill Theorems 2.1–2.3. Moreover, they are related by

$$\check{H}^*(M,\mathcal{B}_{L'}) = \check{H}^*(M,\mathcal{R}) \otimes B_{L'}, \tag{4.3}$$

which, like Eq. (4.2), is obtained by means of the universal coefficient theorem. Then using the de Rham Theorem 2.4 one gets the canonical isomorphisms

$$\check{H}^*(M,\mathcal{B}_{L'}) = \check{H}^*(M,\mathcal{R}) \otimes B_{L'} \simeq H^*_{DR}(M) \otimes B_{L'} = H^*_{DR}(M,B_{L'}), \tag{4.4}$$

which can be regarded as morphisms of $B_{L'}$ modules.

As far as the Čech cohomologies $\check{H}^*(M,\mathcal{G}^p)$ are concerned, one should remark that in general they do not vanish, contrary to what happens in the case of $\check{H}^*(M,\mathcal{C}^p)$. More generally, one can prove the following result.

Theorem 4.2: Assume that $\check{H}^k(M,\mathcal{G}^p) = 0$ for $0 \leqslant p \leqslant q-1$ and $1 \leqslant k \leqslant q$. Then $H^k_{SDR}(M) \simeq H^k_{DR}(M) \otimes B_{L'}$ for $1 \leqslant k \leqslant q$.

Proof: Following Ref. 11 (p. 44), one gets $H^k_{SDR}(M) \simeq \check{H}^k(M,\mathcal{B}_{L'})$. Then the isomorphism (4.4) completes the proof. □

Since we shall see examples where $H^*_{SDR}(M) \neq H^*_{DR}(M) \otimes B_{L'}$, Theorem 4.2 shows that $\check{H}^*(M,\mathcal{G}^p) \neq 0$ in general.

V. COHOMOLOGY OF DE WITT SUPERMANIFOLDS

In this section we investigate the case of De Witt supermanifolds (see Sec. III) for the various cohomologies we have introduced in the preceding section. We shall see that the $B_{L'}$-valued de Rham cohomology, the Čech cohomology of $\mathcal{B}_{L'}$, and the SDR cohomology are isomorphic, and that they all coincide with the $B_{L'}$-valued de Rham cohomology of the base real manifold. We consider a De Witt supermanifold M topologized with the fine topology; however, if we endow M with the coarse topology, the result is the same, even though the proof must be slightly modified (see remark at the end of this section).

Let M be an (m,n)-dimensional De Witt supermanifold over the real manifold M_0 with projection Φ: $M \to M_0$. $\mathcal{R}$ denotes again the locally constant sheaf on M with stalk $\mathbb{R}$, and $\mathcal{R}_0$ is the analogous object on M_0. Moreover, in this section $\mathcal{U} = \{U_\alpha, \alpha \in J\}$, with J an ordered set, denotes a good cover of M_0, namely, an open cover such that all nonvoid finite intersections of its members are diffeomorphic to open balls in $\mathbb{R}^m$.

Lemma 5.1: For all $p \geqslant 0$,

$$\check{H}^p(M,\mathcal{R}) \simeq \check{H}^p(M_0,\mathcal{R}_0). \tag{5.1}$$

Proof: As a consequence of Theorem 2.1, we have

$$\check{H}^p(\mathcal{U},\mathcal{R}_0) \simeq \check{H}^p(M_0,\mathcal{R}_0), \quad p \geqslant 0. \tag{5.2}$$

On the other hand, $\mathcal{W} = \{W_\alpha = \Phi^{-1}(U_\alpha), \alpha \in J\}$ is an open cover of M, and it is obvious that

$$\check{H}^p(\mathcal{W},\mathcal{R}) \simeq \check{H}^p(\mathcal{U},\mathcal{R}_0), \quad p \geqslant 0. \tag{5.3}$$

Now, if V is a nonvoid intersection of members of $\mathcal{W}$, V is C^∞ homotopic to the fiber of M, which is a vector space and hence has a vanishing de Rham cohomology. Since $\check{H}^*(V,\mathcal{R}|V)$ is isomorphic to the de Rham cohomology of V, this implies $\check{H}^p(V,\mathcal{R}|V) = 0$, $p \geqslant 1$. Then Theorem 2.1 implies

$$\check{H}^p(\mathcal{W},\mathcal{R}) \simeq \check{H}^p(M,\mathcal{R}), \quad p \geqslant 0. \tag{5.4}$$

Collecting Eqs. (5.2)–(5.4) one gets the proof. □

Corollary 5.1:

$$H^p_{DR}(M) \simeq H^p_{DR}(M_0), \quad p \geqslant 0. \tag{5.5}$$

□

Let $\{\lambda_\alpha, \alpha \in J\}$ be a C^∞ partition of unity on M_0 subordinate to $\mathcal{U}$. Applying to the coordinate expression to each λ_α the mapping $Z_{L',L}$ given in Eq. (3.1), one gets a "tubular" GH^∞ partition of unity $\{\Lambda_\alpha, \alpha \in J\}$ on M subordinate to $\mathcal{W} = \Phi^{-1}(\mathcal{U})$. Then the classical proof of the vanishing of the Čech cohomology of the sheaves of differential forms on a real manifold, obtained by means of a partition of unity argument,[5] can be adapted to show the following result.

Lemma 5.2: Let $\mathcal{W} = \{W_\alpha, \alpha \in J\}$ be the open cover of M obtained as above. For all $q \geqslant 0$ the long sequence of Abelian groups

$$0 \to \mathcal{G}^q(M) \overset{\text{restr}}{\to} \prod_\alpha \mathcal{G}^q(W_\alpha) \overset{\delta}{\to} \prod_{\alpha<\beta} \mathcal{G}^q(W_\alpha \cap W_\beta) \overset{\delta}{\to} \cdots$$

is exact, that is, $\check{H}^p(\mathcal{W},\mathcal{G}^q) = 0$, $p \geqslant 1$. □

We can now prove the main result of this section.

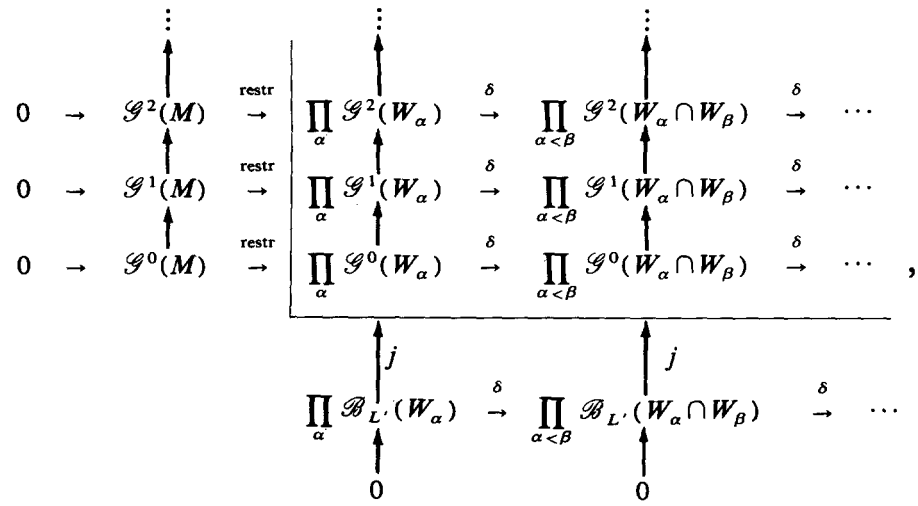
Theorem 5.1: For all $p \geqslant 0$ there is an isomorphism of

graded right $B_{L'}$ modules

$$H^p_{\mathrm{SDR}}(M) \simeq H^p_{\mathrm{DR}}(M_0) \otimes B_{L'}, \tag{5.6}$$

where the right-hand side is given a structure of graded right $B_{L'}$ module by setting $(\omega \otimes a)b = \omega \otimes (ab)$.

Proof: Let us consider the double complex



where $\mathscr{B}_{L'}$ is the locally constant sheaf on M with stalk $B_{L'}$ and the vertical arrows above the horizontal line are given by the exterior differential d. The columns on the right of the vertical line are exact as a consequence of Poincaré's Lemma for GH^∞ forms (see the Appendix), and the rows above the horizontal line are exact as a consequence of Lemma 5.2. Then a general result in homological algebra[3] implies that the cohomologies of the initial column and of the bottom line are isomorphic, namely,

$$H^p_{\mathrm{SDR}}(M) \simeq \check{H}^p(M, \mathscr{B}_{L'}), \quad p \geqslant 0.$$

This isomorphism, together with Eqs. (4.4) and (5.5), establish Eq. (5.6) as a group isomorphism. That (5.6) is also a morphism of graded $B_{L'}$ modules is proved by direct computation. □

Summing up, we have shown that, given a De Witt supermanifold M with body M_0, the following cohomologies are all isomorphic: (i) de Rham cohomology of GH^∞ differential forms of M; (ii) de Rham cohomology of $B_{L'}$-valued C^∞ differential forms on M; (iii) de Rham cohomology of $\mathscr{B}_{L'}$-valued C^∞ differential forms on M_0; (iv) Čech cohomology of the locally constant sheaf $\mathscr{B}_{L'}$ with stalk $B_{L'}$ on M. Moreover, these isomorphisms are actually isomorphisms of graded $B_{L'}$ modules.

Remark: If the De Witt supermanifold M is endowed with the coarse topology, not all the results of this section apply, due to the fact that M is not Hausdorff and therefore not paracompact. However, Eq. (5.3) still holds. Since in the coarse topology all the open covers are obtained by pulling back open covers of M_0, the direct limit involved in the definition of $\check{H}^*(M,\mathscr{R})$ can be taken over covers of the type of $\mathscr{W}$, so that Eq. (5.1) holds. Lemma 5.2 being still valid, Theorem 5.1 follows again.

VI. EXAMPLES

We proceed now to the explicit computation of the SDR cohomology of three supermanifolds having nontrivial topologies.

Example 6.1: $M = S^1 \times \mathbb{R}$ endowed with a structure of (1,0)-dimensional GH^∞ supermanifold. We take $L = L' = 2$; B_L has a basis $\{1, \beta_1, \beta_2, \beta_3 = \beta_1\beta_2\}$. We choose two charts $(x, U_1 \times \mathbb{R})$ and $(y, U_2 \times \mathbb{R})$, where U_1 (U_2) is S^1 without the north pole (south pole), x and y are given in terms of $z \in \mathbb{R}$ and the stereographic angles θ, ϕ, respectively, from the north and south pole, as follows:

$$x = e^z + \tan\frac{\theta}{2}\beta_3, \quad -\frac{\pi}{2} < \theta < \frac{\pi}{2};$$

$$y = e^{-2z}\left[e^z - \tan\left(\frac{\pi}{4} - \frac{\phi}{2}\right)\beta_3\right], \quad -\frac{\pi}{2} < \phi < \frac{\pi}{2}.$$

It is easily shown that x and y are C^∞ diffeomorphisms and that the transition functions $x(y)$ and $y(x)$ are GH^∞.

A direct calculation shows that a global GH^∞ function on M has the form

$$f = K + f^i\beta_i, \quad i = 1,2,3,$$

where the constant K and the C^∞ functions f^i of z are real valued. Then

$$\mathscr{G}(M) = \mathscr{G}^0(M) = \mathbb{R} \oplus [C^\infty(\mathbb{R}) \otimes N], \tag{6.1}$$

where N is the nilpotent ideal of B_L and $C^\infty(\mathbb{R})$ is the vector space of C^∞ real functions on $\mathbb{R}$. Moreover, denoting by $\mathscr{Z}^q$ ($\mathscr{B}^q$) the sheaf of closed (exact) q-forms, one has

$$\mathscr{G}^1(M) = \mathscr{Z}^1(M) = [C^\infty(\mathbb{R}) \otimes N] \oplus [\mathbb{R} \otimes \beta_3],$$

$$\mathscr{B}^1(M) = C^\infty(\mathbb{R}) \otimes N,$$

whence

$$H^1_{\mathrm{SDR}}(M) = \mathscr{Z}^1(M)/\mathscr{B}^1(M) = \mathbb{R} \otimes \beta_3.$$

The unique $B_{L'}$ module structure of $H^1_{\mathrm{SDR}}(M)$ which makes the projection onto the quotient a morphism of $B_{L'}$ modules is given by $(r \otimes \beta_3)a = r \otimes (\beta_3 a)$. Clearly, $H^1_{\mathrm{SDR}}(M)$ is not free as a $B_{L'}$ module.

Obviously, since $\mathscr{G}^q = 0$ and $q > 1$, we have

$H^p_{\mathrm{SDR}}(M) = 0$ for $p > 1$. On the other hand,

$$H^1_{\mathrm{DR}}(M) \otimes B_{L'} = B_{L'},$$

so that in this case the SDR and the $B_{L'}$-valued de Rham cohomology are different. Then Theorem 4.2 implies $\check{H}^1(M, \mathscr{G}^0) \neq 0$.

Example 6.2: Here $M = T^2 \times \mathbb{R}^2$, where T^2 is the two-dimensional torus. Here M is endowed with a structure of $(1,1)$-dimensional GH^∞ supermanifold, with $L = 2$, $L' = 1$. If z, ξ are local real coordinates in T^2, and u, t, real coordinates in $\mathbb{R}^2$, we put in M local B_L-valued coordinates $x = z + u\beta_3$, $y = \xi\beta_1 + t\beta_2$. A direct computation shows that the global GH^∞ functions on M have the form

$$f = \alpha + \gamma\beta_1 + [u\alpha' - t\mu]\beta_3, \tag{6.2}$$

where a, γ, and μ are periodic real valued functions of z and $'$ denotes differentiation. So we have

$$\mathscr{G}^0(M) = C^\infty(S^1) \oplus [C^\infty(S^1) \otimes \beta_1] \oplus [C^\infty(S^1) \otimes \beta_3]$$

equipped with the structure of $B_{L'}$ module given by the wedge product of elements of the form (6.2) by elements of $B_{L'}$. Standard computations show that

$$\mathscr{Z}^1(M) = C^\infty(S^1) \oplus [C^\infty(S^1) \otimes \beta_1] \oplus [\mathbb{R} \otimes \beta_2] \oplus [C^\infty(S^1) \oplus \mathbb{R}] \otimes \beta_3,$$

$$\mathscr{B}^1(M) = C^\infty_0(S^1) \oplus [C^\infty_0(S^1) \otimes \beta_1] \oplus [C^\infty(S^1) \otimes \beta_3],$$

where $C^\infty_0(S^1)$ denotes the space of functions in $C^\infty(S^1)$ whose integral over S^1 vanishes. Taking the quotient gives

$$H^1_{\mathrm{SDR}}(M) = B_L,$$

where B_L has its canonical structure of $B_{L'}$ module. On the other hand, one has

$$H^1_{\mathrm{DR}}(M) \otimes B_{L'} = B_{L'} \oplus B_{L'}, \tag{6.3}$$

so that, acording to Theorem 4.2, we must have $\check{H}^1(M, \mathscr{G}^0) \neq 0$ again.

Example 6.3: The same underlying real manifold as in Example 6.2 but with a different GH^∞ structure, obtained by letting $x = z + \xi\beta_3, y = u\beta_1 + t\beta_2$. Now a global function on M has the form

$$f = K_1 + [\alpha + K_2 u]\beta_1 + K_2 t\beta_2 + t\gamma\beta_3,$$

where K_1, K_2 are real constants and α, γ are real valued periodic functions of z, so that

$$\mathscr{G}^0(M) = \mathbb{R} \oplus [C^\infty(S^1) \otimes \beta_1] \oplus [\mathbb{R} \otimes \beta_2] \oplus [C^\infty(S^1) \otimes \beta_3].$$

An explicit computation shows that $\mathscr{Z}^1(M) \simeq \mathscr{G}^0(M)$ as $B_{L'}$ modules and that

$$\mathscr{B}^1(M) = [C^\infty_0(S^1) \otimes \beta_1] \oplus [\mathbb{R} \otimes \beta_2] \oplus [C^\infty(S^1) \otimes \beta_3],$$

whence

$$H^1_{\mathrm{SDR}}(M) = B_{L'}$$

with its natural structure of $B_{L'}$ module. Obviously, the $B_{L'}$-valued de Rham cohomology of M is given by Eq. (6.3) again.

The supermanifolds in Examples 6.2 and 6.3 have the same underlying real manifold, but their SDR cohomologies are different. Therefore SDR cohomology is neither a topological nor a real differentiable invariant, while, according to Theorem 4.1, it is a superdifferentiable invariant; so it carries information about the superdifferentiable structure of a supermanifold.

VII. FINAL REMARK

In this paper we have considered the category of supermanifolds defined in terms of GH^∞ functions, because in that case the modules of derivations of functions are free and the differential geometry of supermanifolds can be studied in terms of local coordinates. However, one can stick to the choice of G^∞ functions originally introduced by Rogers,[6] and all the results of this paper are still true, provided that L' is everywhere replaced by L.

ACKNOWLEDGMENT

This research was partly supported by the National Group for Mathematical Physics of the Italian Research Council (CNR) and by the Italian Ministry for Public Education through the research project "Geometria e Fisica."

APPENDIX: GH^∞ POINCARÉ LEMMA

Though the statement and the proof of Poincaré's lemma for GH^∞ forms are a straightforward adaptation of the classical result, for the sake of completeness we report it here.

Lemma A.1: Let U be a star-shaped open subset of $B^{m,n}_L$, and let ω be a closed GH^∞ p-form on U, $p \geqslant 1$. There is a GH^∞ $(p-1)$-form η on U such that $\omega = d\eta$.

Proof: We may assume that U is star-shaped with respect to 0. Define a homotopy operator K: $\mathscr{G}^p(U) \to \mathscr{G}^{p-1}(U)$ as follows: if $\omega = dx^{A_p} \wedge \cdots \wedge dx^{A_1} \times \omega_{A_1 \cdots A_p}(x)$, with $x \in U$, and $A_i = 1 \cdots m + n$,

$$K\omega = (-1)^p p\, dx^{A_{p-1}} \wedge \cdots \wedge dx^{A_1} x^B \times \int_0^1 t^{p-1} \omega_{BA_1 \cdots A_{p-1}}(tx)\, dt.$$

A direct computation yields $dK\omega + Kd\omega = \omega$; moreover it is easily shown that $K\omega$ is GH^∞. Then setting $\eta = K\omega$ one gets the proof. □

[1] B. Kostant, in *Lecture Notes in Mathematics*, Vol. 570 (Springer, Berlin, 1977), p. 177.
[2] J. Rabin, Commun. Math. Phys. **108**, 375 (1987).
[3] R. Godement, *Théorie des Faisceaux* (Hermann, Paris, 1964).
[4] F. W. Warner, *Foundations of Differential Geometry and Lie Groups* (Scott, Foresman, Glenview, IL 1971).
[5] R. Bott, and L. W. Tu, *Differential Forms in Algebraic Topology* (Springer, Berlin, 1982).
[6] A. Rogers, J. Math. Phys. **21**, 1352 (1980).
[7] A. Rogers, Commun. Math. Phys. **105**, 375 (1986).
[8] M. J. Rothstein, Trans. Am. Math. Soc. **297**, 159 (1986).
[9] A. Rogers, J. Math. Phys. **22**, 443 (1981).
[10] B. De Witt, *Supermanifolds* (Cambridge U. P., London, 1984).
[11] P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (Wiley, New York, 1978).

Canonoid transformations and constants of motion

Luiz J. Negri, Luimar C. Oliveira, and Julio M. Teixeira
Departamento de Física, Universidade Federal da Paraíba-João Pessoa, PB-Brazil 58.000

(Received 24 February 1987; accepted for publication 13 May 1987)

The necessary and sufficient conditions for a canonoid transformation with respect to a given Hamiltonian are obtained in terms of the Lagrange brackets of the transformation. The relation of these conditions with the constants of motion is discussed.

I. INTRODUCTION

The usual Hamiltonian description of mechanical systems with N degrees of freedom is formulated in a $2N$-dimensional space, the phase space, where (q_i,p_i), $i=1,...,N$, specify the canonical variables. Alternatively one can introduce a compact notation in which (q_i,p_i) are treated as components ξ_μ; $\mu=1,...,2N$, of a single entity $\xi=(\xi_1,...,\xi_N, \xi_{N+1},...,\xi_{2N})\equiv(q_1,...,q_N,p_1,...,p_N)$. In this notation the canonical equations of motion corresponding to a given Hamiltonian $H(q,p,t)=H(\xi,t)$ are written as

$$\dot{\xi}_\alpha=\gamma_{\alpha\beta}H_{,\beta}, \tag{1.1}$$

where $H_{,\beta}=\partial H/\partial\xi_\beta$ and

$$||\gamma_{\nu\beta}||=\begin{vmatrix} 0_N & 1_N \\ -1_N & 0_N \end{vmatrix} \tag{1.2}$$

such that[1,2]

$$\gamma_{\alpha\beta}\cdot\gamma_{\alpha\nu}=\delta_{\beta\nu}, \tag{1.3}$$

$$\gamma_{\alpha\beta}+\gamma_{\beta\alpha}=0. \tag{1.4}$$

Also, given any two dynamical variables $R(\xi,t)$, $S(\xi,t)$ the Poisson bracket (PB) is defined as

$$[S,R]_\xi=S_{,\alpha}\,\gamma_{\alpha\beta}\,R_{,\beta}. \tag{1.5}$$

In this equation and everywhere in this paper we denote $\partial\psi/\partial\xi_\alpha$ by $\psi_{,\alpha}$ for any function ψ.

Now, given a Hamiltonian description (ξ,H) let

$$\eta_\alpha=\eta_\alpha(\xi,t) \tag{1.6}$$

be an invertible transformation on phase space. This transformation is called canonoid with respect to $H(\xi)$ if there exists a function $K(\eta,t)$ such that[1,3]

$$\dot{\eta}_\alpha=\gamma_{\alpha\beta}\frac{\partial K}{\partial\eta_\beta}. \tag{1.7}$$

For $N=1$ it was recently shown[4] that (1.6) represents a canonoid transformation with respect to $H(\xi)$ if and only if the PB $[\eta_\alpha,\eta_\beta]_\xi$, $\alpha,\beta=1,2$, is a constant of motion for the (ξ,H) system.

The main goal of the present paper is to discuss the generalization of this result for systems with $N>1$ degrees of freedom. As we shall prove, the foregoing result is a peculiarity of one-dimensional systems not being necessarily correct in the general case. In Sec. II we show that a generalization to $N\geqslant 1$ of the result from Ref. 4 leads to a necessary but not sufficient condition for a canonoid transformation. The necessary and sufficient conditions are obtained in Sec. III in which we also establish a systematic procedure of constructing a canonoid map for a given Hamiltonian function. The main results of this section are resumed in the form of two theorems. In Sec. IV we give some applications.

II. NECESSARY CONDITIONS FOR A CANONOID TRANSFORMATION

Given a Hamiltonian system (ξ,H) we can look at Eqs. (1.6) as defining a set of $2N$ dynamical variables. Hence setting

$$A_{\alpha\beta}=[\eta_\alpha,\eta_\beta]_\xi \tag{2.1}$$

we can use the Poisson bracket theorem[1] writing

$$\dot{A}_{\alpha\beta}=[\dot{\eta}_\alpha,\eta_\beta]_\xi+[\eta_\alpha,\dot{\eta}_\beta]_\xi, \tag{2.2}$$

where the dot over a letter has its usual meaning of indicating time derivative and Greek indices $\alpha,\beta,\mu,\nu,...$ will be assumed to range from 1 to $2N$. In what follows we shall also assume the summation convention for repeated indices. Using definition (1.5) we have

$$\dot{A}_{\alpha\beta}=\frac{\partial\dot{\eta}_\alpha}{\partial\eta_\nu}A_{\nu\beta}-\frac{\partial\dot{\eta}_\beta}{\partial\eta_\nu}A_{\nu\alpha}$$

from which we obtain

$$\gamma_{\alpha\beta}\,\dot{A}_{\alpha\beta}=B_{\beta\nu}\,A_{\nu\beta}, \tag{2.3}$$

where

$$B_{\beta\nu}\equiv\gamma_{\alpha\beta}\frac{\partial\dot{\eta}_\alpha}{\partial\eta_\nu}-\gamma_{\alpha\nu}\frac{\partial\dot{\eta}_\alpha}{\partial\eta_\beta}. \tag{2.4}$$

Now let us assume that Eqs. (1.6) stand for a canonoid transformation (CT), with respect to $H(\xi,t)$. In this case there will exist a function, say $K(\eta,t)$, the new Hamiltonian, such that Eqs. (1.7) hold, and from Eq. (2.3) we obtain

$$\gamma_{\alpha\beta}\dot{A}_{\alpha\beta}=0, \tag{2.5}$$

which means that the trace of $\|\gamma_{\alpha\beta}A_{\beta\nu}\|$ is a constant of motion for the (ξ,H) system, i.e.,

$$\gamma_{\alpha\beta}\,A_{\alpha\beta}=\text{constant of motion}. \tag{2.6}$$

Thus (2.6) is a necessary condition for the map $\xi\rightleftarrows\eta$ to be a CT. But it is not sufficient: the validity of (2.6) does not imply (for $N>1$) the existence of $K(\eta)$. Note that for $N=1$ we obtain $B_{12}=B_{21}$ as a consequence of (2.6) and from the definitions (2.4) the existence of some $K(\eta)$ function follows. Thus for one-dimensional systems conditions (2.6) are necessary and sufficient for a CT. Indeed, in the most common usage $\xi_\alpha\to(q_i,p_i)$, $\eta_\alpha\to(Q_i,P_i)$, $i=1,...,N$, condition (2.6) reads

 0022-2488/87/102369-04$02.50 © 1987 American Institute of Physics

$$\sum_{j=1}^{N} [Q_j,P_j]_{(q,p)} = \text{constant of motion}, \tag{2.7}$$

which for $N=1$ reduces to the previously mentioned result of Ref. 4.

Summing up the foregoing results we can say that condition (2.6), although necessary, is not sufficient to guarantee the "canonoidicity" of a transformation and the result of Ref. 4 cannot immediately be extended for the $N>1$ case. This will be done in the next section.

III. NECESSARY AND SUFFICIENT CONDITIONS FOR A CT

In order to discuss the characterization of a CT let us consider more fully the objects $B_{\alpha\beta}$ defined in (2.4). For simplicity we restrict ourselves to time-independent invertible maps,

$$\xi \rightleftarrows \eta, \quad \eta_\alpha = \eta_\alpha(\xi). \tag{3.1}$$

We start by considering (3.1) as a set of CT for a given Hamiltonian $H(\xi)$. In this case there will exist some $K(\eta)$ such that Eqs. (1.7) hold. Using those equations we obtain $B_{\mu\nu}=0$. For the converse, i.e., starting with $B_{\mu\nu}=0$, it follows from (2.4) that

$$\frac{\partial \psi_\beta}{\partial \eta_\nu} = \frac{\partial \psi_\nu}{\partial \eta_\beta},$$

where $\psi_\beta = \gamma_{\alpha\beta}\dot{\eta}_\alpha$. Hence, as is well known, there exists a function, say $K(\eta)$, such that

$$\psi_\beta = \frac{\partial K}{\partial \eta_\beta},$$

and Eqs. (1.7) follow immediately. Thus we have shown that for a CT we have $B_{\mu\nu}=0$ and conversely when $B_{\mu\nu}=0$ Eqs. (3.1) stand for a CT.

Now, in contrast to the canonical transformations a CT requires the specification of a Hamiltonian, and so this Hamiltonian is an important piece in the analysis of a CT on phase space. Indeed, if one looks at Eqs. (3.1) as defining $2N$ dynamical variables for a given (ξ,H) system one can write, from (1.1),

$$\dot{\eta}_\alpha = \eta_{\alpha,\mu}\gamma_{\mu\rho}H_{,\rho}. \tag{3.2}$$

Using these relations it is not difficult to obtain the result

$$B_{\mu\nu}\eta_{\nu,\alpha}\eta_{\mu,\beta} = t_{\alpha\beta} - t_{\beta\alpha}, \tag{3.3}$$

where

$$t_{\alpha\beta} = P_{\alpha,\beta}, \tag{3.4}$$

$$P_\alpha = l_{\alpha\mu}\gamma_{\mu\nu}H_{,\nu}, \tag{3.5}$$

$$l_{\alpha\beta} = \{\xi_\alpha,\xi_\beta\}_\eta = \eta_{\mu,\alpha}\gamma_{\mu\nu}\eta_{\nu,\beta} = -l_{\beta\alpha}. \tag{3.6}$$

Notice that $l_{\alpha\beta}$ defined by (3.6) are the so-called[1,2] Lagrange brackets of the ξ's with respect to the η's. Now, due to the assumption of invertibility of (3.1) we can rewrite (3.3) as

$$B_{\mu\nu} = (t_{\alpha\beta} - t_{\beta\alpha})\frac{\partial \xi_\alpha}{\partial \eta_\mu}\frac{\partial \xi_\beta}{\partial \eta_\nu}. \tag{3.7}$$

Thus from (3.3) and (3.7) it is easily verified that the symmetry of the objects $t_{\alpha\beta}$ implies a CT and vice versa. Hence Eqs. (3.1) will correspond to a CT for a given (ξ,H) system if and only if

$$t_{\mu\nu} = t_{\nu\mu}. \tag{3.8}$$

This result is related to the one obtained in the previous section. In fact, the general result expressed by Eq. (2.3) can be rewritten as

$$\gamma_{\alpha\beta}\dot{A}_{\alpha\beta} = \gamma_{\beta\alpha}(t_{\beta\alpha} - t_{\alpha\beta}) \tag{3.9}$$

after using (3.7). Thus, as before, but now based on conditions (3.8), we see that for a CT the trace of $\|\gamma_{\alpha\beta}A_{\beta\nu}\|$ is a constant of motion but not vice versa, i.e., condition (2.6) did not necessarily imply a CT.

Conditions (3.8) can be exploited to shed some light upon the question of relating constants of motion of a given mechanical system to CT. To this end, using (3.4) and (3.5) we first rewrite (3.8) as

$$(l_{\mu\rho,\nu} + l_{\rho\nu,\mu})\gamma_{\rho\beta}H_{,\beta} + l_{\mu\rho}\gamma_{\rho\beta}H_{,\beta\nu} - l_{\nu\rho}\gamma_{\rho\beta}H_{,\beta\mu} = 0. \tag{3.10}$$

From definitions (3.6) it is easy to verify the following property for the Lagrange bracket $l_{\alpha\beta}$:

$$l_{\alpha\beta,\nu} + l_{\beta\nu,\alpha} + l_{\nu\alpha,\beta} = 0. \tag{3.11}$$

This result permits us to conclude the existence of $2N$ functions $g_\mu(\xi)$ such that

$$l_{\alpha\beta} = g_{\alpha,\beta} - g_{\beta,\alpha}. \tag{3.12}$$

Then, using (3.11) and (3.12) in (3.10), we obtain after some manipulation,

$$(\dot{g}_\mu)_{,\nu} - (\dot{g}_\nu)_{,\mu} + R_{\mu,\nu} - R_{\nu,\mu} \equiv P_{\mu,\nu} - P_{\nu,\mu} = 0, \tag{3.13}$$

where

$$R_\mu = g_\rho\gamma_{\rho\beta}H_{,\beta\mu}. \tag{3.14}$$

Equations (3.13) are more than only a new version of conditions (3.8) in the sense that they allow us to more easily treat the problem of constructing a CT for a given $H(\xi)$. We have the following procedure: for any set of $2N$ dynamical variables $g_\mu(\xi)$ satisfying (3.13) the corresponding $\xi \rightleftarrows \eta$ mapping is obtained by solving the $N(2N-1)$ first-order partial differential equations for $\eta_\alpha(\xi)$ which, in turn, is obtained from (3.6) and (3.12). Of course this could still be rather complicated in some cases but we can restrict ourselves to transformations under which the coordinates in configuration space are preserved, the so-called fouling transformations,[5] which considerably simplify the problem. Indeed, for fouling transformations we have

$$\eta_i = \xi_i, \tag{3.15}$$

$$\eta_{i+N} = f_i(\xi), \tag{3.16}$$

with $i=1,2,\ldots,N$, thus reducing to N the number of unknown quantities.

We also note that for any suitable set $g_\mu(\xi)$ satisfying (3.13) we can associate infinitely many other sets letting

$$g_\mu \rightarrow \bar{g}_\mu = g_\mu + X_{,\mu}, \tag{3.17}$$

with $X = X(\xi)$ arbitrarily chosen functions. This result comes from the invariance of Eqs. (3.12) and (3.13) with respect to (3.17) and, following the usual notion, defines a gauge for the problem. It is also interesting to observe that the gauge-transformation set is the intersection between fouling transformations and canonical transformations (which implies $l_{\alpha\beta}$ equals numerical constants), both constituting subclasses of the CT.[5]

Equations (3.13) admit a simpler and more interesting class of solutions than those corresponding to dynamical variables. In fact if we select $g_\mu(\xi)$ as constants of the motion for the (ξ,H) system in such a way that

$$R_{\mu,\nu} = R_{\nu,\mu}, \tag{3.18}$$

the necessary and sufficient conditions for a CT will be accomplished. Hence we have the following theorem.

Theorem I: Let ξ_α be a set of general coordinates on phase space and (ξ,H) a Hamiltonian description of some mechanical system with N degrees of freedom. Let $g_\mu(\xi)$ be a set of $2N$ constants of motion for this system and $\eta_\alpha(\xi)$ invertible time-independent transformation on phase space constructed so that

$$l_{\alpha\beta} = g_{\alpha,\beta} - g_{\beta,\alpha}.$$

This transformation will be canonoid with respect to $H(\xi)$ if the conditions

$$R_{\mu,\nu} = R_{\nu,\mu}$$

are fulfilled.

The procedure of explicitly obtaining the CT map after the $g_\mu(\xi)$ family of constants is determined follows the same steps as the one described for dynamical variables. On the other hand our technique limits the g_μ family to constants which do not depend explicitly on time. This, in some cases, could become a severe restriction but, fortunately, there is no need of keeping $g_\mu = g_\mu(\xi)$. The more general $g_\mu(\xi,t)$ can be used in writing Eqs. (3.12). In this case conditions (3.13) are changed to

$$\begin{aligned}&(\dot{g}_\mu)_{,\nu} - (\dot{g}_\nu)_{,\mu} + g_{\nu,\mu t} - g_{\mu,\nu t} + R_{\mu,\nu} - R_{\nu,\mu}\\ &\quad = P_{\mu,\nu} - P_{\nu,\mu} = 0\end{aligned} \tag{3.19}$$

so that in addition we must impose the conditions

$$g_{\nu,\mu t} = g_{\mu,\nu t} \tag{3.20}$$

in order to constitute the g_μ family of Theorem I. These conditions mean that $l_{\alpha\beta}$ will not explicitly depend on time thus agreeing with our initial assumption, namely Eqs. (3.1).

There exists an alternative procedure of treating the present problem in which we need not worry about the functions $g_\mu(\xi)$, but may directly use the elements $l_{\alpha\beta}$. Actually, defining the quantities

$$m_{\mu\nu} = l_{\mu\rho}\gamma_{\rho\beta}H_{,\beta\nu} \tag{3.21}$$

it is not difficult to set (3.13) in the following form:

$$\dot{l}_{\mu\nu} + m_{\mu\nu} - m_{\nu\mu} = P_{\mu,\nu} - P_{\nu,\mu} = 0.$$

Hence we have the following theorem.

Theorem II: Let ξ_α be a set of general coordinates on phase space and (ξ,H) a Hamiltonian description of some mechanical system with N degrees of freedom.

Let $\eta_\alpha(\xi)$ be an invertible time-independent transformation on phase space so that

$$l_{\alpha\beta} = \eta_{\mu,\alpha}\gamma_{\mu\nu}\eta_{\nu,\beta}, \quad m_{\mu\nu} = l_{\mu\rho}\gamma_{\rho\beta}H_{,\beta\nu}.$$

This transformation is canonoid with respect to $H(\xi)$ if and only if

$$\dot{l}_{\mu\nu} + m_{\mu\nu} - m_{\nu\mu} = 0. \tag{3.22}$$

It is Theorem II that corresponds to the generalization to the $N > 1$ case of the result presented in Ref. 4 for $N = 1$. Indeed for the particular case $N = 1$ we have $m_{12} = m_{21}$ so that it follows from (3.22) that l_{12} is necessarily a constant of motion. We also have $l_{12} = A_{12}$ for $N = 1$.

We point out that if $l_{\mu\nu}$ are numerical constants [i.e., $g_\mu(\xi)$ are linear on the ξ's] the corresponding map $\xi \rightleftarrows \eta$ will be canonoid with respect to $H(\xi)$ if and only if $m_{\mu\nu} = m_{\nu\mu}$. The particular choice $l_{\mu\nu} = \gamma_{\mu\nu}$ is a solution of (3.22) independent of the initial Hamiltonian function $H(\xi)$: the corresponding map is a canonical transformation which is a subclass of the CT.

Another class of particular solutions is obtained when $l_{\mu\nu}$ are constants of the motion for the (ξ,H) system and $m_{\mu\nu} = m_{\nu\mu}$. The resulting CT map turns out to be rather cumbersome. This and the foregoing results are discussed in the next section where some examples are presented.

IV. EXAMPLES

To avoid unimportant calculations which only obscure the main point we shall restrict ourselves to the case $N = 2$. As a first example consider

$$H = \alpha\xi_1 + \beta\xi_2 + (1/2m)(\xi_3^2 + \xi_4^2), \tag{4.1}$$

where α, β are numerical constants. In this case we have $R_1 = 0, R_2 = 0, R_3 = m^{-1}g_1, R_4 = m^{-1}g_2$. A set of g_μ constants of motion satisfying (3.13) is

$$g_1 = \beta\xi_4 + \beta^2 t, \quad g_2 = \beta\xi_3 - \alpha\xi_4, \quad g_3 = 0, \quad g_4 = 0.$$

Restricting ourselves to fouling transformations, $\eta_1 = \xi_1$, $\eta_2 = \xi_2, \eta_3 = f_1(\xi), \eta_4 = f_2(\xi)$, the corresponding differential equations for the unknown $f_i(\xi)$ are

$$f_{1,2} - f_{2,1} = 0, \quad f_{1,3} = 0,$$
$$f_{1,4} = \beta, \quad f_{2,3} = \beta, \quad f_{2,4} = -\alpha.$$

Thus one possible CT is

$$\eta_1 = \xi_1, \quad \eta_2 = \xi_2, \quad \eta_3 = \beta\xi_4 + \tfrac{1}{2}\xi_2^2,$$
$$\eta_4 = \beta\xi_3 - \alpha\xi_4 + \xi_1\xi_2,$$

and the fouled Hamiltonian is easily found to be

$$\begin{aligned}K &= (1/2m\beta)(2m\beta^3 - 2\eta_2\eta_3 + \eta_2^3)\eta_1\\ &\quad + (1/8m\beta^2)(\alpha\eta_2^2 - 4\alpha\eta_3 - 4\beta\eta_4)\eta_2^2\\ &\quad + (1/2m\beta^2)(\alpha\eta_3 + 2\beta\eta_4)\eta_3.\end{aligned}$$

As a second example consider the two-dimensional isotropic simple harmonic oscillator (2 DISHO). The Hamiltonian is

$$H = \tfrac{1}{2}(\xi_1^2 + \xi_2^2 + \xi_3^2 + \xi_4^2).$$

For $g_1 = 3\xi_3, g_2 = \xi_4, g_3 = \xi_2, g_4 = \xi_1$, conditions (3.13) are satisfied and $l_{\mu\nu}$ are numerical constants. A fouling canonoid transformation corresponding to this choice is

$$\eta_1 = \xi_1, \quad \eta_2 = \xi_2, \quad \eta_3 = 2\xi_3 - \xi_4 + \xi_1\xi_2,$$
$$\eta_4 = \xi_4 - \xi_3 + \tfrac{1}{2}\xi_1^2,$$

and the fouled Hamiltonian is found to be

$$K = \tfrac{1}{4}(\eta_1 + 2\eta_2)\eta_1^3 + \tfrac{1}{2}(2 - \eta_3 - 2\eta_4 + \eta_2^2)\eta_1^2 + \tfrac{1}{2}(\eta_2^2 + \eta_3^2 + 2\eta_4^2) - (\eta_3 + \eta_4 + 1)\eta_1\eta_2 + \eta_3\eta_4.$$

A rather complicated fouling transformation is obtained when one decides to specify $l_{\mu\nu}$ as constants of motion. For instance, with $l_{12} = 0$, $l_{13} = \frac{1}{2}(\xi_1^2 + \xi_3^2)$, $l_{14} = 0$, $l_{23} = 0$, $l_{24} = \frac{1}{2}(\xi_2^2 + \xi_4^2)$, $l_{34} = 0$, all the requirements of Theorem II are fulfilled and a fouling CT for this case is

$$\eta_1 = \xi_1, \quad \eta_2 = \xi_2, \quad \eta_3 = \tfrac{1}{2}\xi_1^2\xi_3 + \tfrac{1}{6}\xi_3^3,$$
$$\eta_4 = \tfrac{1}{2}\xi_2^2\xi_4 + \tfrac{1}{6}\xi_4^3,$$

for which the inverse map is

$$\xi_1 = \eta_1, \quad \xi_2 = \eta_2, \quad \xi_3 = b\,[(u+2)^{1/3} - u^{1/3}],$$
$$\xi_4 = a[(v+2)^{1/3} - v^{1/3}],$$

where

$$a^3 = 3\eta_4, \quad b^3 = 3\eta_3, \quad a^3 v = (a^6 + \eta_2^6)^{1/2} - a^3,$$
$$b^3 u = (b^6 + \eta_1^6)^{1/2} - b^3.$$

The corresponding fouled Hamiltonian is given by

$$K = \tfrac{3}{8}(\eta_1^4 + \eta_2^4) - (a^4/8) \times [v^{1/3}(v+4) + (v+2)^{1/3}(v-2)] - (b^4/8)[u^{1/3}(u+4) + (u+2)^{1/3}(u-2)].$$

Despite its complicated form this Hamiltonian function also describes the 2 DISHO.

[1] E. J. Saletan and A. H. Cromer, *Theoretical Mechanics* (Wiley, New York, 1971).
[2] E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974).
[3] D. G. Currie and E. J. Saletan, Nuovo Cimento B **9**, 143 (1972).
[4] C. Leubner and Monika A. M. Marte, Phys. Lett. A **101**, 179 (1984).
[5] Y. Gelman and E. J. Saletan, Nuovo Cimento B **18**, 53 (1973).

Decomposition of Lorentz transformations

D. Han
National Aeronautics and Space Administration, Goddard Space Flight Center (Code 636), Greenbelt, Maryland 20771

Y. S. Kim
Department of Physics and Astronomy, University of Maryland, College Park, Maryland 20742

D. Son
Department of Physics, Kyungpook National University, Taegu 635, Korea

(Received 6 March 1987; accepted for publication 10 June 1987)

It is shown that every Lorentz transformation can be decomposed into a helicity-preserving transformation that changes the momentum of a free particle and a helicity-changing transformation that leaves the momentum invariant. Since momentum-preserving transformations constitute a subgroup of the Lorentz group, helicity-preserving transformations form a coset space. It is shown further that, for massive particles, every Lorentz transformation can be decomposed into the Wigner rotation and helicity-preserving transformations. For massless particles, every Lorentz transformation can be decomposed into the gauge transformation and helicity-preserving transformation. The gauge transformation in this case is a Lorentz-boosted Wigner rotation.

I. INTRODUCTION

In his 1957 paper on relativistic invariance and quantum phenomena,[1] Wigner noted that there are Lorentz transformations that preserve helicity and those that do not.[2] He suggested that the difference between these two different sets of transformations may play an important role in understanding the internal space-time symmetries of elementary particles, particularly the symmetry of massless particles as a limiting case of the space-time symmetry of massive particles.

In his earlier work,[3] Wigner studied systematically the subgroups of the Lorentz group that leave the four-momentum of a given particle invariant. These subgroups, which are called the little groups, have been extensively discussed in the literature.[4,5] The transformations of the little group do not leave the helicity invariant.

Since the little group is a subgroup of the Lorentz group, it is of interest to study the cosets of this subgroup. We are particularly interested in the physical quantity that remains invariant under the transformations of these cosets.

We shall show in this paper that the transformations of these cosets leave the helicity invariant while changing the momentum. We shall therefore establish the following theorem. Every Lorentz transformation can be decomposed into a momentum-preserving transformation and a helicity-preserving transformation. This theorem is applicable to both massive and massless particles *and* to the case in which the massless limit is taken from a massive case.

As Wigner pointed out in 1957,[1] a boost along the direction perpendicular to the momentum does not preserve the helicity. We shall show in this paper that this transformation can be decomposed into a helicity-preserving transformation and a momentum-preserving transformation. The helicity-preserving transformation in this case consists of a boost along the direction of momentum and a rotation around the axis perpendicular to both the momentum and the direction of boost. The momentum-preserving transformation is an element of the little group. This can be done for both massive and massless particles, and the massless case is a special case of the massive case.

The organization of this paper is very similar to that of our previous paper,[6] but the Lorentz kinematics is different. The kinematics of the present paper is designed to illustrate fully the set of helicity-preserving transformations.

In Sec. II, we construct a Lorentz kinematics that enables us to write an arbitrary Lorentz transformation as a product of a helicity-preserving transformation and a momentum-preserving transformation. In Sec. III, the kinematics constructed in Sec. II is compared with the traditional approach to the O(3)-like little group for massive particles. The role of the Wigner rotation is studied in detail. We study also the role of the Wigner rotation in the zero-mass limit. It is shown that, in this limit, the little group becomes a group of gauge transformations applicable to massless particles with spin 1. In Sec. IV, we study the conclusions of Secs. II and III using the SL(2,c) formalism for spin-$\frac{1}{2}$ particles.

II. DECOMPOSITION OF LORENTZ TRANSFORMATIONS

If we perform a Lorentz transformations on a free particle with definite helicity, this applies to the helicity as well as the four-momentum. Lorentz boosts along the direction of momentum changes the magnitude of momentum but leaves the helicity and the direction of momentum unchanged. Rotations around the momentum leave both the helicity and the momentum invariant. Other rotations change the direction of momentum, while preserving the helicity and the magnitude of momentum. These transformations form a set of helicity-preserving transformations. They are capable of transforming the momentum to every possible value in the three-dimensional momentum space.[1,2]

Let us start with a particle at rest with mass m and its

 0022-2488/87/102373-06$02.50 © 1987 American Institute of Physics

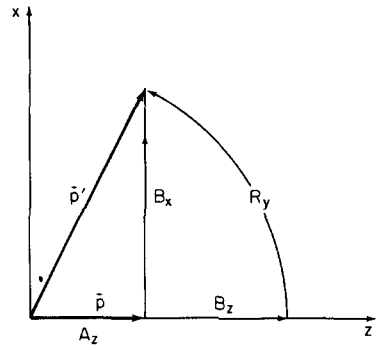


FIG. 1. Lorentz boost along the x direction. The four-momentum p can be boosted to p' either directly by B_x or through the rotation R_y preceded by B_z along the z direction. These operations produce two different effects when applied to the internal space-time coordinates. This figure appears identical to Fig. 1 of Ref. 10, but there is one important difference. This figure is applicable also to massive particles, and allows the Lorentz boost $A_z(\alpha)$.

spin in the z direction, and then boost this particle along the z direction with velocity parameter α, as is illustrated in Fig. 1. In the four-vector convention: $x^\mu = (x,y,z,t)$, the resulting four-momentum p is

$$p^\mu = (0,0,\alpha/a,1/a), \text{ with } a = (1-\alpha^2)^{1/2}. \tag{1}$$

The matrix that boosts the rest state four-momentum to the above form is

$$A_z(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/a & \alpha/a \\ 0 & 0 & \alpha/a & 1/a \end{pmatrix}. \tag{2}$$

After this boost, the particle is in the positive helicity state.

We can boost this particle with nonzero momentum along the z direction without changing the helicity. We can also rotate the system without affecting the helicity. We cannot, however, boost the system along the direction perpendicular to the momentum without changing the helicity. In this case, both the momentum and helicity become changed. We propose to write this transformation as a product of helicity-preserving and momentum-preserving transformations.

Let us take this perpendicular direction to be the x direction, and boost the four-momentum of Eq. (1) along this direction with velocity parameter β,

$$p' = B_x(\beta)p, \tag{3}$$

where

$$B_x(\beta) = \begin{pmatrix} 1/b & 0 & 0 & \beta/b \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \beta/b & 0 & 0 & 1/b \end{pmatrix}, \tag{4}$$

with $b = (1-\beta^2)^{1/2}$. This is not a helicity-preserving transformation.

The boost $B_x(\beta)$ is not the only transformation that changes the four-momentum p to p'. As is illustrated in Fig. 1, we can boost p along the z axis first so that its speed (or magnitude of momentum) is the same as that of p', and then rotate this boosted vector until its direction coincides with that of p',

$$p' = (R_y(\alpha,\beta)B_z(\alpha,\beta))p. \tag{5}$$

Since rotations and boosts along the direction of momentum preserve the helicity, the above transformation is a helicity-preserving transformation. The explicit forms for the matrices are

$$B_z(\alpha,\beta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & (1-\alpha f)/a^2 b & (f-a)/a^2 b \\ 0 & 0 & (f-\alpha)/a^2 b & (1-\alpha f)/a^2 b \end{pmatrix}, \tag{6}$$

and

$$R_y(\alpha,\beta) = \begin{pmatrix} \alpha b/f & 0 & \beta/f & 0 \\ 0 & 1 & 0 & 0 \\ -\beta/f & 0 & \alpha b/f & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{7}$$

where

$$f = (1-a^2b^2)^{1/2} = (\alpha^2+\beta^2-\alpha^2\beta^2)^{1/2}.$$

The boost velocity of B_z is $(f-a)/(1-\alpha f)$. The rotation angle of R_y is

$$\theta = \sin^{-1}(\beta/f). \tag{8}$$

We have seen above that p can be transformed to p' in two different ways. However, these two transformations do not produce the same result when applied to the internal space-time symmetry space. The best way to see this difference is to construct the closed-loop transformation,

$$D_x(\alpha,\beta) = [B_x(\beta)]^{-1}R_y(\alpha,\beta)B_z(\alpha,\beta), \tag{9}$$

described in Fig. 2. The result of the above matrix multiplication is

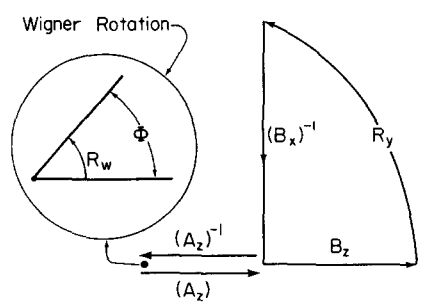


FIG. 2. The difference between the two transformations illustrated in Fig. 1. The difference can best be described by the closed-loop transformation $[(B_x(\beta))^{-1}R_y(\alpha,\beta)B_z(\alpha,\beta)]$. This closed-loop transformation leaves the four-momentum p invariant, and is therefore an element of Wigner's O(3)-like little group if the particle mass does not vanish. According to Wigner's original version of kinematics, we bring the particle to the rest state by applying the inverse of the boost operator $A_z(\alpha)$. We can then perform a Wigner rotation without changing the momentum of the rest particle. We then apply $A_z(\alpha)$ to the rest particle in order to increase its momentum to $\mathbf{p}$. This procedure does not change the four-momentum, but performs nontrivial transformations on the internal space-time structure of the particle. However, this traditional kinematics is possible only for particles with nonzero mass. On the other hand, the closed-loop kinematics is possible for both massive and massless particles.

$$D_x(\alpha,\beta)=\begin{pmatrix} \alpha/f & 0 & \beta/f & -\alpha\beta/f \\ 0 & 1 & 0 & 0 \\ -\beta/f & 0 & \alpha(1-\alpha f)/a^2 f & \alpha(f-\alpha)/a^2 f \\ -\alpha\beta/f & 0 & \alpha(\alpha-f)/a^2 f & (f-\alpha^3)/a^2 f \end{pmatrix}. \tag{10}$$

When applied to the four-momentum p, this matrix leaves it invariant,

$$p = D_x(\alpha,\beta)p. \tag{11}$$

Therefore, the above four-by-four matrix is a representative of the little group that leaves the four-momentum p invariant.

Let us now write B_x of Eq. (4) as

$$B_x = [R_y B_z (B_z)^{-1}(R_y)^{-1}]B_x. \tag{12}$$

Then the right-hand side of the above equation can be rearranged, and

$$\begin{aligned} B_x &= R_y B_z [(B_x)^{-1} R_y B_z]^{-1} \\ &= (R_y(\alpha,\beta)B_z(\alpha,\beta))(D_x(\alpha,\beta))^{-1}. \end{aligned} \tag{13}$$

The transformation $(R_y B_z)$ is a helicity-preserving transformation, but changes the momentum. D^{-1} is also a representative of the little group, but it can change the helicity. Therefore, B_x can be decomposed into a helicity-preserving transformation which changes the momentum and a momentum-preserving tranformation which changes the helicity.

In Eq. (13), $(D_x)^{-1}$ is a representative of the little group, and $(R_y B_z)$ is an element of the left coset consisting of a helicity-preserving transformation. Equation (13) can be written in terms of the right coset,

$$\begin{aligned} B_x &= (R_y B_z)(D_x)^{-1}(R_y B_z)^{-1}(R_y B_z) \\ &= (R_y B_z D_x^{-1} B_z^{-1} R_y^{-1})(R_y B_z). \end{aligned} \tag{14}$$

The transformation $(R_y B_z D^{-1} B_z^{-1} R_y^{-1})$ is a representative of the little group which leaves the four-momentum p' invariant.

III. MASSIVE AND MASSLESS PARTICLES

The kinematics presented in Sec. II is applicable to both massive and massless particles. For massless particles, this kinematics has been discussed in the literature.[5] For a massive particle, the D matrix of Eq. (10) is a representative of the O(3)-like little group. Wigner's original kinematics for this little group is the three-dimensional rotation in the Lorentz frame where the particle is at rest,[3]

$$p = A_z(\alpha)R_W(\Phi)[A_z(\alpha)]^{-1}p, \tag{15}$$

where $R_W(\Phi)$ is a rotation matrix. This means that the particle is brought to its rest frame and then is rotated before it is brought back to its original frame, as is indicated in Fig. 2. The rotation at the rest frame is called the *Wigner rotation*.[6,7]

If the transformations are performed on the xz plane, as in the case of Sec. II, $R_W(\Phi)$ represents a rotation matrix around the y axis,

$$R_W=\begin{pmatrix} \cos\Phi & 0 & \sin\Phi & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\Phi & 0 & \cos\Phi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{16}$$

This rotation matrix leaves the four-momentum invariant in the Lorentz frame in which the momentum is zero. However, this rotation changes the direction of spin. Thus the Wigner rotation is not a helicity-preserving transformation.

If the transformation of Eq. (11) is equivalent to that of Eq. (15), we should be able to write

$$D_x(\alpha,\beta) = A_z(\alpha)R_W(\Phi)[A_z(\alpha)]^{-1}, \tag{17}$$

and the following similarity transformation on $D_x(\alpha,\beta)$ should produce the Wigner rotation matrix:

$$R_W(\alpha,\beta) = [A_z(\alpha)]^{-1}D(\alpha,\beta)A_z(\alpha). \tag{18}$$

The resulting matrix is indeed of the form of Eq. (16), and the Wigner angle is determined from the parameters α and β,

$$\begin{aligned} \Phi &= \sin^{-1}(\beta a/f) \\ &= \sin^{-1}(\beta(1-\alpha^2)^{1/2}/[\alpha^2+\beta^2-\alpha^2\beta^2]^{1/2}). \end{aligned} \tag{19}$$

The D transformation is therefore a Lorentz-boosted Wigner rotation.

For the case of massive particles, every Lorentz transformation can be written in terms of the Wigner rotation and helicity-preserving transformations, since we can replace D^{-1} in Eq. (13) by the inverse of the expression given in Eq. (17),

$$\begin{aligned} B_x &= (R_y B_z)(A_z(\alpha)R_W(-\Phi)A_z^{-1}(a)) \\ &= (R_y(B_z A_z))R_W(-\Phi)A_z^{-1}. \end{aligned} \tag{20}$$

The transformation $(B_z A_z)$ is a boost along the direction of momentum. Every transformation, except R_W, on the right-hand side of the above expression is a helicity-preserving transformation.

Let us next address the question of whether the O(3)-like little group becomes the E(2)-like little group in the $\alpha \to 1$ limit. This has been discussed in terms of the singular transformation known as the Inonu–Wigner group contraction.[8,9] The parameter α that we use here is not the parameter in the Lie group upon which the group contraction method is based. It is therefore not surprising to see that every element in D_x of Eq. (10) is analytic in α at and near $\alpha = 1$.[6]

For this reason, we do not have to make any special effort to take the limiting process. At $\alpha = 1$, the D matrix takes the form

$$D_x(1,\beta)=\begin{pmatrix} 1 & 0 & \beta & -\beta \\ 0 & 1 & 0 & 0 \\ -\beta & 0 & 1-\beta^2/2 & \beta^2/2 \\ -\beta & 0 & -\beta^2/2 & 1+\beta^2/2 \end{pmatrix}. \tag{21}$$

This matrix as an element of the E(2)-like little group was given in Wigner's original paper,[3] and discussed repeatedly in the literature since then as a gauge transformation matrix.[5,10] However, there is one difference. The magnitude of β cannot exceed one in Eq. (21), while the parameters in the

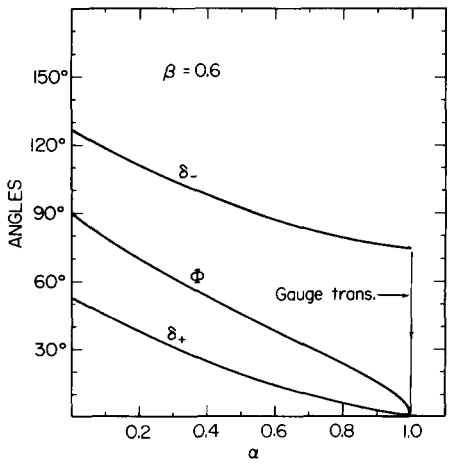


FIG. 3. The Wigner angle and the angular separation between the direction of the momentum and that of the spin in the SL(2,c) regime as functions of β for a given value of α. As $\beta \to 1$, the particle speed approaches that of light, and δ_+ vanishes, but δ_- becomes 180°.

E(2)-like little group can be made arbitrarily large. We can bridge this gap by observing the fact that the D transformation can be repeated in the following manner:

$$D_x(1,\beta_2)D_x(1,\beta_1) = D_x(1,\beta_1+\beta_2). \tag{22}$$

The parameter β in Eq. (21) can be replaced by $(\beta_1+\beta_2+\cdots)$ which can become arbitrarily large.[11]

The expression given for D in Eq. (17) is still valid, and the Lorentz boosted Wigner rotation is a gauge transformation.[6] As is indicated in Fig. 3, the Wigner rotation angle Φ vanishes as $\alpha \to 1$. However, the parameters in the boost matrix A_z become infinite to make the elements of the D matrix remain finite.

IV. PARTICLES WITH SPIN ½

The purpose of this section is to study what we did in Secs. II and III in terms of SL(2,c) for spin-½ particles. While the generators of rotations in SL(2,c) are $S_i = \frac{1}{2}\sigma_i$, the boost generators can take two different signs: $K_i = (\pm)(i/2)\sigma_i$.[5,6]

Let us start with a massive particle at rest, and the usual normalized Pauli spinors χ_+ and χ_- for the spin in the positive and negative z directions, respectively. If we take into account Lorentz boosts, there are four spinors. We shall use the notation $\chi_\pm$ for which the boost generators $K_i = (i/2)\sigma_i$ are applicable, and $\dot{\chi}_\pm$ to which $K_i = -((i/2)\sigma_i)$ are applicable.

The boost matrix that brings the spinor $\chi_\pm$ and $\dot{\chi}_\pm$ from the zero-momentum state to that of p is

$$A_z^{(\pm)}(\alpha) = \begin{pmatrix} N_\pm & 0 \\ 0 & N_\mp \end{pmatrix}, \tag{23}$$

with

$$N_\pm = [(1\pm\alpha)/(1\mp\alpha)]^{1/4}.$$

We use the superscripts $(+)$ and $(-)$ for the undotted and dotted spinors, respectively.

If this matrix is applied to the spinors at rest $\chi_\pm$ and $\dot{\chi}_\pm$,

$$\chi_\pm(p) = A_z^{(+)}(\alpha)\chi_\pm, \text{ and } \dot{\chi}_\pm(p) = A_z^{(-)}(\alpha)\dot{\chi}_\pm, \tag{24}$$

it produces the spinors for the particle with the four-momentum p,

$$\chi_\pm(p) = N_\pm\chi_\pm, \quad \dot{\chi}_\pm(p) = N_\mp\dot{\chi}_\pm. \tag{25}$$

The subscripts $+$ and $-$ denote in this case positive and negative helicities, respectively.

Let us next boost the above spinors along the x direction,

$$\chi'_\pm(p') = B_x^{(+)}(\beta)\chi_\pm(p), \quad \dot{\chi}'(p') = B_x^{(-)}(\beta)\dot{\chi}_\pm(p). \tag{26}$$

The boost matrix takes the form

$$B_x^{(\pm)}(\beta) = \begin{pmatrix} ((1+b)/2b)^{1/2} & \pm((1-b)/2b)^{1/2} \\ \pm((1-b)/2b)^{1/2} & ((1+b)/2b)^{1/2} \end{pmatrix}, \tag{27}$$

where b is defined in Eq. (4). The new spinors of Eq. (26) can be written as

$$\chi'_\pm(p') = N_\pm\chi'_\pm, \quad \dot{\chi}'_\pm(p) = N_\mp\dot{\chi}'_\pm, \tag{28}$$

where

$$\chi'_\pm = \begin{pmatrix} ((1\pm b)/2b)^{1/2} \\ ((1\mp b)/2b)^{1/2} \end{pmatrix}, \quad \dot{\chi}'_\pm = \begin{pmatrix} \pm((1+b)/2b)^{1/2} \\ \mp((1\mp b)/2b)^{1/2} \end{pmatrix}.$$

This boost is not a helicity-preserving transformation. The spin directions represented by the above spinors are $\pm\sin^{-1}(\beta)$. These new spinors do not represent the spins parallel or antiparallel to the new momentum $\mathbf{p}'$. The angle between $\mathbf{p}'$ and the z axis is given in Eq. (8). The angle between the momentum and the spin direction is

$$\delta_\pm = \sin^{-1}(\beta/[\alpha^2+\beta^2-\alpha^2\beta^2]^{1/2}) \pm \sin^{-1}(\beta), \tag{29}$$

as is described in Fig. 4.

The angles δ_+ and δ_- are plotted in Fig. 3 against β for a fixed value of α. Since we are starting with spins that are parallel and antiparallel to the momentum, these angles are zero for $\beta = 0$. When $\beta \to 1$, one of the spins become parallel to the momentum ($\delta_+ \to 0$), but the other becomes antiparallel ($\delta_- \to 180°$).

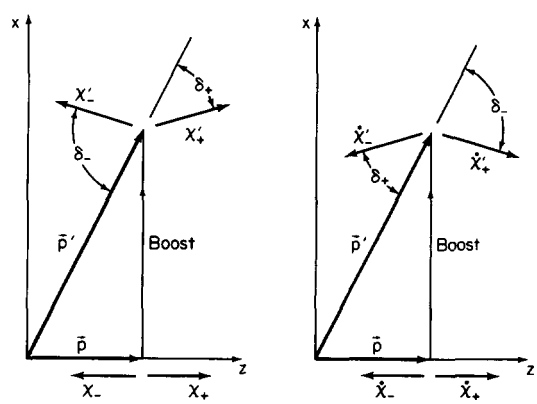


FIG. 4. The Wigner angle and the angular separation between the direction of the momentum and that of the spin in the SL(2,c) regime as functions of α for a given value of β. As $\alpha \to 1$, δ_+ becomes zero, but δ_- does not. This nonvanishing angular separation is the source of gauge degrees of freedom.

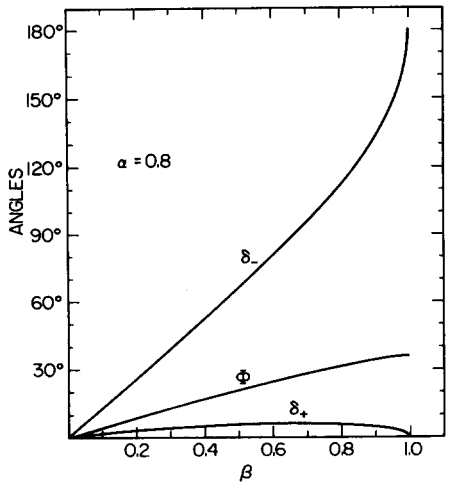


FIG. 5. The angular separation between the direction of momentum and that of the spin in the SL(2,c) regime. As we boost the spinor along the x direction, the spin of orientation angle depends only on β, while the direction of the momentum depends on both α and β. All the spin directions are rotated clockwise from those of the helicity states. This is due to the clockwise Wigner rotation in the rest frame.

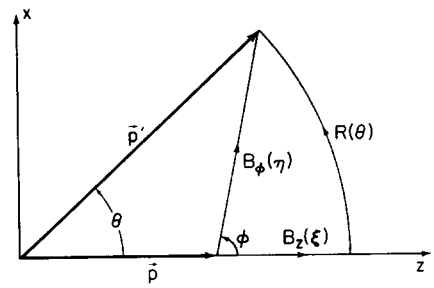


FIG. 6. Lorentz boost along an arbitrary direction. The procedure developed in the present paper is applicable to this general kinematics.

In order to study the case for massless particles, let us go to Fig. 3 where δ_+ and δ_- are plotted as functions of α for a fixed value of β. Here again, the δ_+ becomes zero in the $\alpha \to \pm 1$ limit, while δ_- does not. Indeed, for one of the two spin orientations, every Lorentz transformation is a helicity-preserving transformation as the momentum/mass becomes infinite, as was pointed out by Wigner.[1] However, for the other spin orientation, the spin direction never coincides with the direction of momentum. This is illustrated also in Fig. 5 where the angles are plotted against β for a fixed value of α.

This lack of spin alignment is the origin of the gauge degrees of freedom. In order to see this, let us calculate the D matrix from the closed-loop kinematics of Fig. 2. Its form is

$$D_x^{(\pm)}(\alpha,\beta) = \begin{pmatrix} \left(\frac{f+|\alpha|}{2f}\right)^{1/2} & \left(\frac{(1\pm\alpha)(f-|\alpha|)}{2(1\mp\alpha)f}\right)^{1/2} \\ \left(\frac{(1\mp\alpha)(f-|\alpha|)}{2(1\pm\alpha)}\right)^{1/2} & \left(\frac{f+|\alpha|}{2f}\right)^{1/2} \end{pmatrix}. \tag{30}$$

The boost matrix $B_x^{(\pm)}(\beta)$ can now be written as

$$B_x^{(\pm)}(\beta) = [R_y(\alpha,\beta)B_z^{(\pm)}(\alpha,\beta)]D^{(\pm)}(\alpha,\beta), \tag{31}$$

where $[R_y(\alpha,\beta)B_z^{(\pm)}(\alpha,\beta)]$ is a helicity-preserving transformation. Therefore, the spin disalignment is caused by the D matrix. In terms of the Wigner rotation, the D matrix is

$$D_x^{(\pm)}(\alpha,\beta) = A_z^{(\pm)}(\alpha)W(-\Phi)(A_z(\alpha))^{-1}, \tag{32}$$

where

$$W(\Phi) = \begin{pmatrix} \cos(\Phi/2) & -\sin(\Phi/2) \\ \sin(\Phi/2) & \cos(\Phi/2) \end{pmatrix}. \tag{33}$$

In the present kinematics, the rotation angle given in Eq. (19) is positive. Therefore, the matrix $W(-\Phi)$ performs a clockwise rotation. Figure 4 clearly indicates the effect of this rotation.

In the limit $\alpha \to 1$,

$$D_x^{(+)}(1,\beta) = \begin{pmatrix} 1 & -\beta \\ 0 & 1 \end{pmatrix}, \quad D_x^{(-)}(1,\beta) = \begin{pmatrix} 1 & 0 \\ \beta & 1 \end{pmatrix}. \tag{34}$$

Here again the parameter β can be replaced by u that can become arbitrarily large.

In Sec. III, we noted that the D transformation applied to a free electromagnetic four-potential performs a gauge transformation. Thus the SL(2,c) spinors are gauge invariant in the sense that

$$D_x^{(+)}(1,u)\chi_+ = \chi_+, \quad D_x^{(-)}(1,u)\dot{\chi}_- = \dot{\chi}_-. \tag{35}$$

On the other hand, the SL(2,c) spinors are gauge dependent in the sense that

$$\begin{aligned} D_x^{(+)}(1,u)\chi_- &= \chi_- + u\chi_+, \\ D_x^{(-)}(1,u)\dot{\chi}_+ &= \dot{\chi}_+ - u\dot{\chi}_-. \end{aligned} \tag{36}$$

The gauge-invariant spinors of Eq. (35) appear as polarized neutrinos in the real world. As was discussed in the literature,[5,6] it is possible to construct the four-potential from the above SL(2,c) spinors. These gauge-dependent spinors give rise to the gauge dependence of the four-potential.

V. CONCLUDING REMARKS

The starting point of the present work is Wigner's 1957 paper,[1] in which he suggests the possibility of exploiting the difference between helicity-preserving and nonpreserving transformations. Indeed, Wigner's suggestion leads to a physical embodiment of the coset expansion of the Lorentz group with respect to its little groups.

In this paper, we specialized in the Lorentz transformation perpendicular to the momentum. In this case, the set of helicity-preserving transformations includes a boost along the direction of momentum and a rotation around the axis perpendicular to the momentum and to the direction of the boost. This is the reason why we had to introduce Lorentz

kinematics different from that of Ref. 6. The kinematics of Ref. 6, while designed for its own purpose, cannot accommodate boosts along the direction of momentum.

The kinematics of the present paper is a special case of the most general kinematics described in Fig. 6. This general case includes the Lorentz boost along with an arbitrary direction, in addition to the boost along the perpendicular direction. We can use the procedure established in the present paper in order to study the general case with more complicated formulas. However, it will not alter the conclusion of the present paper.

ACKNOWLEDGMENT

We are grateful to Professor Eugene P. Wigner for very helpful discussions on the contents of his 1957 paper.

[1]E. P. Wigner, Rev. Mod. Phys. **29**, 255 (1957).
[2]Chou Kuang-Chao and L. G. Zastavenco, Zh. Eksp. Teor. Fiz. **35**, 1417 (1958) [Sov. Phys. JETP **8**, 990 (1959)]; M. Jacob and G. C. Wick, Ann. Phys. **7**, 404 (1959).
[3]E. P. Wigner, Ann. Math. (NY) **40**, 149 (1939).
[4]E. P. Wigner, Z. Phys. **124**, 665 (1948); A. S. Wightman, in *Dispersion Relations and Elementary Particles*, edited by C. De Witt and R. Omnes (Hermann, Paris, 1960); E. P. Wigner, in *Theoretical Physics*, edited by A. Salam (International Atomic Energy Agency, Vienna, 1962); M. Hamermesh, *Group Theory* (Addison–Wesley, Reading, MA, 1962); H. van Dam, Y. J. Ng, and L. C. Biedenharn, Phys. Lett. B **158**, 227 (1985).
[5]Y. S. Kim and M. E. Noz, *Theory and Applications of the Poincaré Group* (Reidel, Dordrecht, The Netherlands, 1986).
[6]D. Han, Y. S. Kim, and D. Son, J. Math. Phys. **27**, 2228 (1986).
[7]A. Chakrabarti, J. Math. Phys. **5**, 1747 (1964).
[8]E. Inonu and E. P. Wigner, Proc. Natl. Acad. Sci. USA **39**, 510 (1953); D. W. Robinson, Helv. Phys. Acta **35**, 98 (1962); D. Korff, J. Math. Phys. **5**, 869 (1964); S. Weinberg, in *Lectures on Particles and Field Theory, Brandeis 1964*, edited by S. Deser and K. W. Ford (Prentice–Hall, Englewood Cliffs, NJ, 1965), Vol. 2; J. D. Talman, *Special Functions, A Group Theoretical Approach Based on Lectures by E. P. Wigner* (Benjamin, New York, 1968); S. P. Misra and J. Maharana, Phys. Rev. D **14**, 133 (1976).
[9]D. Han, Y. S. Kim, and D. Son, Phys. Lett. B **131**, 327 (1983); D. Han, Y. S. Kim, M. E. Noz, and D. Son, Am. J. Phys. **52**, 1037 (1984); Y. S. Kim and E. P. Wigner, J. Math. Phys. **28**, 1175 (1987).
[10]S. Weinberg, Phys. Rev. **135**, B1049 (1964); A. Janner and T. Jenssen, Physica (Utrecht) **53**, 1 (1971); **60**, 292 (1972); J. L. Richard, Nuovo Cimento A **8**, 485 (1972); J. Kuperzstych, Nuovo Cimento B **31**, 1 (1976); Phys. Rev. D **17**, 629 (1978); J. J. van der Bij, H. van Dam, and Y. J. Ng, Physica (Utrecht) A **116**, 307 (1982); D. Han, Y. S. Kim, and D. Son, Phys. Rev. D **26**, 3717 (1982).
[11]D. Han and Y. S. Kim, Am. J. Phys. **49**, 348 (1981).

Generalized Lorentz transformation for an accelerated, rotating frame of reference

Robert A. Nelson
Department of Physics and Astronomy, University of Maryland, College Park, Maryland 20742

(Received 31 October 1986; accepted for publication 10 June 1987)

An exact, explicit coordinate transformation between an inertial frame of reference and a frame of reference having an arbitrary time-dependent, nongravitational acceleration and an arbitrary time-dependent angular velocity is given. This transformation is a generalization of the Lorentz transformation and is obtained in two steps. First, the Minkowski metric is transformed under an intermediate coordinate transformation to obtain a new set of noninertial metric coefficients in which one can easily identify the Thomas precession, as well as the expression for the acceleration of the moving frame with respect to the instantaneous rest frame in terms of the acceleration as seen from a stationary inertial frame. Second, a rotation of axes is performed to absorb the Thomas precession and to add an ordinary spatial rotation. The coordinate transformation obtained by combining these effects is nonlinear, since certain terms involve time integrals, and leads to the appropriate space-time metric for an accelerated, rotating frame of reference. It is shown that the usual forms of the Lorentz transformation are contained as special cases of this result.

I. INTRODUCTION

In special relativity it is customary to represent an accelerated frame of reference by an infinite sequence of comoving inertial frames. Each successive comoving inertial frame, or rest frame, is connected to the previous one by an infinitesimal Lorentz boost.[1,2] As is well known, when the velocity and acceleration are not collinear the axes of the accelerated frame do not remain parallel to axes in the stationary frame but rather rotate at the Thomas precession frequency.[3] In the general case the accelerated frame may also have an ordinary spatial rotation. The space-time metric in the accelerated, rotating frame (x^i,t) with Cartesian spatial coordinates x^i and time t $(x^0 \equiv ct)$ is[4]

$$g_{ij} = \delta_{ij}, \tag{1a}$$

$$g_{0j} = \omega_{kj}x^k/c = (\boldsymbol{\omega}\times\mathbf{r})_j/c, \tag{1b}$$

$$-g_{00} = (1 + \mathbf{W}\cdot\mathbf{r}/c^2)^2 - (\boldsymbol{\omega}\times\mathbf{r})^2/c^2, \tag{1c}$$

where $\mathbf{W}$ is the time-dependent, nongravitational acceleration of the observer's frame of reference relative to the instantaneous rest frame, $\mathbf{r}$ is the position vector locating a spatial point with respect to the origin of the observer's accelerated frame, and $\boldsymbol{\omega}$ is the time-dependent angular velocity of the observer's ordinary spatial rotation with respect to the rest frame. Historically, accelerated reference frames and the Thomas precession have been studied by approximate methods.[1,2] The coordinates of an accelerated, rotating observer have also been studied by the method of Fermi–Walker transport.[5] In this paper an exact, explicit, nonlinear coordinate transformation that incorporates the Thomas precession and leads to the metric above will be given.

II. ACCELERATED FRAME OF REFERENCE

For simplicity, consider first a frame of reference whose origin moves along an arbitrary path with velocity and acceleration that vary arbitrarily with time but whose axes have no ordinary spatial rotation. The metric in this frame is then[6]

$$g_{ij} = \delta_{ij}, \tag{2a}$$

$$g_{0j} = 0, \tag{2b}$$

$$-g_{00} = (1 + \mathbf{W}\cdot\mathbf{r}/c^2)^2, \tag{2c}$$

and the coordinate transformation that leads to this metric is to be sought.

Suppose the instantaneous rest system is defined with respect to the stationary frame by a pure Lorentz boost rather than a boost plus a rotation. Then the transformation from the stationary inertial frame S'' with coordinates (x''^i,t'') to a comoving inertial frame S' with coordinates (x'^i,t'), whose origin has constant velocity $V^i \equiv dx''^i/dt''$ in S'' and is instantaneously at rest with respect to the observer's accelerated frame S with coordinates (x^i,t) at the observer's proper time $t = \tau$, is given by (notice that unprimed coordinates are reserved for the accelerated frame)

$$x'^\alpha = \Lambda^\alpha{}_\beta x''^\beta, \tag{3}$$

where

$$\Lambda^i_j = \delta^i_j + (1/V^2)(\gamma - 1)V^iV_j, \tag{4a}$$

$$\Lambda^i{}_0 = -(1/c)\gamma V^i, \tag{4b}$$

$$\Lambda^0{}_j = -(1/c)\gamma V_j, \tag{4c}$$

$$\Lambda^0{}_0 = \gamma, \tag{4d}$$

and

$$\gamma \equiv (1 - V^2/c^2)^{-1/2}. \tag{5}$$

If the velocity V^i is regarded as a function of proper time τ, the transformation (3) defines a family of comoving inertial frames, each of whose axes are parallel to the axes of the stationary inertial frame. However, if the velocity and acceleration are not collinear the axes of successive comoving

 0022-2488/87/102379-05$02.50 © 1987 American Institute of Physics

frames will not appear parallel to one another as seen from the observer's accelerated frame due to the Thomas precession.

In the stationary inertial frame S'' the observer's four-velocity is $U''^{\alpha} \equiv dx''^{\alpha}/d\tau = (\gamma V^i, c\gamma)$. Therefore, the components of the four-acceleration are[7]

$$\frac{dU''^i}{d\tau} = \gamma^2\left[W''^i + \frac{1}{c^2}\gamma^2(V_m W''^m)V^i\right], \tag{6a}$$

$$\frac{dU''^0}{d\tau} = \frac{1}{c}\gamma^4(V_m W''^m), \tag{6b}$$

where $W''^i \equiv dV^i/dt'' = \gamma^{-1}\, dV^i/d\tau$ is the acceleration of S in S''. Since $U''_{\alpha}U''^{\alpha} = -c^2$ the four-velocity and the four-acceleration are orthogonal, i.e., $U''_{\alpha}\, dU''^{\alpha}/d\tau = 0$.

In the instantaneous rest frame S' the four-velocity is $U'^{\alpha} = (0,c)$. By the orthogonality of the four-velocity and four-acceleration, the four-acceleration is of the form $dU'^{\alpha}/d\tau = (W^i,0)$. In this frame the observer's three-dimensional acceleration W^i is thus the spatial part of the four-acceleration. The observer's four-acceleration in the instantaneous rest frame is obtained by applying a Lorentz boost

$$\frac{dU'^i}{d\tau} = \Lambda^i_{\ \alpha}\frac{dU''^{\alpha}}{d\tau} = W^i, \tag{7a}$$

$$\frac{dU'^0}{d\tau} = \Lambda^0_{\ \alpha}\frac{dU''^{\alpha}}{d\tau} = 0. \tag{7b}$$

Substituting Eqs. (4) and (6) into Eq. (7a) one obtains

$$W^i = \gamma^2[W''^i + (1/V^2)(\gamma - 1)(V_m W''^m)V^i], \tag{8}$$

which gives the acceleration W^i of S in the instantaneous rest frame S' in terms of the acceleration W''^i of S in the stationary inertial frame S''.

As an intermediate step toward the desired coordinate transformation, consider the transformation from the stationary inertial frame (x''^i,t'') to an accelerated frame (X^i,T),

$$x''^i = X^i + \int_0^T \gamma V^i\, dT + \frac{1}{V^2}(\gamma - 1)(V_m X^m)V^i, \tag{9a}$$

$$t'' = \int_0^T \gamma dT + \frac{1}{c^2}\gamma(V_m X^m), \tag{9b}$$

where V^i and γ are functions of time $T = t = \tau$. The metric transforms as

$$g_{\mu\nu} = \frac{\partial x''^{\alpha}}{\partial X^{\mu}}\frac{\partial x''^{\beta}}{\partial X^{\nu}}\eta_{\alpha\beta}, \tag{10}$$

where $\eta_{\alpha\beta}$ is the Minkowski metric. We wish to investigate the components of the metric given by Eq. (10) in the frame defined by the coordinate transformation given by Eq. (9). For $\mu = i$, $\nu = j$,

$$g_{ij} = \frac{\partial x''^k}{\partial X^i}\frac{\partial x''^l}{\partial X^j}\delta_{kl} + \frac{\partial x''^0}{\partial X^i}\frac{\partial x''^0}{\partial X^j}(-1)$$

$$= \left[\delta^k_i + \frac{1}{V^2}(\gamma - 1)V_i V^k\right]\left[\delta^l_j + \frac{1}{V^2}(\gamma - 1)V_j V^l\right]\delta_{kl} - \left(\frac{1}{c}\gamma V_i\right)\left(\frac{1}{c}\gamma V_j\right) = \delta_{ij}. \tag{11a}$$

For $\mu = 0$, $\nu = j$,

$$g_{0j} = \frac{\partial x''^k}{\partial X^0}\frac{\partial x''^l}{\partial X^j}\delta_{kl} + \frac{\partial x''^0}{\partial X^0}\frac{\partial x''^0}{\partial X^j}(-1)$$

$$= \frac{1}{c}\left[\gamma V^k + \frac{1}{V^2}(\gamma - 1)(V_m X^m)\frac{dV^k}{d\tau} + \frac{1}{V^2}(\gamma - 1)\left(\frac{dV_m}{d\tau}X^m\right)V^k\right.$$

$$\left. + \frac{1}{c^2}\frac{1}{V^2}\gamma^3 V\frac{dV}{d\tau}(V_m X^m)V^k - 2\frac{1}{V^4}V\frac{dV}{d\tau}(\gamma - 1)(V_m X^m)V^k\right]\left[\delta^l_j + \frac{1}{V^2}(\gamma - 1)V_j V^l\right]\delta_{kl}$$

$$- \left[\gamma + \frac{1}{c^2}\gamma\frac{dV_m}{d\tau}X^m + \frac{1}{c^4}\gamma^3 V\frac{dV}{d\tau}(V_m X^m)\right]\left(\frac{1}{c}\gamma V_j\right)$$

$$= -\frac{1}{c}\gamma\left[-\frac{1}{V^2}(\gamma - 1)(V_k W''_j - V_j W''_k)X^k\right]. \tag{11b}$$

For $\mu = \nu = 0$,

$$-g_{00} = -\frac{\partial x''^k}{\partial X^0}\frac{\partial x''^l}{\partial X^0}\delta_{kl} - \frac{\partial x''^0}{\partial X^0}\frac{\partial x''^0}{\partial X^0}(-1)$$

$$= -\frac{1}{c^2}\left[\gamma V^k + \frac{1}{V^2}(\gamma - 1)(V_m X^m)\frac{dV^k}{d\tau} + \frac{1}{V^2}(\gamma - 1)\left(\frac{dV_m}{d\tau}X^m\right)V^k\right.$$

$$\left. + \frac{1}{c^2}\frac{1}{V^2}\gamma^3 V\frac{dV}{d\tau}(V_m X^m)V^k - 2\frac{1}{V^4}V\frac{dV}{d\tau}(\gamma - 1)(V_m X^m)V^k\right]$$

$$\times\left[\gamma V^l + \frac{1}{V^2}(\gamma - 1)(V_n X^n)\frac{dV^l}{d\tau} + \frac{1}{V^2}(\gamma - 1)\left(\frac{dV_n}{d\tau}X^n\right)V^l\right.$$

$$+\frac{1}{c^2}\frac{1}{V^2}\gamma^3 V\frac{dV}{d\tau}(V_n X^n)V^i - 2\frac{1}{V^4}V\frac{dV}{d\tau}(\gamma-1)(V_n X^n)V^l\Big]\delta_{kl}$$

$$+\left[\gamma+\frac{1}{c^2}\gamma\frac{dV_m}{d\tau}X^m+\frac{1}{c^4}\gamma^3 V\frac{dV}{d\tau}(V_m X^m)\right]^2$$

$$=\left\{1+\frac{1}{c^2}\gamma^2\left[W''_k+\frac{1}{V^2}(\gamma-1)(V_m W''^m)V_k\right]X^k\right\}^2-\frac{1}{c^2}\gamma^2\left[-\frac{1}{V^2}(\gamma-1)(V_k W''^m - V^m W''_k)X^k\right]$$

$$\times\left[-\frac{1}{V^2}(\gamma-1)(V_n W''_m - V_m W''_n)X^n\right], \tag{11c}$$

where $W''_k=\gamma^{-1}dV_k/d\tau$.

One finds that W''^i from $dV^i/d\tau$ appears inside the first term of Eq. (11c) in the combination of Eq. (8). It also appears in Eq. (11b) and in the second term of Eq. (11c) in the combination

$$\Omega''_{kj}=-(1/V^2)(\gamma-1)(V_k W''_j - V_j W''_k), \tag{12}$$

which is the well-known Thomas precession frequency. The metric can therefore be written

$$g_{ij}=\delta_{ij}, \tag{13a}$$

$$g_{0j}=-\Omega_{kj}X^k/c=-(\boldsymbol{\Omega}\times\mathbf{r})_j/c, \tag{13b}$$

$$-g_{00}=(1+\mathbf{W}\cdot\mathbf{r}/c^2)^2-(\boldsymbol{\Omega}\times\mathbf{r})^2/c^2, \tag{13c}$$

where $\Omega_{kj}=\gamma\Omega''_{kj}$. This metric reduces to the metric given by Eq. (2) for the observer's accelerated frame (x^i,t) under the coordinate transformation

$$X^i=x^i+\int_0^t(\boldsymbol{\Omega}\times\mathbf{r})^i\,dt, \tag{14a}$$

$$T=t. \tag{14b}$$

The appearance of $\boldsymbol{\Omega}\times\mathbf{r}$ in g_{0j} and g_{00} is furthermore characteristic of a rotation. From Eq. (12) it follows that there is no precession phenomenon when the velocity and acceleration are collinear, as is well known.

The origin of the frame (X^i,T) defined by Eq. (9) coincides with the origin of the observer's accelerated reference frame (x^i,t). The origin moves along an arbitrary path in the stationary inertial frame (x''^i,t'') with velocity V^i that varies arbitrarily with proper time. The axes of (X^i,T) appear to remain parallel to the axes of (x''^i,t''). However, with respect to the axes of (x^i,t) they appear to rotate with proper angular velocity $-\Omega^i=-\frac{1}{2}\epsilon^{imn}\Omega_{mn}$ opposite in sense to the Thomas precession. At any given instant of proper time, the origin and axes of (X^i,T) coincide with those of the instantaneous rest frame (x'^i,t') that moves with constant velocity V^i relative to (x''^i,t'').

The desired coordinate transformation from the stationary inertial frame to the observer's accelerated frame must include the rotation of axes (14) to account for the Thomas precession. Therefore, combining the transformations (9) and (14), and taking care to include coefficients of X^i in Eq. (9) inside the integral in Eq. (14a), one obtains

$$x''^i=x^i+\int_0^t(\boldsymbol{\Omega}\times\mathbf{r})^i\,dt+\int_0^t\gamma V^i\,dt$$
$$+\frac{1}{V^2}(\gamma-1)(V_m x^m)V^i$$
$$+\int_0^t\frac{1}{V^2}(\gamma-1)V_m(\boldsymbol{\Omega}\times\mathbf{r})^m V^i\,dt, \tag{15a}$$

$$t''=\int_0^t\gamma\,dt+\frac{1}{c^2}\gamma(V_m x^m)+\int_0^t\frac{1}{c^2}\gamma V_m(\boldsymbol{\Omega}\times\mathbf{r})^m\,dt, \tag{15b}$$

where V^i and γ are functions of time $t=\tau$. It is understood that the integrals are functions of time only. The metric transforms as

$$g_{\mu\nu}=\frac{\partial x''^\alpha}{\partial x^\mu}\frac{\partial x''^\beta}{\partial x^\nu}\eta_{\alpha\beta}. \tag{16}$$

In carrying out the details of the calculation of the metric components directly with Eq. (15) one finds that $dV^i/dt=dV^i/d\tau=\gamma W''^i$ again appears in the combinations of Eqs. (8) and (12). The kinematics of the acceleration and rotation are thus automatically incorporated into the transformation. Equation (15) is the exact, explicit, nonlinear coordinate transformation that transforms the Minkowski metric for an inertial frame of reference into the metric given by Eq. (2) for an accelerated frame of reference.

When the velocity V^i is constant, Eq. (15) reduces to the usual Lorentz transformation

$$x''^i=x^i+\gamma V^i t+(1/V^2)(\gamma-1)(V_m x^m)V^i, \tag{17a}$$

$$t''=\gamma t+(1/c^2)\gamma(V_m x^m), \tag{17b}$$

where the axes of S'' and S remain parallel but the direction of V^i is arbitrary. Also, as is well known, for motion with time-dependent acceleration along the x'' axis one may write $V=c\tanh\theta$, $\gamma=\cosh\theta$, $W=\gamma^3 W''=\gamma^2\,dV/dt=c\,d\theta/dt$, and $\Omega=0$, where θ is the velocity parameter and is an arbitrary function of time $t=\tau$. The transformation (15) then becomes[8]

$$x''=\int_0^t c\sinh\theta\,dt+x\cosh\theta;\quad y''=y;\quad z''=z, \tag{18a}$$

$$t''=\int_0^t\cosh\theta\,dt+\frac{x}{c}\sinh\theta. \tag{18b}$$

In particular, if the acceleration W is constant the motion is hyperbolic. In this case, with the substitution $\bar{x}=x+(c^2/W)$, the metric of Eq. (2) reduces to the Rindler

metric[9] $ds^2 = d\bar{x}^2 + dy^2 + dz^2 - c^{-2}(W\bar{x})^2\,dt^2$ for the space-time geometry of a uniformly accelerated reference frame.

III. ACCELERATED, ROTATING FRAME OF REFERENCE

Suppose now that the accelerated frame of reference has an ordinary spatial rotation with time-dependent angular velocity ω with respect to the instantaneous rest frame in addition to its Thomas precession. Then since the origins of the accelerated, rotating frame and the instantaneous rest frame coincide, it is only necessary to make the substitution $\Omega \to \Omega + \omega$ in Eq. (14a). One therefore obtains the following general coordinate transformation that leads to the metric of Eq. (1):

$$x''^i = x^i + \int_0^t [(\Omega+\omega)\times\mathbf{r}]^i\,dt + \int_0^t \gamma V^i\,dt + \frac{1}{V^2}(\gamma-1)(V_m x^m)V^i + \int_0^t \frac{1}{V^2}(\gamma-1)V_m[(\Omega+\omega)\times\mathbf{r}]^m V^i\,dt, \tag{19a}$$

$$t'' = \int_0^t \gamma\,dt + \frac{1}{c^2}\gamma(V_m x^m) + \int_0^t \frac{1}{c^2}\gamma V_m[(\Omega+\omega)\times\mathbf{r}]^m\,dt, \tag{19b}$$

where again V^i and γ are functions of time $t = \tau$. Upon transforming the Minkowski metric with this coordinate transformation one finds

$$g_{ij} = \delta_{ij}, \tag{20a}$$

$$g_{0j} = -(1/c)\gamma[-(1/V^2)(\gamma-1)(V_k W''_j - V_j W''_k)x^k] + (1/c)[(\Omega+\omega)\times\mathbf{r}]_j = (1/c)(\omega\times\mathbf{r})_j, \tag{20b}$$

$$\begin{aligned} -g_{00} &= \{1 + (1/c^2)\gamma^2[W''_k + (1/V^2)(\gamma-1)(V_m W''^m)V_k]x^k\}^2 \\ &\quad - (1/c^2)\gamma^2[-(1/V^2)(\gamma-1)(V_k W''^m - V^m W''_k)x^k][-(1/V^2)(\gamma-1)(V_n W''_m - V_m W''_n)x^n] \\ &\quad - (1/c^2)[(\Omega+\omega)\times\mathbf{r}]^2 + 2(1/c^2)\gamma[-(1/V^2)(\gamma-1)(V_k W''^m - V^m W''_k)x^k][(\Omega+\omega)\times\mathbf{r}]_m \\ &= (1 + \mathbf{W}\cdot\mathbf{r}/c^2)^2 - (\omega\times\mathbf{r})^2/c^2 \end{aligned} \tag{20c}$$

as required.

The coordinate transformation given by Eq. (19) represents a boost followed by a rotation, as may be made clear by defining the rotation operators

$$\mathscr{R}^i{}_j = \delta^i_j + \int_0^t dt(\Omega+\omega)_j{}^i, \tag{21a}$$

$$\mathscr{R}''^i{}_j = \delta^i_j - \int_0^t dt(\Omega+\omega)_j{}^i. \tag{21b}$$

For infinitesimal rotations $\mathscr{R}'' = \mathscr{R}^{-1}$. If V^i is the velocity of the accelerated, rotating frame S relative to the stationary, inertial frame S'' and V''^i is the velocity of S'' relative to S, then

$$V''^i = -\mathscr{R}''^i{}_j V^j, \tag{22a}$$

$$V^i = -\mathscr{R}^i{}_j V''^j. \tag{22b}$$

Therefore, Eq. (19) may be expressed as

$$x''^i = \mathscr{R}^i{}_m x^m + \int_0^t \gamma V^i\,dt - V''^m{}_m\left[\frac{1}{V^2}(\gamma-1)x^m V^i\right], \tag{23a}$$

$$t'' = \int_0^t \gamma\,dt - V''^m{}_m\left(\frac{1}{c^2}\gamma x^m\right), \tag{23b}$$

where the identity $[(\Omega+\omega)\times\mathbf{V}]\cdot\mathbf{r} = -[(\Omega+\omega)\times\mathbf{r}]\cdot\mathbf{V}$ has been used. Equation (23) has the form of the most general Lorentz transformation including rotation.[10]

IV. EQUATION OF MOTION

For completeness, one may derive the exact equation of motion of a particle as seen from the accelerated, rotating frame of reference. The differential equation for a space-time geodesic implies

$$\begin{aligned} \frac{d^2x^i}{dt^2} &= \left(\frac{dt}{d\tau}\right)^{-2}\left(\frac{d^2x^i}{d\tau^2} - \frac{d^2t}{d\tau^2}\frac{dx^i}{dt}\right) \\ &= -\left(\Gamma^i_{jk}\frac{dx^j}{dt}\frac{dx^k}{dt} + 2c\Gamma^i_{j0}\frac{dx^j}{dt} + c^2\Gamma^i_{00}\right) \\ &\quad + \left(\frac{1}{c}\Gamma^0_{jk}\frac{dx^j}{dt}\frac{dx^k}{dt} + 2\Gamma^0_{j0}\frac{dx^j}{dt} + c\Gamma^0_{00}\right)\frac{dx^i}{dt}. \end{aligned} \tag{24}$$

The Christoffel symbols for the metric given by Eq. (1) are

$$\Gamma^i_{jk} = 0, \tag{25a}$$

$$\Gamma^i_{j0} = (1/c)[\omega_j{}^i - (1/c^2)(1 + \mathbf{W}\cdot\mathbf{r}/c^2)^{-1}W_j(\omega\times\mathbf{r})^i], \tag{25b}$$

$$\begin{aligned} \Gamma^i_{00} &= (1/c^2)\{1 + \mathbf{W}\cdot\mathbf{r}/c^2)W^i + (\dot{\omega}\times\mathbf{r})^i \\ &\quad + [\omega\times(\omega\times\mathbf{r})]^i - (1/c^2)(1 + \mathbf{W}\cdot\mathbf{r}/c^2)^{-1} \\ &\quad \times[(\omega\times\mathbf{r})\cdot\mathbf{W} + (\dot{\mathbf{W}}\cdot\mathbf{r})](\omega\times\mathbf{r})^i\}, \end{aligned} \tag{25c}$$

$$\Gamma^0_{jk} = 0, \tag{25d}$$

$$\Gamma^0_{j0} = (1/c^2)(1 + \mathbf{W}\cdot\mathbf{r}/c^2)^{-1}W_j, \tag{25e}$$

$$\Gamma^0_{00} = (1/c^3)(1 + \mathbf{W}\cdot\mathbf{r}/c^2)^{-1}[(\omega\times\mathbf{r})\cdot\mathbf{W} + (\dot{\mathbf{W}}\cdot\mathbf{r})]. \tag{25f}$$

Therefore, the equation of motion is

$$a^i = -(1+\mathbf{W}\cdot\mathbf{r}/c^2)W^i - (\dot{\omega}\times\mathbf{r})^i - 2(\omega\times\mathbf{v})^i - [\omega\times(\omega\times\mathbf{r})]^i + (1/c^2)(1+\mathbf{W}\cdot\mathbf{r}/c^2)^{-1}[2(\mathbf{v}+\omega\times\mathbf{r})\cdot\mathbf{W} + (\dot{\mathbf{W}}+\omega\times\mathbf{W})\cdot\mathbf{r}][v^i+(\omega\times\mathbf{r})^i], \quad (26)$$

where in Cartesian coordinates $v^i = dx^i/dt$ is the velocity measured by the accelerated, rotating observer and $a^i = d^2x^i/dt^2$ is the coordinate acceleration.

Equation (26) agrees with an exact result of DeFacio, Dennis, and Retzloff[11] obtained by using a general coordinate-independent treatment of special relativity. An approximate form of this equation was obtained using local coordinate methods by Ni and Zimmermann[12] who were investigating general relativistic corrections to special relativity; they included gravitational tidal effects expressed in terms of the Riemann tensor evaluated along the world line of the observer but found no coupling between the gravitational terms and inertial terms to the order calculated. Reference 11 and the special relativity part of Ref. 12 were later shown to be consistent.[13] Li and Ni[14] extended the method of Ref. 12 and found that the gravitational-inertial coupling only occurs at the next order; they also found an exact expression for the inertial terms in agreement with the earlier result of Ref. 11.

V. CONCLUSION

An exact, explicit coordinate transformation has been presented that yields the exact space-time metric for a flat space noninertial reference frame having an arbitrary, time-dependent translational acceleration and angular velocity. It was shown that the usual forms of the Lorentz transformation are contained as special cases of this result. The expression for the Thomas precession frequency and the relation between the reference frame's acceleration as measured in the instantaneous rest frame and in the stationary inertial frame are contained implicitly in the calculation of the metric. Therefore all of the relevant special relativity kinematics appear in a self-consistent manner.

The results of this paper should be useful for practical numerical calculations of special relativistic effects, such as in an accelerated spacecraft or on the rotating Earth. The simple metric coefficients of Eq. (1) and the explicit coordinate transformation of Eq. (19) may be more convenient than some of the more formal works referenced here.

[1] H. Goldstein, *Classical Mechanics* (Addison–Wesley, Reading, MA, 1980), 2nd ed., p. 286.
[2] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), 2nd ed., p. 545.
[3] L. H. Thomas, Nature **117**, 514 (1926); Philos. Mag. **3**, 1 (1927).
[4] Latin indices assume the range 1–3 while Greek indices assume the range 0–3. Repeated indices are summed over their respective ranges. The signature of the space-time metric is $(-+++)$.
[5] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), Chap. 6.
[6] C. Møller, *The Theory of Relativity* (Oxford, New York, 1952), p. 254.
[7] See Ref. 6, p. 103.
[8] See Ref. 6, p. 255.
[9] W. Rindler, Am. J. Phys. **34**, 1174 (1966).
[10] See Ref. 6, p. 43.
[11] B. DeFacio, P. W. Dennis, and D. G. Retzloff, Phys. Rev. D **18**, 2813 (1978).
[12] W.-T. Ni and M. Zimmermann, Phys. Rev. D **17**, 1473 (1978).
[13] B. DeFacio, P. W. Dennis, and D. G. Retzloff, Phys. Rev. D **20**, 570 (1979).
[14] W.-Q. Li and W.-T. Ni, J. Math. Phys. **20**, 1473 (1979).

Jauch–Piron states in W^*-algebraic quantum mechanics

Anton Amann[a)]
Naturwissenschaftlich Theoretisches Zentrum, Karl-Marx Universität Leipzig, DDR-7010 Leipzig, German Democratic Republic

(Received 13 February 1987; accepted for publication 10 June 1987)

A state ϕ on a W^*-algebra $\mathcal{M}$ is said to fulfill the *Jauch–Piron condition* if $\phi(p) = \phi(q) = 1$ for projections $p,q \in \mathcal{M}$ implies $\phi(p \wedge q) = 1$. Here $p \wedge q$ denotes the infimum of p and q in the projection lattice of $\mathcal{M}$. The Jauch–Piron condition is a compatibility condition between the algebraic and the lattice-theoretic approach for the description of physical systems. Normal (i.e., σ-weakly continuous) states always fulfill the Jauch–Piron condition. It is argued that states not fulfilling this condition should be regarded as unphysical. It is shown that a state ϕ on a σ-finite *factor* $\mathcal{M}$ is singular if and only if projections $e, f \in \mathcal{M}$ exist such that $\phi(e) = \phi(f) = 1$ and $e \wedge f = 0$. In particular, any *pure* state ϕ on $\mathcal{M}$ fulfilling the Jauch–Piron condition is normal, which implies that the underlying factor $\mathcal{M}$ is of type I. Furthermore, the following result is proved: Let ϕ be a *pure* Jauch–Piron state on W^*-algebra $\mathcal{M}$ with separable predual and without any commutative summand. Then ϕ is normal and a central projection $z_0 \in \mathcal{M}$ exists such that $\phi(z_0) = 1$ and $z_0 \mathcal{M} z_0$ is a factor of type I. Thus, *cum grano salis*, *pure* Jauch–Piron states exist only on commutative W^*-algebras and type I factors. The former case corresponds to classical theories, the latter to Hilbert-space quantum mechanics. The implications of these results on the interpretation of quantum mechanics are discussed.

I. INTRODUCTION

Different axiomatic formulations have been found to widen the original quantum mechanical formalism and to embed quantum mechanics *and* classical theories into a broader structural setting. One may distinguish three main approaches: quantum logics, algebraic quantum mechanics, and the "convex state approach." In quantum logics and algebraic quantum mechanics one starts with "observations" or "observables" forming a lattice or an algebra (C^*- or W^*-algebra), respectively. In the convex state approach the physical states of a system to be described are the central object of interest. In the former situation an appropriate notion of "state" has to be introduced, in the latter an appropriate notion of observable.

In algebraic quantum mechanics physical states are commonly introduced as *normalized positive linear functionals* on the respective algebra. The critical point of such a definition is the *linearity* of the functional. Various attempts for its justification have been developed. The most famous one is the theorem of Gleason.[1]

In the following, attention will be focused on a particular algebraic approach, namely *W^*-algebraic quantum mechanics*.[2,3] It is interesting to note that this approach is intimately connected with quantum logics. The set $\mathcal{P}(\mathcal{M}) \overset{\text{def}}{=} \{p \in \mathcal{M} \mid p^2 = p = p^*\}$ of projections of a W^*-algebra $\mathcal{M}$ is a (complete, orthomodular) lattice. Its elements can be interpreted as the "propositions" or "yes–no questions" of quantum logics. It is therefore possible to adopt a characterization of physical states given by Jauch[4] who says: "*A state of a system is the set of all 'true' propositions of the system.*"

Recall that a lattice $(L, \leqslant)$ is a partially ordered set L with a least element 0 and a greatest element 1 such that the supremum $p \vee q$ and the infimum $p \wedge q$ of two arbitrary elements $p,q \in L$ with respect to the partial order $\leqslant$ exist. An *orthocomplemented lattice* or simply *ortholattice* $(L, \leqslant, \perp)$ is a lattice $(L, \leqslant)$ together with an involutive mapping $\perp: L \to L$ such that $p \wedge p^\perp = 0$ and $p \vee p^\perp = 1$, $\forall p \in L$, and such that $p \leqslant q$ implies $q^\perp \leqslant p^\perp$ for arbitrary $p,q \in L$.

In the projection lattice $\mathcal{P}(\mathcal{M})$ of a W^*-algebra $\mathcal{M}$, 0 is given by the zero operator and 1 by the unit operator in $\mathcal{M}$. The orthocomplement $p^\perp$ of $p \in \mathcal{P}(\mathcal{M})$ is defined as $p^\perp \overset{\text{def}}{=} 1 - p$. Furthermore $p \leqslant q$ holds for $p,q \in \mathcal{P}(\mathcal{M})$ if $pq = qp = p$.

The definition of physical states (with respect to a W^*-algebra $\mathcal{M}$) as given above is formalized as follows[5]: Consider an orthosublattice T of $\mathcal{P}(\mathcal{M})$ and an ortholattice homomorphism $\tau: T \to \{0,1\}$ into the lattice $\{0,1\}$ consisting only of a least element 0 and a greatest element 1:

$$\tau(p \wedge q) = \tau(p) \wedge \tau(q),$$
$$\tau(p \vee q) = \tau(p) \vee \tau(q), \quad p,q, \in \mathcal{P}(\mathcal{M}),$$
$$\tau(p^\perp) = \tau(p)^\perp, \quad p \in \mathcal{P}(\mathcal{M}).$$

The couple (T,τ) will be called a *truth-functional*. The propositions $p \in T$ with $\tau(p) = 1$ $[\tau(p) = 0]$ are considered as true (false). A truth functional (T,τ) is *maximal* if there is no truth functional (T',τ') with $T \subsetneq T'$ and $\tau'|_T = \tau$. *Henceforth physical states of a system described by a W^*-algebra $\mathcal{M}$ will be represented by maximal truth functionals* (T,τ), $T \subseteq \mathcal{P}(\mathcal{M})$, $\tau: T \to \{0,1\}$. Their existence is guaranteed by the axiom of choice.

[a)] Permanent address: Laboratory of Physical Chemistry, ETH-Zentrum, CH-8092 Zürich, Switzerland.

 0022-2488/87/102384-06$02.50 © 1987 American Institute of Physics

II. THE JAUCH–PIRON PROPERTY

In concordance with mathematical literature the technical term *state* will be used for a "normalized positive linear functional." To prevent misunderstandings, the word "physical state" will henceforth be omitted and replaced by (maximal) truth functional. Recall the following definitions.[6]

Definition: A state ϕ on a W^*-algebra $\mathcal{M}$ is called *normal* if $\phi(\sup_{\beta\in J} x_\beta) = \sup_{\beta\in J}\phi(x_\beta)$ holds for every bounded increasing net $(x_\beta)_{\beta\in J}$ of positive operators in $\mathcal{M}$ with supremum $\sup_{\beta\in J} x_\beta$.

Definition: A state ϕ on a W^*-algebra $\mathcal{M}$ is called *singular* if for every $p\in\mathcal{P}(\mathcal{M})$ with $\phi(p)\neq 0$ there exists a $q\in\mathcal{P}(\mathcal{M})$, $q\neq 0$, $q\leqslant p$ such that $\phi(q) = 0$.

Starting from a pure state ϕ on a W^*-algebra $\mathcal{M}$, one can eventually construct a maximal truth functional by setting

$$T_\phi \overset{\text{def}}{=} \{p\in\mathcal{P}(\mathcal{M})\,|\,\phi(p)\in\{0,1\}\} \quad \text{and} \quad (T,\tau) \overset{\text{def}}{=} (T_\phi, \phi|_{T_\phi}).$$

Consider two examples, referring to classical mechanics and Hilbert space quantum mechanics, respectively.

Example 1: Let ϕ be a pure state on a commutative W^*-algebra $\mathcal{M}$. Then ϕ is a character, i.e., $\phi(xy) = \phi(x)\cdot\phi(y)$, $\forall x, y\in\mathcal{M}$. Consequently, $\phi(p) = \phi(p^2) = \{\phi(p)\}^2$, $\forall p\in\mathcal{P}(\mathcal{M})$, and therefore $\phi(p)\in\{0,1\}$, $\forall p\in\mathcal{P}(\mathcal{M})$. For $p,q\in\mathcal{P}(\mathcal{M})$ one has $p\wedge q = pq$ (since $\mathcal{M}$ is commutative) and $\phi(p\wedge q) = \phi(pq) = \phi(p)\cdot\phi(q) = \phi(p)\wedge\phi(q)$. Thus $(\mathcal{P}(\mathcal{M}), \phi|_{\mathcal{P}(\mathcal{M})})$ is a maximal truth functional.

Example 2: Consider a Hilbert space $\mathcal{H}$, $\xi\in\mathcal{H}$, $\|\xi\| = 1$. Let $\mathcal{M} \overset{\text{def}}{=} \mathcal{B}(\mathcal{H})$ consist of all bounded linear operators on $\mathcal{H}$. Set $\phi(x) \overset{\text{def}}{=} \langle\xi\,|x\xi\rangle$, $x\in\mathcal{B}(\mathcal{H})$. Then ϕ is a pure normal state on $\mathcal{B}(\mathcal{H})$ and $(T_\phi, \phi|_{T_\phi})$ is a maximal truth functional (Ref. 5: Appendix 1, Corollary 1).

Conversely, one can start from a truth functional (T,τ) and find a linear functional ϕ with $\phi(p) = \tau(p)$, $\forall p\in T$. This is done in the following theorem and justifies the use of *linear* functionals in algebraic quantum mechanics.

Theorem 1 (Ref. 5: Theorem II. 2.3): Let (T,τ) be a truth functional in a W^*-algebra $\mathcal{M}$. Then there exists a *pure* state ϕ on $\mathcal{M}$, such that $\phi(p) = \tau(p)$, $\forall p\in T$.

Note that in this theorem the truth functional need not be maximal. If it is maximal, one may expect that the state ϕ extending τ is unique. This is at least the case if (T,τ) is normal, i.e., if T is a complete orthosublattice of $\mathcal{P}(\mathcal{M})$ and $\tau(\wedge_{i\in J} p_i) = \wedge_{i\in J}\tau(p_i)$ holds for every family $\{p_i\,|i\in J\}$ in T [(T,τ) is then extended by a *normal* state ϕ].

If (T,τ) is a maximal truth functional and ϕ an extending pure state, it might happen that $T \overset{\subset}{\neq} T'$. This means that projections $e,f\in\mathcal{M}$ exist with $\phi(e) = \phi(f) = 1$ but $\phi(e\wedge f)\neq 1$. It is even possible to sharpen this statement to $\phi(e) = \phi(f) = 1, e\wedge f = 0$. This means that ϕ has expectation value 1 on e and f but nevertheless e and f are not true propositions, since they cannot be incorporated within a truth-definite orthosublattice T. Such states ϕ—and the corresponding truth functionals—should in fact be excluded in physics.

It has been argued above (see Examples 1 and 2) that states ϕ without those unpleasant properties exist. They will be called states with the Jauch–Piron property or simply Jauch–Piron states (cf. Ref. 7).

Definition: A state ϕ on a W^*-algebra $\mathcal{M}$ is said to have the *Jauch–Piron property* if $\phi(p) = \phi(q) = 1$, $p,q\in\mathcal{P}(\mathcal{M})$ implies $\phi(p\wedge q) = 1$.

Theorem 2: Let ϕ be a pure state on a W^*-algebra $\mathcal{M}$ with the Jauch–Piron property. Then ϕ is the *only* state on $\mathcal{M}$ which extends the maximal truth functional $(T_\phi, \phi|_{T_\phi})$.

Proof: The assertion follows from the results in Ref. 8, Chaps. 1 and 2. The maximality of $(T_\phi, \phi|_{T_\phi})$ is a consequence of Ref. 9.

Theorem 3 (Ref. 5: Appendix 1, Corollary 1): Every *normal* state ϕ on a W^*-algebra $\mathcal{M}$ satisfies the Jauch–Piron condition.

III. JAUCH–PIRON STATES ON FACTORS

In the present chapter, $\mathcal{M}$ is assumed to be a *σ-finite* W^*-algebra. Recall that $\mathcal{M}$ is σ-finite if it has a separable predual.

In the following lemma a class of states *not* satisfying the Jauch–Piron condition is introduced.

Lemma 1: Consider the W^*-algebra

$$\mathcal{M} \overset{\text{def}}{=} \mathcal{L}_\infty(\mathbf{N})\,\bar{\otimes}\,\mathcal{M}_2,$$

where $\mathbf{N}$ denotes the natural numbers, $\mathcal{L}_\infty(\mathbf{N})$ the W^*-algebra of bounded complex-valued functions on $\mathbf{N}$, and $\mathcal{M}_2$ the algebra of 2×2 matrices. Let $\phi = \Psi_1\otimes\Psi_2$ be a product state on $\mathcal{M}$. Assume that Ψ_2 is a (normal) pure state on $\mathcal{M}_2$ and Ψ_1 is a state on $\mathcal{L}_\infty(\mathbf{N})$ such that

$$\Psi_1|_{C_0(\mathbf{N})} \equiv 0 \left(C_0(\mathbf{N}) \overset{\text{def}}{=} \{m\in L_\infty(\mathbf{N})\,|\lim_{j\to\infty} m(j) = 0\}\right).$$

Then there are projections e and f in $\mathcal{M}$ with $\phi(e) = \phi(f) = 1$ and $e\wedge f = 0$.

Proof: Here $\mathcal{L}_\infty(\mathbf{N})\,\bar{\otimes}\,\mathcal{M}_2$ is naturally isomorphic to the direct sum algebra $\oplus_{j\in\mathbf{N}}\mathcal{R}_j$, where $\mathcal{R}_j\cong\mathcal{M}_2$, $\forall j\in\mathbf{N}$ (cf. Ref. 6: Chap. IV.7). The support r of Ψ_2 is an atom in $\mathcal{P}(\mathcal{M}_2)$. In appropriate coordinates one has $r = \left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right)\in\mathcal{M}_2$. Also $e \overset{\text{def}}{=} 1\otimes r$ corresponds, to the direct sum $\oplus_{j\in\mathbf{N}} e_j$, $e_j = r$, $\forall j\in\mathbf{N}$. For f one takes the projection $\oplus_{j\in\mathbf{N}} f_j$,

$$f_j \overset{\text{def}}{=} \frac{1}{j}\begin{pmatrix} j-1 & \sqrt{j-1} \\ \sqrt{j-1} & 1\end{pmatrix}, \quad j\in\mathbf{N}.$$

The projections f_j are atoms in $\mathcal{P}(\mathcal{M}_2)$; $e\wedge f = \oplus_{j\in\mathbf{N}}(r\wedge f_j) = 0$ since $r\neq f_j$ implies $r\wedge f_j = 0$, $\forall j\in\mathbf{N}$,

$$\phi(e) = (\Psi_1\otimes\Psi_2)(1\otimes r) = 1,$$

$$\phi(f) = \phi((1\otimes r) f (1\otimes r)) = \phi(\{\mathbf{N}\ni j\to(j-1)/j\}\otimes r)$$
$$= \Psi_1(\{\mathbf{N}\ni j\to(j-1)/j\}) = 1,$$

for

$$\lim_{j\to\infty}[(j-1)/j] = 1. \qquad \text{Q.E.D}$$

Theorem 4: Let ϕ be a *singular* state on a W^*-algebra $\mathcal{M}$. Assume a projection r to exist in $\mathcal{M}$ with $\phi(r) = 1$ and $r\sim r^\perp$ (i.e., r and $r^\perp = 1 - r$ are Murray–von Neumann

equivalents). Then projections $e, f \in \mathcal{M}$ exist with $\phi(e) = \phi(f) = 1$ and $e \wedge f = 0$.

Proof: There exists a W^*-algebra $\tilde{\mathcal{M}}$ and a *-isomorphism I: $\tilde{\mathcal{M}} \bar{\otimes} \mathcal{M}_2 \to \mathcal{M}$ such that $I\left(1 \otimes \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)\right) = r$ (use Ref. 6: Proposition V. 1.22). The Cauchy–Schwarz inequality together with $\phi(r^\perp) = 0$ implies that $\phi \circ I$ is a product state $\Psi_1 \otimes \Psi_2$ on $\tilde{\mathcal{M}} \bar{\otimes} \mathcal{M}_2$. Here Ψ_2 is a (normal) pure state on $\mathcal{M}_2$. Also Ψ_1 is a *singular* state on $\tilde{\mathcal{M}}$ since ϕ is singular on $\mathcal{M}$. Using the singularity of Ψ_1 and the axiom of choice one can construct a family $(q_k)_{k \in J}$ (J an index set) of pairwise orthogonal projections in $\tilde{\mathcal{M}}$ with the properties $\Sigma_{k \in J} q_k = 1$, $\Psi_1(q_k) = 0$, $\forall k \in J$. The set J cannot be finite. This would imply $\Psi_1(1) = 0$. Since $\mathcal{M}$ was supposed to be σ-finite, the set J is countable. Identify $J = \mathbb{N}$.

The W^*-algebra generated by $\{q_k \mid k \in \mathbb{N}\}$ is *-isomorphic to $\mathcal{L}_\infty(\mathbb{N})$ in such a way that the q_k, $k \in \mathbb{N}$, correspond to the atoms in $\mathcal{P}(\mathcal{L}_\infty(\mathbb{N}))$. Since $\Psi_1(q_k) = 0$, $\forall k \in \mathbb{N}$, it follows that $\Psi_1|_{C_0(\mathbb{N})} \equiv 0$, where $C_0(\mathbb{N}) \subseteq \mathcal{L}_\infty(\mathbb{N})$ is considered as a subalgebra of $\tilde{\mathcal{M}}$. The assertion is then an immediate consequence of Lemma 2.

Remark: For type I factors the following theorem has been proved by Anderson (cf. Ref. 5: pp. 49 and 50).

Theorem 5: Let ϕ be a *singular* state on a *factor* $\mathcal{M}$. Then projections $e, f \in \mathcal{M}$ exist with $\phi(e) = \phi(f) = 1$ and $e \wedge f = 0$.

Proof: A factor $\mathcal{M}$ is of type I_n, $n \in \mathbb{N}$, I_∞, II_1, II_∞, or III. These cases are studied separately.

(i) Let $\mathcal{M}$ be of type III. Since ϕ is singular there exists $r \in \mathcal{P}(\mathcal{M})$, $r \notin \{0,1\}$, such that $\phi(r) = 1$. Since any two non-zero projections in a σ-finite type III factor are equivalent (Ref. 6: Proposition V. 1.39), $r \sim r^\perp$ holds and the assertion follows from Theorem 4.

(ii) Let $\mathcal{M}$ be of type II_1. Since ϕ is singular there exist pairwise orthogonal projections (see proof of Theorem 4) q_k, $k \in \mathbb{N}$, in $\mathcal{M}$ with $\Sigma_{k \in \mathbb{N}} q_k = 1$, $\phi(q_k) = 0$, $\forall k \in \mathbb{N}$. Let Tr be the canonical normalized trace on $\mathcal{M}$. Then $\Sigma_{k \in \mathbb{N}} \mathrm{Tr}(q_k) = 1$. Therefore $k_0 \in \mathbb{N}$ exists such that $\mathrm{Tr}(\Sigma_{k=1}^{k_0} q_k) > \frac{1}{2}$. Set

$$q \overset{\text{def}}{=} \sum_{k=1}^{k_0} q_k .$$

$$\Rightarrow \mathrm{Tr}(q) > \mathrm{Tr}(q^\perp).$$

$$\Rightarrow q^\perp \lesssim q.$$

$$\Rightarrow \exists \text{ a projection } p \lesssim q \text{ such that } q^\perp \sim p.$$

Set $s \overset{\text{def}}{=} p + q^\perp$ and consider the factor $s\mathcal{M}s$. The restriction $\phi|_{s\mathcal{M}s}$ of ϕ to $s\mathcal{M}s$ is a singular state with $\phi|_{s\mathcal{M}s}(s) = \phi|_{s\mathcal{M}s}(q^\perp) = 1$. Furthermore within $s\mathcal{M}s$ the projection p is the orthogonal complement of $q^\perp$ and $q^\perp \sim p$. The assertion then follows from Theorem 4.

(iii) Let $\mathcal{M}$ be of type I_∞ or II_∞. It will be shown first that an *infinite* projection s exists in $\mathcal{M}$ such that $\phi(s) = 0$: Consider pairwise orthogonal q_k, $k \in \mathbb{N}$, with $\Sigma_{k \in \mathbb{N}} q_k = 1$ and $\phi(q_k) = 0$, $\forall k \in \mathbb{N}$ (cf. the proof of Theorem 4). The q_k's may be supposed to be finite (if q_{k_0} is infinite, set $s = q_{k_0}$). Since $\mathrm{Tr}(1) = \infty$, one can suppose that $\mathrm{Tr}(q_k) > 1$, $\forall k \in \mathbb{N}$. The W^*-algebra generated by the operators q_k, $k \in \mathbb{N}$, is *-isomorphic to $\mathcal{L}_\infty(\mathbb{N})$ [see (ii)]. If I: $\mathcal{L}_\infty(\mathbb{N}) \to W^*\{q_k \mid k \in \mathbb{N}\} \subseteq \mathcal{M}$ is such a *-isomorphism, it follows from $\phi(q_k) = 0$, $\forall k \in \mathbb{N}$, that $\phi \circ I|_{C_0(\mathbb{N})} \equiv 0$, $C_0(\mathbb{N}) \subseteq \mathcal{L}_\infty(\mathbb{N})$.

By use of the axiom of choice one can construct a maximal family $(p_\kappa)_{\kappa \in J}$ (J an index set) of projections in $\mathcal{L}_\infty(\mathbb{N})$ such that $p_\kappa \cdot p_\delta \in C_0(\mathbb{N})$ if $\kappa \neq \delta$, the subsets A_κ of $\mathbb{N}$ corresponding to the projections p_κ, $\kappa \in J$ have infinitely many elements, the complement of a union $\cup_{j=1}^n A_{\kappa_j}$ of finitely many A_κ's is a set with infinitely many elements.

Such a family *cannot be countable.* If this were the case, $J = \mathbb{N}$, one could set $B_k = \cup_{j=1}^k A_k$ and take elements $n_k \in B_k \setminus B_{k-1}$, $k = 2,3,\ldots$. Then the projection p_∞ corresponding to the subset $A_\infty \overset{\text{def}}{=} \cup_{k=2,3,\ldots} n_k$ of $\mathbb{N}$ would extend the family $(p_k)_{k \in \mathbb{N}}$ contradicting its maximality.

Consider a finite number of projections p_{κ_j}, $j = 1,..,n$. Then

$$\begin{aligned}
\phi\left(\sum_{j=1}^n I(p_{\kappa_j})\right) &= \phi(I(p_{\kappa_1}) + I(p_{\kappa_2} \wedge p_{k_1}^\perp) + I(p_{k_2} \wedge p_{\kappa_1}) + \cdots \\
&\quad + I(p_{\kappa_j} \wedge (p_{\kappa_{j-1}} \vee \cdots \vee p_{\kappa_1})^\perp) + I(p_{\kappa_j} \wedge (p_{k_{j-1}} \vee \cdots \vee p_{\kappa_1})) + \cdots \\
&\quad + I(p_{\kappa_n} \wedge (p_{\kappa_{n-1}} \vee \cdots \vee p_{\kappa_1})^\perp) + I(p_{\kappa_n} \wedge (p_{\kappa_{n-1}} \vee \cdots \vee p_{\kappa_1}))) \\
&\overset{\#}{=} \phi(I(p_{\kappa_1}) + I(p_{\kappa_2} \wedge p_{\kappa_1}^\perp) + \cdots + I(p_{\kappa_n} \wedge (p_{k_{n-1}} \vee \cdots \vee p_{\kappa_1})^\perp)) \leqslant \phi(I(1)) \leqslant 1 .
\end{aligned}$$

At (#) the fact was used that $p_{\kappa_j} \wedge (p_{\kappa_{j-1}} \vee \cdots \vee p_{\kappa_1}) \in C_0(\mathbb{N})$, $\forall j = 2,...,n$. Consequently, the set $\{\kappa \in J \mid \phi(I(p_\kappa)) \neq 0\}$ has only countably many elements. In particular, there exists $\kappa_0 \in J$ such that $\phi(I(p_{\kappa_0})) = 0$. Here $I(p_{\kappa_0})$ is an infinite projection [because $\mathrm{Tr}(q_k) > 1$, $\forall k \in \mathbb{N}$, and $I(p_{\kappa_0})$ is an infinite sum of different q_k's]. Thus $s \overset{\text{def}}{=} I(p_{\kappa_0})$ is an infinite projection with $\phi(s) = 0$. If $s^\perp$ is infinite, too, the assertion follows just as in (i), because in this situation $s \sim s^\perp$ holds (Ref. 6: Proposition 1.39). If $s^\perp$ is finite, a reasoning similar to that of (ii) proves the assertion.

(iv) Since factors of type I_n, $n \in \mathbb{N}$, do not admit singular states, the theorem is proved. Q.E.D.

Corollary 1: Let ϕ be a *pure* state on a *factor* $\mathcal{M}$ fulfilling the Jauch–Piron property. Then ϕ is normal and $\mathcal{M}$ is a factor of type I.

Proof: A pure state is either normal or singular. Only type I factors possess pure normal states. Q.E.D.

Corollary 2: Let (T,τ) be a maximal *non-normal* truth functional in a factor $\mathcal{M}$. Consider an arbitrary (T,τ)-ex-

tending state ϕ (which might be nonpure if the extension procedure is not unique!). Then ϕ does not fulfill the Jauch–Piron condition.

Proof: If (T,τ) is maximal and non-normal, any extending state ϕ is singular (Ref. 5: Theorem II.3.2). The corollary then is a consequence of Theorem 5. Q.E.D.

Corollary 3: A state ϕ on a *factor* $\mathcal{M}$ is singular if and only if projections $e, f \in \mathcal{M}$ exist with $\phi(e) = \phi(f) = 1$ and $e \wedge f = 0$.

Proof: Assume ϕ is nonsingular, i.e., the decomposition $\phi = \lambda\phi_n + (1-\lambda)\phi_s$ into a normal state ϕ_n and a singular state ϕ_s is nontrivial, $0 < \lambda < 1$. Let e, f be projections with $\phi(e) = \phi(f) = 1$. This implies $\phi_n(e) = \phi_n(f) = 1$ and (by Theorem 3) $\phi_n(e \wedge f) = 1$. In particular, $e \wedge f \neq 0$. Q.E.D.

Corollary 4: Let $\mathcal{M}$ be a *factor* and consider a representation $\pi: \mathcal{M} \to \mathcal{B}(\mathcal{H})$ of $\mathcal{M}$ on a Hilbert space $\mathcal{H}$ with the property

$$\pi(p \wedge q) = \pi(p) \wedge \pi(q), \quad \forall p,q \in \mathcal{P}(\mathcal{M}) . \tag{1}$$

Then π is σ-weakly continuous.

Proof: There exists a central projection $z \in \pi(\mathcal{M})''$ such that $\pi_n: \mathcal{M} \ni x \to (\pi(x)z) \in \mathcal{B}(z\mathcal{H})$ is σ-weakly continuous and $\pi_s: \mathcal{M} \ni x \to (\pi(x)(1-z)) \in \mathcal{B}((1-z)\mathcal{H})$ is singular, i.e., $\mathcal{M} \ni x \to \langle \xi | \pi(x)(1-z)\xi \rangle \in \mathbb{C}$ is singular for every $\xi \in \mathcal{H}$ (Ref. 6: Theorem II.2.14). Suppose $z \neq 1$. From (1) it can be inferred that the singular states ω: $\mathcal{M} \ni x \to \langle \xi | \pi(x)\xi \rangle \in \mathbb{C}$, $\xi \in (1-z)\mathcal{H}$, $\|\xi\| = 1$, have the Jauch–Piron property. This contradicts Theorem 5. Thus $z = 1$ and $\pi = \pi_n$ is σ-weakly continuous. Q.E.D.

The results of this section may be regarded as an *a posteriori* explanation of the fact that only *normal* states on factors are considered in physics. Corollary 1 was conjectured in 1981 by Raggio and the author. A preliminary version of this corollary—not comprising factors of type II_1—was derived in July of 1986 by Zsido together with the author.

IV. JAUCH–PIRON STATES ON ARBITRARY W*-ALGEBRAS

In this section $\mathcal{M}$ is assumed to be a W^*-algebra with *separable* predual. Recall that a W^*-algebra has separable predual iff it is *-isomorphic to a von Neumann algebra on a *separable* Hilbert space. Attention will be focused on *pure* states with the Jauch–Piron property. The Gel'fand–Naimark–Segal (in short GNS) representation with respect to such a state will be an essential tool.

Theorem 6: Let ϕ be a pure state with the Jauch–Piron property on a W^*-algebra $\mathcal{M}$. Let $(\mathcal{H}_\phi, \pi_\phi, \Omega_\phi)$ denote the GNS representation of $\mathcal{M}$ with respect to ϕ. Then

$$\pi_\phi(p \wedge q) = \pi_\phi(p) \wedge \pi_\phi(q), \quad \forall p,q \in \mathcal{P}(\mathcal{M})$$

holds true.

Proof: For $p,q \in \mathcal{P}(\mathcal{M})$, $\pi_\phi(p \wedge q) \leqslant \pi_\phi(p) \wedge \pi_\phi(q)$ is obviously fulfilled. Suppose $\pi_\phi(p \wedge q) \neq \pi_\phi(p) \wedge \pi_\phi(q)$. Consider a vector $\xi \in \mathcal{H}_\phi$, $\|\xi\| = 1$, $\pi_\phi(p \wedge q)\xi = 0$, $\{\pi_\phi(p) \wedge \pi_\phi(q)\}\xi = \xi$. Since $\pi_\phi(\mathcal{M})$ acts irreducibly on $\mathcal{H}_\phi$ (Ref. 10: 10.2.3), there exists (Ref. 10: Theorem 5.4.5) a self-adjoint operator $\tilde{H} \in \pi_\phi(\mathcal{M})$ such that for the unitary $\tilde{U} \overset{\text{def}}{=} \exp i\tilde{H}$ one has $\tilde{U}\xi = \Omega_\phi$. Due to Ref. 10: 4.6.2 a self-adjoint operator $H \in \mathcal{M}$ exists with $\pi_\phi(H) = \tilde{H}$. Setting $U \overset{\text{def}}{=} \exp iH$, one has $\pi_\phi(U) = \tilde{U}$. For an arbitrary $s \in \mathcal{P}(\mathcal{M})$ the following holds:

$$\begin{aligned}
&\langle \xi | \pi_\phi(s)\xi \rangle \\
&\quad = \langle \xi | \pi_\phi(U^*U)\pi_\phi(s)\pi_\phi(U^*U)\xi \rangle \\
&\quad = \langle \tilde{U}\xi | \pi_\phi(UsU^*)\tilde{U}\xi \rangle = \langle \Omega_\phi | \pi_\phi(UsU^*)\Omega_\phi \rangle \\
&\quad = \phi(UsU^*) .
\end{aligned}$$

Due to the definition of ξ we have $\phi(UpU^*) = \phi(UqU^*) = 1$, $\phi(U(p \wedge q)U^*) = 0$. Since ϕ has the Jauch–Piron property, this leads to a contradiction $\phi(U(p \wedge q)U^*) = \phi((UpU^*) \wedge (UqU^*)) = 1$. Q.E.D.

Remark: Theorem 6 together with the observation of Anderson (quoted in Ref. 5: pp. 49 and 50) that singular states on a type I factor do not have the Jauch–Piron property, can be used to prove Corollary 1 for factors of type III and II_∞. Compare the remark at the end of Sec. III.

Lemma 2: Let ϕ be a pure state on the W^*-algebra

$$\mathcal{M} \overset{\text{def}}{=} \mathcal{L}_\infty([0,1]) \bar{\otimes} \mathcal{M}_2$$

[$[0,1] = \{x \in \mathbb{R} | 0 \leqslant x \leqslant 1\}$; $\mathcal{L}_\infty([0,1])$ is the algebra of (classes of) Borel measurable functions, which are essentially bounded with respect to Lebesgue measure]. Then projections $e, f \in \mathcal{M}$ exist with $\phi(e) = \phi(f) = 1$ and $e \wedge f = 0$.

Proof: Let $C([0,1])$ denote the algebra of continuous functions on the interval $[0,1]$. Here $C([0,1]) \subseteq \mathcal{L}_\infty([0,1])$. Considering the (irreducible!) GNS-representation π_ϕ of $\mathcal{M}$ with respect to ϕ, one sees that ϕ is a product state, that $\phi|_{C([0,1]) \otimes 1}$ is a character and $\phi|_{1 \otimes \mathcal{M}_2}$ is a (normal) pure state. Therefore $\phi|_{C([0,1]) \otimes 1}$ is represented by a point $x_0 \in [0,1]$: $\phi(m \otimes 1) = m(x_0)$, $\forall m \in C([0,1])$ (cf. Ref. 6: Proposition I.4.5). The support of $\phi|_{1 \otimes \mathcal{M}_2}$ is given by an atom r in $1 \otimes \mathcal{M}_2$. In appropriate coordinates $r = 1 \otimes \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Here $\mathcal{L}_\infty([0,1]) \bar{\otimes} \mathcal{M}_2 \cong \mathcal{L}_\infty([0,1], \mathcal{M}_2)$ (cf. Ref. 6: Chap. IV.7), i.e., every element of $\mathcal{M}$ can be represented by a measurable essentially bounded function from $[0,1]$ into $\mathcal{M}_2$. Define

$$e(x) \overset{\text{def}}{=} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \forall x \in [0,1] ,$$

$$f(x) \overset{\text{def}}{=} \begin{pmatrix} 1 - |x - x_0| & \sqrt{|x - x_0| - |x - x_0|^2} \\ \sqrt{|x - x_0| - |x - x_0|^2} & |x - x_0| \end{pmatrix},$$

$$x \in [0,1] .$$

Note that f is an element of $C([0,1]) \otimes \mathcal{M}_2 \cong C([0,1], \mathcal{M}_2) \subseteq \mathcal{L}_\infty([0,1], \mathcal{M}_2)$. Since $e(x) \wedge f(x) = 0$, $\forall x \neq x_0$, and $\{x_0\}$ is a Lebesgue null set, $e \wedge f = 0$ follows,

$$\begin{aligned}
\phi(e) &= \phi\left(1 \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right) = 1 , \\
\phi(f) &= \phi\left(\left(1 \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right) f \left(1 \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right)\right) \\
&= \phi\left(\{[0,1] \ni x \to (1 - |x - x_0|)\} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right) = 1 .
\end{aligned}$$

Q.E.D.

Theorem 7: Let ϕ be a *pure* state on a W^*-algebra $\mathcal{M}$, which does not admit a commutative direct summand. Suppose the center $\mathcal{Z}(\mathcal{M})$ of $\mathcal{M}$ does not have any minimal projections. Then ϕ does not fulfill the Jauch–Piron condition.

Proof: The W^*-algebra $\mathcal{M}$ is a direct sum $\mathcal{M} = \mathcal{M}_{\mathrm{I}} \oplus \mathcal{M}_{\mathrm{II}} \oplus \mathcal{M}_{\mathrm{III}}$ (Ref. 6: Theorem V.1.19) of W^*-algebras of type I, II, and III. Since ϕ is pure, it vanishes on all but one summand of this decomposition. Thus one can suppose $\mathcal{M}$ to be of a fixed type. Let $\pi_\phi: \mathcal{M} \to \mathcal{B}(\mathcal{H}_\phi)$ be the GNS representation of $\mathcal{M}$ with respect to ϕ.

(i) Let $\mathcal{M}$ be of type II or III. Then a W^*-algebra $\tilde{\mathcal{M}}$ exists such that $\mathcal{M} \cong \tilde{\mathcal{M}} \otimes \mathcal{M}_2$ (Ref. 6: Proposition V.1.35 and Proposition V.1.22). Here $\mathcal{Z}(\mathcal{M}) \cong \mathcal{Z}(\tilde{\mathcal{M}}) \cong \mathcal{L}_\infty([0,1])$ (Ref. 6: Theorem III.1.22). Consider a vector $\xi \in \mathcal{H}_\phi$, $\|\xi\| = 1$ with $\pi_\phi\left(1 \otimes \left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right)\right)\xi = \xi$. Such a vector always exists since every representation of $\mathcal{M}_2$ is faithful. The state $\Psi(x) \overset{\text{def}}{=} \langle \xi | \pi_\phi(x)\xi \rangle$, $x \in \mathcal{M}$, and its restriction to $\mathcal{Z}(\tilde{\mathcal{M}}) \bar{\otimes} \mathcal{M}_2 \cong \mathcal{L}_\infty([0,1]) \bar{\otimes} \mathcal{M}_2$ fulfills the Jauch–Piron condition (this is a consequence of Theorem 6). Here $\Psi|_{\mathcal{L}_\infty([0,1]) \bar{\otimes} \mathcal{M}_2}$ is a pure state and the assertion follows from Lemma 2 by contradiction.

(ii) Let $\mathcal{M}$ be of type I. Since ϕ is pure, it can be supposed that $\mathcal{M}$ is either of type I_∞ or a direct sum $\mathcal{M} \cong \bigotimes_{j \in J} \{\mathcal{A}_j \otimes \mathcal{M}_j\}$, $J \subseteq \{2,3,\ldots\}$, where $\mathcal{A}_j$, $j \in J$, is a commutative W^*-algebra and $\mathcal{M}_j$ is the $j \times j$-matrix algebra (Ref. 6: Theorem V.1.27). The former case can be handled just as in (i). In the latter one considers projections $q_j \in \mathcal{M}_j$ of even dimension

$$\dim(q_j) = \begin{cases} j \text{ if } j \text{ is even}, \\ j-1 \text{ if } j \text{ is odd}, \end{cases} \quad j \in J,$$

$$q \overset{\text{def}}{=} \left(\bigoplus_{j \in J} 1 \otimes q_j\right) \in \left\{\bigoplus_{j \in J} \mathcal{A}_j \bar{\otimes} \mathcal{M}_j\right\}.$$

(α) Assume $\pi_\phi(q) \neq 0$. Since $\mathcal{Z}(\mathcal{M})$ has no atoms, there cannot exist atoms in the commutative W^*-algebras $\mathcal{A}_j$, $j \in J \Rightarrow \mathcal{A}_j \cong \mathcal{L}_\infty([0,1])$, $\forall j \in J$. All projections q_j are the sum of two equivalent orthogonal projections

$$\Rightarrow q_j \mathcal{M}_j q_j \cong \begin{cases} \mathcal{M}_{j/2} \bar{\otimes} \mathcal{M}_2 & \text{if } j \text{ is even}, \\ \mathcal{M}_{(j-1)/2} \bar{\otimes} \mathcal{M}_2 & \text{if } j \text{ is odd}. \end{cases}$$

In particular, $\mathcal{M}_q \cong \oplus_{j \in J}\{\mathcal{A}_j \otimes q_j \mathcal{M}_j q_j\}$ contains a W^*-subalgebra which is *-isomorphic to $\mathcal{M}_2$. Since $\pi_\phi|_{\mathcal{M}_q}$: $\mathcal{M}_q \to \mathcal{B}(\pi_\phi(q)\mathcal{H}_\phi)$ has the property (1), $\mathcal{Z}(\mathcal{M}_q) = q\mathcal{Z}(\mathcal{M})q \cong \mathcal{L}_\infty([0,1])$, and $\pi_\phi|_{\mathcal{M}_q}(\mathcal{Z}(\mathcal{M}_q)) \subseteq \mathbb{C} \cdot \pi_\phi(q)$, a pure state with the Jauch–Piron property on $\mathcal{L}_\infty([0,1]) \bar{\otimes} \mathcal{M}_2$ can be constructed just as in (i). This contradicts Lemma 2.

(β) Assume $\pi_\phi(q) = 0 \Rightarrow \pi_\phi(q^\perp) = 1$. Here $q_j^\perp$ vanishes for j even and is an atom in $\mathcal{M}_j$ if j is odd. A projection p can be constructed out of projections $p_j \geq q_j^\perp$, $j \in J$, such that $\dim p_j = 2$, $\forall j \in J$, $\pi_\phi(p) \neq 0$. In the same manner as under (α) this leads to a contradiction to Lemma 2. Q.E.D.

Theorem 8: Let $\mathcal{M}$ be a W^*-algebra with an atomic center $\mathcal{Z}(\mathcal{M})$ and without any commutative summand. Consider a *pure* state ϕ on $\mathcal{M}$ with the Jauch–Piron property. Then ϕ is *normal* and an atom $z_0 \in \mathcal{Z}(\mathcal{M})$ exists with

(i) $\phi(z_0) = 1$,

(ii) $z_0 \mathcal{M} z_0$ is a factor of type 1.

Proof: $\mathcal{M}$ may be supposed to be of a fixed type (cf. proof of Theorem 7). If there is an atom z in $\mathcal{Z}(\mathcal{M})$ with $\phi(z) \neq 0$, the assertion follows from Corollary 1. Therefore $\mathcal{Z}(\mathcal{M})$ can be supposed to be infinite dimensional, $\mathcal{Z}(\mathcal{M}) \cong \mathcal{L}_\infty(\mathbb{N})$, and $\phi|_{C_0(\mathbb{N})} \equiv 0 \cdot \Rightarrow \mathcal{M} \cong \oplus_{j \in \mathbb{N}} \mathcal{R}_j$, where the $\mathcal{R}_j$'s are factors. Similarly as in the proof of Theorem 7, a projection $q \in \mathcal{M}$ can be found ($q = 1$, if $\mathcal{M}$ is of type II or III) with $\pi_\phi(q) \neq 0$ such that $q\mathcal{M}q$ contains a W^*-subalgebra $\mathcal{L}_\infty(\mathbb{N}) \bar{\otimes} \mathcal{M}_2$. This contradicts Lemma 1 (cf. again the proof of Theorem 7). Q.E.D.

Let $\mathcal{M}$ be a W^*-algebra with a commutative summand $\mathcal{A}$. Then $\mathcal{M}$ is *-isomorphic to the direct sum $\tilde{\mathcal{M}} \oplus \mathcal{A}$ for an appropriate W^*-algebra $\tilde{\mathcal{M}}$. If Ψ is a pure state on $\mathcal{A}$, the state $\phi(x,y) \overset{\text{def}}{=} \Psi(y)$, $x \in \tilde{\mathcal{M}}$, $y \in \mathcal{A}$, defines a Jauch–Piron state on $\tilde{\mathcal{M}} \oplus \mathcal{A} \cong \mathcal{M}$. Therefore the assumption in Theorems 7 and 8—that $\mathcal{M}$ has no commutative summand—cannot be omitted.

Corollary 5: Let $\mathcal{M}$ be a W^*-algebra without a commutative summand and consider a *pure* state ϕ on $\mathcal{M}$ obeying the Jauch–Piron property. Then ϕ is normal and an atom $z_0 \in \mathcal{Z}(\mathcal{M})$ exists with

(i) $\phi(z_0) = 1$,

(ii) $z_0 \mathcal{M} z_0$ is a factor of type I.

Corollary 5 summarizes all results of this paper on *pure* states. A remaining question of interest concerns the existence of truth functionals (T,τ) in a W^*-algebra $\mathcal{M}$ such that $T = \mathcal{P}(\mathcal{M})$. The unique extending state ϕ [$\phi(p) = \tau(p)$, $\forall p \in \mathcal{P}(\mathcal{M})$; see Theorem 1] then is a character, i.e., $\phi(xy) = \phi(x) \cdot \phi(y)$, $\forall x,y \in \mathcal{M}$. Investing little effort one can prove the following (essentially known) result (cf. Refs. 11 and 12).

Theorem 9: Let $\mathcal{M}$ be a W^*-algebra and ϕ be a multiplicative state on $\mathcal{M}$ [i.e., $\phi(x \cdot y) = \phi(x) \cdot \phi(y)$, $\forall x,y \in \mathcal{M}$]. Then a central projection c exists with $\phi(c) = 1$ and such that $c\mathcal{M}c$ is commutative.

V. CONCLUDING REMARKS

An individual interpretation of quantum mechanics presupposes the existence of "suitable" *pure* states (in contrast to mixed ones for an ensemble interpretation). In the present context, suitable means of course that the Jauch–Piron condition is fulfilled.

The results of this paper show that pure Jauch–Piron states exist essentially only on commutative W^*-algebras and type I factors. More precisely, let ϕ be a pure Jauch–Piron state on a W^*-algebra $\mathcal{M}$. Then W^*-algebras $\mathcal{R}$ and $\mathcal{L}$ exist, such that $\mathcal{M} \cong \mathcal{R} \oplus \mathcal{L}$ where $\mathcal{R}$ is arbitrary, $\mathcal{L}$ is either commutative or *-isomorphic to the algebra $\mathcal{B}(\mathcal{H})$ of bounded linear operators on an appropriate Hilbert space $\mathcal{H}$, and ϕ restricted to $\mathcal{R}$ vanishes. Note that ϕ may be a singular state if $\mathcal{L}$ is commutative. For $\mathcal{L} \cong \mathcal{B}(\mathcal{H})$, ϕ is always normal and represented by a vector $\xi \in \mathcal{H}$, $\|\xi\| = 1$: $\phi(x) = \langle \xi | x\xi \rangle$, $\forall x \in \mathcal{B}(\mathcal{H})$.

Therefore an individual interpretation of quantum mechanics is only possible in classical theories (whose corresponding "algebra of observables" is commutative) and Hilbert space quantum mechanics. The interpretation of systems whose respective W^*-algebra is of type III—such as algebraic statistical mechanics—has to be nonindividual or at least partly nonindividual: The classical properties of such a system—corresponding to the center of the respective W^*-algebra $\mathcal{M}$—could behave as belonging to an individual system if the state on the whole of $\mathcal{M}$ is a *factor* state (with the Jauch–Piron property) and thus its restriction to the center of $\mathcal{M}$ is pure.

ACKNOWLEDGMENTS

My thanks go to Dr. G. Raggio, who laid the foundations of this paper in his thesis of 1981. Furthermore, I am very much indebted to Professor L. Zsido for clarifying and catalytic discussions at the IAMP Congress (Marseille) in July 1986. Finally, I thank Professor Lassner and Dr. Alberti from the University of Leipzig for their hospitality. Most of the results of this paper have been worked out during my stay there in August 1986.

[1]A. M. Gleason, J. Math. Mech. **6**, 885 (1957).
[2]H. Primas, *Chemistry, Quantum Mechanics, and Reductionism. Perspectives in Theoretical Chemistry* (Springer, Berlin, 1983), 2nd corrected edition.
[3]A. Amann, Fortschr. Phys. **34**, 167 (1986).
[4]J. M. Jauch, "Foundations of quantum mechanics," in *Proceedings of the International School of Physics Enrico Fermi, Course 49*, edited by B. d'Espagnat (Academic, New York 1971).
[5]G. Raggio, Ph.D. thesis, ETH No. 6824, Eidgenössische Technische Hochschule, Zürich, 1981.
[6]M. Takesaki, *Theory of Operator Algebras I* (Springer, New York, 1979).
[7]G. T. Rüttimann, J. Math. Phys. **18**, 189 (1977).
[8]J. Anderson, Trans. Am. Math. Soc. **249**, 303 (1979).
[9]E. Størmer, Proc. Am. Math. Soc. **19**, 1100 (1968).
[10]R. V. Kadison and J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras* (Academic, New York, 1983, 1986), 2 volumes.
[11]B. Misra, Nuovo Cimento A **47**, 841 (1967).
[12]R. J. Plymen, Nuovo Cimento A **54**, 862 (1968).

The Wigner transformation is of finite order

Joseph C. Várilly and José M. Gracia-Bondía
Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica

(Received 20 November 1986; accepted for publication 20 May 1987)

The Wigner integral transformation, which intertwines the twisted product and the composition of kernels, is of order 24. Indeed, it commutes with, and its sixth power equals, the Fourier cotransformation.

I. INTRODUCTION

In the Weyl–Wigner–Moyal formulation of quantum mechanics,[1–4] observables are functions on phase space and the composition of observables is given by the twisted product or Moyal product of functions. It is well known that this operation is equivalent to composition of $2n$-variable kernels,[5–8] and this equivalence is implemented by an integral transformation known as the Wigner transformation[8] (or alternatively the Weyl transformation[5]). The Wigner transforms of kernels corresponding to positive operators have been extensively studied under the name of "Wigner distribution functions."[3,9] They are also widely used in the theory of radar signal processing as "radar autocorrelation functions."[10,11] Thus it is of great interest to know as much as possible about the intrinsic properties of the Wigner transformation.

If one considers the Wigner transformation as an isomorphism of $L^2(\mathbb{R}^{2N})$ onto itself, it is easily seen to be unitary. Moreover, as Pool[12] and Cressman[13] noted long ago, the Wigner transformation factorizes into a reflection on phase space (of order 2) and a partial Fourier transformation (of order 4); however, these factors do not commute. In this article we establish that the Wigner transformation is itself of finite order, namely, of order 24, and indeed its sixth power is just the ($2n$-variable) Fourier cotransformation. The proof we give involves merely a transfer of context to the so-called Bargmann representation, wherein the proof reduces to an elementary computation.

In Sec. II we recall the definition of the Wigner transformation and some of its properties. In Sec. III we briefly review the Bargmann spaces of analytic functions. Using these we show in Sec. IV that the Wigner transformation is of order 24.

II. THE WIGNER TRANSFORMATION

The Wigner–Weyl–Moyal formulation of quantum mechanics starts from the Weyl quantization which establishes a correspondence between "symbols" f, i.e., functions on phase space, and operators A of the conventional formulation. For our present purposes, it suffices to consider that "flat" phase space $\mathbb{R}^{2N} = T^*(\mathbb{R}^N)$. Here the Weyl quantization rule may be formally written as

$$A = (2\pi)^{-2N} \int_{\mathbb{R}^N}\int_{\mathbb{R}^N} (Ff)(a,b) \times \exp[i(a\cdot Q + b\cdot P)]\,da\,db,$$

where F denotes the Fourier transform and $Q_1,\ldots,Q_N$ and $P_1,\ldots,P_N$ denote the usual position and momentum operators.

Composition of operators corresponds then to the "twisted product" of symbols,

$$(f\times g)(u) := (2\pi)^{-2N} \int_{\mathbb{R}^{2N}}\int_{\mathbb{R}^{2N}} f(v)g(w) \times \exp(i(uJv + vJw + wJu))\,dv\,dw \qquad (1)$$

[where $u,v,w\in\mathbb{R}^{2N}$, and $uJv := \Sigma_{k=1}^{N}(u_k v_{N+k} - u_{N+k}v_k)$ is the symplectic product of two vectors in $\mathbb{R}^{2N}$], and, at a formal level, the usual operator calculus is replaced by a "twisted product calculus" whose objects are functions on phase space. The twisted product is noncommutative, as is the "kernel product" on $\mathbb{R}^{2N}$, namely the product $f\circ g$ given by

$$(f\circ g)(x,y) := (2\pi)^{-N/2} \int_{\mathbb{R}^N} f(x,z)g(z,y)\,dz$$

(where $x,y,z\in\mathbb{R}^N$). In fact, these two products are isomorphic; the Wigner transformation W, which we will now recall, satisfies $W(f\times g) = Wf\circ Wg$.

To make the above discussion rigorous, we observe that (1) makes sense whenever $f,g\in\mathcal{S}(\mathbb{R}^{2N})$, the space of rapidly decreasing smooth functions over $\mathbb{R}^{2N}$; then also $f\times g\in\mathcal{S}(\mathbb{R}^{2N})$ and the product is continuous on $\mathcal{S}(\mathbb{R}^{2N})$. The same is true for the kernel product, or if $\mathcal{S}(\mathbb{R}^{2N})$ is replaced by the space of square-integrable functions $L^2(\mathbb{R}^{2N})$. Let $\mathcal{S}'(\mathbb{R}^{2N})$ denote the space of tempered distributions over $\mathbb{R}^{2N}$. Define the isomorphisms (of Fréchet spaces) R, Φ, W from $\mathcal{S}(\mathbb{R}^{2N})$ onto $\mathcal{S}(\mathbb{R}^{2N})$ by

$$(Rf)(x,y) := f(2^{-1/2}(x+y),\,2^{-1/2}(x-y)),$$

$$(\Phi f)(x,y) := (2\pi)^{-N/2} \int_{\mathbb{R}^N} f(x,t)\exp(iyt)\,dt,$$

$$(Wf)(x,y) := (R\Phi f)(x,y) = (2\pi)^{-N/2} \int_{\mathbb{R}^N} f(2^{-1/2}(x+y),t) \times \exp(2^{-1/2}i(x-y)t)\,dt,$$

noting that these extend to unitary operators on $L^2(\mathbb{R}^{2N})$. The Wigner transformation is the operator W.

We will write x, y, t, p, q, r for vectors in $\mathbb{R}^N$ and u, v, w for vectors in $\mathbb{R}^{2N}$. We also find it convenient to use the Haar measures $d\hat{}x := (4\pi)^{-N/2}\,dx$ on $\mathbb{R}^N$, $d\hat{}u := (4\pi)^{-N}\,du$ on $(\mathbb{R}^{2N})$ (where dx, du denote the usual Lebesgue measures).

The Fourier transformation F on $\mathcal{S}(\mathbb{R}^{2N})$ is given by

 0022-2488/87/102390-03$02.50 © 1987 American Institute of Physics

$$(Ff)(u) := (2\pi)^{-N}\int_{\mathbf{R}^{2N}} f(v)e^{-iuv}\,dv.$$

We note also that

$$\int_{\mathbf{R}^N}\exp(-x^2+2^{1/2}xt)d\hat{}t = 2^{-N}\exp\left(\frac{t^2}{2}\right),$$
$$\int_{\mathbf{R}^{2N}}\exp(-u^2+2^{1/2}uv)d\hat{}u = 2^{-2N}\exp\left(\frac{v^2}{2}\right). \tag{2}$$

We may now verify that $W(f\times g) = Wf\circ Wg$ by a direct calculation. By duality (see Ref. 8) this relation holds also for f or g in $\mathcal{S}'(\mathbf{R}^{2N})$.

Standard treatments of the one-dimensional harmonic oscillator involve the Hermite functions (suitably normalized for our purposes) $h_m\in\mathcal{S}(\mathbf{R})$ given by

$$h_m(x) := (2^{m-1}m!)H_m(x)\exp(-x^2/2)$$
$$(x\in\mathbf{R},\ m=0,1,2,\ldots).$$

Let us set $h_{mn}(x,y) := h_m(x)h_n(y)$; then in the case $N=1$, one computes that $h_{mn} = W(f_{mn})$, where the f_{mn} are the Wigner functions corresponding to the harmonic oscillator transitions between states[8,10,11,14]:

$$f_{mn}(q,p) := 2(-1)^n(n!/m!)^{1/2}(q-ip)^{m-n}L_n^{m-n}(q^2+p^2)\times\exp(-(q^2+p^2)/2)$$

for $m\geqslant n$; $f_{mn} := f_{nm}^*$ (complex conjugate) if $m<n$. For $N>1$, we also have $h_{mn} = W(f_{mn})$, where we define, for given multi-indices $m,n\in\mathbf{N}^N$, h_m, h_{mn}, and f_{mn} as direct products of the corresponding one-variable or two-variable functions h_{m_i}, $h_{m_i n_i}$, $f_{m_i n_i}$.

Both h_{mn} and f_{mn} are eigenfunctions of the Fourier transformation F on $\mathcal{S}(\mathbf{R}^{2n})$, with common eigenvalue $(-i)^{m+n}$; thus F and W commute.[8] This suggests a possible relationship between the Wigner and Fourier transformations. On the other hand, both h_{mn} and f_{rs} are eigenfunctions of the Hermite operator with the same eigenvalue if and only if $m+n=r+s$ (Ref. 8); this would rather suggest that W is of infinite order. This question is resolved by the following somewhat surprising theorem, which is proved in Sec. IV.

Theorem: $W^3(h_{mn}) = \exp(i\pi(m+n)/4)h_{mn}$ for all $m,n\in\mathbf{N}^N$.

Corollary 1: $W^6 = F^{-1}$.

Corollary 2: $W^{24} = \mathrm{Id}$.

III. THE BARGMANN REPRESENTATION

The simplest approach we know to proving the above theorem is to pass to the Bargmann representation. We briefly review what is involved.[15–17] Let μ denote the Gaussian measure on $\mathbf{C}^N$:

$$d\mu(z) := \pi^{-N}\exp(-zz^*)dz = \pi^{-N}\exp(-(x^2+y^2))dx\,dy,$$

where $z = x+iy$. Let $\mathcal{F}(\mathbf{C}^N)$ denote the Hilbert space of entire analytic functions in $L^2(\mathbf{C}^N,\mu)$, and let

$$\mathcal{E}(\mathbf{C}^N) := \{f\in\mathcal{F}(\mathbf{C}^N):\ (1+zz^*)^m\exp(-zz^*/2)f(z)\ \text{is bounded, for all } m\in\mathbf{N}^N\}.$$

As shown in Ref. 16, $\mathcal{E}(\mathbf{C}^N)$ can be given the topology of a Fréchet space, such that the maps A_N: $\mathcal{S}(\mathbf{R}^N)\to\mathcal{E}(\mathbf{C}^N)$, V_N: $\mathcal{E}(\mathbf{C}^N)\to\mathcal{S}(\mathbf{R}^N)$, given by

$$(A_N h)(z) := \int_{\mathbf{R}^N} A(z,t)h(t)d\hat{}t,$$
$$(V_N f)(t) := \int_{\mathbf{C}^N} A(z^*,t)f(z)d\mu(z), \tag{3}$$

are Fréchet-space isomorphisms, with $V_N = A_N^{-1}$, the kernel $A(z,t)$ being given by

$$A(z,t) := 2^{N/2}\exp(-z^2/2 - t^2/2 + 2^{1/2}zt).$$

We recall[15] that $A_N h_m(z) = (m!)^{-1/2}z^m$ for $m\in\mathbf{N}^N$.

The spaces $\mathcal{E}(\mathbf{C}^N)$ and $\mathcal{F}(\mathbf{C}^N)$ have a reproducing kernel

$$\int_{\mathbf{C}^N}\exp(w^*z)f(w)d\mu(w) = f(z) \tag{4}$$

(see Refs. 15 and 16), from which we obtain the following formula:

$$\int_{\mathbf{C}^N}\exp(az+bz^*)d\mu(z) = \exp(ab). \tag{5}$$

The advantage of working with these spaces is that operators T on $\mathcal{S}(\mathbf{R}^N)$ or $L^2(\mathbf{R}^N)$ transfer to operators $A_N T V_N$ on $\mathcal{E}(\mathbf{C}^N)$ or $\mathcal{F}(\mathbf{C}^N)$ given by integral kernels.

As F denotes the Fourier transformation on $\mathcal{S}(\mathbf{R}^{2N})$, one shows easily using (2)–(4) that $(A_{2N}FV_{2N}f)(z) = f(-iz)$, for $f\in\mathcal{E}(\mathbf{C}^{2N})$, in agreement with Ref. 15.

IV. PROOF OF THE THEOREM

Since $(A_{2N}h_{mn})(z_1,z_2) = (m!n!)^{-1/2}z_1^m z_2^n$, it suffices to show that

$$(A_{2N}W^3h_{mn})(z_1,z_2) = (A_{2N}h_{mn})(\exp(i\pi/4)z_1,\exp(i\pi/4)z_2).$$

By the reproducing kernel property (4), this reduces to showing that the operator $A_{2N}W^3V_{2N}$ is given by the integral kernel $\exp(2^{-1/2}(1+i)w^*z)$. This we do in a few steps.

Step 1: For $f\in\mathcal{E}(\mathbf{C}_{2N})$, we obtain

$$(A_{2N}\Phi V_{2N}f)(z) = \int A(z,u)(\Phi V_{2N}f)(u)d\hat{}u = \iint 2^{N/2}A(z,u)e^{ipt}(V_{2N}f)(q,t)d\hat{}t\,d\hat{}u$$
$$= \iiint 2^{N/2}A(z,u)e^{ipt}A(w^*;q,t)f(w)d\mu(w)d\hat{}t\,d\hat{}u = \int M(z,w^*)f(w)d\mu(w)$$

[with $u=(q,p)$], where the kernel $M(z,w^*)$ is given by

$$M(z,w^*) := 2^{N/2}\int\int\int A(z;q,p)e^{ipt}A(w^*;q,t)d\hat{}q\,d\hat{}p\,d\hat{}t$$

$$= 2^{5N/2}\int\int\int \exp\left\{-\frac{z^2}{2}-\frac{w^{2*}}{2}-q^2+2^{1/2}q(z_1+w_1^*)-\frac{p^2}{2}+p(2^{1/2}z_2+it)-\frac{t^2}{2}+2^{1/2}tw_2^*\right\}d\hat{}q\,d\hat{}p\,d\hat{}t$$

$$= 4^N\int\int \exp\left\{-\frac{z_1^2+z_2^2+w_1^{2*}+w_2^{2*}}{2}+\frac{(z_1+w_1^*)^2}{2}-r^2+2^{1/2}r(2^{1/2}z_2+it)-\frac{t^2}{2}+2^{1/2}tw_2^*\right\}d\hat{}r\,d\hat{}t$$

$$= 2^N\int \exp\left\{-\frac{z_2^2+w_2^{2*}}{2}+z_1w_1^*-\frac{t^2}{2}+2^{1/2}tw_2^*+\frac{(2^{1/2}z_2+it)^2}{2}\right\}d\hat{}t$$

$$= 2^N\int \exp\left\{\frac{z_2^2-w_2^{2*}}{2}+z_1w_1^*-t^2+2^{1/2}t(w_2^*+iz_2)\right\}d\hat{}t$$

$$= \exp\left\{\frac{z_2^2-w_2^{2*}}{2}+z_1w_1^*+\frac{(w_2^*+iz_2)^2}{2}\right\} = \exp(z_1w_1^*+iz_2w_2^*)$$

on using (2) repeatedly.

Step 2: Again with $f\in\mathscr{E}(\mathbb{C}^{2N})$, we find that

$$(A_{2N}RV_{2N}f)(z) = \int L(z,w^*)f(w)d\mu(w),$$

where the kernel L is given by

$$L(z,w^*) := \exp\{2^{-1/2}(z_1w_1^* + z_1w_2^* + z_2w_1^* - z_2w_2^*)\}$$

by a similar calculation.

Step 3: Hence for $f\in\mathscr{E}(\mathbb{C}^{2N})$, we get

$$(A_{2N}WV_{2N}f)(z)$$
$$= (A_{2N}RV_{2N}A_{2N}\Phi V_{2N}f)(z)$$
$$= \int N(z,w^*)f(w)d\mu(w),$$

where the kernel $N(z,w^*)$ is given by

$$N(z,w^*)$$
$$:= \int L(z,t^*)M(t,w^*)d\mu(t)$$
$$= \int\int \exp\{2^{-1/2}[(z_1+z_2)t_1^* + (z_1-z_2)t_2^*)$$
$$+ w_1^*t_1 + iw_2^*t_2\}d\mu(t_1)d\mu(t_2)$$
$$= \exp\{2^{-1/2}[(z_1+z_2)w_1^* + i(z_1-z_2)w_2^*]\}$$

upon using (5) twice.

Step 4: Thus $A_{2N}W^2V_{2N}$ is given by the kernel

$$N^2(z,w^*) := \int N(z,t^*)N(t,w^*)d\mu(t)$$
$$= \exp\{2^{-1}(1+i)[(z_1-iz_2)w_1^*$$
$$+ (z_1+iz_2)w_2^*]\}$$

again by (5).

Finally, $A_{2N}W^3V_{2N}$ is given by the kernel

$$N^3(z,w^*) := \int N^2(z,t^*)N(t,w^*)d\mu(t)$$
$$= \exp\{2^{-3/2}(1+i)(2z_1w_1^* + 2z_2w_2^*)\}$$
$$= \exp(2^{-1/2}(1+i)zw^*)$$

as claimed.

Proof of Corollary 1: From (4) we see that $(A_{2N}W^3V_{2N}f)(z) = f(\exp(i\pi/4)z)$ for $f\in\mathscr{E}(\mathbb{C}_{2N})$, so that

$$(A_{2N}W^6V_{2N}f)(z) = f((\exp(i\pi/4))^2z)$$
$$= f(iz) = (A_{2N}F^{-1}V_{2N}f)(z);$$

hence $W^6 = F^{-1}$ on $\mathscr{S}(\mathbb{R}_{2N})$.

Remark: The Fourier transformation is often replaced in computational problems by its discrete analog, the finite Fourier transform, a matrix in $\mathbb{C}^{n\times n}$ of order 4. Using its action on the basis functions h_{mn} as a guide, one may construct a discrete analog of the Wigner transformation also, whose sixth power is the finite Fourier cotransform. This could provide a useful tool for the analysis of signal processing.

ACKNOWLEDGMENT

We gratefully acknowledge support from the Vicerrectoría de Investigación of the Universidad de Costa Rica.

[1] J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 99 (1949).
[2] F. Bayen, M. Flato, C. Fronsdal, A. Lichnérowicz, and D. Sternheimer, Ann. Phys. (NY) **111**, 61, 111 (1978).
[3] M. Hillery, R. F. O'Connell, M. Scully, and E. P. Wigner, Phys. Rep. **106**, 121 (1984).
[4] J. M. Gracia-Bondía, Phys Rev. A **30**, 691 (1984).
[5] I. E. Segal, Math. Scand. **13**, 31 (1963).
[6] J.-B. Kammerer, J. Math. Phys. **27**, 529 (1986).
[7] L. Hörmander, Commun. Pure Appl. Math. **32**, 359 (1979).
[8] J. M. Gracia-Bondía and J. C. Várilly, "Algebras of distributions suitable for phase-space quantum mechanics, I.," to appear.
[9] E. P. Wigner, Phys. Rev. **40**, 749 (1932).
[10] W. Schempp, Proc. Am. Math. Soc. **92**, 103 (1984).
[11] W. Schempp, "Lie methods in optics," in *Lecture Notes in Physics*, Vol. 250 (Springer, Heidelberg, 1986).
[12] J. C. T. Pool, J. Math. Phys. **7**, 66 (1966).
[13] R. Cressman, J. Func. Anal. **22**, 405 (1976).
[14] M. S. Bartlett and J. E. Moyal, Proc. Cambridge Philos. Soc. **45**, 545 (1949).
[15] V. Bargmann, Commun. Pure Appl. Math. **14**, 187 (1961).
[16] V. Bargmann, Commun. Pure Appl. Math. **20**, 1 (1967).
[17] J. C. Várilly, E. de Faria, and J. M. Gracia-Bondía, Cienc. Tec. (C. R.) **10**, 81 (1986).

Geometric properties of transition amplitude spaces

Sylvia Pulmannová
Mathematics Institute, Slovak Academy of Sciences, 814 73 Bratislava, Czechoslovakia

Stanley Gudder
Department of Mathematics and Computer Science, University of Denver, Denver, Colorado 80208

(Received 16 October 1986; accepted for publication 27 May 1987)

The investigation of transition amplitude spaces (tas's) introduced by Gudder and Pulmannová [J. Math. Phys. **28**, 376 (1987)] is continued. In particular, ordered structures related to tas's are considered. Under some conditions, which are analogous to the conditions obtained for transition probability spaces by Pulmannová [J. Math. Phys. **27**, 1791 (1986)], the ordered structure related to a tas can be represented by the orthocomplemented lattice of all f-closed subspaces of a generalized Hilbert space $(\mathscr{X},\mathscr{D},\theta,f)$. It is shown that, provided that the above representation takes place for a total tas, the division ring $\mathscr{D}$ must be isomorphic with a subfield $\mathbb{C}_1$ of the field of complex numbers $\mathbb{C}$. Sufficient conditions are also given under which the ordered structure of a tas can be represented by the lattice of all closed subspaces of a complex Hilbert space.

I. ORDERED STRUCTURES RELATED TO A tas

Contrary to the traditional Hilbert space formulation of quantum mechanics, it is our opinion that transition amplitudes should play the primary role. This idea is basic to the early work of Feynman and we have begun to develop it into an axiomatic foundation for quantum mechanics.[1] Since the Hilbert space structure is physically unmotivated and is only the result of fairly restrictive *ad hoc* assumptions, this approach has the advantage of placing the foundations of quantum mechanics at a more basic level.

In Ref. 1 we have given some strong physical reasons why a quantum system acts like a Markov process at the amplitude level. The basic property of a transition amplitude then follows from the Chapman–Kolmogorov equation for a Markov process.[1] We have also given a second justification for this property in terms of a transmission amplitude interpretation. Moreover, relationships between the present framework, the algebraic approach, the operational statistics and quantum logic approaches, and traditional Hilbert space quantum mechanics were presented. It was shown in Ref. 1 that a transition amplitude space always admits a Hilbert space representation and that sums and tensor products of such spaces can be formulated in a natural way.

Recently, the concept of a *transition amplitude space* has been introduced.[1] Let us recall basic definitions. Let S be a nonempty set and let $A: S\times S\to\mathbb{C}$. We say that $x,y\in S$ are *orthogonal* ($x\perp y$) if $x\neq y$ and $A(x,y)=0$. Let us denote by $\mathscr{M}_A$ the collection of maximal orthogonal sets in S. We call a set $M\subset S$ an A *set* if for every $x,y\in S$, we have

$$\sum_{z\in M}|A(x,z)\overline{A}(y,z)|<\infty,$$

and

$$A(x,y)=\sum_{z\in M}A(x,z)\overline{A}(y,z),$$

where an overbar denotes the complex conjugate.

Denote the collection of A sets by $\mathscr{N}_A$. We call $A: S\times S\to\mathbb{C}$ a *transition amplitude* if (i) $\mathscr{N}_A\neq\emptyset$, and (ii) $A(x,x)=1$ for all $x\in S$. If A is a transition amplitude on S we call (S,A) a *transition amplitude space* (tas). A *strong* (*ultrastrong*) tas is a tas (S,A) that satisfies (iii) $A(x,y)=1$ [(iv) $|A(x,y)|=1$] implies $x=y$. Although an ultrastrong tas is clearly strong, it is shown in Ref. 1 that the converse need not hold. A tas (S,A) is *total* if $\mathscr{N}_A=\mathscr{M}_A$. To every tas we can associate a strong tas if we introduce a relation $\approx$ by $x\approx y$ if $A(x,y)=1$. Then $\approx$ is an equivalence relation and the function $\mathring{A}: \mathring{S}\times\mathring{S}\to\mathbb{C}$ defined by $\mathring{A}(\mathring{x},\mathring{y})=A(x,y)$ (where $\mathring{S}=S/\approx$) is well defined and $(\mathring{S},\mathring{A})$ is a tas. Similarly, we define a relation $\sim$ on S by $x\sim y$ if $|A(x,y)|=1$. Then $\sim$ is also an equivalence relation. Denote by $\hat{x}$ the class containing x and let $\hat{S}=S/\sim$. If (S,A) is a total tas, then the pair $(\hat{S},T)$, where $T(\hat{x},\hat{y})=|A(x,y)|^2$, is a transition probability space (see Theorem 2.2 in Ref. 1).

Let (S,A) be a tas. We shall use the orthogonality relation on S defined by $x\perp y$ if $A(x,y)=0$ to introduce

$$X^0=\{y\in S: y\perp x \text{ for all } x\in X\},$$

where X is any subset of S. It is easy to check that the map $X\mapsto X^{00}$ has the following properties:

$$X\subset X^{00},\quad X^{00}=(X^{00})^{00},$$

$$X\subset Y \text{ implies } X^{00}\subset Y^{00},$$

i.e., it is a closure operation (see Ref. 2, p. 148). We shall write $\overline{X}=X^{00}$ and denote by $\mathscr{F}(S)$ the set of all "closed" subsets of S, i.e.,

$$\mathscr{F}(S)=\{X\subset S: X=\overline{X}\}.$$

If $X=\{x\}$, we shall write x^0 instead of $\{x\}^0$, and $\bar{x}$ instead of $\{x\}^-$.

Proposition 1.1: Let (S,A) be a tas. Let the following condition be fulfilled:

$$x^0\subset y^0 \text{ implies } x^0=y^0 \text{ for any } x,y\in S. \tag{1}$$

Then $\mathscr{F}(S)$ is a complete, orthocomplemented, atomic lattice with the set of atoms $\{\bar{x}: x\in S\}$. Specifically, if (S,A) is a total tas, then condition (1) is satisfied.

Proof: Let (S,A) be a tas satisfying condition (1). Let $Q\subset S$, $Q\neq\emptyset$, be such that $Q\subset\bar{x}$ for $x\in S$. Then there is $y\in Q$,

 0022-2488/87/102393-07$02.50 © 1987 American Institute of Physics

and $\bar{y}\subset\bar{x}$ implies $x^0\subset y^0$. By (1), $x^0=y^0$, so that $\bar{y}=\bar{x}$. From this we obtain that $\bar{x}=\bar{y}\subset Q\subset\bar{x}$, i.e., $Q=\bar{x}$. The fact that $X\mapsto\overline{X}$ is a closure operation implies that $\mathscr{F}(S)$ is a complete lattice with the operations $\wedge X_i=\cap X_i$ and $\vee X_i=(\cup X_i)^-$ ($X_i\subset S$, $i\in I$). As the orthogonality relation $\perp$ is symmetric and antireflexive, it is easy to show that $X\mapsto X^0$ is an orthocomplementation on $\mathscr{F}(S)$. We also have $S^0=\varnothing$ and $\varnothing^0=S$.

If (S,A) is total, then for every $x\in S$ there is $M\in\mathscr{N}_A$ such that $x\in M$. Let $y^0\subset x^0$ and let $y\in M$, $M\in\mathscr{N}_A$. Then $M-y\subset y^0\subset x^0$. Therefore

$$1=A(x,x)=\sum_{z\in M}|A(x,z)|^2=|A(x,y)|^2.$$

By Corollary 3.3(a) in Ref. 1, it is implied that $A(y,z)=A(y,z)A(x,z)$ for all $x\in S$. From this we obtain that $x^0=y^0$. □

Let $x\simeq y$ if $x^0\subset y^0$ ($x,y\in S$). If (1) is satisfied, then $\simeq$ is an equivalence relation. Let $\tilde{x}$ be the class containing x ($x\in S$). It is easy to see that $\tilde{x}=\bar{x}$. Let $\bar{x}\perp\bar{y}$ if $x\perp y$ for any representative $x\in\bar{x}$, $y\in\bar{y}$. It is straightforward that this relation is well defined. If (S,A) is total, then $x\simeq y$ if and only if $|A(x,y)|=1$.

Recall that $E\subset S$ is an *event* if $E\subset M$ for some $M\in\mathscr{N}_A$. For an event E, let A_E be the *A-transition amplitude conditioned by E*, i.e., A_E is defined by

$$A_E(x,y)=\sum_{z\in E}A(x,z)A(z,y).$$

Proposition 1.2: Let (S,A) be a tas. If E is an event, then

$$\overline{E}=\{x\in S: A_E(x,x)=1\}.$$

Proof: Let $x\in\overline{E}$. Then $x\perp M-E$, where $E\subset M$ and $M\in\mathscr{N}_A$. Therefore

$$1=A(x,x)=\sum_{z\in E}|A(x,z)|^2+\sum_{z\in M-E}|A(x,z)|^2$$

$$=\sum_{z\in E}|A(x,z)|^2=A_E(x,x).$$

Now let $A_E(x,x)=1$. Then

$$A_E(x,x)=\sum_{z\in E}|A(x,z)|^2=1$$

implies that

$$\sum_{z\in M-E}|A(x,z)|^2=0,$$

i.e., $A(x,z)=0$ for all $z\in M-E$. Therefore for any $y\in E^0$ we get

$$A(x,y)=\sum_{z\in M}A(x,z)A(z,y)$$

$$=\sum_{z\in E}A(x,z)A(z,y)+\sum_{z\in M-E}A(x,z)A(z,y)=0.$$

This shows that $x\in\overline{E}$. □

Proposition 1.3: Let E,F be events. Then (i) $A_EA_F=A_E$ if and only if $\overline{E}\subset\overline{F}$ (cf. Ref. 1); and (ii) $A-A_E=A_{M-E}$, where $E\subset M$ and $M\in\mathscr{N}_A$.

Proof: (i) Let $A_EA_F=A_E$ and let $x\in\overline{E}$. For $M\in\mathscr{N}_A$ such that $E\subset M$, we obtain

$$A_E(x,y)=\sum_{z\in E}A(x,z)A(z,y)=\sum_{z\in M}A(x,z)A(z,y)=A(x,y)$$

and

$$1=A_E(x,x)=A_EA_F(x,x)=\sum_{z\in M}A_E(z,x)A_F(x,z)$$

$$=\sum_{z\in M}A(z,x)A_F(x,z)=A_F(x,x).$$

The last equality holds by (4.3) in Ref. 1. Hence $x\in\overline{F}$.

Now let $\overline{E}\subset\overline{F}$. Then $F^0\subset E^0$ and hence

$$A_EA_F(x,y)$$

$$=\sum_{z\in M}A_E(z,y)A_F(x,z)$$

$$=\sum_{z\in E}A_E(z,y)A_F(x,z)+\sum_{z\in M-F}A_E(x,y)A_F(x,z)$$

$$=\sum_{z\in E}A(z,y)A_F(x,z)$$

$$=\sum_{z\in E}A(z,y)\sum_{z\in F}A(x,z')A(z',z)$$

$$=\sum_{z\in E}A(z,y)\sum_{z'\in M'}A(x,z')A(z',z)$$

$$=\sum_{z\in E}A(z,y)A(x,z)=A_E(x,y).$$

This proves that $A_EA_F=A_E$.

The proof of the second statement is straightforward. □

Corollary 1.4: Let (S,A) be a total tas. Then the map $A_E\mapsto\overline{E}$ is an orthoisomorphism between the atomic, σ-orthocomplete orthomodular posets $\mathscr{E}=\{A_E: E \text{ is an event}\}$ and $\mathscr{O}(S)=\{\overline{E}: E \text{ is an event}\}$. (See Refs. 1 and 3.)

In the next proposition we show a relation between the sets $\mathscr{F}(S)$ and $\mathscr{O}(S)$ for a total tas. To introduce it, we need some definitions. Let F be a partially ordered set. For a subset G of F set

$$G^\nabla=\{a\in F: a\geqslant b \text{ for all } b\in G\},$$

$$G^\Delta=\{a\in F: a\leqslant b \text{ for all } b\in G\}.$$

The map $G\mapsto G^{\nabla\Delta}$ is a closure operation and the set of all subsets G of F such that $G=G^{\nabla\Delta}$ is called a *completion by cuts* of the set F (see Ref. 2, p. 167).

Proposition 1.5: If (S,A) is a total tas, then $\mathscr{F}(S)$ is a completion by cuts of $\mathscr{O}(S)$.

Proof: $\mathscr{O}(S)$ is an orthocomplemented partially ordered set and the set $\hat{S}$ of all atoms is join dense in $\mathscr{O}(S)$. For a subset $B\subset\mathscr{O}(S)$ let

$$B^\perp=\{a\in\mathscr{O}(S): b\perp a \text{ for all } b\in B\},$$

where $b\perp a$, $a,b\in\mathscr{O}(S)$ if $a\subset b^0$. Clearly, for $x,y\in S$ we have $\hat{x}\perp\hat{y}$ if and only if $T(\hat{x},\hat{y})=|A(x,y)|^2=0$. A subset B of $\mathscr{O}(S)$ is closed if $B^{\perp\perp}=B$. By Theorem 2.5 in Ref. 4, the set of all closed subsets of $\mathscr{O}(S)$ is orthoisomorphic with the set of all closed subsets of $\hat{S}$, but the latter is orthoisomorphic with $\mathscr{F}(S)$. It is easy to see that for any $B\subset\mathscr{O}(S)$ the equality $B^{\nabla\Delta}=B^{\perp\perp}$ is satisfied, so that $\mathscr{F}(S)$ is the completion by cuts of $\mathscr{O}(S)$. □

It is easy to see that if E is an event, then $(\overline{E},A_E)$ is a tas. It is a strong (ultrastrong) tas if (S,A) is strong (ultra-

strong). If (S,A) is total, then $(\overline{E},A_E)$ is total, too. Indeed, let F be a maximal orthogonal set in $\overline{E}$. Then $F\subset\overline{E}$, $\overline{E},\overline{F}\in\mathcal{O}(S)$, and the orthomodularity of $\mathcal{O}(S)$ implies that $\overline{F}=\overline{E}$. By Proposition 1.3 then $A_E=A_F$, i.e., $F\in\mathcal{N}_{A_E}$.

The following definitions are analogous to those introduced in Ref. 3 for transition probability spaces.

Let (S,A) be a tas. We say that an element $x\in S$ is a *superposition* of a subset F of S if $A(z,y)=0$ for all $y\in F$ and for $z\in S$ implies $A(z,x)=0$. It is easy to see that x is a superposition of F if and only if $x\in F^{00}=\overline{F}$.

An element $x\in S$ is called a *minimal superposition* of F $(F\subset S)$ if x is a superposition of F, but x is not a superposition of any proper subset of F.

We say that a *minimal superposition postulate* (MSP) holds for (S,A) if for any finite subset $F\subset S$ and any minimal superposition x of F there holds

$$\{x,F_1\}^-\cap\overline{F}_2\neq\emptyset,$$

where $\{F_1,F_2\}$ is any partition of the set F (i.e., F_1 and F_2 are nonempty disjoint subsets of F such that $F_1\cup F_2=F$).

We say that a *superposition principle* holds for a tas (S,A) if for any $x,y\in S$ such that $x\notin\overline{y}$, $y\notin\overline{x}$, there is $z\in S$ such that $z\notin\overline{x}$, $z\notin\overline{y}$, and $z\in\{x,y\}^-$ (in other words, z is a minimal superposition of x and y).

A physical motivation of the above notions can be found in Refs. 3 and 5. An advantage of this approach is that superpositions can be defined without assuming any underlying linear structure. In this way the properties of superpositions follow directly from those of the transition amplitude. If a state x is a superposition of a subset F, then x is orthogonal to every state that is orthogonal to F. In a certain sense, this means that as far as its transition amplitudes are concerned, x is determined by the elements of F. For example, if F is a subset of an A set, it follows that for every $y\in S$ we have

$$A(x,y)=\sum_{z\in F}A(x,z)A(z,y).$$

Recall that a tas (S,A) is a *direct sum* of two tas's (S_1,A_1) and (S_2,A_2) if $S_1\cap S_2=\emptyset$, $S=S_1\cup S_2$, and A: $S\times S\to\mathbb{C}$ is defined by

$$A(x,y)=\begin{cases}A_i(x,y), & \text{if } x,y\in S_i,\ i=1,2,\\ 0, & \text{otherwise.}\end{cases}$$

If the superpositon principle holds for a tas (S,A) then (S,A) cannot be isomorphic to a direct sum of two tas's. Indeed, let the superpositon principle hold and let (S,A) be isomorphic with a direct sum $(S_1\oplus S_2,A_1\oplus A_2)$ of the TAS's (S_1,A_1) and (S_2,A_2). Without any loss of generality, we may assume that $(S,A)=(S_1\oplus S_2,A_1\oplus A_2)$. If $x\in S_1$, then $S_2\subset x^0$ implies $\overline{x}\subset S_2^0$. This implies that $\overline{x}\cap S_2=\emptyset$, i.e., $\overline{x}\subset S_1$. Similarly, if $y\in S_2$, then $\overline{y}\subset S_2$. Now let $x\in S_1$ and $y\in S_2$. Clearly, $x\notin\overline{y}$ and $y\notin\overline{x}$. Let z be a minimal superposition of x and y. We have $z\in S_1\cup S_2$. Suppose that $z\in S_1$. Let $u\in x^0$. If $u\in S_1$, then $u\in y^0$, i.e., we have $A(x,u)=A(y,u)=0$. As z is a superposition of $\{x,y\}$, we obtain that $A(z,u)=0$, so that $u\in z^0$. If $u\in S_2$, then we get again that $A(z,u)=0$. Therefore $x^0\subset z^0$, i.e., $z\in\overline{x}$, which contradicts the supposition that z is a minimal superposition of x and y.

The proof of the following theorem can be obtained by essentially the same manner as has been used in Ref. 3.

Theorem 1.6: Let (S,A) be a tas with dimension of at least 4 which satisfies condition (1). Let the superposition principle and postulate of minimal superposition hold. Then there is a division ring $\mathcal{D}$ with an involutive antiautomorphism θ: $\mathcal{D}\to\mathcal{D}$, and a vector space $\mathcal{V}$ over $\mathcal{D}$ with a Hermitian form f: $\mathcal{V}\times\mathcal{V}\to\mathcal{D}$, such that the set $\mathcal{F}(S)$ is orthoisomorphic with the set $\mathcal{L}_f(\mathcal{V})$ of all f-closed subspaces of $\mathcal{V}$. (For the details concerning Theorem 1.6 see Ref. 3.) We recall that a subspace $\mathcal{M}$ of $\mathcal{V}$ is *f closed* if $\mathcal{M}^{00}=\mathcal{M}$, where $\mathcal{M}^0=\{u\in\mathcal{V}: f(u,v)=0 \text{ for all } v\in\mathcal{M}\}$.

To conclude this section, we show some relations between the elements of $\mathcal{F}(S)$ and a representation of a tas. Let H be a complex Hilbert space. We say that a map ϕ: $S\to H$ is a *representation* of a tas (S,A), if $A(x,y)=\langle\phi(x),\phi(y)\rangle$ for all $x,y\in S$ and $\phi(M)\in\mathcal{N}_H$ for some $M\in\mathcal{N}_A$ (see Ref. 1). Every tas admits a representation (see Ref. 1, Theorem 3.2). A tas (S,A) is strong if and only if all its representations are injective. If ϕ: $S\to H$ is a representation, then for any set $M\in\mathcal{N}_A$, $\phi(M)\in\mathcal{N}_H$ [i.e., $\phi(M)$ is a base for H].

Let (S,A) be a tas and let ϕ: $S\to H$ be a representation. For any subset X of S we have

$$\begin{aligned}\phi(X^0)&=\phi\{y\in S: A(x,y)=0 \text{ for every } x\in x\}\\ &=\phi\{y\in S: \langle\phi(x),\phi(y)\rangle=0 \text{ for every } x\in X\}\\ &=\phi(x)^\perp\cap\phi(S),\end{aligned}$$

where $\phi(X)^\perp$ is the orthogonal complement of $\phi(X)$ in H. Therefore $\phi(\overline{X})=(\phi(X)^\perp\cap\phi(S))^\perp\cap\phi(S)$. From $\phi(X)^\perp\cap\phi(S)\subset\phi(X)^\perp$ we obtain $\phi(\overline{X})\supset\phi(X)^{\perp\perp}\cap\phi(S)$.

Proposition 1.7: Let (S,A) be a tas and let ϕ: $S\to H$ be a representation. Then for any event $E\subset S$,

$$\phi(\overline{E})=\phi(E)^{\perp\perp}\cap\phi(S).$$

Proof: Let $E\subset M$, $M\in\mathcal{N}_A$. Then $\phi(M)\in\mathcal{N}_H$, i.e., any $f\in H$ has the form

$$f=\sum_{z\in M}\langle f,\phi(z)\rangle\phi(z).$$

Now let $f\in\phi(E)^\perp$. Then

$$f=\sum_{z\in M-E}\langle f,\phi(z)\rangle\phi(z).$$

If $y\in\overline{E}$, then $y\perp z$ for all $z\in M-E$, so that for any $f\in\phi(E)^\perp$,

$$\langle\phi(y),f\rangle=\sum_{z\in M-E}\langle f,\phi(z)\rangle\langle\phi(y),\phi(z)\rangle=0.$$

Hence $\phi(y)\in\phi(E)^{\perp\perp}\cap\phi(S)$. □

Corollary 1.8: If $E=\{z_1,z_2,\ldots,z_n\}$ is a finite event of a tas (S,A), then $x\in\overline{E}$ if and only if there are complex numbers $c_1,c_2,\ldots,c_n$ such that $\{(-1,x),(c_1,z_1),\ldots,(c_n,z_n)\}\in n_A$.

Proof: We have previously shown that the subset $\{(c_1,z_1),\ldots,(c_n,z_n)\}$ of $\mathbb{C}\times S$ belongs to n_A if $\sum_{i=1}^n c_iA(z_i,x)=0$ for all $x\in S$ (see Ref. 1). Now the closed subspace of H generated by the elements $\phi(z_1),\phi(z_2),\ldots,\phi(z_n)$ is just the set of all linear combinations of these elements. If $x\in\overline{E}$, then Proposition 1.7 implies that $\phi(x)\in\phi(E)^{\perp\perp}$, but this implies that

$$\phi(x)=\sum_{i=1}^n c_i\phi(z_i)$$

for some $c_1,c_2,\ldots,c_n\in\mathbb{C}$. Then for every $y\in S$ we obtain

$$A(x,y) - \sum_{i=1}^{n} c_i A(z_i,y)$$
$$= \langle \phi(x),\phi(y)\rangle - \sum_{i=1}^{n} c_i \langle \phi(z_i),\phi(y)\rangle$$
$$= \left\langle \phi(x) - \sum_{i=1}^{n} c_i \phi(z_i),\phi(y)\right\rangle = 0,$$

i.e.,

$$\{(-1,x),(c_1,z_1),\ldots,(c_n,z_n)\}\in n_A. \qquad \square$$

II. DIVISION RINGS RELATED TO A tas

In this section we investigate the division rings which can be related to tas's by Theorem 1.6. We shall show that only the subrings of the field of complex numbers can be realized.

Theorem 2.1: Let $\mathscr{D}$ and $\mathscr{D}'$ be division rings and $\mathscr{V}$ and $\mathscr{V}'$ be vector spaces over $\mathscr{D}$ and $\mathscr{D}'$, respectively. Let the dimension of $\mathscr{V}$ be n $(n\geqslant 3)$. Let $\mathscr{L}(\mathscr{V},\mathscr{D})$ and $\mathscr{L}(\mathscr{V}',\mathscr{D}')$ be the lattices of all linear subspaces of $\mathscr{V}$ and $\mathscr{V}'$, respectively, and let there be an injective lattice morphism ξ: $\mathscr{L}(\mathscr{V},\mathscr{D})\to\mathscr{L}(\mathscr{V}',\mathscr{D}')$ which maps atoms to atoms. Then there is a subdivision ring $\mathscr{D}_1'$ of $\mathscr{D}'$ such that $\mathscr{D}$ and $\mathscr{D}_1'$ are isomorphic.

Proof: Let $x\in\mathscr{V}$. We write

$$\mathscr{D}\cdot x = \{dx: d\in\mathscr{D}\}, \tag{2}$$

and $\mathscr{D}'\cdot x'$ is defined similarly for $x'\in\mathscr{V}'$. If $c\in\mathscr{D}$, and $c\neq 0$, then $\mathscr{D}\cdot x = \mathscr{D}\cdot(cx)$. Suppose $x\in\mathscr{V}, x'\in\mathscr{V}'$. We write $x\sim x'$ when and only when $x\neq 0$, $x'\neq 0$, $\xi(\mathscr{D}\cdot x) = \mathscr{D}'\cdot x'$ (as ξ maps atoms to atoms, to any $x\in\mathscr{V}$, $x\neq 0$ there is $x'\in\mathscr{V}'$ such that $x\sim x'$). To continue the proof, we need a lemma.

Lemma 2.2: Let $x\in\mathscr{V}$, $x'\in\mathscr{V}'$, and let $x\sim x'$. Then, for any $y\in\mathscr{V}$ with $y\neq 0$ and $\mathscr{D}\cdot y\neq\mathscr{D}\cdot x$, there exists a unique $y'\in\mathscr{V}'$ such that

$$y\sim y' \quad\text{and}\quad x-y\sim x'-y'. \tag{3}$$

Proof: Let y'' be some nonzero vector in $\xi(\mathscr{D}\cdot y)$. Since ξ is a lattice morphism that maps atoms to atoms, it follows that for some $a,b\in\mathscr{D}'$,

$$\xi(\mathscr{D}\cdot(x-y)) = \mathscr{D}'\cdot(ax'+by'').$$

Since $\mathscr{D}\cdot y\neq\mathscr{D}\cdot x$, $\mathscr{D}\cdot(x-y)$ is distinct from both $\mathscr{D}\cdot x$ and $\mathscr{D}\cdot y$, and since ξ is injective, $a\neq 0$, $b\neq 0$. Define y' by

$$y' = -(a^{-1}b)y''.$$

Then $y'\neq 0$, $y\sim y'$, and $x-y\sim x'-y'$.

To prove the uniqueness of y', suppose that $z'\neq 0$ is in $\mathscr{V}'$ such that $y\sim z'$, $x-y\sim x'-z'$. Then for $a,b\in\mathscr{D}'$, which are both nonzero,

$$z' = ay', \quad x'-z' = b(x'-y'),$$

from which it follows that

$$x' = bx' + (a-b)y'.$$

Since $\mathscr{D}'\cdot x'$ and $\mathscr{D}'\cdot y'$ are distinct, this implies that $b=1$ and $a-b=0$, so that $a=1$. Hence $y'=z'$ and this finishes the proof of Lemma 2.2. $\square$

Let $x\in\mathscr{V}$, $x'\in\mathscr{V}'$, and let $x\sim x'$. We define the mapping $T_{x,x'}$ by

$$T_{x,x'}(0)=0, \quad T_{x,x'}(y)=y' \quad \text{if } y\in\mathscr{V},\ y\neq 0,\ \mathscr{D}\cdot x\neq\mathscr{D}\cdot y, \tag{4}$$

where $y'\in\mathscr{V}'$ satisfies (3). Note that $T_{x,x'}$ is not defined for a nonzero y unless $\mathscr{D}\cdot y\neq\mathscr{D}\cdot x$.

Since the dimension of $\mathscr{V}$ is at least 3, we can choose three vectors u_1,u_2,u_3 of $\mathscr{V}$ that are independent. Let $u_1'\in\mathscr{V}'$ be such that $u_1\sim u_1'$ and let $u_2',u_3'\in\mathscr{V}'$ be defined by

$$u_2' = T_{u_1,u_1'}(u_2), \quad u_3' = T_{u_1,u_1'}(u_3).$$

It can be shown that the mappings $T_{x,x'}$ have all the properties proved in Ref. 5 (Lemmas 3.4–3.8). As the proofs are literally the same; we shall not repeat them. We note that all that one needs to prove these lemmas is the fact that x',y',z' are independent provided x,y,z are independent. This is guaranteed by the properties of ξ.

We can now define a mapping L of $\mathscr{V}$ into $\mathscr{V}'$. We set

$$L\,0 = 0. \tag{5}$$

If $x\neq 0$ there will exist an integer i such that $1\leqslant i\leqslant 3$ and $\mathscr{D}\cdot x\neq\mathscr{D}\cdot u_i$. We then set

$$Lx = T_{u_i,u_i'}(x). \tag{5'}$$

Again, similarly as in Ref. 5, we prove that L is well defined and $Lu_i = u_i'$ for $i=1,2,3$. Also, L is additive, i.e., if y and z are vectors of $\mathscr{V}$, then

$$L(y+z) = Ly + Lz. \tag{6}$$

(See Lemmas 3.9 and 3.10 in Ref. 5.)

We now examine the properties of L relative to scalar multiplication. Similarly as in Lemma 3.12 in Ref. 5, we prove that

$$L(cx) = c^\sigma Lx$$

for every $x\in\mathscr{V}$. Then $\sigma(c\mapsto c^\sigma)$ is a well-defined map of $\mathscr{D}$ into $\mathscr{D}'$.

Lemma 2.3: Let $\sigma(\mathscr{D})$ be the image of $\mathscr{D}$ under σ. Then $\sigma(\mathscr{D})$ is a division ring and $\sigma(c\to c^\sigma)$ is an isomorphism of $\mathscr{D}$ onto $\sigma(\mathscr{D})$.

Proof: First we show that σ: $\mathscr{D}\to\mathscr{D}'$ is a morphism of $\mathscr{D}$ into $\mathscr{D}'$. Let $x\neq 0$ be in $\mathscr{V}$. Let $c_1,c_2\in\mathscr{D}$. Then $Lx\neq 0$ and

$$(c_1+c_2)^\sigma Lx = L((c_1+c_2)x) = L(c_1x+c_2x)$$
$$= L(c_1x)+L(c_2x) = (c_1^\sigma + c_2^\sigma)Lx,$$

so that

$$(c_1+c_2)^\sigma = c_1^\sigma + c_2^\sigma.$$

Further

$$(c_1c_2)^\sigma Lx = L(c_1(c_2x)) = c_1^\sigma L(c_2x) = c_1^\sigma c_2^\sigma Lx,$$

so that

$$(c_1c_2)^\sigma = c_1^\sigma c_2^\sigma.$$

Now if $c_1^\sigma = c_2^\sigma$, then for $x\neq 0$ in $\mathscr{V}$, $L(c_1x) = L(c_2x)$, i.e., $L(c_1x-c_2x)=0$. But $Ly=0$ if and only if $y=0$. Therefore $c_1-c_2=0$, i.e., $c_1=c_2$. This proves that $\sigma(\mathscr{D})$ is an isomorphic image of $\mathscr{D}$, and consequently $\sigma(\mathscr{D})$ is a division ring. The proof of Lemma 2.3 is finished. $\square$

To conclude the proof of Theorem 2.1, it is enough to set $\mathscr{D}_1' = \sigma(\mathscr{D})$. $\square$

In what follows, we shall use Theorem 2.1 to prove that the only division rings which can be related to a total tas

in the sense of Theorem 1.6 are the subdivision rings of the field of complex numbers.

Theorem 2.4: Let (S,A) be a strong total tas with dimension of at least 4. Let there be a division ring $\mathscr{D}$ with an involutive antiautomorphism $\theta: \mathscr{D} \to \mathscr{D}$, a vector space $\mathscr{W}$ over $\mathscr{D}$, and a Hermitian form $f: \mathscr{W} \times \mathscr{W} \to \mathscr{D}$ such that the set $\mathscr{L}_f(\mathscr{W})$ of all f-closed subspaces of $\mathscr{W}$ is orthoisomorphic with the set $\mathscr{F}(S)$. Then there is a subfield $\mathbb{C}_1$ of the field of complex numbers $\mathbb{C}$ such that $\mathscr{D}$ and $\mathbb{C}_1$ are isomorphic.

Proof: Let $\mathscr{V}$ be a subspace of $\mathscr{W}$ such that $\dim \mathscr{V} = n$ $(n \geqslant 3)$. Let $S_1 \in \mathscr{F}(S)$ correspond to $\mathscr{V}$. Since S_1 is finite dimensional, there is an orthogonal set $E \subset S$ such that $S_1 = \overline{E}$ (see Ref. 2). Since (S,A) is total, E is an event. Then (S_1, A_E) is a strong, total tas. Without any loss of generality, we may assume that $(S,A) = (S_1, A_E)$, and that $\mathscr{F}(S)$ is orthoisomorphic with the orthocomplemented lattice $\mathscr{L}(\mathscr{V})$ of all linear subspaces of $\mathscr{V}$. Let $\phi: S \to H$ be any representation. For a linear subspace $\mathscr{M} \subset \mathscr{V}$ set

$$\xi(\mathscr{M}) = \phi(X)^{\perp\perp}, \tag{7}$$

where $X \in \mathscr{F}(S)$ corresponds to M. We shall show that ξ is an injective lattice morphism from the subspaces of $\mathscr{V}$ into the subspaces of H, which maps atoms to atoms. Then, when we set $\mathscr{D}' = \mathbb{C}$ and $\mathscr{V}' = H$ in Theorem 2.1, the proof of Theorem 2.4 will follow.

Let $x \in S$. Since (S,A) is total, x is an event. By Proposition 1.7,

$$\phi(\bar{x}) = \phi(x)^{\perp\perp} \cap \phi(S) \subset \phi(x)^{\perp\perp} \subset \phi(\bar{x})^{\perp\perp}.$$

Therefore

$$\phi(\bar{x})^{\perp\perp} = \phi(x)^{\perp\perp} = \mathbb{C} \cdot \phi(x).$$

Now if $v \in \mathscr{V}$, then to the atom $\mathscr{D} \cdot v$ of $\mathscr{L}(\mathscr{V})$ there corresponds an atom $\bar{x}$ of the $\mathscr{F}(S)$, so that

$$\xi(\mathscr{D} \cdot v) = \phi(\bar{x})^{\perp\perp} = \mathbb{C} \cdot \phi(x).$$

This shows that ξ maps atoms to atoms.

Let $\mathscr{M}_1, \mathscr{M}_2$ be linear subspaces of $\mathscr{V}$. Let X_1, X_2 be the elements of $\mathscr{F}(S)$ corresponding to $\mathscr{M}_1, \mathscr{M}_2$, respectively. Evidently,

$$\xi(\mathscr{M}_1) \vee \xi(\mathscr{M}_2) = \phi(X_1)^{\perp\perp} \vee \phi(X_2)^{\perp\perp} \subset \phi(X_1 \vee X_2)^{\perp\perp} = \xi(\mathscr{M}_1 \vee \mathscr{M}_2).$$

Now let $x \in X_1 \vee X_2$, $x \notin X_1$, $x \notin X_2$. As X_1 and X_2 are finite dimensional, there are $x_1 \in X_1$ and $x_2 \in X_2$ such that $x \in \{x_1, x_2\}^-$ (see Ref. 3). Let z_1, z_2 be an orthogonal base for $\{x_1, x_2\}^-$, i.e., $\{z_1, z_2\}^- = \{x_1, x_2\}^-$ and $z_1 \perp z_2$ (such an orthogonal base exists by Ref. 3). By Corollary 1.8 we have

$$\phi(x_1) = a_1\phi(z_1) + a_2\phi(z_2), \quad a_1, a_2 \in \mathbb{C},$$
$$\phi(x_2) = b_1\phi(z_1) + b_2\phi(z_2), \quad b_1, b_2 \in \mathbb{C},$$
$$\phi(x) = c_1\phi(z_1) + c_2\phi(z_2), \quad c_1, c_2 \in \mathbb{C}.$$

As $\phi(x_1)$ and $\phi(x_2)$ are independent vectors in H, we can express $\phi(z_1)$ and $\phi(z_2)$ as linear combinations of $\phi(x_1)$ and $\phi(x_2)$, so that for some $d_1, d_2 \in \mathbb{C}$,

$$\phi(x) = d_1\phi(x_1) + d_2\phi(x_2) \in \phi(X_1)^{\perp\perp} \vee \phi(X_2)^{\perp\perp}.$$

Thus we obtain that

$$\phi(X_1 \vee X_2) \subset \phi(X_1)^{\perp\perp} \vee \phi(X_2)^{\perp\perp},$$

and hence

$$\xi(\mathscr{M}_1 \vee \mathscr{M}_2) = \xi(\mathscr{M}_1) \vee (\mathscr{M}_2).$$

For $\mathscr{M} \in \mathscr{L}(\mathscr{V})$ we have $\mathscr{M}^0 = \{u \in \mathscr{V}: f(u,v) = 0 \text{ for all } v \in \mathscr{M}\}$. If $X \in \mathscr{F}(S)$ corresponds to $\mathscr{M}$, then X^0 corresponds to $\mathscr{M}^0$. Then we obtain

$$\xi(\mathscr{M}^0) = \phi(X^0)^{\perp\perp}.$$

Let B be an orthogonal base for X. As (S,A) is total, there is $M \in \mathscr{N}_A$ such that $B \subset M$. Then $X^0 = B^0 = (M - B)^-$, and

$$\phi(X^0) = \phi(X)^{\perp} \cap \phi(S) \subset \phi(X)^{\perp} \subset \phi(B)^{\perp}.$$

As $M \in \mathscr{N}_A$, $\phi(M) \in \mathscr{N}_H$, so that for every $f \in H$,

$$f = \sum_{z \in B} \langle f, \phi(z) \rangle \phi(z) + \sum_{z \in M - B} \langle f, \phi(z) \rangle \phi(z).$$

Now if $f \perp \phi(z)$ for all $z \in B$, then

$$f = \sum_{z \in M - B} \langle f, \phi(z) \rangle \phi(z),$$

i.e., $f \in \phi(M - B)^{\perp\perp}$. Therefore $\phi(B)^{\perp} \subset \phi(M - B)^{\perp\perp} \subset \phi(B^0)^{\perp\perp} = \phi(X^0)^{\perp\perp}$. From this we obtain that $\phi(X^0)^{\perp\perp} = \phi(X)^{\perp}$, i.e., $\xi(\mathscr{M}^0) = \phi(X)^{\perp} = \xi(\mathscr{M})^{\perp}$. Now by de Morgan's law we obtain that

$$\xi(\mathscr{M}_1 \wedge \mathscr{M}_2) = \xi(\mathscr{M}_1) \wedge \xi(\mathscr{M}_2).$$

This proves that ξ is a lattice morphism.

Let $X = \overline{B}$, where B is an event. By Proposition 1.7 we have

$$\phi(X) = \phi(E)^{\perp\perp} \cap \phi(S) \subset \phi(E)^{\perp\perp},$$

i.e., $\phi(X)^{\perp\perp} \subset \phi(E)^{\perp\perp}$, and hence $\phi(E)^{\perp\perp} = \phi(X)^{\perp\perp}$. Now if $\xi(\mathscr{M}) = \xi(\mathscr{N})$, then $\phi(X)^{\perp\perp} = \phi(Y)^{\perp\perp}$, where X and Y correspond to $\mathscr{M}$ and $\mathscr{N}$, respectively. As every finite-dimensional element of $\mathscr{F}(S)$ belongs to $\mathscr{O}(S)$, we obtain that $\phi(X) = \phi(X)^{\perp\perp} \cap \phi(S) = \phi(Y)^{\perp\perp} \cap \phi(S) = \phi(Y)$. Since ϕ is injective, we get $X = Y$, i.e., $\mathscr{M} = \mathscr{N}$. This proves that ξ is injective. □

Example 1: Let H be a complex Hilbert space and let M $(M \subset H)$ be an orthonormal base. Let us take all finite real linear combinations of the elements of M, and complete this set in H. We obtain a real Hilbert space, which we donote by H^r. Let $S = \{x \in H^r: \|x\| = 1\}$ and $A(x,y) = \langle x,y \rangle$. Then (S,A) is a strong total tas. It is not difficult to check that the set $\mathscr{F}(S)$ is orthoisomorphic with the set of all closed subspaces of H^r. Clearly, $\mathscr{D}$ is the field of real numbers, $\theta: R \to R$ is the identity, and the Hermitian form f is identical with the scalar product in H^r.

Example 2: Let H be a complex Hilbert space and let M be an orthonormal base in H. Let $\mathscr{D} = \{a + bi: a,b \text{ are rational numbers}\}$. Evidently, $\mathscr{D}$ is a division ring and the complex conjugation restricted to $\mathscr{D}$ is an involutive antiautomorphism. Take all finite linear combinations of the elements of M over $\mathscr{D}$ and denote it by $\mathscr{V}$. Let $S = \{x \in \mathscr{V}: x \neq 0\}$ and let $A(x,y) = \langle x,y \rangle / \|x\| \, \|y\|$. Then (A,S) is a tas. If the dimension of S is finite, we can use the orthogonalization method to prove that any maximal orthogonal set in S is at the same time a maximal linearly independent set in $\mathscr{V}$. This implies that (S,A) is total.

The map $\phi: S \to H$ defined by $\phi(x) = x/\|x\|$ is a representation. If we set $f(x,y) = \langle x,y \rangle$, then $f: \mathscr{V} \times \mathscr{V} \to \mathscr{D}$ is a Hermitian form, and $A(x,y) = 0$ if and only if $f(x,y) = 0$.

III. STRONG SUPERPOSITION PRINCIPLE

In this section we shall investigate the conditions under which the set $\mathscr{F}(S)$ of all closed subsets of S, where (S,A) is a tas, is orthoisomorphic with the lattice $\mathscr{L}(H)$ of all closed subspaces of a Hilbert space H. Evidently, we can concentrate our attention to total tas's. In what follows, we shall need the concept of an automorphism of a tas. A map $J: S\to S$ is an *automorphism* if J is a bijection and $A(x,y) = A(Jx,Jy)$ for all $x,y\in S$.

We say that a *strong superposition principle* (SSP) holds in a total tas (S,A) if (i) for every x,y and u,v in S, such that $\bar{x}\neq\bar{y}$ and $\bar{u}\neq\bar{v}$ there is an automorphism $J: S\to S$ such that $\{u,v\}^- = \{Jx,Jy\}^-$; and (ii) there is a representation $\phi: S\to H$ such that for some $x,y\in S$, $\bar{x}\neq\bar{y}$, we have

$$\mathbb{C}\cdot\phi(\{x,y\}^-) = \{\phi(x),\phi(y)\}^{\perp\perp}$$

[i.e., for every $c_1,c_2\in\mathbb{C}$ there exists a $c\in\mathbb{C}$ and $z\in\{x,y\}^-$ such that $c_1\phi(x) + c_2\phi(y) = c\phi(z)$].

Lemma 3.1: Let (S,A) be a tas and let $J: S\to S$ be an automorphism. Then for any $X\subset S$ we have $J(X^0) = J(X)^0$.

Proof: Let $y\in J(X^0)$. Then there is $x\in X^0$ such that $Jx = y$. Since $A(x,z) = 0$ for all $z\in X$, we obtain $A(y,Jz) = 0$ for all $z\in X$, and hence $y\in J(X)^0$. On the other hand, if $y\in J(X)^0$, then $A(y,Jx) = 0$ for every $x\in X$. Let $z\in S$ be such that $y = Jz$. Then $A(Jz,Jx) = 0$ implies that $A(z,x) = 0$ for all $x\in X$, and hence $z\in X^0$. From this we obtain that $y\in J(X^0)$, and together with the first part of the proof this proves that $J(X^0) = J(X)^0$. □

Lemma 3.1 implies that for any $x,y\in S$, $J(\{x,y\}^-) = \{Jx,Jy\}^-$.

Lemma 3.2: If SSP holds for a total tas (S,A) then $\mathbb{C}\cdot\phi(\{u,v\}^-) = \{\phi(u),\phi(v)\}^{\perp\perp}$ for every $u,v\in S$ such that $\bar{u}\neq\bar{v}$.

Proof: Let $x,y\in S$ be such that for them (ii) of SSP holds, and let $\phi: S\to H$ be the corresponding representation. Let $u,v\in S$ be such that $\bar{u}\neq\bar{v}$. By (i) of SSP, there is an automorphism $J: S\to S$ such that $\{u,v\}^- = \{Jx,Jy\}^-$. Then we obtain

$$\phi(\{u,v\}^-) = \phi(\{Jx,Jy\}^-) = \phi(J(\{x,y\}^-)).$$

By Corollary 4.3 in Ref. 1, there exists a unique unitary operator U on H such that $\phi J = U\phi$. Therefore,

$$\mathbb{C}\cdot\phi(J(\{x,y\}^-)) = U(\mathbb{C}\cdot\phi(\{x,y\}^-)) = U(\{\phi(x),\phi(y)\}^{\perp\perp}).$$

We have $\phi(u),\phi(v)\in U(\{\phi(x),\phi(y)\}^{\perp\perp})$, and since $\phi(u),\phi(v)$ are independent [in the opposite case we would have $|\langle\phi(u),\phi(v)\rangle| = |A(u,v)| = 1$, i.e., $\bar{u} = \bar{v}$, a contradiction], we obtain that

$$\{\phi(u),\phi(v)\}^{\perp\perp} = U(\{\phi(x),\phi(y)\}^{\perp\perp}) = \mathbb{C}\cdot\phi(\{u,v\}^-).$$

□

Corollary 3.3: If SSP holds in a total tas (S,A), then the superposition principle holds.

Proof: Let us suppose that the superposition principle does not hold, and let $u,v\in S$ be such that $\bar{u}\neq\bar{v}$ and $\{u,v\}$ does not admit any minimal superposition. By the preceding statement,

$$\mathbb{C}\cdot\phi(\{u,v\}^-) = \{\phi(u),\phi(v)\}^{\perp\perp}.$$

From the properties of H it follows that there is an element g in $\{\phi(u),\phi(v)\}^{\perp\perp}$, which is contained neither in $\phi(u)^{\perp\perp}$ nor in $\phi(v)^{\perp\perp}$. As $g\in\mathbb{C}\cdot\phi(\{u,v\}^-)$, there is an element w in $\{u,v\}^-$ and $c\in\mathbb{C}$ such that $c\phi(w) = g$. It is easy to see that w is a minimal superposition of $\{u,v\}$. □

Since by Corollary 4.2 in Ref. 1 any two representations of a tas are unitarily equivalent, it follows that if SSP holds for a tas (S,A), then the condition (ii) is satisfied for any representation of (S,A).

Theorem 3.4: If a total tas (S,A) satisfies the strong superposition principle, then the set $\mathscr{F}(S)$ is orthoisomorphic with the set $\mathscr{L}(H)$ of all closed subspaces of a Hilbert space H.

Proof: Let $\phi: S\to H$ be any representation. Set $\mathbb{C}\cdot\phi(S) = \{c\phi(x): c\in\mathbb{C}, x\in S\}$. Here SSP implies that $\mathbb{C}\cdot\phi(S)$ is a linear subspace of H. Moreover, $\mathbb{C}\cdot\phi(S)$ endowed by the scalar product inherited from H becomes an inner product space. Since (S,A) is total, every maximal orthonormal set in $\mathbb{C}\cdot\phi(S)$ is a base, which by Ref. 6 implies that $\mathbb{C}\cdot\phi(S)$ is complete, and therefore $\mathbb{C}\cdot\phi(S) = H$.

For $X\in\mathscr{F}(S)$ let

$$\rho(X) = \phi(X)^{\perp\perp}.$$

Since (S,A) is total, we have $\phi(X) = \phi(X)^{\perp\perp}\cap\phi(S)$, and since $\mathbb{C}\cdot\phi(S) = H$, we obtain that

$$\mathbb{C}\cdot\phi(X) = \phi(X)^{\perp\perp} = \rho(X).$$

For $X,Y\in\mathscr{F}(S)$ we have

$$\rho(X\wedge Y) = \mathbb{C}\cdot\phi(X\wedge Y) = \mathbb{C}\cdot\phi(X)\wedge\phi(Y) = \mathbb{C}\cdot\phi(X)\wedge\mathbb{C}\cdot\phi(Y) = \rho(X)\wedge\rho(Y),$$

and

$$\rho(X^0) = \mathbb{C}\cdot\phi(X^0) = \mathbb{C}\cdot\phi(X)^{\perp}\cap\phi(S) = \phi(X)^{\perp} = (\mathbb{C}\cdot\phi(X))^{\perp} = \rho(X)^{\perp}.$$

If $\rho(X) = \rho(Y)$, then $\phi(X)^{\perp\perp} = \phi(Y)^{\perp\perp}$ implies that $\phi(X) = \phi(X)^{\perp\perp}\cap\phi(S) = \phi(Y)^{\perp\perp}\cap\phi(S) = \phi(Y)$. Now $|A(x,y)| = 1$ implies that $\bar{x} = \bar{y}$. Therefore $\phi(X) = \phi(Y)$ implies that $X = Y$. This shows that ρ is an injective orthoisomorphism of $\mathscr{F}(S)$ into $\mathscr{L}(H)$. It remains to show that ρ is onto. For $V\in\mathscr{L}(H)$ let

$$X = \{x\in S: \mathbb{C}\cdot\phi(x)\subset V\}.$$

Since $\mathbb{C}\cdot\phi(S) = H$, we obtain that $\mathbb{C}\cdot\phi(X) = V$. Moreover,

$$X^0 = \{y\in S: y\perp x \text{ for any } y\in X\} = \{y\in S: \phi(y)\perp\phi(x) \text{ for any } x\in X\} = \{y\in S: \phi(y)\perp V\},$$

i.e.,

$$\phi(X^0) = V^{\perp}\cap\phi(S).$$

This implies that

$$\mathbb{C}\cdot\phi(X^0) = V^{\perp}.$$

Now if $x\in\overline{X}$, then $x\perp y$ for all $y\in X^0$, which implies that $\phi(x)\perp\phi(X^0)$. But then $\phi(x)\perp V^{\perp}$, i.e., $\phi(x)\in V$. This shows that $x\in X$, and hence $X\in\mathscr{F}(S)$. This completes the proof. □

We close with some open questions.

(1) Do we obtain, by applying Theorem 2.1, that $(c^{\theta})^{\sigma} = \overline{c^{\sigma}}$, where an overbar denotes complex conjugation? This problem can be reduced to the following question:

Is $\langle Lx,Ly\rangle$ in $\sigma(\mathscr{D})$ for any $x, y \in V$? If yes, then we define $g(x,y) = \langle Lx,Ly\rangle^{\sigma^{-1}}$. We have $f(x,y) = 0$ if and only if $\mathscr{D}\cdot x \perp \mathscr{D}\cdot y$, and this holds if and only if $\xi(\mathscr{D}\cdot x) \perp \xi(\mathscr{D}\cdot y)$, i.e., if and only if $\langle Lx,Ly\rangle = 0$. From this we obtain that $g(x,y) = 0$ if and only if $f(x,y) = 0$, and g is a Hermitian form with respect to $\overline{\theta}$ defined by $c^{\overline{\theta}} = (\overline{c^{\sigma}})^{\sigma^{-1}}$. We then obtain, from the von Neumann and Birkhoff theorem, that $\overline{\theta} = \theta$ and $f = g$, and hence $(c^{\theta})^{\sigma} = \overline{c^{\sigma}}$.

(2) Is the tas in example 2 total if its dimension is infinite?

[1]S. Gudder and S. Pulmannová, "Transition amplitude spaces," J. Math. Phys. **28**, 376 (1987).

[2]G. Birkhoff, *Lattice Theory* (Am. Math. Soc., Providence, RI, 1967) [in Russian: *Teorija Rešetok* (Nauka, Moscow, 1984)].

[3]S. Pulmannová, "Transition probability spaces," J. Math. Phys. **27**, 1791 (1986).

[4]M. MacLaren, "Atomic orthocomplemented lattices," Pac. J. Math. **14**, 597 (1964).

[5]V. Varadarajan, *Geometry of Quantum Theory I.* (Van Nostrand, Princeton, 1968).

[6]S. Gudder, "Inner product spaces," Am. Math. Mon. **81**, 29 (1974); **82**, 818 (E) (1975).

In general, the less degeneracy the less transition. A principle for time-dependent Hamiltonian systems in quantum mechanics

H. Gingold
119 Eiesland Hall, Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506

(Received 13 May 1986; accepted for publication 17 June 1987)

A principle in quantum mechanics is proposed: "In general, the less degeneracy the less transition." Mathematical support of this principle is given in a setting of a slowly varying time-dependent Schrödinger equation via a theorem of asymptotic decomposition. Formulas that quantitatively relate transition and degeneracy are developed. Ramifications of those formulas are discussed.

I. INTRODUCTION

In this paper we discuss the following question. To what extent does a general setting of quantum mechanics support the following principle: "In general, the less degeneracy the less transition?" Much of our effort revolves about supporting the above somewhat vague statement by an appropriate, concise mathematical analysis. We will derive an innocent mathematical formula that will produce, among other things, the principle mentioned above. It will be shown in the sequel that our analysis could be associated with the Jahn–Teller effect,[1] with the adiabatic approximation theorem in quantum mechanics (see Born and Fock[2]), and with a certain admissibility criterion of self-adjoint time-dependent Hamiltonian systems.

The basic ideas are as follows. We will consider a setting of time-dependent slowly varying Hamiltonian systems that evolve in time according to Schrödinger's equation. The qualitative phenomenon of degeneracy will be associated quantitatively with the "amount" of degeneracy present in our system. Using a method of asymptotic decomposition proposed by Gingold[3] and developed by Gingold and Hsieh,[4,5] we will derive an asymptotic formula for the transition probabilities of evolving states. The "asymptotic size" of the transition probabilities as a function of the amount of degeneracy present in our evolving system will be indicated by some mathematical formulas. The asymptotic size of the transition probabilities will indicate for us the "amount of transition" present in a Hamiltonian system. Their interpretation will lead us to the desired conclusions.

The setting is as follows. Consider the evolution of the system

$$i\hbar y' = H(\tilde{\epsilon}t)y, \quad i=\sqrt{-1}, \quad '=\frac{d}{dt}, \tag{1.1}$$

where $\hbar$ is Planck's normalized constant. We make the following assumption.

Assumption 1.1: $\tilde{\epsilon}$ is a positive smallness parameter, $0<\tilde{\epsilon}<\infty$. Let

$$\tau=\tilde{\epsilon}t, \quad H(\tau):=H(\tilde{\epsilon}t). \tag{1.2}$$

Here $H(\tau)$ is an $n\times n$ Hermitian analytic matrix function on the closed interval $0\leqslant\tau\leqslant\infty$.

Notice that for each finite time t, $0\leqslant t<\infty$ we have $\lim_{\tilde{\epsilon}\to 0^+} H(\tilde{\epsilon}t)=H(0)$. However, we allow $H(0)\neq H(\infty)$. It is in this sense that our system is "slowly varying." Evidently, the Hamiltonian system (1.1) is equivalent to the system

$$i\epsilon y' = H(\tau)y, \quad 0<\epsilon=\tilde{\epsilon}\hbar, \quad 0\leqslant\tau\leqslant\infty, \quad '=\frac{d}{d\tau}. \tag{1.3}$$

Notice that $\epsilon\to 0^+$ if $\tilde{\epsilon}\to 0^+$ or if $\hbar\to 0^+$.

An immediate consequence of our assumption is Rellich's theorem.[6]

Rellich's Theorem 1.2: Let Assumption 1.1 hold. Then $H(\tau)$ possesses n orthonormal analytic eigenvectors $u_1(\tau),\ldots,u_n(\tau)$ on $[0,\infty]$ that correspond to n real analytic eigenvalues (energy levels) $E_1(\tau),\ldots,E_n(\tau)$.

Thus the unitary transformation $U(\tau)=[u_1(\tau),\ldots,u_n(\tau)]$, where $u_1(\tau),\ldots,u_n(\tau)$ are column vectors, satisfies

$$\begin{aligned} &H(\tau)=U(\tau)E(\tau)U^*(\tau),\\ &E(\tau)=\operatorname{diag}[E_1(\tau),\ldots,E_n(\tau)], \quad UU^*=I, \end{aligned} \tag{1.4}$$

where I is the $n\times n$ identity matrix.

The state that evolves will be a column vector solution to (1.3). Specifically we will be concerned with the set of states that evolve from the initial eigenstates.

Apology: We intend to study Hamiltonians with multi-degenerate energy levels. This is in spite of certain works like Hund[7] and Von Neumann and Wigner[8] which argue in a mathematical manner that level crossing or degeneracies of energy levels is an exceptional phenomenon. Let us point out the possible benefits of studying systems with degeneracies. First, if indeed degeneracies are exceptional we expect the exceptional to illuminate the common systems without degeneracies. Second, transition probabilities for systems with close energy levels (even if noncrossing) can be better understood mathematically by assuming a limiting situation of degeneracies. Third, symmetry plays an important role in physics. Certain symmetries show up as degeneracies in certain configurations. Fourth, there is an interest in symmetry and degeneracy by physicists and chemists. The Jahn–Teller work[1] is just one article on this subject area. For more details one can consult Knox and Gold,[9] Nikitin,[10] and Pearson.[11]

The order of events in this article is as follows. In Sec. II we measure quantitatively the amount of degeneracy present in a Hamiltonian system. In Sec. III we classify Hamiltonian systems according to the type of degeneracy present and we

 0022-2488/87/102400-07$02.50 © 1987 American Institute of Physics

mention the availability of asymptotic decompositions. In Sec. IV we provide an asymptotic decomposition theorem which is a refinement of a theorem of Gingold and Hsieh.[4] In Sec. V we derive our principle from certain mathematical relations. In Sec. VI we elaborate on a key mathematical relation and its ramifications.

II. MEASURING DEGENERACIES AT EACH INSTANT

In order to be able to relate transition and degeneracy we need to measure the degeneracies in our system (1.3) and to obtain an asymptotic decomposition for its solutions. To this end we introduce certain indices whose significance will become clear in the sequel.

First we introduce an index $\tilde{d}$ that will be associated with a pair of energy levels $\langle E_j(\tau), E_k(\tau)\rangle$, $j,k = 1,...,n$, $j\neq k$, at each ("scaled") instant $\tilde{\tau}$ of the interval $[0,\infty]$. These indices will help us measure the amount of degeneracy present in a Hamiltonian system (1.3).

The index $\tilde{d}$ is defined as the order of level crossing in the following precise way.

Convention 2.1: Let $\langle j,k\rangle$ be a fixed ordered set of indices $j,k = 1,...,n$ with $j\neq k$. We say that $\tilde{d} = \tilde{d}(j,k,\tilde{\tau})$ is the order of the (turning point) level crossing at $\tilde{\tau}$ for $\langle j,k\rangle$ if in a neighborhood of a finite time $\tilde{\tau}$,

$$E_j(\tau) - E_k(\tau) = (\tau - \tilde{\tau})^{\tilde{d}} h(\tau). \tag{2.1}$$

The mapping $h(\tau)$ is analytic at $\tilde{\tau}$ and

$$h(\tilde{\tau})\neq 0. \tag{2.2}$$

We say that $\tilde{d}$ is the order of the (turning point ∞) level crossing at $\tilde{\tau} = \infty$ if

$$E_j(\tau) - E_k(\tau) = \tau^{-\tilde{d}} h(\tau). \tag{2.3}$$

The mapping $h(\tau)$ is analytic at $\tilde{\tau} = \infty$ and

$$h(\infty)\neq 0. \tag{2.4}$$

In other words $\tilde{d}$ is the order of zero of $[E_j(\tau) - E_k(\tau)]$ at $\tilde{\tau}$.

If

$$E_j(\tau) - E_k(\tau) \equiv 0, \tag{2.5}$$

we set

$$\tilde{d}(j,k,\tilde{\tau}) = \infty \tag{2.6}$$

for all times $\tilde{\tau}$, $0\leqslant\tilde{\tau}\leqslant\infty$.

Notice that by the above convention $\tilde{d} = 0$ at $\tilde{\tau}$ implies that at $\tilde{\tau}$ no level crossing occurs. Let

$$R = (r_{jk}) = -U^*U', \quad j,k = 1,...,n. \tag{2.7}$$

For a given ordered pair $\langle j,k\rangle$, we denote by $\tilde{e} = \tilde{e}(j,k,\tilde{\tau})$ the order of zero of $r_{jk}(\tau)$ at $\tilde{\tau}$. If $r_{jk}(\tilde{\tau})\neq 0$ we set $\tilde{e} = 0$. If $r_{jk}(\tau)\equiv 0$ on $[0,\infty]$ we set $\tilde{e} = \infty$ for all points $\tilde{\tau}$ of $[0,\infty]$.

It is an easy exercise to verify that $r_{jk}(\tau) = -\overline{r_{kj}}(\tau)$, $j,k = 1,...n$, since U is a unitary operator. Therefore $\tilde{e}(j,k,\tilde{\tau}) = \tilde{e}(k,j,\tilde{\tau})$.

Finally we denote by $\tilde{m}$ an "index of local perturbation" for each pair $\langle j,k\rangle$ at each time $\tilde{\tau}$, $0\leqslant\tilde{\tau}\leqslant\infty$, as follows:

$$\tilde{m} = (\tilde{e}+1)/(\tilde{d}+1). \tag{2.8}$$

If $\tilde{e} = \infty$ and $\tilde{d} = \infty$ then we define $\tilde{m}$ to be $\tilde{m} = \infty$. Thus our index of local perturbation is defined for all $\tilde{e}$ and $\tilde{d}$ in the range

$$0\leqslant\tilde{e}\leqslant\infty,\ 0\leqslant\tilde{d}\leqslant\infty. \tag{2.9}$$

III. CLASSIFICATION OF SELF-ADJOINT HAMILTONIAN SYSTEMS

It is natural to distinguish among a few classes in the family of systems (1.3) with discrete energy levels. The distinction could be based on the index $\tilde{d}$ defined in Sec. II. Given a Hamiltonian system (1.3) together with (1.4) we distinguish among the following classes.

Convention 3.1: Class I: Nondegenerate. Namely $\tilde{d} = 0$ for all times $0\leqslant\tilde{\tau}\leqslant\infty$ and all $j,k = 1,...,n, j\neq k$.

Class II: Accidentally degenerate. Namely, for at least one pair of energy levels $E_j(\tau)$ and $E_k(\tau), j,k = 1,..,n, j\neq k$ we have at least one time $0\leqslant\tilde{\tau}\leqslant\infty$ such that $\tilde{d} > 0$. However, we never have $\tilde{d} = \infty$.

Class III: Partially totally degenerate. Namely, at least one pair of energy levels $E_j(\tau)$ and $E_k(\tau), j,k = 1,...,n, j\neq k$ are identical for all times. At least one pair of energy levels are not identical.

Class IV: Totally degenerate. Namely, all energy levels are identical to one value $E_1(\tau)$ for all times. It is easily verified that then $H(\tau) = E_1(\tau)I$, where I is the identity operator.

Even though one may believe that classes III and IV rarely occur in applications, the above classification is useful for the sake of the completeness of a mathematical discussion. It is evident that for dimensions $n\geqslant 2$ the classification above divides all Hamiltonian systems into four mutually exclusive classes.

An asymptotic decomposition theorem for solutions of Hamiltonian systems (1.3) as $\epsilon\to 0^+$ could be instrumental to the understanding of the evolution of time dependent self-adjoint Hamiltonian systems. However, such a theorem, which comprehensively covers *all* four classes of Hamiltonians, has not been seen in the literature until recently.

If a Hamiltonian system (1.3) belongs to the class of nondegenerate or the class of totally degenerate Hamiltonians then a comprehensive asymptotic decomposition for their solutions can be extracted from the available literature. Moreover, even in the case that $H(\tau)$ is infinite dimensional and is twice continuously differentiable such that

$$\int_0^\infty \|H''(\tau)\|d\tau < \infty, \tag{3.1}$$

a combination of methods that includes a tool of Kato[12] can be used to obtain a comprehensive asymptotic decomposition. Compare also with Messiah,[13] Chap. XVII. Evidently the case of a totally degenerate Hamiltonian is trivial. All solutions of

$$i\epsilon y' = E_1(\tau)Iy \tag{3.2}$$

have the form

$$y = \left(\exp(i\epsilon)^{-1}\int_0^\tau E_1(\eta)d\eta\right)Ic, \tag{3.3}$$

where c is a constant initial vector. An asymptotic decomposition for special cases of Hamiltonians that pertain to class II, namely of accidental degeneracy, can be extracted from

Born and Fock,[2] Kato,[12] and Friedrichs.[14] The most difficult class that remains is the class of partially totally degenerate Hamiltonians.

Recently Gingold[3,15] provided a comprehensive asymptotic decomposition for two-dimensional Hamiltonian systems. "Invariant" asymptotic formulas and asymptotic decompositions in a generalized sense were developed. Gingold and Hsieh[4] managed to obtain a complete asymptotic decomposition for most general accidentally degenerate n dimensional self-adjoint Hamiltonian systems. A formulation of an asymptotic decomposition theorem that covers *all* four classes of Hamiltonians including the class of partially totally degenerate Hamiltonians can be found in Ref. 16. The details of the proof are given in Gingold and Hsieh.[5] The analysis in Refs. 3, 4, 5, and 15 reinforces the fact that asymptotic expansions in *fractional powers* of ϵ play a basic role in systems with accidental degeneracies. It also points out the fact that traditional asymptotic decompositions could be impossible to get in systems which belong to the partially totally degenerate class. An alternative to the traditional method of stationary phase is also developed in those articles. This alternative method does not resort to integration in the complex plane.

A refinement of the theorem in Gingold and Hsieh[4] will be elaborated upon in the next section. It will serve us in our present study.

IV. A THEOREM OF ASYMPTOTIC DECOMPOSITION

Given a system (1.3) we have the following theorem.

Theorem 4.1: Let the slowly varying Hamiltonian system

$$i\epsilon y' = H(\tau)y \tag{4.1}$$

satisfy Assumption 1.1 and be accidentally (or non-) degenerate. Then the general solution of (4.1) is given by

$$y = U(\tau)\left[\exp(i\epsilon)^{-1}\int_0^\tau D(\eta)d\eta\right](I + P(\tau,\epsilon))c, \tag{4.2}$$

where $U(\tau)$ is a unitary analytic matrix function on $[0,\infty]$ which satisfies the relations (1.4). Here $D(\eta)$ is a certain real valued diagonal matrix to be elaborated upon in the sequel, and $I + P(\tau,\epsilon)$ is an $n\times n$ invertible and continuous matrix function in the domain $0\leqslant\tau\leqslant\infty$, $0\leqslant\epsilon<\infty$, such that

$$P(0,\epsilon) = 0, \quad \|P(\tau,\epsilon)\|\leqslant K\epsilon^m \quad \text{for } 0<\epsilon\ll 1, \tag{4.3}$$

where $\|\ \|$ is the induced Euclidean norm. Here K is a non-negative fixed number independent of τ and ϵ, and m is characterized by

$$m = \inf[(\tilde{e}+1)/(\tilde{d}+1)] = \inf(\tilde{m}). \tag{4.4}$$

The infimum is taken over all pairs of indices $j,k = 1,...,n$, $j\neq k$, and all points $\tilde{\tau}$, $0\leqslant\tilde{\tau}\leqslant\infty$, according to Sec. II.

Proof: The proof follows by a slight modification of the treatment in Gingold and Hsieh.[4] Rather than repeating the details we will outline here the basic features. One may distinguish two stages in the process of asymptotic decomposition. In the first stage a linear transformation is applied. Thanks to Rellich's theorem[6] the existence of a linear unitary and analytic transformation $U(\tau)$, which satisfies (1.4), is guaranteed. Then the transformation

$$y = U(\tau)v \tag{4.5}$$

takes the differential system (4.1) into

$$i\epsilon v' = [E(\tau) - i\epsilon U^*U']v, \tag{4.6}$$

where v is an n column vector. The coefficient matrix of the differential system (4.6) is rearranged into a diagonal and an off-diagonal part as follows. The diagonal part $D(\tau)$ is given by

$$D(\tau) = D = E + i\epsilon\ \mathrm{Diag}[g_1,...,g_n], \tag{4.7}$$

where

$$\mathrm{Diag}\ [g_1,...,g_n] := -\mathrm{Diag}(U^*U'). \tag{4.8}$$

The off-diagonal part is given by

$$R = (r_{jk}) = -U^*U' + \mathrm{Diag}(U^*U'), \quad j,k = 1,...,n. \tag{4.9}$$

Notice that by this arrangement

$$E - i\epsilon U^*U' = D + i\epsilon R, \quad r_{jj}\equiv 0, \quad j = 1,...,n. \tag{4.10}$$

In the second stage of an asymptotic decomposition we first solve for an approximate $n\times n$ fundamental diagonal matrix solution V,

$$i\epsilon V' = DV, \quad V = \exp(i\epsilon)^{-1}\int_0^\tau D(\eta)d\eta. \tag{4.11}$$

Then we set

$$v = V(I + P)c, \tag{4.12}$$

where c is a constant vector and P is a certain $n\times n$ "small perturbation matrix" as $\epsilon\to 0^+$. Consequently P satisfies the matrix differential equation

$$(I + P)' = V^{-1}RV(I + P). \tag{4.13}$$

Moreover, it can be shown that the existence of a solution to the initial value problem determined by (4.13) and $P(0,\epsilon) = 0$ is guaranteed if the integral equation

$$P = LI + L^2I + L^2P, \tag{4.14}$$

with

$$LP(\tau) = \int_0^\tau V^{-1}RVP\,d\eta \tag{4.15}$$

possesses a solution with the desired properties. Indeed this is the case under our assumptions. It turns out that L^2 is a contraction and that LI and L^2I tend to zero *uniformly* for $0\leqslant\tau\leqslant\infty$ as $\epsilon\to 0^+$. Those conclusions can be shown by estimates on terms of the form

$$f(\tau) = \int_a^\tau r_{jk}(s)\left\{\exp(i\epsilon^{-1})\int_a^s[(E_j(u) - E_k(u)) + i\epsilon(g_j(u) - g_k(u))]du\right\}ds, \tag{4.16}$$

$j,k = 1,...,n$, $j\neq k$, as $\epsilon\to 0^+$. It is from such terms that the role of the indices $\tilde{d}$, $\tilde{e}$, and $\tilde{m}$ is revealed. For more details see Ref. 3 and Gingold and Hsieh.[4] The entries of the matrix P can be approximated to any level of accuracy by $\Sigma_{\nu=0}^{\nu=N}L^\nu I$, where N is a non-negative integer.

A good estimate on the *size* of the transition probabilities depends on estimates on the entries of the matrix P. They are provided by Theorem 4.1.

We stress again that the importance of relation (4.3) lies in the fact that the bound $K\epsilon^m$ is uniform for all $0\leqslant\tau\leqslant\infty$.

V. SMOOTHNESS OF EIGENSTATES, TRANSITION, AND DEGENERACY

The purpose of this section is twofold. The first task is to show rigorously that although we need to assume an amount of smoothness on the entries of $H(\tau)$ in order to carry out an asymptotic analysis, we need not assume from the outset that the eigenstates themselves are smooth. This is relevant to our second task, which is to relate quantitatively in an asymptotic sense the concepts of transition and degeneracy.

Assume that the n normalized eigenstates of $H(\tau)$ are the column vectors of the unitary matrix

$$\tilde{U}(\tau) = [\tilde{u}_1(\tau), \tilde{u}_2(t),...,\tilde{u}_n(\tau)] \tag{5.1}$$

which satisfies

$$H(\tau) = \tilde{U}(\tau)E(\tau)\tilde{U}^*(\tau). \tag{5.2}$$

Then, the solution to the initial value problem

$$i\epsilon y_j' = H(\tau)y_j, \quad y_j(0) = \tilde{u}_j(0) \tag{5.3}$$

is given by

$$y_j(\tau) = U(\tau)\left[\exp\int_0^\tau (i\epsilon)^{-1}D(s)ds\right] \times (I+P)U^*(0)\tilde{U}(0)e_j, \tag{5.4}$$

where e_j is the j column of the identity operator.

It is an easy exercise to verify that the matrix function $\tilde{U}^*(\tau)U(\tau)$ is unitary and diagonal. This is thanks to the assumption that $H(\tau)$ belongs to the accidental degeneracy class. Thus

$$U(\tau) \equiv \tilde{U}(\tau)\exp[i\theta(\tau)], \tag{5.5}$$

where

$$\theta(\tau) = \operatorname{diag}[\theta_1(\tau),...,\theta_n(\tau)]. \tag{5.6}$$

The mappings $\theta_j(\tau), j = 1,...,n$, are certain real-valued mappings.

Let us combine (5.5) with (5.4). Then the solution $y_j(\tau)$ to the initial value problem (5.3) is given by

$$y_j(\tau) = \tilde{U}(t)[\exp\{i\theta(\tau)\}]\left[\exp(i\epsilon)^{-1}\int_0^\tau D(s)ds\right] \times [I+P][\exp -i\theta(0)]e_j. \tag{5.7}$$

One may question why it is that from the outset we did not produce the transformation

$$y = \tilde{U}(\tau)v \tag{5.8}$$

rather then the transformation (4.5). To answer this we need to remember that $U(\tau)$ was guaranteed to be analytic. This was crucial for the derivation of (4.6). However, it will turn out that we need not restrict ourselves from the outset with the assumption that the eigenstates themselves are smooth or analytic.

From (5.7) we conclude that for certain scalar coefficients $c_{jl}(\tau,\epsilon)$, $l = 1,...,n$, we have

$$y_j(\tau) = \sum_{l=1}^n c_{jl}(\tau,\epsilon)\tilde{u}_l(\tau), \quad j = 1,...,n. \tag{5.9}$$

Let us calculate the probability q_j^j of the state j to continue to evolve in the state j. Let us also calculate the transition probability q_j^k of the state j to evolve into the state k. Combining the superposition principle (see, e.g., Liboff[17] Chap. 5) with our formulas (5.9) we obtain

$$q_j^j = |c_{jj}(\tau,\epsilon)|^2 = \frac{|\tilde{\lambda}_j|^2|1+p_{jj}|^2}{|\tilde{\lambda}_j|^2|1+p_{jj}|^2 + \sum_{\substack{l=1\\ l\neq j}}^n |\tilde{\lambda}_l p_{lj}|^2} \equiv |\tilde{\lambda}_j|^2|1+p_{jj}|^2 = |1+p_{jj}|^2, \tag{5.10}$$

$$q_j^k = |c_{jk}(\tau,\epsilon)|^2 = \frac{|\tilde{\lambda}_k p_{kj}|^2}{|\tilde{\lambda}_j|^2|1+p_{jj}|^2 + \sum_{\substack{l=1\\ l\neq j}}^n |\tilde{\lambda}_l p_{lj}|^2} \equiv |\tilde{\lambda}_k p_{kj}|^2 = |p_{kj}|^2, \quad k \neq j. \tag{5.11}$$

The denominators in (5.10) and in (5.11) are identically 1 because

$$(y_j^*(\tau),y_j(\tau)) \equiv (y_j^*(0),y_j(0)) = (\tilde{u}_j^*(0),\tilde{u}_j(0)) \equiv 1. \tag{5.12}$$

The entries $\tilde{\lambda}_1,...,\tilde{\lambda}_n$ are given by

$$\operatorname{diag}[\tilde{\lambda}_1,...,\tilde{\lambda}_n] = \exp(i\epsilon)^{-1}\int_0^\tau D(s)ds. \tag{5.13}$$

Notice that by our assumptions $[E(\tau) - i\epsilon U^*(\tau)U'(\tau)]$ is also a Hermitian operator and therefore

$$|\tilde{\lambda}_j| = 1, \quad j = 1,...,n. \tag{5.14}$$

By (4.3) we conclude that each entry p_{jk} of the matrix P satisfies

$$|p_{jk}|^2 = O(\epsilon^{2m}). \tag{5.15}$$

It is evident from formulas (5.10), (5.11), and (5.15) that the *larger m is* the closer to 1 is the probability q_j^j, in an asymptotic sense as $\epsilon \to 0^+$, for each state $y_j(\tau)$ to stay in the eigenstate $\tilde{u}_j(\tau)$. This is so because (5.15) implies that

$$q_j^j = 1 - O(\epsilon^{2m}), \quad q_j^k = O(\epsilon^{2m}), \quad k \neq j, \quad j,k = 1,...,n. \tag{5.16}$$

Evidently at $\tau = 0$,

$$q_j^j = 1, \quad q_j^k = 0. \tag{5.17}$$

It is the size of the transition probabilities and their variation with τ and their dependence on ϵ which we intend to utilize as a measure of the transition in a slowly varying time-dependent quantum mechanical system. Thus we can refer to the transition of one particular eigenstate $\tilde{u}_j$ by analyzing q_j^j. We can refer to the transition of the Hamiltonian system as a whole by refering to all the eigenstates $\tilde{u}_j(\tau)$ and to their probabilities $q_j^j, j = 1,...,n$. We discuss the simultaneous transition of all eigenstates. Before we continue with our discussion it is worthwhile to point out that a straightforward well-known calculation reveals that the values of the transition probabilities q_j^j, q_j^k are independent of the value of $\theta(\tau)$ in (5.5). Since, from now on, we will mainly be concerned with the values of the transition probabilities of nondegenerate or accidently degenerate Hamiltonians, we will assume without loss of generality that

$$\theta(\tau) \equiv 0, \quad U = \tilde{U}. \tag{5.18}$$

We are ready now to justify our principle: "In general, the less degeneracy the less transition."

Consider first the class of Hamiltonians (1.3) for which

the off-diagonal entries of $U^*(\tau)U'(\tau)$ are identically 0. In such a class, regardless of the amount and order of level crossings, it is an easy exercise to verify that the transition probabilities are independent of time and that

$$q_j^j \equiv 1, \quad q_j^k \equiv 0, \quad j \neq k, \quad j,k = 1,2,\ldots . \tag{5.19}$$

"In general" means that in the *larger* family of Hamiltonian systems (1.3) the above class of systems is exceptional.

We turn to the "general case." We view now a Hamiltonian $H(\tau)$ given by (1.4) as composed of two independent parts. The one is a set of n orthonormal eigenvectors that make up the unitary operator $U(\tau)$. The second part is the matrix of energy levels $E(\tau)$.

Consider two Hamiltonian systems (1.3) with coefficient matrices $H_1(\tau)$ and $H_2(\tau)$, respectively, such that

$$H_1(\tau) = U(\tau)E_1(\tau)U^*(\tau), \quad H_2(\tau) = U(\tau)E_2(\tau)U^*(\tau). \tag{5.20}$$

Then $H_1(\tau)$ and $H_2(\tau)$ share the same set of eigenstates but they could differ in their energy levels in a manner to be specified below. Also, the jth states that evolve are, respectively,

$$y_j = U(\tau)\left[\exp(i\epsilon)^{-1}\int_0^\tau D(\eta)d\eta\right](I+P_1)e_j, \tag{5.21}$$

$$y_j = U(\tau)\left[\exp(i\epsilon)^{-1}\int_0^\tau D(\eta)d\eta\right](I+P_2)e_j. \tag{5.22}$$

The matrix $P_1 = (p_{jk}^1)$ pertains to the initial value problem

$$i\epsilon y_j' = H_1 y_j, \quad y_j(0) = u_j(0). \tag{5.23}$$

The corresponding transition probabilities will be denoted by q_{j1}^j, q_{j1}^k. In an analogous and obvious manner we will have the matrix $P_2 = (p_{jk}^2), j,k = 1,\ldots,n$, together with the transition probabilities q_{j2}^j, q_{j2}^k related to

$$i\epsilon y_j' = H_2 y_j, \quad y_j(0) = u_j(0). \tag{5.24}$$

Moreover, with $\|\ \|$ denoting the induced Euclidean norm we have

$$|p_{jk}^1|^2 \leqslant \|P_1\|^2 \leqslant K_1^2\epsilon^{2m_1}, \quad |p_{jk}^2|^2 \leqslant \|P_2\|^2 \leqslant K_2^2\epsilon^{2m_2}, \tag{5.25}$$

where m_1 and m_2, K_1 and K_2, have, respectively, the same meaning as in (4.4). Denote by $\tilde{d}_1(j,k,\tilde{\tau})$ and by $\tilde{d}_2(j,k,\tilde{\tau})$ the order of level crossing in (5.23) and (5.24), respectively. Assume that for all $j,k,\tilde{\tau}$ we have

$$d_1(j,k,\tilde{\tau}) \leqslant d_2(j,k,\tilde{\tau}), \tag{5.26}$$

and that for one specific triplet $j,k,\tilde{\tau}$, we have

$$d_1(j,k,\tilde{\tau}) < d_2(j,k,\tilde{\tau}), \quad e(j,k,\tilde{\tau}) \neq \infty.$$

By the definition of the indices m_1, m_2 we have $m_2 < m_1$ and therefore for $0 < \epsilon \ll 1$ we have $\epsilon^{m_2} > \epsilon^{m_1}$. Obviously,

$$\sum_{\substack{j=1 \\ j\neq k}}^n q_{j1}^k = O(\epsilon^{2m_1}), \quad j,k = 1,\ldots,n, \tag{5.27}$$

$$\sum_{\substack{j=1 \\ j\neq k}}^n q_{j2}^k = O(\epsilon^{2m_2}), \quad j,k = 1,\ldots,n, \tag{5.28}$$

$$q_{j1}^j = 1 - O(\epsilon^{2m_1}), \quad q_{j2}^j = 1 - O(\epsilon^{2m_2}), \quad \text{as } \epsilon \to 0^+. \tag{5.29}$$

It is in this sense that we propose the principle: "In general, the less degeneracy the less transition." It is in this sense that Gingold and Hsieh[4] proposed: "In general, the less degeneracy the closer is the state which evolves to its initial eigenstate."

The index m in (4.3) could be too crude a yardstick to relate degeneracy and transition quantitatively. The fact that $P = 0$ in (4.2) implies that the transition probabilities satisfy $q_j^j \equiv 1$ and $q_j^k \equiv 0, j,k = 1,\ldots,n, j \neq k$, and consequently $S(\epsilon)$ defined by

$$S(\epsilon) = \epsilon^{-2}\sum_{j,k=1}^n \int_0^\infty |p_{jk}(\tau,\epsilon)|^2 d\tau \tag{5.30}$$

is a good measure of the amount of transition in a Hamiltonian system. However, $P(\tau,\epsilon)$ is a solution of an integral equation and in general is obtained via a laborious process of approximations. Mathematical considerations show that $LI = 0$ in (4.14) implies $P = 0$. Moreover, all of the degeneracy indices $\tilde{d}$, the indices $\tilde{e}$, $\tilde{m}$, and the index m can be obtained (read off) from LI. This can be seen in detail in Refs. 3 and 4. Therefore we suggest the following measure of degeneracy $S_1(\epsilon)$:

$$S_1(\epsilon) = \epsilon^{-2}\sum_{\substack{j,k=1 \\ j\neq k}}^n \left|\int_0^\infty r_{jk}(\tau,\epsilon)\left[\exp(i\epsilon)^{-1}\int_0^\tau [(E_j(\eta) - E_k(\eta)) + i\epsilon(g_j(\eta) - g_k(\eta))]d\eta\right]d\tau\right|^2. \tag{5.31}$$

Recall that the entries $r_{jk}(\tau,\epsilon)$ are defined by (2.7) and that the entries g_j are defined by (4.8). We remark on the following. If $H(\tau)$ is nondegenerate then $m \geqslant 1$ in (4.3) and $S_1(\epsilon)$ is a bounded function of ϵ as $\epsilon \to 0$. In general, if the off diagonal entries of U^*U' are not all zero and $H(\tau)$ is nondegenerate then $m = 1$ and $S_1(\epsilon)$ is a bounded function of ϵ as $\epsilon \to 0^+$. If $H(\tau)$ is degenerate then in general $S_1(\epsilon)$ is an unbounded function of ϵ as $\epsilon \to 0^+$. If the off-diagonal entries U^*U' are identically zero then $m = \infty$ and $S_1(\epsilon) \equiv 0$.

Our analysis seems to reinforce the Jahn–Teller effect[1] if we are willing to accept that less transition implies more stability. Jahn and Teller[1] investigated the conditions under which a polyatomic molecule can have a stable equilibrium configuration when its electronic state has orbital degeneracy, i.e., not arising from the spin. They applied group theory to perturbation calculations and concluded that "orbital electronic degeneracy and stability of the nuclear configuration are incompatible unless all the atoms of a molecule lie on a straight line." It is interesting to note that our principle is obtained in a general and different setting using different methods. In addition we produced a quantitative relation which associates transition and degeneracy.

VI. RAMIFICATIONS OF A MATHEMATICAL FORMULA

The asymptotic decomposition (4.2) is associated with the mathematical relations

$$\|P\| \leqslant K\epsilon^m, \quad P = (p_{jk}), \quad |p_{jk}| \leqslant K\epsilon^m, \quad j,k = 1,\ldots,n, \tag{6.1}$$

where P may be considered in (4.2) as a transition probability matrix by virtue of the relations

$$q_j^j = |c_{jj}(\tau,\epsilon)|^2 = |1 + p_{jj}|^2 = 1 + O(\epsilon^{2m}), \tag{6.2}$$

$$q_j^k = |c_{jk}(\tau,\epsilon)|^2 = |p_{jk}|^2 = O(\epsilon^{2m}), \quad j \neq k. \tag{6.3}$$

The clue to the derivation of the principle "In general, the less degeneracy the less transition" is an innocent looking relation

$$m = \mathrm{Inf}[(\tilde{e}+1)/(\tilde{d}+1)], \tag{6.4}$$

where the infimum is taken over all pairs of energy levels $\langle E_j(\tau), E_k(\tau)\rangle$ and entries r_{jk},

$$(r_{jk}) = -U^*U' + \mathrm{Diag}(U^*U'), \quad j \neq k, \quad j,k = 1,\ldots,n, \tag{6.5}$$

at all points $\tilde{\tau}$, $0 \leqslant \tilde{\tau} \leqslant \infty$, as explained in Secs. II–V. It seems that the relation (6.4) [*together with* (6.2) and (6.3)] is capable of generating several additional clues.

(i) The relation (6.4) indicates by the presence of $\tilde{e}$ that "the transition in a slowly varying system (1.3) depends on the *smoothness properties* of *one* set of orthonormal eigenvectors of $H(\tau)$." This statement is better understood by recalling that $\tilde{e}$ denotes an order of zero of r_{jk}, $j \neq k$ in (6.5). Notice that the r_{jk} depend on the smoothness of U.

(ii) The exceptional case, which is termed in our principle "*in general*," is also adequately described by (6.4). Because then we have $r_{jk} \equiv 0$ for $j \neq k$ and by the convention of Sec. II we have $\tilde{e}(j,k,\tilde{\tau}) = \infty$ for all $j,k,\tilde{\tau}$. Consequently, $m = \infty$ and $\epsilon^m = 0$ is the right interpretation for $0 < \epsilon \ll 1$. This indeed implies that for all times $q_j^j \equiv 1$ and $q_j^k \equiv 0, j \neq k$.

(iii) The formula (6.4) establishes the validity of the adiabatic approximation theorem in quantum mechanics for nondegenerate and accidentally degenerate Hamiltonians $H(\tau)$. This theorem can be traced back to Ehrenfest[18] and to Born and Fock.[2] We recall that a version of the adiabatic approximation theorem in quantum mechanics states the following: "In a slowly varying time-dependent Hamiltonian, a state y_j which evolves from an eigenstate $u_j(0)$ will continue to evolve asymptotically as $\epsilon \to 0^+$ in the eigenstate $u_j(\tau)$ for all times." Compare, e.g., with Mesiah,[13] Chap. XVII and with Liboff,[17] Chap. 5. Proof of the theorem for special cases of degeneracies were given by Born and Fock,[2] Kato[12] and Friedrichs.[14] A complete proof in the case that $H(\tau)$ is finite dimensional and belongs to the class of accidentally degenerate Hamiltonians was given in Ref. 3 and Gingold and Hsieh.[4] The proof boils down to showing that asymptotical-ly $q_j^k \sim 0$, $j \neq k$, $q_j^j \sim 1$ as $\epsilon \to 0^+$. The proof is indicated by (6.4) where we have under the above circumstances

$$M = \mathrm{Inf}\left(\frac{\tilde{e}+1}{\tilde{d}+1}\right) = \mathrm{Min}\left(\frac{\tilde{e}+1}{\tilde{d}+1}\right) = \frac{\tilde{e}^\dagger + 1}{\tilde{d}^\dagger + 1} > 0 \tag{6.6}$$

for a certain pair of numbers $\tilde{d}^\dagger$ and $\tilde{e}^\dagger$.

(iv) But is the adiabatic approximation theorem true for *all* four classes of Hamiltonians? A hint towards its invalidity is given again by our innocent relation (6.4). If not all $\tilde{e}(j,k,\tilde{\tau})$ satisfy $\tilde{e} = \infty$ we could obtain $m = 0$ if some $\tilde{d}(j,k,\tilde{\tau}) = \infty$. In other words we may have $q_j^k \sim O(1)$ as $\epsilon \to 0^+$ rather than $q_j^k \sim o(1)$ as $\epsilon \to 0^+$ if for some choice of the parameters $j,k,\tilde{\tau}$ we have $\tilde{d}(j,k,\tilde{\tau}) = \infty$. By Sec. II, $\tilde{d} = \infty$ if for $j \neq k$ we have $E_j(\tau) - E_k(\tau) \equiv 0$. Indeed, Gingold[16] took up this hint and produced a counterexample to the adiabatic approximation theorem for a Hamiltonian that possesses two identical energy levels. A generalized version of the theorem was also offered in Ref. 16. It is then natural to ask what indications we can get from (6.4) if $H(\tau)$ is an infinite-dimensional Hermitian operator operating on a Hilbert space? It will be shown elsewhere that then if $\mathrm{Sup}\,\tilde{d} < \infty$ (namely the total order of degeneracy is bounded) and if certain additional conditions hold then we will still have $m > 0$ in (4.6). Otherwise, we would speculate that $m = 0$ and that the adiabatic approximation theorem may not hold even if $H(\tau)$ belongs to the class of accidentally degenerate Hamiltonians.

(v) Admissibility of $H(\tau)$. Assume that we deal with an evolving Hamiltonian system that pertains to a quantum mechanical system which satisfies the following.

Postulate 6.1: The quantum mechanical system is such that it tends after a long time to settle into a "most stable" configuration.

In other words our principle "In general, the less degeneracy the less transition" is not a consequence of mathematical manipulations but is a result of the characteristics of mother nature. What use can we make of (6.4) then? Assume that our Hamiltonian is such that $\tilde{e}(j,k,\tilde{\tau}) \neq \infty$ for $j,k,\tilde{\tau}$. Let the matrix $U(\tau)$ be fixed in advance. Assume that we are willing to accept "more stability" as "less transition." Then, in order to guarantee more stability after a long time the combination of Postulate 6.1 with (6.4) indicates a restriction on the values of $\tilde{d}(j,k,\tilde{\tau})$ at times $\tilde{\tau}$ large enough. For $\tilde{\tau}$ large enough, $\tilde{d}$ need not exceed the values of $\tilde{d}$ for $0 \leqslant \tilde{\tau} = \tilde{\epsilon}t \leqslant \gamma$, where γ is a certain finite number. Thus (6.4) could be used for obtaining admissibility conditions on $H(\tau)$.

(vi) *Design* of quantum mechanical systems can be aided by (6.4) and in particular by the principle proposed. Assume that we need to design certain quantum mechanical systems that evolve according to (1.3). Suppose that our systems are such that we can write the Hamiltonians $H(\tau)$ for various possible designs. If we are interested in a most stable design we would choose an $H(\tau)$ that is either nondegenerate or possesses accidental degeneracies and is such that the off-diagonal entries of U^*U' are identically 0. If this is impossible, we will create a design with energy levels that are separated as much as possible, to make m in (6.4) the largest. This is in accordance with the principle proposed.

It is worthwhile to mention that Theorem 4.1 and (6.4) could be related to problems regarding adiabatic invariants discussed in Wasow.[19] A recent research text on linear turning point theory is Ref. 20.

ACKNOWLEDGMENTS

Acknowledgment is due to Dr. Trutzer and to Professor Gould who improved considerably the style in this paper.

This research was supported in part by the NASA Research Grant No. NAG-1-741.

[1] H. A. Jahn and E. Teller, "Stability of polyatomic molecules in degenerate electronic states," Proc. R. Soc. London, Ser. A **161**, 220 (1937).

[2]M. Born and V. Fock, "Beweis des Adiabatensatzes," Z. Phys. **5**, 165 (1928).
[3]H. Gingold, "An asymptotic decomposition method applied to multi-turning point problems," SIAM J. Math. Anal. **15**, 7 (1985).
[4]H. Gingold and P. F. Hsieh, "Global decomposition of a Hamiltonian system with multi-degenerate energy levels," SIAM J. Math. Anal. **18**, 1 (1987).
[5]H. Gingold and P. F. Hsieh, "Asymptotic solution of a Hamiltonian system with several turning points," to be published in SIAM J. Math. Anal.
[6]F. Rellich, "Störungstheorie der Spektralzerlegung, I," Math. Ann. **113**, 600 (1937).
[7]F. Hund, Z. Phys. **40**, 742 (1927).
[8]von Neumann and E. Wigner, "On the behavior of eigenvalues in adiabatic processes," Phys. Z. **30**, 467 (1929).
[9]R. S. Knox and A. Gold, *Symmetry in Solid State* (Benjamin, New York 1964).
[10]E. E. Nikitin, *Theory of Elementary Atomic and Molecular Processes in Gases* (Clarendon, Oxford, 1974).
[11]R. G. Pearson, *Symmetry Rules for Chemical Reactions* (Wiley, New York 1976).
[12]T. Kato, "On the adiabatic theorem of quantum mechanics," J. Phys. Soc. Jpn. **5**, 435 (1955).
[13]A. Messiah, *Quantum Mechanics, Vol. II* (Interscience, New York, 1961).
[14]K. O. Friedrichs, *Special Topics in Analysis, Lecture Notes* (New York Univ., New York, 1953).
[15]H. Gingold, "Asymptotic decompositions on an entire interval for two by two first order linear differential systems with multi-coalescing turning points," WVU preprint, pp. 1–228.
[16]H. Gingold, "A counter-example to the adiabatic approximation theorem in quantum mechanics," WVU preprint.
[17]R. L. Liboff, *Introductory Quantum Mechanics* (Holden–Day, San Francisco, 1980).
[18]P. Ehrenfest, "On adiabatic changes of a system in connection with the quantum theory." Proc. R. Acad. Amsterdam **XIX**, 576 (1916).
[19]W. Wasow, "Adiabatic invariants in the asymptotic theory of ordinary linear differential equations," *Asymptoic Methods and Singular Perturbations*, SIAM-AMS Proceedings, edited by R. E. O'Malley, Jr. (SIAM, Philadelphia, 1976), Vol. 10, pp. 131–144.
[20]W. Wasow, *Linear Turning Point Theory* (Springer, New York, 1985).

Tunneling in one-dimensional ideal barriers

Jacek M. Kowalski
Department of Physics, University of Dallas, Irving, Texas 75062-4799

John L. Fry
Department of Physics, University of Texas at Arlington, Arlington, Texas 76019

(Received 13 February 1987; accepted for publication 3 June 1987)

General properties of the transmission coefficient of an ideal, one-dimensional potential barrier of arbitrary shape are studied. It is proved that an arbitrary symmetric barrier is perfectly transparent for at least one energy in each energy band of the related band problem, where the barrier potential is periodically continued on the whole real axis. Recursion relations are obtained for transmission coefficients of barriers consisting of 2^k structural units. They are used in a simple proof showing that transmission coefficients of finite barriers composed of m identical arbitrary structural units have chaotic behavior for almost all energies for $m \to \infty$ in each energy band. There exists, however, becoming more dense with m, a countable set of energies in each energy band where finite repeated barriers are perfectly transparent. The results are illustrated by a numerical example.

I. INTRODUCTION

The problem of quantum particle passage through a potential barrier still receives considerable attention, mostly due to the continued progress in tunneling spectroscopy methods and the possibility of creating a large variety of artificial layered microstructures (quantum devices).[1] Besides the well-known studies of tunneling through some model random potentials originated by Lifshitz and coworkers,[2] recent theoretical investigations have been concentrated on tunneling in superconductors[3] and the related general tunneling problem with energy dissipation.[4] Much attention has also been paid to the tunneling dynamics in the presence of external electric and magnetic fields.[5–8]

It may seem that the simplest, nondissipative stationary tunneling problem with its mathematics essentially coinciding with the more than 100 year old Sturm–Louville problem is a closed subject to be presented in introductory quantum-mechanical texts, discussing standard exact solutions and equally well-known approximate methods like quasiclassical approximation. However, the group structures related to ideal one-dimensional barriers were only very recently investigated[9] with interesting applications. The one-dimensional Schrödinger equation also remains a subject of continuous mathematical investigations,[10] with many quite recent important results.

In this paper we prove five simple lemmas describing analytic properties of the transmission coefficient for a general class of one-dimensional potentials with compact support. Section II contains necessary preliminaries. Although its contents can hardly be claimed new, it still gives, in our opinion, the most concise description of the tunneling problem (see also Sec. VI where some other pedagogical advantages of this approach are clearly seen). In particular, we emphasize the relation between the tunneling problem and the band problem for a periodic potential with a "unit cell" coinciding with the single barrier shape. Two comparison lemmas in Sec. III describe analytic properties of the transmission coefficient at low energies (below the lowest potential energy in the barrier region). Section IV is devoted to an elementary discussion of the necessary and sufficient conditions for a complete barrier transparency at some energies. The main result (Lemma 3) shows that the localization of the complete resonances (transparency equal to 1) on the energy axis furnishes some information concerning the band structure of the related infinite periodic solid. The complete resonances are also always present for barriers composed of 2^k identical structure units. They are discussed in Sec. V, where a two-variable iterative map is obtained for such a sequence of barriers, with one component being just a logistic map in the chaotic and ergodic regime. This implies that the approach to the perfect transparency in the allowed energy bands for such sequences of barriers is quite specific, with a dense set of complete resonances interlocked by local minima. The concluding Sec. VI contains some applications and numerical illustrations for some model potentials.

II. PRELIMINARIES

A. Transfer operator in terms of fundamental solutions

Consider the one-dimensional Schrödinger equation

$$\psi'' + (\epsilon - U(x))\psi = 0 \quad (\hbar^2/2m = 1,\ x \in \mathbb{R}), \tag{1}$$

where the potential U is an arbitrary, piecewise continuous, real, bounded function with compact support ($U(x) \equiv 0$ outside of a given interval $[0,L]$). The tunneling problem consists of finding all C^1 solutions of Eq. (1), parametrically dependent on $\epsilon \in (0, +\infty)$, and behaving as $\exp(i\epsilon^{1/2}x) + a\exp(-i\epsilon^{1/2}x)$ for $x < 0$ and as $c\exp(i\epsilon^{1/2}x)$ for $x > L$ (the case of a normalized particle particle beam incident from the left). Introducing the variable: $\xi := \psi'$, one may consider the equivalent canonical system

$$\psi' = \xi, \quad \xi' = (U(x) - \epsilon)\psi, \tag{2}$$

or, in the matrix notation,

 0022-2488/87/102407-09$02.50 © 1987 American Institute of Physics

$$\psi' = G(x;\epsilon)\psi, \quad \psi = \begin{pmatrix} \psi \\ \psi' \end{pmatrix}, \tag{3}$$

where the generator G is the traceless 2×2 operator

$$G(x;\epsilon) = \begin{pmatrix} 0 & 1 \\ U(x) - \epsilon & 0 \end{pmatrix}. \tag{4}$$

Together with G one may also consider "evolution" or "transfer" operators defined on the solution set by

$$\psi(x) := M(x,x_0;\epsilon)\psi(x_0). \tag{5}$$

[Below we will often set $x_0 = 0$, $x = L$ and then use an abbreviated notation $M(L,0;\epsilon) \equiv M(\epsilon)$ for this type of transfer operator and other related quantities.] Operators $M(x,x_0,\epsilon)$ are *real* and unimodular. The simplest way to prove it is to consider the fundamental solutions of Eq. (3) obeying the initial conditions

$$\psi_1(x_0;\epsilon) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \psi_2(x_0;\epsilon) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6}$$

[and thus having the Wronskian $W(\psi_1,\psi_2) = 1$]. It is well known that these solutions are both real, for real ϵ and U. As ψ_1 and ψ_2 form the natural, canonical basis in the solution space at the point x_0, the transfer operator can be written in terms of these solutions

$$M(x,x_0;\epsilon) = \begin{pmatrix} \psi_1(x;\epsilon) & \psi_2(x;\epsilon) \\ \psi_1'(x;\epsilon) & \psi_2'(x;\epsilon) \end{pmatrix}, \tag{7}$$

which completes the proof.

Representation (7) reduces the tunneling problem to the solution of the Cauchy problem at $x = 0$ or, if convenient, for some other intermediate points, with a product transfer operator and an automatic solution matching at these points. The resulting group structure of transfer operators (in several different forms from the above-presented representations) has been thoroughly investigated in Ref. 9. Below we will show that this representation is extremely useful in studying the analytic properties of the transmission coefficient $|c|^2$ (also called the "barrier transparency"), as well as in the numerical calculations.

B. Transfer operator in terms of arbitrary solution basis

Let ϕ_1, ϕ_2 be two arbitrary, linearly independent solutions of Eq. (1) in the $[0,L]$ interval. The fundamental solutions ψ_1, ψ_2 and their derivatives (matrix elements of the transfer operator) can be easily expressed in terms of ϕ_1 and ϕ_2:

$$\begin{aligned}
\psi_1(x;\epsilon) &= W^{-1}(\phi_1,\phi_2)(\phi_2'(0,\epsilon)\phi_1(x;\epsilon) \\
&\quad - \phi_1'(0;\epsilon)\phi_2(x;\epsilon)), \\
\psi_2(x;\epsilon) &= W^{-1}(\phi_1,\phi_2)(\phi_1(0,\epsilon)\phi_2(x;\epsilon) \\
&\quad - \phi_2(0;\epsilon)\phi_1'(x;\epsilon)), \\
\psi_1'(x;\epsilon) &= W^{-1}(\phi_1,\phi_2)(\phi_2'(0,\epsilon)\phi_1'(x;\epsilon) \\
&\quad - \phi_1'(0;\epsilon)\phi_2(x;\epsilon)), \\
\psi_2'(x;\epsilon) &= W^{-1}(\phi_1,\phi_2)(\phi_1(0,\epsilon)\phi_2'(x;\epsilon) \\
&\quad - \phi_2(0;\epsilon)\phi_1'(x;\epsilon)).
\end{aligned} \tag{8}$$

Formulas (8) are useful when ϕ_1, ϕ_2 are originally known or have some known additional properties (like Bloch solutions). They can also be used in examining the properties of the transfer operator at the energies corresponding to the allowed and forbidden energy intervals (see below, Sec. II D).

C. Barrier transparency

The transmission coefficient of the barrier, $\sigma := |c|^2$, is obviously a function of the energy parameter ϵ and a functional of U. Here σ can be obtained by solving the equation

$$M(\epsilon)\begin{pmatrix} 1+a \\ i\epsilon^{1/2}(1-a) \end{pmatrix} = c\exp(i\epsilon^{1/2}L)\begin{pmatrix} 1 \\ i\epsilon^{1/2} \end{pmatrix}. \tag{9}$$

This immediately leads to the following expression for σ in terms of fundamental solutions:

$$\sigma(\epsilon) = 1/[\mu^2(\epsilon) + \nu^2(\epsilon)], \tag{10}$$

where

$$\mu(\epsilon) := \tfrac{1}{2}(\psi_1(L;\epsilon) + \psi_2'(L;\epsilon)), \tag{11}$$

and

$$\nu(\epsilon) := \tfrac{1}{2}(\epsilon^{1/2}\psi_2(L;\epsilon) - \epsilon^{-1/2}\psi_1'(L;\epsilon)). \tag{12}$$

In some considerations below, an expanded version of the formula (10) will be useful:

$$\begin{aligned}
\sigma(\epsilon) = 4(&\psi_1^2(L;\epsilon) + \psi_2'^2(L;\epsilon) + \epsilon\psi_2^2(L;\epsilon) \\
&+ \epsilon^{-1}\psi_1'^2(L;\epsilon) + 2)^{-1}.
\end{aligned} \tag{13}$$

It follows that σ is an analytic function of ϵ for $\epsilon > 0$. Due to the current [Wronskian $W(\psi^*,\psi)$] conservation we have $|a|^2 + |c|^2 = 1$, and hence $\mu^2(\epsilon) + \nu^2(\epsilon) \geqslant 1$ and $\sigma(\epsilon) \leqslant 1$, as expected. Both functions μ and ν have obvious high energy asymptotics for bounded potentials U: $\mu(\epsilon) \sim \cos\epsilon^{1/2}L$, $\nu(\epsilon) \sim \sin\epsilon^{1/2}L$, and the transparency is arbitrarily close to unity for sufficiently high energies.

D. Equivalence classes of the transfer operator and relations with the band structure problem

Representation (7) allows immediate listing of all the equivalence classes of the transfer operators $M(\epsilon)$ (with respect to complex transformations). For any set of real unimodular matrices M, these are determined by the solution type of the eigenvalue problem,

$$\lambda^2 - (\mathrm{Tr}\, M)\lambda + 1 = 0, \tag{14}$$

where, in our case,

$$\mathrm{Tr}\, M = 2\mu(\epsilon) = \psi_1(L;\epsilon) + \psi_2'(L;\epsilon). \tag{15}$$

For the reader's convenience we list once again these well-known classes: (a) for $|\mu| > 1$, Eq. (14) has real distinct roots and

$$M(\epsilon) \sim \begin{pmatrix} \lambda(\epsilon) & 0 \\ 0 & \lambda^{-1}(\epsilon) \end{pmatrix}; \tag{16}$$

(b) for $|\mu| < 1$, Eq. (14) has complex roots on the unit circle, $\lambda_{1,2} = \exp(\pm i\theta(\epsilon))$, and

$$M(\epsilon) \sim \begin{pmatrix} \exp(i\theta(\epsilon)) & 0 \\ 0 & \exp(-i\theta(\epsilon)) \end{pmatrix}; \tag{17}$$

and (c) for $\mu = 1$ ($\mu = -1$), Eq. (14) has double root $\lambda = 1$ ($\lambda = -1$) and $M(\epsilon)$ is either diagonal matrix [$M(\epsilon) = \pm I$] or equivalent to

$$\begin{pmatrix} \pm 1 & \gamma \\ 0 & \pm 1 \end{pmatrix} \quad (\gamma \neq 0). \tag{18}$$

The properties of μ as a function of ϵ were investigated in classical papers by Kramers,[11] James,[12] and Kohn,[13] in the context of the band structure problem for an arbitrary periodic potential with period L. There the solution can be constructed from "cellular" solutions in an arbitrary interval of length L, with the eigenvalues of the transfer operator uniquely determining a class of "self-matching" (in the sense of James) solutions obeying the condition

$$\phi(L;\epsilon) = \lambda(\epsilon)\phi(0;\epsilon). \tag{19}$$

These are the solutions of Eq. (3) on the $[0,L]$ interval with an eigenvector of the transfer operator $M(\epsilon)$ taken as a vector of initial conditions. For complex λ these are the usual Bloch solutions. Solutions with real eigenvalues λ lying outside the unit circle are subsequently excluded in the periodic problem as physically unacceptable (exponentially increasing for $x \to +\infty$ or $x \to -\infty$).

The most important property of the μ function is that the zeros of its derivative $\partial\mu/\partial\epsilon$ may only occur[13] for $|\mu| \geqslant 1$. It follows that for any energy ϵ such that $|\mu(\epsilon)| \leqslant 1$, there exists a whole interval (an energy band) on the energy axis where μ is a one-to-one function of energy, mapping this interval onto $[-1,+1]$ interval (see Fig. 1). Different energy bands may still have common end points. This may happen if and only if $|\mu| = 1$ and $\partial\mu/\partial\epsilon = 0$ for some energy. As usual, for a fixed energy band one can parametrize complex roots setting

$$\cos k(\epsilon)L = \mu(\epsilon) \tag{20}$$

and choosing, e.g., $k(\epsilon) \in [0,\pi/L]$, $\lambda_1 = \exp(ik(\epsilon)L)$, $\lambda_2 = \exp(-ik(\epsilon)L)$. So defined, $k(\epsilon)$ is a one-to-one function of energy for a given band. As for $k \neq 0$ and $k \neq \pi/L$ there are always two linearly independent solutions of the Bloch type, in the standard approach one considers k as a quasimomentum varying within $[-\pi/L, \pi/L]$ interval and $\epsilon(k) = \epsilon(-k)$ by definition. Then there is exactly one Bloch-type solution for each $k \in [-\pi/L, \pi/L]$ (with $k = 0$ corresponding to the band center and $k = \pm\pi/L$ to the band edges in the k-ϵ plane).

It is obvious that all these general considerations remain relevant in the tunneling problem, for a potential barrier consisting of a single or any finite number of identical structural units. Of course, real exponential solutions are physical here and contribute to the barrier characteristics. It is also clear in this context that frequently discussed resonant tunneling (usually associated with the existence of an almost localized, or indeed localized in a "shifted" barrier, energy level, see, e.g., Refs. 1 and 2) can also be considered in terms of the band limit, where intuitively one may expect a perfect transparency for band energies and a vanishing one for forbidden energy intervals. Below we investigate the resonances and the existence of the band limit for arbitrary finite size structures, without any additional approximations, like the quasiclassical one used in Ref. 14. At present, an immediate consequence of representation (10) is that $\sigma(\epsilon) < 1$ inside the forbidden energy intervals. Hence a complete resonance $[\sigma(\epsilon) = 1]$ may only happen at some energies from the allowed energy bands. Before further discussion of complete resonances (Secs. IV and V) we will consider first the behavior of ϵ in the low energy limit for some general subclass of barrier potentials.

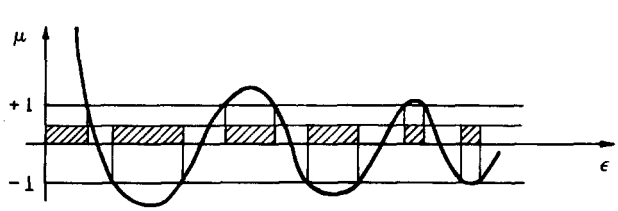


FIG. 1. Qualitative behavior of the μ function for nonoverlapping energy bands. Hatched regions indicate forbidden energy intervals.

III. NONRESONANT TUNNELING: COMPARISON LEMMAS

Representation (10) yields two simple lemmas for the specific case of potential with $U_1 = \inf_{[0,L]} U(x) > 0$.

Lemma I: Given the two tunneling problems

$$\psi'' + (\epsilon - U(x))\psi = 0, \quad \psi'' + (\epsilon - \tilde{U}(x))\psi = 0, \tag{21}$$

where $\tilde{U}(x) \geqslant U(x)$ on $[0,L]$. Then for $0 < \epsilon < U_1$ (see Fig. 2),

$$\sigma(\epsilon;[\tilde{U}]) \leqslant \sigma(\epsilon;[U]). \tag{22}$$

Proof: Let ψ_1, ψ_2, and $\tilde{\psi}_1$, $\tilde{\psi}_2$ be fundamental solutions of Eqs. (1) for, respectively, potentials U and $\tilde{U}$. For $0 < \epsilon < U_1$ all these solutions are increasing on the $[0,L]$ interval. (It follows from the fact that they have the same sign as their second derivatives in the indicated energy interval and specific initial conditions.) From the Wronskian theorem we have

$$W(\psi_i(x), \tilde{\psi}_i(x)) = \int_0^x (\tilde{U}(t) - U(t))\psi_i(t)\tilde{\psi}_i(t)dt \qquad (i = 1,2).$$

Hence

$$\frac{\tilde{\psi}_i'(x)}{\tilde{\psi}_i(x)} \geqslant \frac{\psi_i'(x)}{\psi_i(x)}, \quad x > 0.$$

Integrating the last inequality[15] we obtain $\tilde{\psi}_i(x) > \psi_i(x)$ and hence $\tilde{\psi}_i'(x) > \psi_i'(x)$, as well. Inspection of the expression for transparency in the form (13) completes the proof.

Lemma II: Given the tunneling problem

$$\psi'' + (\epsilon - U(x)) = 0, \quad U_1 = \inf_{[0,L]} U(x) > 0, \tag{23}$$

σ is an increasing function of ϵ for $0 < \epsilon < U_1$.

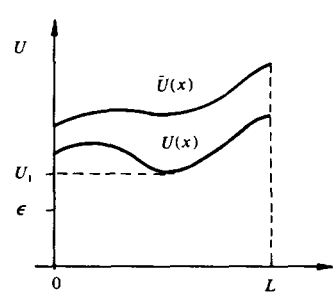


FIG. 2. First comparison lemma. Potential barrier U and its majorant $\tilde{U}$.

Proof: It can be carried out along the same lines as for Lemma I. Alternatively, for two arbitrary energies $\epsilon_1 > \epsilon_2$, from the indicated interval and such that $\epsilon_2 > \epsilon_1$, one may set $\epsilon = \epsilon_2$, $\tilde{U}(x) = U(x) + (\epsilon_2 - \epsilon_1)$, and use Lemma I.

Lemmas I and II characterize the so-called[2] "nonresonant" regime in tunneling, where local extrema in transparency are absent. Lemma I can be used in uniform estimations of the barrier transparency using, e.g., piecewise constant majorants and minorants of a given potential U.

IV. COMPLETE RESONANCES

Let us investigate the necessary and sufficient conditions for a complete barrier transparency at some resonant energy ϵ_r. We already noticed that it may happen only for energies from the allowed bands, where $\mu(\epsilon) = \cos k(\epsilon)L$ with real $k(\epsilon) \in [-\pi/L, \pi/L]$. Hence at a complete resonance there exists such a wave vector $k(\epsilon_r)$ that

$$\psi_1(L;\epsilon_r) + \psi_2'(L;\epsilon_r) = 2\cos k(\epsilon_r)L, \tag{24a}$$

$$\epsilon^{1/2}\psi_2(L;\epsilon_r) - \epsilon^{-1/2}\psi_1'(L;\epsilon_r) = 2\sin k(\epsilon_r)L, \tag{24b}$$

and, as always,

$$\psi_1(L;\epsilon_r)\psi_2'(L;\epsilon_r) - \psi_2(L;\epsilon_r)\psi_1'(L;\epsilon_r) = 1. \tag{24c}$$

It is easy to check that the only real solution to the system (24) is

$$\psi_1(L;\epsilon_r) = \psi_2'(L;\epsilon_r) = \cos k(\epsilon_r)L,$$

$$\psi_2(L;\epsilon_r) = \epsilon^{-1/2}\sin k(\epsilon_r)L,$$

$$\psi_1'(L;\epsilon_r) = -\epsilon^{1/2}\sin k(\epsilon_r)L.$$

Thus at a complete resonance the transfer matrix must be of the form

$$M(\epsilon_r) = \begin{pmatrix} \cos k(\epsilon_r)L & \epsilon_r^{-1/2}\sin k(\epsilon_r)L \\ -\epsilon_r^{1/2}\sin k(\epsilon_r)L & \cos k(\epsilon_r)L \end{pmatrix}. \tag{25}$$

Conversely, for any $M(\epsilon)$ of the form (25), $\sigma(\epsilon) = 1$.

The same result can be obtained more simply if we note that at a complete resonance the vector

$$\begin{pmatrix} 1 \\ i\epsilon_r^{1/2} \end{pmatrix} \tag{26}$$

is an eigenvector of the *real* transfer matrix $M(\epsilon_r)$, belonging to an eigenvalue on the unit circle ($a = 0$ and $|c| = 1$).

For an arbitrarily shaped potential U Eqs. (24) need not be satisfied and complete resonances may be totally absent. However, for *symmetric* potentials, $U(x + L/2) = U(x - L/2)$, and one can prove that $\psi_1(L;\epsilon) = \psi_2'(L;\epsilon)$ (see Refs. 12 and 13). Then as the only nontrivial condition of the complete resonance at some band energy we have

$$\psi_1'(L;\epsilon_r)/\psi_1(L;\epsilon_r) = -\epsilon_r^{1/2}\tan k(\epsilon_r)L. \tag{27}$$

The rhs of Eq. (27) considered as a function of the wave vector assumes all values from the $(-\infty, +\infty)$ interval for $k \in (-\pi/2L, \pi/2L)$. We obtain the following lemma.

Lemma III: A symmetric barrier is completely transparent for at least one energy from each allowed energy band of the related periodic structure.

It is clear from the above considerations that at a complete resonance the transfer matrix need not be of the form $\pm I$, where I is the identity matrix. If it happens, however, the barrier is certainly completely transparent at this energy. A transfer matrix of this type may occur, on the other hand, only for the energies corresponding to the band edges on the energy axis, if additionally the matrix has *two* linearly independent eigenvectors [i.e., case of the Jordan form (17) is excluded]. Kramers,[13] in his analysis of the band structures, has shown that it may happen if and only if $|\mu(\epsilon)| = 1$ and $\partial\mu/\partial\epsilon = 0$ at the given energy, i.e., for the case of overlapping energy bands. He also pointed out that this type of situation will certainly happen if one works with the μ function defined not for the original unit cell but for a doubled one, and gave the relation between these two types of μ functions (which is nothing but the "logistic map" see Sec. V below). These observations are even more relevant in the tunneling problem, where the period doubling means working with subsequences of barriers composed of 2^k identical structural units (k-positive integer) and where $|\mu(\epsilon)| = 1$ and $\partial\mu/\partial\epsilon = 0$ simultaneously satisfied always lead to a complete resonance. It follows also that $\sigma(\epsilon) < 1$ at the "nonoverlapping" band edges, as matrix (18) is not of type (25).

V. FINITE SEQUENCES OF IDENTICAL BARRIERS

In this section we consider barriers composed of m identical structural units, each of them having the same potential shape and length L. In particular, we will study the behavior of the $\{\sigma_m(mL;\epsilon)\}$ sequences for fixed values of energy ϵ.

Let us consider first the case when the energy ϵ belongs to a forbidden energy interval. In accordance with intuitive expectations one may prove the following lemma.

Lemma IV: For any energy from a forbidden energy interval

$$\lim_{m\to\infty} \sigma_m(mL;\epsilon) = 0. \tag{28}$$

Moreover, sequences σ_m become monotonically decreasing for sufficiently large m.

Proof: Here we exploit the properties of the self-matching solutions in a forbidden energy interval where the transfer operator has real, distinct eigenvalues λ and λ^{-1}. For any energy from this interval there exist exactly two linearly independent self-matching solutions $\phi_1(x;\epsilon)$, $\phi_2(x;\epsilon)$, which can be continued on the $[0,mL]$ interval to give

$$\begin{aligned} &\phi_1(mL;\epsilon) = \lambda^m\phi_1(0;\epsilon), \\ &\phi_2(mL,\epsilon) = \lambda^{-m}\phi_2(0;\epsilon) \quad (|\lambda| > 1). \end{aligned} \tag{29}$$

Using representation (8) we obtain

$$\begin{aligned} &\psi_i^2(mL) = a_i\lambda^{2m} + b_i\lambda^{-2m} + c_i, \\ &\psi_i'^2(mL) = a_i'\lambda^{2m} + b_i'\lambda^{-2m} + c_i' \quad (i = 1,2), \end{aligned} \tag{30}$$

where primed and unprimed coefficients a_i, b_i, and c_i can be expressed in terms of $\phi_i(0,\epsilon)$ components and energy ϵ, and where non-negative coefficients a_1, a_1', a_2, a_2' are not all zero (otherwise using their explicit form one may show that at least one of the solutions ϕ_i would be trivial, which is impossible). Statement (28) follows then from Eq. (13). One can also easily show that a sequence $\alpha\lambda^{2m} + \beta\lambda^{-2m} + \gamma$, with $\alpha > 0$, $\beta > 0$, is monotonically increasing for sufficiently large m, which completes the proof. It essentially reduces to

the observation that a transfer matrix equivalent to

$$\begin{pmatrix} \lambda^m & 0 \\ 0 & \lambda^{-m} \end{pmatrix}$$

must have at least one term proportional to λ^m. This type of reasoning will be exploited below also, for energies from the allowed energy bands, where the behavior of $\{\sigma_m(mL;\epsilon)\}$ happens to be much more complex. Passing to the investigation of this case, we begin with an elementary remark.

Remark I: If an arbitrary barrier of length L is completely transparent at some energy $\epsilon=\epsilon_r$, then all finite barriers composed of its $m\geqslant 2$ replicas are also completely transparent at this energy. Indeed, if vector (26) is an eigenvector of the transfer matrix $M(\epsilon_r)$, it will also be an eigenvector of the $M^m(\epsilon_r)$ matrix. A constructive version of the proof starts with the transfer operator at complete resonance written as

$$\begin{aligned} M(\epsilon_r) &= \cos k(\epsilon_r)L\cdot I + i\sigma_x(-i\epsilon^{1/2})\sin k(\epsilon_r)L \\ &= \exp[i\sigma_x(-i\epsilon_r^{1/2})k(\epsilon_r)L], \end{aligned} \tag{31}$$

where the generalized Pauli matrix $\sigma_x(d)$ is defined as

$$\sigma_x(d):=\begin{pmatrix} 0 & d \\ d^{-1} & 0 \end{pmatrix}, \tag{32}$$

and still $\sigma_x^2(d)=I$. It follows that

$$M^m(\epsilon_r)=\exp[im\sigma_x(-i\epsilon_r^{1/2})k(\epsilon_r)L]$$

and hence $\sigma_m(\epsilon_r)=1$ for any $m\geqslant 2$.

As the next step, we investigate the $\{\sigma_m(mL;\epsilon)\}$ sequences at the end points of allowed energy intervals. If two such intervals have a common end point ϵ_c, $M(\epsilon_c)=\pm I$ (see above, Sec. II) and trivially $\sigma_m(mL;\epsilon_c)=1$ for all m. For an end point not overlapping with end points of other bands one has

$$M^m(\epsilon)\sim\begin{pmatrix} (\pm 1)^m & (\pm 1)^{m+1}m\gamma \\ 0 & (\pm 1)^m \end{pmatrix}$$

and $M^m(\epsilon)$ always contains an element linearly increasing with m. Thus $\lim_{m\to\infty}\sigma_m=0$ there.

For energies within a given energy band let us notice first that for all wave vectors of the form

$$k(\epsilon)=n\pi/mL, \quad n=1,2,\ldots,mL-1, \tag{33}$$

the transfer matrix is equivalent to

$$\begin{pmatrix} \exp(ik(\epsilon)L) & 0 \\ 0 & \exp(-ik(\epsilon)L) \end{pmatrix}. \tag{34}$$

Hence after m steps it will become equivalent and thus coinciding with the matrix $\pm I$. When rephrased this observation is equivalent to the following remark.

Remark II: For sufficiently large m a repeated barrier will become completely transparent for all energies corresponding to Bloch waves $\exp[ik(\epsilon)x]u_k(x)$ with the half-period $\pi/k(\epsilon)$ of the phase factor *commensurate* with the barrier length. All these energies form a countable, dense set in each energy band. In particular, discrete sets of energies obtained via imposing the Born–von Karman conditions belong to it. However, the measure of this set is zero in each energy band and it is interesting to investigate how sequences behave for all other allowed energies. We will see that they are *divergent* for almost all energies in each band. Here it is sufficient to consider the subsequences of $\{\sigma_m(mL;\epsilon)\}$, corresponding to the barrier size doubling at each step.

Recursion relations for transparencies: Let us consider a barrier composed of two identical structural units, each of them having the same potential shape and length L. Notice, first, that the simple algebraic identity $(\lambda^{-1}+\lambda)^2=\lambda^{-2}+\lambda^2-2$ can be generalized to an identity for an arbitrary unimodular matrix A,

$$\operatorname{Tr}A^2=(\operatorname{Tr}A)^2-2. \tag{35}$$

This, in turn, leads to the recursion relation for the Kramers functions μ

$$\mu(2L;\epsilon)=2\mu^2(L;\epsilon)-1, \tag{36}$$

or more generally, to

$$\begin{aligned} &\tilde{\mu}_{l+1}(\epsilon)=2\tilde{\mu}_l^2(\epsilon)-1, \\ &\tilde{\mu}_l(\epsilon):=\mu(2^{l-1}L;\epsilon), \quad l=1,2,\ldots, \end{aligned} \tag{37}$$

if we continue the doubling procedure.

Interesting enough, the recursion relation can also be obtained for the correspondingly defined functions $\tilde{\nu}_l(\epsilon)=\nu(2^{l-1}L;\epsilon)$. Indeed, for any 2×2 matrix we have

$$(A^2)_{12}=A_{12}\operatorname{Tr}A, \quad (A^2)_{21}=A_{21}\operatorname{Tr}A, \tag{38}$$

which immediately leads to

$$\tilde{\nu}_{l+1}(\epsilon)=2\tilde{\mu}_l(\epsilon)\tilde{\nu}_l(\epsilon). \tag{39}$$

A straightforward algebra with σ_l defined as $(\tilde{\mu}_l^2+\tilde{\nu}_l^2)^{-1}$ gives the following lemma.

Lemma V: For a size-doubling sequence of barriers, the barrier transparencies can be calculated from the two-variable iterative map

$$\tilde{\sigma}_{l+1}(\epsilon)=\tilde{\sigma}_l(\epsilon)/[4\tilde{\mu}_l^2(\epsilon)(1-\tilde{\sigma}_l(\epsilon))+\tilde{\sigma}_l(\epsilon)], \tag{40}$$

$$\tilde{\mu}_{l+1}(\epsilon)=2\tilde{\mu}_l^2(\epsilon)-1, \tag{41}$$

assuming that $\tilde{\mu}_1(\epsilon):=\mu(\epsilon)$ (the "band structure") and $\tilde{\sigma}_1(\epsilon):=\sigma(\epsilon)$ (single barrier transparency) are known.

For arbitrary finite structure, the transparency is always greater than zero for all $\epsilon>0$. (This follows from the expression for σ used in the proof of Lemma I.) For positive energies one may then define the resistances $\tilde{\rho}_l(\epsilon)$ as

$$\tilde{\rho}_l(\epsilon):=\tilde{\sigma}_l^{-1}(\epsilon)-1, \tag{42}$$

being just the ratio of the reflection to transmission coefficients. In terms of the $\tilde{\rho},\mu$ variables the map (40) and (41) simplifies to

$$\tilde{\rho}_{l+1}(\epsilon)=4\tilde{\mu}_l^2(\epsilon)\tilde{\rho}_l(\epsilon), \tag{43}$$

$$\tilde{\mu}_{l+1}(\epsilon)=2\tilde{\mu}_l^2(\epsilon)-1. \tag{44}$$

Previous general statements can be easily checked here: $\sigma=1$ is an obvious fixed point of the map (40) (compare Remark 1) and in any forbidden energy interval where $|\mu_1|>1$ the map for μ along obviously diverges; Eqs. (43) and (44) show then that ρ_l tends monotonically to infinity for $l\to\infty$ (σ_l tends monotonically to zero—compare Lemma IV). At the band end points with $\sigma_1<1$ ($\rho_1\neq 0$) we have $\rho_l\to+\infty$.

To investigate the $\{\sigma_l\}$ sequences for all other, internal band energies we note that the transformation (37) coin-

cides, after the linear transformation $\mu_l = -2x_l + 1$, with the logistic map in its canonical form: $x_{l+1} = qx_l(1-x_l)$ for $q = 4$. Hence it exhibits chaotic and ergodic behavior on the [0,1] interval. For a given value of q, it is also a doubling transformation in terms of the angular variable θ defined by

$$\mu(\epsilon) = \cos\theta(\epsilon), \quad \theta\in[0,\pi]. \tag{45}$$

So defined, θ is a single-valued function of energy in each band, which coincides in the $L = 1$ length scale with the nonnegative quasimomentum in the reduced-zone scheme. We have

$$\tilde{\mu}_l(\epsilon) = \cos 2^l\theta(\epsilon). \tag{46}$$

Maps (40) and (41) or, equivalently, (43) and (44) in angular parametrization (46) allow reconstruction of all complete resonances at internal band energies occurring in barriers composed of 2^{l-1} structural units. We already know that they have to occur for all θ of the form

$$n\pi/2^{l-1}, \quad n = 1,2,...,2^{l-1}-1. \tag{47}$$

The same result may be obtained by noticing that $\tilde{\mu}_{k-1} = 0$ necessarily leads to $\tilde{\sigma}_k = 1$. Moreover, all angles $\theta\in(0,\pi)$, leading to $\tilde{\mu}_{k-1} = 0$ under the transformation (46), are given by

$$\theta_{s,k} = (2s+1)\pi/2^{k-1}, \quad s = 0,1,...,2^{k-2}-1. \tag{48}$$

It is clear that the sum of the sets (48) coincides with the set (47), as each element of the set (47) can be written $\theta_{s,k}$ for some $k < L$ and vice versa.

A more interesting observation is related to the existence of the other (besides $\mu = -1$) fixed point of the map (44), $\mu = -\frac{1}{2}$. Sets of all $\theta\in(0,\pi)$ that will give $\tilde{\mu}_k(\theta) = -\frac{1}{2}$ for some k have elements

$$\theta^{(1)}_{s,k} = 2^{1-k}(2s+2/3)\pi, \quad \theta^{(2)}_{s,k} = 2^{1-k}(2s+4/3)\pi,$$
$$s = 0,1,...,2^{k-2}-1. \tag{49}$$

For all such θ the resistance will become frozen at its $\tilde{\rho}_{k-1}(\epsilon(\theta))$ value equal to

$$\tilde{\rho}_{k-1}(\epsilon(\theta)) = \prod_{s=1}^{k-2} 2\tilde{\mu}_s(\epsilon(\theta))^2\rho(\epsilon(\theta)). \tag{50}$$

Once again, the sum of all sets (47) is a countable, dense set on $[0,\pi]$. The transparencies for the corresponding energies will preserve some fixed value, smaller than unity, for all sufficiently long doubled barriers.

Finally, it is not difficult to prove, using a representation like (50), that $\tilde{\rho}$ sequences are divergent for almost all energies from a given band. Indeed, the infinite product

$$\prod_{s=1}^{\infty} (2\tilde{\mu}_s(\epsilon))^2 \tag{51}$$

(where we can now assume that all $\tilde{\mu}_s \neq 0$) diverges for almost all ϵ. This is because the sequence $\{\ln|2\mu_s(\epsilon)|\}$ does not converge to zero for almost all energies, due to the ergodic character of the map (36).

To summarize, for almost all energies from an allowed energy band $\{\sigma_m(mL;\epsilon)\}$ sequences will have an erratic, oscillatory behavior for increasing m, in contrast to naive expectations. However, the "amplitude" of these oscillations should, in general, decrease with energy for bounded potentials (compare Sec. II C).

VI. EXAMPLES, NUMERICAL ILLUSTRATIONS, AND FINAL REMARKS

A. Rectangular barriers

For the simplest "rectangular" barrier with constant potential U on $[0,L]$ the generator G does not depend on x:

$$G(\epsilon) = \begin{pmatrix} 0 & 1 \\ \kappa^2 & 0 \end{pmatrix}, \quad \kappa^2 := U - \epsilon, \tag{52}$$

and Eq. (3) can be integrated leading to

$$M(x,x_0;\epsilon) = \exp[G(\epsilon)(x-x_0)]. \tag{53}$$

Writing $G(\epsilon)$ as

$$G(\epsilon) = \kappa\sigma_x(\kappa) \tag{54}$$

with, as before, $\sigma_x(\kappa)$ belonging to the equivalence class of the Pauli matrix σ_x,

$$\sigma_x(\kappa) := \begin{pmatrix} 0 & \kappa^{-1} \\ \kappa & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \kappa^{1/2} & 0 \\ 0 & \kappa^{-1/2} \end{pmatrix} \sigma_x \begin{pmatrix} \kappa^{-1/2} & 0 \\ 0 & \kappa^{1/2} \end{pmatrix},$$
$$\sigma_x^2(\kappa) = I, \tag{55}$$

we obtain

$$M(x,x_0;\epsilon) = \exp(\kappa(x-x_0)\sigma_x(\kappa))$$
$$= \cosh\kappa(x-x_0)\cdot I + \sigma_x(\kappa)\sinh\kappa(x-x_0). \tag{56}$$

The fundamental solutions can be read from (56):

$$\psi_1(x,x_0;\epsilon) = \cosh\kappa(x-x_0),$$
$$\psi_2(x,x_0;\epsilon) = \sinh\kappa(x-x_0). \tag{57}$$

The unimodularity property reduces here to the simple trigonometric identity, which, conversely, may be considered as a very special case of Wronksian-type identities for pairs of fundamental solutions of second-order, linear differential equations. The relationship between the unimodularity and the traceless character of the generator G is particularly clear here, as for any matrix A, $\det A = \exp(\mathrm{Tr}\,A)$.

One can immediately write the transfer operator for a finite array of rectangular potential barriers, i.e., for a barrier with piecewise constant potential U: $U = U_k$ for $x\in(x_{j-1}, x_k)$, $0 < x_0 < x_1,...,x_{n-1} < x_n = L$. Continuously extending potentials U_k onto closed intervals $[x_{k-1}, x_k]$ we have

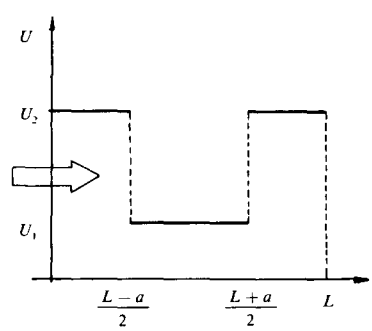


FIG. 3. Symmetric double barrier. Here $\kappa_i^2 := U_i - \epsilon$, $i = 1,2$.

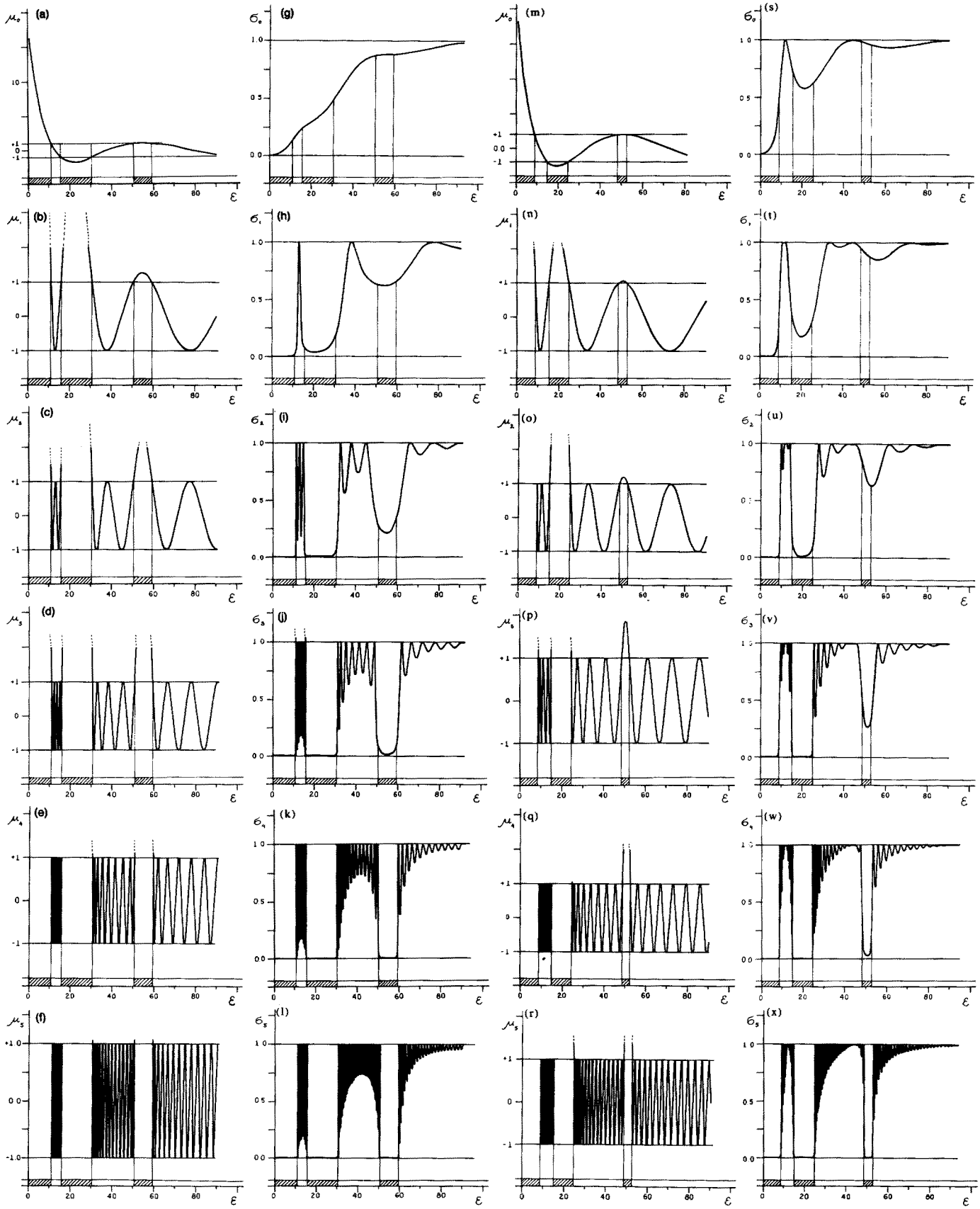


FIG. 4. Solving numerically the Cauchy problem on $[0,L]$ one can find fundamental solutions ψ_1 and ψ_2 and determine $\mu(\epsilon)$ and $\nu(\epsilon)$. Then $\tilde{u}_l$ and $\tilde{\sigma}_l$ can be found using recurrence relations (40). (a)–(l) numerically calculated $\tilde{\mu}_l$, $\tilde{\sigma}_l$ ($l=0,1,...,5$) for the asymmetric model potential (61) ($x_0=0.7$). Note that each zero of the $\tilde{\mu}_l$ function generates one complete resonance after period doubling [compare Eq. (40)]. At given energy scale fast $\tilde{\mu}_l$ oscillations soon cause the loss of graphical resolution in the first narrow energy band and then close to the band edges in higher-energy bands (shadowed regions). Here σ_0 does not have much structure for a single asymmetric barrier and complete resonance are absent in the considered energy intervals. However, the σ_5 plot (for a barrier composed of 64 structural units) allows one to localize, quite precisely, the band-edge energies and shows the peculiar behavior inside each band. (m)–(x) the same for the symmetric potential (61). Note the presence of a complete resonance in each indicated energy band (Lemma III) for a single-unit barrier.

$$M(L;\epsilon)=\prod_{k-1}^{N} M_k(l_k;\epsilon),\quad M_k(l_k;\epsilon_k)$$
$$=\cosh\kappa_k l_k+\sigma_x(\kappa_k)\sinh\kappa_k l_k,$$
$$l_k:=x_k-x_{k-1},\quad \kappa_k:=(U_k-\epsilon)^{1/2}. \tag{58}$$

Products of this type can be used to approximate uniformly transfer operators for arbitrary potential. This also furnishes an independent proof of the unimodularity property of transfer operators for traceless generators.

Representation (58) reduces to a minimum the algebraic calculations for rectangular barriers (still rather lengthy in most quantum-mechanical textbooks), and may be used in obtaining systematic density expansions for disordered barriers of the A, B, etc. type. As a simple illustration let us consider a symmetric double barrier structure (Fig. 3). Just multiplying three binomials

$$(\cosh\kappa_2[(L-a)/2]+\sigma_x(\kappa_2)\sinh\kappa_2[(L-a)/2])$$
$$\times(\cosh\kappa_1 a+\sigma_x(\kappa_1)\sinh\kappa_1 a)$$
$$\times(\cosh\kappa_2[(L-a)/2]+\sigma_x(\kappa_2)\sinh\kappa_2[(L-a)/2]),$$

we have

$$M_{11}(\epsilon)=M_{22}(\epsilon)=\mu(\epsilon)=\cosh\kappa_1 a\cosh\kappa_2(L-a)$$
$$+\tfrac{1}{2}(\kappa_1/\kappa_2+\kappa_2/\kappa_1)\sinh\kappa_1 a\sinh\kappa_2(L-a) \tag{59}$$

and

$$M_{12}(\epsilon)=\kappa_2^{-1}\Big[\cosh\kappa_1 a\sinh\kappa_2(L-a)+\sinh\kappa_1 a$$
$$\times\Big(\frac{\kappa_1}{\kappa_2}\sinh^2\kappa_1\frac{L-a}{2}+\frac{\kappa_2}{\kappa_1}\cosh^2\kappa_1\frac{L-a}{2}\Big)\Big],$$
$$M_{21}(\epsilon)=\kappa_2\Big[\cosh\kappa_1 a\sinh\kappa_2(L-a)+\sinh\kappa_1 a$$
$$\times\Big(\frac{\kappa_2}{\kappa_1}\sinh^2\kappa_2\frac{L-a}{2}+\frac{\kappa_1}{\kappa_2}\cosh^2\kappa_2\frac{L-a}{2}\Big)\Big]. \tag{60}$$

Explicitly known matrix elements of the transfer operator allow one to write an expression for the transparency of a double barrier [Eq. (13)]. The equation $\mu(\epsilon)=\cos k(\epsilon)L$ with $\mu(\epsilon)$ given by (59) coincides, of course, with the implicit dispersion relation of the Kronig–Penney model (with the U_2 barriers of length $L-a$ separated by U_1 wells of width a).

Another almost trivial application is a kind of Ramsauer effect for finite disordered sequences of rectangular barriers of two types, where tunneling phenomena depend on four parameters: L_A, L_B and κ_A, κ_B. From representations (59) and (60) it is obvious that a *disordered* barrier will behave as a uniform, shorter barrier of the A type for all energies such that $\kappa_A L_A=in\pi$.

B. Numerical example

For graphical and numerical illustration we have chosen a barrier of the form

$$U(x)=a(x-x_0)^2+b,\quad x\in[0,1], \tag{61}$$

with parameters $a=100$, $b=2$, and $x_0=0.7$ and 0.5 (for asymmetric and symmetric barriers, respectively). Results for single and double barriers (up to 2^5 structural units) are presented in Figs. 4 and 5.

C. Final remarks

The above-presented band-type characterization of transparencies for single barriers and for sequences of identical barriers may have some practical application in electron spectroscopy and in designing quantum tunneling devices. Here we stress again the localization of complete resonances in symmetric barriers (Lemma III), and erratic transparency behavior in allowed energy bands.

The numerical examples clearly show that chaotic transmission fluctuations have some lower-bound envelope, most likely analytic in each energy band. It will be interesting to investigate this problem closer.

All described phenomena should be common for other systems with similar mathematics such as transmission lines or layers of inhomogeneous dielectric transmitting electromagnetic waves. The "perfect transparency" Lemma III can also be used there. In particular, the wave equation in a layer region with symmetric permittivity profile can be trans-

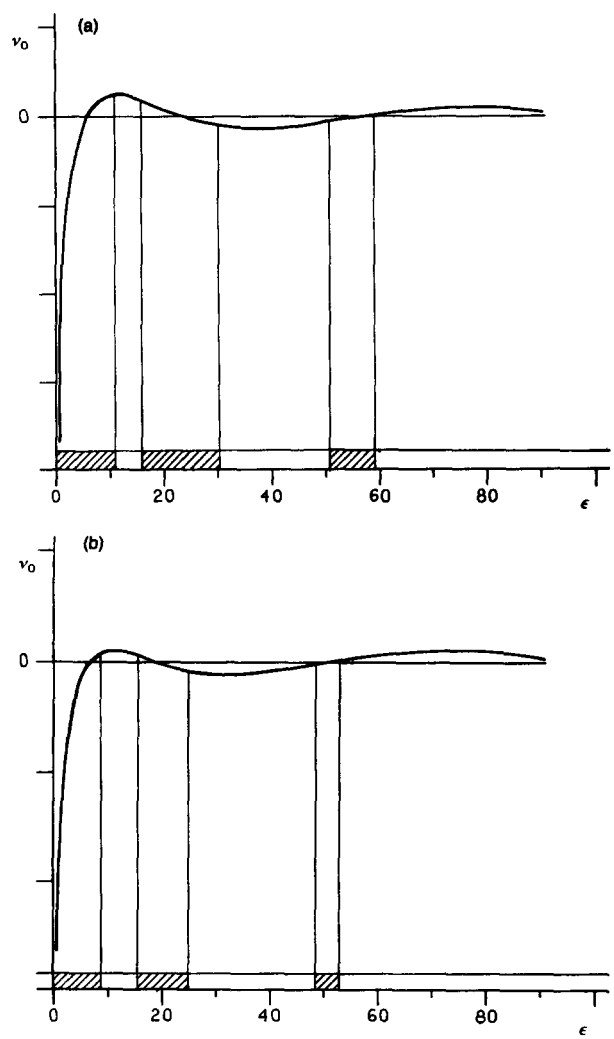


FIG. 5. Here ν functions are for a single barrier (61): (a) asymmetric case, (b) symmetric case.

formed into a Schrödinger-like equation with an effective symmetric potential. This problem and some related questions (the case of an oblique incidence, curvilinear dielectric layers) will be considered in a separate paper.

Realistic sequences of barriers can hardly be considered as made of identical units. Therefore, it will be interesting to investigate the influence of some barrier shape or energy noise, e.g., on the two variable map (43) and (44). Random effects are usually studied in the $L \to \infty$ limit. They are certainly important, however, for finite microscopic layers and lead to interesting mathematics. As a preliminary result in this direction we can announce another simple lemma stating that the potential noise about some given average potential U can only decrease the average barrier transparency.

ACKNOWLEDGMENTS

We thank Dr. W. R. Frensley and Dr. M. A. Reed from Texas Instruments for an informative discussion. Family members were also engaged—Miss K. Kowalski did the drawings proving once again human superiority over (some) automatic plotters.

This work was partially supported by R. A. Welch Grant No. Y-707.

[1]E. L. Wolf, *Principles of Electron Tunneling Spectroscopy* (Oxford U. P., New York, 1985).
[2]I. M. Lifshitz and V. Ya. Kirpichenkov, Sov. Phys. JETP **50** (3), 499 (1979). See also I. M. Lifshitz, S. A. Gredeskul, and L. A. Pastur, *Introduction into Theory of Disordered Systems* (Nauka, Moscow, 1982), Chap. VII (in Russian).
[3]A. O. Caldeira and A. J. Leggett, Ann. Phys. (NY) **149**, 374 (1983); **153**, 445(E) (1984).
[4]A. J. Leggett, S. Chakravarty, A. T. Dorsey, M. P. A. Fisher, A. Garg, and W. Zwerger, "Dynamics of the dissipative two-state systems," preprint, Univ. of Illinois at Urbana-Champaign, P/85/10/185, 1985.
[5]B. Ricco and M. Ya. Azbel, Phys. Rev. B **29**, 1970 (1984).
[6]A. D. Stone, M. Ya. Azbel, and P. A. Lee, Phys. Rev. B **13**, 1707 (1985).
[7]E. Cota, J. V. Jose, and M. Ya. Azbel, Phys. Rev. B **32**, 6157 (1985).
[8]R. L. Greene and K. K. Bajaj, Phys. Rev. B **31**, 913 (1985).
[9]A. Peres, J. Math. Phys. **24**, 1110 (1983).
[10]See, e.g., M. Schechter, *Operator Methods in Quantum Mechanics* (North-Holland, Amsterdam, 1981); H. Kalf, "Schroedinger-type operators with continuous spectra," in *Research Notes in Mathematics*, Vol. 65 (Pitman, Bolton, 1982).
[11]H. A. Kramers, Physica **2**, 483 (1935).
[12]H. M. James, Phys. Rev. **76**, 1602 (1949).
[13]W. Kohn, Phys. Rev. **115**, 809 (1959).
[14]E. A. Pshenichnov, Sov. Phys. Solid State **4** (5), 819 (1962).
[15]For $i = 2$ one should integrate in the $[\delta, x]$ interval and then take the limit $\delta \to 0+$ noticing that

$$\lim_{\delta \to 0+} \frac{\tilde{\psi}_2(x)}{\psi_2(x)} = \lim_{\delta \to 0+} \frac{\tilde{\psi}_2'(x)}{\psi_2'(x)} = 1.$$

Inverse scattering for standing wave solutions of the Schrödinger equation

Erkki Somersalo[a)]
Physics Department, Indiana University, Bloomington, Indiana 47405

(Received 12 February 1987; accepted for publication 10 June 1987)

In this paper a reformulation of the inverse scattering theory for a three-dimensional Schrödinger equation is given in terms of the standing wave solutions and the reactance matrix. A counterpart of the generalized Marchenko equation is given as well as a trace-type formula for the potential. Derivations are based on the $\bar{\partial}$ method in the complex k plane.

I. INTRODUCTION

The standard formulation of the inverse scattering problem for the three-dimensional Schrödinger equation is to reconstruct the underlying scattering potential from the asymptotic amplitude of the outgoing scattered wave. As was shown by Newton,[1] a generalized Marchenko equation can be derived from the fact that the outgoing and incoming solutions of the Schrödinger operator are related to each other via the scattering amplitude.

The aim of the present paper is to show that the inverse scattering theory can also be formulated in terms of the so-called standing wave solutions and the corresponding reactance matrix. The direct scattering theory for standing waves was developed by Kouri and Levin.[2] (See also Newton[3] and references therein.) The main tool in this paper will be the $\bar{\partial}$ equations, introduced to scattering theory by Beals and Coifman[4,5] and later successfully developed by Ablowitz and Nachman.[6,7]

In Sec. II, we shall summarize some results from the standing wave scattering theory. In Sec. III, the main results are given. We derive a counterpart of the generalized Marchenko equation and give a reconstruction formula for the potential from the standing wave solution and the K matrix. Finally, the necessary $\bar{\partial}$ equations are briefly summarized in the Appendix.

II. STANDING WAVE SOLUTIONS

To fix the notations, we shall consider the Schrödinger equation

$$(-\Delta + V)\psi = K^2\psi,$$

in $\mathbb{R}^3$. Generalizations to other dimensions are straightforward. In this paper K means complex wave number and k its real part, i.e., $K = k + iq \in \mathbb{C}$, $k,q \in \mathbb{R}$. The scattering potential V is real and assumed to satisfy

$$\|V\|_{L^2} + \sup_{x\in\mathbb{R}^3} \int_{\mathbb{R}^3} \left[\frac{|x| + |y| + a}{|x-y|}\right]^2 |V(y)|dy < \infty, \tag{2.1}$$

for some constant $a > 0$. This class of potentials was introduced by Newton.[8]

The principal value Green's function G^p is defined through the formula

$$G^p(x,k) = \left(\frac{1}{2\pi}\right)^3 \text{P.V.} \int_{\mathbb{R}^3} \frac{e^{ix\cdot\zeta}}{k^2 - |\zeta|^2} d\zeta$$

$$= \frac{1}{2}\left(\frac{1}{2\pi}\right)^3 \int_{S^2} d\hat{\theta} \fint_{-\infty}^{\infty} \frac{t}{k-t} e^{it\hat{\theta}\cdot x} dt, \tag{2.2}$$

where $k \in \mathbb{R}$ and $\zeta = |\zeta|\hat{\theta} \in \mathbb{R}^3$, and $\fint$ stands for the Cauchy principal value integral. It is very well known[3] that

$$G^p(x,k) = G^{\pm}(x,k) \pm i\Delta(x,k), \tag{2.3}$$

where $G^{\pm}$ are outgoing (+) and incoming (−) Green's functions, respectively, i.e.,

$$G^{\pm}(x,k) = \left(\frac{1}{2\pi}\right)^3 \int_{\mathbb{R}^3} \frac{e^{ix\cdot\zeta}}{k^2 \pm i0 - |\zeta|^2} d\zeta$$

$$= -\frac{1}{4\pi} \frac{e^{\pm ik|x|}}{|x|},$$

and $\Delta(x,k)$ is given by

$$\Delta(x,k) = \operatorname{sgn} k\pi \left(\frac{1}{2\pi}\right)^3 \int_{\mathbb{R}^3} e^{ix\zeta} \delta(|\zeta|^2 - k^2) d\zeta$$

$$= \left(\frac{1}{4\pi}\right)^2 k \int_{S^2} e^{ik\hat{\theta}\cdot x} d\hat{\theta}. \tag{2.4}$$

Formula (2.2) gives also the simple relation

$$G^p(x,k) = \tfrac{1}{2}(G^+(x,k) + G^-(x,k)) \tag{2.5}$$

between Green's functions.

To discuss the principal value Green's function further, we recall some basic facts from the standard scattering theory.

Let $\psi^+(x,K,\hat{\theta})$ denote the outgoing scattering solution of the Schrödinger equation with plane wave incidence, i.e., ψ^+ satisfies the Lippman–Schwinger equation

$$\psi^+(x,K,\hat{\theta}) = \psi_0(x,K,\hat{\theta}) + \mathcal{G}^+(V\psi^+)(x,K,\theta). \tag{2.6}$$

Here $\psi_0(x,K,\hat{\theta}) = \exp(iK\theta\cdot x)$ denotes the plane wave, and $\mathcal{G}^+$ is the convolution operator with kernel G^+. The solvability of (2.6) has been discussed by several authors under various assumptions on the potential V. (See, e.g., Refs. 8–10.) Here we shall refer to Ref. 8, where it was shown that under assumption (2.1), Eq. (2.6) has a unique solution for all $K \in \overline{\mathbb{C}}^+ = \{K \in \mathbb{C} \mid \operatorname{Im} K \geqslant 0\}$ except possibly for a finite number of values on the imaginary axis, $K = i\varkappa_1, \ldots, i\varkappa_m$, $0 \leqslant \varkappa_1 < \cdots < \varkappa_m$. Throughout the paper, we shall assume that $K = 0$ is nonexceptional, i.e., $0 < \varkappa_1$.

Similarly, the incoming scattering solution $\psi^-(x,K,\theta)$, $K \in \overline{\mathbb{C}}^-$, is defined in terms of G^-.

For fixed $k \in \mathbb{R}$, the scattering operator $\mathcal{S}$: $L^2(S^2) \to L^2(S^2)$ is defined by

a) Permanent address: Department of Mathematics, University of Helsinki, Hallituskatu 15, 00100 Helsinki, Finland.

 0022-2488/87/102416-04$02.50 © 1987 American Institute of Physics

$$\mathcal{S}(\psi)(\hat{\theta}) = \psi(\hat{\theta}) + \frac{ik}{2\pi}\int_{S^2} A(k,\hat{\theta},\hat{\theta}')\psi(\hat{\theta}')d\hat{\theta}', \tag{2.7}$$

where the kernel $A(k,\hat{\theta},\hat{\theta}')$ is obtained as the scattering amplitude of the outgoing wave ψ^+:

$$A(k,\hat{\theta},\hat{\theta}') = -\frac{1}{4\pi}\int_{S^2} e^{-ik\hat{\theta}\cdot x} V(x)\psi^+(x,k,\hat{\theta}')dx. \tag{2.8}$$

The following theorem relates the existence of the standing wave solutions of the Schrödinger equation to the properties of the S operator and the underlying potential.

We shall use the notation $\|V\|_R$ for the Rollnik norm of V, i.e.,

$$\|V\|_R^2 = \int_{R^6} \frac{|V(x)|\,|V(y)|}{|x-y|^2}\,dx\,dy.$$

Further, we use the notation $\mathcal{G}^p(\phi)(x) = \int_{R^3} G^p(x-y,k)\phi(y)dy$, and $\psi_0(x,k,\hat{\theta}) = \exp(ik\hat{\theta}\cdot x)$ denotes the plane wave.

Theorem 2.1: The equation

$$\psi^p(x,k,\hat{\theta}) = \psi_0(x,k,\hat{\theta}) + \mathcal{G}^p(V\psi^p)(x,k,\hat{\theta}) \tag{2.9}$$

has a unique solution ψ^p if the potential satisfies (2.1) and the corresponding scattering operator does not have the eigenvalue -1. Especially, a unique solution exists if one of the following conditions is fulfilled: (i) $V\in L_1\cap R$, R denoting the Rollnik class and $\|V\|_R < 4\pi$; (ii) $|V(x)|\leqslant C(1+|x|)^{-\mu}$ for some constants $C>0$ and $\mu>1$, and $|k|$ is large enough.

Proof: Factorizing first the right-hand side of Eq. (2.3) one gets

$$1 - \mathcal{G}^p V = 1 - \mathcal{G}^+ V - i\Delta V$$
$$= (1 - \mathcal{G}^+ V)(1 + (ik/4\pi)\mathcal{M}),$$

where $\mathcal{M}$ is an integral operator with kernel

$$M(x,y,k) = -\frac{1}{4\pi}\int_{S^2} \psi^+(x,k,\hat{\theta})e^{-ik\hat{\theta}\cdot y} V(y)d\hat{\theta}.$$

Following the argument in Ref. 8 we notice that

$$\mathrm{Tr}\,\mathcal{M}^n = \mathrm{Tr}\,\mathcal{A}^n,$$

$\mathcal{A}$ being the integral operator $L^2(S^2)\to L^2(S^2)$ with the kernel $A(k,\hat{\theta},\hat{\theta}')$. Hence, using the product formula for modified Fredholm determinants,[3] denoted as $\det_2$, we have

$$\det_2(1-\mathcal{G}^p V) = \det_2(1-\mathcal{G}^+V)\det(1+(ik/4\pi)\mathcal{M})$$
$$\times\exp(-\mathrm{Tr}\,(1-\mathcal{G}^+V)(ik/4\pi)\mathcal{M})$$
$$= \det_2(1-\mathcal{G}^+V)\det(1+(ik/4\pi)\mathcal{A})$$
$$\times\exp\left(\frac{ik}{4\pi}\int_{R^3} V(x)dx\right),$$

which proves the first part of the theorem.

The special cases (i) and (ii) are treated as follows: In the standard way, Eq. (2.9) is transformed to a Fredholm equation by multiplication with $|V(x)|^{1/2}$. Result (i) follows by Theorem XI 43 in Ref. 9 and representation (2.5). Result (ii) is a consequence of a theorem of Saitō[9]: The operators $\mathcal{G}^\pm$ are compact operators

$$\mathcal{G}^\pm: L^2_\delta(\mathbf{R}^3)\to L^2_{-\delta}(\mathbf{R}^3),\quad \tfrac{1}{2}<\delta\leqslant 1,$$

with norm less than $C/|k|$ for large $|k|$. Here

$$L^2_\delta(\mathbf{R}^3) = \left\{ f\in L^2_{\mathrm{loc}}(\mathbf{R}^n)\,\middle|\, \int |f(x)|^2(1+|x|)^{2\delta}dx < \infty\right\}.$$

We shall skip the details of the proof. □

The solution ψ^p is called the standing wave solution. The connection between ψ^p and ψ^+ is given in terms of the scattering amplitude. We have

$$\psi^p(x,k,\hat{\theta}) = \psi^+(x,k,\hat{\theta}) - \frac{ik}{4\pi}\int_{S^2}\psi^p(x,k,\hat{\theta}')A(k,\hat{\theta}',\hat{\theta})d\hat{\theta}'. \tag{2.10}$$

This result can be found in Refs. 2 and 3.

Defining the reactance matrix $K(k,\hat{\theta},\hat{\theta}')$ by the formula

$$K(k,\hat{\theta},\hat{\theta}') = -\frac{1}{4\pi}\int e^{-ik\hat{\theta}\cdot x}V(x)\psi^p(x,k,\hat{\theta}')dx,$$

we also have

$$\psi^+(x,k,\hat{\theta}) = \psi^p(x,k,\hat{\theta}) + \frac{ik}{4\pi}\int_{S^2}\psi^+(x,k,\hat{\theta}')K(k,\hat{\theta}',\hat{\theta})d\hat{\theta}'. \tag{2.11}$$

Integrating (2.10) with $e^{-ik\hat{\theta}\cdot x}V(x)$ we arrive at Heitler's integral equation

$$K(k,\hat{\theta},\hat{\theta}') = A(k,\hat{\theta},\hat{\theta}') - \frac{ik}{4\pi}\int_{S^2} K(k,\hat{\theta},\hat{\theta}'')A(k,\hat{\theta}'',\hat{\theta}')d\hat{\theta}''.$$

This equation takes a simple form, if written in terms of the scattering operator and the reactance operator $\mathcal{K}$: $L^2(S^2)\to L^2(S^2)$,

$$\mathcal{K}\phi(\hat{\theta}) = -\frac{ik}{4\pi}\int_{S^2} K(k,\hat{\theta},\hat{\theta}')\phi(\hat{\theta}')d\hat{\theta}'.$$

We have

$$\mathcal{K}(1+\mathcal{S}) = (1-\mathcal{S}),$$

i.e., $\mathcal{K}$ is the Cayley transformation of the unitary operator $\mathcal{S}$:

$$\mathcal{K} = (1-\mathcal{S})(1+\mathcal{S})^{-1}.$$

Remark 2.2: The above discussion gives the following denseness result as a corollary: If $\{\psi_n\,|n\in\mathbf{N}\}$ is a dense set in $L^2(S^2)$ and assumptions (i) or (ii) of Theorem 2.1 hold, then $\{\psi_n + \mathcal{S}\psi_n\,|n\in\mathbf{N}\}$ also form a dense set. This type of denseness result may have some significance in fixed energy inversion schemes.

III. THE $\bar{\partial}$ EQUATIONS

The definition and a brief summary of the $\bar{\partial}$ operator in $\mathbf{C}$ is given in the Appendix.

We start by defining the Green's function for $K = k + iq\in\mathbf{C}$ by the formula

$$G(x,K) = \left(\frac{1}{2\pi}\right)^3\int_{R^3}\frac{e^{ix\cdot\zeta}}{K^2-|\zeta|^2}d\zeta.$$

Obviously, $G(x,K) = G^{\pm}(x,K)$ for $K \in \mathbb{C}^{\pm} = \{K = k + iq | \pm q > 0\}$ and $G(x,K) = G^{P}(x,k)$ as $q = 0$.

Using Eq. (A2) we get

$$\begin{aligned}\bar{\partial} G(x,K) &:= \frac{\partial}{\partial \bar{K}} G(x,K) \\ &= \left(\frac{1}{2\pi}\right)^3 \int_{\mathbb{R}^3} \frac{d\zeta}{2|\zeta|} \bar{\partial}\left(\frac{1}{K-|\zeta|} - \frac{1}{K+|\zeta|}\right) e^{ix\cdot\zeta} \\ &= \left(\frac{1}{4\pi}\right)^2 \int_{\mathbb{R}^3} \frac{d\zeta}{|\zeta|} (\delta_{\mathbb{C}}(K-|\zeta|) \\ &\quad - \delta_{\mathbb{C}}(K+|\zeta|)) e^{ix\cdot\zeta} = \delta(q)\Delta(x,k),\end{aligned}$$

where $\Delta(x,k)$ is defined in (2.4). This is simply a manifestation of the fact that $\bar{\partial}$ of a sectionally holomorphic function results in a density $i/2$ times the jump of the function along the cut.

Next, let $\psi(x,K,\hat{\theta})$ be the solution of the equation

$$\psi = \psi_0 + \mathcal{G}(V\psi), \tag{3.1}$$

where we assume that V is a potential that allows a unique solution of the above equation for all $K \in \mathbb{C}$. The bound states are especially excluded. Clearly, $\psi = \psi^{\pm}$ for $K \in \mathbb{C}^{\pm}$, $\psi^{\pm}$ denoting the outgoing and incoming solutions, respectively, and $\psi = \psi^{P}$ for K real. Applying $\bar{\partial}$ on both sides of the equation one gets

$$\begin{aligned}&\bar{\partial}\psi(x,K,\hat{\theta}) - \mathcal{G}(V\bar{\partial}\psi)(x,K,\hat{\theta}) \\ &\quad = \bar{\partial}\mathcal{G}(V\psi)(x,K,\hat{\theta}) \\ &\quad = \delta(q)\left(\frac{1}{4\pi}\right)^2 k \int_{S^2} d\hat{\theta}' \int_{\mathbb{R}^3} dy\, e^{2k\hat{\theta}'\cdot(x-y)} V(y)\psi(y,k,\hat{\theta}) \\ &\quad = -\delta(q)\frac{k}{4\pi}\int_{S^2} e^{ik\hat{\theta}'\cdot x} K(k,\hat{\theta}',\hat{\theta})\, d\hat{\theta}',\end{aligned}$$

which yields, by the assumption of V,

$$\bar{\partial}\psi(x,K,\hat{\theta}) = -\delta(q)\frac{k}{4\pi}\int_{S^2} \psi^{P}(x,k,\hat{\theta}')K(k,\hat{\theta}',\hat{\theta})\, d\hat{\theta}'.$$

To reconstruct ψ we need information of the large $|K|$ behavior of the solution ψ. Let us assume that the potential V satisfies

$$\int_{\mathbb{R}^3} \left| \frac{V(x+y)}{1+\hat{\theta}\cdot y} \right|^2 dy < \infty \tag{3.2}$$

for all $x \in \mathbb{R}^3$, $\hat{\theta} \in S^2$. Then, with a minor modification of the proof of Newton,[11] we see that the mappings $K \to e^{-iK\hat{\theta}\cdot x}\psi^{\pm}(x,K,\theta) - 1$ are in $H^2(\mathbb{C}^{\pm})$, the Hardy classes of the upper and lower half-spaces, respectively. [In Ref. 12, it is shown that $e^{-iK\hat{\theta}\cdot x}\psi(x,K,\hat{\theta}) - 1$ behaves like $(1/|K|)$ under more restrictive conditions on V than (3.2).] Therefore we can use formula (A1) with $\Omega = \{|K| < R\}$ with $R \to \infty$. However, one has to be careful on the real axis, since the reconstruction formula holds originally only for $C^1(\bar{\Omega})$ functions. We shall denote by $\chi(x,K,\theta)$ the outcome of (A1), i.e.,

$$\begin{aligned}&\chi(x,K,\hat{\theta})e^{-iK\hat{\theta}\cdot x} - 1 \\ &\quad = \frac{1}{2\pi i}\int_{\mathbb{C}} \frac{\bar{\partial}(e^{-iK'\hat{\theta}\cdot x}\psi(x,K,\hat{\theta}) - 1)}{K-K'} d\bar{K}' \wedge dK' \\ &\quad = -\frac{1}{4\pi^2}\int_{-\infty}^{\infty} \frac{k'}{K-k'} e^{-ik'\hat{\theta}\cdot x} \\ &\qquad \times \int_{S^2} \psi^{P}(x,k',\hat{\theta}')K(k',\hat{\theta}',\hat{\theta})\, d\hat{\theta}'\, dk'.\end{aligned} \tag{3.3}$$

If $q \neq 0$, we get $\chi(x,K,\hat{\theta}) = \psi(x,K,\hat{\theta})$. On the real axis, the k' integral has to be interpreted as a principal value integral; therefore

$$\begin{aligned}\chi(x,k,\hat{\theta}) &= \tfrac{1}{2}(\psi^{+}(x,k,\hat{\theta}) + \psi^{-}(x,k,\hat{\theta})) \\ &\neq \psi^{P}(x,k,\hat{\theta}).\end{aligned}$$

The solution χ appears also in the paper of Kouri and Levin.[2]

They give the relation

$$\begin{aligned}&\psi^{P}(x,k,\hat{\theta}) \\ &\quad = \chi(x,k,\hat{\theta}) + \frac{k^2}{4\pi}\int_{S^2} \chi(x,k,\hat{\theta}')K^2(x,\hat{\theta}',\hat{\theta})\, d\hat{\theta}',\end{aligned} \tag{3.4}$$

which is a direct consequence of (2.11). Here K^2 is understood in the operational sense. Combining the results, we have the following theorem.

Theorem 3.1: Assume that the potential admits a unique solution of (3.1) for all $K \in \mathbb{C}$ and satisfies the condition (3.2). Then we have the inverse scattering equations

$$\begin{aligned}&\chi(x,k,\theta)e^{-ik\theta\cdot x} - 1 \\ &\quad = -\frac{1}{4\pi^2}\int_{-\infty}^{\infty} \frac{k'}{k-k'} e^{-ik'\hat{\theta}\cdot x} \\ &\qquad \times \int_{S^2} \psi^{P}(x,k',\hat{\theta}')K(k',\hat{\theta}',\hat{\theta})\, d\hat{\theta}'\, dk', \\ &\psi^{P}(x,k,\hat{\theta}) = \chi(x,k,\hat{\theta}) \\ &\qquad + \frac{k^2}{4\pi}\int_{S^2} \chi(x,k,\hat{\theta}')K^2(k,\hat{\theta}',\hat{\theta})\, d\hat{\theta}',\end{aligned}$$

for the reactance matrix K.

Equation (3.3) may be viewed as a counterpart of the generalized Marchenko equation of Newton. This interpretation has the following ground: if the jump of ψ on the real K axis were expressed in terms of A matrix and ψ^{+}, as usually in inverse scattering theory, i.e.,

$$\begin{aligned}\bar{\partial}\psi(x,K,\theta) = &-\frac{ik}{8\pi}\delta(q)\int_{S^2} A(-k,\hat{\theta},\hat{\theta}') \\ &\times \psi^{+}(x,k,\hat{\theta}')\, d\hat{\theta}',\end{aligned}$$

the reconstruction formula (A1) would yield exactly the Fourier transform of the Marchenko equation.

Equation (3.3) gives the following reconstruction formulas for the potential.

Theorem 3.2: Assume that V satisfies the assumptions of Theorem 3.1. Then V admits a representation

$$\begin{aligned}V(x) = &-\frac{i}{2\pi}\hat{\theta}\cdot\nabla\int_{-\infty}^{\infty} k' \int_{S^2} e^{-ik'\hat{\theta}\cdot x}\psi^{P}(x,k',\hat{\theta}') \\ &\times K(k',\hat{\theta}',\hat{\theta})\, d\hat{\theta}'\, dk'.\end{aligned} \tag{3.5}$$

Proof: We use the fact that χ satisfies the Schrödinger equation. Applying $\Delta + K^2$ to χ in (3.3) we get

$$V(x)\chi(x,K,\hat{\theta}) = (\Delta + K^2)\chi(x,K,\hat{\theta})$$
$$= -\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\frac{k'}{K-k'}\int_{S^2}K(k',\hat{\theta}',\hat{\theta})(\Delta+K^2)(e^{i(K-k')\hat{\theta}\cdot x}\psi^p(x,k',\hat{\theta}'))dk'\,d\hat{\theta}'$$
$$= -\frac{i}{2\pi^2}e^{iK\hat{\theta}\cdot x}\hat{\theta}\cdot\nabla\int_{-\infty}^{\infty}k'\int_{S^2}e^{-ik'\hat{\theta}\cdot x}\psi^p(x,k',\hat{\theta}')K(k',\hat{\theta}',\hat{\theta})d\theta'\,dk'$$
$$-\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\frac{k'}{K-k'}\int_{S^2}e^{i(K-k')\hat{\theta}\cdot x}(\Delta+k'^2)\psi^p(x,k',\hat{\theta}')K(k',\hat{\theta}',\hat{\theta})d\hat{\theta}'\,dk'. \tag{3.6}$$

In the above equation, the identity

$$(\Delta+K^2)(e^{-i(K-k')\hat{\theta}\cdot x}f(x))$$
$$= e^{i(K-k')\hat{\theta}\cdot x}(\Delta+k'^2)f(x)$$
$$+ 2i(K-k')e^{iK\hat{\theta}\cdot x}\hat{\theta}\cdot\nabla(e^{-ik'\hat{\theta}\cdot x}f(x))$$

was used. Comparing the large $|k|$ behavior of both sides in (3.5) we get the desired representation. □

This trace-type formula should be compared to the one given by Newton in Ref. 13, which is obtainable from the generalized Marchenko equation in the same way as done above. Note that Eq. (3.5) is "miraculous," of course, in the sense that the apparent $\hat{\theta}$ dependence of the right-hand side does not show up on the left-hand side.

Equation (3.5) together with the integral equation (2.9) may be viewed as an alternative formulation of the inverse problem for standing waves.

ACKNOWLEDGMENTS

The author wishes to thank Professor Roger G. Newton and Professor Adrian Nachman for useful discussions, and Indiana University for its kind hospitality.

Financial support for this research was provided by the Academy of Finland and the Finnish Cultural Foundation.

APPENDIX: THE $\bar{\partial}$ OPERATOR

This short Appendix provides the basic $\bar{\partial}$ equations used in the text. For further details and proofs see, e.g., Hörmander.[14]

Let $K = k + iq \in \mathbb{C}$, k and q real. The antiholomorphic derivative $\bar{\partial}$ with respect to K is defined as

$$\bar{\partial} = \frac{\partial}{\partial\bar{K}} = \frac{1}{2}\left(\frac{\partial}{\partial k} + i\frac{\partial}{\partial q}\right).$$

Here $\bar{K}$ denotes the complex conjugate of K.

Let $\Omega\in\mathbb{C}$ be a bounded domain with C^1 boundary $\partial\Omega$ and $f\in C^1(\bar{\Omega})$. Then for $K\in\Omega$ we have the generalization of the Cauchy integral formula,

$$f(K) = -\frac{1}{2\pi i}\int_{\partial\Omega}\frac{f(K')}{K-K'}dK'$$
$$+\frac{1}{2\pi i}\int_{\Omega}\frac{\bar{\partial}f(K')}{K-K'}d\bar{K}'\wedge dK'. \tag{A1}$$

Furthermore, since $\bar{\partial}$ applied to any holomorphic function vanishes, we get in Ω the equation

$$\bar{\partial}f(K) = \frac{1}{2\pi i}\int_{\Omega}\bar{\partial}_{K'}f(K')\bar{\partial}_K((K-K')^{-1})d\bar{K}'\wedge dK'.$$

On the other hand, the $\bar{\partial}$ equation $\bar{\partial}f(K) = u(K)$, $K\in\Omega$, has a solution for all $u\in C^\infty(\Omega)$ and $d\bar{K}'\wedge dK' = 2i\,dk'\,dq'$, we may identify

$$\bar{\partial}_K((K-K')^{-1}) = \pi\delta_{\mathbb{C}}(K-K')$$
$$\equiv\pi\delta(k-k')\delta(q-q'). \tag{A2}$$

[1] R. G. Newton, "The Marchenko and Gel'fand–Levitan methods in the inverse scattering problem in one and three dimensions," in *Conference on Inverse Scattering: Theory and Application*, edited by J. D. Bednar, R. Redner, E. Robinson, and A. Weglein (SIAM, Philadelphia, 1983), and references therein.

[2] D. J. Kouri and F. S. Levin, "On standing wave solutions to the Schrödinger equation," Ann. Phys. (NY) **83**, 316 (1974).

[3] R. G. Newton, *Scattering Theory of Waves and Particles* (Springer, New York, 1982).

[4] R. Beals and R. R. Coifman, "Scattering, transformations spectrales et equations d'evolution nonlineaire I, II," presented at Seminaire Goulaouic–Meyer–Schwartz, Ecole Polytechnique, Palaiseau, 22 (1980 and 1981); 21 (1981 and 1982).

[5] R. Beals and R. R. Coifman, "Multidimensional inverse scattering and nonlinear partial differential equations," Proc. Symp. Pure Math. **43**, 45 (1985).

[6] M. Ablowitz and A. Nachman, "A multi-dimensional inverse-scattering method," Stud. Appl. Math. **71**, 243 (1984).

[7] M. Ablowitz and A. Nachman, "Multidimensional inverse scattering for first-order systems," Stud. Appl. Math. **71**, 251 (1984).

[8] R. G. Newton, "Noncentral potentials: The generalized Levinson theorem and the structure of the spectrum," J. Math. Phys. **7**, 1348 (1977).

[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics III: Scattering Theory* (Academic, New York, 1979).

[10] Y. Saito, "The principle of limiting absorption for the non-selfadjoint Schrödinger operator in $\mathbf{R}^N$ ($N\neq 2$)," Publ. RIMS, Kyoto Univ. **9**, 397 (1974).

[11] R. G. Newton, "Inverse scattering II. Three dimensions," J. Math. Phys. **21**, 1698 (1980).

[12] M. Cheney, "A rigorous derivation of the 'miracle' identity of three-dimensional inverse scattering," J. Math. Phys. **25**, 2988 (1984).

[13] R. G. Newton, "Representation of the potential in the Schrödinger equation," Phys. Rev. Lett. **20**, 1863 (1984).

[14] L. Hörmander, *Introduction to Complex Analysis in Several Variables* (North-Holland, Amsterdam, 1973).

Ambiguities in the Debye expansion of the elastic S matrix

J. B. Galvan
Departamento de Fisica Teorica, Facultad de Ciencias, 50009 Zaragoza, Spain
J. Sesma[a)]
Max-Planck-Institut für Kernphysik, Heidelberg, West Germany

(Received 3 March 1986; accepted for publication 6 May 1987)

The possibility of expanding the S matrix in a multiple-reflection series (Debye expansion) is shown for an arbitrary-shape short-range potential. Ambiguity in the definition of outgoing and incoming waves inside the potential leads to an infinity of expansions. These are analyzed in a simple example (square well or barrier) in order to characterize the correct choice of outgoing and incoming waves.

I. INTRODUCTION

Classical and semiclassical methods have turned out to be very useful in the description of heavy-ion elastic scattering.[1] They fail, however, in the explanation of heavy-ion phenomena, like rainbow and glory, where interference effects are important. These phenomena can, nevertheless, be analyzed without abandoning the appealing picture of trajectories. The procedure consists in writing the exact quantum mechanical S matrix as a series, known as Debye expansion, whose terms can be viewed as due to multiple reflections of the incident wave in the interaction region: an incoming spherical wave partially transmitted to the interior of the potential is totally reflected from the origin, then partially reflected to the interior at the potential surface, and so on, to be finally transmitted, in part, to the exterior.

That procedure, applied by Debye[2] to the scattering of electromagnetic waves from a circular cylinder and by van der Pol and Bremmer[3] to the case of a sphere, has been successfully utilized by Nussenzveig[4] in the treatment of light scattering from water droplets. Its application to heavy-ion elastic scattering has been considered by Anni, Renna, and Taffara in a series of papers[5] dealing with analytically solvable potentials or with JWKB approximations, and by Agassi and Avishai[6] in the "staircase" approximation to the potential. The second term of the Debye expansion, for several realistic ion–ion potentials, was considered by Brink and Takigawa[7] in their analysis of the barrier penetration effects in the semiclassical theory of elastic scattering.

The usefulness of the Debye expansion in the explanation of rainbow and glory effects was already emphasized by Nussenzveig.[4] Our purpose in a current research was to apply the Debye expansion in the analysis of the phenomenon, in heavy-ion physics, known as ALAS (anomalous large angle scattering) and reviewed in a report by Braun-Munzinger and Barrette.[8] However, as we were trying to develop our program, we encountered an ambiguity in the definition of the Debye expansion.

We show in Sec. II that a Debye expansion of the S matrix is always possible for a spherically symmetric potential, independently of its shape: one needs only to specify which solutions of the wave equation are to be interpreted as outgoing and incoming spherical waves. Their radial parts, which we shall denote, respectively, by ψ^{I} and ψ^{II}, can be expressed as linear combinations of ψ^r and ψ^i, the regular (at the origin) and the irregular solutions of the radial Schrödinger equation, with the only limitation that the sum of the outgoing and incoming waves must produce the stationary one, i.e., the regular solution. Therefore,

$$\psi^{\mathrm{I}} = \tfrac{1}{2}\psi^r + ia\psi^i, \quad \psi^{\mathrm{II}} = \tfrac{1}{2}\psi^r - ia\psi^i, \tag{1.1}$$

where a is a parameter measuring the "component" of the irregular solution in the traveling waves. The ambiguity mentioned above stems from the fact that the parameter a can be arbitrarily chosen. An infinity of Debye-like expansions can be obtained in this way.

In the case of a step potential, as considered by Nussenzveig,[4] one infers which is the "orthodox" choice of traveling waves inside and outside the potential from the motion of their wave fronts. In the more realistic case (in heavy-ion elastic scattering) of an attractive nuclear potential with a Coulombian tail, the behavior at large distances makes it possible to recognize the outgoing and incoming waves in the outside region; but there is no *a priori* guidance to distinguish the orthodox choice from the heterodox ones inside the potential.

The continuity of the radial current as one passes from the external to the internal region does not help to eliminate the ambiguity in the choice of ψ^{I} and ψ^{II}. We are going to show [see Eqs. (2.11)–(2.14) below] that the continuity of the radial current is guaranteed by the continuity of the logarithmic derivative of the scattering solution of the Schrödinger equation, that is assured if the correct expression of the S matrix is taken. But the S matrix [Eq. (2.2) below] does not depend at all on the outgoing and incoming waves in the internal region.

To illustrate the preceding considerations, let us assume that ψ^r and ψ^i in Eq. (1.1) have been taken to be real (always possible for energies above the barrier) and the parameter a also real. The radial current associated to ψ^{I} is then given by

$$J = (\hbar/2m)aW(\psi^r,\psi^i), \tag{1.2}$$

where W means the Wronskian. Obviously, ψ^r and ψ^i could be redefined by multiplying them by arbitrary factors. In particular one could choose the same factor, let us say 2, for both ψ^r and ψ^i. The Wronskian would then become multi-

[a)] On leave of absence from Universidad de Zaragoza, Zaragoza, Spain.

 0022-2488/87/102420-07$02.50 © 1987 American Institute of Physics

plied by 4 and the parameter a should be divided by 4 to obtain the same current. The new ψ^{I} so obtained would be different (i.e., linearly independent) than the previous one, so making evident the existence of the ambiguity.

In order to obtain some criteria which are helpful in removing such ambiguity, we consider in Sec. III the case of a step potential, for which the orthodox choice of outgoing and incoming waves is evident. An analysis of the convergence of orthodox and heterodox Debye expansions is made as well as a comparison of the corresponding reflection coefficients.

Finally, some conclusions are mentioned in Sec. IV. As long as our step potential is an oversimplification of the realistic heavy-ion ones, the possibility that some of our conclusions are valid only in the case under consideration cannot be discarded. The ambiguity in the Debye expansion, seen in evidence in this paper, should be analyzed for more general (complex) potentials, preferably analytically solvable, in order to draw more general conclusions.

II. DEBYE-LIKE EXPANSION

Let us consider scattering by a spherically symmetric potential such that two regions, interior and exterior (labeled, respectively, by 1 and 2), are clearly distinguished. The spherical surface of radius b separating the two regions will be referred to as the potential surface. In the external region the potential is assumed to be constant or purely Coulombian, so as to yield an unambiguous definition of outgoing and incoming waves in that region. Due to the spherical symmetry of the potential, the scattering can be analyzed in terms of partial waves of definite angular momentum. In what follows, a label l denoting the angular momentum is to be understood although it is not explicitly displayed.

The wave function inside and outside the potential surface is given by

$$\psi(r)=\begin{cases}\psi_1^{\mathrm{regular}}(r), & r<b,\\ \psi_2^{\mathrm{incoming}}(r)-S\psi_2^{\mathrm{outgoing}}(r), & r>b,\end{cases} \tag{2.1}$$

where S represents the l-wave component of the S matrix. By imposing continuity at $r=b$ of the logarithmic derivative of the wave function one obtains

$$S=\frac{\psi_2^{\mathrm{incoming}}}{\psi_2^{\mathrm{outgoing}}}\frac{\ln'\psi_2^{\mathrm{incoming}}-\ln'\psi_1^{\mathrm{regular}}}{\ln'\psi_2^{\mathrm{outgoing}}-\ln'\psi_1^{\mathrm{regular}}}, \tag{2.2}$$

where all intervening functions are to be taken at $r=b$. For brevity we write

$$\psi_2^{\mathrm{I}}\equiv\psi_2^{\mathrm{outgoing}},\quad \psi_2^{\mathrm{II}}\equiv\psi_2^{\mathrm{incoming}}. \tag{2.3}$$

These two functions are well known in both cases of potential with constant or Coulombian tail.

Let us now consider in the internal region "outgoing" and "incoming" waves, $\psi_1^{\mathrm{I}}(r)$ and $\psi_1^{\mathrm{II}}(r)$, arbitrarily chosen with the only restriction

$$\psi_1^{\mathrm{I}}(r)+\psi_1^{\mathrm{II}}(r)=\psi_1^{\mathrm{regular}}(r). \tag{2.4}$$

The S matrix given by Eq. (2.2) can then be written, after some straightforward manipulations, in the form

$$S=-\frac{\psi_2^{\mathrm{II}}}{\psi_2^{\mathrm{I}}}\left\{R_{22}+T_{21}\frac{\psi_1^{\mathrm{I}}}{\psi_1^{\mathrm{II}}}\sum_{p=0}^{\infty}\left(\frac{R_{11}\psi_1^{\mathrm{I}}}{\psi_1^{\mathrm{II}}}\right)^p T_{12}\right\}, \tag{2.5}$$

with the notation

$$R_{11}\equiv-(\ln'\psi_2^{\mathrm{I}}-\ln'\psi_1^{\mathrm{I}})/(\ln'\psi_2^{\mathrm{I}}-\ln'\psi_1^{\mathrm{II}}), \tag{2.6}$$

$$T_{12}\equiv(\ln'\psi_1^{\mathrm{I}}-\ln'\psi_1^{\mathrm{II}})/(\ln'\psi_2^{\mathrm{I}}-\ln'\psi_1^{\mathrm{II}}), \tag{2.7}$$

$$R_{22}\equiv-(\ln'\psi_2^{\mathrm{II}}-\ln'\psi_1^{\mathrm{II}})/(\ln'\psi_2^{\mathrm{I}}-\ln'\psi_1^{\mathrm{II}}), \tag{2.8}$$

$$T_{21}\equiv(\ln'\psi_2^{\mathrm{I}}-\ln'\psi_2^{\mathrm{II}})/(\ln'\psi_2^{\mathrm{I}}-\ln'\psi_1^{\mathrm{II}}), \tag{2.9}$$

and provided the geometrical series in the right-hand side of Eq. (2.5) is convergent, i.e.,

$$|R_{11}\psi_1^{\mathrm{I}}/\psi_1^{\mathrm{II}}|<1. \tag{2.10}$$

If one associates the coefficients R_{11}, T_{12}, R_{22}, and T_{21}, respectively, with internal reflection on, transmission to the exterior through, external reflection on, and transmission to the interior through the potential surface, the expansion on the right-hand side of Eq. (2.5) can be interpreted in terms of multiple reflections, just like the Debye expansion,[4] no matter whether ψ_1^{I} and ψ_1^{II} are the *true* outgoing and incoming waves or not. An infinity of Debye-like expansions can, therefore, be considered.

The reflection and transmission coefficients are not independent. They satisfy relations that are formally the same for orthodox and heterodox choices of ψ_1^{I} and ψ_1^{II}. From Eqs. (2.6) and (2.7) one immediately obtains

$$1+R_{11}=T_{12}, \tag{2.11}$$

$$\ln'\psi_1^{\mathrm{I}}+R_{11}\ln'\psi_1^{\mathrm{II}}=T_{12}\ln'\psi_2^{\mathrm{I}}. \tag{2.12}$$

These relations merely express the continuity at $r=b$ of the function and its derivative for a wave incident from the interior on the potential surface. Analogously, for a wave incident from the exterior, one has

$$1+R_{22}=T_{21}, \tag{2.13}$$

$$\ln'\psi_2^{\mathrm{II}}+R_{22}\ln'\psi_2^{\mathrm{I}}=T_{21}\ln'\psi_1^{\mathrm{II}}, \tag{2.14}$$

trivially deduced from Eqs. (2.8) and (2.9).

Until now no limitations, apart from Eq. (2.4), have been imposed to $\psi_1^{\mathrm{I}}(r)$ and $\psi_1^{\mathrm{II}}(r)$. It seems, however, convenient to restrict these functions to be solutions of the wave equation, since they are interpreted as traveling waves. The Schrödinger equation in the internal region has two independent solutions, for instance $\psi_1^{\mathrm{regular}}$ and $\psi_1^{\mathrm{irregular}}$, in terms of which ψ_1^{I} and ψ_1^{II} can be written as

$$\psi_1^{\mathrm{I}}(r)=\tfrac{1}{2}\psi_1^{\mathrm{regular}}(r)+\alpha\psi_1^{\mathrm{irregular}}(r), \tag{2.15a}$$

$$\psi_1^{\mathrm{II}}(r)=\tfrac{1}{2}\psi_1^{\mathrm{regular}}(r)-\alpha\psi_1^{\mathrm{irregular}}(r), \tag{2.15b}$$

the parameter α being complex for the moment. In the case of a real nuclear potential $\psi_1^{\mathrm{regular}}(r)$ and $\psi_1^{\mathrm{irregular}}(r)$ are solutions of a differential equation with real coefficients and can be taken as real functions.

In the case of a constant potential the outgoing and incoming waves are complex conjugates of each other. It seems, therefore, plausible to impose in our heavy-ion problem the condition

$$\overline{\psi_1^{\mathrm{I}}(r)}=\psi_1^{\mathrm{II}}(r), \tag{2.16}$$

at least at energies above the barrier. This condition is satis-

fied if the parameter α in Eqs. (2.15) is taken to be purely imaginary.

Finally, another requirement in analogy with what happens for free waves could be

$$\psi_1^{\mathrm{II}}(-r) = (-1)^l \psi_1^{\mathrm{I}}(r), \tag{2.17}$$

associated with the fact that an incoming wave is totally reflected from the origin to give an outgoing wave.

As stated above, the validity of the Debye-like expansion in Eq. (2.5) is conditioned by the convergence of the geometrical series

$$\sum_{p=0}^{\infty} \rho^p, \tag{2.18}$$

where

$$\rho = R_{11}\, \psi_1^{\mathrm{I}}/\psi_i^{\mathrm{II}}. \tag{2.19}$$

We shall refer to this parameter ρ as the Debye parameter. In order to analyze its magnitude, we are going to obtain some useful relations.

The product of Eq. (2.12) times the complex conjugate of Eq. (2.11) gives

$$\ln' \psi_1^{\mathrm{I}} + R_{11} \ln' \psi_1^{\mathrm{II}} + \overline{R_{11}} \ln' \psi_1^{\mathrm{I}} + |R_{11}|^2 \ln' \psi_1^{\mathrm{II}} = |T_{12}|^2 \ln' \psi_2^{\mathrm{I}}. \tag{2.20}$$

If the condition expressed in Eq. (2.16) is satisfied, one obtains from the real and imaginary parts of Eq. (2.20),

$$|T_{12}|^2 \operatorname{Re}\{\ln' \psi_1^{\mathrm{I}} - \ln' \psi_2^{\mathrm{I}}\} + 2 \operatorname{Im}\{R_{11}\} \operatorname{Im}\{\ln' \psi_1^{\mathrm{I}}\} = 0, \tag{2.21}$$

$$1 - |R_{11}|^2 - |T_{12}|^2 \operatorname{Im}\{\ln' \psi_2^{\mathrm{I}}\}/\operatorname{Im}\{\ln' \psi_1^{\mathrm{I}}\} = 0. \tag{2.22}$$

Analogously, from Eqs. (2.13) and (2.14), it follows that

$$\ln' \psi_2^{\mathrm{II}} + R_{22} \ln' \psi_2^{\mathrm{I}} + \overline{R_{22}} \ln' \psi_2^{\mathrm{II}} + |R_{22}|^2 \ln' \psi_2^{\mathrm{I}} = |T_{21}|^2 \ln' \psi_1^{\mathrm{II}}. \tag{2.23}$$

In the case where

$$\overline{\psi_2^{\mathrm{I}}(r)} = \psi_2^{\mathrm{II}}(r) \tag{2.24}$$

(as it happens for constant or Coulombian potential tails), Eq. (2.23) gives

$$|T_{21}|^2 \operatorname{Re}\{\ln' \psi_2^{\mathrm{I}} - \ln' \psi_1^{\mathrm{II}}\} - 2 \operatorname{Im}\{R_{22}\} \operatorname{Im}\{\ln' \psi_2^{\mathrm{I}}\} = 0, \tag{2.25}$$

$$1 - |R_{22}|^2 + |T_{21}|^2 \operatorname{Im}\{\ln' \psi_1^{\mathrm{II}}\}/\operatorname{Im}\{\ln' \psi_2^{\mathrm{I}}\} = 0. \tag{2.26}$$

Equations (2.22) and (2.26) are what we need to analyze the convergence of the Debye-like expansion. Bearing in mind the definitions of T_{12} and T_{21} and Eqs. (2.16) and (2.24), these equations can be written in the form

$$1 - |R_{11}|^2 = \left| \frac{W\{\psi_1^{\mathrm{II}}, \psi_1^{\mathrm{I}}\}}{W\{\psi_1^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}} \right|^2 \frac{W\{\psi_2^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}}{W\{\psi_1^{\mathrm{II}}, \psi_1^{\mathrm{I}}\}}, \tag{2.27}$$

$$1 - |R_{22}|^2 = \left| \frac{W\{\psi_2^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}}{W\{\psi_1^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}} \right|^2 \frac{W\{\psi_1^{\mathrm{II}}, \psi_1^{\mathrm{I}}\}}{W\{\psi_2^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}}, \tag{2.28}$$

where W stands for the value of the Wronskian of the two corresponding functions at $r = b$. Except for the trivial case of the parameter α in Eqs. (2.15) being zero, the outgoing and incoming waves are independent in both the internal and external regions and, therefore, the Wronskians in the numerators of the right-hand sides of Eqs. (2.27) and (2.28) do not vanish. We can then conclude that

$$|R_{11}| < 1, \quad |R_{22}| < 1, \tag{2.29}$$

if the parameter α has been chosen such that

$$\operatorname{sgn}(iW\{\psi_2^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}) = \operatorname{sgn}(iW\{\psi_1^{\mathrm{II}}, \psi_1^{\mathrm{I}}\}), \tag{2.30}$$

whereas

$$|R_{11}| > 1, \quad |R_{22}| > 1, \tag{2.31}$$

in the case

$$\operatorname{sgn}(iW\{\psi_2^{\mathrm{II}}, \psi_2^{\mathrm{I}}\}) = \operatorname{sgn}(-iW\{\psi_1^{\mathrm{II}}, \psi_1^{\mathrm{I}}\}). \tag{2.32}$$

Since, due to Eq. (2.16), the Debye parameter and the internal reflection coefficient have the same modulus,

$$|\rho| = |R_{11}|, \tag{2.33}$$

the convergence or divergence of the Debye-like expansion depends on the choice of the parameter α in Eqs. (2.15) so as to give, respectively, Eq. (2.30) or Eq. (2.32).

Some other interesting relations among the reflection and transmission coefficients can be written whenever Eqs. (2.16) and (2.24) are satisfied. From Eqs. (2.22) and (2.26) it is immediate to obtain

$$1 - |R_{11}|^2 - \overline{T_{12}} T_{21} = 0, \tag{2.34}$$

$$1 - |R_{22}|^2 - T_{12}\, \overline{T_{21}} = 0. \tag{2.35}$$

Equations (2.34) and (2.35) imply that the product $\overline{T_{12}} T_{21}$ is real and, therefore,

$$\overline{T_{12}} T_{21} = T_{12}\, \overline{T_{21}}. \tag{2.36}$$

Comparison of Eqs. (2.34) and (2.35) allows us to conclude that

$$|R_{11}| = |R_{22}|. \tag{2.37}$$

To end this section a brief comment concerning bound states and resonances is in order. As is well known, both bound states and resonances are associated with poles of the S matrix at negative or complex energies. In both cases one has

$$\ln' \psi_1^{\mathrm{regular}} = \ln' \psi_2^{\mathrm{I}} \tag{2.38}$$

and, in consequence,

$$\rho = 1. \tag{2.39}$$

Notice that this result is compatible with the inequalities (2.29) or (2.31). The latter have been obtained by assuming the validity of Eqs. (2.16) and (2.24), which obviously do not hold at negative or complex energies.

III. AN EXAMPLE: STEP POTENTIAL

In order to illustrate the ambiguity in the Debye expansion pointed out in the preceding section, we consider here the scattering of a particle of mass m and energy E by a short-range spherically symmetric potential of radius b and constant intensity V (positive for a barrier, negative for a well). The internal and external dimensionless wave numbers are, respectively,

$$z_1 = b\,[2m(E - V)]^{1/2}/\hbar, \quad z_2 = b(2mE)^{1/2}/\hbar. \tag{3.1}$$

In the external region the outgoing and incoming l waves are, obviously,

$$\psi_2^{\mathrm{I}}(r) = h_l^1(z_2 r/b), \quad \psi_2^{\mathrm{II}}(r) = h_l^2(z_2 r/b). \tag{3.2}$$

In the internal region we take

$$\psi_1^{\mathrm{I}}(r) = j_l(z_1 r/b) + i a y_l(z_1 r/b), \tag{3.3a}$$

$$\psi_1^{\mathrm{II}}(r) = j_l(z_1 r/b) - i a y_l(z_1 r/b). \tag{3.3b}$$

In these expressions h_l, j_l, and y_l represent the spherical Bessel functions.[9] The parameter a could be an arbitrary function of the energy. For our purposes it is enough to consider it as a constant. Moreover, we restrict it to real values in order to have Eq. (2.16) satisfied. Obviously, the orthodox choice is $a = 1$. The Wronskians of the internal and external waves at $r = b$ turn out to be

$$W\{\psi_1^{\mathrm{II}}, \psi_1^{\mathrm{I}}\} = 2ia/z_1 b, \tag{3.4}$$

$$W\{\psi_2^{\mathrm{II}}, \psi_2^{\mathrm{I}}\} = 2i/z_2 b, \tag{3.5}$$

and, therefore, z_1 and z_2 being real and positive, the Debye-like expansion is convergent for $a > 0$ and divergent for $a < 0$.

In our subsequent discussion of the reflection and transmission coefficients we do not need to consider both signs of a, in view of the existing relations among coefficients corresponding to opposite values of a. If we label with a superindex $(+)$ the waves and coefficients for positive a and with $(-)$ those for the opposite a, we have

$$\psi_1^{\mathrm{I}(-)} = \psi_1^{\mathrm{II}(+)}, \quad \psi_1^{\mathrm{II}(-)} = \psi_1^{\mathrm{I}(+)}, \tag{3.6}$$

and, in consequence,

$$R_{11}^{(-)} = 1/R_{11}^{(+)}, \tag{3.7a}$$

$$R_{22}^{(-)} = -(R_{22}^{(+)} + R_{11}^{(+)} + 1)/R_{11}^{(+)},$$

$$T_{12}^{(-)} = T_{12}^{(+)}/R_{11}^{(+)}, \quad T_{21}^{(-)} = -T_{21}^{(+)}/R_{11}^{(+)}, \tag{3.7b}$$

$$\rho^{(-)} = 1/\rho^{(+)}. \tag{3.7c}$$

In what follows we shall consider only positive values of a.

By substitution of the explicit form of the outgoing and incoming waves, given in Eqs. (3.2) and (3.3), into the corresponding definitions, one obtains for the Debye parameter

$$\rho = \frac{z_2 h^{1\prime}(z_2)\{h^1(z_2) + cj(z_1)\} - z_1 h^1(z_2)\{h^{1\prime}(z_1) + cj'(z_1)\}}{z_2 h^{1\prime}(z_2)\{h^2(z_1) + cj(z_1)\} - z_1 h^1(z_2)\{h^{2\prime}(z_1) + cj'(z_1)\}}, \tag{3.8}$$

and for the reflection coefficients

$$R_{11} = \rho\{h^2(z_1) + cj(z_1)\}/\{h^1(z_1) + cj(z_1)\}, \tag{3.9}$$

$$R_{22} = -\frac{z_2\{h^{2\prime}(z_2)/h^2(z_2)\}\{h^2(z_1) + cj(z_1)\} - z_1\{h^{2\prime}(z_1) + cj'(z_1)\}}{z_2\{h^{1\prime}(z_2)/h^1(z_2)\}\{h^2(z_1) + cj(z_1)\} - z_1\{h^{2\prime}(z_1) + cj'(z_1)\}}, \tag{3.10}$$

where

$$c \equiv (1 - a)/a. \tag{3.11}$$

For simplicity of notation, the subscript l for the Bessel functions has been omitted. Approximate expressions of those coefficients can be obtained in the limits of high and low energies.

A. High energies

Let us assume that $E \to +\infty$. This implies

$$z_1 = z_2(1 - V/E)^{1/2} \to \infty, \quad z_2 \to \infty. \tag{3.12}$$

By defining

$$\delta \equiv V/2E, \tag{3.13}$$

approximating

$$z_1 \simeq z_2(1 - V/2E), \tag{3.14}$$

retaining only the first terms of the Taylor expansion of the Bessel functions $h(z_1)$ and $j(z_1)$ about z_2, and making use of the differential equation they satisfy,[9] we find

$$\rho \simeq -\frac{cW\{j,h^1\} - z_2\delta\{h^{1\prime}(h^{1\prime} + cj') + h^1(h^1 + cj)\}}{W\{h^2,h^1\} + cW\{j,h^1\} - z_2\delta\{h^{1\prime}(h^{2\prime} + cj') + h^1(h^2 + cj)\}}, \tag{3.15}$$

$$R_{11} \simeq \rho\{h^2 + cj - z_2\delta(h^{2\prime} + cj')\}/\{h^1 + cj - z_2\delta(h^{1\prime} + cj')\}, \tag{3.16}$$

$$R_{22} \simeq -\frac{h^1}{h^2}\frac{cW\{j,h^2\} - z_2\delta[h^{2\prime}(h^{2\prime} + cj') + h^2(h^2 + cj)]}{W\{h^2,h^1\} + cW\{j,h^1\} - z_2\delta[h^{1\prime}(h^{2\prime} + cj') + h^1(h^2 + cj)]}, \tag{3.17}$$

where all Bessel function are to be taken at z_2. To the lowest order, the above expressions can be approximated in the form

$$\rho \simeq (a - 1)/(a + 1), \tag{3.18}$$

$$R_{11} \simeq \frac{(1 + a)/(1 - a) + \exp[i2\theta]}{(1 + a)/(1 - a) + \exp[-i2\theta]}\exp[-i2\theta]\rho, \tag{3.19}$$

$$R_{22} \simeq [(1 - a)/(1 + a)]\exp\{i2\theta\}, \tag{3.20}$$

with

$$\theta \equiv z_2 - (l+1)\pi/2. \tag{3.21}$$

All these expressions have been obtained under the assumption that $a \neq 1$, i.e., in the case of a heterodox choice of the outgoing and incoming waves in the inner region. For the orthodox choice one has

$$\rho \simeq \exp\{i(2z_1 - l\pi)\} V/4E, \tag{3.22}$$

$$R_{11} \simeq -V/4E, \tag{3.23}$$

$$R_{22} \simeq V/4E. \tag{3.24}$$

The transmission coefficients are trivially obtained from the reflection ones in view of Eqs. (2.11) and (2.13).

B. Low energies

Let us consider now the case $E \to 0$, i.e.,

$$z_1 \to z_0 \equiv b(-2mV)^{1/2}/\hbar, \quad z_2 \to 0. \tag{3.25}$$

By retaining only dominant terms in the expressions of the Bessel functions and their derivatives,[9] one has from Eqs. (3.8)–(3.10),

$$\rho \simeq -\frac{(1+l)\{h^1(z_0) + cj(z_0)\} + z_0\{h^{1\prime}(z_0) + cj'(z_0)\}}{(1+l)\{h^2(z_0) + cj(z_0)\} + z_0\{h^{2\prime}(z_0) + cj'(z_0)\}}, \tag{3.26}$$

$$R_{11} \simeq -\frac{1 + l + z_0\{h^{1\prime}(z_0) + cj'(z_0)\}/\{h^1(z_0) + cj(z_0)\}}{1 + l + z_0\{h^{2\prime}(z_0) + cj'(z_0)\}/\{h^2(z_0) + cj(z_0)\}}, \tag{3.27}$$

$$R_{22} \simeq -1. \tag{3.28}$$

In the case of a potential well (z_0 real) it is immediately obvious that $|R_{11}| \to 1$ as $E \to 0$, whereas for potential barrier (z_0 purely imaginary) R_{11} tends to a real constant.

C. Intermediate energies

We have evaluated numerically the expressions given in the right-hand side of Eqs. (3.8)–(3.10) as functions of the energy E for the orthodox value of the parameter $a (a = 1)$ and for two heterodox ones ($a = 2$ and $a = \frac{1}{2}$). Two different values of the angular momentum ($l = 0$ and $l = 5$) have been considered for both cases of well ($V = -25\hbar^2/2mb^2$) and barrier ($V = 25\hbar^2/2mb^2$). The results are shown in Figs. 1–8.

Equations (3.8)–(3.11) allow us to define reflection and transmission coefficients at energies at which one of the regions becomes classically forbidden. It is interesting to remark that, the wave number being purely imaginary, the reduced logarithmic derivative of the wave function is real in a classically forbidden region. Therefore, in the case of a potential well one has

$$|\rho| = |R_{11}| = 1 \quad \text{for } V < E < 0, \tag{3.29}$$

the particular value $\rho = 1$ corresponding to bound states, whereas in the case of potential barrier

$$|R_{22}| = 1 \quad \text{for } 0 < E < V. \tag{3.30}$$

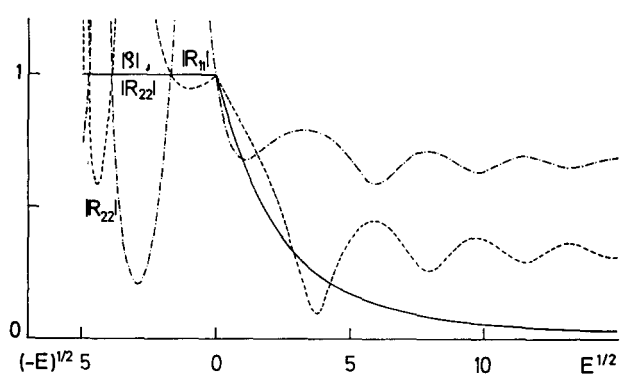


FIG. 1. Moduli of the Debye parameter and the internal and external reflection coefficients for S wave ($l = 0$) scattering by a square well potential of range and intensity as given in the text. The continuous line corresponds to the orthodox choice of the outgoing and incoming internal waves. The dashed and dash–dotted lines result for two different heterodox choices, corresponding, respectively, to values $a = 2$ and $a = \frac{1}{2}$ for the parameter on the right-hand side of Eqs. (3.3) in the text.

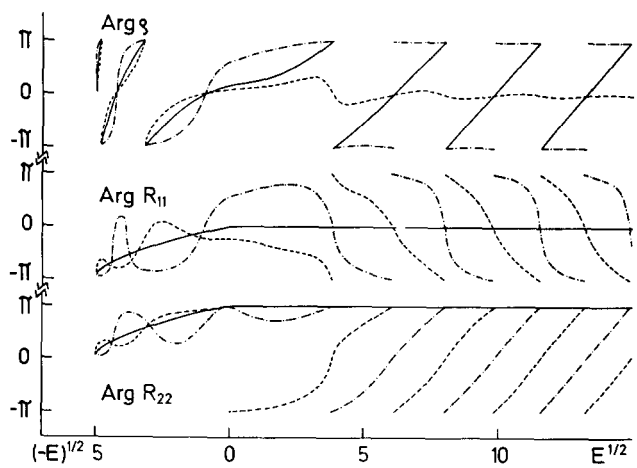


FIG. 2. Phases of the Debye parameter and the internal and external reflection coefficients whose moduli have been shown in Fig. 1. The continuous, dashed, and dash–dotted lines correspond to the same cases as before. The arguments have been reduced to the interval $[-\pi, \pi]$ by adding or subtracting a multiple of 2π.

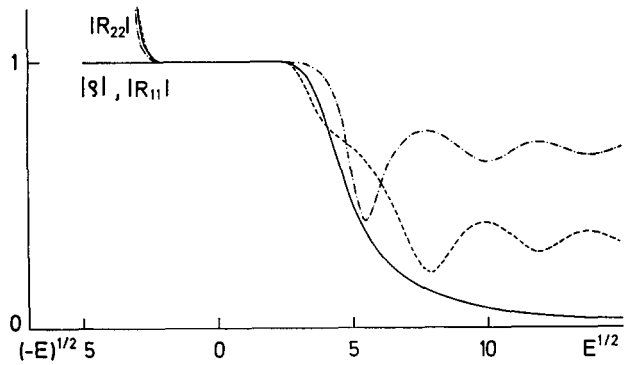


FIG. 3. Moduli of the Debye parameter and the reflection coefficients for H-wave ($l = 5$) scattering by the same square well potential as in Fig. 1.

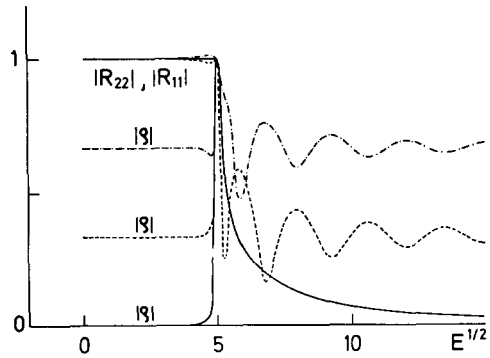


FIG. 5. Moduli of the Debye parameter and the reflection coefficients for S-wave scattering by the square barrier mentioned in the text. The comments in caption for Fig. 1 are also valid here.

IV. CONCLUSIONS

As we have seen in Sec. II, there are an infinity of possible choices of outgoing and incoming waves in the nuclear region, all of them leading to convergent Debye-like expansions of the S matrix. Although these expansions are correct from a mathematical point of view, a physical meaning could hardly be assigned to their successive terms if the choice of outgoing and incoming waves is not the correct one. It is, therefore, interesting to find signatures of the orthodox election.

From the example considered in Sec. III, it turns out that the most relevant features of the orthodox choice appear in the high-energy behavior of the Debye parameter and the reflection coefficients. In the orthodox case, the moduli $|\rho| = |R_{11}| = |R_{22}|$ decrease monotonically to zero as the energy increases, and $\mathrm{Arg}\{\rho\}$ increases monotonically whereas $\mathrm{Arg}\{R_{11}\}$ and $\mathrm{Arg}\{R_{22}\}$ remain nearly constant and equal to a multiple of π. In the heterodox cases, the moduli tend, with oscillations of decreasing amplitude, to a finite constant, and $\mathrm{Arg}\{\rho\}$ varies slightly around a multiple of π whereas $\mathrm{Arg}\{R_{11}\}$ decreases and $\mathrm{Arg}\{R_{22}\}$ increases monotonically.

Another feature that allows us to distinguish the orthodox Debye parameter from the heterodox ones is the behavior at resonances. All parameters become real at resonant energies, but, whereas the phase of the orthodox ρ varies rapidly as the energy passes the resonant value, the phases of the heterodox ones remain nearly stationary, and, whereas the modulus of the correct ρ decreases uniformly, those of the incorrect ones present relative maxima or minima.

The potential well considered in our example has two S-wave, one P-wave, and one D-wave bound states. At the corresponding energies, $\rho = 1$ independently of the choice of outgoing and incoming waves. However, we can see in Fig. 2 that, for negative energies and $l = 0$, $\mathrm{Arg}\{R_{11}\}$ and $\mathrm{Arg}\{R_{22}\}$ have a smooth dependence on the energy in the orthodox case and an oscillatory behavior in the heterodox ones.

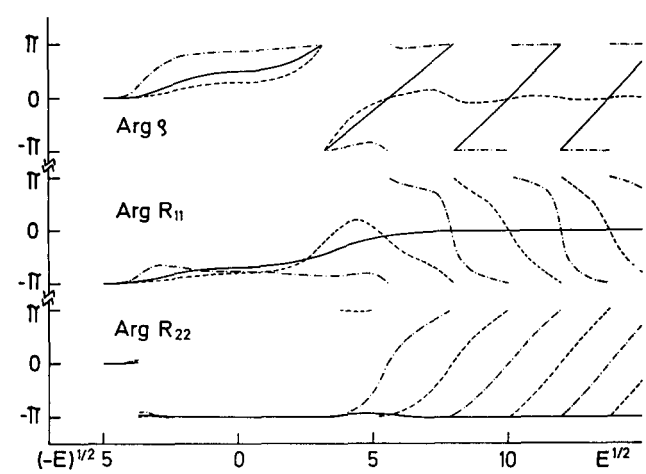


FIG. 4. Phases of the Debye parameter and the reflection coefficients whose moduli are shown in Fig. 3. The same convention as in Fig. 2 has been adopted. The discontinuity (of value $-\pi$) in Arg $\{R_{22}\}$ corresponds to a pole of R_{22}.

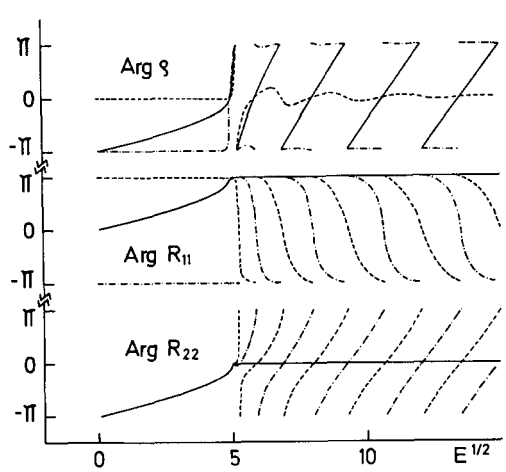


FIG. 6. Phases of the Deby parameter and the reflection coefficients whose moduli can be seen in Fig. 5.

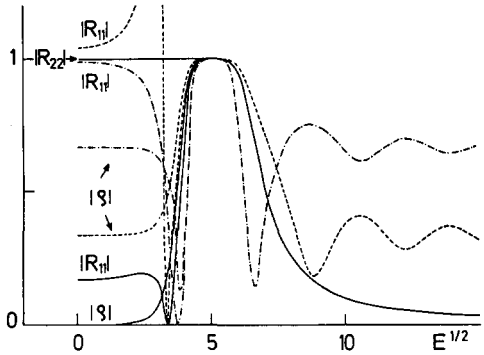


FIG. 7. Moduli of the Debye parameter and the reflection coefficients for the H-wave scattering by the same square barrier as in Fig. 5.

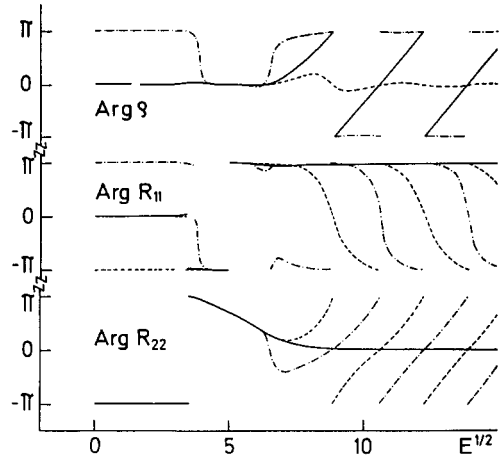


FIG. 8. Phases of the Debye parameter and the reflection coefficients whose moduli have been represented in Fig. 7. The discontinuities (of value $\pm\pi$) in Arg $\{R_{11}\}$ correspond to a pole and a zero of R_{11}.

ACKNOWLEDGMENTS

The authors are very grateful to Professor H. A. Weidenmüller for many helpful suggestions and a critical reading of the manuscript, and to the referee for very useful comments in order to improve the presentation of this paper.

One of the authors (J. S.) acknowledges also the hospitality of Professor Weidenmüller and the support of the Max–Planck Society during his visit to Heidelberg. The work has been partially supported by CAICYT.

[1]W. Nörenberg and H. A. Weidenmüller, "Introduction to the theory of heavy-ion reactions," *Lecture Notes in Physics*, Vol. 51 (Springer, Heidelberg, 1976); K. W. McVoy, "Semiclassical methods and surface creep waves in heavy-ion reactions," Lectures presented at the Escuela Latino-Americana de Fisica, Mexico, August, 1977, University of Wisconsin preprint; J. N. L. Connor, in "Semiclassical methods in molecular scattering and spectroscopy," *Proceedings of the NATO Advanced Study Institute, Cambridge, England, September 1979*, edited by M. S. Child (Reidel, Dordrecht, 1980), pp. 45–107; R. A. Broglia and A. Winther, *Heavy-ion reactions: Elastic and Inelastic Reactions* (Benjamin/Cummings, Reading, MA, 1981).

[2]P. J. Debye, Phys. Z. **9**, 775 (1908).

[3]B. van der Pol and H. Bremmer, Philos. Mag. **24**, 825 (1937).

[4]H. M. Nussenzveig, J. Math. Phys. **10**, 82, 125 (1969).

[5]R. Anni and L. Taffara, Nuovo Cimento A **31**, 321 (1976); R. Anni, L. Renna, and L. Taffara, *ibid.* **39**, 403 (1977); **45**, 123 (1978); R. Anni and L. Renna, *ibid.* **65**, 311 (1981); R. Anni, Lett. Nuovo Cimento **33**, 449 (1982).

[6]D. Agassi and Y. Avishai, Nucl. Phys. A **272**, 215 (1976); Phys. Lett. B **74**, 18 (1978).

[7]D. M. Brink and N. Takigawa, Nucl. Phys. A **279**, 159 (1977).

[8]P. Braun-Munzinger and J. Barrette, Phys. Rep. **87**, 209 (1982).

[9]H. A. Antosiewicz, in *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. Stegun (Dover, New York, 1965), pp. 435–478.

A note about closed timelike curves in Gödel space-time

David B. Malament
Department of Philosophy, The University of Chicago, 1050 East 59th Street, Chicago, Illinois 60637

(Received 9 April 1987; accepted for publication 3 June 1987)

A greatest lower bound for the total (integrated) energy of closed timelike curves in Gödel space-time is derived. (Here "energy" is determined relative to the velocity field of the major mass points of the universe.) The derivation is then used to reconstruct and extend a remark of Gödel's concerning total (integrated) acceleration requirements for "time travel" in his model universe.

I. INTRODUCTION

Gödel space-time,[1] of course, is not a live candidate for describing our universe. But it is an interesting geometric structure, and a source of insight into the possibilities allowed by relativity theory.

In this paper we present an elementary, but perhaps somewhat curious, proposition concerning the geometry of closed timelike curves in Gödel space-time (Proposition 2). It establishes a greatest lower bound for the total (integrated) energy of such curves (where "energy" is determined relative to the velocity field of the major mass points of the Universe). The proposition turns on the possibility of reducing questions about total energy (of closed timelike curves in Gödel space-time) to more tractable questions about area enclosure by curves in the hyperbolic plane (Proposition 1).

By way of application, we also invoke the proposition to reconstruct and extend a remark of Gödel's[2] concerning total (integrated) acceleration requirements for "time travel" in his model universe. It was this remark that first suggested our question about total energy. We close with a brief discussion of a conjecture on minimal total acceleration requirements.

II. PRELIMINARIES

In this section we recall several basic facts about Gödel space-time and introduce some notation.[3]

We take Gödel space-time to be the pair (M, g_{mn}) where M is $\mathbf{R}^4$ and g_{mn} is a Lorentz metric on M characterized by the condition that for some point (and hence, by homogeneity, any point) p in M, there is a global adapted (cylindrical) coordinate system t, r, φ, y on M in which $t(p) = r(p) = y(p) = 0$ and

$$g_{mn} = 4\mu^2[(dt)_m(dt)_n - (dr)_m(dr)_n - (dy)_m(dy)_n + (\mathrm{sh}^4 r - \mathrm{sh}^2 r)(d\varphi)_m(d\varphi)_n + 2\sqrt{2}\,\mathrm{sh}^2 r(d\varphi)_{(m}(dt)_{n)}].$$

(We use sh r and ch r, respectively, to abbreviate sinh r and cosh r.) Here $-\infty < t < \infty$, $-\infty < y < \infty$, $0 \leqslant r < \infty$, and $0 \leqslant \varphi < 2\pi$ with $\varphi = 0$ identified with $\varphi = 2\pi$. The metric g_{mn} is a solution to Einstein's equation

$$R_{mn} - \tfrac{1}{2} g_{mn} R = 8\pi\kappa[\rho\eta_m\eta_n - p(g_{mn} - \eta_m\eta_n)]$$

for a perfect fluid source with four-velocity $\eta^m = (\partial/\partial t)^m/2\mu$, mass density $\rho = 1/(16\pi\kappa\mu^2)$, and pressure $p = 1/(16\pi\kappa\mu^2)$.[4]

Here, η^m is a unit timelike Killing field, and defines a temporal orientation on (M, g_{mn}). The integral curves of the field, characterized by constant values for r, φ, and y, will be called *matter lines*. The $(\partial/\partial\varphi)^m$ is a rotational Killing field with squared norm $4\mu^2(\mathrm{sh}^4 r - \mathrm{sh}^2 r)$. Its (closed) integral curves, characterized by constant values for t, r, and y, will be called *Gödel circles*. Gödel circles with critical radius $r_c = \ln(1+\sqrt{2})$ are closed null curves (since $\mathrm{sh}\, r_C = 1$). Those with radius $r > r_C$ are closed timelike curves. Here $(\partial/\partial y)^m$ is a covariantly constant field with squared norm $-4\mu^2$.

Let S be a $t = \text{const}$, $y = \text{const}$ submanifold of M. Orthogonal projection of g_{mn} induces a (negative definite) metric

$$h_{mn} = g_{mn} - \left(\frac{1}{4}\mu^2\right)\left[\left(\frac{\partial}{\partial t}\right)_m\left(\frac{\partial}{\partial t}\right)_n - \left(\frac{\partial}{\partial y}\right)_m\left(\frac{\partial}{\partial y}\right)_n\right]$$

on S.[5] Now

$$\left(\frac{\partial}{\partial t}\right)_m = 4\mu^2[(dt)_m + \sqrt{2}\,\mathrm{sh}^2 r(d\varphi)_m]$$

and

$$\left(\frac{\partial}{\partial y}\right)_m = -4\mu^2(dy)_m.$$

So

$$h_{mn} = -4\mu^2[(dr)_m(dr)_n + \tfrac{1}{4}\mathrm{sh}^2 2r(d\varphi)_m(d\varphi)_n].$$

Once h_{mn} is presented in this form it is not difficult to verify that the pair $(S, -h_{mn})$ is a complete two-dimensional Riemannian manifold with constant curvature $-1/\mu^2$.[6]

In what follows we use the following notation. Given a timelike curve[7] γ in (M, g_{mn}), we take its four-velocity (i.e., unit tangent vector field) to be ξ^m, and set

$\alpha^m = \xi^n\nabla_n\xi^m$ (the acceleration of γ),

$a = (-\alpha^m\alpha_m)^{1/2}$ (the magnitude of γ's acceleration)

$E = \xi^m\eta_m$

(γ's energy with respect to the unit Killing field η^m).

We also use the parameter s for arc length (= elapsed proper time) along γ, and set

$$PT(\gamma) = \int_\gamma ds \quad \text{(total elapsed proper time of } \gamma),$$

 0022-2488/87/102427-04$02.50 © 1987 American Institute of Physics

$$TA(\gamma) = \int_\gamma a\, ds \quad \text{(total acceleration of } \gamma\text{)},$$

$$TE(\gamma) = \int_\gamma E\, ds \quad \text{(total energy of } \gamma\text{)}.$$

Note that $E \geqslant 1$ (since ξ^m and η^m are both future directed, unit timelike vectors), and that $E = 1/(1-v^2)^{1/2}$, where v is the speed of γ relative to matter lines. In terms of the coordinates above, E is given by

$$E = 2\mu\left[\left(\frac{dt}{ds}\right) + \sqrt{2}\,\mathrm{sh}^2 r\left(\frac{d\varphi}{ds}\right)\right].$$

In the special case where γ is a Gödel circle of radius $r > r_C$, we have

$$\xi^m = \left(\frac{d\varphi}{ds}\right)\left(\frac{\partial}{\partial\varphi}\right)^m$$

where

$$\frac{d\varphi}{ds} = \frac{1}{2\mu(\mathrm{sh}^4 r - \mathrm{sh}^2 r)^{1/2}},$$

and hence,[8]

$$E = \sqrt{2}\,\mathrm{sh}^2 r/(\mathrm{sh}^4 r - \mathrm{sh}^2 r)^{1/2},$$

$$a = (1/4\mu)\mathrm{sh}\, 2r(2\,\mathrm{sh}^2 r - 1)/(\mathrm{sh}^4 r - \mathrm{sh}^2 r),$$

$$PT(\gamma) = 4\pi\mu(\mathrm{sh}^4 r - \mathrm{sh}^2 r)^{1/2},$$

$$TA(\gamma) = \pi\,\mathrm{sh}\, 2r(2\,\mathrm{sh}^2 r - 1)/(\mathrm{sh}^4 r - \mathrm{sh}^2 r)^{1/2},$$

$$TE(\gamma) = 4\sqrt{2}\pi\mu\,\mathrm{sh}^2 r.$$

III. ENERGY AND AREA

Clearly, $4\sqrt{2}\pi\mu$ is the (unrealized) greatest lower bound of $TE(\gamma)$ as γ ranges over Gödel circles of radius $r > r_C$. In this section we prove that it is actually the greatest lower bound as γ ranges over *all* closed timelike curves. The first step in the argument is to give $TE(\gamma)$ an intuitive geometric interpretation.

In what follows let γ be some closed timelike curve, let S be some $t = \text{const}$, $y = \text{const}$ submanifold of M, and let γ^* be the closed (at least piecewise smooth) curve that results from projecting γ into S. Notice first that since γ is closed, we have (using our coordinate expression for E)

$$TE(\gamma) = 2\sqrt{2}\mu \int_\gamma \mathrm{sh}^2 r\, d\varphi.$$

The integrand on the right depends only on r and φ. So we may perform the integration over γ^* rather than γ. Thus

$$TE(\gamma) = 2\sqrt{2}\mu \int_{\gamma^*} \mathrm{sh}^2 r\, d\varphi.$$

We can evaluate the right-hand integral using Stokes' theorem. Let S be assigned the orientation, say, determined by the field $(\partial/\partial\varphi)^m$. Assume for the moment that γ^* is a simple (i.e., non-self-intersecting) curve. Then it forms the boundary of an (oriented) region G in S, and we have

$$\int_{\gamma^*} \mathrm{sh}^2 r\, d\varphi = \int_G d(\mathrm{sh}^2 r\, d\varphi)$$

$$= \int_G \mathrm{sh}\, 2r\, dr\, d\varphi = \frac{1}{2\mu^2}\int_G dA,$$

where dA is the area element $2\mu^2\,\mathrm{sh}\, 2r\, dr\, d\varphi$ on S. Now notice that the formula, suitably interpreted, holds even in the case where γ is allowed to be self-intersecting. For in this case γ^* can be decomposed as a "sum" of simple closed curves, and we can associate with it a corresponding sum G of oriented regions bounded by these curves. To extend the formula we simply apply it to each element in the sum, and add.

In what follows, "area" should be understood in the extended sense of "signed, summed area." On that understanding we can formulate our conclusion as follows.

Proposition 1: Let γ be a closed timelike curve, and let G be the (oriented, summed) region obtained by projecting γ into any $t = \text{const}$, $y = \text{const}$ submanifold S. Then

$$TE(\gamma) = (\sqrt{2}/\mu) \cdot \text{the area of } G.$$

Now we determine a greatest lower bound for the right-hand side of the equation. We do so using an "isoperimetric inequality." Consider any complete two-dimensional Riemannian manifold of constant curvature k. Let L and A, respectively, be the length of, and area enclosed by, a (possibly self-intersecting) closed curve in the manifold. Then

$$L^2 \geqslant (4\pi - kA)A,$$

and equality holds iff the curve is a circle.[9] (It follows that of all closed curves of given length, area is strictly maximized by circles.) The case of interest to us is that in which $k = -1/\mu^2$.

Let γ, γ^*, and G be as above, let ξ^m be the four-velocity of γ, and let σ^m be the component of ξ^m orthogonal to $(\partial/\partial t)^m$ and $(\partial/\partial y)^m$. Then

$$-\sigma^m\sigma_m = E^2 - E_y{}^2 - 1,$$

where $E_y = \xi^m(\partial/\partial y)_m/2\mu$. So if L is the length of γ^* and A is the area of G, we have (by Proposition 1)

$$L = \int_\gamma (E^2 - E_y{}^2 - 1)^{1/2}\, ds < \int_\gamma E\, ds = (\sqrt{2}/\mu)A.$$

Combining our two inequalities (with $k = -1/\mu^2$) we arrive at our principal result.

Proposition 2: Let γ, γ^*, and G be as above. Let L be the length of γ^*, and let A be the area of G. Then

(a) $A > 4\pi\mu^2$ and $L > 4\sqrt{2}\pi\mu$.

Hence (by Proposition 1),

(b) $TE(\gamma) > 4\sqrt{2}\pi\mu$.

Given our previous remarks about Gödel circles, it follows that $4\sqrt{2}\pi\mu$ is the greatest lower bound of $TE(\gamma)$ as γ ranges over all closed timelike curves. It also follows that the two lower bounds in (a) are greatest. For this we need only observe that Gödel circles γ of radius $r > r_C$ have area and length

$$A = (\mu/\sqrt{2})TE(\gamma) = 4\pi\mu^2\,\mathrm{sh}^2 r,$$

$$L = \int_\gamma (E^2 - 1)^{1/2}\, ds = (E^2 - 1)^{1/2} PT(\gamma) = 2\pi\mu\,\mathrm{sh}\, 2r.$$

We can think of clause (a) as asserting, simply, that no closed timelike curve has an associated area, after projection, that is as small as the area of a disk of critical radius r_C (or a

length, after projection, as small as the circumference of that disk).

IV. GÖDEL'S REMARK

In a paper devoted to a discussion of the philosophical significance of his discoveries in general relativity, Gödel cites a calculation of "fuel requirements" for travel along closed timelike curves in his universe:

"Basing the calculation on a mean density of matter equal to that observed in our world, and assuming one were able to transform matter completely into energy, the weight of the "fuel" of the rocket ship, in order to complete the voyage in t years (as measured by the traveler), would have to be of the order of magnitude of $10^{22}/t^2$ times the weight of the ship (if stopping, too, is effected by recoil). This estimate applies to $t \ll 10^{11}$ yr. Irrespective of the value of t, the velocity of the ship must be at least $1/\sqrt{2}$ of the velocity of light."[2]

It seems likely that Gödel was considering time travel along Gödel circles, and calculated the fuel required to accelerate from zero velocity to the velocities characteristic of those circles, and then back again.[10] (Here "velocity" is understood to mean "speed relative to matter lines.") That is why he can refer to *the* (unchanging) velocity of the ship.[11] Using Proposition 2, it will be possible for us to recover Gödel's numbers without assuming that the time traveler traverses Gödel circles (or sections thereof).

We make use of a lemma[12] that connects total acceleration to changes in energy value.

Lemma 3: Let γ be a timelike curve connecting points p and q. Then

$$TA(\gamma) > |\ln E(q) - \ln E(p)|.$$

[Here, of course, $E(q)$ is the value of E that γ assumes at q.]

Proof: Let $g'_{mn} = g_{mn} - \xi_m \xi_n$ be the (negative definite) metric that results from projecting g_{mn} orthogonal to ξ^m. Since η^m is a Killing field, we have

$$\frac{dE}{ds} = \xi^n \nabla_n E = \xi^n \xi^m \nabla_n \eta_m + \eta_n \alpha^n = \eta_n \alpha^n = g'_{mn} \eta^m \alpha^n.$$

Hence, by the Schwarz inequality (applied to $-g'_{mn}$),

$$\begin{aligned}\left|\frac{dE}{ds}\right| &= |-g'_{mn} n^m \alpha^n| \\ &\leqslant (-g'_{mn}\alpha^m\alpha^n)^{1/2}(-g'_{mn}\eta^m\eta^n)^{1/2} \\ &= a(E^2-1)^{1/2} < aE.\end{aligned}$$

So $a > |d(\ln E)/ds|$, and therefore

$$TA(\gamma) = \int_\gamma a\, ds > |\ln E(q) - \ln E(p)|. \quad \blacksquare$$

The corollary we now state concerns closed timelike curves that have initial (and perhaps final) four-velocity η^m. They represent the trajectories of time travelers who start out (and perhaps end up) at rest relative to the major mass points of the Universe.

Corollary 4: Let γ be a closed timelike curve.

(a) If γ has initial four-velocity η^m, then

$$TA(\gamma) > |\ln(4\sqrt{2}\pi\mu/PT(\gamma))|.$$

(b) If γ has *both* initial and final four-velocity η^m, then

$$TA(\gamma) > 2|\ln(4\sqrt{2}\pi\mu/PT(\gamma))|.$$

Proof: Let p be the initial (= terminal) point of γ, and let q be a point on γ at which E achieves its *average* value. By Proposition 2,

$$E(q)\cdot PT(\gamma) = \int_\gamma E\, ds > 4\sqrt{2}\pi\mu.$$

Let $E(p^+)$ and $E(p^-)$ be the initial and terminal values of E at p. In case (a) we have $E(p^+) = 1$, and the assertion follows immediately if we apply the lemma to that stretch of γ running from p to q. In case (b) we have $E(p^-) = 1$ as well, and so we can apply the lemma, in addition, to the return stretch of γ running from q back to p. $\blacksquare$

Now we establish the connection between total acceleration and "fuel consumption."[12] Suppose γ represents the trajectory of a point particle "rocket ship." Let m be its mass, and J^n the energy momentum of its exhaust. Let us assume that the rocket is suitably isolated during its trip. (It is not refueled, nor hit by meteors.) Then the energy momentum of the rocket's exhaust must balance precisely the rate at which the rocket itself loses energy momentum, i.e.,

$$J^n = -\xi^p \nabla_p (m\xi^n) = -(\xi^n \xi^p \nabla_p m + m\alpha^n).$$

And the mass of the rocket must be nonincreasing (i.e., $\xi^p \nabla_p m \leqslant 0$) since the rocket is consuming fuel. Hence, since J^n is causal (i.e., $J^n J_n \geqslant 0$),

$$a \leqslant (-\xi^n \nabla_n m)/m = -d(\ln m)/ds.$$

Let m_p be the mass of the rocket's payload (the rocket with empty fuel tanks), and let m_f be the mass of the fuel with which it starts. Assuming that the rocket arrives with empty fuel tanks, we have (by integration)

$$(m_p + m_f)/m_p \geqslant e^{TA(\gamma)}.$$

Now let us insert some numbers. Recall that the parameter μ is correlated with cosmic mass density ρ by the relation $\rho = 1/(16\pi\kappa\mu^2)$. If we take for ρ the value 10^{-30} g/cm^3 (the estimated mass density of our universe), then $\mu \approx 10^{10}$ yr, and $4\sqrt{2}\pi\mu \approx 10^{11}$ yr. Hence, in our two cases (a) and (b), assuming $PT(\gamma) \ll 10^{11}$ yr,

case (a): $m_f/m_p \geqslant 10^{11}/PT(\gamma)$,

case (b): $m_f/m_p \geqslant 10^{22}/(PT(\gamma))^2$

[where $PT(\gamma)$ is given in years].

V. A CONJECTURE

Corollary 4 applies only to closed timelike curves γ that are initially tangent to matter lines. And even within this restricted class, it places no lower bound on $TA(\gamma)$. It leaves open the possibility that $TA(\gamma)$ can be made arbitrarily small if $PT(\gamma)$ is allowed to be arbitrarily large. (A sufficiently patient time traveler might not need much fuel for his rocket ship.)

It seems natural to ask what the greatest lower bound of $TA(\gamma)$ is as γ ranges over *all* closed timelike curves. Let GLB be this number.[13] In earlier work we showed that GLB > 0.[3] It now seems to us overwhelmingly likely that GLB $= 2\pi(9 + 6\sqrt{3})^{1/2} \approx 28$. (This would yield a fuel consumption ratio m_f/m_p larger than 10^{12}.)

One arrives at that particular number by considering Gödel circles. As noted in Sec. II., Gödel circles γ of radius $r > r_C$ have total acceleration

$$TA(\gamma) = \pi \operatorname{sh} 2r(2 \operatorname{sh}^2 r - 1)/(\operatorname{sh}^4 r - \operatorname{sh}^2 r)^{1/2}.$$

This expression assumes a minimal value of $2\pi(9 + 6\sqrt{3})^{1/2}$ when $\operatorname{sh}^2 r = (1 + \sqrt{3})/2$.

One might hope to prove the conjecture using ideas related to those in Sec. III, i.e., by reducing it to an assertion about closed curves in the hyperbolic plane, and then invoking the "isoperimetric inequality" (or something similar). But we have not been able to do so. The best we have done so far,[14] is to show that *Gödel circles of the required special radius are the only closed timelike curves that minimize total acceleration against local variation*. So if the value GLB is realized by any closed timelike curve, the conjecture must be true. It seems overwhelmingly likely that the value *is* realized (because of the nature of the sectional curvatures of the Gödel metric).

ACKNOWLEDGMENTS

We wish to thank David Garfinkle, Lee Lindblom, and especially Robert Geroch for many helpful discussions.

This work was supported by the National Science Foundation (SES85-21343).

[1]See, for example, K. Gödel, Rev. Mod. Phys. **21**, 447 (1949); W. Kundt, Z. Phys. **145**, 611 (1956); S. Chandrasekhar and J. P. Wright, Proc. Natl. Acad. Sci. USA **47**, 341 (1961); H. Stein, Philos. Sci. **37**, 589 (1970); J. Pfarr, Gen. Relativ. Gravit. **13**, 1073 (1981); and other references cited by Pfarr.

[2]K. Gödel, "A remark about the relationship between relativity theory and idealistic philosophy," in *Albert Einstein: Philosopher–Scientist*, edited by P. Schilpp (Open Court, La Salle, IL, 1949).

[3]To make the paper self-contained, we here (and in Sec. V) include some material previously presented in D. Malament, J. Math. Phys. **26**, 774 (1985).

[4]In Gödel's original paper he interpreted his model as a solution to Einstein's equation

$$R_{mn} - \tfrac{1}{2}g_{mn}R + \lambda g_{mn} = 8\pi\kappa\rho'\eta_m\eta_n$$

with cosmological constant $\lambda = 1/(2\mu^2)$, for a pressureless fluid source with mass density $\rho' = 1/(8\pi\kappa\mu^2)$.

[5]This statement is slightly delicate. The Killing field η^m is not hypersurface orthogonal. So strictly speaking, the metric h_{mn} does not live *on* S (or on any other submanifold of M). However, we can here invoke a canonical one-to-one (tensor operation preserving) correspondence between tensor fields on S and fields on M that are (i) orthogonal to $(\partial/\partial y)^m$ and η^m in all indices, and (ii) Lie derived by $(\partial/\partial y)^m$ and η^m. [See the Appendix in R. Geroch, J. Math Phys. **12**, 918 (1971).] That h_{mn} is Lie derived by $(\partial/\partial y)^m$ and η^m follows from the fact that η^m is a Killing field of constant length and $(\partial/\partial y)^m$ is covariantly constant.

[6]One way to see this is the following. Consider new coordinates on S defined by

$$x_1 = \mu \operatorname{ch} 2r, \quad x_2 = \mu \operatorname{sh} 2r \cos\varphi, \quad x_3 = \mu \operatorname{sh} 2r \sin\varphi.$$

Clearly, $x_1 > 0$ and $x_1^2 - x_2^2 - x_3^2 = \mu^2$ for all r and φ. Furthermore, in these coordinates the metric $-h_{mn}$ assumes the form

$$-h_{mn} = -(dx_1)_m(dx_1)_n + (dx_2)_m(dx_2)_n + (dx_3)_m(dx_3)_n.$$

Thus $(S, -h_{mn})$ is isometric to the upper half of a two-sheeted hyperboloid of radius μ in $\mathbb{R}^3$, with respect to the metric induced on the latter by a background flat metric of signature $(-,+,+)$. It is a standard result that this hyperboloid (under the induced metric) is a complete Riemannian manifold with constant curvature $-1/\mu^2$. [See, for example, B. O'Neill, *Semi-Riemannian Geometry* (Academic, New York, 1983).]

Correction: In the paper cited in Ref. 3, we worked throughout with the value $\mu = \frac{1}{2}$, and mistakenly asserted a curvature of $-\frac{1}{4}$, rather than -4. This slip did not affect our argument.

[7]"Timelike curves" will be understood to be future directed and smooth (everywhere) unless they are closed, in which case smoothness will be allowed to fail at initial (= terminal) points.

[8]We derive, e.g., the expression for a. Let $f = 1/[2\mu(\operatorname{sh}^4 r - \operatorname{sh}^2 r)^{1/2}]$. Then $\xi^n = f(\partial/\partial\varphi)^n$. Clearly, $\xi^n\nabla_n f = 0$. Hence, since $(\partial/\partial\varphi)^n$ is a Killing field,

$$\begin{aligned} a_n &= f^2\left(\frac{\partial}{\partial\varphi}\right)^m \nabla_m \left(\frac{\partial}{\partial\varphi}\right)_n \\ &= -f^2\left(\frac{\partial}{\partial\varphi}\right)^m \nabla_n \left(\frac{\partial}{\partial\varphi}\right)_m \\ &= -(f^2/2)\nabla_n[4\mu^2(\operatorname{sh}^4 r - \operatorname{sh}^2 r)] \\ &= -2f^2\mu^2 \operatorname{sh} 2r(2 \operatorname{sh}^2 r - 1)\nabla_n r. \end{aligned}$$

Our expression for a now follows from the fact that $(\nabla_n r) = (-1/4\mu^2)(\partial/\partial r)_n$.

[9]See, for example, C. Bandle, *Isoperimetric Inequalities and Applications* (Pitman, London, 1980), p. 35. The inequality is usually proved only for non-self-intersecting curves. But it can easily be extended to the more general case we are considering (if area is interpreted in the sense explained).

[10]This is the way J. Pfarr (see Ref. 1) reconstructs Gödel's remark.

[11]It is clear on this construal how Gödel arrives at the number $1/\sqrt{2}$. As noted in Sec. II., a Gödel circle with radius $r > r_c$ has energy $E = \sqrt{2} \operatorname{sh}^2 r/(\operatorname{sh}^4 r - \operatorname{sh}^2 r)^{1/2}$, and so velocity $v = \operatorname{ch} r/(\sqrt{2} \operatorname{sh} r) > 1/\sqrt{2}$.

[12]S. Chakrabarti, R. Geroch, and S. Liang, J. Math. Phys. **24**, 597 (1983).

[13]Note that GLB is a dimensionless constant. The scale dependencies of acceleration and elapsed proper time cancel each other.

[14]Joint work with R. Geroch and L. Lindblom (unpublished).

Anisotropic fluid with SU(2)-type structure in general relativity: A model of localized matter

Patricio S. Letelier
Departamento de Física Universidade de Brasília, 70910 Brasília, Brazil

Enric Verdaguer
Departament de Física Teòrica, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain

(Received 25 February 1987; accepted for publication 10 June 1987)

A model of anisotropic fluid with three perfect fluid components in interaction is studied. Each fluid component obeys the stiff matter equation of state and is irrotational. The interaction is chosen to reproduce an integrable system of equations similar to the one associated to self-dual SU(2) gauge fields. An extension of the Belinsky–Zakharov version of the inverse scattering transform is presented and used to find soliton solutions to the coupled Einstein equations. A particular class of solutions that can be interpreted as lumps of matter propagating in empty space-time is examined.

I. INTRODUCTION

Anisotropic fluids are found in general relativity when electromagnetic fields or a viscous term are present.[1] But they may also be found using two perfect fluid components[2–4] or even more components.[5]

Models with multifluid components are increasingly being used in cosmology,[6,7] in the description of collapsing spheres,[8] and in the problem of halo and hole formation[9,10] in expanding universes to represent inhomogeneous zones that develop galaxies and voids.[11]

In the present paper we study a model of anisotropic fluid with three perfect fluid components in interaction. Each fluid component obeys the stiff matter equation of state and is irrotational. The interaction is chosen to reproduce an integrable system of equations similar to the one associated to the Yang equations[12] for self-dual SU(2) gauge fields with axial symmetry. For instance, these equations can be solved using a simple generalization of the Belinsky–Zakharov solution generating technique[13] (BZSGT).

The application of the BZSGT opens the possibility of finding solitonlike solutions for the fluid. In particular, we can describe lumps of matter coupled to gravity propagating in empty space. The description of lumps is greatly simplified in the three-fluid model, since we only need a two-soliton solution, i.e., a scattering matrix with two complex poles.[13] For the two-fluid model we need four complex poles, a fact that makes the analysis of the solutions quite complicated.

These solutions may be generalized to represent the collision of cylindrical lumps which may show some features of the merging of galaxies. These generalizations will be treated in a future paper.

In Sec. II we present a summary of the main formulas for the model of anisotropic fluid with multifluid components, of Ref. 5. In Sec. III we examine a three-fluid model with potentials interacting via a Yang-type of equation. In Sec. IV we study the Einstein equations coupled to the three-fluid model for cylindrically symmetric space-times. In Sec. V we present a class of two-soliton solutions for the self-gravitating anisotropic fluid of Sec. III and analyze a particular subclass that describes the propagation of a lump of matter on a flat space-time. In the Appendix we extend the usual BZSGT valid for symmetric matrices to the case of Hermitian matrices.

II. A MODEL OF ANISOTROPIC FLUID WITH MULTIFLUID COMPONENTS

The stress-energy tensor for the anistropic fluid is formed from the sum of three tensors, each of which is the energy-momentum tensor of a perfect fluid, in the particular case that the fluids' four-velocities are linearly dependent,[5] i.e.,

$$T^{\mu\nu}=\sum_{i=1}^{3} t^{\mu\nu}_{(i)}, \tag{2.1}$$

$$t^{\mu\nu}_{(i)}\equiv(p_i+\rho_i)u^{\mu}_{(i)}u^{\nu}_{(i)}-p_i g^{\mu\nu}. \tag{2.2}$$

The four-velocities $u^{\mu}_{(i)}$ are restricted by

$$u^{\mu}_{(i)}u_{(i)\mu}=1, \tag{2.3}$$

and the existence of quantities b_i different from 0 such that

$$\sum_{i=1}^{3} b_i u^{\mu}_{(i)}=0. \tag{2.4}$$

The functions ρ_i and p_i are the fluids' rest energies and pressures, respectively.

With the transformations[5]

$$\mu^{\mu}_{(1)}\to u^{*\mu}_{(1)}=\cos\phi u^{\mu}_{(1)}+\left(\frac{\beta+\gamma\epsilon_{12}}{\beta+\alpha\epsilon_{12}}\right)^{1/2}\sin\phi u^{\mu}_{(2)}, \tag{2.5a}$$

$$u^{\mu}_{(2)}\to u^{*\mu}_{(2)}=-\left(\frac{\beta+\alpha\epsilon_{12}}{\beta+\gamma\epsilon_{12}}\right)^{1/2}\sin\phi u^{\mu}_{(1)}+\cos\phi u^{\mu}_{(2)}, \tag{2.5b}$$

where

$$\tan(2\phi)=2[(\beta+\alpha\epsilon_{12})(\beta+\gamma\epsilon_{12})]^{1/2}/(\alpha-\beta), \tag{2.6}$$

and the condition (2.4), we find that the energy-momentum tensor (2.1) can be cast in the form

$$T^{\mu\nu}=(\rho+\pi)U^{\mu}U^{\nu}+(\sigma-\pi)\chi^{\mu}\chi^{\nu}-\pi g^{\mu\nu}. \tag{2.7}$$

 0022-2488/87/102431-06$02.50 © 1987 American Institute of Physics

The quantities ϵ_{12}, α, γ, and β are related to the fluid components by

$$\epsilon_{ij} = \epsilon_{ji} \equiv u^{\mu}_{(i)} u_{(j)\mu}, \quad i,j = 1,2,3, \tag{2.8}$$

$$\alpha \equiv (p_1 + \rho_1) + (p_3 + \rho_3)\left(\frac{\epsilon_{13} - \epsilon_{23}\epsilon_{12}}{1 - \epsilon_{12}^2}\right)^2, \tag{2.9}$$

$$\gamma \equiv (p_2 + \rho_2) + (p_3 + \rho_3)\left(\frac{\epsilon_{23} - \epsilon_{12}\epsilon_{13}}{1 - \epsilon_{12}^2}\right)^2, \tag{2.10}$$

$$\beta \equiv (p_3 + \rho_3)\frac{(\epsilon_{13} - \epsilon_{23}\epsilon_{12})(\epsilon_{23} - \epsilon_{12}\epsilon_{13})}{(1 - \epsilon_{12}^2)^2}. \tag{2.11}$$

The symbols U^{μ}, χ^{μ}, ρ, σ, and π represent the fluid flux velocity, the direction of anistropy, the fluid rest energy, the pressure along the anisotropy direction, and the pressure on the "perpendicular directions" to χ^{μ}, respectively. These quantities are related to the perfect fluid components by

$$U^{\mu} = u^{*\mu}_{(1)} / (u^{*\alpha}_{(1)} u^{*}_{(1)\alpha})^{1/2}, \tag{2.12}$$

$$\chi^{\mu} = u^{*\mu}_{(2)} / (-u^{*\alpha}_{(2)} u^{*}_{(2)\alpha})^{1/2}, \tag{2.13}$$

$$\rho = \tfrac{1}{2}(\alpha + \gamma + 2\beta\epsilon_{12} - 2\pi) + \tfrac{1}{2}[(\alpha - \gamma)^2 + 4(\beta + \alpha\epsilon_{12})(\beta + \gamma\epsilon_{12})]^{1/2}, \tag{2.14}$$

$$\sigma = -\tfrac{1}{2}(\alpha + \gamma + 2\beta\epsilon_{12} - 2\pi) + \tfrac{1}{2}[(\alpha - \gamma)^2 + 4(\beta + \alpha\epsilon_{12})(\beta + \gamma\epsilon_{12})]^{1/2}, \tag{2.15}$$

$$\pi = p_1 + p_2 + p_3. \tag{2.16}$$

Also, we have that

$$U^{\mu}U_{\mu} = -\chi_{\mu}\chi^{\mu} = 1, \quad U^{\mu}\chi_{\mu} = 0, \tag{2.17}$$

$$\rho = T^{\mu\nu}U_{\mu}U_{\nu}, \quad \sigma = T^{\mu\nu}\chi_{\mu}\chi_{\nu}. \tag{2.18}$$

In general, it is necessary to add supplementary conditions to close the model; this point was treated in some detail in Ref. 2.

III. A SPECIAL CASE OF MULTIFLUID WITH IRROTATIONAL COMPONENTS

A simple closed model of fluid is obtained by assuming that each fluid four-velocity component is irrotational, i.e.,

$$u^{\mu}_{(i)} = \phi'^{\mu}_{(i)} / (\phi_{(i),\alpha}\phi'^{\alpha}_{(i)})^{1/2}, \tag{3.1}$$

where, as usual, $\phi'^{\mu}_{(i)} = g^{\mu\alpha}\phi_{(i),\alpha}$ and $\phi_{(i),\alpha} = \partial_{\alpha}\phi_{(i)}$; and obeys the simple equation of state

$$p_i = \rho_i = \tfrac{1}{2}\phi_{(1)}^{-2}\phi_{(i),\alpha}\phi'^{\alpha}_{(i)}, \tag{3.2}$$

i.e., each fluid obeys the stiff matter equation of state. Note that for the first component we recover the well-known Tabensky–Taub[14] relations for the stiff fluid in terms of the potential $\Lambda \equiv \ln \phi_{(1)}$. Thus the multifluid with fluid components obeying (3.2) can be considered as the interaction of a Tabensky–Taub fluid with other two irrotational fluids. The case $\phi_{(3)} = 0$ corresponds to the fluid studied in Ref. 4.

The condition of linear dependence of the fluids' four-velocities in this case reads

$$\mathbf{b}\cdot\boldsymbol{\phi}'^{\mu} = 0, \tag{3.3}$$

where we have introduced the notation

$$\mathbf{A}\cdot\mathbf{B} = \sum_{i=1}^{3} A_{(i)}B_{(i)}. \tag{3.4}$$

From (2.1), (3.1), and (3.2) we find

$$T_{\mu\nu} = \phi_{(1)}^{-2}(\boldsymbol{\phi}_{,\mu}\cdot\boldsymbol{\phi}_{,\nu} - \tfrac{1}{2}g_{\mu\nu}\boldsymbol{\phi}_{,\alpha}\cdot\boldsymbol{\phi}^{,\alpha}). \tag{3.5}$$

The energy-momentum tensor (3.5) can also be derived, in the usual way, from the Lagrangian density

$$\mathscr{L} = \tfrac{1}{2}\sqrt{-g}\,\phi_{(1)}^{-2}\boldsymbol{\phi}_{,\alpha}\cdot\boldsymbol{\phi}^{,\alpha}. \tag{3.6}$$

The simpler condition to determine the fields $\phi_{(i)}$ is to impose that they satisfy the Euler–Lagrange equations obtained from (3.6), i.e.,

$$(\sqrt{-g}\,g^{\alpha\beta}\phi_{(1)}^{-2}\phi_{(1),\beta})_{,\alpha} + \sqrt{-g}\,\phi_{(1)}^{-3}\boldsymbol{\phi}_{,\alpha}\cdot\boldsymbol{\phi}^{,\alpha} = 0, \tag{3.7a}$$

$$(\sqrt{-g}\,g^{\alpha\beta}\phi_{(1)}^{-2}\phi_{(2),\beta})_{,\alpha} = 0, \tag{3.7b}$$

$$(\sqrt{-g}\,g^{\alpha\beta}\phi_{(1)}^{-2}\phi_{(3),\beta})_{,\alpha} = 0. \tag{3.7c}$$

The energy-momentum tensor (3.5) obeys

$$\nabla_{\mu}T^{\mu\nu} = 0 \tag{3.8}$$

as a consequence of (3.7). And for each fluid component we have

$$\nabla_{\mu}t^{\mu\nu}_{(i)} \neq 0. \tag{3.9}$$

The relations (3.8) and (3.9) tell us that the whole fluid is a closed system with "internal" fluid components in interaction. Also, the anisotropic fluid is completely determinated by the fields $\phi_{(i)}$ and their evolution equations (3.7), i.e., no other extra equation like an equation of state is needed. As a matter of fact, the anisotropic fluid is completely determined by the quantities α, β, γ, π, and ϵ_{12} that in terms of $\phi_{(i)}$ can be written as

$$\alpha = \frac{\lambda_{11}}{\phi_{(1)}^2}\left[1 + \left(\frac{\lambda_{13}\lambda_{22} - \lambda_{12}\lambda_{23}}{\lambda_{11}\lambda_{22} - \lambda_{12}^2}\right)^2\right], \tag{3.10}$$

$$\gamma = \frac{\lambda_{22}}{\phi_{(1)}^2}\left[1 + \left(\frac{\lambda_{11}\lambda_{23} - \lambda_{12}\lambda_{13}}{\lambda_{11}\lambda_{22} - \lambda_{12}^2}\right)^2\right], \tag{3.11}$$

$$\beta = \sqrt{\lambda_{11}\lambda_{22}}\,\frac{(\lambda_{22}\lambda_{13} - \lambda_{12}\lambda_{23})(\lambda_{11}\lambda_{23} - \lambda_{12}\lambda_{13})}{\phi_{(1)}^2(\lambda_{11}\lambda_{22} - \lambda_{12}^2)^2}, \tag{3.12}$$

$$\pi = (1/2\phi_{(1)}^2)\boldsymbol{\phi}_{,\mu}\cdot\boldsymbol{\phi}^{,\mu}, \tag{3.13}$$

$$\epsilon_{ij} = \lambda_{ij}/\sqrt{\lambda_{ii}\lambda_{jj}}, \tag{3.14}$$

where

$$\lambda_{ij} = \boldsymbol{\phi}_{(i)}{}^{,\mu}\boldsymbol{\phi}_{(j),\mu}. \tag{3.15}$$

IV. EINSTEIN EQUATIONS COUPLED TO MATTER

The Einstein equations

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = -T_{\mu\nu} \tag{4.1}$$

coupled to the energy-momentum tensor (3.5) are equivalent to

$$R_{\mu\nu} = -\phi_{(1)}^{-2}\boldsymbol{\phi}_{,\mu}\cdot\boldsymbol{\phi}_{,\nu}. \tag{4.2}$$

The integrability of (4.1) is guaranteed by the field equations (3.7).

We shall consider a space-time with the cylindrically symmetric metric

$$ds^2 = e^{\omega}(dt^2 - dr^2) - \gamma_{ab}\,dx^a\,dx^b, \tag{4.3}$$

where the sum convention is assumed in the indices a and b

that take the values 2 and 3; $(x^0,x^1,x^2,x^3)=(t,r,\theta,z)$, γ_{ab} and ω are functions of t and r only, and

$$\gamma\equiv(\gamma_{ab})=\gamma^T, \tag{4.4}$$

$$\det\gamma=t^2. \tag{4.5}$$

From (4.2)–(4.5) and the fact that cylindrical symmetry implies that the $\phi_{(i)}$ are functions of t and r only, we get

$$-(R_{00}+R_{11})$$

$$=\omega_{,t}/t+1/t^2+\tfrac{1}{4}\operatorname{tr}(\gamma_{,t}\gamma_{,t}^{-1}+\gamma_{,r}\gamma_{,r}^{-1}), \tag{4.6a}$$

$$=\phi_{(1)}^{-2}(\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,t}+\boldsymbol{\phi}_{,r}\cdot\boldsymbol{\phi}_{,r}), \tag{4.6b}$$

$$-2R_{01}=\omega_{,r}/t+\tfrac{1}{2}\operatorname{tr}(\gamma_{,t}\gamma_{,r}^{-1}), \tag{4.7a}$$

$$=2\phi_{(1)}^{-2}\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,r}, \tag{4.7b}$$

and

$$(t\gamma_{,t}\gamma^{-1})_{,t}-(t\gamma_{,r}\gamma^{-1})_{,r}=0, \tag{4.8}$$

where $\gamma^{-1}{}_{,\mu}\equiv(\gamma^{-1})_{,\mu}$.

The field equations (3.7) in space-time with the metric (4.3) reduce to

$$\phi_{(1),tt}+\phi_{(1),t}/t-\phi_{(1),rr}+\phi_{(1)}^{-1}(\phi_{(2),t}^2-\phi_{(2),r}^2+\phi_{(3),t}^2-\phi_{(3),r}^2-\phi_{(1),t}^2+\phi_{(1),r}^2)=0, \tag{4.9a}$$

$$(t\phi_{(1)}^{-2}\phi_{(2),t})_{,t}-(t\phi_{(1)}^{-2}\phi_{(2),r})_{,r}=0, \tag{4.9b}$$

$$(t\phi_{(1)}^{-2}\phi_{(3),t})_{,t}-(t\phi_{(1)}^{-2}\phi_{(3),r})_{,r}=0. \tag{4.9c}$$

These three equations can be written in a completely equivalent form as the single matrix equation

$$(tQ_{,t}Q^{-1})_{,t}-(tQ_{,r}Q^{-1})_{,r}=0, \tag{4.10}$$

where

$$Q\equiv\frac{t}{\phi_{(1)}}\begin{pmatrix}1 & \phi_{(2)}-i\phi_{(3)}\\ \phi_{(2)}+i\phi_{(3)} & \boldsymbol{\phi}\cdot\boldsymbol{\phi}\end{pmatrix}. \tag{4.11}$$

Note that

$$Q=Q^\dagger, \tag{4.12}$$

$$\det Q=t^2. \tag{4.13}$$

By using definition (4.11) it is not difficult to prove the following useful identities:

$$\operatorname{tr}(Q_{,r}Q_{,r}^{-1})=-2\phi_{(1)}^{-2}\boldsymbol{\phi}_{,r}\cdot\boldsymbol{\phi}_{,r}, \tag{4.14a}$$

$$\operatorname{tr}(Q_{,t}Q_{,t}^{-1})=-2(t^{-2}+\phi_{(1)}^{-2}\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,t}), \tag{4.14b}$$

$$\operatorname{tr}(Q_{,r}Q_{,t}^{-1})=-2\phi_{(1)}^{-2}\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,r}. \tag{4.14c}$$

From (4.6), (4.7), and (4.14) we get

$$\omega_{,t}=-2/t-(t/4)\operatorname{tr}(\gamma_{,t}\gamma_{,t}^{-1}+\gamma_{,r}\gamma_{,r}^{-1})-(t/2)\operatorname{tr}(Q_{,r}Q_{,r}^{-1}+Q_{,t}Q_{,t}^{-1}), \tag{4.15a}$$

$$\omega_{,r}=-(t/2)\operatorname{tr}(\gamma_{,t}\gamma_{,r}^{-1})-t\operatorname{tr}(Q_{,t}Q_{,r}^{-1}). \tag{4.15b}$$

The existence of ω, i.e., $\omega_{,tr}=\omega_{,rt}$, is a direct consequence of Eqs. (4.8) and (4.10). Thus the solution of the Einstein equations (4.2) for the metric (4.3) reduces to the solution of (4.8) and (4.10), and the computation of a quadrature for the coefficient ω. [Compare Eqs. (4.15).]

In the case under consideration we have that the condition (3.3) is automatically satisfied as a consequence of the dependence of the function $\phi_{(i)}$ in only two variables, t and r. Then, we can have the anisotropic fluid interpretation of the field equations (4.2). We find, after some algebra, that the quantities that appear in (2.8)–(2.12) can be written as

$$\rho=\sigma=\tfrac{1}{2}\phi_{(1)}^{-2}e^{-\omega}|\boldsymbol{\phi}_{,t}-\boldsymbol{\phi}_{,r}|\,|\boldsymbol{\phi}_{,t}+\boldsymbol{\phi}_{,r}|, \tag{4.16}$$

$$\pi=\tfrac{1}{2}\phi_{(1)}^{-2}e^{-\omega}(\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,t}-\boldsymbol{\phi}_{,r}\cdot\boldsymbol{\phi}_{,r}), \tag{4.17}$$

$$U^\mu=\frac{e^{-\omega/2}}{\sqrt{2\theta_0}}\left[\frac{2\theta_2}{(\theta_1-\theta_0)^{1/2}},-(\theta_1-\theta_0)^{1/2},0,0\right], \tag{4.18}$$

$$\chi^\mu=\frac{e^{-\omega/2}}{\sqrt{2\theta_0}}\left[\frac{2\theta_2}{(\theta_1+\theta_0)^{1/2}},-(\theta_1+\theta_0)^{1/2},0,0\right], \tag{4.19}$$

where the vertical bars indicate the usual Euclidean norm and

$$\theta_0=|\boldsymbol{\phi}_{,t}-\boldsymbol{\phi}_{,r}|\,|\boldsymbol{\phi}_{,t}+\boldsymbol{\phi}_{,r}|, \tag{4.20a}$$

$$\theta_1=\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,t}+\boldsymbol{\phi}_{,r}\cdot\boldsymbol{\phi}_{,r}, \tag{4.20b}$$

$$\theta_2=\boldsymbol{\phi}_{,t}\cdot\boldsymbol{\phi}_{,r}. \tag{4.20c}$$

Two useful identities are

$$\alpha+\gamma+2\beta\epsilon_{12}-2\pi=0, \tag{4.21}$$

$$\theta_0^2+4\theta_2^2-\theta_1^2=0. \tag{4.22}$$

Equation (4.21) is a consequence of (3.10)–(3.15) and (4.3), and (4.22) follows from (4.20).

Also, we have

$$\phi_{(1)}=t/Q_{11}, \tag{4.23a}$$

$$\phi_{(2)}=\operatorname{Re}(Q_{12})/Q_{11}, \tag{4.23b}$$

$$\phi_{(3)}=-\operatorname{Im}(Q_{12})/Q_{11}. \tag{4.23c}$$

Expressions (4.16)–(4.19) can also be obtained directly by solving the eigenvalue problem for the tensor (3.5) with the metric (4.3) and $\phi_{(i)}$ functions of t and r only. In other words, the anisotropic fluid interpretation of (4.3) is independent of the identifications (3.1) and (3.2). Thus, even though (3.1) is meaningless in the case that $\phi_{(i),\alpha}\phi_{(i)}{}^{,\alpha}<0$ the anisotropic fluid interpretation of (4.3) is still valid. The only problem that one has is that $\pi<0$. Note that the same problem occurs in the one-fluid case (Tabensky–Taub fluid).[14,15]

Equations (3.7) can also be cast as a matrix equation in terms of the matrix Q whose elements are space-time scalars [cf. Eq. (4.11)],

$$\nabla^\mu(Q_{,\mu}Q^{-1})=0. \tag{4.24}$$

In the case of Euclidean metric $g_{\mu\nu}=\delta_{\mu\nu}$, Eq. (4.24) is closely related to the equation for self-dual SU(2) gauge fields in the Yang gauge.[12] For a cylindrically symmetric Euclidean space-time (4.24) reduces exactly to the Yang equation for axially symmetric instantons.[16] Hence (4.10) can be considered as the hyperbolic version of the Yang equation for self-dual SU(2) gauge fields.

The real case, i.e., $Q=Q^T$ is equivalent to the case studied in Ref. 4. Note that in this case we also have $\rho=\sigma$, i.e., a stiff equation of state in the direction of anisotropy.

V. PARTICULAR SOLITARY WAVE SOLUTION

There are many different techniques used to solve Eq. (4.10), the most commonly used are the Bäcklund transformations and the inverse scattering method. In this section we study a particular solution obtained using an extension of the

Belinsky–Zakharov[13] version of the inverse scattering transform that we present in the Appendix.

We shall focus our attention mainly in the matter content of the solution, for this reason we take the metric as being diagonal along the complete evolution of the spacetime, i.e.,

$$\gamma_0 = t\begin{pmatrix} e^{\Omega} & 0 \\ 0 & e^{-\Omega} \end{pmatrix}. \tag{5.1}$$

The solution for the matter is generated using the Belinsky–Zakharov solution generating technique with the diagonal seed solution

$$Q_0 = t\begin{pmatrix} e^{-\Lambda} & 0 \\ 0 & e^{\Lambda} \end{pmatrix}. \tag{5.2}$$

In this case we have that the Einstein equations (4.8), (4.10), and (4.15) and the metric (4.3) can be written as

$$ds^2 = e^{\omega_{00}}(dt^2 - dr^2) - t(e^{\Omega}\, d\theta^2 + e^{-\Omega}\, dz^2), \tag{5.3}$$

$$\omega_{00} = -\tfrac{1}{2}\ln t + \nu[\Omega] + 2\nu[\Lambda], \tag{5.4}$$

$$\nu[\Omega] \equiv \frac{1}{2}\int t\,[(\Omega_{,t}^2 + \Omega_{,r}^2)dt + 2\Omega_{,t}\Omega_{,r}\, dr], \tag{5.5}$$

$$\Omega_{,tt} + \Omega_{,t}/t - \Omega_{,rr} = 0, \tag{5.6}$$

$$\Lambda_{,tt} + \Lambda_{,t}/t - \Lambda_{,rr} = 0. \tag{5.7}$$

This particular solution to the Einstein equation for a single fluid component, i.e.,

$$\phi_{(1)} = e^{\Lambda}, \tag{5.8a}$$

$$\phi_{(2)} = \phi_{(3)} = 0, \tag{5.8b}$$

obeying the stiff equation of motion $p_1 = \rho_1$ is studied in Refs. 15 and 17. Note that one recovers the vacuum solution (Einstein–Rosen[18] solution) in the case $\Lambda = 0$ ($\phi_{(1)} = 1$) and $\phi_{(2)} = \phi_{(3)} = 0$.

The application of the one-soluton BZSGT to the seed solution (5.2), i.e., to the matter only, yields the solution (see the Appendix)

$$\omega_{01} = \omega_{00} + 2\ln\frac{t^{1/2}|\mu_1|(|p_1|^2|Y_1|^2 + |q_1|^2|Y_1|^{-2})}{[(|\mu_1|^2 - t^2)|\mu_1^2 - t^2|]^{1/2}}, \tag{5.9}$$

$$\phi_{(1)} = e^{\Lambda}\frac{|p_1|^2|Y_1|^2 + |q_1|^2|Y_1|^{-2}}{|p_1|^2|Y_1/X_1| + |q_1|^2|Y_1/X_1|^{-2}}, \tag{5.10a}$$

$$\phi_{(2)} = -e^{\Lambda}\frac{(|X_1|^2 - |X_1|^{-2})\mathrm{Re}(\bar{p}_1 q_1 \bar{Y}_1/Y_1)}{|p_1|^2|Y_1/X_1|^2 + |q_1|^2|Y_1/X_1|^{-2}}, \tag{5.10b}$$

$$\phi_{(3)} = e^{\Lambda}\frac{(|X_1|^2 - |X_1|^{-2})\mathrm{Im}(\bar{p}_1 q_1 \bar{Y}_1/Y_1)}{|p_1|^2|Y_1/X_1|^2 + |q_1|^2|Y_1/X_1|^{-2}}, \tag{5.10c}$$

where

$$Y_k \equiv \exp(F_k - \Lambda/2), \tag{5.11}$$

$$X_k \equiv (\mu_k/t)^{1/2}. \tag{5.12}$$

By doing $\Lambda = q_1 = 0$ we have the solution characterized by

$$\omega_{01} = \frac{1}{2}\ln t + \nu[\Omega] + \ln\frac{|\mu_1 - \bar{\mu}_1|}{|\mu_1^2 - t^2|}, \tag{5.13}$$

$$\phi_{(1)} = |\mu_1/t|, \tag{5.14a}$$

$$\phi_{(2)} = \phi_{(3)} = 0. \tag{5.14b}$$

To derive (5.13) we have assumed that $\mathrm{Im}\,\mu_1 \neq 0$ used the identity

$$(\mu_2 - \mu_1)(\mu_1\mu_2 - t^2) = 2(\alpha_2 - \alpha_1)\mu_1\mu_2, \tag{5.15}$$

and disregarded a constant of integration, a practice that we shall follow without warning.

A more interesting solution is obtained by applying the same one-soliton BZSGT to the diagonal one-soliton already found. We find the two-soliton solution

$$\omega_{02} = \frac{3}{2}\ln t + \nu[\Omega]1 + \ln\frac{|\mu_1 - \bar{\mu}_1||\mu_2 - \bar{\mu}_2|}{|\mu_1^2 - t^2||\mu_2^2 - t^2|} + 2\ln(|p_2|^2|Y_2|^2 + |q_2|^2|Y_2|^{-2}), \tag{5.16}$$

$$\phi_{(1)} = \frac{|\mu_1|}{t}\frac{|p_2|^2|Y_2|^2 + |q_2|^2|Y_2|^{-2}}{|p_2|^2|Y_2/X_2|^2 + |q_2||Y_2/X_2|^{-2}}, \tag{5.17a}$$

$$\phi_{(2)} = -\frac{|\mu_1|}{t}\frac{(|X_2|^2 - |X_2|^{-2})\mathrm{Re}(\bar{p}_2 q_2 \bar{Y}_2/Y_2)}{|p_2|^2|Y_2/X_2|^2 + |q_2|^2|Y_2/X_2|^{-2}}, \tag{5.17b}$$

$$\phi_{(3)} = \frac{|\mu_1|}{t}\frac{(|X_2|^2 - |X_2|^{-2})\mathrm{Im}(\bar{p}_2 q_2 \bar{Y}_2/Y_2)}{|p_2|^2|Y_2/X_2|^2 + |q_2|^2|Y_2/X_2|^{-2}}, \tag{5.17c}$$

where

$$Y_2 = \left(\frac{t(\mu_1 - \mu_2)(\bar{\mu}_1 - \mu_2)}{\mu_2|\mu_1|}\right)^{1/2},$$

and X_2 is given by (5.12).

The two-soliton solution (5.16) and (5.17) is a particular case of the complete two-soliton obtained from the vacuum as seed solution.

The solution can be used to represent localized distributions of matter with cylindrical symmetry propagating in empty space. This is not possible with the one-soliton solution (5.13) and (5.14) because the fluid potential $\phi_{(1)}$ diverges when $r \to \infty$.

From the application of the BZSGT to some cosmological solutions we know[19] that two independent complex pole trajectories are needed in order to obtain localized solutions (gravitational solitons). Moreover since the metric coefficients have to be real, for each complex pole its complex conjugate is also a pole[13]; so that we need four pole trajectories in all. However in the present case, since the matrix Q describing the fluid potentials is not real, localized solutions can be obtained by using two complex pole trajectories only.

The way by which localized solutions are obtained is by taking opposite signs in the square roots of the two pole trajectories μ_1 and μ_2 [see Eq. (A6)]. With such a prescription it is easy to see that the fluid potentials $\phi_{(i)}$ ($i = 1,2,3$), from (5.17), approach the seed values ($\phi_{(1)} = 1$, $\phi_{(2)} = \phi_{(3)} = 0$) in the asymptotic regions in the following way. They approach the seed decreasing like r^{-1} at $t \ll r \to \infty$, they decrease like t^{-1} at $r \ll t \to \infty$, and decrease like $t^{-1/2}$ along the light cone $r \sim t \to \infty$ This behavior is typical of the gravitational solitons in cosmological[20–23] or cylindrical models[24] and is an indication that the anistropic fluid is localized around the light cone $r = t$.

In Fig. 1 the fluid density ρ and pressure along the radial direction $\sigma(=\rho)$ of Eq. (4.16) is represented for the fluid potentials (5.17). We take a negative sign for the square root of μ_1 and a positive sign for that of μ_2. The density is mainly localized around a cylinder that expands at the speed of

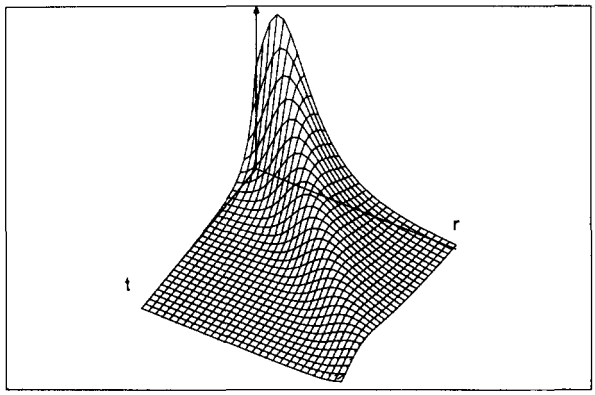


FIG. 1. Density ρ and pressure along the radial direction are given in Eq. (4.16) with the fluid potentials (5.17). The parameters taken are $p_2 = q_2 = 1$ and $\alpha_1 = -0.2i$, $\alpha_2 = -0.1i$. The time and radial coordinates, t and r, are both represented in the range (0.1,1). The relatively large value of $|\alpha_1 - \alpha_2|$ has been taken to avoid sharp picks and to obtain a smooth figure. The background outside the wave is $\rho = \sigma \sim 0$, the space-time axes are drawn on the origin of the vertical axis: $\rho = 0$. The maximum value of ρ is 7.5.

light. The shape of the wave (its amplitude and width) is controlled by the imaginary part of $\alpha_1 - \alpha_2$. The amplitude of the density wave decreases as the wave gets far from the origin. The wave propagates on an essentially empty background ($\rho = 0$).

In Fig. 2 the pressure π, Eq. (4.17), tangent to the cylinder is shown. As for the density we have a wave essentially localized along the light cone which propagates on an empty background ($\pi = 0$). The peculiarity here is that π takes negative values on the region $r \gtrsim t$. As mentioned in Sec. IV the interpretation in terms of the fluid (3.1) is not possible although a fluid interpretation is still valid (see Ref. 15).

This model can be used to represent lumps of matter coupled to gravity propagating on empty space. The qualitative similarity of the waves of matter with the gravitational solitons,[20,21] which are similar to the hydrodynamical solitons, suggests that the collision of lumps of matter might also have a solitonlike character. Models representing such collisions are being considered by the authors.

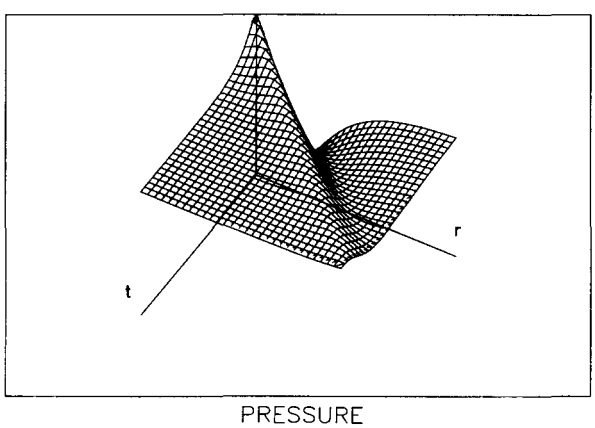


FIG. 2. Pressure π is tangent to the cylinder of Eq. (4.17), with the same parameters as in Fig. 1. The pressure takes positive and negative values in different regions of the space-time. The background value outside the "perturbation" is $\pi \sim 0$; the space-time axes are drawn in the negative region of the vertical axis: $\pi = -5$.

A word should be said about the equality $\rho = \sigma$. Although this is not verified by the general three-fluid case (2.12)–(2.18) it is a feature of the cylindrical case, already present in the two-fluid case.[4] The stiff equation of state verified along the propagation direction seems to be a feature of systems which admit solitons.[22] The reason may be traced to the fact that there is a unique velocity on the system in the direction of propagation of the wave,[24] in this case the speed of sound and the speed of the gravitational field are the same: It is known, for instance,[25] that an initial perturbation with cylindrical symmetry on a perfect fluid coupled to gravity will disperse and form shock waves unless the fluid is stiff.

ACKNOWLEDGMENTS

We want to thank the following institutions for financial support: "Comisión Asesora para la Investigación de España" (EV), "Ministerio de Educación y Ciencia de España" (PSL), "Departament de Física Teòrica de la Universitat Autònoma de Barcelona" (PSL), and "Conselho Nacional de Pesquisas do Brasil" (PSL).

APPENDIX: EXTENSION OF THE BZSGT

In this Appendix we present an extension of the BZSGT for the use of Hermitian matrices. This extension is already known for the elliptic case[26] [SU(2) case]. Since the hyperbolic case can be treated in a completely similar way, we shall give only the results.

The n-soliton solution constructed from the seed solution Q_0 is

$$(Q_n)_{ab} = \prod_{k=1}^{n} \left|\frac{\mu_k}{t}\right| \left((Q_0)_{ab} - \sum_{k,l=1}^{n} \frac{\overline{N}_a^{(l)} (\Gamma^{-1})_{lk} N_b^{(k)}}{\mu_k \bar{\mu}_l} \right) \tag{A1}$$

where, now the indices a and b take the values 1 and 2, and

$$\Gamma_{kl} \equiv \frac{m_a^{(k)} (Q_0)_{ab} \overline{m}_b^{l}}{\mu_k \bar{\mu}_l - t^{(2)}} = \overline{\Gamma}_{lk}, \tag{A2}$$

$$N_a^{(k)} \equiv m_b^{(k)} (Q_0)_{ab}, \tag{A3}$$

$$m_a^{(k)} \equiv m_{0b}^{(k)} M_{ba}^{(k)}, \tag{A4}$$

$$M^{(k)} \equiv \psi_0^{-1}|_{\lambda = \mu_k}, \tag{A5}$$

$$\mu_k = \alpha_k - r \pm [(\alpha_k - r)^2 - t^2]^{1/2}. \tag{A6}$$

The bar denotes complex conjugation, $m_0^{(k)}$ and α_k are sets of complex constants. Here $\psi_0 = \psi_0(t,r,\lambda)$ is the solution to the equations

$$D_t \psi_0 = ((tU_0 + \lambda V_0)/(t^2 - \lambda^2))\psi_0, \tag{A7}$$

$$D_r \psi_0 = ((tV_0 + \lambda U_0)/(t^2 - \lambda^2))\psi_0, \tag{A8}$$

where

$$D_t \equiv \partial_t + (2\lambda t/(t^2 - \lambda^2))\partial_\lambda, \tag{A9}$$

$$D_r \equiv \partial_r + (2\lambda^2/(t^2 - \lambda^2))\partial_\lambda, \tag{A10}$$

$$U \equiv t Q_{0,t} Q_0^{-1}, \quad V = t Q_{0,r} Q_0^{-1}. \tag{A11}$$

The coefficient ω can be explicitly computed, we find

$$\omega_m = \hat{\omega}_0 + 2\ln\left\{ t^{-m^2/2}\left(\prod_{k=1}^{m} |\mu_k|^{m+1}\right) \times\left(\prod_{\substack{k,l \\ k>l}} (|\mu_k-\mu_l||\mu_k-\bar{\mu}_l|)^{-1}\right) \times\left(\prod_{k=1}^{m} \frac{|\mu_k|^2-t^2}{|\mu_k^2-t^2|}\right)^{1/2} \det\Gamma\right\}, \tag{A12}$$

where $\hat{\omega}_0$ denotes the ω function of the seed solution (metric and matter).

For the diagonal seed (5.2) Eqs. (A7) and (A8) can be solved along the pole trajectories, we get[27]

$$\psi_0|_{\lambda=\mu_k} = (2\alpha_k\mu_k)^{1/2}\begin{pmatrix} \exp(-F_k) & 0 \\ 0 & \exp F_k \end{pmatrix}, \tag{A13}$$

where

$$F_k \equiv \int \frac{t}{2\mu_k}\left((\mu_{k,t}\Lambda_{,t} + \mu_{k,r}\Lambda_{,r})dt + (\mu_{k,t}\Lambda_{,r} + \mu_{k,r}\Lambda_{,t})dr\right). \tag{A14}$$

The existence of F_k is a consequence of (5.7) and that $\ln\mu_k$ satisfies the same Eq. (5.7).

For the diagonal seed solution case[27] the expressions (A2)–(A5) take the simple form,

$$m_1^{(k)} = p_k(\mu_k)^{-1/2}\exp F_k, \tag{A15}$$

$$m_2^{(k)} = q_k(\mu_k)^{-1/2}\exp(-F_k), \tag{A16}$$

$$N_1^{(k)} = p_k t(\mu_k)^{-1/2}\exp(F_k-\Lambda), \tag{A17}$$

$$N_2^{(k)} = q_k t(\mu_k)^{-1/2}\exp(-(F_k-\Lambda)), \tag{A18}$$

$$\Gamma_{kl} = \frac{t\left[p_k\bar{p}_k\exp(F_k+\bar{F}_l-\Lambda) + q_k\bar{q}_k\exp(-(F_k+\bar{F}_l-\Lambda))\right]}{(\mu_k\bar{\mu}_l - t^2)(\mu_k\bar{\mu}_l)^{1/2}}, \tag{A19}$$

where

$$q_k \equiv m_{01}^{(k)}/(2\alpha_k)^{1/2}, \tag{A20}$$

$$q_k \equiv m_{02}^{(k)}/(2\alpha_k)^{1/2}. \tag{A21}$$

Note that the usual BZSGT relations, valid for real as well as for complex poles are obtained by letting $\bar{\mu}_k \to \mu_k$, $\bar{m}_{0b}^{(k)} \to m_{0b}^{(k)}$, $\bar{F}_l \to F_l$, $p_k \to \bar{p}_k$, and $q_k \to \bar{q}_k$ in (A1), (A2), (A12), and (A19).

[1] A. Lichnerowicz, *Relativistic Hydrodynamics and Magnetohydrodyamics* (Benjamin, New York, 1967).
[2] P. S. Letelier, Phys. Rev. D **22**, 807 (1980).
[3] P. S. Letelier and R. Machado, J. Math. Phys. **22**, 827 (1981); P. S. Letelier, Nuovo Cimento B **69**, 145 (1982).
[4] P. S. Letelier, Phys. Rev. D **26**, 2623 (1982).
[5] P. S. Letelier and P. S. C. Alencar, Phys. Rev. D **34**, 343 (1986).
[6] S. S. Bayin, Phys. Rev. D **26**, 1262 (1982); Astrophys. J. **303**, 101 (1986).
[7] M. S. Madsen and G. Bailler, Astrophys. Space Sci. **114**, 203 (1985).
[8] L. Herrera, G. J. Ruggeri, and L. Witten, Astrophys. J. **234**, 1094 (1979); M. Cosenza, L. Herrera, M. Esculpi, and L. Witten, Phys. Rev. D **25**, 2527 (1982).
[9] D. W. Olson and J. Silk, Astrophys. J. **233**, 395 (1979).
[10] P. J. E. Peebles, Astrophys. J. **257**, 438 (1982).
[11] M. A. Hausman, D. W. Olson, and B. D. Roth, Astrophys. J. **270**, 351 (1983).
[12] C. N. Yang, Phys. Rev. Lett. **38**, 1377 (1977).
[13] V. A. Belinsky and V. E. Zakharov, Zh. Eksp. Teor. Fiz. **75**, 1955 (1978); **77**, 3 (1979) [Sov. Phys. JETP **48**, 985 (1978); **50**, 1 (1979)].
[14] R. Tabensky and A. H. Taub, Commun. Math. Phys. **29**, 61 (1973).
[15] P. S. Letelier and R. Tabensky, Nuovo Cimento B **28**, 407 (1975); P. S. Letelier, J. Math. Phys. **16**, 1488 (1975).
[16] L. Witten, Phys. Rev. D **19**, 718 (1979).
[17] J. Wainwright, W. C. W. Ince, and B. J. Marshman, Gen. Relativ. Gravit. **10**, 259 (1979); P. S. Letelier, J. Math. Phys. **20**, 2078 (1979).
[18] See, for instance, J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1966).
[19] B. J. Carr and E. Verdaguer, Phys. Rev. D **28**, 2995 (1983).
[20] J. Ibañez and E. Verdaguer, Phys. Rev. Lett. **51**, 1313 (1983).
[21] J. Ibañez and E. Verdaguer, Phys. Rev. D **31**, 251 (1985).
[22] J. Ibañez and E. Verdaguer, Astrophys. J. **306**, 401 (1986).
[23] E. Verdaguer, in *Observational Theoretical Aspects of Relativistic Astrophysics and Cosmology*, edited by J. L. Sanz and L. J. Goicoechea (World Scientific, Singapore, 1985).
[24] X. Fustero and E. Verdaguer, Gen. Relativ. Gravit. **18**, 1141 (1986).
[25] E. P. Liang, Astrophys. J. **204**, 235 (1976).
[26] P. S. Letelier, J. Math. Phys. **23**, 1175 (1982); **27**, 615 (1986).
[27] P. S. Letelier, J. Math. Phys. **25**, 2675 (1984); **26**, 467 (1985); **27** 564 (1986).

A geometric interpretation for the Dirac field in curved space

D. Ranganathan
Department of Physics, Indian Institute of Technology, Delhi, New Delhi, 110 029, India

(Received 7 August 1986; accepted for publication 10 June 1987)

The imposition of the condition of length invariance on a Weyl manifold that does not lead uniquely to general relativity is shown. Rather, in this limit, the Weyl vector field can be interpreted as a Dirac current. The action is also the same as the Einstein Dirac one, if and only if, the spinor field is anticommuting. The allowed interactions are greatly restricted. They are only minimal gauge couplings and Yukawa interactions with a scalar field transforming according to the rules of Utiyama [Prog. Theor. Phys. **53**, 565 (1975)].

I. INTRODUCTION

Recently, Tavakol and Van den Bergh[1] suggested that the postulates underlying general relativity[2] permit more than just general relativity when the postulate of length invariance under translation is imposed. We give an explicit example of this that is also of physical relevance.

One of the earliest attempts to unify electromagnetism and general relativity was that of Weyl.[3] In his model, invariance of lengths under translations was given up to be replaced by covariance of four-vector lengths under translations. This weakening of restrictions allows extra structure to be associated with the manifold in the form of an intrinsic vector field. Weyl sought to identify this with the electromagnetic potential. There are some difficulties with this interpretation.

(a) As pointed out by Einstein,[3] lengths of measuring rods or clock rates will depend on their history.

(b) There is an ambiguity in the choice of sign of this field.[4] With one choice, polar vectors transported clockwise around a loop will contract while, if they traverse the same path anticlockwise, they expand. With the opposite choice of sign of the intrinsic vector field, the above statement is reversed, clockwise transport causes expansion, etc.

(c) An essential ingredient of scale covariant theories is that of invariance under the conformal group. This implies that only massless fields can exist in such a space-time.

Subsequently, work has been done to circumvent these problems. Utiyama[5] suggested incorporating a scalar field in a special fashion to maintain gauge invariance. Nishioka[6] has shown how to generate mass for the vector field using this. Lucey[4] also independently discovered the scalar field method and has shown that if it is chosen to be a doublet, the parity ambiguity (b) mentioned above can also be circumvented. Other interpretations have also been suggested.[7,8]

The common feature of all these approaches is that scale covariance is retained and additional terms are introduced in order to achieve it.

If instead scale ("length") invariance under translations is imposed on a Weyl manifold, then it does not trivially reduce to general relativity. The vector field (a doublet) remains and is shown to be equivalent to a Dirac current in the next section. The additional terms in the Einstein–Hilbert action due to the intrinsic vector field are shown to reproduce the Dirac action in Sec. III. It is also shown there that the interactions are greatly restricted by the condition of length invariance. In the last section we summarize our results and remark on some possible consequences.

II. EQUIVALENCE OF THE INTRINSIC WEYL FIELD AND THE DIRAC CURRENT

The fundamental additional assumption of Weyl was that lengths are covariant under translation. That is, a four-vector of magnitude l under translation through dx^μ changes in magnitude (length) by dl, where

$$dl = l\phi_\mu\, dx^\mu, \quad A^\mu A_\mu = l^2. \tag{1}$$

Here, A^μ is any four-vector field and ϕ_μ is the Weyl vector field intrinsic to the manifold. Such a manifold is no longer pseudo-Riemannian and the connection on it is given by[3]

$$\bar{\Gamma}^{\mu}_{\nu\lambda} = \Gamma^{\mu}_{\nu\lambda} + g^{\mu\sigma}(g_{\nu\sigma}\phi_\lambda + g_{\lambda\sigma}\phi_\nu - g_{\nu\lambda}\phi_\sigma), \tag{2}$$

where $\Gamma^{\mu}_{\nu\lambda}$ is the connection on the corresponding pseudo-Riemannian manifold (i.e., if $\phi_\nu = 0$). We adopt the convention from now on of denoting the quantities pertaining to the Weyl manifold by an overbar and omitting it in the corresponding pseudo-Riemannian case. Thus to obtain length invariance, we must have

$$g^{\mu\sigma}(g_{\nu\sigma}\phi_\lambda + g_{\lambda\sigma}\phi_\nu - g_{\nu\lambda}\phi_\sigma) = 0. \tag{3}$$

The trivial solution $\phi_\nu = 0$, of course, always exists. Nontrivial solutions, if they exist, must be representable as spinors, as these form the fundamental representation of the symmetry group on the tangent space at every point if the equivalence principle is to be satisfied. Thus after projection onto the local tetrad,[9] we have

$$g \to \eta, \quad \phi_\alpha = \sigma_\alpha^{AB'}\zeta_A\bar{\zeta}_{B'}, \quad \delta^\alpha{}_\beta = \tfrac{1}{2}\sigma^\alpha{}_{AB'}\sigma_\beta{}^{AB'}. \tag{4}$$

Note that we have restricted ourselves to product representations as objects on a Weyl manifold are necessarily lightlike. Substituting from the above in Eq. (3),

$$\sigma^\alpha_{AB'}\epsilon^{AC}\epsilon^{B'D'}(\tfrac{1}{2}\sigma_{\beta CD'}\sigma_{\gamma EF'} + \tfrac{1}{2}\sigma_{\gamma CD'}\sigma_{\beta EF'} - \eta_{\beta\gamma}\epsilon_{CE}\epsilon_{D'F'})\zeta^E\bar{\zeta}^{F'} = 0. \tag{5}$$

This is satisfied for nontrivial ϕ_μ if the factor within the brackets identically vanishes. Let

$$T_{\beta\gamma AB'CD'} = \tfrac{1}{2}\sigma_{\beta AB'}\sigma_{\gamma CD'} + \tfrac{1}{2}\sigma_{\gamma AB'}\sigma_{\beta CD'} - \eta_{\beta\gamma}\epsilon_{AC}\epsilon_{B'D'}. \tag{6}$$

We use the identity[9]

 0022-2488/87/102437-03$02.50 © 1987 American Institute of Physics

$$T^{\cdots\ \cdots}_{\cdots AB\cdots} = T^{\cdots\ \cdots}_{\cdots (AB)\cdots} + \tfrac{1}{2}\epsilon_{AB} T^{\cdots C\cdots}_{\cdots C\cdots} \tag{7}$$

to rewrite Eq. (6) as

$$T_{ABCD'} = T_{(AC)(B'D')} + \tfrac{1}{2}\epsilon_{AC} T^{E}{}_{EB'D'} + \tfrac{1}{2}\epsilon_{B'D'} T_{AC}{}^{F'}{}_{F'} + \tfrac{1}{4}\epsilon_{AC}\epsilon_{B'D'} T^{EF'}{}_{EF'}. \tag{8}$$

The last term identically vanishes. For the symmetric term to vanish, the Weyl current must necessarily be the direct sum of two independent intrinsic fields of opposite parity. In this case the middle two terms also vanish as they can be rewritten as the Dirac algebra as follows:

$$\phi_\alpha = [\xi_C \quad \bar{\xi}_{B'}] \begin{bmatrix} 0 & \sigma_\alpha{}^{CD'} \\ \eta_{\alpha\beta}\sigma^{\beta}{}_{AB'} & 0 \end{bmatrix} \begin{bmatrix} \xi_A \\ \bar{\xi}_{D'} \end{bmatrix} = \bar{\psi}\gamma_\alpha\psi. \tag{9}$$

Thus we see that nontrivial solutions are indeed possible in the intuitively obvious case. Namely, this is when the Weyl manifold possesses two intrinsic fields of opposite parity, so that one promptly undoes any changes in length brought about by the other. This example may seem rather trivial but we see that the effects of this action lead to a possible physical interpretation.

III. THE ACTION

As we have imposed length invariance, the simplest possible action with the full symmetry of the tangent space group is linear in the Ricci scalar. From Ref. 1 this is

$$S = \int \bar{R}\sqrt{-\bar{g}}\, d^4x = \int (R - \phi^{\mu}{}_{;\mu} - \phi^{\mu}\phi_\mu)\sqrt{-g}\, d^4x. \tag{10}$$

First consider the covariant derivative term for the Weyl field. When we project onto the local tetrad the derivative in terms of spinor components is[10]

$$\partial^\mu = V^\mu{}_\alpha\, \partial^\alpha = V^\mu{}_\alpha \sigma^\alpha{}_{AB'}\, \partial^{AB'}. \tag{11}$$

As we have two Weyl fields of opposite handedness, the derivative operators acting on them must have opposite sign (due to the opposite orientation of the basis when projected back from spinor to tetrad consistently). Thus we see that instead of vanishing as a total derivative from the action, we have

$$\phi^{\mu}{}_{;\mu} \to \bar{\psi}\overleftarrow{\gamma^\mu\partial_\mu}\psi. \tag{12}$$

Next consider the quadratic term in the Weyl fields. Keeping in mind the fact that we wish to quantize the field at some stage, we rewrite this as

$$\int \phi^\mu\phi_\mu\sqrt{-g}\, d^4x = \int\int \rho(\Delta x)\phi^\mu(x+\Delta x) \times \phi_\mu(x)\sqrt{-g}\, d^4x\, d^4(\Delta x), \tag{13}$$

where $\rho(\Delta x)$ is an appropriate smearing function normalized to unity. We now confine ourselves to equal time separations and substitute from (9),

$$\int \phi^\mu\phi_\mu\sqrt{-g}\, d^4x = \int\int \rho(\Delta x)\bar{\psi}(x+\Delta x)\gamma^\mu\psi(xt\Delta x) \times \bar{\psi}(x)\gamma_\mu\psi(x)\sqrt{-g}\, d^4x\, d^4(\Delta x). \tag{14}$$

Imposing the canonical anticommutation relations

$$\{\psi(\mathbf{x}+\Delta\mathbf{x},t), \bar{\psi}(\mathbf{x},t)\} = C\delta^3(\Delta x), \tag{15}$$

we see that Eq. (13) becomes

$$\int \phi^\mu\phi_\mu\sqrt{-g}\, d^4x = 4C\int \bar{\psi}\psi\sqrt{-g}\, d^4x. \tag{16}$$

Thus substituting from (12) and (15) in (10), the action is

$$S = \int (R - (\bar{\psi}\overleftarrow{\gamma^\mu\partial_\mu}\psi + 4C\bar{\psi}\psi))\sqrt{-g}\, d^4x. \tag{17}$$

We see, as promised, that under length invariance the Weyl action is equivalent to the Einstein–Dirac one. We can now use the appropriate Green's function in (15) to consistently extend our derivation to arbitrary infinitesimal separations.

Next consider the case of possible interactions. All these must obey our cardinal principle of length invariance, Eq. (3). We can easily see that (3) is preserved under local unitary transformations

$$\psi \to U(x)\psi = \left[S\exp\left(i\int A(x)dx\right)\right]\psi, \tag{18}$$

where S is a constant unitary matrix and A is a Hermitian matrix field. This is, of course, the gauge principle and thus gauge interactions are allowed.

The only other permissible interactions are with a scalar field introduced in the fashion of Utiyama. This has a very interesting consequence, the anticommutation relations in the presence of such a scalar field Ω are

$$\{\psi(\mathbf{x}+\Delta\mathbf{x},t), \bar{\psi}(\mathbf{x},t)\} = \Omega(\mathbf{x},t)\delta^3(\Delta x). \tag{19}$$

When using this instead of Eq. (14), we obtain a Yukawa interaction in the Lagrangian and the spinor field remains massless. This is the standard prescription for a Higgs mechanism and indeed as Nishioka[6] has already shown the scalar field possesses the necessary nonlinearity.

IV. CONCLUSIONS

We have shown here that the condition of length invariance usually imposed on Weyl space[1] has nontrivial consequences. As this extra structure (the Dirac field) is observed it seems reasonable to retain it and the corresponding geometrical interpretation. If the rule of introducing two Weyl fields seems artificial it does have some justification. In the path integral quantization of gravity we take a sum over manifolds. Here if CPT is to be a good symmetry, we expect the wave function to be a coherent sum over Weyl spaces of opposite parity contributing equally.

For this it is normally convenient to add an extra boundary term to the action.[11] An extension of the methods used here leads to these giving rise to Weyl spinors. Details of this and extensions of these results to space-times of higher dimensions will be demonstrated elsewhere.

ACKNOWLEDGMENTS

Support of the Department of Electronics, Government of India during the course of this work is gratefully acknowledged.

[1] R. K. Tavakol and N. Van den Bergh, Phys. Lett. A **112**, 23 (1985).
[2] J. Ehlers, F. A. E. Pirani, and A. Schild, in *General Relativity*, edited by L. O. Raifeartaigh (Oxford U.P., Oxford, 1972).

[3]H. Weyl and S. Preuss, Akad. Wiss. Math. Kl., 465 (1918). A summary of this, Einstein's remarks, and other early work are given by R. Adler, M. Bazin, and M. Schiffer, in *Introduction to General Relativity* (McGraw-Hill, New York, 1965).
[4]C. A. Lucey, Phys. Rev. D **33**, 346 (1986).
[5]R. Utiyama, Prog. Theor. Phys. **53**, 565 (1975).
[6]M. Nishioka, Fortschr. Phys. **88**, 241 (1985).
[7]E. Santamato, Phys. Rev. D **29**, 216 (1984).
[8]R. Godfrey, Phys. Rev. Lett. **52**, 1365 (1984).
[9]S. A. Hugget and K. P. Tod, *An Introduction to Twistor Theory* (Cambridge U.P., Cambridge, 1985), p. 15.
[10]N. D. Birrel and P. C. W. Davies, *Quantum Fields on Curved Space* (Cambridge U.P., Cambridge, 1984), p. 85.
[11]G. W. Gibbons and S. W. Hawking, Phys. Rev. D **15**, 2738 (1977).

Bell's inequalities and quantum field theory. I. General setting

Stephen J. Summers
Department of Mathematics, University of Rochester, Rochester, New York 14627
Reinhard Werner
Fachbereich Physik, Universität Osnabrück, D-4500 Osnabrück, Federal Republic of Germany

(Received 22 July 1986; accepted for publication 27 May 1987)

Bell's inequalities are briefly presented in the context of order-unit spaces and then studied in some detail in the framework of C^*-algebras. The discussion is then specialized to quantum field theory. Maximal Bell correlations $\beta(\phi,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))$ for two subsystems localized in regions $\mathscr{O}_1$ and $\mathscr{O}_2$ and constituting a system in the state ϕ are defined, along with the concept of maximal Bell violations. After a study of these ideas in general, properties of these correlations in vacuum states of arbitrary quantum field models are studied. For example, it is shown that in the vacuum state the maximal Bell correlations decay exponentially with the product of the lowest mass and the spacelike separation of $\mathscr{O}_1$ and $\mathscr{O}_2$. This paper is also preparation for the proof in Paper II [S. J. Summers and R. Werner, J. Math. Phys. **28**, 2448 (1987)] that Bell's inequalities are maximally violated in the vacuum state.

I. INTRODUCTION

Since Bell stated[1,2] the first special case of what has come to be called Bell's inequalities, quite a bit of theoretical and experimental work has been invested in the attempt (1) to clarify the content of the inequalities, i.e., to find the proper framework within which to formulate Bell's inequalities and to determine the consequences of their violation or nonviolation, and (2) to design and carry out experimental tests of these inequalities. For partial reviews of this work, see Refs. 3–5. This paper has the following objectives. (a) We wish to briefly present a formulation of Bell's inequalities in the context of order-unit spaces and then to study them in some detail in the framework of C^*-algebras, in preparation for the discussion of Bell's inequalities and quantum field theory. (b) Then we specialize to a study of Bell's inequalities in relativistic quantum field theory, formulated in the most general axiom system known to us—the so-called algebraic quantum field theory of Haag, Kastler, and Araki[6,7] (which subsumes, at least modulo certain regularity conditions,[8] standard quantum field theories satisfying the Wightman axioms[9,10]). Although we shall have something to say about Bell's inequalities in arbitrary states of the system, our main discussion here will concern the vacuum state. (c) Finally, we shall prove that Bell's inequalities are maximally violated in the vacuum state by suitable observables localized in spacelike separated regions of space-time for both Bose and Fermi free quantum field theories. Some of these results have been announced in Ref. 11. Points (a) and (b) will be presented in this paper while (c) constitutes Paper II.[12]

Bell's inequalities concern results of correlation experiments, and in Sec. II we begin with that which one is confronted with in correlation experiments—preparations, measurements (observables) on two subsystems, and the relative frequency of their outcomes. Following the approach due to Ludwig,[13,14] we model these with what we call correlation dualities $(\hat{p},\mathscr{A},\mathscr{B})$, which are comprised of two order unit spaces $\mathscr{A}$ and $\mathscr{B}$ (real vector spaces with a vector ordering $\geqslant$ and a unit 1), and a bilinear function $\hat{p}$: $\mathscr{A}\times\mathscr{B}\to\mathbb{R}$. The observables of one subsystem are represented by partitions $\{a_i\,|\,i\in I\}$ of the unit in $\mathscr{A}$: $\Sigma_i a_i = 1$ with $a_i \geqslant 0$ for each $i\in I$, and similarly for the other subsystem and $\mathscr{B}$. Each $i\in I$ is interpreted as a possible outcome of the measurement of the observable. The probability (relative frequency) of the joint occurrence of the outcomes $i\in I$ and $j\in J$ in the two subsystems, respectively, is then $\hat{p}(a_i,b_j)$.

In the course of Sec. II we introduce what we call the maximal Bell correlation $\beta(\hat{p},\mathscr{A},\mathscr{B})$ for an arbitrary given correlation duality $(\hat{p},\mathscr{A},\mathscr{B})$ as mentioned above. It is defined by

$$\beta(\hat{p},\mathscr{A},\mathscr{B}) \equiv \tfrac{1}{2}\sup(\hat{p}(x_1,y_1)+\hat{p}(x_1,y_2) + \hat{p}(x_2,y_1)-\hat{p}(x_2,y_2)),$$

where the supremum is taken over all $x_i\in\mathscr{A}$, $y_i\in\mathscr{B}$ with $-1_{\mathscr{A}}\leqslant x_i\leqslant 1_{\mathscr{A}}$ and $-1_{\mathscr{B}}\leqslant y_i\leqslant 1_{\mathscr{B}}$, $i=1,2$. Then the Clauser–Horne[15] (CH) form of Bell's inequalities can be rewritten as

$$\beta(\hat{p},\mathscr{A},\mathscr{B}) = 1. \tag{1.1}$$

In this form an implicit symmetry in the CH-Bell's inequalities is made explicit and some additional calculational advantages accrue, as well. We also determine inequalities on $\beta(\hat{p},\mathscr{A},\mathscr{B})$ that serve the same metatheoretical purpose as (1.1), but for larger classes of theories. In particular,

$$\beta(\hat{p},\mathscr{A},\mathscr{B}) \leqslant 2 \tag{1.2}$$

must hold for any triple $(\hat{p},\mathscr{A},\mathscr{B})$ as described above, and

$$\beta(\hat{p},\mathscr{A},\mathscr{B}) \leqslant \sqrt{2} \tag{1.3}$$

must hold for any theory (such as quantum mechanics or quantum field theory) in which the order unit spaces $\mathscr{A},\mathscr{B}$ modeling the observables of the subsystems are actually C^*-algebras. Theorem 2.1 gives some general results of this nature and characterizes the elements $x_i\in\mathscr{A}$, $y_i\in\mathscr{B}$ for which the maximum $\sqrt{2}$ in (1.3) can actually be attained.

Thus $\beta(\hat{p},\mathscr{A},\mathscr{B}) = \sqrt{2}$ is the maximum possible value

 0022-2488/87/102440-08$02.50 © 1987 American Institute of Physics

for the Bell correlations in quantum mechanics, and evidently its attainment would be a violation of Bell's inequalities (1.1). It is known[3] that quantum mechanics predicts triples $(\hat{p},\mathscr{A},\mathscr{B})$ for which $\sqrt{2}$ is attained, and it is known, up to experimental error, that in nature[4] $\sqrt{2}$ is attained.

In Sec. III we discuss the triples $(\hat{p},\mathscr{A},\mathscr{B})$ that arise in quantum field theory. There the observables that can be measured in a space-time region $\mathscr{O}\subset\mathbb{R}^4$ are modeled by self-adjoint elements of a C^*-algebra $\mathscr{A}(\mathscr{O})$, and a given quantum field model provides a net of such algebras $\{\mathscr{A}(\mathscr{O})\}_{\mathscr{O}\subset\mathbb{R}^4}$ satisfying axioms that are naturally motivated by general physical principles. Section IV occupies itself with a general study of the maximal Bell correlation $\beta(\hat{p},\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))$ when $\hat{p}$ arises from a vacuum state. Among other things, we give *a priori* bounds on $\beta(\hat{p},\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))$ in the vacuum in terms of the space-like distance between $\mathscr{O}_1$ and $\mathscr{O}_2$.

This paper is preparation for the proof that also quantum field theory predicts the maximal violation of Bell's inequalities, i.e., it predicts $\beta(\hat{p},\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))=\sqrt{2}$ for certain localization regions $\mathscr{O}_1$ and $\mathscr{O}_2\subset\mathbb{R}^4$ and certain states $\hat{p}$. In Paper II (Ref. 12) we show, among other things, that if $\mathscr{O}_1$ and $\mathscr{O}_2$ are so-called complementary wedge regions in space-time and if $\hat{p}$ arises from the vacuum state for a Bose or Fermi free quantum field theory [$\mathscr{A}(\mathscr{O}_1)$ and $\mathscr{A}(\mathscr{O}_2)$ are then the observable algebras for the corresponding free field theory], then $\beta(\hat{p},\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))=\sqrt{2}$, indeed. In further work in progress, we intend to show such a prediction holds for theories with interaction and for states other than the vacuum state, as well (see Note added in proof).

II. BELL'S INEQUALITIES

The aim of this section is to establish notation and the basic results on Bell's inequalities in general that we shall need in our discussion of Bell's inequalities in quantum field theory. We are obliged here to assume that the reader has prior familiarity with the discourse on Bell's inequalities in the literature. However, for a detailed discussion of the most general formulation and derivation of Bell's inequalities known to us (which is embedded in an approach to statistical theories due to Ludwig[13,14]) and the connection with that with which one is presented in the experimental situation—preparations, measurements, and relative frequencies of their outcomes—see Refs. 16 and 17, the latter being the preprint of the original version of this paper.

Bell's inequalities are a constraint on the statistical correlations between measurements performed at two sites A and B or on two "parts" A and B of the same system. Following Ludwig[13,14,16,17] we shall assume that the possible measurements at site A are described by an order-unit space[18,19] $(\mathscr{A},\geqslant,1)$, abbreviated by $\mathscr{A}$, which is a vector space $\mathscr{A}$ ordered by a convex cone $\mathscr{A}_+\equiv\{a\in\mathscr{A}\,|\,a\geqslant 0\}$ with a distinguished element $1\in\mathscr{A}_+$ whose multiples eventually dominate every other element of $\mathscr{A}_+$. Preparations correspond to positive, normalized linear functionals on $\mathscr{A}$, called (statistical) states on $\mathscr{A}$. An important subclass of theories with this structure is constituted by classical theories, for which $\mathscr{A}$ is the set of continuous, real-valued functions on a compact space X with pointwise ordering [i.e., $f\geqslant g$ if $f(x)\geqslant g(x)$ for all $x\in X$]. The points $x\in X$ are in one-to-one correspondence with the pure (i.e., extremal states in the convex set of) states on $\mathscr{A}$. In quantum mechanics $\mathscr{A}$ is the set of bounded Hermitian operators on a Hilbert space with the usual operator ordering, and in the quantum field theoretical setting described below in Sec. III $\mathscr{A}$ will be the Hermitian part of a C^*-algebra[20] with identity and with its usual ordering $\mathscr{A}_+\equiv\{a^*a\,|\,a\in\mathscr{A}\}$. In this general setting a measurement with finitely many possible outcomes $i\in I$ is formalized by a finite family $\{a_i\}_{i\in I}$ with $a_i\in\mathscr{A}_+$ and $\Sigma_{i\in I}a_i=1$. A preparation is represented by a statistical state ω on $\mathscr{A}$, and $\omega(a_i)$ is the probability for obtaining the result i in an experiment with preparing and measuring devices represented by ω and $\{a_i\}_{i\in I}$, respectively.

A set of correlation experiments is then described by the following structure.

Definition: A *correlation duality* consists of two order-unit spaces $\mathscr{A}$ and $\mathscr{B}$ together with a bilinear functional $\hat{p}$: $\mathscr{A}\times\mathscr{B}\to\mathbb{R}$ such that $a\in\mathscr{A}$, $b\in\mathscr{B}$, and $a,b\geqslant 0$ imply $\hat{p}(a,b)\geqslant 0$ and $\hat{p}(1,1)=1$.

Then $\hat{p}(a_i,b_j)$ is the probability for obtaining both the result i at site A and j at site B for measuring devices described by $\{a_i\}_{i\in I}\subset\mathscr{A}_+$ and $\{b_j\}_{j\in J}\subset\mathscr{B}_+$ and a preparing procedure described by $\hat{p}$. In the C^*-algebraic setting (which actually subsumes all the examples mentioned above) $\mathscr{A}$ and $\mathscr{B}$ are typically elementwise commuting subalgebras of a larger algebra $\mathscr{C}$ and $\hat{p}$ is given by a state ω on $\mathscr{C}$ by $\hat{p}(a,b)\equiv\omega(ab)$.

A crucial assumption about the correlation experiment being modeled is built into this structure. Consider two measuring devices $\{a_i\}_{i\in I}$ and $\{a'_j\}_{j\in J}$ at A. Then by definition $1=\Sigma a_i=\Sigma a'_j$, so that for any $b\in\mathscr{B}$, $\Sigma_i\hat{p}(a_i,b)=\Sigma_j\hat{p}(a'_j,b)=\hat{p}(1,b)$. Thus the probability for a certain outcome at B does not depend on the measuring devices chosen at A. This is the typical "locality" assumption in derivations of Bell's inequalities. We emphasize that this assumption is not to be confused with locality in the sense of relativistic causality. For the correlation dualities studied in this paper Bell's locality is indeed a consequence of Einsteinian causality, because the full causal structure of Minkowski space is built into algebraic quantum field theory (see Sec. III). But the locality assumption in the definition of correlation dualities is valid in a much broader context. In particular, both the classical and quantum mechanical schemes for describing composite systems lead naturally to correlation dualities.

The measurements considered in the standard versions[3,15,21] admit two outcomes, say $\{+,-\}$, and are thus given by pairs $\{a_+,a_-\}\subset\mathscr{A}$ with $a_+\geqslant 0$, $a_-\geqslant 0$, and $a_++a_-=1$. Clearly such pairs are in one-to-one correspondence with the elements $a\in\mathscr{A}$ with $-1\leqslant a\leqslant 1$, by setting $a_\pm=\frac{1}{2}(1\pm a)$. Two such measurements are considered for each of the two sites A,B. We shall say that (a_1,a_2,b_1,b_2) is an admissible quadruple if $a_1,a_2\in\mathscr{A}$, $b_1,b_2\in\mathscr{B}$ and $-1_{\mathscr{A}}\leqslant a_i\leqslant 1_{\mathscr{A}}$, $i=1,2$ and $-1_{\mathscr{B}}\leqslant b_j\leqslant 1_{\mathscr{B}}$, $j=1,2$.

Definition: If $(\hat{p},\mathscr{A},\mathscr{B})$ is a correlation duality and (a_1,a_2,b_1,b_2) is an admissible quadruple, the latter is said to satisfy Bell's inequality if

$$|\hat{p}(a_1,b_1)+\hat{p}(a_1,b_2)+\hat{p}(a_2,b_1)-\hat{p}(a_2,b_2)|\leqslant 2. \quad (2.1)$$

It is a triviality to see that (2.1) is equivalent to the version of Bell's inequality presented in Ref. 15. This inequality is usually derived assuming both $\mathscr{A}$ and $\mathscr{B}$ are classical.

The following Theorem is the basic result of this section.

Theorem 2.1: Let $(\hat{p},\mathscr{A},\mathscr{B})$ be a correlation duality, let $\omega\in\mathscr{A}^*$ be the state $\omega(a)\equiv\hat{p}(a,1)$, and let (a_1,a_2,b_1,b_2) be an admissible quadruple. Setting

$$\chi=\tfrac{1}{2}|\hat{p}(a_1,b_1+b_2)+\hat{p}(a_2,b_1-b_2)|,$$

one has the following: (1) $\chi\leqslant 2$. (2a) If $\mathscr{A}$ is the Hermitian part of a C^*-algebra, then $\chi\leqslant\sqrt{2}$. (2b) If $\chi=\sqrt{2}$ in this case, the following identities hold for all $a\in\mathscr{A}$ and $i=1,2$: $\omega([a_i,a])=0, \omega(a_i^2a)=\omega(a), \omega((a_1a_2+a_2a_1)a)=0$. (3) If any one of the following conditions holds, then $\chi\leqslant 1$. (a) $\mathscr{A}$ is classical. (b) ω is pure on $\mathscr{A}$. (c) There are states $\xi_\alpha\in\mathscr{A}^*$, $\eta_\alpha\in\mathscr{B}^*$ and positive reals λ_α such that for all $a\in\mathscr{A}$, $b\in\mathscr{B}$, $\hat{p}(a,b)=\Sigma\lambda_\alpha\xi_\alpha(a)\eta_\alpha(b)$.

Proof: (1) If $-1\leqslant a\leqslant 1$ and $-1\leqslant b\leqslant 1$, then $1-\hat{p}(a,b)=\frac{1}{2}\hat{p}(1+a,1-b)+\frac{1}{2}\hat{p}(1-a,1+b)\geqslant 0$, and similarly $1+\hat{p}(a,b)\geqslant 0$, so that each of the four summands in χ is bounded by $\frac{1}{2}$.

(2a) Let $\{\pi_\omega,\mathscr{H}_\omega,\Omega\}$ be the cyclic Gel'fand–Naimark–Segal (GNS) representation[20] of $\mathscr{A}$ associated with the state ω. Then for each $b\in\mathscr{B}$ with $-1\leqslant b\leqslant 1$, the equation $\omega_b(a)\equiv\hat{p}(a,b)$ defines a linear functional ω_b on $\mathscr{A}$ with $-\omega\leqslant\omega_b\leqslant\omega$. Thus there exists a unique $\tilde{b}\in\pi_\omega(\mathscr{A})'$ [the commutant of $\pi_\omega(\mathscr{A})$ in $\mathscr{B}(\mathscr{H}_\omega)$] with $-I\leqslant\tilde{b}\leqslant I$ (I is the identity operator on $\mathscr{H}_\omega$) such that $\hat{p}(a,b)=\omega_b(a)=\langle\Omega,\pi_\omega(a)\tilde{b}\Omega\rangle$ for all $a\in\mathscr{A}$. Let $A\equiv\frac{1}{2}\pi_\omega(a_1+ia_2)$ and $B\equiv(1/2\sqrt{2})(\tilde{b}_1+\tilde{b}_2+i(\tilde{b}_1+\tilde{b}_2))$. Then

$$A^*A+AA^*=\tfrac{1}{2}\pi_\omega(a_1^2+a_2^2)\leqslant 1$$

and

$$B^*B+BB^*=\tfrac{1}{2}(\tilde{b}_1^2+\tilde{b}_2^2)\leqslant 1.$$

Moreover,

$$\begin{aligned}\sqrt{2}\chi&=4\,\mathrm{Re}\langle\Omega,A^*B\Omega\rangle\\&=2\,\mathrm{Re}\langle A\Omega,B\Omega\rangle+2\,\mathrm{Re}\langle B^*\Omega,A^*\Omega\rangle\\&=\|A\Omega\|^2+\|B\Omega\|^2-\|(A-B)\Omega\|^2+\|B^*\Omega\|^2\\&\quad+\|A^*\Omega\|^2-\|(B^*-A^*)\Omega\|^2\\&\leqslant\langle\Omega,(A^*A+AA^*+B^*B+BB^*)\Omega\rangle\leqslant 2. \quad (2.2)\end{aligned}$$

(2b) Suppose that equality obtains in (2.2). Then $A\Omega=B\Omega$, $A^*\Omega=B^*\Omega$, and $\langle\Omega,\frac{1}{2}\pi_\omega(a_1^2+a_2^2)\Omega\rangle=\langle\Omega,(A^*A+AA^*)\Omega\rangle=1$. Since $a_i^2\leqslant 1$, this implies $\pi_\omega(a_i^2)\Omega=\Omega$ for $i=1,2$ as well as $(A^*A+AA^*)\Omega=\Omega$. Hence for arbitrary $a\in\mathscr{A}$,

$$\omega(aa_i^2)=\langle\Omega,\pi_\omega(a)\pi_\omega(a_i^2)\Omega\rangle=\omega(a)$$

and

$$\begin{aligned}\omega(a(a_1+ia_2))&=2\langle\Omega,\pi_\omega(a)A\Omega\rangle=2\langle\Omega,\pi_\omega(a)B\Omega\rangle\\&=2\langle B^*\Omega,\pi_\omega(a)\Omega\rangle=2\langle A^*\Omega,\pi_\omega(a)\Omega\rangle\\&=\omega((a_1+ia_2)a).\end{aligned}$$

Since $\tilde{b}_1^2\Omega=\tilde{b}_2^2\Omega=\Omega$, we have

$$\begin{aligned}\pi_\omega(a_1a_2+a_2a_1)\Omega&=(2/i)(A^2-A^{*2})\Omega\\&=(2/i)(B^2-B^{*2})\Omega\\&=(\tilde{b}_1^2-\tilde{b}_2^2)\Omega=0,\end{aligned}$$

which implies the remaining, final assertion of (2b).

(3a) Since a is Abelian, the four elements

$$a_{\epsilon_1\epsilon_2}\equiv\tfrac{1}{4}(1+\epsilon_1a_1)(1+\epsilon_2a_2)$$

with $\epsilon_1,\epsilon_2\in\{+,-\}$ are positive. By direct computation

$$\begin{aligned}\chi&=\hat{p}(a_{++},b_1)+\hat{p}(a_{+-},b_2)-\hat{p}(a_{-+},b_2)-\hat{p}(a_{--},b_1)\\&\leqslant\hat{p}(a_{++}+a_{+-}+a_{-+}+a_{--},1)=\hat{p}(1,1)=1.\end{aligned}$$

(3c) Since $\hat{p}(1,1)=1$ and $\xi_\alpha(1_{\mathscr{A}})=\eta_\alpha(1_{\mathscr{B}})=1$, one must have $\Sigma\lambda_\alpha=1$. Hence $\chi=\Sigma\lambda_\alpha\chi_\alpha$ with

$$\chi_\alpha\equiv\tfrac{1}{2}(\xi_\alpha(a_1)\eta_\alpha(b_1+b_2)+\xi_\alpha(a_2)\eta_\alpha(b_1-b_2)),$$

and it suffices to show $\chi_\alpha\leqslant 1$ for each α. This is readily done by introducing the four numbers

$$\tilde{a}_{\epsilon_1\epsilon_2}\equiv\tfrac{1}{4}(1+\epsilon_1\xi_\alpha(a_1))(1+\epsilon_2\xi_\alpha(a_2))$$

and proceeding as in (3a).

(3b) Purity of ω entails that the functionals $\omega_b\in\mathscr{A}$ with $-\omega\leqslant\omega_b\leqslant\omega$ given by $\omega_b(a)=\hat{p}(a,b)$ are of the form $\omega_b(a)=\eta(b)\omega(a)$ with $-1\leqslant\eta(b)\leqslant 1$ (see Ref. 20). Hence $\hat{p}(a,b)=\omega(a)\eta(b)$ factorizes and one may apply the proof of (3c). □

Part (3) of Theorem 2.1 gives three different conditions on a correlation duality $(\hat{p},\mathscr{A},\mathscr{B})$ such that Bell's inequality is satisfied by *all* admissible quadruples in the correlation duality. By (3c) Bell's inequalities are satisfied even for quantum systems whenever the correlations are produced by a mechanism which can be simulated by a classical random generator (producing the "outcome" α with probability λ_α).

If ω as defined in Theorem 2.1 is a faithful state on $\mathscr{A}$ [i.e., if $x\in\mathscr{A}$, $x\geqslant 0$ and $\omega(x)=0$ imply $x=0$], then the second and third equations in part (2b) are equivalent to $a_i^2=1$ and $a_1a_2+a_2a_1=0$. Thus, if $\chi=\sqrt{2}$ when $\mathscr{A}$ is the Hermitian part of a C^*-algebra, the corresponding elements a_1,a_2, and $a_3\equiv-(i/2)[a_1,a_2]$ form a realization of the Pauli spin matrices in $\mathscr{A}$. The first equation in part (2b) then implies that the state ω restricted to the 2×2 matrix algebra $M_2(\mathbb{C})$ spanned by $1,a_1,a_2,a_3$ is the normalized trace. This is precisely the case realized in the well known idealized description[3] of the Aspect experiment in terms of a singlet state on $M_2(\mathbb{C})\otimes M_2(\mathbb{C})$. Within experimental error (now very small), the maximal Bell correlation (and violation of Bell's inequalities) $\sqrt{2}$ has indeed been found in nature.[4]

Note that since classical, quantum mechanical, and quantum field theoretical models are all subsumed in the C^*-algebraic framework, part (2a) informs us that $\chi=\sqrt{2}$ really is the maximal possible correlation. The bound $\chi\leqslant\sqrt{2}$ (which has also been noted by Cirel'son[22] and others) constrains "local" quantum theoretical descriptions in the same way that Bell's inequality $\chi\leqslant 1$ constrains local classical descriptions. Thus a correlation experiment reliably yielding a result $\chi>\sqrt{2}$ would have to be taken as a falsification of standard quantum mechanics just as Aspect's experiments ex-

clude "local hidden variable theories." Theories predicting $\chi > \sqrt{2}$ in some experiments are not obviously absurd, as can be shown by the straightforward construction of a correlation duality and admissible quadruple with $\chi = 2$. Thus the bound in part (1) is optimal without further assumptions on $(\hat{p},\mathscr{A},\mathscr{B})$.

It is clearly natural to make the following definition.

Definition: The maximal Bell correlation $\beta(\hat{p},\mathscr{A},\mathscr{B})$ in a correlation duality $(\hat{p},\mathscr{A},\mathscr{B})$ is

$$\beta(\hat{p},\mathscr{A},\mathscr{B}) \equiv \tfrac{1}{2}\sup(\hat{p}(a_1,b_1) + \hat{p}(a_1,b_2) + \hat{p}(a_2,b_1) - \hat{p}(a_2,b_2)),$$

where the supremum is taken over all $a_i \in \mathscr{A}$, $b_j \in \mathscr{B}$ with $-1_{\mathscr{A}} \leqslant a_i \leqslant 1_{\mathscr{A}}$ and $-1_{\mathscr{B}} \leqslant b_j \leqslant 1_{\mathscr{B}}$.

Thus, $\beta(\hat{p},\mathscr{A},\mathscr{B}) = 1$ (resp. $\sqrt{2}$,2) means that *every* admissible quadruple in the duality satisfies the CH version of Bell's inequality (resp. the upper bound $\sqrt{2}$, 2 is arbitrarily well approximated by χ's for some admissible quadruples in the correlation duality).

In the remainder of this paper we shall study the existence of admissible quadruples of observables violating Bell's inequalities in correlation dualities naturally arising in algebraic quantum field theory. If $\beta(\hat{p},\mathscr{A},\mathscr{B}) > 1$ we shall say that Bell's inequalities are violated in $(\hat{p},\mathscr{A},\mathscr{B})$, and since we are working henceforth in the C^*-algebraic context, we shall say that the inequalities are *maximally* violated if $\beta(\hat{p},\mathscr{A},\mathscr{B}) = \sqrt{2}$. We shall see in Paper II that quantum field theory predicts the attainment of the maximal violation $\sqrt{2}$. But first we shall discuss in Sec. III the appropriate formulation of quantum field theory and in Sec. IV the special properties of $\beta(\hat{p},\mathscr{A},\mathscr{B})$ when $\hat{p}$ corresponds to the preparation of the vacuum state.

III. BELL'S INEQUALITIES IN ALGEBRAIC QUANTUM FIELD THEORY

In this section we shall specialize the discussion to statistical dualities $(\hat{p},\mathscr{A},\mathscr{B})$ coming from relativistic quantum field theory (QFT). In our view the proper framework within which this can be accomplished is so-called algebraic QFT,[6,7] which has the advantages over the standard QFT[9,10] of being more general and of dealing directly with observables and states.

As already mentioned, the basic structure is an assignment to each open space-time region $\mathscr{O} \subset \mathbb{R}^4$ a C^*-algebra $\mathscr{A}(\mathscr{O})$ (which one can think of as a norm-closed, *-algebra of bounded operators on some Hilbert space), and this assignment must satisfy certain axioms, motivated by physical principles.

(1) Isotony: if $\mathscr{O}_1 \subseteq \mathscr{O}_2$, then $\mathscr{A}(\mathscr{O}_1) \subseteq \mathscr{A}(\mathscr{O}_2)$; with this assumption one can think of each $\mathscr{A}(\mathscr{O})$ as a subalgebra of the C^*-algebra $\mathscr{A}$ generated by $\cup_{\mathscr{O}\subset\mathbb{R}^4}\mathscr{A}(\mathscr{O})$. It is assumed that $\mathscr{A}$ has an identity 1 and that $1 \in \mathscr{A}(\mathscr{O})$ for each $\mathscr{O}$. Here $\mathscr{A}$ is called the quasilocal algebra.

(2) Poincaré covariance: there exists a representation $\{\alpha_\lambda \mid \lambda \in \mathscr{P}^\uparrow_+\}$ of the identity-connected component $\mathscr{P}^\uparrow_+$ of the Poincaré group by a group of automorphisms on $\mathscr{A}$, such that

$$\alpha_\lambda(\mathscr{A}(\mathscr{O})) = \mathscr{A}(\mathscr{O}_\lambda),$$

where $\mathscr{O}_\lambda$ is the image of $\mathscr{O}$ under the transformation corresponding to λ.

(3) Locality: if $\mathscr{O}_1$ is spacelike separated from $\mathscr{O}_2$, then every element of $\mathscr{A}(\mathscr{O}_1)$ commutes with every element of $\mathscr{A}(\mathscr{O}_2)$.

If one considers heuristically each algebra $\mathscr{A}(\mathscr{O})$ as the algebra "generated" by all observables measurable in $\mathscr{O}$, then these assumptions are perfectly natural for a relativistically covariant theory over Minkowski space.

Before proceeding further we wish to remark that axioms (1)–(3) with the algebras $\{\mathscr{A}(\mathscr{O})\}$ Abelian would naturally be fulfilled by classical field theories and hidden-variable theories which have the idea of relativistic locality built into them. It should also be pointed out that axiom (3) can be viewed, at least heuristically, as a consequence of Einsteinian causality, but, in fact, it is strictly weaker (in particular, the locality axiom says nothing about the impossibility of superluminal signals). Axiom (3) is exactly what is required in the algebraic framework in order to satisfy the requirement that observables in spacelike separated regions be jointly measurable in correlation experiments (see Ref. 23).

Continuing now with the axioms of algebraic QFT, the following is also assumed.

(4) Existence of a physical representation: there exists a faithful (i.e., one-to-one) representation π of $\mathscr{A}$ on a separable Hilbert space $\mathscr{H}$ such that on $\mathscr{H}$ there is a nontrivial, strongly continuous, unitary representation $U(\mathscr{P}^\uparrow_+)$ of (the universal covering group of) the Poincaré group $\mathscr{P}^\uparrow_+$ satisfying (a) $U(\lambda)\pi(A)U(\lambda)^{-1} = \pi(\alpha_\lambda(A))$, for each $A \in \mathscr{A}$, $\lambda \in \mathscr{P}^\uparrow_+$, (b) the generators $\{P_\mu\}_{\mu=0}^3$ of the translation subgroup $U(\mathbb{R}^4) \subset U(\mathscr{P}^\uparrow_+)$ satisfy the spectrum condition $P_0^2 - P_1^2 - P_2^2 - P_3^2 \geqslant 0$ and $P_0 \geqslant 0$, where P_0 is the generator of the time translations.

Self-adjoint elements $A \in \mathscr{A}(\mathscr{O})$ of the local algebras are interpreted as observables measurable in the corresponding space-time region $\mathscr{O} \subset \mathbb{R}^4$. In particular, self-adjoint A with $O \leqslant A \leqslant 1$ can be viewed as yes–no observables, i.e., observables corresponding to (equivalence classes of) measuring devices that have only two outcomes. A mathematical state (a positive, normalized linear functional[20]) ϕ on the C^*-algebra $\mathscr{A}$ is supposed to correspond to a physical state of the system whose local observables are represented in the net $\{\mathscr{A}(\mathscr{O})\}$ (although it is not necessary to assume that every such mathematical state is physically realizable). For such a state ϕ and an observable $A \in \mathscr{A}(\mathscr{O})$, $\phi(A)$ is interpreted as the expected value of the observable A of the (statistical) system that has been prepared in the state ϕ. In the case of an observable satisfying $0 \leqslant A \leqslant 1$, $\phi(A)$ is the probability of the outcome "yes" and $\phi(1-A)$ that of the outcome "no" in the state ϕ. Self-adjoint projectors are special cases of such "yes–no" observables.

A C^*-algebra $\mathscr{A}$ with identity is in a natural way an order unit space. The ordering $\geqslant$ is defined as follows: $A \geqslant B$ if and only if $A - B \geqslant O$, and the latter inequality means that there exists a $C \in \mathscr{A}$ such that $A - B = C^*C$. If $\mathscr{A}$ and $\mathscr{B}$ are commuting C^*-algebras and ϕ is a state on a C^*-algebra $\mathscr{C}$ containing both $\mathscr{A}$ and $\mathscr{B}$, then $(\phi,\mathscr{A},\mathscr{B})$ determines a

correlation duality by $\hat{p}(A,B) \equiv \phi(AB)$, for each $A \in \mathscr{A}$, $B \in \mathscr{B}$. Thus, if ϕ is a state on the quasilocal algebra $\mathscr{A}$ generated by a net of local algebras $\{\mathscr{A}(\mathscr{O})\}$, and if $\mathscr{O}_1$ and $\mathscr{O}_2$ are any two spacelike separated regions in Minkowski space, then $(\phi, \mathscr{A}(\mathscr{O}_1), \mathscr{A}(\mathscr{O}_2))$ is a correlation duality, by the axiom of locality.

In Ref. 8 can be found necessary and sufficient conditions that a net of local algebras can be in any way associated to a standard QFT satisfying a weak regularity condition, and the precise sense in which this association can be made is identified. It should be emphasized that it was found that either there is no net associated in any way to the quantum field or this association is very tight. As these matters are technically involved, we shall not try to give any details here. Suffice it to say that necessary and sufficient conditions are determined such that given a standard QFT $\{\mathscr{H}, \varphi(\cdot), U(\mathscr{P}^{\uparrow}_{+}), \Omega\}$ satisfying them there exists a net of local algebras $\{\mathscr{A}(\mathscr{O})\}$ such that for any $f \in \mathscr{S}(\mathbb{R}^4)$ with $\mathrm{supp}(f) \subset \mathscr{O}$, the operator $\varphi(f)$ (on the usual Wightman domain) is affiliated, in the sense of von Neumann, with the algebra $\mathscr{A}(\mathscr{O})$, i.e., the operator $\varphi(f)$ commutes with all elements of $\mathscr{A}(\mathscr{O})'$ [this means, in a well-determined mathematical sense, that all bounded functions of the operators $\varphi(f)$ are in $\mathscr{A}(\mathscr{O})$]. It is shown that, with the assumption of a weak regularity condition, the said net of local algebras is associated in the same way with every element of the Borchers class[9,24] of the field $\varphi(\cdot)$. Thus, typically, to each net $\{\mathscr{A}(\mathscr{O})\}$ are associated many quantum fields, which can be viewed as alternative descriptions of the same physical situation. This fact has also emerged in work[25–29] where, starting from a representation of a net of local algebras, quantum fields are constructed as limits in certain topologies of elements of the algebras $\mathscr{A}(\mathscr{O})$. We remark that when nets of local algebras and quantum fields are found to be associated in the manner suggested, both the nets[8,30] and the fields[31] manifest desirable properties that are not present in the general situation.

After the discussion above and in Sec. II, it should be clear why we consider Bell's inequalities in quantum field theory in the form

$$\beta(\phi, \mathscr{A}(\mathscr{O}_1), \mathscr{A}(\mathscr{O}_2)) = 1, \tag{3.1}$$

for $\phi \in \mathscr{A}_1^{*(+)}$ (the set of states on $\mathscr{A}$), $\mathscr{O}_1$ spacelike separated from $\mathscr{O}_2$, and with $\mathscr{A}(\mathscr{O}_1)$ and $\mathscr{A}(\mathscr{O}_2)$ the corresponding von Neumann algebras (weakly closed C^*-algebras[20]) in a net of local algebras $\{\mathscr{A}(\mathscr{O})\}$. Here $\mathscr{A}(\mathscr{O}_1)$ and $\mathscr{A}(\mathscr{O}_2)$ are interpreted as the algebras generated by the observables measurable in the space-time regions $\mathscr{O}_1$ and $\mathscr{O}_2$ for the two subsystems in a correlation experiment, and ϕ is viewed as the state of the total system as prepared in the given experiment. $\mathscr{O}_1$ and $\mathscr{O}_2$ will be taken to be causally reflexive in the sense that

$$\mathscr{O}_i = \mathscr{O}_i'', \quad i = 1,2. \tag{3.2}$$

Here $\mathscr{O}'$ is defined to be the interior of the set of all points in Minkowski space that are spacelike separated from $\mathscr{O}$, and $\mathscr{O}'' \equiv (\mathscr{O}')'$. In particular, if a measurement is made in $\mathscr{O}$, then $\mathscr{O}''$ is necessarily causally reflexive and is regarded as the causal shadow of the measurement. $\mathscr{O}''$ is the largest (open) space-time region that is spacelike separated from all points spacelike separated from $\mathscr{O}$. We shall typically consider regions from two classes: the wedges $\mathscr{W}$ and the double cones $\mathscr{K}$. The set $\mathscr{W}$ of wedges is defined to be the set of all Poincaré transforms of

$$W_R \equiv \{x \in \mathbb{R}^4 | x_1 > |x_0|\},$$

and the set $\mathscr{K}$ of double cones is defined to be the set of all nonempty intersections of a forward light cone with a backward light cone.

We are interested in finding states ϕ and regions $\mathscr{O}_1$ and $\mathscr{O}_2$ such that $\beta(\phi, \mathscr{A}(\mathscr{O}_1), \mathscr{A}(\mathscr{O}_2)) = \sqrt{2}$, i.e., Bell's inequalities are not only violated, but maximally violated by suitable observables in $\mathscr{A}(\mathscr{O}_1)$ and $\mathscr{A}(\mathscr{O}_2)$ in the state ϕ. In Paper II (Ref. 12) we show that for free quantum field theories, if ϕ is a vacuum state and $\mathscr{O}_1$ and $\mathscr{O}_2$ are wedges satisfying $\mathscr{O}_1 = \mathscr{O}_2'$, then β is indeed equal to $\sqrt{2}$. (In work in progress we have the beginnings of more general results of the type desired.) Thus, with the reservations already mentioned about interpreting elements of local algebras as physical observables, quantum field theory predicts, just as quantum mechanics does, maximal violations of Bell's inequalities.

In the following section we study $\beta(\phi, \mathscr{A}(\mathscr{O}_1), \mathscr{A}(\mathscr{O}_2))$ in some generality when ϕ is a vacuum state and $\mathscr{O}_1 \subset \mathscr{O}_2'$, before going on to Paper II.

IV. BELL'S INEQUALITIES AND THE VACUUM STATE

In this section we specialize the discussion even further—here we shall present results concerning Bell's inequalities in arbitrary vacuum states. The setting we shall work in is as follows. We shall consider a net $\{\mathscr{A}(\mathscr{O})\}_{\mathscr{O} \subset \mathbb{R}^4}$ of concrete C^*-algebras acting on a separable Hilbert space $\mathscr{H}$, on which there exists a strongly continuous, unitary representation $U(\mathbb{R}^4)$ of the translation group satisfying the spectrum condition and acting covariantly upon the elements of $\{\mathscr{A}(\mathscr{O})\}_{\mathscr{O} \subset \mathbb{R}^4}$, i.e.,

$$U(x)\mathscr{A}(\mathscr{O})U(x)^{-1} = \mathscr{A}(\mathscr{O}_x), \quad x \in \mathbb{R}^4, \quad \mathscr{O} \subset \mathbb{R}^4.$$

Moreover, there exists, up to a factor, a unique vacuum vector $\Omega \in \mathscr{H}$, by which we mean a translation-invariant unit vector which is cyclic for $\mathscr{A}$. This entails[32] that $\mathscr{A}$ acts irreducibly on $\mathscr{H}$. We comment that under weak technical assumptions,[30] if the subspace of translation-invariant vectors were more than one dimensional, a "central decomposition" could be performed to reduce the problem to the situation assumed above.

Let ϕ_0 be the state on $\mathscr{A}$ defined by $\phi_0(A) = \langle \Omega, A\Omega \rangle$, $A \in \mathscr{A}$. The crucial characteristic of this vacuum state ϕ_0 is its clustering properties. That is to say, it is known that if $\mathscr{O}_1, \mathscr{O}_2$ are bounded space-time regions and $a \in \mathbb{R}^4$ is any spacelike vector, then

$$\lim_{t \to \infty} \phi_0(U(ta)AU(ta)^{-1}B) = \phi_0(A)\phi_0(B),$$

for any $A \in \mathscr{A}(\mathscr{O}_1)$ and $B \in \mathscr{A}(\mathscr{O}_2)$. In fact, one has the following theorem. Part (a) gives an upper bound on the rate of clustering in massless theories, while part (b) provides the (best possible) bound in theories with a mass gap.

Theorem 4.1: (a) (Ref. 33) Under the assumptions above,

$$|\phi_0(U(x)AU(x)^{-1}B)-\phi_0(A)\phi_0(B)|$$
$$\leqslant C(\mathcal{O}_1,\mathcal{O}_2)[x]^{-2}[\|A\Omega\|\cdot\|B^*\Omega\|+\|B\Omega\|\cdot\|A^*\Omega\|$$
$$+(|x_0|/[x]^2)(\|HA\Omega\|\cdot\|B^*\Omega\|$$
$$+\|HB\Omega\|\cdot\|A^*\Omega\|)],$$

for any $A\in\mathcal{A}(\mathcal{O}_1)$ and $B\in\mathcal{A}(\mathcal{O}_2)$ with $A\Omega$ and $B\Omega$ in the domain of H, the self-adjoint generator of the time-translation subgroup of $U(\mathbb{R}^4)$, where $C(\mathcal{O}_1,\mathcal{O}_2)$ is a constant proportional to the volume of the convex closure D_1 of D_0, $x\in D_1'$ and $[x]$ is the shortest spacelike distance between x and D_1, and D_0 is the complement in the hyperplane $\{x\in\mathbb{R}^4|x_0=0\}$ of $(\mathcal{O}_1-\mathcal{O}_2)'\cap\{x\in\mathbb{R}^4|x_0=0\}$; i.e., if $\mathcal{O}_1=\mathcal{O}_2\in\mathcal{K}$ then D_1 is the double cone with base centered at the center of $\mathcal{O}_1$ and with twice the diameter.

(b) (Ref. 34) If, in addition to the assumptions above, the spectrum of H is contained in $\{0\}\cup[m,\infty)$, with $m>0$, then for x a spacelike vector,

$$|\phi_0(U(x)AU(x)^{-1}B)-\phi_0(A)\phi_0(B)|$$
$$\leqslant e^{-md(x,\mathcal{O}_1,\mathcal{O}_2)}\{\|A^*\Omega\|\cdot\|B\Omega\|\cdot\|A\Omega\|\cdot\|B^*\Omega\|\}^{1/2},$$

for any $A\in\mathcal{A}(\mathcal{O}_1)$ and $B\in\mathcal{A}(\mathcal{O}_2)$, where $\mathcal{O}_1$ and $\mathcal{O}_2$ are not restricted to be bounded and $d(x,\mathcal{O}_1,\mathcal{O}_2)$ is the maximal timelike distance $\mathcal{O}_{1,x}$ can be translated before $\mathcal{O}_{1,x}\not\subset\mathcal{O}_2'$.

Thus, roughly speaking, in the massless case clustering goes like R^{-2} and the massive case like e^{-mR}, where R is the spacelike distance between $\mathcal{O}_{1,x}$ and $\mathcal{O}_2$. An immediate consequence of these clustering properties is given in the following corollary for the massive case; the analogous result for the massless case should then be clear.

Corollary 4.2: Under the assumptions of Theorem 4.1(b)

$$\beta(\phi_0,\mathcal{A}(\mathcal{O}_1),\mathcal{A}(\mathcal{O}_2))\leqslant 1+2e^{-md(0,\mathcal{O}_1,\mathcal{O}_2)},$$

where $\mathcal{O}_1,\mathcal{O}_2\subset\mathbb{R}^4$ are arbitrary.

Proof: For any $A_1,A_2\in\mathcal{A}(\mathcal{O}_1)$ and $B_1,B_2\in\mathcal{A}(\mathcal{O}_2)$ with $-1\leqslant A_i\leqslant 1$ and $-1\leqslant B_i\leqslant 1$, $i=1,2$, Theorem 4.1(b) entails

$$|\tfrac{1}{2}\phi_0(A_1(B_1+B_2)+A_2(B_1-B_2))|$$
$$\leqslant 2e^{-md(0,\mathcal{O}_1,\mathcal{O}_2)}+\tfrac{1}{2}|\phi_0(A_1)\phi_0(B_1)+\phi_0(A_1)\phi_0(B_2)$$
$$+\phi_0(A_2)\phi_0(B_1)-\phi_0(A_2)\phi_0(B_2)|.$$

But the supremum over all self-adjoint contractions $A_i\in\mathcal{A}(\mathcal{O}_1)$, $B_j\in\mathcal{A}(\mathcal{O}_2)$, of the expression

$$\phi_0(A_1)\phi_0(B_1)+\phi_0(A_1)\phi_0(B_2)+\phi_0(A_2)\phi_0(B_1)$$
$$-\phi_0(A_2)\phi_0(B_2)$$

[which is a product state over $\mathcal{A}(\mathcal{O}_1)\otimes\mathcal{A}(\mathcal{O}_2)$ evaluated on $A_1\otimes(B_1+B_2)+A_2\otimes(B_1-B_2)$] is 2, by Theorem 2.1. □

Since $\beta(\phi_0,\mathcal{A}(\mathcal{O}_1),\mathcal{A}(\mathcal{O}_2))\geqslant 1$, we have $0\leqslant\beta(\phi_0,\mathcal{A}(\mathcal{O}_1),\mathcal{A}(\mathcal{O}_2))-1\leqslant 2e^{-md(0,\mathcal{O}_1,\mathcal{O}_2)}$ (this slightly improves the estimate given in Ref. 11). Hence, if $d(0,\mathcal{O}_1,\mathcal{O}_2)$ is much larger than a few Compton wavelengths of the lightest particle in the theory, then the maximal Bell violation (if any) in the vacuum state of measurements made in $\mathcal{O}_1$ and $\mathcal{O}_2$ will *necessarily* be too small to be observed. In a theory with massless particles the clustering "rate" can be much smaller, so that $d(0,\mathcal{O}_1,\mathcal{O}_2)$ could in principle be allowed to get much larger before the maximal violation (if any) could be unobservable. However, in that case it would be necessary to have efficient counters for arbitrarily soft photons, because a lower bound ϵ on the photon energies that can be efficiently read serves as a lowest mass ϵ, leading to a similar bound on the maximal possible violation that can be detected in the vacuum by the said counters. The next theorem formalizes this statement. With techniques parallelling those used in the proof of the theorem in Ref. 34, one can prove the following result.

Proposition 4.3: Under the assumptions of Theorem 4.1(a), for any $\epsilon>0$ and $\delta\leqslant e^{-\epsilon\tau\lambda}$, λ to be freely chosen from $\mathbb{R}_+$, and for any $A_i\in\mathcal{A}(\mathcal{O}_1)$ such that $-1\leqslant A_i\leqslant 1$ and $\|E_{[0,\epsilon]}A_{i,0}\Omega\|/\|A_{i,0}\Omega\|\leqslant\delta$, $i=1,2$, then

$$|\tfrac{1}{2}\phi_0(A_1(B_1+B_2)+A_2(B_1-B_2))|$$
$$\leqslant 1+2e^{-\epsilon\tau c(\lambda)}(1+\delta e^{\epsilon\tau\lambda}),$$

where $c(\lambda)=(2/\pi)\tan^{-1}\lambda$, $\tau=d(0,\mathcal{O}_1,\mathcal{O}_2)$, and $B_1,B_2\in\mathcal{A}(\mathcal{O}_2)$ satisfy $-1\leqslant B_i\leqslant 1$.

For any self-adjoint $A\in\mathcal{A}(\mathcal{O})$, $\mathcal{O}\subset\mathbb{R}^4$, and unit vector $\psi\in\mathcal{H}$,

$$\langle\psi,E_{[0,\epsilon]}AE_{[0,\epsilon]}\psi\rangle=\langle E_{[0,\epsilon]}\psi,AE_{[0,\epsilon]}\psi\rangle$$

gives the expectation of A in that part of the state $\langle\psi,\cdot\psi\rangle$ that involves energies less then ϵ. Thus, $\chi_\epsilon(A)\equiv\|E_{[0,\epsilon]}A\Omega\|/\|A\Omega\|$ is an approximate indication of the low-energy response of the measuring device represented by A in the vacuum state $\phi_0(\cdot)$ [note that if $\|E_{[0,\epsilon]}AE_{[0,\epsilon]}\|$ is sufficiently small, then $\chi_\epsilon(A)$ is also small]. If A responds inefficiently to soft photons, then $\chi_\epsilon(A)$ should be small (recall, $A\Omega\neq 0$ unless $A=0$, so χ_ϵ is well defined). If, in particular, for some $\epsilon>0$ one has $\chi_\epsilon(A_i)=0$ for both A_1 and A_2 in $\mathcal{A}(\mathcal{O}_1)$, then from Proposition 4.3,

$$|\tfrac{1}{2}\phi_0(A_1(B_1+B_2)+A_2(B_1-B_2))|\leqslant 1+2e^{-\epsilon d(0,\mathcal{O}_1,\mathcal{O}_2)},$$

for any $B_1,B_2\in\mathcal{A}(\mathcal{O}_2)$ with $-1\leqslant B_i\leqslant 1$. This is exactly analogous to the estimate in Corollary 4.2. If, however, $\chi_\epsilon(A_i)\neq 0$, then the bound given by Proposition 4.3 does *not* decrease to 1 as $\tau\to\infty$.

In the light of the theorems presented in this section, we certainly do not suggest that someone look for violations of Bell's inequalities in the vacuum. But we shall prove in Paper II that, at least in some quantum field models, Bell's inequalities are indeed maximally violated in the vacuum state. Thus, maximal violation is a prediction of such theories, just as it is of quantum mechanics.

Before we close this section, we wish to point out a few additional facts about Bell's inequalities in the vacuum state for theories that are dilatation invariant. The dilatations $\mathbb{R}^4\ni x\to\lambda x,\lambda>0$, form a group of transformations of Minkowski space. A model $\{\{\mathcal{A}(\mathcal{O})\}_{\mathcal{O}\subset\mathbb{R}^4},\mathcal{H},U(\mathbb{R}^4)\}$ of the kind discussed in this section is said to be dilatation invariant if there exists a strongly continuous, unitary representation $\{U(\lambda)\}_{\lambda>0}$ of the dilatation group that acts on $\mathcal{H}$ and satisfies $U(\lambda)\Omega=\Omega$, for all $\lambda>0$, $U(\lambda)\mathcal{A}(\mathcal{O})U(\lambda)^{-1}=\mathcal{A}(\lambda\mathcal{O})$, where $\lambda\mathcal{O}\equiv\{\lambda x|x\in\mathcal{O}\}$, and $U(\lambda)U(x)=U(\lambda x)U(\lambda)$, $\lambda>0$, $x\in\mathbb{R}^4$. The following theorem relates the maximal correlation $\beta(\phi_0,\mathcal{A}(\mathcal{O}_1),\mathcal{A}(\mathcal{O}_2))$ for algebras

associated to certain regions to the maximal correlation for algebras associated to certain other regions, assuming dilatation invariance. We shall need a mild technical assumption. A net of local algebras $\{\mathscr{A}(\mathscr{O})\}_{\mathscr{O}\subset\mathbb{R}^4}$ will be said to have wedge algebras that are locally generated if for each wedge $W\subset\mathbb{R}^4$, $\mathscr{A}(W)$ is equal to the C^*-algebra generated by $\{\mathscr{A}(\mathscr{O})\,|\,\mathscr{O}\in\mathscr{K}$ and $\mathscr{O}\subset W\}$. Nets of local algebras coming from standard quantum fields are known to satisfy this property.[8]

Theorem 4.4: In a dilatation-invariant quantum field theory in which the wedge algebras are locally generated, $\beta(\phi_0,\mathscr{A}(\lambda\mathscr{O}_1),\mathscr{A}(\lambda\mathscr{O}_2))$ is independent of $\lambda>0$ for any $\mathscr{O}_1,\mathscr{O}_2\subset\mathbb{R}^4$. Thus, if W_1 and W_2 are spacelike separated wedges, then

$$\beta(\phi_0,\mathscr{A}(W_1),\mathscr{A}(W_2))=\beta(\phi_0,\mathscr{A}(W_i),\mathscr{A}(W_i')),\qquad i=1,2.$$

Moreover, if $\mathscr{O}_1$ and $\mathscr{O}_2$ are tangent double cones (i.e., spacelike separated double cones whose closures intersect at one point), then for any wedge W such that $\mathscr{O}_1\subset W$ and $\mathscr{O}_2\subset W'$, $\beta(\phi_0,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))=\beta(\phi_0,\mathscr{A}(W),\mathscr{A}(W'))$.

Proof: Since $\mathscr{A}(\lambda\mathscr{O})=U(\lambda)\mathscr{A}(\mathscr{O})U(\lambda)^{-1}$ and $U(\lambda)\Omega=\Omega$ for all $\lambda>0$, it is obvious that $\beta(\phi_0,\mathscr{A}(\lambda\mathscr{O}_1),\mathscr{A}(\lambda\mathscr{O}_2))$ is independent of $\lambda>0$. [Similarly, $\beta(\phi_0,\mathscr{A}(\mathscr{O}_{1,x}),\mathscr{A}(\mathscr{O}_{2,x}))$ is independent of $x\in\mathbb{R}^4$.] Thus,

$$\beta(\phi_0,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))=\lim_{\lambda\downarrow 0}\beta(\phi_0,\mathscr{A}(\lambda\mathscr{O}_1),\mathscr{A}(\lambda\mathscr{O}_2))$$
$$=\lim_{\lambda\uparrow\infty}\beta(\phi_0,\mathscr{A}(\lambda\mathscr{O}_1),\mathscr{A}(\lambda\mathscr{O}_2)).$$

But if W_1 and W_2 are spacelike separated wedges,

$$\beta(\phi_0,\mathscr{A}(W_1),\mathscr{A}(W_2))=\beta(\phi_0,\mathscr{A}(W_{1,x_i}),\mathscr{A}(W_{2,x_i})),$$

$i=1,2$, where x_i is the translation that puts the apex of W_i at the origin. Thus,

$$\beta(\phi_0,\mathscr{A}(W_1),\mathscr{A}(W_2))=\lim_{\lambda\downarrow 0}\beta(\phi_0,\mathscr{A}(\lambda(W_{1,x_i})),\mathscr{A}(\lambda(W_{2,x_i}))).$$

But $\lim_{\lambda\downarrow 0}\lambda(W_{i,x_i})=W_{i,x_i}$ and $\lim_{\lambda\downarrow 0}\lambda(W_{j,x_i})=W'_{i,x_i}$, $i\neq j$. Therefore,

$$\beta(\phi_0,\mathscr{A}(W_1),\mathscr{A}(W_2))=\beta(\phi_0,\mathscr{A}(W_1),\mathscr{A}(W_1'))$$
$$=\beta(\phi_0,\mathscr{A}(W_2'),\mathscr{A}(W_2)),$$

using the assumption that the wedge algebras are locally generated, which implies that, e.g., the inductive limit[20] $\lim_{\lambda\downarrow 0}\mathscr{A}(\lambda(W_{j,x_i}))=\mathscr{A}(W'_{i,x_i})$. A similar argument proves the final assertion of the theorem after one notes that $\lim_{\lambda\uparrow\infty}\lambda(\mathscr{O}_{1,x_0})=W_{x_0}$ and $\lim_{\lambda\to\infty}\lambda(\mathscr{O}_{2,x_0})=W'_{x_0}$, where x_0 is the translation that takes the point in the intersection of the closures of the tangent double cones $\mathscr{O}_1$ and $\mathscr{O}_2$ to the origin. □

The free, massless, scalar field and the pure electromagnetic field[35,36] are examples of such dilatation-invariant theories. It is known[37] that any dilatation-invariant theory in which there are massless particles must have a trivial S matrix.

Finally, we remark that Theorem 2.1 (2b) entails that if $A_i\in\mathscr{A}$ and $B_i\in\mathscr{B}$ are admissible, $i=1,2$, and satisfy

$$\tfrac{1}{2}\phi_0(A_1(B_1+B_2)+A_2(B_1-B_2))=\sqrt{2},\qquad(4.1)$$

then the A_i (resp. B_i) are in the centralizer (see Theorem 5.3.28 in Ref. 38) of $\mathscr{A}$ (resp. $\mathscr{B}$) in the state ϕ_0. But it is known[39] that the centralizer of any wedge $\mathscr{A}(W)$ in a pure vacuum state ϕ_0 is trivial, i.e., consists only of multiples of the identity. Thus, at least if $\mathscr{A}$ and $\mathscr{B}$ are commuting wedge algebras, it follows that there is no quadruple $\{A_i,B_j\}$ that would satisfy (4.1), even though it is possible, and that will be shown in Paper II, that admissible quadruples can be found so that the left-hand side of (4.1) comes arbitrarily close to $\sqrt{2}$.

Note added in proof: See our paper in Commun. Math. Phys. **110**, 247 (1987).

ACKNOWLEDGMENTS

Most of this work was carried out when one of us (S. J. S.) was either resident in (until March 1985) or visiting the Fachbereich Physik, Universität Osnabrück (Summer 1986).

S. J. S. wishes to thank his friends and colleagues at Osnabrück for their hospitality and acknowledges the financial support of a University of Rochester Faculty Research Grant during the summer, 1986.

[1] J. S. Bell, "On the Einstein-Podolsky-Rosen paradox," Physics (NY) **1**, 195 (1964).
[2] J. S. Bell, "On the problem of hidden variables in quantum mechanics," Rev. Mod. Phys. **38**, 447 (1966).
[3] J. F. Clauser and A. Shimony, "Bell's theorem: Experimental tests and implications," Rep. Prog. Phys. **41**, 1881 (1978).
[4] A. Aspect, "Experimental tests of Bell's inequalities in atomic physics," in *Atomic Physics 8*, edited by I. Lindgren, A. Rosén, and S. Svanberg (Plenum, New York, 1983), p. 103.
[5] A. Aspect, "Experiences basées sur les inégalités de Bell," J. Phys. (Paris) Suppl. **42**, C2 (1981).
[6] R. Haag and D. Kastler, "An algebraic approach to quantum field theory," J. Math. Phys. **5**, 848 (1964).
[7] H. Araki, "Einführung in die axiomatische Quantenfeldtheorie," lectures given at the ETH, Zürich, 1961–62, unpublished.
[8] W. Driessler, S. J. Summers, and E. H. Wichmann, "On the connection between quantum fields and von Neumann algebras of local operators," Commun. Math. Phys. **105**, 49 (1986).
[9] R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin, New York, 1964).
[10] R. Jost, *The General Theory of Quantized Fields* (American Mathematical Society, Providence, RI, 1965).
[11] S. J. Summers and R. Werner, "The vacuum violates Bell's inequalities," Phys. Lett. A **110**, 257 (1985).
[12] S. J. Summers and R. Werner, "Bell's inequalities and quantum field theory, II: Bell's inequalities are maximally violated in the vacuum," J. Math. Phys. **28**, 2448 (1987).
[13] G. Ludwig, *Foundations of Quantum Mechanics, I, II* (Springer, New York, 1983).
[14] G. Ludwig, *An Axiomatic Basis for Quantum Mechanics, I* (Springer, New York, 1985).
[15] J. F. Clauser and M. A. Horne, "Experimental consequences of objective local theories," Phys. Rev. D **10**, 526 (1974).
[16] R. Werner, "Bell's inequalities and the reduction of statistical theories," in *Reduction in Science*, edited by W. Balzer, D. A. Pearce, and H.-J. Schmidt (Reidel, Amsterdam, 1984).
[17] S. J. Summers and R. Werner, "Bell's inequalities and quantum field theory, I. General setting," preprint, 1986.
[18] B. Z. Vulikh, *Introduction to the Theory of Partially Ordered Spaces* (Wolters–Nordhoff, Groningen, 1967).
[19] *Foundations of Quantum Mechanics and Ordered Linear Spaces*, edited by A. Hartkämper and H. Neumann (Springer, New York, 1974).

[20]S. Sakai, *C*-Algebras and W*-Algebras* (Springer, New York, 1971).
[21]J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, "Proposed experiment to test local hidden-variable theories," Phys. Rev. Lett. **23**, 880 (1969).
[22]B. S. Cirel'son, "Quantum generalizations of Bell's inequalities," Lett. Math. Phys. **4**, 93 (1980).
[23]H. Neumann and R. Werner, "Causality between preparation and registration processes in relativistic quantum theory," Int. J. Theor. Phys. **22**, 781 (1983).
[24]H.-J. Borchers, "Über die Mannigfaltigkeit der interpolierenden Felder zu einer kausalen S-matrix," Nuovo Cimento **15**, 784 (1960).
[25]K. Fredenhagen and J. Hertel, "Local algebras of observables and pointlike localized fields," Commun. Math. Phys. **80**, 551 (1981).
[26]J. Hertel, "Lokale Quantentheorie und Felder am Punkt," Ph.D. thesis, Universität Hamburg, 1980.
[27]J. Rehberg and M. Wollenberg, "Quantum fields as pointlike localized objects," Math. Nachr. **125**, 259 (1986).
[28]M. Wollenberg, "Quantum fields as pointlike localized objects, II," preprint, Akademie der Wissenschaften der DDR, 1985.
[29]S. J. Summers, "From algebras of local observables to quantum fields: Generalized H-bounds," preprint, University of Rochester, 1987.
[30]W. Driessler and S. J. Summers, "Central decomposition of Poincaré-invariant nets of local field algebras and absence of spontaneous breaking of the Lorentz group," Ann. Inst. H. Poincaré **43**, 147 (1985).
[31]W. Driessler and S. J. Summers, "On the decomposition of relativistic quantum field theories into pure phases," Helv. Phys. Acta **59**, 331 (1986).
[32]H. Araki, "On the algebra of all local observables," Prog. Theor. Phys. **32**, 844 (1964).
[33]H. Araki, K. Hepp, and D. Ruelle, "Asymptotic behaviour of Wightman functions," Helv. Phys. Acta **35**, 164 (1962).
[34]K. Fredenhagen, "A remark on the cluster theorem," Commun. Math. Phys. **97**, 461 (1985).
[35]P. Leyland, J. E. Roberts, and D. Testard, "Duality for quantum free fields," preprint, CNRS CPT 78/1016, Marseille, 1978.
[36]G. Benfatto and F. Nicolo, "The local von Neumann algebras for the massless scalar free field and the free electromagnetic field," J. Math. Phys. **19**, 653 (1978).
[37]D. Buchholz and K. Fredenhagen, "Dilatations and interaction," J. Math. Phys. **18**, 1107 (1977).
[38]O. Brattelli and D. W. Robinson, *Operator Algebras and Quantum Statistical Mechanics II* (Springer, New York, 1981).
[39]W. Driessler, "Comments on lightlike translations and applications in relativistic quantum field theory," Commun. Math. Phys. **44**, 133 (1975).

Bell's inequalities and quantum field theory. II. Bell's inequalities are maximally violated in the vacuum

Stephen J. Summers
Department of Mathematics, University of Rochester, Rochester, New York 14627
Reinhard Werner
Fachbereich Physik, Universität Osnabrück, D-4500 Osnabrück, Federal Republic of Germany

(Received 22 July 1986; accepted for publication 27 May 1987)

In the context of the study of Bell's inequalities carried out in Paper I [J. Math. Phys. **28**, 2440 (1987)], it is proven that Bell's inequalities are maximally violated in the vacuum state by suitable spacelike separated observables for both Bose and Fermi free quantum field theories.

I. INTRODUCTION

In this paper we continue our study[1] of Bell's inequalities in quantum field theory with the proof that in both Bose and Fermi free quantum field theories these inequalities are maximally violated in the vacuum state by suitable spacelike separated observables. As explained in Paper I, the form of Bell's inequalities with which we work is as follows:

$$\beta(\phi,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))=1\,, \tag{1.1}$$

where

$$\beta(\phi,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2)) \equiv \tfrac{1}{2}\sup \phi(A_1(B_1+B_2)+A_2(B_1-B_2))\,,$$

and the supremum is taken over all $A_i\in\mathscr{A}(\mathscr{O}_1)$, $B_i\in\mathscr{A}(\mathscr{O}_2)$, with $A_i=A_i^*$, $B_i=B_i^*$, $-1\leqslant A_i\leqslant 1$ and $-1\leqslant B_i\leqslant 1$, $i=1,2$. Here $\mathscr{O}_1$ and $\mathscr{O}_2$ are spacelike separated regions of Minkowski space and ϕ is a state on $\mathscr{A}$, the C^*-algebra of quasilocal observables generated by a given net of local observable algebras $\{\mathscr{A}(\mathscr{O})\}$ (see Sec. III of Paper I for notation and background). For each quantum field model to be considered here we shall explicitly define the algebras $\{\mathscr{A}(\mathscr{O})\}$ at the appropriate place.

From Theorem 2.4 in Paper I it follows that if $\mathscr{O}_1\subseteq\mathscr{O}_2'$, then

$$\beta(\phi,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))\leqslant\sqrt{2}\,, \tag{1.2}$$

for any state $\phi\in\mathscr{A}_1^{*(+)}$. If the equality holds in (1.2), we shall say, for obvious reasons, that Bell's inequalities (1.1) have been maximally violated. It is precisely this equality when ϕ is the vacuum state of a free quantum field theory and $\mathscr{O}_1$ and $\mathscr{O}_2$ are certain space-time regions (e.g., $\mathscr{O}_1$ and $\mathscr{O}_2$ are complementary wedge regions) that will be proven. In work in progress we intend to show that the equality in (1.2) holds for more general classes of quantum field models and states (and regions $\mathscr{O}_1,\mathscr{O}_2$). But in a sense, it is maximal violation in the vacuum for free fields that could be regarded as the least expected of such results, since the very strong correlations between certain spacelike separated observables that are implied by the maximal violation of Bell's inequalities can then neither be attributed to a special preparation of the system nor to some nontrivial interaction of the fields under consideration. The point to be made is that already vacuum fluctuations manifest correlations that are too large to be modeled by "local hidden-variable theories" (see Sec. II of Paper I).

In Sec. II we present the main results of this paper, which are then proven in the subsequent sections. Some of these results were previously announced in Ref. 2.

II. MAIN RESULTS

To facilitate an overview of the theorems proven in this paper, we start by collecting here the main results. We follow the notation established in Paper I.

If $\varphi(\cdot)$ denotes the free Bose quantum field of mass $m\geqslant 0$ and D is its standard domain in the Bose Fock space $\mathscr{H}$ (see Ref. 3), then it is well known that for every real-valued tempered test function $f\in\mathscr{S}(\mathbb{R}^4)$ the operator $\varphi(f)$ is essentially self-adjoint on D. Moreover, if for each open $\mathscr{O}\subset\mathbb{R}^4$ $\mathscr{A}(\mathscr{O})$ is defined to be the von Neumann algebra

$$\mathscr{A}(\mathscr{O})\equiv\{e^{i\overline{\varphi(f)}}\,|\,f\in\mathscr{S}_{\mathbf{R}}(\mathbb{R}^4),\ \mathrm{supp}(f)\subset\mathscr{O}\}'' \tag{2.1}$$

generated by the self-adjoint closure of $\varphi(f)\restriction D$ for all real-valued $f\in\mathscr{S}(\mathbb{R}^4)$ with support in $\mathscr{O}$, then $\{\mathscr{A}(\mathscr{O})\}$ is a net of local algebras satisfying the axioms (1)–(4) in Sec. III of I (see, e.g., Ref. 4) transforming covariantly under the action of the representation $U(\mathscr{P}_+^\uparrow)$ of the (covering group of the) Poincaré group $\mathscr{P}_+^\uparrow$ associated with the field $\varphi(\cdot)$ (Ref. 3). If Ω is the vacuum vector in $\mathscr{H}$, then

$$\phi_0(A)\equiv\langle\Omega,A\Omega\rangle,\quad A\in\mathscr{A}\,, \tag{2.2}$$

defines a state on the algebra of quasilocal observables generated by the net $\{\mathscr{A}(\mathscr{O})\}$ just defined. We remark that in order to keep notation within bounds and the reader's attention on the essential points, we have tacitly assumed that the field $\varphi(\cdot)$ is neutral and has spin 0. But by making use of Refs. 5 and 6, for example, the methods of this paper can be easily extended to fields of any spin, but with finitely many components.

Theorem 2.1: With the above definitions,

$$\beta(\phi_0,\mathscr{A}(W),\mathscr{A}(W'))=\sqrt{2}\,,$$

for any wedge region $W\in\mathscr{W}$.

Here, W and W' are each other's causal complement and are called complementary wedges. See Theorem 2.4 for similar results for regions other than complementary wedges in the case $m=0$.

 0022-2488/87/102448-09$02.50 © 1987 American Institute of Physics

If $\psi(\cdot)$ represents the free Fermi quantum field[3] of mass $m \geqslant 0$ and spin $s = n/2$, $n \in \mathbb{N}$ (but finitely many components), then the anticommutation relations entail that $\psi(f)$ is a bounded operator for every test function f. For that reason one can directly define the free Fermi field algebras as

$$\mathcal{F}(\mathcal{O}) \equiv \{\psi(f) | \operatorname{supp}(f) \subset \mathcal{O}\}'' . \tag{2.3}$$

Here, $\{\mathcal{F}(\mathcal{O})\}$ is a net of local algebras satisfying axioms (1), (2), and (4) in Sec. III of Sec. I. Because field operators $\psi(f)$ and $\psi(g)$ with $\operatorname{supp}(f)$ spacelike separated from $\operatorname{supp}(g)$ anticommute, it is necessary to modify axiom (3) for the field algebras. This is done by introducing a unitary Klein transformation $\mathcal{F}(\mathcal{O}) \to \mathcal{F}(\mathcal{O})'$ that yields $\mathcal{F}(\mathcal{O}) \subseteq \mathcal{F}(\mathcal{O}')''$ for every region $\mathcal{O} \subset \mathbb{R}^4$ as the appropriate expression for the locality of the field operators (see Sec. IV for more details). If, once again, Ω is the vacuum vector in the Fermi Fock space, then

$$\phi_0(F) \equiv \langle \Omega, F\Omega \rangle, \quad F \in \mathcal{F} , \tag{2.4}$$

defines a state on the algebra $\mathcal{F}$ of quasilocal field operators generated by the net $\{\mathcal{F}(\mathcal{O})\}$.

Theorem 2.2: With the above definitions,

$$\beta(\phi_0, \mathcal{F}(W), \mathcal{F}(W')') = \sqrt{2} ,$$

for any wedge region $W \in \mathcal{W}$.

Once again, Theorem 2.4 has similar results for different regions when the field has zero mass.

The fact that there are anticommuting elements of algebras associated with spacelike separated regions is a consequence of the fact that there are nonobservable operations in $\mathcal{F}(\mathcal{O})$. In particular, there are local operators carrying charge. With the reservations made in Sec. III of Paper I in mind, the standard way[7] to choose the "observable algebras" for the free Fermi fields is to take the fixed point subalgebras under the gauge group induced by the free charge operator Q;

$$\alpha_t(F) \equiv e^{itQ} F e^{-itQ} , \tag{2.5}$$

which defines a unitary group of automorphisms on $\mathcal{F}$. Then the observable algebras are given by

$$\mathcal{A}(\mathcal{O}) \equiv \{F \in \mathcal{A}(\mathcal{O}) | \alpha_t(F) = F \text{ for all } t \in \mathbb{R}\} . \tag{2.6}$$

Elements in $\mathcal{A}(\mathcal{O})$ clearly do not carry charge. Also for these Fermi observable algebras we prove maximal violation of Bell's inequalities.

Theorem 2.3: With the above definitions,

$$\beta(\phi_0, \mathcal{A}(W), \mathcal{A}(W')) = \sqrt{2} ,$$

for any wedge region $W \in \mathcal{W}$.

These results, together with Theorem 4.4. in Paper I, entail the following additional maximal violations. Note we are considering only free fields and not generalized free fields, so there is only one mass in each theory. If the free field theory is massless, it is dilatation invariant.

Theorem 2.4: If the mass of the free quantum field theory is zero, then

$$\beta(\phi_0, \mathcal{A}(W_1), \mathcal{A}(W_2)) = \sqrt{2} \quad [= \beta(\phi_0, \mathcal{F}(W_1), \mathcal{F}(W_2)')]$$

and

$$\beta(\phi_0, \mathcal{A}(\mathcal{O}_1), \mathcal{A}(\mathcal{O}_2)) = \sqrt{2} \quad [= \beta(\phi_0, \mathcal{F}(\mathcal{O}_1), \mathcal{F}(\mathcal{O}_2)')]$$

for any two spacelike separated wedges $W_1, W_2 \in \mathcal{W}$ and any two tangent double cones $\mathcal{O}_1, \mathcal{O}_2 \in \mathcal{K}$ (tangent double cones are spacelike separated double cones whose closures intersect at one point).

We shall say a few words about the strategy employed in the proof of these results. In Sec. III (resp. Secs. IV and V) we prove Theorem 2.1 (resp. 2.2 and 2.3). But all three have much in common. First, the local algebras and the vacuum state can be constructed explicitly in each case in terms of suitable test function spaces. We look in all three cases for large Bell correlations in subalgebras generated by finitely many field operators. Thus, we study first an analogous problem for finite-dimensional test function spaces. Under suitable conditions on the test functions we find almost maximal correlations for these finite-dimensional test function spaces. We then obtain the maximal violations by taking certain limits that, however, remain in the original algebras. The fact, in all three cases, that this limit can indeed be taken in the manner we require rests upon the result of Bisognano and Wichmann[5,8] that the modular automorphism group $\{\Delta^{it}\}_{t \in \mathbb{R}}$ of a wedge algebra generated by a standard quantum field in the manner we have indicated above coincides with the subgroup of Lorentz velocity transformations leaving the corresponding wedge invariant, so that the modular operator δ in the test function space has absolutely continuous spectrum containing the point 1.

We advise the reader that any undefined notation in this paper has already been established in Paper I.

III. MAXIMAL VIOLATION FOR FREE BOSE QUANTUM FIELDS

We begin by defining the local algebras of a free Bose field theory in terms of spaces of test functions. Formally the field is a linear assignment of a symmetric operator $\varphi(f)$ on a Hilbert space $\mathcal{H}$ to each element f of a test function space $\mathcal{T}$. Here $\mathcal{T}$ is a real vector space (possibly the real part of a space of $\mathbb{C}^N$-valued functions, where N is the number of components of the field). Two real bilinear forms on $\mathcal{T}$ determine the structure of the theory. One is the antisymmetric (or "symplectic") form σ that determines the canonical commutation relations (CCR), which we write in terms of the unitary Weyl operators $W(f) = \exp(i\varphi(f))$ as

$$W(f)W(g) = \exp((i/2)\sigma(f,g))W(f+g) . \tag{3.1}$$

The C^*-algebra generated by the Weyl operators is called the CCR algebra over $(\mathcal{T}, \sigma)$. The second bilinear form q is symmetric and determines the vacuum state ϕ_0 on the CCR algebra through the relation

$$\phi_0(W(f)) = \exp(-\tfrac{1}{4} q(f,f)), \quad f \in \mathcal{T} . \tag{3.2}$$

We assume that there is a vector $\Omega \in \mathcal{H}$ that is cyclic for $\{W(f) | f \in \mathcal{T}\}$ and satisfies $\langle \Omega, W(f)\Omega \rangle = \phi_0(W(f))$. The positivity of the state ϕ_0 is equivalent to the inequalities

$$(\sigma(f,g))^2 \leqslant q(f,f)q(g,g) \tag{3.3}$$

(see, e.g., Ref. 9). In particular, σ is continuous with respect to the norm on $\mathcal{T}$ given by q. Consequently, the form σ and the commutation relations (3.1) can be extended by continuity to all f,g in the completion of $\mathcal{T}$ with respect to that

norm. Henceforth, we shall assume, without loss of generality, that $\mathscr{T}$ is complete with respect to q. Inequality (3.3) also implies the existence of a bounded operator A on $\mathscr{T}$ with $\sigma(f,g) = q(f,Ag)$ and $0 \leqslant -A^2 \leqslant 1$. A state ϕ_0 determined by (3.2) on the CCR algebra is pure if and only if $A^2 = -1$. The vacuum state of a free Bose field is indeed pure, and we shall denote in that case the operator A relating q and σ by i, thus making $\mathscr{T}$ a complex Hilbert space with inner product $\langle f,g\rangle = q(f,g) - i\sigma(f,g)$. For a free scalar Bose field of mass $m \geqslant 0$ this inner product is given directly by

$$\langle f,g\rangle = \tfrac{1}{2}\langle \Omega,\varphi(f)\varphi(g)\Omega\rangle$$

$$= \int d^4p\, \delta(p^2 - m^2)\, \overline{\tilde{f}(p)}\tilde{g}(p)\,, \qquad (3.4)$$

for $f,g \in S_{\mathbf{R}}(\mathbf{R}^4)$ and $\tilde{f}$ the Fourier transform of $f(x)$.

We shall denote by $M(\mathscr{O})$ the closure in $\mathscr{T}$ of the real linear space of test functions with support contained in the open space-time region $\mathscr{O}$. Then the usual locality condition for Bose fields requires that for $f_k \in M(\mathscr{O}_k)$ and $\mathscr{O}_1$ and $\mathscr{O}_2$ spacelike separated, the field operators $\varphi(f_1)$ and $\varphi(f_2)$ commute, i.e., $\sigma(f_1,f_2) = 0$. Thus, if

$$M' \equiv \{f \in \mathscr{T} \mid \sigma(f,g) = 0 \text{ for all } g \in M\} \qquad (3.5)$$

denotes the symplectic complement of a subset $M \subset \mathscr{T}$, then locality may be stated in terms of the test function spaces as $M(\mathscr{O}') \subset M(\mathscr{O})'$. (See Ref. 10; also Refs. 4 and 11 for the "equal time formulation.") We then define the observable algebra $\mathscr{A}(\mathscr{O})$ associated to a region $\mathscr{O}$ as $\mathscr{A}(M(\mathscr{O}))$, where $\mathscr{A}(M)$ is defined for each subspace $M \subset \mathscr{T}$ as the von Neumann algebra generated by $\{W(f) \mid f \in M\}$. In terms of the symplectic complement (3.5) of a closed real subspace $M \subset \mathscr{T}$, (3.1) implies that $\mathscr{A}(M') \subseteq \mathscr{A}(M)'$, the commutant of $\mathscr{A}(M)$. In Ref. 11 (see also Refs. 10 and 12) it was shown that, in fact, $\mathscr{A}(M)' = \mathscr{A}(M')$, which is called "abstract duality."

By the Reeh–Schlieder theorem[13] the vacuum vector Ω is cyclic and separating for each local algebra $\mathscr{A}(\mathscr{O})$ for which $\mathscr{O}$ and $\mathscr{O}'$ are nonempty. An equivalent condition is that the space $M(\mathscr{O})$ is standard[14] in the sense that $M \cap iM = \{0\}$ and $M + iM$ is dense in $\mathscr{T}$. For standard subspaces Rieffel and van Daele[14] set up a modular theory closely analogous to that of Tomita–Takesaki.[15] Specifically, one defines a closed antilinear involution s on $\mathscr{T}$ by $s(f+ig) = f - ig$ for $f,g \in M$. Then $f \in M$ if and only if $sf = f$. The canonical involution s has a polar decomposition $s = j\delta^{1/2}$ such that the unitary group $t \to \delta^{it}$ leaves M invariant. The canonical involution of the complementary subspace M' is equal to $s^* = \delta^{1/2}j = j\delta^{-1/2}$. The operator δ is related to the modular operator Δ of the von Neumann algebra $\mathscr{A}(M)$ with respect to the vector Ω by the equation $\Delta^{it}W(f)\Delta^{-it} = W(\delta^{it}f)$ for $f \in M$. Similarly, the operator j is related to the modular involution J of $\{\mathscr{A}(M),\Omega\}$ by $JW(f)J = W(jf)$ (see Ref. 12).

We shall utilize this connection to characterize the group $\{\delta^{it}\}_{t\in\mathbf{R}}$ in an important special case, namely when $\mathscr{O} = W \in \mathscr{W}$ is a wedge region in space-time (Paper I). In this case a general result of Bisognano and Wichmann[5,8] states that $\Delta^{it} = V(\pi t)$, where $\{V(t)\}_{t\in\mathbf{R}}$ is the subgroup representing the Lorentz velocity transformations leaving the wedge W (and W') invariant in the representation $U(\mathscr{P}^{\uparrow}_{+})$ of the (covering group of the) Poincaré group $\mathscr{P}^{\uparrow}_{+}$ under which the free field transforms covariantly.[3,8] Since the scalar product in (3.4) is Poincaré covariant, the boosts $\{V(t)\}_{t\in\mathbf{R}}$ are represented by a unitary group $\{v_t\}_{t\in\mathbf{R}}$ on $\mathscr{T}$ [and thus on $M(W)$]. Then we have

$$W(v(\pi t)f) = V(\pi t)W(f)V(\pi t)^*$$

$$= \Delta^{it}W(f)\Delta^{-it} = W(\delta^{it}f)\,,$$

and hence $v(\pi t)f = \delta^{it}f$ for all $f \in M(W)$. Since $M(W) + iM(W)$ is dense, the unitary operators δ^{it} and $v(\pi t)$ must coincide. This may be derived by more direct means, but the above argument has the advantage of being immediately applicable to free fields of any spin.

It should now be clear that the maximal Bell correlation $\beta(\phi_0,\mathscr{A}(\mathscr{O}_1),\mathscr{A}(\mathscr{O}_2))$ can be defined completely in terms of the real linear subspaces $M(\mathscr{O}_1),M(\mathscr{O}_2) \subset \mathscr{T}$. Since the restriction of ϕ_0 to an algebra $\mathscr{A}(M)$ is determined by q restricted to M, it even suffices to know the forms q and σ restricted to $M(\mathscr{O}_1) + M(\mathscr{O}_2)$. Thus we divide the proof of Theorem 2.1 into the two steps of establishing a property of forms q and σ on $M_1 + M_2 \subset \mathscr{T}$ implying $\beta(\phi_0,\mathscr{A}(M_1),\mathscr{A}(M_2)) = \sqrt{2}$ and of demonstrating this property for the concrete spaces $M_k = M(\mathscr{O}_k)$, $k = 1,2$.

Note that in order to define $\beta(\phi_0,\mathscr{A}(M_1),\mathscr{A}(M_2))$ we have to assume that $\mathscr{A}(M_1)$, $\mathscr{A}(M_2)$ commute elementwise, i.e., $\sigma(f_1,f_2) = 0$ for $f_k \in M_k$. If for such pairs we also had $q(f_1,f_2) = 0$, this would imply

$$\phi_0(W(f_1)W(f_2)) = \phi_0(W(f_1)) \cdot \phi_0(W(f_2))$$

and hence $\beta(\phi_0,\mathscr{A}(M_1),\mathscr{A}(M_2)) = 1$ by Theorem 2.1 (3c) of Paper I. In fact, Theorem 2.1 (3c) implies a stronger result: if $\tilde{q}$ is another bilinear form on $M_1 + M_2$ satisfying (3.3) for the same σ, such that $\tilde{q}(f_1,f_2) = 0$ for $f_k \in M_k$ and $\tilde{q} \leqslant q$, then ϕ_0 is a Gaussian average over translates of another state $\tilde{\phi}$, each of which is a product state,[16] hence $\beta(\phi_0,\mathscr{A}(M_1),\mathscr{A}(M_2)) = 1$.

These remarks indicate in which situations one might look for large violations of Bell's inequalities. Here, $\beta(\phi_0,\mathscr{A}(M_1),\mathscr{A}(M_2))$ is obviously convex in ϕ_0, so that large values will be attained if ϕ_0 is a pure state on the algebra $\mathscr{A}(M_1) \vee \mathscr{A}(M_2)$ generated by $\mathscr{A}(M_1)$ and $\mathscr{A}(M_2)$. On the other hand, the elements in $\mathscr{A}(M_2)$ may be viewed as inducing particular decompositions of the state ϕ_0 restricted to $\mathscr{A}(M_1)$. These decompositions are trivial if ϕ_0 restricted to $\mathscr{A}(M_1)$ is pure, implying $\beta(\phi_0,\mathscr{A}(M_1),\mathscr{A}(M_2)) = 1$ [Theorem 2.1 (3b) in Paper I]. Thus β may be expected to be larger if ϕ_0 is almost pure on $\mathscr{A}(M_1) \vee \mathscr{A}(M_2)$ but very impure (or even a tracial state) restricted to each $\mathscr{A}(M_k)$. This intuition is essential for the proofs to follow, because we choose the test functions generating subalgebras of the original algebras in such a way that ϕ_0 does manifest said behavior on these subalgebras.

This now said, we proceed to the proofs.

Proposition 3.1: Let $\mathscr{T}$ be a complex Hilbert space with real subspaces M, N such that $M \subseteq N'$. Let $0 < \lambda < 1$ and suppose that for each $\epsilon > 0$ there are test functions $f_1, f_2 \in M$ and $g_1,g_2 \in N$ (the dependence of these test functions on ϵ will be suppressed in the notation) such that

(i) $\|f_k\|^2 \approx (1+\lambda^2)/(1-\lambda^2) \approx \|g_k\|^2, \quad k=1,2,$

(ii) $\langle f_1, f_2\rangle \approx i \approx \langle g_1, g_2\rangle$,

(iii) $\langle f_1, g_1\rangle \approx 2\lambda/(1-\lambda^2) \approx -\langle f_2, g_2\rangle$,

(iv) $\langle f_1, g_2\rangle \approx 0 \approx \langle f_2, g_1\rangle$,

where $x \approx y$ means $|x-y| \leqslant \epsilon$. Then

$$\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(N)) \geqslant \sqrt{2}[2\lambda/(1+\lambda^2)].$$

Proof: (1) With $f_1, f_2 \in M$ define $\tilde{M}(f_1, f_2)$ as the real linear span of f_1 and f_2. Then $\mathscr{A}(\tilde{M}(f_1,f_2)) \subset \mathscr{A}(M)$ and using similar notation for N, one has $\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(N)) \geqslant \beta(\phi_0, \mathscr{A}(\tilde{M}(f_1,f_2)), \mathscr{A}(\tilde{N}(g_1,g_2)))$. The latter quantity involves only CCR algebras over finite-dimensional test function spaces. In fact, by a trivial rescaling, one may assume that $\mathrm{Im}\langle f_1, f_2\rangle = \mathrm{Im}\langle g_1, g_2\rangle = i$, which fixes the commutation relations for all operators $W(h)$, with $h = \Sigma_{j=1,2}(\alpha_j f_j + \beta_j g_j)$. Thus, by von Neumann's uniqueness theorem,[17] $\mathscr{A}(\tilde{M}(f_1,f_2)) \cong \mathscr{B}(\mathscr{H})$, $\mathscr{A}(\tilde{N}(g_1,g_2)) \cong \mathscr{B}(\mathscr{K})$, and $\mathscr{A}(\tilde{M}(f_1,f_2)) \cdot \mathscr{A}(\tilde{N}(g_1,g_2)) \simeq \mathscr{B}(\mathscr{H}) \otimes \mathscr{B}(\mathscr{K})$, where $\mathscr{H}$ and $\mathscr{K}$ are Hilbert spaces carrying a fixed representation of the CCR for one degree of freedom. The real parts of the inner products in (i)–(iv) determine the restriction of the quadratic form $\|\cdot\|^2$ to $\tilde{M}+\tilde{N}$ and hence a density matrix $p \in \mathscr{T}_1(\mathscr{H}\otimes\mathscr{K})$ with $\mathrm{tr}(pW(h)) = \exp(-\frac{1}{4}\|h\|^2)$.

Let $f_1, \ldots, g_2$ be functions for which (i)–(iv) are satisfied with equality, and let $\bar{p}$ denote the density matrix obtained from $f_1, \ldots, g_2$ in the manner described above. Moreover, let $\{p^{(\nu)}\}_{\nu\in\mathbb{N}}$ be a sequence of density matrices obtained in this way from a sequence $\{f_j^{(\nu)}, g_j^{(\nu)}\}_{\nu\in\mathbb{N}}^{j=1,2}$ satisfying (i)–(iv) elementwise for $\{\epsilon^{(\nu)}\}_{\nu\in\mathbb{N}}$ a sequence of positive numbers converging to 0. Then it is asserted that

$$\overline{\lim_{\nu\to\infty}}\, \beta(p^{(\nu)}, \mathscr{B}(\mathscr{H}), \mathscr{B}(\mathscr{K})) \geqslant \beta(\bar{p}, \mathscr{B}(\mathscr{H}), \mathscr{B}(\mathscr{K})) \equiv \bar{\beta},$$

so that $\bar{\beta}$ is a lower bound for $\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(N))$. To see this, suppose that $A_1, A_2 \in \mathscr{B}(\mathscr{H})$ and $B_1, B_2 \in \mathscr{B}(\mathscr{K})$ are self-adjoint contractions such that $\frac{1}{2}\mathrm{tr}(\bar{p}[A_1(B_1+B_2) + A_2(B_1-B_2))) \geqslant \bar{\beta} - \epsilon_1$, $\epsilon_1 > 0$ given but arbitrary. By Kaplansky's density theorem[18] A_j, B_j are (strong limits of) linear combinations of Weyl operators over $\tilde{M}(f_1,f_2)$ and $\tilde{N}(g_1,g_2)$, respectively. Thus, $T \equiv \frac{1}{2}(A_1(B_1+B_2) + A_2(B_1-B_2))$ is a (strong limit of) linear combination of Weyl operators over $\tilde{M}+\tilde{N}$, so that $\mathrm{tr}(p^{(\nu)}T)$ depends continuously on the inner products (i)–(iv). Hence, for any given $\epsilon_2 > 0$ and all sufficiently large ν,

$$\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(N)) \geqslant \mathrm{tr}(p^{(\nu)}T) \geqslant \mathrm{tr}(pT) - \epsilon_2 \geqslant \bar{\beta} - \epsilon_1 - \epsilon_2.$$

Since ϵ_1 and ϵ_2 are arbitrary, the assertion is proven.

(2) By step 1, any number $\frac{1}{2}\mathrm{tr}(\bar{p}[A_1(B_1+B_2) + A_2(B_1-B_2)))$ with self-adjoint contractions $A_j \in \mathscr{B}(\mathscr{H})$ and $B_j \in \mathscr{B}(\mathscr{K})$ is a lower bound for $\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(N))$. In order to construct such operators A_j, B_j to satisfy the lower bound of the proposition, it is necessary to have an explicit representation of the density matrix $\bar{p}$. Thus, write

$$W(\alpha_1 f_1 + \alpha_2 f_2 + \beta_1 g_1 + \beta_2 g_2) = \exp\{i(\alpha_1 Q_1 + \alpha_2 P_1 + \beta_1 Q_2 + \beta_2 P_2)\}$$

with canonical operators satisfying $i[P_j, Q_k] = I\delta_{jk}$, $j,k = 1,2$. Let $\{|n\rangle_j\}_{n=0}^{\infty}$ denote the set of eigenstates of $\frac{1}{2}(P_j^2 + Q_j^2)$ and let $|nm\rangle \equiv |n\rangle_1 \otimes |m\rangle_2 \in \mathscr{H}\otimes\mathscr{K}$. Then

$$\mathrm{tr}(\bar{p}W(h)) = \langle \Omega, W(h)\Omega\rangle,$$

with

$$\Omega = (1-\lambda^2)^{1/2} \sum_{n=0}^{\infty} \lambda^n |nn\rangle.$$

This is verified most conveniently by noting that Ω is the ground state of

$$H = [(1+\lambda^2)/(1-\lambda^2)]\,\tfrac{1}{2}(P_1^2 + Q_1^2 + P_2^2 + Q_2^2) + [2\lambda/(1-\lambda^2)](P_1P_2 - Q_1Q_2).$$

Thus, in particular, $\bar{p}$ determines a pure state on $\mathscr{B}(\mathscr{H}) \otimes \mathscr{B}(\mathscr{K})$ whose restriction to each factor is given by a density matrix with eigenvalues $(1-\lambda^2)\lambda^{2n}$, which is therefore very impure when λ is close to 1.

Define

$$A_j|2n\rangle_1 \equiv e^{i\alpha_j}|2n+1\rangle_1, \quad B_j|2n\rangle_2 \equiv e^{i\beta_j}|2n+1\rangle_2,$$

$$A_j|2n+1\rangle_1 \equiv e^{-i\alpha_j}|2n\rangle_1, \quad B_j|2n+1\rangle_2 \equiv e^{-i\beta_j}|2n\rangle_2,$$

with $\alpha_j, \beta_j \in \mathbb{R}$. Then

$$\langle \Omega, A_j \otimes B_k \Omega\rangle = 2(1+\lambda^2)^{-1}\,\mathrm{Re}(\lambda e^{i(\alpha_j+\beta_k)}).$$

With the particular choices $\alpha_1 = 0$, $\alpha_2 = \pi/2$, $\beta_1 = -\pi/4$, $\beta_2 = \pi/4$, one has

$$\tfrac{1}{2}\langle \Omega, [A_1 \otimes (B_1+B_2) + A_2 \otimes (B_1-B_2))]\Omega\rangle = [4\lambda/(1+\lambda^2)]\cos(\pi/4).$$

The proposition is thus proven. □

We now utilize this result to take the next step towards proving Theorem 2.1.

Corollary 3.2: Let M be a standard real subspace of a complex Hilbert space with canonical involution $s = j\delta^{1/2}$. Suppose that $0 < \lambda^2 < 1$ is in the spectrum of δ. Then

$$\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(M)') \geqslant \sqrt{2}[2\lambda/(1+\lambda^2)].$$

In particular, if 1 is not an isolated eigenvalue of δ, then

$$\beta(\phi_0, \mathscr{A}(M), \mathscr{A}(M)') = \sqrt{2}.$$

Proof: Pick $\epsilon > 0$ so that $0 < \lambda^2 - \epsilon < \lambda^2 + \epsilon < 1$ and let ϕ be a unit vector in the spectral subspace of δ for $[\lambda^2 - \epsilon, \lambda^2 + \epsilon]$. ϕ must therefore be in the domain of definition of δ, thereby also in the domain of definition of $s = j\delta^{1/2}$ and $s^* = j\delta^{-1/2}$, since j is bounded. Because $j\delta j = \delta^{-1}$, j must exchange the eigenspaces of δ above, resp. below, 1; hence $\langle \phi, s\phi\rangle = \langle \phi, s^*\phi\rangle = 0$. Let $\mu_+ \equiv \langle \phi, \delta\phi\rangle$ and $\mu_- \equiv \langle \phi, \delta^{-1}\phi\rangle$, and define

$$f_1 \equiv (1+s)(1-\mu_+)^{-1/2}\phi,$$

$$f_2 \equiv (1+s)(1-\mu_+)^{-1/2} i\phi,$$

$$g_1 \equiv (1+s^*)(\mu_- - 1)^{-1/2}\phi,$$

$$g_2 \equiv -(1+s^*)(\mu_- - 1)^{-1/2} i\phi.$$

Then since $s^2 = 1$, one has $sf_k = f_k$ for $k = 1,2$, and $s^*g_k = g_k$ for $k = 1,2$. Hence $f_k \in M$ and $g_k \in M'$, since s^* is the canonical involution for the subspace M'. Moreover,

$$\langle f_1, f_2\rangle = (1-\mu_+)^{-1}(i\|\phi\|^2 + \langle j\delta^{1/2}\phi, j\delta^{1/2} i\phi\rangle) = i(1-\mu_+)^{-1}(1-\mu_+) = i,$$

and, by similar computations,

$$\|f_k\|^2 = (1+\mu_+)/(1-\mu_+)\,,$$
$$\|g_k\|^2 = (1+\mu_-)/(\mu_- - 1), \quad k=1,2\,,$$
$$\langle g_1,g_2\rangle = i, \quad \langle f_1,g_2\rangle = \langle f_2,g_1\rangle = 0\,,$$

and

$$\langle f_1,g_1\rangle = -\langle f_2,g_2\rangle = 2(1-\mu_+)^{-1/2}(\mu_- - 1)^{-1/2}\,.$$

As ϵ tends to zero, $\mu_+ \to \lambda^2$, $\mu_- \to \lambda^{-2}$ so that conditions (i)–(iv) of Proposition 3.1 are satisfied. Thus the assertions of the corollary follow directly from Proposition 3.1. □

We have now collected all that is necessary for a proof of Theorem 2.1.

Proof of Theorem 2.1: As already discussed, for a wedge region W, $t \to \delta^{it}$ is a one-parameter boost subgroup in a nontrivial representation of the Poincaré group. Hence the spectrum of its generator $\ln \delta$ is equal to $\mathbf{R}$. In particular, 1 is a nontrivial accumulation point of the spectrum of δ. Thus, Corollary 3.2 implies $\beta(\phi_0, \mathscr{A}(W), \mathscr{A}(W)') = \sqrt{2}$. But by the duality of the wedge algebras,[4,5,8] $\mathscr{A}(W)' = \mathscr{A}(W')$, completing the proof. □

IV. MAXIMAL VIOLATION FOR FREE FERMI QUANTUM FIELDS: FIELD ALGEBRA CASE

The definition of the local field algebras of a free Fermi field theory in terms of spaces of test functions is in many points quite similar to that presented in Sec. III for bosons. Once again let the test function space $\mathscr{T}$ be a real Hilbert space with q a positive symmetric bilinear form on $\mathscr{T}$, and let $f \to \psi(f)$ be a real linear mapping satisfying

$$[\psi(f),\psi(g)]_+ = q(f,g), \quad f,g \in \mathscr{T}\,. \tag{4.1}$$

They generate an abstract C^*-algebra $\mathscr{A}(\mathscr{T},q)$, called the canonical anticommutation relation (CAR) algebra over $(\mathscr{T},q)$. Quasifree states ϕ on $\mathscr{A}(\mathscr{T},q)$ are determined by bounded operators A on $\mathscr{T}$ by

$$\phi_A(1) = 1, \quad \phi_A(\psi(f)\psi(g)) = \tfrac{1}{2}(q(f,g) + iq(Af,g))\,, \tag{4.2}$$

where A satisfies $q(Af,g) = -q(f,Ag)$ and $\|A\| \leqslant 1$ (see Ref. 19). If A satisfies, in addition, $A^2 = -1$, then it induces a complex structure on $\mathscr{T}$ by

$$(\lambda_1 + i\lambda_2)f \equiv \lambda_1 f + \lambda_2 Af, \quad \lambda_1,\lambda_2 \in \mathbf{R}, \quad f \in \mathscr{T}\,, \tag{4.3}$$

with a complex inner product

$$\langle f,g\rangle \equiv q(f,g) + iq(Af,g)\,. \tag{4.4}$$

In terms of this inner product the CAR (4.1) becomes

$$[\psi(f),\psi(g)]_+ = \mathrm{Re}\langle f,g\rangle\,. \tag{4.5}$$

It is known[19] that ϕ_A is pure on $\mathscr{A}(\mathscr{T},q)$ precisely when $A^2 = -1$. Under such circumstances one can define creation and annihilation operators and ϕ_A is a Fock state.[19] One knows that $A = 0$ determines the unique tracial state ϕ_A on $\mathscr{A}(\mathscr{T},q)$.

Let $\mathscr{K}_0$ be such a complexification of $(\mathscr{T},q)$ and $M \subset \mathscr{K}_0$. Then the symplectic complement M' of M is defined to be

$$M' = \{f \in \mathscr{K}_0 \mid \mathrm{Im}\langle f,g\rangle = 0, \text{ all } g \in M\}\,. \tag{4.6}$$

Let $\Gamma(\mathscr{K}_0)$ be the Fock space associated to the complexification $\mathscr{K}_0$. If $M \subset \mathscr{K}_0$ is a real, closed subspace of $\mathscr{K}_0$ and $\psi(h)$ satisfies (4.5) and

$$\langle \Omega,\psi(f)\psi(g)\Omega\rangle = \tfrac{1}{2}\langle f,g\rangle\,, \tag{4.7}$$

where Ω is the Fock vacuum vector corresponding to the given Fock state ϕ_A, then define

$$\mathscr{F}(M) = \{\psi(f) \mid f \in M\}''\,.$$

If $N \subseteq iM'$, then

$$[\psi(f),\psi(g)]_+ = 0, \quad \text{all } f \in M, \quad g \in N\,, \tag{4.8}$$

by (4.5). We now define a Klein transformation in order to express this anticommutation as commutation.

The map $\psi(f) \to -\psi(f)$ determines a unique automorphism γ on $\mathscr{F}(M)$ that leaves the Fock state invariant. Thus, there exists a unitary involution U implementing γ,

$$U\psi(f)U^* = -\psi(f)\,. \tag{4.9}$$

Let $V \equiv (1/\sqrt{2})(I - iU)$ and $\mathscr{F}(M)' \equiv V\mathscr{F}(M)V^*$. Then it is easy to verify that

$$\mathscr{F}(M) \subseteq \mathscr{F}(iM')''\,. \tag{4.10}$$

In fact, it has been proven in Refs. 6 and 20 that equality holds in (4.10) (abstract twisted duality). The definition of standard subspaces, canonical involutions s, and the polar decomposition $s = j\delta^{1/2}$, along with the identities

$$\Delta^{it}\psi(f)\Delta^{-it} = \psi(\delta^{it}f), \quad t \in \mathbf{R}, \quad f \in M\,, \tag{4.11}$$

and

$$J\psi(f)J = V^*\psi(ijf)V \tag{4.12}$$

are given (resp. proven) in Refs. 14 and 20.

This, then, is the abstract setting for free Fermi fields. It is now necessary to specifiy the concrete test function spaces such that the following desiderata are obtained: specify $\mathscr{T}$, q, and the Fock state ϕ_A such that subspaces $M(\mathscr{O}) \subset \mathscr{K}_0$ exist where (i) $\{\mathscr{F}(M(\mathscr{O}))\}$ is unitarily equivalent to the usual set of free Fermi algebras and this unitary equivalence intertwines the Fock vacua and also the representations of the Poincaré group in the obvious manner, (ii) $M(W') = iM(W)'$, all $W \in \mathscr{W}$. In the Fermi case this is more involved than for bosons, and unfortunately it has been carried out in detail only in Ref. 6. We feel therefore obliged to summarize the main points again here. To minimize notation we present the construction for spin $s = \frac{1}{2}$ and mass $m > 0$, but all other cases are explicitly dealt with in Ref. 6, which yields the desiderata (i) and (ii) for them, as well.

Let $\mathscr{T}_0 = \mathscr{S}(\mathbf{R}^4)^{\oplus 2}$ and let $\varphi(f)$ be the usual two-component Fermi spin-$\frac{1}{2}$ field. Double the test function space $\mathscr{T}_0 \oplus \mathscr{T}_0$ and define $\Phi(\underline{f}) = \varphi(f_1) + \varphi(\sigma_2\bar{f}_2)$, $\underline{f} = f_1 \oplus f_2 \in \mathscr{T}_0 \oplus \mathscr{T}_0$, where $\sigma_1,\sigma_2,\sigma_3$ are the Pauli matrices and $\bar{f}$ is the complex conjugate of f. One can verify that $\Phi(\underline{f})^* = \Phi(J\underline{f})$, where

$$J\underline{f} \equiv \begin{pmatrix} 0 & \sigma_2 \\ -\sigma_2 & 0 \end{pmatrix}\begin{pmatrix} \bar{f}_1 \\ \bar{f}_2 \end{pmatrix}, \quad J^2 = I\,,$$

is an antilinear involution on $\mathscr{T}_0 \oplus \mathscr{T}_0$. If Ω_F is the usual Fermi Fock vector, then

$$\langle \underline{f},\underline{g}\rangle_+ \equiv \langle \Omega_F, \Phi(\underline{f})^*\Phi(\underline{g})\Omega_F\rangle$$
$$= \int \frac{d^3\mathbf{p}}{2\omega_p}\left\langle \underline{f}_+(\mathbf{p}), \begin{pmatrix} \tilde{p}^+/m & 0 \\ 0 & \underline{p}_+/m \end{pmatrix}\tilde{\underline{g}}_+(\mathbf{p})\right\rangle, \tag{4.13}$$

where

$$\tilde{p}_+ = p_0 I - \mathbf{p}\cdot\boldsymbol{\sigma}|_{p_0=\omega_p}, \quad \underline{p}_+ = p_0 I + \mathbf{p}\cdot\boldsymbol{\sigma}|_{p_0=\omega_p},$$
$$\omega_p = \sqrt{\mathbf{p}^2+m^2},$$

and $\tilde{f}_\pm(\mathbf{p}) = \tilde{f}(\pm\omega_p,\mathbf{p})$, $\tilde{f}$ being the Fourier transform of f. Moreover,

$$\langle \underline{f},\underline{g}\rangle_- \equiv \langle \Omega_F, \Phi(\underline{g})\Phi(\underline{f})^*\Omega_F\rangle$$
$$= \int \frac{d^3\mathbf{p}}{2\omega_p}\left\langle \tilde{\underline{f}}_-(\mathbf{p}), \begin{pmatrix} \underline{p}_+/m & 0 \\ 0 & \tilde{p}^+/m \end{pmatrix}\tilde{\underline{g}}_-(\mathbf{p})\right\rangle. \tag{4.14}$$

One has $\langle J\underline{f},J\underline{g}\rangle_+ = \langle \underline{g},\underline{f}\rangle_-$. Let $\mathscr{K}_\pm$ be the Hilbert space completion of $\mathscr{T}_0\oplus\mathscr{T}_0$ under the norm $\|\cdot\|_\pm \equiv \langle\cdot,\cdot\rangle_\pm^{1/2}$, and let $\alpha_\pm : \mathscr{T}_0\oplus\mathscr{T}_0 \to \mathscr{K}_\pm$ be the canonical injection. Then, of course, there exists a unitary operator W: $\mathscr{F}\to\Gamma(\mathscr{K}_+)$, where $\mathscr{F}$ is the standard Fermi Fock space and $\Gamma(\mathscr{K}_+)$ is the Fermi Fock space with $\mathscr{K}_+$ as its one-particle subspace, such that (i) $W\Omega_F = \Omega$, the Fock vacuum in $\Gamma(\mathscr{K}_+)$, (ii) if $\underline{f}=J\underline{f}$, then $\Phi(\underline{f})$ is self-adjoint on $\mathscr{F}$ and $\psi(\alpha_+(\underline{f})) = W\Phi(\underline{f})W^*$, where for each $h\in\mathscr{K}_+$, $\psi(h) = (1/\sqrt{2})(A(h)^* + A(h))$ and $A(\cdot)^*$, $A(\cdot)$ form the irreducible Fock representation of the CAR in $\Gamma(\mathscr{K}_+)$ over $\mathscr{K}_+$. Thus,

$$[\psi(h),\psi(k)]_+ = \mathrm{Re}\langle h,k\rangle_+, \quad \text{all } h,k\in\mathscr{K}_+. \tag{4.15}$$

The representation $U_+(a,A)$ of the Poincaré group on $\mathscr{K}_+$ is

$$U_+(a,A)\alpha_+(\underline{f}) = \alpha_+((A\oplus A^{*-1})\underline{f}_{\{a,\Lambda(A)\}}),$$

where

$$\underline{f}_{\{a,\Lambda(A)\}}(x) = \underline{f}(\Lambda(A^{-1})(x-a)).$$

Here, $\Gamma(U_+(a,A))$ gives the representation on $\Gamma(\mathscr{K}_+)$ and W intertwines $\Gamma(U_+(a,A))$ and the usual representation on $\mathscr{F}$.

Unfortunately, in order to describe the local structure of algebras via support properties of the test functions in the manner we need, it is necessary to double the test function space again. To this end, let $\mathscr{K}_0\equiv\mathscr{K}_+\oplus\mathscr{K}_-$ and let α: $\mathscr{T}_0\oplus\mathscr{T}_0\to\mathscr{K}_0$ be defined by

$$\alpha(\underline{f}) = \alpha_+(\underline{f}) + \alpha_-(\underline{f}), \quad \underline{f}\in\mathscr{T}_0\oplus\mathscr{T}_0.$$

A unique antilinear involution Γ on $\mathscr{K}_0$ is induced by J via $\Gamma\alpha(\underline{f}) \equiv \alpha(J\underline{f})$. Let $\mathrm{Re}\,\mathscr{K}_0 \equiv \{h\in\mathscr{K}_0 | \Gamma h = h\}$, and let P: $\mathscr{K}_0\to\mathscr{K}_+$ be the projection onto $\mathscr{K}_+$. Since $\Gamma P = (I-P)\Gamma$, the map $\mathrm{Re}\,\mathscr{K}_0 \ni k\to\sqrt{2}Pk\in\mathscr{K}_+$ is an isomorphism between (real) Hilbert spaces.[21] A complex structure is thus induced on $\mathrm{Re}\,\mathscr{K}_0$ via this isomorphism

$$ik \equiv i(2P-I)k, \quad k\in\mathrm{Re}\,\mathscr{K}_0,$$

where the right-hand side is understood in $\mathscr{K}_0$. The complex scalar product on $\mathrm{Re}\,\mathscr{K}_0$ is then given by

$$\langle k,k'\rangle = \langle\sqrt{2}Pk,\sqrt{2}Pk'\rangle_{\mathscr{K}_+}, \quad k,k'\in\mathrm{Re}\,\mathscr{K}_0. \tag{4.16}$$

(This is equivalent to picking a particular Fock state; see Ref. 21.) The unitary representation $U(a,A)$ of the (covering group of the) Poincaré group is induced on $\mathscr{K}_0$ by

$$U(a,A)\alpha(\underline{f}) \equiv \alpha((A\oplus A^{*-1})\underline{f}_{\{a,\Lambda,(A)\}}),$$

and then induced on $\mathrm{Re}\,\mathscr{K}_0$ by restriction [note $U(a,A)$ commutes with Γ and P].

The net of local field algebras $\{\mathscr{F}(\mathscr{O})\}$ for the free Fermi field[5,7] is unitarily equivalent to the net of local algebras defined by

$$\mathscr{F}(\mathscr{O}) \equiv \{\psi(\alpha(\underline{f}))|\,\underline{f}\in\mathscr{T}_0\oplus\mathscr{T}_0,\ \underline{f}=J\underline{f},\ \mathrm{supp}(\underline{f})\subset\mathscr{O}\}'',$$

where $\psi(\alpha(\underline{f})) = \psi(\alpha_+(\underline{f}))\oplus\psi(\alpha_-(f))$ on $\Gamma(\mathscr{K}_0)$ $\cong\Gamma(\mathscr{K}_+\oplus\mathscr{K}_-)\cong\Gamma(\mathscr{K}_+)\oplus\Gamma(\mathscr{K}_-)$. Because of the norm continuity of the CAR, $\mathscr{F}(\mathscr{O}) = \{\psi(\alpha(\underline{f}))|\,\underline{f}\in M(\mathscr{O})\}''$, where $M(\mathscr{O})\subset\mathrm{Re}\,\mathscr{K}_0$ is the closure of $\{\alpha(\underline{f})|\,\underline{f}\in\mathscr{T}_0\oplus\mathscr{T}_0,\ \underline{f}=J\underline{f},\ \mathrm{supp}\,(f)\subset\mathscr{O}\}$ in $\mathrm{Re}\,\mathscr{K}_0$. From Satz II. 2.2 and Lemma II. 3.6 in Ref. 6 it follows that for any wedge region $W\in\mathscr{W}$,

$$iM(W')' = M(W). \tag{4.17}$$

The unitary that intertwines the field algebras as mentioned above also maps the usual Fock vacuum of the free Fermi field onto the vector $\Omega\equiv\Omega\oplus\Omega\in\Gamma(\mathscr{K}_+)\oplus\Gamma(\mathscr{K}_-)$ and intertwines the representations of the Poincaré group in the proper manner. These claims are straightforward consequences of the construction above and the existence of the unitary intertwiner W previously discussed.

In order to reduce the notational complexity, we shall for each $\underline{f}\in\mathscr{T}_0\oplus\mathscr{T}_0$ identify a vector $h=\alpha(\underline{f})$ in $\mathscr{K}_0$. We have, then, for each $h,k\in\mathrm{Re}\,\mathscr{K}_0$,

$$[\psi(h),\psi(k)]_+ = \mathrm{Re}\langle h,k\rangle \tag{4.18}$$

and

$$\phi_0(\psi(h)\psi(k)) = \langle\Omega,\psi(h)\psi(k)\Omega\rangle = \tfrac{1}{2}\langle h,k\rangle, \tag{4.19}$$

where $\langle\cdot,\cdot\rangle$ is the (complex) scalar product induced on $\mathrm{Re}\,\mathscr{K}_0$ in (4.16). Note that for each $h\in\mathrm{Re}\,\mathscr{K}_0$, $\psi(h)=\psi(h)^*$. Now, with $\mathscr{T}=\mathrm{Re}\,\mathscr{K}_0$, $q(\cdot,\cdot)=\mathrm{Re}\langle\cdot,\cdot\rangle$, $\phi_A(\cdot)$ the state $\langle\Omega,\cdot\,\Omega\rangle$, and the subspace $M(\mathscr{O})\subset\mathrm{Re}\,\mathscr{K}_0$ as given above, we have the desiderata (i), and (ii) in the abstract context presented at the outset of this section. We can now, using much the same tactics as in Sec. III, make an explicit calculation to prove Theorem 2.2.

Proof of Theorem 2.2: (1) From (4.18) it follows that $2\psi(h)^2 = \|h\|^2 I$, so that for $h\in\mathrm{Re}\,\mathscr{K}_0$ such that $\|h\|=\sqrt{2}$, $\psi(h)$ is a self-adjoint unitary. From Theorem 2.1 of Paper I only such operators are candidates for maximal violators. By using the mentioned unitary equivalence of the usual formulation of free Fermi theories with the one shown above, it follows that

$$\beta(\phi_0,\mathscr{F}(W),\mathscr{F}(W')')$$
$$\geq \tfrac{1}{2}(\langle\Omega,\psi(h_1)V\psi(k_1)V^*\Omega\rangle + \langle\Omega,\psi(h_1)V\psi(k_2)V^*\Omega\rangle$$
$$+ \langle\Omega,\psi(h_2)V\psi(k_1)V^*\Omega\rangle$$
$$- \langle\Omega,\psi(h_2)V\psi(k_2)V^*\Omega\rangle), \tag{4.20}$$

for any $h_i\in M(W)$ and $k_i\in M(W')$ satisfying $\|h_i\|^2 = 2 = \|k_i\|^2$, $i=1,2$. Using (4.9) and the fact that $U\Omega=\Omega$, the right-hand side of (4.20) is seen to be equal to

$$- \tfrac{1}{4}(\mathrm{Im}\langle h_1,k_1\rangle + \mathrm{Im}\langle h_1,k_2\rangle + \mathrm{Im}\langle h_2,k_1\rangle - \mathrm{Im}\langle h_2,k_2\rangle),$$

where we have employed the fact that $\mathrm{Re}\langle h_i,k_j\rangle = 0$, $i,j,=1,2$. It remains to choose the test functions appropriately. To that end let s be the canonical involution for the (standard[6]) subspace $M(W)$ and let $s = j\delta^{1/2}$ be its polar decomposition. Then, of course, $j^2 = 1$, $j\delta^{1/2} = \delta^{-1/2}j$, $j(M(W)) = M(W)'$, and $s^* = \delta^{1/2}j$ is the canonical involution for $M(W)'$. This is all the same as in the Bose calculation in Sec. III. The i in (4.17) is the only basic difference at this stage.

(2) Let ϕ, μ_+, and μ_- be as in the proof of Corollary 3.2 for some fixed $\epsilon > 0$ and define

$$\hat{h}_1 \equiv (1+s)(1-\mu_+)^{-1/2}\phi,$$
$$\hat{h}_2 \equiv (1+s)(1-\mu_+)^{-1/2}i\phi,$$
$$\hat{k}_1 \equiv i(1+s^*)(\mu_- - 1)^{-1/2}\phi,$$
$$\hat{k}_2 \equiv i(1+s^*)(\mu_- - 1)^{-1/2}i\phi.$$

Then $\hat{h}_i \in M(W)$ and $\hat{k}_i \in M(W')$, $i = 1,2$, by (4.17) and the argument given in Corollary 3.2. The same calculations as in Corollary 3.2. yield

$$\langle \hat{h}_1,\hat{h}_2\rangle = i = \langle \hat{k}_1,\hat{k}_2\rangle,$$
$$\|\hat{h}_l\|^2 = \langle 1+\mu_+\rangle/(1-\mu_+),$$
$$\|\hat{k}_l\|^2 = (1+\mu_-)/(\mu_- - 1), \quad l = 1,2,$$
$$\langle \hat{h}_1,\hat{k}_2\rangle = \langle \hat{h}_2,\hat{k}_1\rangle = 0,$$

and

$$\langle \hat{h}_1,\hat{k}_1\rangle = -\langle \hat{h}_2,\hat{k}_2\rangle = 2i(1-\mu_+)^{-1/2}(\mu_- - 1)^{-1/2}.$$

Then define

$$h_l \equiv [\sqrt{2}(1-\mu_+)^{1/2}/(1+\mu_+)^{1/2}]\hat{h}_l, \quad l = 1,2,$$
$$k_1 \equiv -[(\mu_- - 1)^{1/2}/(1+\mu_-)^{1/2}](\hat{k}_1 - \hat{k}_2),$$

and

$$k_2 \equiv -[(\mu_- - 1)^{1/2}/(1+\mu_-)^{1/2}](\hat{k}_1 + \hat{k}_2).$$

One sees that $\|h_l\|^2 = 2 = \|k_l\|^2$, that $h_l \in M(W)$, $k_l \in M(W')$, and that

$$\mathrm{Im}\langle h_1,k_1\rangle = -2\sqrt{2}/[(1+\mu_+)^{1/2}(1+\mu_-)^{1/2}] = \mathrm{Im}\langle h_2,k_1\rangle = \mathrm{Im}\langle h_1,k_2\rangle = -\mathrm{Im}\langle h_2,k_2\rangle.$$

Thus, one has from (4.20)

$$\beta(\phi_0,\mathscr{F}(W),\mathscr{F}(W')^t) \geqslant 2\sqrt{2}/[(1+\mu_+)^{1/2}(1+\mu_-)^{1/2}]. \tag{4.21}$$

(3) Since the spectrum of δ is $\mathbb{R}_+$ by the results of Bisognano and Wichmann[5] and by the unitary equivalence established above, for any $\epsilon > 0$ as in Corollary 3.2 there exists a unit vector ϕ with the stated properties. Letting $\epsilon \downarrow 0$ again as in Corollary 3.2, μ_+ and μ_- tend to 1, and the right-hand side of (4.21) tends to $\sqrt{2}$. Therefore, Theorem 2.2. is proven. □

V. MAXIMAL VIOLATION FOR FREE FERMI QUANTUM FIELDS: OBSERVABLE ALGEBRA CASE

In this section we can make use of the formalism already established in Sec. IV. The only new element to enter here is the group of gauge transformations generated by the free Fermi charge operator Q, but the additional requirement that the observables be invariant under this group causes some complications.

First of all, if $\varphi(f)$ is that part of the Fermi field on the usual Fock space that has charge $+$, then

$$e^{itQ}\varphi(f)e^{-itQ} = \varphi(e^{it}f).$$

Similarly, that part of the Fermi field that has charge $-$ transforms to $\varphi(e^{-it}f)$. In other words, the gauge transformations can also be expressed as unitary maps on the one-particle spaces, and tracing the connection of the usual field with that we constructed in Sec. IV, one sees that the gauge transformations can be expressed on the one-particle space $\mathrm{Re}\,\mathscr{K}_0$ as

$$f \to u_t f, \quad u_t \equiv \cos t + v \sin t,$$

where v is an orthogonal operator satisfying $v^2 = -1$. The gauge transformations commute with representation of the Poincaré group on $\mathrm{Re}\,\mathscr{K}_0$. Moreover, v and u_t commute with the complex structure $\mathrm{Re}\,\mathscr{K}_0$ defined in Sec. IV (since the vacuum state is gauge invariant), thus u_t and v are unitary.

Proof of Theorem 2.3: Let $M(W)$, $M(W')$, s, j, and δ be as in Sec. IV, and let $u_t = \cos t + v \sin t$ be the gauge transformation on the one-particle space. v commutes with s, j, δ. Since $v^2 = -1$, v has the eigenvalues $\pm i$. Here u_t and v commute with the representation of the Poincaré group on the one-particle space; thus in each of the eigenspaces for v there is a nontrivial representation of the Poincaré group. Since the boosts leaving W and W' invariant are represented by δ^{it}, this implies that the spectrum of δ equals $\mathbb{R}_+$ for both subspaces.

(1) Fix $N \in \mathbb{N}$ and $\epsilon > 0$. Then pick N test functions $\varphi_1,\ldots,\varphi_N$ of norm one with φ_ν in the spectral subspace of δ for the interval $[1 + (\nu/N)\epsilon, 1 + ((\nu+1)/N)\epsilon]$ and belonging to one of the spectral subspaces of v, each $\nu = 1,\ldots,N$. Thus φ_ν is in the domain of definition of $\delta^{1/2}$ and $\delta^{-1/2}$ and one can set

$$f_\nu \equiv (1/\sqrt{2})(1+j\delta^{1/2})\varphi_\nu, \quad g_\nu \equiv (i/\sqrt{2})(1+j\delta^{-1/2})\varphi_\nu.$$

Then $f_\nu \in M(W)$ since $\frac{1}{2}(1+s)$ projects onto $M(W)$ and $g_\nu \in iM(W)' = M(W')$ since $\frac{1}{2}(1+s^*)$ projects onto $M(W)'$. The inner products between the elements of $\{f_1,\ldots,f_N,g_1,\ldots,g_N\}$ are sums of expressions of the form $\langle \varphi_\nu, C\varphi_\mu\rangle$, where C is some product of the operators j, $\delta^{\pm 1/2}$, and v. Because v commutes with δ and $j\,\delta j = \delta^{-1}$, φ_ν and $C\varphi_\mu$ always belong to orthogonal subspaces if $\nu \neq \mu$; and if C contains an odd number of factors j, φ_ν and $C\varphi_\nu$ are also orthogonal. Using

$$vF_\nu = \pm (i/\sqrt{2})(1-j\delta^{1/2})\varphi_\nu,$$
$$vg_\nu = \pm (i/\sqrt{2})(1-j\delta^{-1/2})\varphi_\nu,$$

one calculates

$$\langle f_\nu,f_\nu\rangle = \tfrac{1}{2}(1+C_\nu^+), \quad \langle g_\nu,g_\nu\rangle = \tfrac{1}{2}(1+C_\nu^-), \tag{5.1}$$

$$\langle f_\nu,g_\nu\rangle = i, \quad \langle f_\nu,vg_\nu\rangle = 0, \tag{5.2}$$

$$\langle f_\nu, vf_\nu\rangle = \pm (i/2)(1 - C_\nu^+),$$
$$\langle g_\nu, vg_\nu\rangle = \pm (i/2)(1 - C_\nu^-), \tag{5.3}$$

with $C_\nu^\pm \equiv \langle \varphi_\nu, \delta^{\pm 1}\varphi_\nu\rangle$. Since C_ν^+ and $(C_\nu^-)^{-1}$ are contained in $[1, 1 + ((\nu+1)/N)\epsilon]$, one sees that for small ϵ they are very close to one.

(2) Pick N in step (1) such that $N = 2^k$, and let

$$f_{\nu+N} \equiv vf_\nu \text{ and } g_{\nu+N} \equiv vg_\nu, \quad 1 \leqslant \nu \leqslant N.$$

Consider the real linear subspace $M_\epsilon \subset M(W)$ [resp. $N_\epsilon \subset M(W')$] spanned by $f_1, \ldots, f_{2N}$ (resp. $g_1, \ldots, g_{2N}$). Here, M_ϵ and N_ϵ are real Hilbert spaces of dimension $2N$ with respect to $\mathrm{Re}\langle\cdot,\cdot\rangle$ and on M_ϵ, N_ϵ v is an orthogonal transformation with $v^2 = -1$. By step 1, if ϵ is small enough there must exist an orthonormal basis $\{\tilde f_\nu\}_{\nu=1}^{2N} \subset M_\epsilon$ with $\|f_\nu - \tilde f_\nu\| = 0(\epsilon)$ and $v\tilde f_\nu = \tilde f_{\nu+N}$ for $\nu \leqslant N$ (similarly for $\{\tilde g_\nu\}_{\nu=1}^{2N} \subset N_\epsilon$). As $\epsilon \downarrow 0$ the algebras $\mathscr{F}(M_\epsilon)$ and $\mathscr{F}(N_\epsilon)$ move around in $\mathscr{F}(W)$ and $\mathscr{F}(W')$. Since what one is interested in here are expectations of certain operators in $\mathscr{F}(M_\epsilon) \vee \mathscr{F}(N_\epsilon) \cong \mathscr{F}(M_\epsilon \oplus N_\epsilon)$ in the quasifree state ϕ_0, for the sake of technical convenience it is preferable to identify these algebras for different ϵ and to consider different states ϕ_ϵ on this one (identified) algebra. In particular, by identifying for different ϵ but the same ν, $1 \leqslant \nu \leqslant 2N$, the basis vectors $\tilde f_\nu$ (similarly for $\tilde g_\nu$), the algebras $\mathscr{F}(M_\epsilon \oplus N_\epsilon)$ are all isomorphic for different ϵ. Thus, $\phi_0 \restriction \mathscr{F}(M_\epsilon \oplus N_\epsilon)$ is given by some state $\phi_\epsilon \restriction \mathscr{F}(M_{\epsilon_0} \oplus N_{\epsilon_0})$ $[\cong \mathscr{F}(M_{\epsilon_0}) \vee \mathscr{F}(N_{\epsilon_0})]$ for some fixed, sufficiently small $\epsilon_0 > 0$ and all $0 < \epsilon \leqslant \epsilon_0$. The value of ϕ_ϵ on any monomial in $\psi(\tilde f_\nu), \psi(\tilde g_\nu)$ is a polynomial in the inner products $\langle \tilde f_\nu, \tilde f_\mu\rangle$, $\langle \tilde g_\nu, \tilde g_\mu\rangle$, $\langle \tilde f_\nu, \tilde g_\mu\rangle$. Hence, as $\epsilon \downarrow 0$ the states ϕ_ϵ on $\mathscr{F}(M_{\epsilon_0} \oplus N_{\epsilon_0})$ converge to a quasifree state $\tilde\phi_0$ and

$$\beta(\phi_0, \mathscr{A}(M(W)), \mathscr{A}(M(W')))$$
$$\geqslant \sup_{\epsilon_0 > \epsilon > 0} \beta(\tilde\phi_\epsilon, \mathscr{A}(N_{\epsilon_0}), \mathscr{A}(N_{\epsilon_0}))$$
$$\geqslant \beta(\tilde\phi_0, \mathscr{A}(M_{\epsilon_0}), \mathscr{A}(N_{\epsilon_0})), \tag{5.4}$$

where $\mathscr{A}(M)$ is the fixed-point subalgebra of $\mathscr{F}(M)$ under the gauge automorphism group induced on $\mathscr{F}(M(\mathbb{R}^4))$ by the action of the group $\{\cos t + v \sin t\}_{t\in\mathbb{R}}$ on $\mathrm{Re}\,\mathscr{K}_0$. [Recall that $\mathscr{A}(M) \subset \mathscr{F}(M) \cap \mathscr{F}(M)'$.]

(3) In the above limiting process none of the subspaces $M_\epsilon \oplus N_\epsilon$ is in general invariant under multiplication by i. Thus, in the limit state $\tilde\phi_0$ one no longer has the complex structure associated to the state ϕ_0. Rather, one has a real Hilbert space $\mathscr{T}_{\epsilon_0} \equiv M_{\epsilon_0} \oplus N_{\epsilon_0}$ with inner product $q(\cdot,\cdot)$ and the orthogonal operator v, and a quasifree state $\tilde\phi_0$ determined by a real bilinear form $\sigma(\cdot,\cdot)$. Define now an operator $\tilde i$ on $\mathscr{T}_{\epsilon_0}$ by $\tilde i \tilde f_\nu \equiv \tilde g_\nu$, $\tilde i \tilde g_\nu \equiv -\tilde f_\nu$. Then one verifies that $\sigma(h_1, h_2) = q(\tilde i h_1, h_2)$ and that $\tilde i^2 = -1$ (the latter by definition, the former by considering ϕ_ϵ and taking the limit $\epsilon \downarrow 0$). It follows that the quasifree state $\tilde\phi_0$ is pure[19] on the algebra $\mathscr{F}(T_{\epsilon_0})$. Moreover, $\tilde i v = v\tilde i$, so that v is unitary in the complexification of T_{ϵ_0} given by $\tilde i$ [see (4.3)].

(4) Since $\tilde\phi_0$ is pure and $\mathscr{F}(T_{\epsilon_0})$ is irreducible in the corresponding Gel'fand–Naimark–Segal (GNS) representation on a Hilbert space $\tilde{\mathscr{H}}$, one has $\mathscr{F}(\mathscr{T}_{\epsilon_0}) \cong \mathscr{B}(\tilde{\mathscr{H}})$ and $\tilde\phi_0(F) = \langle\tilde\Omega, F\tilde\Omega\rangle$ for some unit vector $\tilde\Omega \in \tilde{\mathscr{H}}$. The gauge group $\{\alpha_t\}_{t\in\mathbb{R}}$ induced on $\mathscr{F}(\mathscr{T}_{\epsilon_0})$ by the group $\{\cos t + v \sin t\}_{t\in\mathbb{R}}$ is implementable in $\tilde{\mathscr{H}}$ as

$$\alpha_t(F) = e^{i\tilde Q t} F e^{-i\tilde Q t},$$

with $\tilde Q = \tilde Q^* \in \mathscr{B}(\tilde{\mathscr{H}})$ determined up to a constant. Here, $\tilde Q$ has a spectral resolution $\tilde Q = \Sigma_{k=0}^d (q_0 + k) P_k$ with $q_0 \in \mathbb{R}$, $d = \frac{1}{2}\dim_{\mathbb{R}} \mathscr{T}_{\epsilon_0} = 2N$, and $\mathrm{tr}\, P_k = \binom{2N}{k}$. Since $\tilde\phi_0$ is gauge invariant, $\tilde\Omega$ must be an eigenvector of $\tilde Q$. The corresponding value of k is determined by studying the unitary operator $v\tilde i$, which satisfies $(v\tilde i)^2 = 1$ and hence has eigenvalues ± 1. In the case at hand, one verifies that the two eigenspaces have equal dimension, so there is a symmetry leaving $\tilde\Omega$ invariant but taking α_t to α_{-t}. Hence $\tilde Q\tilde\Omega = (q_0 + d/2)\tilde\Omega = (q_0 + N)\Omega$, and following the standard convention, one can choose $q_0 = -N$.

(5) One notes that $\mathscr{A}(M_{\epsilon_0})$ can be alternatively described as the subalgebra of $\mathscr{F}(M_{\epsilon_0})$ that is fixed under the automorphism $\alpha_t^{(1)}$ of $\mathscr{F}(T_{\epsilon_0})$ induced by

$$\alpha_t^{(1)}(\psi(f \oplus g)) \equiv e^{iQ_1 t}\psi(f \oplus g) e^{-iQ_1 t}$$
$$\equiv \psi([(\cos t + v \sin t) f] \oplus g)$$

for $f \in M_{\epsilon_0}$ and $g \in N_{\epsilon_0}$. Thus, $\mathscr{A}(M_{\epsilon_0}) = \Sigma_{k=0}^{N_\oplus} \mathscr{A}_k(M_{\epsilon_0})$, where k labels the eigenspaces of Q_1 and $\mathscr{A}_k(M_{\epsilon_0}) = \mathscr{B}(\mathbb{C}^{d_k})$, $d_k = \binom{N}{k}$. Note that since $\mathrm{Im}\langle\cdot,\cdot\rangle = \sigma(\cdot,\cdot)$ vanishes on M_{ϵ_0}, $\tilde\phi_0$ is the unique normalized tracial state on $\mathscr{F}(M_{\epsilon_0})$. In particular, $\langle\Omega, \cdot\Omega\rangle$ is the trace on each summand $\mathscr{A}_k(M_{\epsilon_0})$, normalized to $d_k \cdot 2^{-N}$. If $\alpha_t^{(2)}$ and Q_2 are defined analogously for $\mathscr{F}(N_{\epsilon_0})$, there is a similar decomposition here, as well. Moreover, $\alpha_t^{(1)} \cdot \alpha_t^{(2)} = \alpha_t$, so that $Q_1 + Q_2 = \tilde Q$ up to some constant. Consequently, the algebra $\mathscr{A}(M_{\epsilon_0}) \cdot \mathscr{A}(N_{\epsilon_0})$, which is strictly contained in $\mathscr{A}(M_{\epsilon_0} \oplus N_{\epsilon_0})$, is decomposed as

$$\sum_{k_1, k_2 = 0}^{N_\oplus} \mathscr{A}_{k_1}(M_{\epsilon_0}) \mathscr{A}_{k_2}(N_{\epsilon_0}).$$

The state $\tilde\phi_0$ vanishes on every summand with $k_1 + k_2 \neq N$, is pure on each summand $\mathscr{A}_k(M_{\epsilon_0}) \mathscr{A}_{N-k}(N_{\epsilon_0})$, and restricts to the trace in each factor.

(6) An operator $A \in \mathscr{A}(M_{\epsilon_0})$ decomposed as $A = \Sigma_k A^{(k)}$ with $A^{(k)} \in \mathscr{A}_k(M_{\epsilon_0})$ is a self-adjoint contraction if and only if each of the summands $A^{(k)}$ is a self-adjoint contraction. Hence maximizing the expression

$$\tfrac{1}{2}\tilde\phi_0(A_1(B_1 + B_2) + A_2(B_1 - B_2))$$
$$= \sum_k \frac{1}{2}\tilde\phi_0(A_1^{(N-k)}(B_1^{(k)} + B_2^{(k)}) + A_2^{(N-k)}(B_1^{(k)} - B_2^{(k)})$$

over all self-adjoint contractions $A_i \in \mathscr{A}(M_{\epsilon_0})$, $B_i \in \mathscr{A}(N_{\epsilon_0})$ is equivalent to maximizing each summand. Since $\tilde\phi_0$ is normalized on $\mathscr{A}_{N-k}(M_{\epsilon_0}) \mathscr{A}_k(N_{\epsilon_0})$ to $2^{-N}\binom{N}{k}$, one can conclude that

$$\beta(\tilde\phi_0, \mathscr{A}(M_{\epsilon_0}), \mathscr{A}(N_{\epsilon_0}))$$
$$= \sum_{k=0}^{N} 2^{-N}\binom{N}{k} \beta(\tilde\phi_0^{(k)}, \mathscr{A}_{N-k}(M_{\epsilon_0}), \mathscr{A}_k(N_{\epsilon_0})), \tag{5.5}$$

with

$$\tilde{\phi}_0^{(k)} = 2^N \binom{N}{k}^{-1} \tilde{\phi}_0 \upharpoonright \mathscr{A}_{N-k}(M_{\epsilon_0}) \cdot \mathscr{A}_k(N_{\epsilon_0}) .$$

Recall that $\tilde{\phi}_0^{(k)}$ is a pure state on $\mathscr{A}_{N-k}(M_{\epsilon_0})\mathscr{A}_k(N_{\epsilon_0})$ and that $\tilde{\phi}_0^{(k)}$ restricted to each factor is the normalized trace. Since $\mathscr{A}_{N-k}(M_{\epsilon_0}) \cong \mathscr{A}_k(N_{\epsilon_0}) \cong \mathscr{B}(\mathbb{C}^{d_k})$ with $d_k = \binom{N}{k}$, it is possible to determine β from facts already established in this paper. Namely, for $k = 0$ or $k = N$, $\dim \mathscr{A}_k(M_{\epsilon_0}) = 1$, so that the corresponding β in (5.5) is equal to 1. For the other terms, note that since N is a power of 2, $\binom{N}{k}$ is even for $1 \leqslant k \leqslant N-1$. Using matched decompositions of the factors $\mathbb{C}^{d_k}$ in $\mathbb{C}^{d_k} \otimes \mathbb{C}^{d_k}$ into two-dimensional subspaces, and picking self-adjoint contractions in each of these subspaces in much the same way as in step 2 of the proof of Proposition 3.1, one sees that

$$\beta(\tilde{\phi}_0^{(k)}, \mathscr{A}_{N-k}(M_{\epsilon_0}), \mathscr{A}_k(N_{\epsilon_0})) = \sqrt{2}, \quad \text{for } 1 \leqslant k \leqslant N-1 .$$

Hence, by (5.5)

$$\beta(\tilde{\phi}_0, \mathscr{A}(M_{\epsilon_0}), \mathscr{A}(N_{\epsilon_0})) = 2 \cdot 2^{-N} + \sqrt{2}(1 - 2 \cdot 2^{-N}) . \tag{5.6}$$

From (5.6) and (5.4), Theorem 2.3 follows by taking $N \to \infty$. □

ACKNOWLEDGMENTS

Most of this work was carried out when one of us (S.J.S.) was either resident in (until March 1985) or visiting the Fachbereich Physik, Universität Osnabrück (Summer, 1986).

S.J.S. wishes to thank his friends and colleagues at Osnabrück for their hospitality and acknowledges the financial support of a University of Rochester Faculty Research Grant during the summer, 1986.

[1] S. J. Summers and R. Werner, "Bell's inequalities and quantum field theory, I: General setting," J. Math. Phys. **28**, 2440 (1987).
[2] S. J. Summers and R. Werner, "The vacuum violates Bell's inequalities," Phys. Lett. A **110**, 257 (1985).
[3] R. Jost, *The General Theory of Quantized Fields* (American Mathematical Society, Providence, RI, 1965).
[4] H. Araki, "Von Neumann algebras of local observables for free scalar field," J. Math. Phys. **5**, 1 (1964).
[5] J. J. Bisognano and E. H. Wichmann, "On the duality condition for quantum fields," J. Math. Phys. **17**, 303 (1976).
[6] J. J. Foit, "Twisted Dualität für freie Fermifelder," Ph.D. thesis, Universität Osnabrück, 1983.
[7] S. Doplicher, R. Haag, and J. E. Roberts, "Fields, observables and gauge transformations, I," Commun. Math. Phys. **13**, 1 (1969).
[8] J. J. Bisognano and E. H. Wichmann, "On the duality condition for a Hermitian scalar field," J. Math. Phys. **16**, 985 (1975).
[9] J. Manuceau and A. Verbeure, "Quasi-free states of the C.C.R.-algebra and Bogoliubov transformations," Commun. Math. Phys. **9**, 293 (1968).
[10] P. Leyland, J. E. Roberts, and D. Testard, "Duality for quantum free fields," preprint, CNRS CPT 78/1016, Marseille, 1978.
[11] H. Araki, "A lattice of von Neumann algebras associated with the quantum theory of a free Bose field," J. Math. Phys. **4**, 1343 (1963).
[12] J.-P. Eckmann and K. Osterwalder, "An application of Tomita's theory of modular Hilbert algebras: Duality for free Bose fields," J. Funct. Anal. **13**, 1 (1973).
[13] H. Reeh and S. Schlieder, "Bemerkungen zur Unitäräquivalenz von Lorentzinvarianten Feldern," Nuovo Cimento **22**, 1051 (1961).
[14] M. Rieffel and A. van Daele, "A bounded operator approach to Tomita-Takesaki theory," Pac. J. Math. **69**, 187 (1977).
[15] M. Takesaki, "Tomita's theory of modular Hilbert algebras and its applications," *Lecture Notes in Mathematics*, Vol. 128 (Springer, Berlin, 1970).
[16] R. Werner, "Quantum harmonic analysis on phase space," J. Math. Phys. **25**, 1404 (1984).
[17] J. von Neumann, "Die Eindeutigkeit der Schrödingerschen Operatoren," Math. Ann. **104**, 570 (1931).
[18] S. Sakai, *C*-Algebras and W*-Algebras* (Springer, New York, 1971).
[19] E. Balslev, J. Manuceau, and A. Verbeure, "Representations of anticommutation relations and Bogoliubov transformations," Commun. Math. Phys. **8**, 315 (1968).
[20] J. J. Foit, "Abstract twisted duality for quantum free Fermi fields," Publ. RIMS, Kyoto Univ. **19**, 729 (1983).
[21] H. Araki, "On quasi-free states of CAR and Bogoliubov transformations," Publ. RIMS, Kyoto Univ. **6**, 385 (1970).

Generalization of the $N=1$ supersymmetric effective Lagrangian to arbitrary N in the absence of central charges

A. El Hassouni, E. G. Oudrhiri-Safiani, and E. H. Saidi
Laboratory of Theoretical Physics, Faculty of Sciences, Av. Ibn Batouta, P. O. Box 1014, Rabat, Morocco

(Received 28 March 1986; accepted for publication 25 February 1987)

Using nonlinear realization, extended supersymmetry breaking is studied. The N-extended Volkov–Akulov and standard superfields are constructed and the N generalized Wess constraints in the presence of central charges are given. The extended Volkov–Akulov Lagrangian and the N-extended effective Lagrangian are constructed for arbitrary N in the absence of central charges.

I. INTRODUCTION

Nonlinear realizations and construction of an effective Lagrangian play an important role in understanding symmetry breaking.[1] They were first introduced in the context of chiral dynamics[2] and in the strong interaction phenomenology.[3] These nonlinear realizations can be regarded as a linear theory at energies much lower than the breaking mass scale.

Recently, nonlinear realizations have been used to investigate supersymmetry breaking.[4] The Lagrangian is invariant under nonlinear transformations of the supersymmetry group SP_4^N and under linear transformations of the Poincaré subgroup. When the SP_4^N group is broken down to its subgroup P_4, a set of N fermionic Goldstone fields—goldstinos—emerges. These N goldstinos transform nonlinearly under the coset group SP_4^N/P_4.[5]

Wess showed—in the case of one supersymmetry—that using the transformation laws and one goldstino, one can construct from any Poincaré-(eventually gauge-) invariant Lagrangian an effective supersymmetric (gauge) one invariant under SP_4 modulo some constraints. The part of the new theory that does not depend on the goldstino field is exactly the original theory, a fact that allows us to take into account the supersymmetric effects at low energy.

The aim of this paper is to generalize for an arbitrary N the preceding results. We give the corresponding Wess constraints on the superfields with central charges with $N\geqslant 2$. On the other hand, we establish an N extended supersymmetric Lagrangian when the central charges are set to zero. The presentation is as follows.

In Sec. II the superspace and the superalgebra are recalled together with Wess constraints and the effective $N=1$ supersymmetric Lagrangian using the superspace formalism is given. In Sec. III we start by giving the notation for the N superspace and superalgebra, and then we give the N-extended superfields with the generalized constraints in the presence of central charges.

Section IV is devoted to the construction of the N-extended supersymmetric Lagrangian. We start by discussing the $N=1$ case. We note that, making use of the generalized derivative Δ_μ,[5,6] the Lagrangian $L(\phi,\Delta_\mu\phi)$ satisfies the standard transformation law, hence it allows the construction of a super-Lagrangian. This yields, through the Wess procedure, an $N=1$ effective supersymmetric Lagrangian. This method contains both the Wess Lagrangian based on the superspace technique and the other one[7] based on the ordinary fields. In addition, we show that they are equivalent up to the fourth order in terms of the goldstino fields. The construction of the N-effective theory goes through a natural extension of the above idea of the $N=1$ case. However, in the presence of central charges we came across some difficulties when constructing the transformation laws giving a realization of the SP_4^N algebra for $N\geqslant 2$. We have therefore limited ourselves to the case where the central charges (Z) are set to zero, which implies that the symmetry group of the SP_4^N generators Q reduces down to $SU(N)/Z_N$.

II. $N=1$ SUPERSYMMETRIC ALGEBRA—REALIZATIONS AND THE $N=1$ WESS LAGRANGIAN

It is well known that the superalgebra $SP_4=\{P_\mu, M_{\mu\nu}, Q_a, \bar{Q}^{\dot{a}}\}$ (see Refs. 8–10) is given—using the constant Weyl Grassman variables ξ and η—by the following relations of the Lie algebra:

$$\begin{aligned}&\text{(a)}\ [\xi Q,\bar{\xi}\bar{Q}]=-2\xi\sigma^\mu\bar{\xi}P_\mu,\\ &\text{(b)}\ [\xi Q,\eta Q]=[\bar{\xi}\bar{Q},\bar{\eta}\bar{Q}]=0.\end{aligned}\tag{2.1}$$

In the following we shall use Weyl spinors and our notations are identical to those of Bagger and Wess.[9]

Two realizations of the algebra (2.1) have been constructed.[4,5] They are given by (i) the Volkov–Akulov nonlinear realization

$$\begin{aligned}\delta_\xi\lambda_a(x)&=f\xi_a+\zeta^\mu\partial_\mu\lambda_a(x),\\ \delta_\xi\bar{\lambda}_{\dot{a}}(x)&=f\bar{\xi}_{\dot{a}}+\zeta^\mu\partial_\mu\bar{\lambda}_{\dot{a}}(x);\end{aligned}\tag{2.2}$$

and (ii) the standard realization

$$\delta_\xi\phi(x)=+\zeta^\mu\partial_\mu\phi(x),\tag{2.3}$$

where

$$\partial_\mu\xi_\alpha=0,\quad \zeta^\mu=-(i/f)(\xi\sigma^\mu\bar{\lambda}-\lambda\sigma^\mu\bar{\xi}).\tag{2.4}$$

Here f is a real two-dimensional mass unknown constant that characterizes the supersymmetry breaking, $\lambda_a(x)$ and $\bar{\lambda}_{\dot{a}}(x)$ are two Weyl fields (goldstino),[5] and $\phi(x)$ is a generic field, which can be a matter, Higgs, or gauge field. These transformation laws give a realization of the algebra:

$$\begin{aligned}&[\delta_\xi,\delta_\eta]f(x)=2i(\xi\sigma^\mu\bar{\eta}-\eta\sigma^\mu\bar{\xi})\partial_\mu f(x),\\ &f(x)=\lambda_a(x),\ \bar{\lambda}_{\dot{a}}(x),\ \text{or}\ \phi(x).\end{aligned}\tag{2.5}$$

 0022-2488/87/102457-06$02.50 © 1987 American Institute of Physics

On the other hand, given a realization of (2.1) one can usually construct an $N=1$ superfield as a function defined on the superspace $(x,\theta,\bar{\theta})$ and which can be expressed by its finite expansion in terms of θ and $\bar{\theta}$.[8] Thus (2.2) and (2.3) give the (i) $N=1$ Volkov–Akulov superfield,

$$\begin{aligned}\Lambda_a(x,\theta,\bar{\theta}) &= e^{\delta_\theta}\lambda_a(x),\\ \Lambda_{\dot a}(x,\theta,\bar{\theta}) &= e^{\delta_\theta}\bar{\lambda}_{\dot a}(x);\end{aligned} \tag{2.6}$$

and (ii) $N=1$ standard superfield,

$$\Phi(x,\theta,\bar{\theta}) = e^{\delta_\theta}\phi(x), \tag{2.7}$$

where

$$\delta_\theta = \theta^a Q_a + \theta_{\dot a}\bar{Q}^{\dot a}.$$

These superfields can also be defined as the unique solutions of the following Wess constraints[5,8]:

$$\begin{aligned}&D_a\Lambda_b = \epsilon_{ba} + i\sigma^\mu{}_{a\dot a}\bar{\Lambda}^{\dot a}\partial_\mu\Lambda_b,\\ &\bar{D}_{\dot a}\Lambda_b = -i\Lambda^a\sigma^\mu{}_{a\dot a}\partial_\mu\Lambda_b,\\ &\bar{D}_{\dot a}\bar{\Lambda}_b = \epsilon_{b\dot a} - i\Lambda^a\sigma^\mu{}_{a\dot a}\partial_\mu\bar{\Lambda}_b,\\ &D_a\bar{\Lambda}_b = i\sigma^\mu{}_{a\dot a}\bar{\Lambda}^{\dot a}\partial_\mu\bar{\Lambda}_b,\\ &D_a\Phi = i\sigma^\mu{}_{a\dot a}\bar{\Lambda}^{\dot a}\partial_\mu\Phi,\\ &\bar{D}_{\dot a}\Phi = -i\Lambda^a\sigma^\mu{}_{a\dot a}\partial_\mu\Phi,\end{aligned} \tag{2.8}$$

where D_a and $\bar{D}_{\dot a}$ are the usual $N=1$ spinor covariant derivatives.[8,9] Now let us focus on the construction of the $N=1$ effective Lagrangian. Wess showed that using (2.6) and (2.7) one can generalize any Lorentz (eventually gauge) invariant Lagrangian $L(\phi,\partial\phi)$ to an effective supersymmetric one compatible with low energy phenomenology.[5] This compatibility is due to the fact that the part of the new Lagrangian that does not depend on the Goldstone field is identical to the original Lagrangian:

$$L_{\text{Wess}}^{N=1} = \int d^2\theta\, d^2\bar{\theta} f^{-4}\Lambda^2\bar{\Lambda}^2\left[-\frac{1}{2}f^2 + L(\Phi,\Delta_\mu\Phi)\right]. \tag{2.9}$$

Here $\Lambda^2\bar{\Lambda}^2$ is the Wess weight and Δ_μ is the generalized covariant derivative defined in Refs. 5 and 6 as

$$\Delta_\mu\phi = (\partial_\mu\Phi)\big|_{\substack{\theta=-\lambda\\ \bar{\theta}=-\lambda}} = (\partial_\mu(e^{\delta_\theta}\phi))\big|_{\substack{\theta=-\lambda\\ \bar{\theta}=-\lambda}}.$$

The integration of this weight with respect to θ and $\bar{\theta}$ gives, up to $\mathcal{O}(\lambda^4)$,

$$\begin{aligned}&f^{-4}\int d^2\theta\, d^2\bar{\theta}\,\Lambda^2\bar{\Lambda}^2\\ &\quad = 1 - (i/f^2)(\lambda\sigma^\mu\partial_\mu\bar{\lambda} - \partial_\mu\lambda\sigma\mu\bar{\lambda}) + \mathcal{O}(\lambda^2\bar{\lambda}^2, f^4).\end{aligned} \tag{2.10}$$

The constant value (1) is crucial in this study since it carries the supersymmetry breaking factor and reproduces the original theory:

$$\begin{aligned}&f^{-4}\int d^2\theta\, d^2\bar{\theta}\,\Lambda^2\bar{\Lambda}^2\left[-\frac{1}{2}f^2 + L(\Phi,\Delta_\mu\Phi)\right]\\ &\quad = -\tfrac{1}{2}f^2 + (i/2)(\lambda\sigma^\mu\partial_\mu\bar{\lambda} - \partial_\mu\lambda\sigma^\mu\bar{\lambda})\\ &\qquad + L(\phi,\Delta_\mu\phi) + \mathcal{O}(\lambda^2\bar{\lambda}^2, f^{-2}).\end{aligned} \tag{2.9'}$$

We stress the fact that (2.9) is manifestly supersymmetric since it is expressed in superspace; on the other hand it contains the original Lagrangian.

The term $-\frac{1}{2}f^2$ exhibits the fact that the theory is spontaneously broken from $N=1$ to $N=0$,

$$\langle -L_{\text{Wess}}^{N=1}\rangle = \tfrac{1}{2}f^2 > 0, \tag{2.11}$$

and finally all supersymmetric effects at low energy are carried by the term $\mathcal{O}(\lambda^2\bar{\lambda}^2, f^{-2})$.

Prior to the generalization of Wess's $N=1$ supersymmetric Lagrangian to an arbitrary N, we shall construct the N-extended Volkov–Akulov and standard superfields.

III. N-EXTENDED VOLKOV–AKULOV AND STANDARD SUPERFIELDS

Let us start by fixing our notation. The N-extended superspace is parametrized by $Z=(X,\Theta,\bar{\Theta})$, where Θ and $\bar{\Theta}$ are a set of N variables belonging to the vector representation $\mathbf{N}$ and its conjugate $\bar{\mathbf{N}}$ of the $SU(N)$ group. A contravariant $SU(N)$ index will be used to refer to the vector representation and a covariant one to its conjugate $\bar{\mathbf{N}}$. Therefore the odd part of the superspace Z can be written as

$$\Theta_a^i = \begin{pmatrix}\theta_a^1\\ \vdots\\ \theta_a^N\end{pmatrix} \sim (\mathbf{N};(\mathbf{2},\mathbf{1})), \quad \Theta_1^a = (\theta_1^a\cdots\theta_N^a)\sim(\bar{\mathbf{N}},(\mathbf{2},\mathbf{1})),$$

$$\bar{\Theta}_i^{\dot a} = \begin{pmatrix}\bar{\Theta}_1^{\dot a}\\ \vdots\\ \bar{\Theta}_N^{\dot a}\end{pmatrix} \sim (\bar{\mathbf{N}};(\mathbf{1},\mathbf{2})), \quad \bar{\Theta}_{\dot a}^i = (\bar{\theta}_{\dot a}^i\cdots\bar{\theta}_{\dot a}^N)\sim(\mathbf{N},(\mathbf{1},\mathbf{2})), \tag{3.1}$$

where $(N;(2,1))$, etc., give the corresponding representations of the direct product group $SU(N)\times SL(2,C)$.

A supertranslation in this superspace has the form

$$\Theta' = \Theta + \xi, \quad \xi_a^i = \begin{pmatrix}\xi_a^1\\ \vdots\\ \xi_a^N\end{pmatrix}. \tag{3.2}$$

The infinitesimal variation δ_Θ can be written as

$$\delta_\Theta = \Theta Q + \bar{\Theta}\bar{Q} = \sum_{i=1}^N(\theta_i^a Q_a^i + \bar{\theta}_{\dot a}^i\bar{Q}_i^{\dot a}), \tag{3.3}$$

where the Q's are the usual supersymmetric generators and satisfy the following algebra:

$$\begin{aligned}&[\xi_i^a Q_a^i, \bar{\xi}_{\dot a}^j\bar{Q}_j^{\dot a}] = -2\xi_i\sigma^\mu\bar{\xi}^i P_\mu,\\ &[\xi_i Q^i, \eta_j Q^j] = -\xi_i\eta_j Z^{ij}.\end{aligned} \tag{3.4}$$

In the differential representation the Q's have the form

$$\begin{aligned}Q_{a,i} &= \frac{\partial}{\partial\theta^{a,i}} - i\sigma^\mu{}_{a\dot a}\bar{\theta}_i^{\dot a}\partial_\mu - \frac{1}{2}\theta_{aj}Z^{ji},\\ \bar{Q}_{\dot a}{}^i &= +\frac{\partial}{\partial\bar{\theta}_i^{\dot a}} - i\theta^{a,i}\sigma^\mu{}_{a\dot a}\partial_\mu - \frac{1}{2}\bar{\theta}_{\dot a,j}Z^{*ji},\end{aligned} \tag{3.5}$$

where $Z^{ij} = -Z^{ji}$ constitute the set of complex central charges of the algebra (3.4).

The corresponding covariant derivatives are[10]

$$D_{a,i} = \frac{\partial}{\partial\theta^{a,i}} + i\sigma^{\mu}_{a\dot{a}}\bar{\theta}^{\dot{a}}\partial_{\mu} + \frac{1}{2}\theta_{a,j}Z^{j}_{i},$$
$$\bar{D}_{\dot{a}}{}^{i} = -\frac{\partial}{\partial\bar{\theta}^{\dot{a}}_{i}} - i\theta^{a,i}\sigma^{\mu}_{a\dot{a}}\partial_{\mu} + \frac{1}{2}\bar{\theta}_{\dot{a}}{}^{j}Z^{*i}_{j}. \tag{3.6}$$

They satisfy the following algebra:

$$\{D_{a,i},\bar{D}_{\dot{a}}{}^{j}\} = -2i\sigma^{\mu}_{a\dot{a}}\delta_{i}{}^{j}\partial_{\mu},$$
$$\{D_{a,i},D_{b,j}\} = -\epsilon_{ab}Z_{ij}. \tag{3.7}$$

The N Volkov–Akulov spinors $\lambda^{\ i}_{a(x)}$ and $\bar{\lambda}^{\ i}_{\dot{a}(x)}$ corresponding to the breaking of the N fermionic generators $Q_a{}^i$ and $\bar{Q}_{\dot{a}}{}^i$ are written as SU(N) isovectors [see (3.1)]:

$$\psi_a \equiv (\lambda^{\ i}_{a}) = \begin{pmatrix} \lambda^{\ 1}_{a} \\ \vdots \\ \lambda^{\ N}_{a} \end{pmatrix}, \quad \bar{\psi}^{\dot{a}} \equiv (\bar{\lambda}^{\ \dot{a}}_{i}) = \begin{pmatrix} \bar{\lambda}^{\ \dot{a}}_{1} \\ \vdots \\ \bar{\lambda}^{\ \dot{a}}_{N} \end{pmatrix}. \tag{3.8a}$$

Their Hermitian conjugates are

$$\psi^{a} \equiv (\lambda^{\ a}_{\ i}) = (\lambda^{\ a}_{\ 1}\cdots\lambda^{\ a}_{\ N}), \quad \bar{\psi}_{\dot{a}} \equiv (\bar{\lambda}_{\dot{a}}{}^{i}) = (\bar{\lambda}_{\dot{a}}{}^{1}\cdots\bar{\lambda}_{\dot{a}}{}^{N}). \tag{3.8b}$$

From (3.8a) and (3.8b) we can form an SU(N) and Weyl scalar

$$\psi^{a}\psi_{a} + \bar{\psi}_{\dot{a}}\bar{\psi}^{\dot{a}} = \sum_{i=1}^{N}(\lambda^{\ a}_{\ i}\lambda^{\ i}_{\ a} + \bar{\lambda}_{\dot{a}}{}^{i}\bar{\lambda}^{\ \dot{a}}_{\ i}). \tag{3.9}$$

Now let us focus on the N-extended superfields. They are given, for the Volkov–Akulov and standard superfields, respectively, by

$$\text{(i)}\quad \Psi_{a}(x,\Theta,\bar{\Theta}) \equiv (\Lambda^{i}_{\ a}(x,\Theta,\bar{\Theta})) = e^{\delta_{\Theta}}\psi_{a}(x),$$
$$\bar{\Psi}_{\dot{a}}(x,\Theta,\bar{\Theta}) \equiv (\bar{\Lambda}^{\dot{a}}_{\ i}(x,\Theta,\bar{\Theta})) = e^{\delta_{\Theta}}\bar{\psi}^{\dot{a}}(x); \tag{3.10}$$

$$\text{(ii)}\quad \Phi(x,\Theta,\bar{\Theta}) = e^{\delta_{\Theta}}\phi(x); \tag{3.11}$$

$e^{\delta_{\Theta}}$, where $\delta_{\theta} = \Sigma_{i=1}^{N}\delta_{\theta_i}$ given by (3.3), cannot be reduced to a simple product of $(\exp\delta_{\theta_i})$, unless the central charges are set to zero, which implies that δ_{θ_i} and δ_{θ_j} commute. As an example, let us write down $N=2$ Volkov–Akulov fields:

$$\Psi_{a}(x,\Theta,\bar{\Theta}) = \tfrac{1}{2}\{\exp[-\tfrac{1}{2}\theta_1\theta_2 Z^{12} + \text{H.c.}] \times e^{\delta_{\theta_1}}e^{\delta_{\theta_2}} + \exp[+\tfrac{1}{2}\theta_1\theta_2 Z^{12} + \text{H.c.}] \times e^{\delta_{\theta_2}}e^{\delta_{\theta_1}}\}\psi_{a}(x), \tag{3.10'}$$

where

$$\delta_{\theta_1} = \theta^{a}_{1}Q^{1}_{a} + \bar{\theta}^{1}_{\dot{a}}\bar{Q}^{\dot{a}}_{1},$$
$$\delta_{\theta_2} = \theta^{a}_{2} + Q^{2}_{a} + \bar{\theta}^{2}_{\dot{a}}\bar{Q}^{\dot{a}}_{2}, \tag{3.12}$$
$$[\delta_{\theta_1},\delta_{\theta_2}] = \theta_1\theta_2 Z^{12}.$$

If we switch off the central charges ($Z^{12}=0$), we get

$$\Psi_{a}(x,\Theta,\bar{\Theta}) = \tfrac{1}{2}\{e^{\delta_{\theta_1}},e^{\delta_{\theta_2}}\}\psi_{a}(x) = e^{\delta_{\theta_1}}e^{\delta_{\theta_2}}\psi_{a}(x). \tag{3.10''}$$

As in the case $N=1$, the Volkov–Akulov superfields $\Psi_a(x,\Theta,\bar{\Theta})$ and the standard superfield $\Phi(x,\Theta,\bar{\Theta})$ can also be defined as unique solutions of a set of generalized constraints. Using the form of the covariant derivatives (3.6) we have established these constraints for the case of arbitrary N:

$$\begin{aligned} D^{i}_{a}\Lambda^{j}_{b} &= \epsilon_{ab}\delta^{ij} + i\sigma^{\mu}_{aa}\bar{\Lambda}^{\dot{a}i}\partial_{\mu}\Lambda^{j}_{b} + \tfrac{1}{2}\Lambda^{k}_{a}Z^{i}_{\ k}\Lambda^{j}_{b}, \\ \bar{D}^{i}_{\dot{a}}\Lambda^{j}_{b} &= -i\Lambda^{a,i}\sigma^{\mu}_{a\dot{a}}\partial_{\mu}\Lambda^{j}_{b} + \tfrac{1}{2}\bar{\Lambda}^{k}_{\dot{a}}z^{*}_{k}{}^{i}\Lambda^{j}_{b}, \\ \bar{D}^{i}_{\dot{a}}\bar{\Lambda}^{j}_{b} &= \epsilon_{\dot{a}\dot{b}}\delta^{ij} - i\Lambda^{a,i}\sigma^{\mu}_{a\dot{a}}\partial_{\mu}\bar{\Lambda}^{j}_{b} + \tfrac{1}{2}\bar{\Lambda}_{\dot{a}}{}^{k}Z^{*}_{k}{}^{i}\bar{\Lambda}^{j}_{b}; \end{aligned} \tag{3.13}$$

$$\begin{aligned} D^{i}_{a}\bar{\Lambda}^{j}_{b} &= +i\sigma^{\mu}_{a\dot{a}}\bar{\Lambda}^{\dot{a},i}\partial_{\mu}\Lambda^{j}_{b} + \tfrac{1}{2}\Lambda^{k}_{a}Z^{i}_{\ k}\bar{\Lambda}^{j}_{b}, \\ D^{i}_{a}\Phi &= i\sigma^{\mu}_{a\dot{a}}\bar{\Lambda}^{ai}\partial_{\mu}\phi + \tfrac{1}{2}\Lambda^{k}_{a}Z^{i}_{\ k}\Phi, \\ \bar{D}^{i}_{\dot{a}}\Phi &= -i\Lambda^{a,i}\sigma^{\mu}_{a\dot{a}}\partial_{\mu}\phi + \tfrac{1}{2}\bar{\Lambda}_{\dot{a}}{}^{k}Z^{*}_{k}{}^{i}\Phi. \end{aligned} \tag{3.14}$$

If we set $Z^{ij}=0$, we note that (a) the standard superfield constraints (3.14) become similar to those of the $N=1$ case (2.8), and (b) the diagonal terms ($i=j$) are also similar to the $N=1$ case. However, the remaining terms ($i\neq j$) can be seen as the standard constraints (3.14) and (2.8).

Now we go back to the study of Wess's $N=1$ supersymmetric Lagrangian to an arbitrary N.

IV. N-EFFECTIVE THEORY

For this purpose we start giving a generalized expresssion of an effective $N=1$ supersymmetric Lagrangian whose expansion in terms of the Volkov–Akulov field contains Wess's Lagrangian (2.9) and the one based on ordinary fields.

We know that the derivative of the standard field does not transform as a standard field:

$$\delta(\partial_{\mu}\phi) = \zeta^{\nu}\partial_{\nu}[\partial_{\mu}\phi] + \partial_{\mu}\zeta^{\nu}\partial_{\nu}\phi.$$

Therefore using the generalized covariant derivative introduced before (Sec. II), one obtains a standard realization

$$\delta(\Delta_{\mu}\phi) = \zeta^{\nu}\partial_{\nu}(\Delta_{\mu}\phi). \tag{4.1}$$

The relation between the two derivatives is

$$\Delta_{\mu}\phi = E_{\mu}{}^{-1\nu}\partial_{\nu}\phi, \tag{4.2}$$
$$E_{\mu}{}^{\nu} = \eta^{\nu}_{\mu} + T^{\nu}_{\mu}, \tag{4.3}$$
$$T_{\mu}{}^{\nu} = -(i/f^2)(\lambda\sigma_{\mu}\partial^{\nu}\bar{\lambda} - \partial_{\mu}\lambda\sigma^{\nu}\bar{\lambda}). \tag{4.4}$$

Generally, one can easily check, using (2.3) and (4.3), that any function depending on ϕ and $\Delta_{\mu}\phi$, say a Lagrangian $L(\phi,\Delta_{\mu}\phi)$, satisfies the standard transformation (2.3):

$$\delta_{\xi}[L(\phi,\Delta_{\mu}\phi)] = \zeta^{\nu}\partial_{\nu}L(\phi,\Delta_{\mu}\phi). \tag{4.5}$$

Therefore

$$[\delta_{\xi},\delta_{\eta}]L(\phi,\Delta_{\mu}\phi) = 2i[\xi\sigma^{\nu}\bar{\eta} - \eta\sigma^{\nu}\bar{\xi}]\partial_{\nu}[L(\phi,\Delta_{\mu}\phi)]. \tag{4.6}$$

Similarly to (2.3), Eq. (4.5) carries a realization of the $N=1$ supersymmetric algebra (4.6). Consequently, as for (2.7) one can usually construct a super-Lagrangian:

$$L(\Phi,\Delta_{\mu}\Phi) = e^{\delta_{\Theta}}L(\phi,\Delta_{\mu}\phi) = L(\phi,\Delta_{\mu}\phi) + \cdots. \tag{4.7}$$

Following (2.9), we write down our effective $N=1$ supersymmetric Lagrangian:

$$L = \int d^2\theta\, d^2\bar{\theta}\,\Lambda^2\bar{\Lambda}^2[-\tfrac{1}{2}f^2 + L(\Phi,\Delta_{\mu}\Phi)] \tag{4.8a}$$

$$= -\tfrac{1}{2}f^2 + (i/2)(\lambda\sigma^{\mu}\partial_{\mu}\bar{\lambda} - \partial_{\mu}\lambda\sigma^{\mu}\bar{\lambda}) + L(\phi,\Delta_{\mu}\phi) + \mathcal{O}(\lambda^2\bar{\lambda}^2, f^{-2}), \tag{4.8b}$$

where the weight $\Lambda^2\bar{\Lambda}^2$ is still the one of Wess.

Making use of the inverse of E^{ν}_{μ},

$$E^{-1\nu}_{\mu} = \eta^{\nu}_{\mu} - T^{\nu}_{\mu} + \sigma(T^2), \tag{4.9}$$

we easily see that L contains as a first term the original Lagrangian.

Furthermore, one can easily verify that the integration over θ's of the Wess weight contains, up to $\mathcal{O}(\lambda^4)$, the module of the "vierbein" $|E| = \det(E_\mu^\nu)$ introduced in Ref. 7, i.e.,

$$\int d^2\theta\, d^2\bar{\theta} f^{-4}\Lambda^2\bar{\Lambda}^2 \equiv |E|.$$

This implies that the two approaches based on superfields and on ordinary fields are equivalent up to the fourth order in the goldstino field. It is this form (4.8) that we shall use for the construction of the effective N-extended supersymmetric Lagrangian. The latter has to satisfy the two following requirements: First, it must be reduced to (or at least contain) (4.8) when we set $N=1$. Second, it must be $SU(N)$ symmetric to ensure that the N-global supersymmetries are going to be spontaneously broken simultaneously. This last requirement is somehow a strong constraint since one would like to have a partial supersymmetry breaking in an effective theory. Nevertheless since we are discussing here global supersymmetry, breaking must occur simultaneously and in an $SU(N)$ symmetric way, as is shown by the average of the Hamiltonian:

$$H = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\sum_{a=1}^{2}|Q_a^i|^2\right). \tag{4.10}$$

Now let us deal with the generalization of the Volkov–Akulov and standard realizations. In this case where $N \geqslant 2$ the extension of relations (2.2) and (2.3) need more elaboration. This is due to the presence of central charges that affect these realizations. Their extension requires additional fields associated with the central charges.[7] We shall restrict ourselves below to the case where the central charges are set to zero. Therefore the realizations of the N algebra (3.4) without Z are given for the Volkov–Akulov and standard fields, respectively, by

$$\text{(i)}\quad \delta_\xi \psi_a = f\xi_a - (i/f)[\xi_0{}^\mu\bar{\psi} - \psi\sigma^\mu\bar{\xi}]\partial_\mu\psi_a, \tag{4.11a}$$

$$\delta_\xi \bar{\psi}_{\dot{a}} = f\bar{\xi}_{\dot{a}} - (i/f)[\xi\sigma^\mu\bar{\psi} - \psi\sigma^\mu\bar{\xi}]\partial_\mu\bar{\psi}_{\dot{a}}; \tag{4.11b}$$

$$\text{(ii)}\quad \delta_\xi\phi(x) = -(i/f)[\xi\sigma^\mu\bar{\psi} - \psi\sigma^\mu\bar{\xi}]\partial_\mu\phi(x); \tag{4.12}$$

and

$$[\delta_\xi, \delta_\eta]f(x) = 2i[\xi\sigma^\mu\bar{\eta} - \eta\sigma^\mu\bar{\xi}]\partial_\mu f(x),$$

where $f(x) \equiv \psi_a(x), \bar{\psi}_{\dot{a}}(x)$, and $\phi(x)$ and where the $SU(N)$ indices are understood. Therefore the construction of the N extended super-Lagrangian can be directly obtained by substituting in the relations (4.4)–(4.7), $\lambda(x)$, ξ, and δ_θ, respectively, by $\psi(x)$ [(3.8)], ξ [(3.2)], and δ_Θ [(3.3)]:

$$L^N(\Phi, \Delta_\mu\Phi) = e^{\delta_\Theta}L(\phi, \Delta_\mu\phi). \tag{4.13}$$

Let us note at this level that working in the absence of central charges, the symmetry $SU(N)$ connecting the N generators of the N supersymmetries is reduced to the coset group $SU(N)/Z_N$. Setting $Z=0$ leads to subtracting the $SU(N)$ group center. Furthermore, we note that (4.13) is invariant under the $SU(N)/Z_N$ transformations.

In order to achieve the construction of the N-extended effective Lagrangian, we have to build the N-extended weight, which gives, after integration on the Θ variables, the N-generalized Volkov–Akulov Lagrangian. The weight, which must be a $SU(N)/Z_N$ scalar, must reproduce the kinetic Volkov–Akulov terms, and must contain the field independent constant, is given by

$$\Gamma^2(x,\Theta,\bar{\Theta})\bar{\Gamma}^2(x,\Theta,\bar{\Theta}) = [1/(N!)^4][(\Psi^a\Psi_a)(\bar{\Psi}_{\dot{a}}\bar{\Psi}^{\dot{a}})]^N, \tag{4.14}$$

where Ψ_a and $\bar{\Psi}_{\dot{a}}$ are given by (3.10):

$$\Gamma^2(x,\Theta,\bar{\Theta}) = \frac{1}{(N!)^2}[\Psi^a\Psi_a]^N = \frac{1}{N!}\prod_{i=1}^{N}[\Lambda_i^a(x,\Theta,\bar{\Theta})\Lambda_a^i(x,\Theta,\bar{\Theta})].$$

Since all the powers of the N-extended Weyl spinors Λ_a^i $(x,\Theta,\bar{\Theta})$ greater than 2 for fixed i vanish, we have

$$\Gamma^2(x,\Theta,\bar{\Theta})\bar{\Gamma}^2(x,\Theta,\bar{\Theta}) = \frac{1}{(N!)^2}\prod_{i=1}^{N}(\Lambda_i^a\Lambda_a^i)(\bar{\Lambda}_{\dot{a}}^i\bar{\Lambda}_i^{\dot{a}}). \tag{4.15}$$

Therefore, the N-extended Volkov–Akulov Lagrangian is given by

$$L_{V-A}^N(\lambda_a^i, \lambda_a^i, f) = -\tfrac{1}{2}f^2\int d^2\Theta\, d^2\bar{\Theta} f^{-4N}\Gamma^2(x,\Theta,\bar{\Theta}) \cdot\bar{\Gamma}^2(x,\Theta,\bar{\Theta}), \tag{4.16}$$

where

$$d^2\Theta = \prod_{i=1}^{N}[d\theta_i^a \cdot d\theta_a^i].$$

To be more explicit, let us discuss the $N=2$ case. The weight $\Gamma^2\bar{\Gamma}^2$ becomes

$$\Gamma^2\bar{\Gamma}^2 = [1/(2!)^2][(\Lambda_1^a\Lambda_a^1)(\bar{\Lambda}_{\dot{a}}^1\bar{\Lambda}_1^{\dot{a}})][(\Lambda_2^a\Lambda_a^2)(\bar{\Lambda}_{\dot{a}}^2\bar{\Lambda}_2^{\dot{a}})] \equiv [1/(2!)^2](\Lambda_1^2\bar{\Lambda}_1^2)(\Lambda_2^2\bar{\Lambda}_2^2). \tag{4.17}$$

The $N=2$ extended Volkov–Akulov Lagrangian up to the second order of the λ's is obtained by carrying out the integration with respect to θ_1 and θ_2. We note that the order of the integration over the θ's is arbitrary as a consequence of the existing symmetry $SU(2)/Z_2$. We have then

$$\begin{aligned}
L_{V-A}^{N=2} &= -\frac{1}{2}f^2\int d^2\Theta\, d^2\bar{\Theta} f^{-8}\Gamma^2\bar{\Gamma}^2 \\
&= -\frac{1}{(2)^3}f^2\int d^2\theta_2\, d^2\bar{\theta}_2 f^{-8}\left[\int d^2\theta_1\, d^2\bar{\theta}_1(\Lambda_1^2\bar{\Lambda}_1^2)(\Lambda_2^2\bar{\Lambda}_2^2)\right] \\
&\quad -\frac{1}{(2)^3}f^2\int d^2\theta_2\, d^2\bar{\theta}_2 f^{-8}\left[\left(\int d^2\theta_1\, d^2\bar{\theta}_1\,\Lambda_1^2\bar{\Lambda}_1^2\right)(\tilde{\Lambda}_2^2\tilde{\bar{\Lambda}}_2^2) + \left(\int d^2\theta_1\, d^2\bar{\theta}_1\,\Lambda_2^2\bar{\Lambda}_2^2\right)(\tilde{\Lambda}_1^2\tilde{\bar{\Lambda}}_1^2)\right. \\
&\quad \left. + \left(\int d^2\theta_1\,\Lambda_1^2\right)(d^2\bar{\theta}_1\,\bar{\Lambda}_2^2)\tilde{\bar{\Lambda}}_1^2\tilde{\bar{\Lambda}}_2^2 + \left(\int d^2\theta_1\,\Lambda_2^2\right)\left(\int d^2\bar{\theta}_1\,\bar{\Lambda}_1^2\right)\tilde{\bar{\Lambda}}_2^2\tilde{\bar{\Lambda}}_1^2\right] + \cdots,
\end{aligned} \tag{4.18}$$

Where the ellipses means that we have omitted all the fourth powers of the goldstino fields and terms with higher derivatives for spinors. The tilde ($\sim$) means that the θ_1 variable is set to zero and therefore we are dealing with $(N-1)=(2-1)=1$ superfields. Using the θ_1 expansion of the $N=2$ extended Volkov–Akulov superfields, one can establish

$$
\begin{aligned}
L_{V-A}^{N=2} = & -\frac{1}{(2)^3} f^2 \int d^2\theta_2\, d^2\bar{\theta}_2 f^{-4} [\{1-(i/f^2)(\tilde{\Lambda}_1\sigma^\mu\partial_\mu\tilde{\bar{\Lambda}}^1-\partial_\mu\tilde{\Lambda}_1\sigma^\mu\tilde{\bar{\Lambda}}^1)+\sigma(f^{-4},\tilde{\Lambda}_1^2\tilde{\bar{\Lambda}}_1^2)\}\tilde{\Lambda}_2^2\tilde{\bar{\Lambda}}_2^2 \\
& +\{1-(i/f^2)(\tilde{\Lambda}_2\sigma^\mu\partial_\mu\tilde{\bar{\Lambda}}^2-\partial_\mu\tilde{\Lambda}_2\sigma^\mu\tilde{\bar{\Lambda}}^2)+\sigma(f^{-4},\tilde{\Lambda}_2^2\tilde{\bar{\Lambda}}_2^2)\}\tilde{\Lambda}_1^2\tilde{\bar{\Lambda}}_1^2 \\
& +\{1+(i/f^2)\partial_\mu\tilde{\Lambda}_1\sigma^\mu\tilde{\bar{\Lambda}}^1+\sigma(f^{-4},\tilde{\Lambda}_1^2\tilde{\bar{\Lambda}}_1^2)\}\{1-(i/f^2)\tilde{\Lambda}_2\sigma^\mu\partial_\mu\tilde{\bar{\Lambda}}^2+\sigma(f^{-4},\tilde{\Lambda}_2^2\tilde{\bar{\Lambda}}_2^2)\}\tilde{\bar{\Lambda}}_1^2\tilde{\Lambda}_2^2 \\
& +\{1-(i/f^2)\tilde{\Lambda}_1\sigma^\mu\partial_\mu\tilde{\bar{\Lambda}}^1+\sigma(f^{-4},\tilde{\Lambda}_1^2\tilde{\bar{\Lambda}}_1^2)\}\{1+(i/f^2)\partial_\mu\tilde{\Lambda}_2\sigma^\mu\tilde{\bar{\Lambda}}^2+\sigma(f^{-4}\tilde{\Lambda}_2^2\tilde{\bar{\Lambda}}_2^2)\}\tilde{\bar{\Lambda}}_2^2\tilde{\Lambda}_1^2].
\end{aligned}
$$

Now the integration over θ_2 can be carried out exactly in the same manner and we get

$$
\begin{aligned}
L_{V-A}^{N=2} = & -[1/(2)^3 f^2[2\{1-(i/f^2)(\lambda_1\sigma^\mu\partial_\mu\bar{\lambda}^1-\partial_\mu\lambda_1\sigma^\mu\bar{\lambda}^1)+\mathcal{O}(f^{-4},\lambda_1^2\bar{\lambda}_1^2)\} \\
& \times\{1-(i/f^2)(\lambda_2\sigma^\mu\partial_\mu\bar{\lambda}^2-\partial_\mu\lambda_2\sigma^\mu\bar{\lambda}^2)+\mathcal{O}(f^{-4},\lambda_2^2\bar{\lambda}_2^2)\} \\
& +2\{(1+(i/f^2)\partial_\mu\lambda_1\sigma^\mu\bar{\lambda}^1+\sigma(f^{-4},\lambda_1^2\bar{\lambda}_1^2))(1-(i/f^2)\lambda_2\sigma^\mu\partial_\mu\bar{\lambda}^2+\mathcal{O}(f^{-4},\lambda_2^2\bar{\lambda}_2^2)) \\
& \times(1-(i/f^2)\lambda_1\sigma^\mu\partial_\mu\bar{\lambda}^1+\sigma(f^{-4},\lambda_1^2\bar{\lambda}_1^2))(1+(i/f^2)\partial_\mu\lambda_2\sigma^\mu\bar{\lambda}^2+\mathcal{O}(f^{-4},\lambda_2^2\bar{\lambda}_2^2))\}] \\
= & -\tfrac{1}{2}f^2[1/(2)^2][(2)^2\{1-(i/f^2)[(\lambda_1\sigma^\mu\partial_\mu\bar{\lambda}^1-\partial_\mu\lambda_1\sigma^\mu\bar{\lambda}^1)+(\lambda_2\sigma^\mu\partial_\mu\bar{\lambda}^2-\partial_\mu\lambda_2\sigma^\mu\bar{\lambda}^2)] \\
& +\mathcal{O}(f^{-4},\lambda_1^2\bar{\lambda}_1^2,\lambda_2^2\bar{\lambda}_2^2,\lambda_1^2\bar{\lambda}_2^2,\bar{\lambda}_1^2\lambda_2^2)\}] \\
= & -\tfrac{1}{2}f^2+\frac{i}{2}\sum_{i=1}^{2}[(\lambda_i\sigma^\mu\partial_\mu\bar{\lambda}^i-\partial_\mu\lambda_i\sigma^\mu\bar{\lambda}^i)]+\mathcal{O}(f^{-2},\lambda_i^2\bar{\lambda}_j^2). \qquad (4.19)
\end{aligned}
$$

Through this $N=2$ example and after integration over $d^2\Theta$ and $d^2\bar{\Theta}$ in (4.16) the N extended Volkov–Akulov Lagrangian expression in the component fields contains as for the $N=1$ case and up to $\mathcal{O}(\lambda)^4$ the constant term and the kinetic term of the goldstinos:

$$
\begin{aligned}
& L_{V-A}^{N}(\lambda_a^i,\bar{\lambda}_{\dot{a}}^i,f) \\
& = -\tfrac{1}{2}f^2+\frac{i}{2}\sum_{i=1}^{N}[(\lambda_i\sigma^\mu\partial_\mu\bar{\lambda}^i-\partial_\mu\lambda_i\sigma^\mu\bar{\lambda}^i)] \\
& \quad +\mathcal{O}(f^{-2},\lambda_i\lambda^i\bar{\lambda}^j\bar{\lambda}_j). \qquad (4.20)
\end{aligned}
$$

Before giving the full Lagrangian, we note that our construction differs from the one in Ref. 4(b) in which Bagger and Wess introduced the breaking of supersymmetry from a given $N(=2)$ to $N-1(=1)$. As a consequence they obtained an $(N-1)$ extended Volkov–Akulov superfield. In our case since we are interested in the breaking of the N-extended supersymmetry directly and simultaneously down to $N=0$ (Poincaré subgroup), which is imposed by the N-extended global supersymmetry (4.10), the weight (4.14) is the most general one satisfying all the necessary requirements mentioned before: it must be $SU(N)/Z_N$ symmetric, reproduce the kinetic Volkov–Akulov terms, and contain the crucial constant term. This presents a double interest: first, it exhibits the spontaneously N-supersymmetry breaking and, second, allows one to obtain the original Lagrangian. Furthermore, we note that (4.14) reduces when we set $N=1$ to Wess's weight [(2.6) and (2.10)].

By analogy with (4.8) and using (4.14), the full Lagrangian of an N-extended supersymmetric theory can be written

$$
\begin{aligned}
L^N = & \int d^2\Theta\, d^2\bar{\Theta} f^{-4N}\Gamma^2\bar{\Gamma}^2\left[-\frac{1}{2}f^2+L^N(\Phi,\Delta_\mu\Phi)\right] \qquad (4.21) \\
= & -\frac{1}{2}f^2+\frac{i}{2}\sum_{i=1}^{N}[\lambda_i\sigma^\mu\partial_\mu\bar{\lambda}^i-\partial_\mu\lambda_i\sigma^\mu\bar{\lambda}^i] \\
& +L(\phi,\Delta_\mu\phi)+\mathcal{O}(\lambda_i\lambda^i\bar{\lambda}^j,f^{-2}). \qquad (4.22)
\end{aligned}
$$

It describes the phenomenology of low energy physics where supersymmetry is necessarily broken down to the Poincaré subgroup; the constant $-\frac{1}{2}f^2$ is a signal of this supersymmetry breaking. The magnitude of the scale f is model dependent,[11,12] but may be thought of order of the M_w mass, if we believe that supersymmetry is the proper tool to avoid the hierarchy problem. The second term on the right-hand side of (4.22) is nothing but the kinetic terms of the N Volkov–Akulov fields occurring in the breaking, and finally the term $\mathcal{O}(\lambda_i\lambda^i\bar{\lambda}^j\bar{\lambda}_j,f^{-4})$, which is highly nonrenormalizable, carries all the supersymmetric effects. However, processes involving such terms are negligible due to the magnitude of the mass scale f.

V. CONCLUSIONS

In this paper we have studied the generalization of the effective Wess Lagrangian to an arbitrary N. We have constructed the N-extended Volkov–Akulov and standard superfields and also the generalized Wess constraints in the presence of the central charges. We have pointed out that a standard transformation can be built for the Lagrangian itself, if we substitute the normal derivative ∂_μ by a "covariant" one Δ_μ (4.2). For $N=1$, we have proposed an $N=1$ effective Lagrangian that contains the Wess one and that of the approach based on ordinary fields. We have also checked

that the two latter approaches are equivalent to the fourth order in the Volkov–Akulov fields.

For the N-extended case, we have restricted ourselves to the limit $Z = 0$. We have imposed two requirements: First it must be reduced to (or at least contain) (4.8) when we set $N = 1$, which is obviously satisfied. Second, we have demanded $SU(N)/Z_N$ symmetry in order to ensure the simultaneous breaking of the N-extended global supersymmetries. We have found that the more general weight consistent with the generalization and satisfying all the requirements is that given by (4.14).

The phenomenological Lagrangian part (matter, gauge, Higgs) turns out to satisfy our requirements also. Therefore, the full N-extended effective theory is given by (4.21). It has all the good features, as we remarked before, as long as we keep the central charges equal to zero. However, it is interesting to study the extension to the central charge case and examine their effects. This requires taking into account the contribution part of the central charges in the realization of the N-extended superalgebra (3.4). This is under study and will be treated elsewhere.

ACKNOWLEDGMENTS

One of the authors (E.H.S) would like to thank J. Wess for many useful discussions and encouragements and J. Strathdee for reading the manuscript and pointing out some remarks.

All of the authors would like also to thank Professor Abdus-Salam, the International Atomic Energy Agency, and UNESCO for their kind hospitality at the International Centre for Theoretical Physics (ICTP) Trieste, Italy where a part of this work has been done.

[1]S. Coleman, J. Wess, and B. Zumino, Phys. Rev. **177**, 2239 (1969); S. Weinberg, Physica A **96**, 327 (1979).
[2]M. Gell-Mann and M. Levy, Nuovo Cimento **16**, 705 (1960).
[3]H. Pagels, Phys. Rep. **16c**, 219 (1975); M. Peskin, SLAC-PUB. 3021, 1983.
[4](a) D. V. Volkov and V. P. Akulov, JETP Lett. **16**, 438 (1972); D. V. Volkov and V. Soroka, *ibid.* **18**, 438 (1973); B. De Witt and D. Freedman, Phys. Rev. Lett. **35**, 827 (1975); (b) J. Bagger and J. Wess, SLAC-PUB 3255, 1983.
[5]J. Wess, "Non linear realization of $N = 1$ supersymmetry," Karlsrube preprint, December 1982; S. Samuel and J. Wess, Nucl. Phys. B **221**, 153 (1983); **226**, 289 (1983); **233**, 488 (1984).
[6]E. Ivanov and A. Kapustnikov, J. Phys. A **11**, 2375 (1978); G **8**, 167 (1982); T. Uematsu and C. Zachos, Nucl. Phys. B **201**, 250 (1982).
[7]S. Ferrara, L. Maiani, and P. West, Z. Phys. C **19** 267 (1983).
[8]J. Wess, "Non linear realization of supersymmetry, Karlsruhe preprint, 1983 (unpublished).
[9]J. Bagger and J. Wess, *Supersymmetry and Supergravity* (Princeton U. P., Princeton, NJ, 1983).
[10]J. Gates, M. Grisaru, M. Rocĕck, and W. Siegel, "Superspace or one thousand and one lessons on supersymmetry," 1983.
[11]S. Dimopoulos and H. Georgi, Nucl. Phys. B **193**, 150 (1981); D. U. Nanopoulos, Phys. Rep. **105**, 71 (1984); E. Witten, Nucl. Phys. B **185**, 513 (1981).
[12]C. Sakai, Z. Phys. C **11**, 153 (1981); J. Ibañez, F.T.V.A.M. preprint 84/7, April 1984; K. Inoue, A. Kakuto, H. Komatsu, and S. Takeshita, Prog. Theor. Phys. **68**, 927 (1982); E. Witten, lectures at the International School on Supersymmetry in Physics, UNAM, Mexico City, December 1981; J. Ellis, "Phenomenology of GUTS," LAPP.Ttl.48-TH 3174, September 1981.

Infinite-dimensional Lie algebras acting on the solution space of various σ models

Michel Jacques[a)]
Institut de Physique Théorique, Université Catholique de Louvain, 2, Chemin du Cyclotron, B-1348 Louvain-la-Neuve, Belgium

Yvan Saint-Aubin
Centre de recherches mathématiques and Département de mathématiques et de statistique, Université de Montréal, Case Postale 6128, succ. A, Montréal, Canada H3C 3J7

(Received 24 February 1987; accepted for publication 3 June 1987)

Infinite-dimensional Lie algebras of infinitesimal transformations acting on the solution space of various two-dimensional σ models are investigated. The main tools are (i) Takasaki's interpretation [Commun. Math. Phys. **94**, 35 (1984)] of the solutions of the associated linear system in terms of points in an infinite-dimensional Grassmann manifold and (ii) Mikhaïlov's reduction procedure [Physica D **3**, 73 (1981)] for linear systems. Takasaki's approach leads, for the σ models with values in a Lie group G, to a set of transformations that has the structure of the loop algebra $\mathfrak{g} \otimes \mathbb{R}[t,t^{-1}]$, where $\mathfrak{g}$ is the Lie algebra of G. (This algebra has already been encountered by Dolan [Phys. Rev. Lett. **47**, 1371 (1981)] and by Wu [Nucl. Phys. B **211**, 160 (1983)] among others.) The σ models with a Wess–Zumino term are also considered; the algebraic structure is found to be the same. Finally, Mikhaïlov's procedure is used to study the σ models with values in a Riemannian symmetric space (RSS) G/H which is not a Lie group. The algebra in these cases is a subalgebra of the loop algebra found for the principal models but it does not seem to be graded. However, it contains two graded infinite-dimensional subalgebras with the following structure: if $\mathfrak{h}$ and $\mathfrak{m}$ are the two eigenspaces of the involution σ defining the RSS G/H, these two graded subalgebras are $\mathfrak{h} \otimes \mathbb{R}[t]$ and $(\oplus_{i\in\mathbb{N}} \mathfrak{h} \otimes t^{2i}) \oplus (\oplus_{i\in\mathbb{N}} \mathfrak{m} \otimes t^{2i+1})$.

I. INTRODUCTION

During the last years, physicists have recognized both the existence and the importance of many new infinite-dimensional Lie algebras in various physical systems. The conformal algebra in two dimensions has been shown to be connected to the Virasoro algebra. Loop algebras and affine (Kac–Moody) algebras have arisen in many other contexts.

The σ models with values in different Riemannian symmetric spaces (RSS) are among the systems where many different infinite-dimensional Lie algebras manifest themselves. The space time symmetry is the conformal algebra; the current algebra closes in a loop algebra and there also exists an infinite-dimensional algebra acting on the solution space of the model. The present paper deals with the latter type of symmetry transformations.

Dolan[1] was the first to investigate the latter algebra for the principal σ model with values in $SU(n)$. She was able, first, to give the explicit action of a set of generators and, second, to identify the structure of the algebra $\hat{\mathfrak{g}}$ spanned by these generators. [This action is nonlocal in the sense that it is defined in terms of integrals of the field $g(\xi,\eta)\in SU(n)$ and its derivatives.] The structure of the algebra $\hat{\mathfrak{g}}$ is $\mathrm{su}(n) \otimes \mathbb{R}[t]$, i.e., the algebra with elements of the form $U \otimes t^m$ with $U\in\mathrm{su}(n)$, $m\in\mathbb{N}$ and commutation relations: $[U \otimes t^m, V \otimes t^n] = [U,V] \otimes t^{m+n}$. It is common to introduce the notation $\hat{\mathfrak{g}}_i \equiv \{U \otimes t^i, U\in\mathrm{su}(n)\}$ to describe the natural gradation: $\hat{\mathfrak{g}} = \oplus_{i\in\mathbb{N}} \hat{\mathfrak{g}}_i$ and $[\hat{\mathfrak{g}}_i, \hat{\mathfrak{g}}_j] \subseteq \hat{\mathfrak{g}}_{i+j}$. Since then, many groups tried to extend this algebra to a larger one. Let us recall that, in an affine algebra, the center manifests itself only in commutation of elements belonging, respectively, to $\hat{\mathfrak{g}}_i$ and $\hat{\mathfrak{g}}_{-i}$. Since Dolan's algebra contains only elements with grading $i\in\mathbb{N}$, hopes were that this algebra could be a subalgebra of such an affine algebra whose elements with $i<0$, $i\in\mathbb{Z}$ were still to be found. (The appeal of affine Lie algebras is due to the fact that both their structure and their representation theory are intimately related to those of finite-dimensional simple Lie algebras.)

Some time after, Wu[2] enlarged the algebra to an $(\mathrm{su}(n) \otimes \mathbb{R}[t,t^{-1}]) \oplus \mathrm{su}(n)$ showing then that the structure was that of a loop algebra with $\mathbb{Z}$ gradation (without the central extension). His work systematically uses generating functions as a tool to define the action of symmetry transformations and to compute their commutation rules. (These generating functions had been introduced earlier by several people.[3]) We shall use them when needed and push their range of applications further.

Parallel to these developments, Ueno and Nakamura[4] (see also Ref. 5) provided the link between Dolan's algebra and the so-called infinitesimal Riemann–Hilbert transformation. Moreover they expressed the action of the generators in a form very similar to a Riccati action.[6,7] (In fact, they reported a larger algebra with structure $\mathrm{su}(n) \otimes \mathbb{R}[t,t^{-1}]$, the gradation being now in $\mathbb{Z}$. They observed however that their generators corresponding to $i<0$, $i\in\mathbb{Z}$ act trivially on the solution space.) In this direction, the next major step was taken by Takasaki[8] for the self-dual Yang–Mills (SDYM)

[a)] Current address: Laboratoire de Physique Nucléaire, Université de Montréal, Case Postale 6128, succ. A, Montréal, Canada H3C 3J7.

 0022-2488/87/102463-17$02.50 © 1987 American Institute of Physics

system. His contribution was to interpret the solution of the linear system associated with SDYM equations as the evolution of a point in an infinite-dimensional Grassmann manifold. With this picture, he was able, first, to integrate formally the equations and, second, to give a finite group action corresponding to the infinitesimal transformations given by Dolan. Even though the analogy is not complete yet, Takasaki's work offers the first clear setting of the SDYM system in a context similar to the one used by the Kyoto school[9,10] to describe other nonlinear systems like the Kadomtsev–Petviashvili equation and the Korteweg–de Vries equation.

The present paper deals with various classical σ models on two-dimensional Minkowski space. More precisely, the models to be considered are the principal σ models (σ models whose fields take their values in a Lie group), the σ models with values in a Riemannian symmetric space [these include, for example, the well-known nonlinear $O(3)$ model whose field lives on the sphere S^2], and σ models with a Wess–Zumino term (to be referred to as WZσ models). The second section provides a definition of these models together with the description of their associated linear system. The goal of this paper is twofold: first, to set these different σ models in the language of infinite-dimensional Grassmannian as Takasaki did for the SDYM system (Sec. III) and, second, to construct an infinite set of symmetry transformations for each of these models. This latter goal is achieved in several steps. In Sec. IV, the starting point is the infinite-dimensional algebra acting naturally on the infinite-dimensional Grassmannian. We show how this action on the solution space of the linear system leads to an (almost) uniquely defined action on the solution space of the Sl(n,$\mathbb{C}$) principal σ model. To characterize the subalgebras acting on the solution space of the other principal models, we take advantage of the reduction procedure introduced first by Mikhaïlov.[11] (See also Ref. 6.) For nonlinear systems, there exist discrete symmetries that allow us to define subsystems by imposing the solutions to be invariant under these symmetries. [A simple example is the discrete symmetry $g(\xi,\eta)\to g^{-1\dagger}(\xi,\eta)$ of the Sl(n,$\mathbb{C}$) principal model. The fixed points of this symmetry are of course the solutions of the SU(n) principal model.] The central idea of Mikhaïlov's reduction procedure is to formulate the content of a constraint (invariance under a discrete symmetry) at the level of the linear system. In Sec. V, the σ models with a Wess–Zumino term are studied. The only difference with the case of principal models lies in the fact that, to identify the algebraic structure spanned by the symmetry transformations, one has to perform a change of basis. (The generating functions *à la* Wu are in this context a very useful tool.) Again the reduction procedure is applied to obtain the subalgebras for the WZσ models with values in $G=$ SU(n), SO(n), and Sp(n). Section VI is devoted to the analysis of the symmetry transformations for σ models with values in a RSS. Again, the construction of these infinitesimal transformations relies heavily upon Mikhaïlov's reduction procedure. Finally, changes of basis lead to the identification of two remarkable subalgebras for each of the models considered.

II. THE MODELS

A. The Lagrangian formulation

We start with a Lagrangian description of the models that we are to consider. Space-time is two-dimensional Minkowski space and will be described by the light-cone coordinates

$$\xi\equiv(t+x)/2,\quad \eta\equiv(t-x)/2. \tag{2.1}$$

Let G be any of the simple Lie groups contained in Cartan's classical series. The fields of the σ models are maps g from Minkowski space into G.

The dynamics of the principal σ model with values in G is specified by the following action:

$$\mathcal{S}=\frac{1}{2}\int d^2x\ \mathrm{tr}\,\partial_\mu g\,\partial^\mu g^{-1}. \tag{2.2}$$

The Euler–Lagrange equation leads to the equation of motion

$$\partial_\eta((\partial_\xi g)g^{-1})+\partial_\xi((\partial_\eta g)g^{-1})=0. \tag{2.3}$$

Defining the right-invariant fields

$$A_R\equiv(\partial_\xi g)g^{-1}, \tag{2.4a}$$

$$B_R\equiv(\partial_\eta g)g^{-1}, \tag{2.4b}$$

the field equation becomes

$$\partial_\eta A_R+\partial_\xi B_R=0. \tag{2.5}$$

Moreover, the new fields satisfy the following identity:

$$\partial_\eta A_R-\partial_\xi B_R+[A_R,B_R]=0 \tag{2.6}$$

ensuring that there exists a g related to A_R and B_R by relations (2.4).

Note that Eq. (2.3) is conformally invariant. Let $g(\xi,\eta)$ be a solution (2.3) and define

$$h(\xi,\eta)\equiv g(\alpha(\xi),\beta(\eta))$$

for any strictly monotonous analytic functions α and β. Then $h(\xi,\eta)$ is still a solution of (2.3). Parity (interchange of ξ and η) and time reversal [transformation of (ξ,η) into $(-\eta,-\xi)$] are discrete external symmetry transformations. Equation (2.3) is obviously invariant under both operations. Moreover, we have the additional symmetries

$$g(\xi,\eta)\to\bar{g}(\xi,\eta), \tag{2.7a}$$

$$g(\xi,\eta)\to g^{-1}(\xi,\eta), \tag{2.7b}$$

$$g(\xi,\eta)\to g^{T}(\xi,\eta). \tag{2.7c}$$

Indeed, for (2.7b),

$$\partial_\eta((\partial_\xi g^{-1})g)+\partial_\xi((\partial_\eta g^{-1})g)$$
$$=-g^{-1}\{\partial_\eta((\partial_\xi g)g^{-1})+\partial_\xi((\partial_\eta g)g^{-1})\}g=0$$

and the same argument applies for (2.7c). We shall make extensive use of (2.7b) in the sequel.

The nonprincipal σ models can be obtained from the principal models by a reduction procedure that will be described below. So, we will not dwell any further on their Lagrangian formulation.

Let us consider the G-valued σ model with a Wess–Zumino term. The action is given by[12]

$$\mathcal{S}_{\mathrm{WZ}} = \frac{1}{4\lambda^2}\int d^2x\ \mathrm{tr}\,\partial_\mu g\,\partial^\mu g^{-1} + \frac{n}{24\pi}\int_B d^3y\,\epsilon^{ijk}\,\mathrm{tr}\,\hat{g}^{-1}\frac{\partial\hat{g}}{\partial y^i}\hat{g}^{-1}\frac{\partial\hat{g}}{\partial y^j}\hat{g}^{-1}\frac{\partial\hat{g}}{\partial y^k}, \tag{2.8}$$

where $\hat{g}$ is any extension of g to B, a solid ball whose boundary is the two-sphere, λ is the coupling constant, and n is an integer, so that the second term is well-defined mod 2π. [Strictly speaking, the second term has to be defined on Euclidean space taken to be S^2. We shall use the field equations obtained from (2.8) on Euclidean space after setting them back on Minkowski space.] The Euler–Lagrange equation gives

$$\partial_\mu(g^{-1}\partial^\mu g) - (n\lambda^2/4\pi)\epsilon^{\mu\nu}\partial_\mu(g^{-1}\partial_\nu g) = 0. \tag{2.9}$$

Setting $\kappa \equiv n\lambda^2/4\pi$, (2.9) is equivalent to

$$(1-\kappa)\partial_\eta((\partial_\xi g)g^{-1}) + (1+\kappa)\partial_\xi((\partial_\eta g)g^{-1}) = 0. \tag{2.10}$$

With the definitions (2.4) of A_R and B_R, this means

$$(1-\kappa)\partial_\eta A_R + (1+\kappa)\partial_\xi B_R = 0. \tag{2.11}$$

Equation (2.10) is still conformally invariant, but it is no longer invariant under parity and time reversal. Similarly, among the symmetries (2.7), only (2.7a) survives. However, the model remains invariant under special combinations of all the previous symmetries. The generators of the finite group of discrete symmetries (both internal and external) are

$$g(\xi,\eta) \to \bar{g}(\xi,\eta), \tag{2.12a}$$

$$g(\xi,\eta) \to g^{-1}(\eta,\xi), \tag{2.12b}$$

$$g(\xi,\eta) \to g^{T-1}(\xi,\eta), \tag{2.12c}$$

$$g(\xi,\eta) \to g(-\xi,-\eta). \tag{2.12d}$$

Again, the discrete symmetry (2.12b) [$g(\xi,\eta) \to h(\xi,\eta) \equiv g^{-1}(\eta,\xi)$] will play a central role in what follows.

B. The reduction procedure

As emphasized by Mikhaïlov,[11] some integrable models can be viewed as subsystems of more general integrable models. The key point is to impose reduction constraints on the general model.

For example, in the case of σ models, it is sufficient to start from the $\mathrm{Sl}(n,\mathbb{C})$ principal model.[6] Let σ be an involution automorphism of $\mathrm{Sl}(n,\mathbb{C})$, which can be taken among the following ones:

$$\sigma_1(g) = IgI^{-1}, \quad \sigma_2(g) = I\bar{g}I^{-1}, \quad \sigma_3(g) = Ig^{T-1}I^{-1}, \quad \sigma_4(g) = Ig^{\dagger-1}I^{-1}, \tag{2.13}$$

where I may be chosen, up to conjugation, as[13]

$$\mathbf{1}_{p,q} = \begin{pmatrix} \mathbf{1}_p & 0 \\ 0 & -\mathbf{1}_q \end{pmatrix}, \quad \mathbf{J}_{2n} = \begin{pmatrix} 0 & \mathbf{1}_n \\ -\mathbf{1}_n & 0 \end{pmatrix},$$

$$\mathbf{K}_{p,q} = \begin{pmatrix} \mathbf{1}_p & & & \\ & -\mathbf{1}_q & & \\ & & \mathbf{1}_p & \\ & & & -\mathbf{1}_q \end{pmatrix}.$$

Then it is always possible to impose that g be in one of the classical groups by a condition of the type

$$\sigma(g) = g. \tag{2.14}$$

Turning now to the nonprincipal σ models, we consider models with values in a Riemannian symmetric space. (One of the reasons for this choice being the so-called dual symmetry.[14]) Let G be one of the classical groups and σ an automorphism of G of the type (2.13). Let H be a subgroup of G such that

$$(G_\sigma)_0 \subseteq H \subseteq G_\sigma,$$

where G_σ is the subgroup of fixed points of σ and $(G_\sigma)_0$ its identity component. Denote by $\mathfrak{g}$ and $\mathfrak{h}$ the Lie algebras of G and H, respectively. Then we have the canonical decomposition

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m} \tag{2.15}$$

with the relations

$$[\mathfrak{h},\mathfrak{h}] \subseteq \mathfrak{h}, \quad [\mathfrak{h},\mathfrak{m}] \subseteq \mathfrak{m}, \quad [\mathfrak{m},\mathfrak{m}] \subseteq \mathfrak{h}. \tag{2.16}$$

The RSS G/H can be embedded in its isometry group G through the Cartan immersion i: $G/H \to G$ defined by

$$i(gH) = \sigma(g)g^{-1}. \tag{2.17}$$

The points in the image $i(G/H)$ have the property to be such that $\sigma(g)g = \mathbf{1}$. Let Σ_0 be the submanifold of $\Sigma = \{g \in G \mid \sigma(g)g = \mathbf{1}\}$ which contains the identity in G. Then every solution $g(\xi,\eta)$ of the G-principal model whose values for all (ξ,η) lie on Σ_0 gives rise to a solution of the G/H model by[14]

$$q(\xi,\eta) = i^{-1} \circ g(\xi,\eta). \tag{2.18}$$

Of course, the condition

$$\sigma(g) = g^{-1} \tag{2.19}$$

does not ensure that g lies on Σ_0. (See Ref. 15.) However, we are interested in defining a transformation law between solutions and we are going to treat the infinitesimal form of that law. Starting with a solution g in Σ_0, we build a new solution g' which is infinitesimally close to g. Since Σ_0 has an empty intersection with the other submanifolds of Σ, it is enough to require that g' is on Σ, i.e., (2.19).

Thus, in the sequel, we shall restrict ourselves to the case $G = \mathrm{Sl}(n,\mathbb{C})$ and impose reduction constraints of the type (2.14) or (2.19) to proceed to σ models with values either in a compact Lie group (principal models) or in a RSS which is not a Lie group. The complete list of irreducible RSS's to be considered here (essentially all the irreducible RSS's whose isometry group is one of the classical simple Lie groups) is to be found in Table I of Ref. 6 together with the algebraic constraint(s) necessary to construct them from the $\mathrm{Sl}(n,\mathbb{C})$ principal model.

C. The linear systems

The models defined above possess the important property of being integrable in the sense that they enjoy a Lax formulation. In this section we specify the linear systems whose integrability conditions are the nonlinear equations under study.

Let λ be a complex (spectral) parameter. We first consider the Sl(n,C) principal σ model. Let $R = R(\xi,\eta;\lambda)$ be an Sl(n,C)-valued function satisfying[16-18]

$$\partial_\xi R = (1+\lambda)^{-1} A_R R, \tag{2.20a}$$

$$\partial_\eta R = (1-\lambda)^{-1} B_R R. \tag{2.20b}$$

The integrability condition in the sense of Frobenius of the linear system (2.20) has to be fulfilled identically in λ and yields precisely (2.5) and (2.6). Taking into account the symmetry (2.7b) we also consider the same system for g^{-1}:

$$A_L \equiv (\partial_\xi g^{-1}) g = -g^{-1} A_R g, \tag{2.21a}$$

$$B_L \equiv (\partial_\eta g^{-1}) g = -g^{-1} B_R g, \tag{2.21b}$$

$$\partial_\xi L = (1+\lambda)^{-1} A_L L, \tag{2.22a}$$

$$\partial_\eta L = (1-\lambda)^{-1} B_L L, \tag{2.22b}$$

for some Sl(n,C)-valued function $L = L(\xi,\eta;\lambda)$. Note that the systems (2.20) and (2.22) do not uniquely fix the solutions R and L, respectively. Starting with solutions R and L, we can build new solutions R' and L' by

$$R'(\xi,\eta;\lambda) = R(\xi,\eta;\lambda) C_1(\lambda),$$

$$L'(\xi,\eta;\lambda) = L(\xi,\eta;\lambda) C_2(\lambda),$$

for arbitrary functions $C_1(\lambda)$ and $C_2(\lambda)$. In order to determine uniquely the solutions R and L, we have to impose a normalization condition. Let (ξ_0,η_0) be an arbitrary fixed point in Minkowski space; we fix $R(\lambda)$ and $L(\lambda)$ by requiring that

$$R(\xi_0,\eta_0;\lambda) = L(\xi_0,\eta_0;\lambda) = \mathbf{1} \tag{2.23}$$

hold identically in λ. With this normalization, the solutions R and L of (2.20) and (2.22) are unique. Moreover the condition (2.23) has important consequences. First of all, taking (2.20) and (2.22) for λ going to infinity, together with (2.23), gives

$$R(\xi,\eta;\lambda=\infty) = L(\xi,\eta;\lambda=\infty) = \mathbf{1}. \tag{2.24}$$

Second, evaluating (2.20) and (2.22) at $\lambda = 0$ gives

$$R(\xi,\eta;\lambda=0) = g(\xi,\eta) D_1,$$

$$L(\xi,\eta;\lambda=0) = g^{-1}(\xi,\eta) D_2,$$

for some constants D_1 and D_2, while (2.23) forces $D_1 = g_0^{-1}$ and $D_2 = g_0$, where

$$g_0 \equiv g(\xi_0,\eta_0). \tag{2.25}$$

Thus

$$R(\xi,\eta;\lambda=0) = g(\xi,\eta) g_0^{-1}, \tag{2.26}$$

$$L(\xi,\eta;\lambda=0) = g^{-1}(\xi,\eta) g_0. \tag{2.27}$$

Finally, consider

$$Y(\xi,\eta;\lambda) \equiv g(\xi,\eta) L(\xi,\eta;\lambda).$$

A direct calculation shows that

$$\partial_\xi Y(\xi,\eta;\lambda) = (1+1/\lambda)^{-1} A_R Y(\xi,\eta;\lambda),$$

$$\partial_\eta Y(\xi,\eta;\lambda) = (1-1/\lambda)^{-1} B_R Y(\xi,\eta;\lambda),$$

i.e., $Y(\xi,\eta;\lambda) = R(\xi,\eta;1/\lambda) C(\lambda)$ while (2.23) yields $C(\lambda) = g_0$ and thus

$$L(\xi,\eta;\lambda) = g^{-1}(\xi,\eta) R(\xi,\eta;1/\lambda) g_0. \tag{2.28}$$

As a last comment for this case, note that the normalization condition (2.23) uniquely determines R and L once g is given. However, the converse statement does not hold. As Eqs. (2.26) and (2.27) clearly show, R and L only determine an equivalence class of solutions g, modulo their value at (ξ_0,η_0). We shall come back to this question later.

Turning now to the Sl(n,C) σ model with a Wess–Zumino term, the associated linear system is simply[17]

$$\partial_\xi R = (1+\lambda)^{-1}(1-\kappa) A_R R, \tag{2.29a}$$

$$\partial_\eta R = (1-\lambda)^{-1}(1+\kappa) B_R R. \tag{2.29b}$$

The symmetry (2.12b) now suggests defining $(h(\xi,\eta) \equiv g^{-1}(\eta,\xi))$:

$$A_L(\xi,\eta) \equiv (\partial_\xi h(\xi,\eta)) h^{-1}(\xi,\eta) = -(g^{-1} B_R g)|_{(\eta,\xi)}, \tag{2.30a}$$

$$B_L(\xi,\eta) \equiv (\partial_\eta h(\xi,\eta)) h^{-1}(\xi,\eta) = -(g^{-1} A_R g)|_{(\eta,\xi)}. \tag{2.30b}$$

We introduce the analog of (2.22):

$$\partial_\xi L = (1+\lambda)^{-1}(1-\kappa) A_L L, \tag{2.31a}$$

$$\partial_\eta L = (1-\lambda)^{-1}(1+\kappa) B_L L. \tag{2.31b}$$

Again, to fix completely R and L, we impose the condition

$$R(\xi_0,\eta_0;\lambda) = L(\xi_0,\eta_0;\lambda) = \mathbf{1}. \tag{2.32}$$

In order to find simple analogs of (2.26)–(2.28), we choose $\xi_0 = \eta_0$. The consequences of (2.32) are derived in the same way as above and we simply list them here:

$$R(\xi,\eta;\lambda=\infty) = L(\xi,\eta;\lambda=\infty) = \mathbf{1}, \tag{2.33}$$

$$R(\xi,\eta;\lambda=-\kappa) = g(\xi,\eta) g_0^{-1}, \tag{2.34}$$

$$L(\xi,\eta;\lambda=-\kappa) = g^{-1}(\eta,\xi) g_0, \tag{2.35}$$

$$L(\xi,\eta;\lambda) = g^{-1}(\eta,\xi) R(\eta,\xi;-(1+\lambda\kappa)/(\lambda+\kappa)) g_0. \tag{2.36}$$

The last step of this section is to implement constraints of the types (2.14) and (2.19) on the field g into constraints on the solutions R and L of the linear systems listed above. The subgroup reduction (2.14) for the σ models with or without a Wess–Zumino term is obtained by the constraint[6]

$$\sigma(R(\lambda)) = R(\tilde{\lambda}), \tag{2.37}$$

where $\tilde{\lambda} = \lambda$ for σ_1 or σ_3 and $\tilde{\lambda} = \bar{\lambda}$ for σ_2 or σ_4 [see (2.13)]. Indeed, evaluating (2.37) and $\lambda = 0$ or at $\lambda = -\kappa$ (depending on the model), we get

$$g^{-1}\sigma(g) = g_0^{-1}\sigma(g_0)$$

implying that, if g_0 is such that $\sigma(g_0) = g_0$, $\sigma(g) = g$ holds for all (ξ,η). The quotient reduction (2.19) is implemented by the constraint[6]

$$\sigma(R(\lambda)) = L(\tilde{\lambda}). \tag{2.38}$$

Once more, evaluating (2.38) at $\lambda = 0$ gives

$$g\sigma(g) = g_0\sigma(g_0)$$

ensuring that, if g_0 lies on Σ, so does g for all values of (ξ,η).

III. FORMULATION OF σ MODELS IN TAKASAKI'S APPROACH

A. Preliminaries

In the case of the self-dual Yang–Mills equations in four (complex) dimensions, Takasaki[8] proposed to encode the

information contained in the associated linear system into an infinite-dimensional matrix, the latter defining affine coordinates in an infinite-dimensional Grassmann manifold. Using the geometry of this manifold, he was able to linearize the equations and formally solve them. Moreover, this interpretation allowed him to give a nice description of a group action on the space of solutions of the self-dual equations.

Our aim in this section is to transpose this interpretation to the models described in Sec. II. First of all, notice that the linear systems (2.20), (2.22), (2.29), and (2.31) all have the same structure, namely,

$$\partial_\xi R = (1+\lambda)^{-1} A\, R, \tag{3.1a}$$

$$\partial_\eta R = (1-\lambda)^{-1} B\, R, \tag{3.1b}$$

with the normalization condition

$$R(\xi_0,\eta_0;\lambda) = \mathbf{1} \tag{3.2}$$

implying

$$R(\xi,\eta;\lambda=\infty) = \mathbf{1}. \tag{3.3}$$

In the present section we consider the generic system (3.1) with conditions (3.2) and (3.3). It is understood that all the results to be derived equally apply to the four linear systems (2.20), (2.22), (2.29), and (2.31). We will come back to these four specific forms in the next sections.

The inverse R^{-1} of R satisfies

$$\partial_\xi R^{-1} = -(1+\lambda)^{-1} R^{-1} A, \tag{3.4a}$$

$$\partial_\eta R^{-1} = -(1-\lambda)^{-1} R^{-1} B, \tag{3.4b}$$

with the conditions

$$R^{-1}(\xi_0,\eta_0;\lambda) = \mathbf{1}, \tag{3.5}$$

$$R^{-1}(\xi,\eta;\lambda=\infty) = \mathbf{1}. \tag{3.6}$$

The conditions (3.3) and (3.6) allow us to perform an expansion of R and R^{-1} in terms of inverse powers of λ around $\lambda=\infty$:

$$R(\xi,\eta;\lambda) \equiv \sum_{j=0}^{\infty} \lambda^{-j} R_j(\xi,\eta), \tag{3.7a}$$

$$R^{-1}(\xi,\eta;\lambda) \equiv \sum_{j=0}^{\infty} \lambda^{-j} R_j^*(\xi,\eta), \tag{3.7b}$$

$$R_0 = R_0^* = \mathbf{1}, \tag{3.7c}$$

$$R_j(\xi_0,\eta_0) = R_j^*(\xi_0,\eta_0) = 0, \quad \text{for } j \geqslant 1. \tag{3.7d}$$

Inserting (3.7) into (3.1) and (3.4) gives

$$\partial_\xi R_j + \partial_\xi R_{j+1} - A R_j = 0, \tag{3.8a}$$

$$\partial_\eta R_j - \partial_\eta R_{j+1} - B R_j = 0, \tag{3.8b}$$

$$\partial_\xi R_j^* + \partial_\xi R_{j+1}^* + R_j^* A = 0, \tag{3.9a}$$

$$\partial_\eta R_j^* - \partial_\eta R_{j+1}^* + R_j^* B = 0, \tag{3.9b}$$

for $j \geqslant 0$. The relations (3.8) for $j=0$, together with (3.7c) give

$$A = \partial_\xi R_1, \tag{3.10a}$$

$$B = -\partial_\eta R_1. \tag{3.10b}$$

Substituting this back into (3.8) and (3.9) gives

$$\partial_\xi R_j + \partial_\xi R_{j+1} - (\partial_\xi R_1) R_j = 0, \tag{3.11a}$$

$$\partial_\eta R_j - \partial_\eta R_{j+1} + (\partial_\eta R_1) R_j = 0, \tag{3.11b}$$

$$\partial_\xi R_j^* + \partial_\xi R_{j+1}^* + R_j^* (\partial_\xi R_1) = 0, \tag{3.12a}$$

$$\partial_\eta R_j^* - \partial_\eta R_{j+1}^* - R_j^* (\partial_\eta R_1) = 0. \tag{3.12b}$$

B. Interpretation in terms of Grassmann manifold

Equations (3.11) and (3.12) can no longer be solved recursively. In order to avoid this difficulty, Takasaki introduced an infinite-dimensional matrix M whose $(n\times n)$ blocks M_{ij} $(-\infty < i < \infty,\ -\infty < j \leqslant -1)$ are defined by

$$M_{ij} \equiv \sum_{k=-\infty}^{-1} R_{i-k}^* R_{k-j}, \tag{3.13}$$

where it is understood that $R_i = R_i^* = 0$ for $i \leqslant -1$. We refer the reader to Ref. 8 for the proof of the following properties:

$$M_{ij} = \delta_{ij}\mathbf{1}, \quad \text{for } i,j<0, \tag{3.14}$$

$$M_{0j} = -R_{-j}, \quad \text{for } j<0, \tag{3.15}$$

$$M_{i+1,j} = M_{i,j-1} + M_{i,-1} M_{0j}, \quad \text{for } i \in \mathbb{Z},\ j<0. \tag{3.16}$$

The last equation shows that the positive rows of M are entirely determined by the zeroth row, i.e., the infinite matrix M contains exactly the same information as the solution R of the linear system (3.1). Note that (3.16) may be rewritten in matrix form. Let Λ be the infinite shift matrix whose $(n\times n)$ blocks Λ_{ij} $(i,j\in\mathbb{Z})$ are

$$\Lambda_{ij} = \delta_{i+1,j}\mathbf{1} \tag{3.17}$$

and C be the M-dependent matrix whose $(n\times n)$ blocks C_{ij} $(i,j<0)$ are defined as

$$C_{ij} = \delta_{i+1,j}\mathbf{1}, \quad \text{for } i<-1,\ j<0, \tag{3.18a}$$

$$C_{ij} = M_{0j}, \quad \text{for } i=-1. \tag{3.18b}$$

With these notations, (3.16) is simply

$$\Lambda M = MC. \tag{3.16'}$$

We can now interpret M as defining affine coordinates for a point in an infinite-dimensional Grassmann manifold.[9] (See also Ref. 19.) Let V be an infinite-dimensional vector space with a decomposition,

$$V = V_- \oplus V_+, \tag{3.19}$$

where V_- is the (formal) linear span of the basis vectors numbered from $-\infty$ to -1 and V_+ is the (formal) linear span of the basis vectors numbered from 0 to $+\infty$. Points $[P]$ in the Grassmannian are equivalence classes of maps from V_- to V, modulo changes of frame in V_-. Homogeneous coordinates can be given in terms of infinite rectangular

$$[P] \equiv \left\{ \begin{bmatrix} P'_- \\ P'_+ \end{bmatrix} \middle| \exists K\colon V_- \to V_- \text{ invertible such that } \begin{bmatrix} P'_- \\ P'_+ \end{bmatrix} = \begin{bmatrix} P_- K \\ P_+ K \end{bmatrix} \right\}. \tag{3.20}$$

The affine part of the Grassmannian is the subset of points such that P_- is invertible and we identify

$$M \equiv P_+ P_-^{-1}. \tag{3.21}$$

Equation (3.16') then means that the point in the Grassmannian associated to M is invariant under the map described by Λ, C representing the change of frame. Note that

under an arbitrary change of frame

$$M' = MK, \tag{3.22}$$

the matrix C transforms according to

$$C' = K^{-1}CK. \tag{3.23}$$

C. The dynamics of M

We are now ready to state the main result of this section, namely reformulate Eqs. (3.11) and (3.12) in terms of the matrix M.

Equations (3.11) *are equivalent to*

$$(1+\Lambda)\partial_\xi M + MS = 0, \tag{3.24a}$$

$$(1-\Lambda)\partial_\eta M + MT = 0, \tag{3.24b}$$

where the matrices S and T are defined by

$$S_{ij} \equiv -\delta_{i,-1}\,\partial_\xi M_{0j}, \quad \text{for } i,j<0, \tag{3.25a}$$

$$T_{ij} \equiv \delta_{i,-1}\,\partial_\eta M_{0j}, \quad \text{for } i,j<0. \tag{3.25b}$$

Proof: We give the proof for (3.24a) only. First, (3.11a) implies (3.24a). Indeed, for $i\in\mathbb{Z}, j<0$:

$$\{(1+\Lambda)M\}_{ij} = \sum_{k=-\infty}^{+\infty}\sum_{n=-\infty}^{-1}(\delta_{ik}+\delta_{i+1,k})\partial_\xi(R^*_{k-n}R_{n-j}) \quad \text{by} \tag{3.13}$$

$$= \sum_{n=-\infty}^{-1}\{(\partial_\xi R^*_{i-n}+\partial_\xi R^*_{i+1-n})R_{n-j} + (R^*_{i-n}+R^*_{i+1-n})\partial_\xi R_{n-j}\}$$

$$= -\sum_{n=-\infty}^{-1} R^*_{i-n}(\partial_\xi R_1)R_{n-j} + \sum_{n=-\infty}^{-1} R^*_{i-n}\,\partial_\xi R_{n-j} + \sum_{n=-\infty}^{-2} R^*_{i-n}\,\partial_\xi R_{n+1-j} \quad \text{by} \tag{3.12a}$$

$$= -\sum_{n=-\infty}^{-1} R^*_{i-n}(\partial_\xi R_1)R_{n-j} + \sum_{n=-\infty}^{-2} R^*_{i-n}(\partial_\xi R_1)R_{n-j} + R^*_{i+1}\,\partial_\xi R_{-1-j} \quad \text{by} \tag{3.11a}$$

$$= R^*_{i+1}\{-(\partial_\xi R_1)R_{-1-j} + \partial_\xi R_{-1-j}\}$$

$$= -R^*_{i+1}\,\partial_\xi R_{-j} \quad \text{by} \tag{3.11a}$$

$$= \sum_{k=-\infty}^{-1}\sum_{n=-\infty}^{-1} R^*_{i-n}R_{n-k}\delta_{k,-1}\,\partial_\xi M_{0j}$$

$$= \sum_{k=-\infty}^{-1} M_{ik}\delta_{k,-1}\,\partial_\xi M_{0j}$$

$$= -(MS)_{ij}.$$

Conversely, (3.24a) implies (3.11a): for $j<0$,

$$0 = \{(1+\Lambda)\partial_\xi M + MS\}_{0j}$$

$$= \partial_\xi M_{0j} + \partial_\xi M_{1j} - M_{0,-1}(\partial_\xi M_{0j})$$

$$= \partial_\xi M_{0j} + \partial_\xi M_{0,j-1} + (\partial_\xi M_{0,-1})M_{0j} \quad \text{by (3.16)}$$

which is (3.11a). ∎

Note that, under a general change of frame (3.22), solutions M of (3.24) are preserved provided S and T transform according to

$$S' = K^{-1}SK - K^{-1}\,\partial_\xi K - K^{-1}C\,\partial_\xi K, \tag{3.26a}$$

$$T' = K^{-1}TK - K^{-1}\,\partial_\eta K + K^{-1}C\,\partial_\eta K. \tag{3.26b}$$

Of course, under (3.22), M' no longer satisfies (3.14) which means that C', S', and T' are no longer related to M' by relations like (3.18) and (3.25). However, assume that we have a solution M' of (3.24) of the form

$$M' = \begin{bmatrix} M'_{(-)} \\ M'_{(+)} \end{bmatrix}$$

with $M'_{(-)}$ invertible and satisfying the constraint $\Lambda M' = M'C'$ for a certain C'. In other words, the point $[M']$ in the infinite Grassmannian is a fixed point of the map Λ. Define M by

$$M = M'(M'_{(-)})^{-1}$$

and C, S, and T correspondingly by using (3.23) and (3.26) with $K = (M'_{(-)})^{-1}$. Then one easily gets that, because M still satisfies (3.16′) and (3.24), the new matrices C, S, and T are now defined in terms of M by (3.18) and (3.25). This remark will be most important for the next section.

Note also that the normalization condition (3.2) or (3.7d), expressed in terms of the matrix M, is

$$M(\xi_0,\eta_0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{3.27}$$

Finally, we want to mention that we have not been able to solve the evolution problem from Cauchy data for (3.24), contrarily to what happens for the self-dual Yang–Mills case itself [8] or for the supersymmetric ($N=3$) Yang–Mills equations in four dimensions.[19]

D. Group action on the space of solutions of (3.24)

Assume we know a solution M_1 of (3.24) (with S_1 and T_1) satisfying the constraint (3.16′) (with C_1). We want to generate a new solution M_2 from M_1. This will be achieved through multiplication of M_1 on the left by a matrix D with blocks D_{ij} ($i,j\in\mathbb{Z}$). The product DM_1 has still to satisfy (3.16′) and (3.24) and this forces D to fulfill

$$[\Lambda, D] = 0, \tag{3.28a}$$

$$[(1+\Lambda)\partial_\xi, D] = [(1-\Lambda)\partial_\eta, D] = 0. \tag{3.28b}$$

The first equation (3.28a) means that the blocks D_{ij} lying on the same diagonal of D (i.e., for $j-i$ fixed) are equal. Hence (3.28b) means that the blocks D_{ij} have to be constant in

(ξ,η). This allows us to introduce an $(n\times n)$ matrix function of a formal parameter λ which we also denote by D:

$$D(\lambda)=\sum_{j=-\infty}^{+\infty}\lambda^{-j}D_j, \tag{3.29}$$

where D_j is the common value of the blocks on the jth diagonal of D.

The quantity DM_1 does not satisfy (3.14). We denote by $(DM_1)_{(-)}$ the upper block of DM_1 and assume it to be invertible. Then M_2 is defined by

$$M_2\equiv(DM_1)(DM_1)_{(-)}^{-1}. \tag{3.30}$$

This being a change of frame, M_2 still satisfies (3.16') and (3.24) with matrices C_2, S_2, and T_2 obtained from C_1, S_1, and T_1 by transformations of the type (3.23) and (3.26). Moreover, M_2 fulfills (3.14) and C_2, S_2, and T_2 are given in terms of M_2 by relations (3.18) and (3.25), as argued at the end of the previous subsection.

Decomposing D into four infinite blocks according to (3.19),

$$D\equiv\begin{pmatrix}d_1 & d_2\\ d_3 & d_4\end{pmatrix}, \tag{3.31}$$

M_2 is explicitly given by

$$M_2\equiv\begin{pmatrix}\mathbf{1} & \\ (d_3+d_4M_{1(+)})(d_1+d_2M_{1(+)})^{-1} & \end{pmatrix}. \tag{3.32}$$

Assume now M_1 satisfies the normalization condition (3.27), i.e., $M_{1(+)}(\xi_0,\eta_0)=0$. Solution M_2 will satisfy (3.27), too, provided that

$$\begin{aligned}0&=M_{2(+)}(\xi_0,\eta_0)\\&=(d_3+d_4M_{1(+)}(\xi_0,\eta_0))(d_1+d_2M_{1(+)}(\xi_0,\eta_0))^{-1}\\&=d_3d_1^{-1},\end{aligned}$$

i.e., $d_3=0$. Due to the structure of D imposed by (3.28), this forces d_1 and d_4 to be (block) upper triangular matrices and thus D has to be (block) upper triangular. (We use *block upper triangular* to characterize a matrix whose blocks under the main diagonal are zero.) This transforms (3.29) into

$$D(\lambda)=\sum_{j=0}^{+\infty}\lambda^{-j}D_j. \tag{3.33}$$

Note that (3.30) provides us with an obvious group law under multiplication of infinite matrices (when this product makes sense). However, this group law is expressed at the level of the infinite matrix M and it does not seem possible to give an explicit form of the group action on the finite matrix R (and thus at the level of the solutions of the nonlinear field equations under study). On the other hand, what is possible is to express for R the infinitesimal action corresponding to this group law, and this is what we are going to describe now.

We assume that D is close to the identity, which means

$$d_i=\mathbf{1}+\epsilon_i,\quad \text{for } i=1,4,$$

$$d_i=\epsilon_i\quad \text{for } i=2,$$

and all ϵ_i infinitesimal. The lower block of M_2 in (3.32) is then given by (at first order in the ϵ's)

$$\begin{aligned}&(\mathbf{1}+\epsilon_4)M_{1(+)}(\mathbf{1}+\epsilon_1+\epsilon_2M_{1(+)})^{-1}\\&\quad\cong(M_{1(+)}+\epsilon_4M_{1(+)})(\mathbf{1}-\epsilon_1-\epsilon_2M_{1(+)})\\&\quad\cong M_{1(+)}+\epsilon_4M_{1(+)}-M_{1(+)}\epsilon_1-M_{1(+)}\epsilon_2M_{1(+)},\end{aligned}$$

which is a Riccati-type transformation law, already well known to play a crucial role in two-dimensional integrable models.[6,7] Looking at the zeroth row of the previous expression, we can define an infinitesimal transformation law for the coefficients R_k of R $(k>0)$:

$$\begin{aligned}-\delta R_k&=(\epsilon_4M_{1(+)}-M_{1(+)}\epsilon_1-M_{1(+)}\epsilon_2M_{1(+)})_{0,-k}\\&=\sum_{n=0}^{+\infty}(\epsilon_4)_{0n}(M_{1(+)})_{n,-k}\\&\quad-\sum_{n=-\infty}^{-1}(M_{1(+)})_{0,n}(\epsilon_1)_{n,-k}\\&\quad-\sum_{n=-\infty}^{-1}\sum_{m=0}^{+\infty}(M_{1(+)})_{0n}(\epsilon_2)_{nm}(M_{1(+)})_{m,-k}\\&=\sum_{n=0}^{+\infty}D_n\sum_{j=1}^{+\infty}R^{*}_{n+j}R_{k-j}+\sum_{n=1}^{+\infty}R_nD_{n-k}\\&\quad+\sum_{n=1}^{+\infty}\sum_{m=0}^{+\infty}R_nD_{n+m}\sum_{j=1}^{+\infty}R^{*}_{m+j}R_{k-j}.\end{aligned}$$

Assume now $D(\lambda)$ to be $D(\lambda)=\lambda^{-i}T$, where $i\geqslant 0$ is fixed and T is a generator of $\mathrm{sl}(n,\mathbb{C})$. Then

$$\begin{aligned}-\delta^{T(i)}R_k&=R_{k+i}T+T\sum_{j=1}^{+\infty}R^{*}_{i+j}R_{k-j}\\&\quad+\sum_{n=1}^{i}\sum_{j=1}^{+\infty}R_nTR^{*}_{i-n+j}R_{k-j}\\&=R_{k+i}T+\sum_{n=0}^{i}\sum_{j=1}^{k}R_nTR^{*}_{i-n+j}R_{k-j}.\end{aligned} \tag{3.34}$$

In order to make expressions more compact, we introduce a formal parameter λ' and a generating function $\delta^T(\lambda')$ defined as

$$\delta^T(\lambda')\equiv\sum_{i=0}^{+\infty}\lambda'^{-i}\delta^{T(i)}. \tag{3.35}$$

The following result links the infinitesimal action (3.34) obtained in the context of infinite Grassmann manifolds to Wu's generating function[2]: *In terms of the generating function, the transformation law* (3.34) *is given by*

$$\begin{aligned}&\{\delta^T(\lambda')R(\lambda)\}R^{-1}(\lambda)\\&\quad=[\lambda'/(\lambda'-\lambda)]\{R(\lambda')TR^{-1}(\lambda')\\&\qquad-R(\lambda)TR^{-1}(\lambda)\}.\end{aligned} \tag{3.36}$$

Proof: Introducing the notation

$$T^m\equiv\sum_{i=0}^{m}R_iTR^{*}_{m-i},$$

one has

$$\begin{aligned}&\{\delta^T(\lambda')R(\lambda)\}R^{-1}(\lambda)\\&\quad=\left[1-\frac{\lambda}{\lambda'}\right]^{-1}\sum_{m=1}^{+\infty}(\lambda'^{-m}-\lambda^{-m})T^m\\&\quad=\sum_{m=1}^{+\infty}(-\lambda^{-m})\frac{[1-(\lambda/\lambda')^m]}{[1-\lambda/\lambda']}T^m\\&\quad=-\sum_{m=1}^{+\infty}\lambda^{-m}\sum_{j=0}^{m-1}\left(\frac{\lambda}{\lambda'}\right)^jT^m\\&\quad=-\sum_{n=1}^{+\infty}\sum_{i=0}^{+\infty}\lambda'^{-i}\lambda^{-n}T^{n+i}.\end{aligned}$$

Then

$$-\delta^T(\lambda')R(\lambda) = -\sum_{i,k=0}^{+\infty}\sum_{n=1}^{+\infty}\lambda'^{-i}\lambda^{-(k+n)}T^{n+i}R_k$$

$$= -\sum_{k=1}^{+\infty}\sum_{i=0}^{+\infty}\sum_{n=0}^{k-1}\lambda^{-k}\lambda'^{-i}T^{k+i-n}R_n,$$

which gives

$$-\delta^{T(i)}R_k = \sum_{n=0}^{k-1}T^{k+i-n}R_n$$

$$= \sum_{n=0}^{k-1}\sum_{m=0}^{k+i-n}R_m TR^*_{k+i-n-m}R_n$$

$$= \left\{\sum_{m=0}^{i}\sum_{n=0}^{k-1} + \sum_{m=i+1}^{i+k}\sum_{n=0}^{i+k-m}\right\} \times R_m TR^*_{k+i-n-m}R_n$$

$$= \sum_{m=0}^{i}\sum_{j=1}^{k}R_m TR^*_{i+j-m}R_{k-j} + \sum_{m=i+1}^{i+k}R_m T\delta_{0,i+k-m}$$

$$= R_{i+k}T + \sum_{n=0}^{i}\sum_{j=1}^{k}R_n TR^*_{i-n+j}R_{k-j},$$

which is relation (3.34). ■

We are now going to apply relation (3.36) to the different linear systems we introduced in Sec. II.

IV. THE PRINCIPAL σ MODELS

A. The transformation laws for the Sl(n,C) model

The aim of this subsection is to give the explicit transformation laws corresponding to (3.36) for the case of the Sl(n,C) principal σ model. Taking the symmetry (2.7b) into account, we apply (3.36) to the solutions $R(\lambda)$ and $L(\lambda)$ of the linear systems (2.20) and (2.22) associated to g and g^{-1}, respectively. We thus get two different types of transformations δ_R and δ_L, defined by

$$\{\delta_R^T(\lambda')R(\lambda)\}R^{-1}(\lambda) = [\lambda'/(\lambda'-\lambda)]\{R(\lambda')TR^{-1}(\lambda') - R(\lambda)TR^{-1}(\lambda)\}, \tag{4.1}$$

$$\{\delta_L^T(\lambda')L(\lambda)\}L^{-1}(\lambda) = [\lambda'/(\lambda'-\lambda)]\{L(\lambda')TL^{-1}(\lambda') - L(\lambda)TL^{-1}(\lambda)\}, \tag{4.2}$$

for $T\in$sl(n,C).

We now want to derive the action of δ_R and δ_L on the field g itself. Evaluating (4.1) and (4.2) at $\lambda = 0$ and using (2.26) and (2.27), we get

$$\{\delta_R^T(\lambda')(gg_0^{-1})\}(g_0g^{-1}) = R(\lambda')TR^{-1}(\lambda') - (gg_0^{-1})T(g_0g^{-1}),$$

$$\{\delta_L^T(\lambda')(g^{-1}g_0)(g_0^{-1}g) = L(\lambda')TL^{-1}(\lambda') - (g^{-1}g_0)T(g_0^{-1}g),$$

which gives

$$g^{-1}\delta_R^T(\lambda')g - g^{-1}R(\lambda')TR^{-1}(\lambda')g = g_0^{-1}\delta_R^T(\lambda')g_0 - g_0^{-1}Tg_0,$$

$$(\delta_L^T(\lambda')g)g^{-1} + gL(\lambda')TL^{-1}(\lambda')g^{-1} = (\delta_L^T(\lambda')g_0)g_0^{-1} + g_0Tg_0^{-1}.$$

The rhs's of these equations being independent of (ξ,η), the lhs's have to be arbitrary functions of λ' only. The problem is now to fix these functions. Clearly, relations (4.1) and (4.2) do not fix them. This is linked to a fact already underlined in Sec. II: the correspondance between g and $R(\lambda)$ [or $L(\lambda)$] is not one-to-one. With normalization (2.23), $R(\lambda)$ and $L(\lambda)$ are uniquely fixed by g, but the converse is not true. Here $R(\lambda)$ [resp. $L(\lambda)$] only fixes an equivalence class of solutions g, the equivalence relation being given by right (resp. left) multiplication of the solution by a constant matrix, as (2.26) [resp. (2.27)] easily shows. We thus now have to choose a class representative in order to determine $\delta_R^T(\lambda')g$ and $\delta_L^T(\lambda')g$. Note that the way we make this choice is irrelevant. Indeed, a different choice would correspond to taking linear combinations of the $\delta_R^{(i)}$ (resp. $\delta_L^{(i)}$) and $\delta_L^{(0)}$ (resp. $\delta_R^{(0)}$) and thus this would be a change of basis in the algebra spanned by them. We take the simplest choice, namely set the aforementioned arbitrary functions of λ' equal to zero and get the transformation laws for g:

$$\delta_R^T(\lambda')g = R(\lambda')TR^{-1}(\lambda')g, \tag{4.3}$$

$$\delta_L^T(\lambda')g = -gL(\lambda')TL^{-1}(\lambda'). \tag{4.4}$$

We ultimately want to compute the commutators between all these δ's. As (4.3) and (4.4) show, we also need to know $\delta_R L$ and $\delta_L R$ in order to proceed further. Using (2.28) and (4.1)–(4.4), we immediately get

$$\{\delta_L^T(\lambda')R(\lambda)\}R^{-1}(\lambda) = [1/(\lambda\lambda'-1)]\{gL(\lambda')TL^{-1}(\lambda')g^{-1} - R(\lambda)g_0Tg_0^{-1}R^{-1}(\lambda)\}, \tag{4.5}$$

$$\{\delta_R^T(\lambda')L(\lambda)\}L^{-1}(\lambda) = [1/(\lambda\lambda'-1)]\{g^{-1}R(\lambda')TR^{-1}(\lambda')g - L(\lambda)g_0^{-1}Tg_0L^{-1}(\lambda)\}. \tag{4.6}$$

B. The commutation rules

We are now ready to compute the commutation rules between all the different δ's. For example, we treat explicitly the case of two δ_R's. Using (4.3) and (4.1), we get for U, V in sl(n,C),

$$g \xrightarrow{\delta_R^V(\lambda')} g + \delta_R^V(\lambda')g \xrightarrow{\delta_R^U(\lambda)} (g + \delta_R^V(\lambda')g) + \delta_R^U(\lambda)(g + \delta_R^V(\lambda')g)$$

$$= g + R(\lambda')VR^{-1}(\lambda')g + \{R(\lambda) + [\lambda'/(\lambda'-\lambda)](R(\lambda')VR^{-1}(\lambda')R(\lambda) - R(\lambda)V)\}U \times\{R^{-1}(\lambda) - [\lambda'/(\lambda'-\lambda)](R^{-1}(\lambda)R(\lambda')VR^{-1}(\lambda') - VR^{-1}(\lambda))\}\{g + R(\lambda')VR^{-1}(\lambda')g\}$$

$$= g + R(\lambda')VR^{-1}(\lambda')g + R(\lambda)UR^{-1}(\lambda)g + R(\lambda)UR^{-1}(\lambda)R(\lambda')VR^{-1}(\lambda')g - [\lambda'/(\lambda'-\lambda)][R(\lambda)UR^{-1}(\lambda), R(\lambda')VR^{-1}(\lambda')]g + [\lambda'/(\lambda'-\lambda)]R(\lambda)[U,V]R^{-1}(\lambda)g$$

up to second order in U, V and thus

$$[\delta_R^U(\lambda),\delta_R^V(\lambda')]g=[\lambda'/(\lambda'-\lambda)]R(\lambda)[U,V]R^{-1}(\lambda)g+[\lambda/(\lambda-\lambda')]R(\lambda')[U,V]R^{-1}(\lambda')g$$
$$=[1/(\lambda'-\lambda)]\{\lambda'\delta_R^{[U,V]}(\lambda)-\lambda\delta_R^{[U,V]}(\lambda')\}g. \tag{4.7}$$

Similary, using (4.4), (4.2) and (4.5), (4.6) we can derive

$$[\delta_L^U(\lambda),\delta_L^V(\lambda')]g=[1/(\lambda'-\lambda)]\{\lambda'\delta_L^{[U,V]}(\lambda)-\lambda\delta_L^{[U,V]}(\lambda')\}g, \tag{4.8}$$

$$[\delta_R^U(\lambda),\delta_L^V(\lambda')]g=[1/(\lambda\lambda'-1)]\{\delta_R^{[U,g_0Vg_0^{-1}]}(\lambda)+\delta_L^{[g_0^{-1}Ug_0,V]}(\lambda')\}g. \tag{4.9}$$

Inserting expansion (3.35) in both sides of (4.7)–(4.9) and collecting identical powers of λ,λ' gives, after some manipulation,

$$[\delta_R^{U(i)},\delta_R^{V(j)}]g=\delta_R^{[U,V](i+j)}g, \tag{4.10}$$

$$[\delta_L^{U(i)},\delta_L^{V(j)}]g=\delta_L^{[U,V](i+j)}g, \quad \text{for } i,j\in\mathbf{N}, \tag{4.11}$$

$$[\delta_R^{U(i)},\delta_L^{V(j)}]g=\begin{cases}0, & \text{if } i \text{ or } j=0,\\ \{\delta_R^{[U,g_0Vg_0^{-1}](0)}+\delta_L^{[g_0^{-1}Ug_0,V](0)}\}g, & \text{if } i=j\neq 0,\\ \delta_R^{[U,g_0Vg_0^{-1}](i-j)}g, & \text{if } i>j\geqslant 1,\\ \delta_L^{[g_0^{-1}Ug_0,V](j-i)}g, & \text{if } j>i\geqslant 1.\end{cases} \tag{4.12}$$

Equations (4.10) and (4.11) indicate that the δ_R's and δ_L's separately span two loop algebras graded over N. The relationship between these two algebras is described by (4.12). The global structure generated by both sets of transformations is identified by performing the following two steps. First define

$$\tilde{\delta}_L^{U(i)}g\equiv\delta_L^{g_0^{-1}Ug_0(i)}g. \tag{4.13}$$

Then, relations (4.11) and (4.12) are transformed into

$$[\tilde{\delta}_L^{U(i)},\tilde{\delta}_L^{V(j)}]g=\tilde{\delta}_L^{[U,V](i+j)}g, \tag{4.11'}$$

$$[\delta_R^{U(i)},\tilde{\delta}_L^{V(j)}]g=\begin{cases}0, & \text{if } i \text{ or } j=0,\\ \{\delta_R^{[U,V](0)}+\tilde{\delta}_L^{[U,V](0)}\}g, & \text{if } i=j\neq 0,\\ \delta_R^{[U,V](i-j)}g, & \text{if } i>j\geqslant 1,\\ \tilde{\delta}_L^{[U,V](j-i)}g, & \text{if } j>i\geqslant 1.\end{cases} \tag{4.12'}$$

Second, introduce generators $\hat{\delta}^{U(i)}$ for $i\in\mathbf{Z}$ by

$$\begin{aligned}&\hat{\delta}^{U(i)}\equiv\delta_R^{U(i)}, && \text{for } i\geqslant 1,\\ &\hat{\delta}^{U(0)}\equiv\delta_R^{U(0)}+\tilde{\delta}_L^{U(0)}, && \text{for } i=0,\\ &\hat{\delta}^{U(i)}\equiv\tilde{\delta}_L^{U(-i)}, && \text{for } i\leqslant -1.\end{aligned} \tag{4.14}$$

Note that there remains one independent combination, namely

$$\check{\delta}^U\equiv\delta_R^{U(0)}-\tilde{\delta}_L^{U(0)}. \tag{4.15}$$

The above commutation relations then become

$$[\hat{\delta}^{U(i)},\hat{\delta}^{V(j)}]g=\hat{\delta}^{[U,V](i+j)}g, \quad \text{for } i,j\in\mathbf{Z}, \tag{4.16}$$

$$[\hat{\delta}^{U(i)},\check{\delta}^V]g=0, \quad \text{for } i\in\mathbf{Z}, \tag{4.17}$$

which shows that the algebra spanned by the symmetry transformations has the structure of the direct sum of a finite-dimensional Lie algebra with a loop algebra graded over $\mathbf{Z}$: $\mathrm{sl}(n,\mathbf{C})\oplus(\mathrm{sl}(n,\mathbf{C})\otimes\mathbf{C}[t,t^{-1}])$.

C. The subgroup reductions

The last step of this section is to consider the principal σ models with values in a subgroup G of $\mathrm{Sl}(n,\mathbf{C})$. We thus start with a solution g satisfying the constraint (2.14) and the associated solution of the linear system $R(\lambda)$ subject to (2.37). In order that the new solution generated by the symmetry transformation also lies in G, we have to impose

$$\begin{aligned}&\sigma(g+\delta_R^T(\lambda)g)=g+\delta_R^T(\tilde{\lambda})g,\\ &\sigma(g+\delta_L^T(\lambda)g)=g+\delta_L^T(\tilde{\lambda})g,\end{aligned} \tag{4.18}$$

the $\tilde{\lambda}$ in the rhs being there so that the constraint be holomorphic in λ. For example, we treat the R case. Let σ_* be the differential of the automorphism σ at the identity in $\mathrm{Sl}(n,\mathbf{C})$. Due to (2.14), (2.37), and (4.3), we have

$$\sigma(g+\delta_R^T(\lambda)g)=\sigma(\mathbf{1}+\delta_R^T(\lambda)g\cdot g^{-1})\sigma(g)$$
$$=g+\sigma_*(\delta_R^T(\lambda)g\cdot g^{-1})g,$$

where

$$\sigma_*(\delta_R^T(\lambda)g\cdot g^{-1})=\sigma_*(\mathrm{ad}_{R(\lambda)}T)=\mathrm{ad}_{\sigma(R(\lambda))}\sigma_*(T)$$
$$=\mathrm{ad}_{R(\tilde{\lambda})}\sigma_*(T).$$

Using (2.37), relation (4.18) is seen to be equivalent to the condition

$$\sigma_*(T)=T, \tag{4.19}$$

which means that T lies in $\mathfrak{g}$, the Lie algebra of the group G.

The main result of this section is then the following: *For the principal σ model with values in the Lie group G, with Lie algebra $\mathfrak{g}$, the generators of the symmetry transformations defined by* (4.14) *and* (4.15) *span an infinite-dimensional Lie algebra with structure* $\mathfrak{g}\oplus(\mathfrak{g}\otimes\mathbf{C}[t,t^{-1}])$ *or* $\mathfrak{g}\oplus(\mathfrak{g}\otimes\mathbf{R}[t,t^{-1}])$ *according to whether* $\mathfrak{g}$ *is a complex or a real algebra.* [Strictly speaking, the only groups which are irreducible RSS are $\mathrm{SU}(n)$, $\mathrm{SO}(n)$, and $\mathrm{Sp}(n)$ and thus we

should only consider the second structure.]

We would like to close this section by showing how the subgroup reduction can fit nicely in the setting of Sec. III. We shall give the example of the reduction from the $Sl(n,\mathbb{C})\sigma$ model to the $SU(n)\sigma$ model. For this reduction, the constraint (2.37) on the solution of the linear system reads

$$R^{\dagger -1}(\bar{\lambda}) = R(\lambda), \tag{4.20}$$

which is simply

$$R_i^{*\dagger} = R_i \quad \text{for } i \geqslant 0. \tag{4.21}$$

This implies on the $n \times n$ blocks of the matrix M representing the point $[P]$ in affine coordinates ($i \geqslant 0, j < 0$):

$$\begin{aligned} M_{ij}^{\dagger} &= \sum_{k=j}^{-1} R^{*}_{k-j} R_{i-k} \\ &= \delta_{ij}\mathbb{1} - \sum_{k=-i-1}^{-1} R^{*}_{-j-k-1} R_{i+k+1} \\ &= -M_{-j-1,-i-1}, \quad \text{for } i \geqslant 0,\ j < 0. \end{aligned} \tag{4.22}$$

This constraint is equivalent to the following (infinite-dimensional) matrix equation:

$$0 = (M^{\dagger} \quad \mathbb{1}) \begin{pmatrix} \mathbb{1} \\ M \end{pmatrix}, \tag{4.23}$$

where $M^{\dagger}$ is defined by (4.22): $(M^{\dagger})_{ij} = (M_{-j-1,-i-1})^{\dagger}$. This definition of the † consists of taking the usual matrix † of each $n \times n$ block and then transposing by blocks the result with respect to the diagonal blocks $M_{i,i-1}$, $i \geqslant 0$. In terms of the geometry of the infinite Grassmann manifold, the condition (4.23) is really eloquent. It means that the points $[P] = \binom{\mathbb{1}}{M}$ representing solutions of the $SU(n)$ σ models are totally isotropic planes with respect to the bilinear form defined in (4.23). [This is analogous to what happens in the construction of soliton solutions with the soliton correlation matrix[6,20]; in that context, the soliton correlation matrix (related to the dressing matrix[16]) represents a point in a finite-dimensional Grassmannian. The subgroup reduction implies also that this point is a totally isotropic plane with respect to a given bilinear form.]

Let $\mathcal{G}$ be the group of infinite matrices D with blocks D_{ij}, $i,j \in \mathbb{Z}$ solving Eqs. (3.28). The group $\mathcal{G}_0$ preserving the constraint (4.20) through the action (3.30) is the subgroup of $\mathcal{G}$ leaving the bilinear form (4.23) constant, i.e., such that

$$D^{\dagger} D = \mathbb{1}_{\infty \times \infty}. \tag{4.24}$$

[Again † should be understood as in (4.23).] At the infinitesimal level, $D = \mathbb{1} + \epsilon \mathcal{D}$ ($\epsilon \ll 1$), this condition reads

$$\mathcal{D}^{\dagger} + \mathcal{D} = 0 \tag{4.25}$$

or, equivalently,

$$\begin{aligned} \mathcal{D}^{\dagger}_{--} + \mathcal{D}_{++} &= 0, \\ \mathcal{D}^{\dagger}_{+-} + \mathcal{D}_{+-} &= 0, \\ \mathcal{D}^{\dagger}_{-+} + \mathcal{D}_{-+} &= 0, \end{aligned} \tag{4.26}$$

if we write $\mathcal{D}$ as

$$\mathcal{D} = \begin{pmatrix} \mathcal{D}_{--} & \mathcal{D}_{-+} \\ \mathcal{D}_{+-} & \mathcal{D}_{++} \end{pmatrix}.$$

[A $-$ or $+$ sign means that the index at that position belongs to $(-\infty, -1]$ or $[0, +\infty) \subset \mathbb{Z}$, respectively.] The above equations (4.26) mean that each block $\mathcal{D}_{ij}$, $i,j \in \mathbb{Z}$ should be an anti-Hermitian matrix

$$\mathcal{D}_{ij} = -\mathcal{D}_{ij}^{\dagger}, \tag{4.27}$$

which is Eq. (4.19). Hence the geometric condition (4.23) is equivalent to the reduction equation (4.20).

V. THE σ MODELS WITH A WESS–ZUMINO TERM

In this section, the σ models with a Wess–Zumino term (WZσ models) will be studied. Only the models with values in $G = SU(n)$, $SO(n)$, and $Sp(n)$ will be considered. (These groups are the only irreducible RSS's that are Lie groups. See the remark in Sec. IV C.) To our knowledge, a Wess–Zumino term for σ models with values in a RSS which is not a Lie group has not yet been devised. (See, however, Ref. 21.) The goal of this section is to define an infinite set of symmetry transformations for these models and to identify the algebraic structure they span. As by-products, another example of the reduction procedure will be presented and a useful tool will be introduced: the changes of basis preserving a gradation over $\mathbb{N}$.

A. The infinitesimal transformations $\delta_R g$ and $\delta_L g$ and the reduction problem

As for the principal models, the starting point of the discussion is the infinitesimal variations $\delta_R^T(\lambda')R(\lambda)$ and $\delta_L^T(\lambda')L(\lambda)$ of the solutions of the associated linear systems

$$\begin{aligned} &\{\delta_R^T(\lambda')R(\lambda)\}R^{-1}(\lambda) \\ &\quad = [\lambda'/(\lambda'-\lambda)]\{R(\lambda')TR^{-1}(\lambda') \\ &\qquad - R(\lambda)TR^{-1}(\lambda)\}, \end{aligned} \tag{5.1}$$

$$\begin{aligned} &\{\delta_L^T(\lambda')L(\lambda)\}L^{-1}(\lambda) \\ &\quad = [\lambda'/(\lambda'-\lambda)]\{L(\lambda')TL^{-1}(\lambda') \\ &\qquad - L(\lambda)TL^{-1}(\lambda)\}, \end{aligned} \tag{5.2}$$

together with Eqs. (2.34) and (2.35) which show how $g(\xi,\eta)$ and $g^{-1}(\eta,\xi)$ are retrieved from $R(\lambda)$ and $L(\lambda)$, respectively:

$$R(\xi,\eta;\ \lambda = -\kappa) = g(\xi,\eta)g_0^{-1}, \tag{2.34}$$

$$L(\xi,\eta;\ \lambda = -\kappa) = g^{-1}(\eta,\xi)g_0. \tag{2.35}$$

Setting $\lambda = -\kappa$ in (5.1) and (5.2), one gets

$$\begin{aligned} &g^{-1}(\xi,\eta)\{\delta_R^T(\lambda')g(\xi,\eta) \\ &\quad - [\lambda'/(\lambda'+\kappa)]R(\xi,\eta;\lambda')TR^{-1}(\xi,\eta;\lambda')g(\xi,\eta)\} \\ &\quad = g_0^{-1}\{\delta_R^T(\lambda')g_0 - [\lambda'/(\lambda'+\kappa)]Tg_0\} \end{aligned} \tag{5.3}$$

and

$$\begin{aligned} &\{\delta_L^T(\lambda')g(\eta,\xi) + [\lambda'/(\lambda'+\kappa)]g(\eta,\xi)L(\xi,\eta;\lambda') \\ &\quad \times TL^{-1}(\xi,\eta;\lambda')\}g^{-1}(\eta,\xi) \\ &\quad = \{\delta_L^T(\lambda')g_0 + [\lambda'/(\lambda'+\kappa)]g_0T\}g_0^{-1}. \end{aligned} \tag{5.4}$$

As in Sec. IV, we note that the right-hand sides of (5.3) and (5.4) depend only on λ' and so should the left-hand sides. A similar argument leads then to the following generating functions for the infinitesimal transformations δ_R and δ_L on g in the WZσ models

$$\delta_R^T(\lambda')g(\xi,\eta) = [\lambda'/(\lambda'+\kappa)]R(\xi,\eta;\lambda')TR^{-1}(\xi,\eta;\lambda')g(\xi,\eta), \tag{5.5}$$

$$\delta_L^T(\lambda')g(\eta,\xi) = -[\lambda'/(\lambda'+\kappa)]g(\eta,\xi)L(\xi,\eta;\lambda')TL^{-1}(\xi,\eta;\lambda'). \tag{5.6}$$

These relations are the analog of Eqs. (4.3) and (4.4) for the principal models. Note that the (ξ,η) dependency is explicitly written because L at (ξ,η) is related to g^{-1} at the parity-transformed point (η,ξ).

The reduction constraints are all of the type group–subgroup and are identical to those of the principal models (without the Wess–Zumino term). [The fact that the solutions $g(\xi,\eta)$ and $g^{-1}(\eta,\xi)$ are obtained, respectively, from $R(\lambda)$ and $L(\lambda)$ evaluated at $\lambda=-\kappa$ (instead of at $\lambda=0$) does not alter the algebraic constraint on $R(\lambda)$ or $L(\lambda)$ because $\kappa\in\mathbb{R}$. For example, $R^\dagger(\bar\lambda)=R^{-1}(\lambda)$ is indeed the unitarity condition for the solution $g(\xi,\eta)$ of the WZσ model. Setting $\lambda=-\kappa$, one gets $g^\dagger g=g_0^\dagger g_0$, i.e., $g(\xi,\eta)$ is unitary at all (ξ,η) if it is at any point, for example, at (ξ_0,η_0).] Hence the reduction conditions

$$\begin{aligned}\sigma(g+\delta_R^T(\lambda)g)&=g+\delta_R^T(\tilde\lambda)g,\\ \sigma(g+\delta_L^T(\lambda)g)&=g+\delta_L^T(\tilde\lambda)g,\end{aligned} \tag{5.7}$$

lead to the same constraint on T as discussed in Sec. IV C,

$$\sigma_*(T)=T, \tag{5.8}$$

where σ_* is the differential map associated to the automorphism σ. Note, finally, that the realization of the reduction conditions as geometric constraints on the point $[P]$ in the Grassmann manifold (presented in Sec. IV C) applies to the present case without any change.

B. The commutation rules $[\delta_R^U(\lambda),\delta_L^V(\lambda')]g$

Even though the symmetry transformations for $\delta_R^T(\lambda')R(\lambda)$ and $\delta_L^T(\lambda')L(\lambda)$ are identical to those of the principal σ models, the infinitesimal $\delta_L^T(\lambda')R(\lambda)$ and $\delta_R^T(\lambda')L(\lambda)$ are different. The reason for that is twofold: first, the induced transformations on g are different and, second, the relationship between $R(\lambda)$ and $L(\lambda)$ is more complex than in the case of the principal models.

A calculation similar to the one performed in Sec. IV using Eqs. (5.1), (5.2), (5.5), (5.6), and (2.36) leads to the following infinitesimal transformations:

$$\begin{aligned}\{\delta_R^T(\lambda')L(\xi,\eta;\lambda)\}L^{-1}(\xi,\eta;\lambda) &= \frac{\lambda'(\kappa^2-1)}{(\lambda+\kappa)(\lambda'+\kappa)(\lambda'+K(\lambda))}\\ &\times\{g^{-1}(\eta,\xi)R(\eta,\xi;\lambda')TR^{-1}(\eta,\xi;\lambda')g(\eta,\xi)\\ &-L(\xi,\eta;\lambda)g_0^{-1}Tg_0L^{-1}(\xi,\eta;\lambda)\},\end{aligned} \tag{5.9}$$

$$\begin{aligned}\{\delta_L^T(\lambda')R(\xi,\eta;\lambda)\}R^{-1}(\xi,\eta;\lambda) &= \frac{\lambda'(\kappa^2-1)}{(\lambda+\kappa)(\lambda'+\kappa)(\lambda'+K(\lambda))}\\ &\times\{g(\xi,\eta)L(\eta,\xi;\lambda')TL^{-1}(\eta,\xi;\lambda')g^{-1}(\xi,\eta)\\ &-R(\xi,\eta;\lambda)g_0Tg_0^{-1}R^{-1}(\xi,\eta;\lambda)\},\end{aligned} \tag{5.10}$$

where

$$K(\lambda)=(1+\lambda\kappa)/(\lambda+\kappa). \tag{5.11}$$

The set of relations (5.1), (5.5), and (5.9) can be transformed in the set (5.2), (5.6), and (5.10) by the following identification:

$$\begin{aligned}&\text{subindex } R\leftrightarrow\text{subindex } L,\\ &R(\xi,\eta;\lambda)\leftrightarrow L(\xi,\eta;\lambda),\\ &g(\xi,\eta)\leftrightarrow g^{-1}(\eta,\xi),\\ &g_0\leftrightarrow g_0^{-1}.\end{aligned} \tag{5.12}$$

With these expressions, the commutator $[\delta_R^U(\lambda),\delta_R^V(\lambda')]g$ can be obtained directly,

$$\begin{aligned}[\delta_R^U(\lambda),\delta_L^V(\lambda')]g &= \frac{(\kappa^2-1)}{(1+(\lambda+\lambda')\kappa+\lambda\lambda')}\\ &\times\left\{\frac{\lambda'}{\lambda'+\kappa}\delta_R^{[U,g_0Vg_0^{-1}]}(\lambda)\right.\\ &\left.+\frac{\lambda}{\lambda+\kappa}\delta_L^{[g_0^{-1}Ug_0,V]}(\lambda')\right\}g.\end{aligned} \tag{5.13}$$

Again the value of g_0 appears explicitly in the right-hand side and can be taken care of as previously by a redefinition of $\delta_L^T\to\tilde\delta_L^T\equiv\delta_L^{g_0^{-1}Tg_0}$. The main problem here is the appearance of this complicated function of λ and λ' in front of the curly brackets which makes the structure of the algebra spanned by δ_R and δ_L unreadable. A change of basis will be necessary to shed light on the structure of this algebra. This is the subject to be discussed in the next paragraph.

As a closing observation, it should be noted that, as $\kappa\to0$, the commutator does not give back the commutator for the principal model but

$$\begin{aligned}[\delta_R^U(\lambda),\delta_L^V(\lambda')]g|_{\kappa=0} &= [-1/(\lambda\lambda'+1)]\{\delta_R^{[U,g_0Vg_0^{-1}]}(\lambda)\\ &+\delta_L^{[g_0^{-1}Ug_0,V]}(\lambda')\}g.\end{aligned} \tag{5.14}$$

The discrepancy comes from the fact that, in the precedent case, L was the solution associated to $g^{-1}(\xi,\eta)$ contrarily to the present case where L is the solution associated to $g^{-1}(\eta,\xi)$. (Recall that the WZσ model is invariant neither under the parity transformation nor under the inversion operation but only when both discrete symmetries are performed simultaneously.) On the solution of the linear system, the parity transformation reads

$$L(\xi,\eta;\lambda)\to L(\eta,\xi;-\lambda). \tag{5.15}$$

Hence the transformations δ_L for the principal models and those for the WZσ models are related by changing the formal parameter λ into $-\lambda$. Changing $\lambda'\to-\lambda'$ in (5.14), one recovers (4.9).

C. The changes of basis preserving a gradation over $\mathbb{N}$

Let $\hat{\mathfrak{g}}=\mathfrak{g}\otimes\mathbb{R}[t]$ be a Lie algebra graded over $\mathbb{N}$ with $\mathfrak{g}$ a finite-dimensional semisimple algebra with $\dim\mathfrak{g}=d$. As before, the elements of the ith graded subspace are denoted $\delta^{U(i)}$ where $U\in\mathfrak{g}$. Hence

$$[\delta^{U(i)},\delta^{V(j)}]=\delta^{[U,V](i+j)},\quad \text{for } U,V\in\mathfrak{g},\ i,j\in\mathbb{N}. \tag{5.16}$$

If $\{e_1,e_2,...,e_d\}$ is a basis of $\mathfrak{g}$, then $\{\delta^{a(i)} \equiv e_a \otimes t^i,\ a=1,...,d$ and $i \in \mathbb{N}\}$ is a basis of $\hat{\mathfrak{g}} = \mathfrak{g} \otimes \mathbb{R}[t]$. Let Δ be a change of basis acting only on the gradation index, i.e., preserving the d subspaces $\mathfrak{g}_a$:

$$\mathfrak{g}_a = \oplus_{i\in\mathbb{N}} e_a \otimes t^i, \quad a=1,...,d. \tag{5.17}$$

The following lemma will be helpful in disentangling the algebra generated by the δ_R and δ_L for the WZσ model.

Lemma: Let $\hat{\delta}^{a(i)}$ be the elements defined by the following generating function:

$$\hat{\delta}^a(\lambda) \equiv \sum_{i\in\mathbb{N}} \lambda^{-i}\hat{\delta}^{a(i)} = \delta^a((\lambda-\beta)/\alpha)/(1-\beta/\lambda), \quad \alpha,\beta\in\mathbb{R}, \quad \alpha\neq 0. \tag{5.18}$$

Then $\{\delta^{a(i)}\} \overset{\Delta(\alpha,\beta)}{\to} \{\hat{\delta}^{a(i)}\}$ are the only changes of basis sharing the two following properties: (i) each element $\hat{\delta}^{a(i)}$ (resp. $\delta^{a(i)}$) is a finite linear combination of $\delta^{a(i)}$ (resp. of $\hat{\delta}^{a(i)}$), and (ii) the elements $\hat{\delta}^{a(i)}$ verify the same graded commutation rules

$$[\hat{\delta}^{a(i)},\hat{\delta}^{b(j)}] = \hat{\delta}^{[a,b](i+j)},$$

where $[a,b]$ is set for $[e_a,e_b]$.

The proof is straightforward. First notice that for any change of basis Δ satisfying (ii) $\hat{\delta}^{a(0)} = \delta^{a(0)}$. For if $\hat{\delta}^{a(0)} = \Sigma_{i=0}^N \Delta_i^0 \delta^{a(i)}$ with $N\neq 0$ and $\Delta_N^0 \neq 0$, then $[\hat{\delta}^{a(0)},\hat{\delta}^{b(0)}]$ will contain a term $(\Delta_N^0)^2\delta^{[a,b](2N)}$ (different from zero for a certain choice of a and b since $\mathfrak{g}$ is semisimple). This would violate (ii). The condition (ii) also implies that if $\hat{\delta}^{a(1)} = \Sigma_i \Delta_i^1 \delta^{a(i)}$ is given, the whole set $\{\hat{\delta}^{a(i)}\}$ is uniquely defined. Indeed the Δ_j^i are defined recursively on the index i by noting that

$$[\hat{\delta}^{a(i)},\hat{\delta}^{b(1)}] = \hat{\delta}^{[a,b](i+1)}, \quad i\geqslant 1,$$

and hence

$$\Delta_j^{i+1} = \sum_{k=0}^{i+1} \Delta_k^i \Delta_{i+1-k}^1. \tag{5.19}$$

It is then sufficient to study $\hat{\delta}^{a(1)}$. The condition (i) forces $\hat{\delta}^{a(1)}$ to be of the form

$$\hat{\delta}^{a(1)} = \sum_{i=0}^N \Delta_i^1 \delta^{a(i)},$$

for a given $1\leqslant N<\infty$ such that $\Delta_N^1 \neq 0$. The elements $\hat{\delta}^{a(i)}$ will then contain a nonvanishing contribution of $\delta^{a(iN)}$. The condition (i) also requires that $\delta^{a(i)}$ be a finite linear combination of the $\hat{\delta}^{a(j)}$. Suppose then that

$$\delta^{a(1)} = \sum_{j=0}^M (\Delta^{-1})_j^1 \hat{\delta}^{a(j)},$$

for a certain finite M. Since $\hat{\delta}^{a(M)}$ is the only term in the sum to contain $\delta^{a(NM)}$, condition (i) implies that both N and M have to be 1. Hence

$$\hat{\delta}^{a(1)} = \alpha\delta^{a(1)} + \beta\delta^{a(0)}. \tag{5.20}$$

As observed before, the $\hat{\delta}^{a(i)}$, $i\geqslant 2$ are then totally determined. They are found to be

$$\hat{\delta}^{a(i)} = \sum_{j=0}^i \binom{i}{j} \alpha^j \beta^{i-j} \delta^{a(j)}. \tag{5.21}$$

Hence the generating function $\hat{\delta}^a(\lambda)$ is

$$\hat{\delta}^a(\lambda) = \sum_{i=0}^\infty \lambda^{-i} \sum_{j=0}^i \binom{i}{j} \alpha^j \beta^{i-j} \delta^{a(j)}$$

$$= \sum_{j=0}^\infty \frac{\lambda^{-j}\alpha^j\delta^{a(j)}}{j!} \sum_{n=0}^\infty \beta^n \lambda^{-n} \frac{(n+j)!}{n!}.$$

Since

$$\left(\frac{d}{d(\beta/\lambda)}\right)^j \frac{1}{1-\beta/\lambda} = \sum_{n=0}^\infty \frac{(n+j)!}{j!}\left(\frac{\beta}{\lambda}\right)^n$$

the generating function is

$$\hat{\delta}^a(\lambda) = \sum_{j=0}^\infty \frac{\lambda^{-j}\alpha^j\delta^{a(j)}}{j!}\left(\frac{d}{d(\beta/\lambda)}\right)^j \frac{1}{1-\beta/\lambda} = \delta^a((\lambda-\beta)/\alpha)/(1-\beta/\lambda).$$

The fact that Δ is indeed a change of basis—that is, invertible—is obvious. One can convince oneself of the invertibility of Δ either by noticing that, in matrix notation, Δ is an upper triangular matrix with powers of α on the diagonal or by inverting the generating function

$$\delta^a(\lambda) = \hat{\delta}^a(\alpha\lambda+\beta)/(1+\beta/\alpha\lambda), \tag{5.22}$$

which also shows that the group of changes of basis of the form (5.18),

$$\delta^a(\lambda) \overset{\Delta(\alpha,\beta)}{\to} \hat{\delta}^a(\lambda) \tag{5.23}$$

is isomorphic to the upper triangular subgroup of $\mathrm{Sl}(2,\mathbb{R})$ imbedded as

$$\Delta(\alpha,\beta) \leftrightarrow \begin{pmatrix} 1/\alpha & -\beta \\ 0 & \alpha \end{pmatrix}. \tag{5.24}$$

This ends the proof of the lemma. ■

[Note that if condition (i) is relaxed, there are still more changes of basis possible. For these new Δ's, the elements of the new basis might be expressed in terms of finite linear combinations of the initial basis but for the inverse Δ^{-1}, this property will not hold.]

The usefulness of this lemma is that we can change the commutation rule (5.13) without altering the gradation of the subalgebras generated by $\{\delta_R^{T(i)}\}$ and by $\{\delta_L^{T(i)}\}$, respectively. Indeed, let us change the basis in the following way:

$$\hat{\delta}_R^T(\lambda) = \delta_R^T((\kappa+1)\lambda-\kappa)(1-\kappa/(\kappa+1)\lambda)^{-1}, \tag{5.25a}$$

$$\hat{\delta}_L^T(\lambda) = \delta_L^T((\kappa-1)\lambda-\kappa)(1-\kappa/(\kappa-1)\lambda)^{-1}. \tag{5.25b}$$

According to the lemma, the following commutation rules remain the same:

$$[\hat{\delta}_R^U(\lambda),\hat{\delta}_R^V(\lambda')]g = [1/(\lambda'-\lambda)]\{\lambda'\hat{\delta}_R^{[U,V]}(\lambda) - \lambda\hat{\delta}_R^{[U,V]}(\lambda')\}g, \tag{5.26}$$

$$[\hat{\delta}_L^U(\lambda),\hat{\delta}_L^V(\lambda')]g = [1/(\lambda'-\lambda)]\{\lambda'\hat{\delta}_L^{[U,V]}(\lambda) - \lambda\hat{\delta}_L^{[U,V]}(\lambda')\}g. \tag{5.27}$$

However, the commutator $[\hat{\delta}_R^U(\lambda),\hat{\delta}_L^V(\lambda')]g$ is

$$[\hat{\delta}_R^U(\lambda),\hat{\delta}_L^V(\lambda')]g$$
$$= \frac{(\kappa+1)\lambda}{[(\kappa+1)\lambda-\kappa]} \cdot \frac{(\kappa-1)\lambda}{[(\kappa-1))\lambda-\kappa]}$$
$$\times [\delta_R^U((\kappa+1)\lambda-\kappa),\delta_L^V((\kappa-1)\lambda-\kappa)]g$$

and, using (5.13), one gets directly

$$= [1/(\lambda\lambda'-1)]$$
$$\times \{\hat{\delta}_R^{[U,g_0Vg_0^{-1}]}(\lambda) + \hat{\delta}_L^{[g_0^{-1}Ug_0,V]}(\lambda')\}g. \qquad (5.28)$$

Hence the commutation rules (5.26)–(5.28) are the same as in the case of the principal models. From this point on, the discussion is identical to the one of Sec. IV. Hence the final result of this section is that *the infinite-dimensional algebra of infinitesimal transformations for the* WZσ *models with values in* $G=\mathrm{SU}(n)$, $\mathrm{SO}(n)$, *and* $\mathrm{Sp}(n)$ *is* $\mathfrak{g}\oplus(\mathfrak{g}\otimes\mathbb{R}[t,t^{-1}])$, *where* $\mathfrak{g}=\mathrm{su}(n)$, $\mathrm{so}(n)$, *and* $\mathrm{sp}(n)$, *respectively.*

VI. THE σ MODELS WITH VALUES IN A RIEMANNIAN SYMMETRIC SPACE G/H

This final section deals with symmetries for σ models with values in Riemannian symmetric spaces G/H that are not Lie groups. Again, we shall restrict our study to the irreducible RSS's of the classical series. (See Table I of Ref. 6.)

To our knowledge, the problem of finding infinitesimal symmetries for these models has been addressed in two different ways. Ueno[5] considered the SO(3) nonlinear σ model, i.e., the σ model with values in S^2. Modifying the linear system for the SU(2) principal model, he was able to find a genuine linear system for the reduced system—genuine in the sense that it is not an algebraically reduced linear system as we shall use. The problem with his contruction is that is does not seem to be generalizable for other RSS's. (It relies heavily on properties that hold only for 2×2 matrices.) Moreover, the solutions of the linear system are not uniquely determined (even though the linear system is supplemented with normalization conditions) and the correspondence between solutions of the linear system and of the nonlinear model is thus more complicated. It is not clear (to us) whether the algebraic structure carried by the infinitesimal transformations acting on the solution space of the linear system remains the same when "projected" on the solution space of the σ model under consideration.

The second approach has been proposed by Uhlenbeck[22] and the present authors.[23] In Ref. 22, Uhlenbeck uses the constraint $g^2=\mathbf{1}$ to characterize the solution space of the σ models with values in complex Grassmannians as a subset of the solution space of the SU(n) principal model and proceeds to find the subalgebra leaving this constraint satisfied. In Ref. 23, we give an example of a similar construction for the $\mathbf{CP}^{n-1}$ σ models, making explicit the use of the Cartan immersion and hence paving the way towards models with values in other RSS's.

The content of this approach can be outlined as follows. As it was shown before, the solution space of the model with values in G/H appears to be a subset of that of the principal model with values in G. The Cartan immersion provides an algebraic constraint defining this subset. The substance of Mikhaïlov's reduction procedure is to translate this algebraic constraint on the solution space of the linear system. Using the latter, the algebra of infinitesimal transformations for the model with values in G/H appears to be a subalgebra of the algebra found in Sec. IV for the G-valued principal model. As this section will show, this algebra is, however, no longer graded with respect to the original gradation. The commutation rules will be computed and two remarkable graded subalgebras identified.

A. The generators satisfying Mikhaïlov's reduction condition

Any RSS is defined through an involutive automorphism σ of the isometry group G. [The involution σ is one of the four described in (2.13). See Sec. II B.] The (necessary) condition for a field $g(\xi,\eta)$ to lie on the image $\Sigma_0\subset G$ of G/H by the Cartan immersion is

$$\sigma(g)=g^{-1}, \quad \text{for all } (\xi,\eta). \qquad (6.1)$$

On the solutions of the linear systems, this equation reads

$$\sigma(R(\tilde{\lambda}))=L(\lambda), \qquad (6.2)$$

where $\tilde{\lambda}=\lambda$ if σ does not involve a complex conjugation [σ_1 and σ_3 in Eq. (2.13)] and $\tilde{\lambda}=\bar{\lambda}$ if it does (σ_2 and σ_4). This is Mikhaïlov's reduction constraint. [See Eq. (5.8) of Ref. 6 where reductions of the type (6.1) are denoted by σ_-.] Note that the reduction (6.2) relates $R(\lambda)$ and $L(\lambda)$ contrarily to the reductions encountered in Sec. IV which were of the type group–subgroup.

The infinitesimal transformations δg in the algebra for the G/H-valued σ models have to preserve the algebraic constraint (6.1), i.e., they are such that

$$\sigma(g+\delta g)=g^{-1}+\delta g^{-1}, \qquad (6.3)$$

or, since σ is an automorphism,

$$\sigma(\mathbf{1}+(\delta g)g^{-1})=\mathbf{1}+(\delta g^{-1})g,$$

since the original solution is taken to solve the G/H-valued model. Using σ_*, the differential of the involution σ at the identity, the reduction condition for δg finally reads

$$\sigma_*((\delta g)g^{-1})=(\delta g^{-1})g. \qquad (6.4)$$

We shall now calculate the action of σ_* on the generators $(\delta_R^{T(i)}g)g^{-1}$ and $(\delta_L^{T(i)}g)g^{-1}$ of the algebra for the G-valued principal model,

$$\begin{aligned}\sigma_*((\delta_R^T(\lambda)g)g^{-1}) &= \sigma_*(R(\lambda)TR^{-1}(\lambda)),\\ \sigma_*((\delta_L^T(\lambda)g^{-1})g) &= \sigma_*(L(\lambda)TL^{-1}(\lambda)).\end{aligned} \qquad (6.5)$$

If the isometry group is a real group G, $R(\lambda)$ does not belong to G but to G^C, the associated complexified group. All the objects in $\mathfrak{g}$ or G can be understood to be in $\mathfrak{g}^C$ or G^C by the inclusion map and the automorphisms σ_* and σ can be extended in a trivial way. Hence

$$\begin{aligned}\sigma_*((\delta_R^T(\lambda)g)g^{-1}) &= \sigma_*(\mathrm{ad}_{R(\lambda)}T)\\ &= \mathrm{ad}_{\sigma(R(\lambda))}\sigma_*(T)\\ &= \mathrm{ad}_{L(\tilde{\lambda})}\sigma_*(T)\\ &= (\delta_L^{\sigma_*(T)}(\tilde{\lambda})g^{-1})g. \qquad (6.6)_R\end{aligned}$$

A similar calculation leads to

$$\sigma_*((\delta_L^T(\lambda)g)g^{-1}) = (\delta_R^{\sigma_*(T)}(\tilde{\lambda})g^{-1})g. \qquad (6.6)_L$$

Hence for solutions whose values are on Σ_0, the variations δ_R and δ_L are related by the involution σ through Eqs. (6.6).

The solution of (6.4) is simplified if we define on the algebra generated by $\{\delta_R^{T(i)}, \delta_L^{T(i)}, T\in\mathfrak{g}, i\in\mathbb{N}\}$ the following automorphism which will also be denoted by σ_* for obvious reasons:

$$\sigma_*(\delta_R^{T(i)}) \equiv \delta_L^{\sigma_*(T)(i)}, \quad \sigma_*(\delta_L^{T(i)}) \equiv \delta_R^{\sigma_*(T)(i)}, \qquad (6.7)$$

where δ_R and δ_L are to be understood as infinitesimal transformations globally defined on the whole solution space. This automorphism is involutive,

$$\sigma_*^2(\delta_R^{T(i)}) = \delta_R^{T(i)} \quad \text{and} \quad \sigma_*^2(\delta_L^{T(i)}) = \delta_L^{T(i)},$$

thanks to the involutiveness of $\sigma_*: \mathfrak{g}\to\mathfrak{g}$. Hence, its eigenvalues are $+1$ and -1 and

$$(\delta_R^{(T)(i)} \pm \delta_L^{\sigma_*(T)(i)}) \qquad (6.8)$$

are the eigenvectors with eigenvalues ± 1, respectively. Equation (6.4) simply picks out the eigenvectors associated to $+1$. Hence a basis for the algebra $\mathfrak{k}$ preserving condition (6.1), i.e., acting on the solution space of the σ model with values in G/H, is

$$\{\Delta^{T(i)} \equiv \delta_R^{T(i)} + \delta_L^{\sigma_*(T)(i)},\ T\in\mathfrak{g},\ i\in\mathbb{N}\}. \qquad (6.9)$$

Since the $\delta_R^{T(i)}$ and $\delta_L^{T(i)}$ are, respectively, identified with subspaces $(+i)$ and $(-i)$, respectively, the subalgebra spanned by (6.9) does not inherit the gradation of the algebra spanned by the $\delta_R^{T(i)}$ and $\delta_L^{T(i)}$ together. In the next subsection, we compute the commutation rules.

B. The commutation rules $[\Delta^U(\lambda),\Delta^V(\lambda')]g$

The generating functions turn out to be the most useful tool to obtain the commutation rules between the $\Delta^{T(i)}$. Using the generating function $\Delta^T(\lambda)$,

$$\Delta^T(\lambda) \equiv \sum_{i=0}^{\infty} \Delta^{T(i)}\lambda^{-i} = \delta_R^T(\lambda) + \delta_L^{\sigma_*(T)}(\lambda), \qquad (6.10)$$

the commutation rules are

$$\begin{aligned}[\Delta^U(\lambda),\Delta^V(\lambda')]g &= [1/(\lambda'-\lambda)]\{\lambda'\Delta^{[U,V]}(\lambda) - \lambda\Delta^{[U,V]}(\lambda')\}g \\ &\quad + [1/(\lambda\lambda'-1)]\{\delta_R^{[U,g_0\sigma_*(V)g_0^{-1}]}(\lambda) \\ &\quad + \delta_L^{[g_0^{-1}Ug_0,\sigma_*(V)]}(\lambda') \\ &\quad + \delta_R^{[g_0\sigma_*(U)g_0^{-1},V]}(\lambda') + \delta_L^{[\sigma_*(U),g_0^{-1}Vg_0]}(\lambda)\}g.\end{aligned}$$

Clearly, for $g_0\in\Sigma_0$,

$$[\sigma_*(U),g_0^{-1}Vg_0] = \sigma_*[U,g_0\sigma_*(V)g_0^{-1}],$$
$$[g_0\sigma_*(U)g_0^{-1},V] = \sigma_*[g_0^{-1}Ug_0,\sigma_*(V)].$$

Now define the automorphism

$$\sigma_0(X) = g_0\sigma_*(X)g_0^{-1}, \qquad (6.11)$$

which depends on the point in the solution space. Observe moreover that σ_0 is itself an involution:

$$\begin{aligned}\sigma_0^2(X) &= \sigma_0(g_0\sigma_*(X)g_0^{-1}) \\ &= g_0\sigma_*(g_0\sigma_*(X)g_0^{-1})g_0^{-1} \\ &= g_0\sigma(g_0)\sigma_*^2(X)\sigma(g_0^{-1})g_0^{-1} \\ &= X.\end{aligned}$$

The commutation rules can be finally rewritten as

$$\begin{aligned}[\Delta^U(\lambda),\Delta^V(\lambda')]g &= [1/(\lambda'-\lambda)]\{\lambda'\Delta^{[U,V]}(\lambda) - \lambda\Delta^{[U,V]}(\lambda')\}g \\ &\quad + [1/(\lambda\lambda'-1)]\{\Delta^{[U,\sigma_0(V)]}(\lambda) \\ &\quad + \Delta^{[\sigma_0(U),V]}(\lambda')\}g. \qquad (6.12)\end{aligned}$$

Recall that, in Sec. IV, the $\delta_L^{T(i)}$ had been redefined to absorb the explicit dependency of the commutation rules on g_0. This trick is not possible in the present context because the constraint (6.4) forces us to mix the generators δ_R and δ_L in precisely the linear combination (6.9).

The commutation rules (6.12) can be spelled out by expanding around $\lambda = \lambda' = \infty$:

$$[\Delta^{U(i)},\Delta^{V(j)}] = \begin{cases}\Delta^{[U,V](i+j)} + \Delta^{[U,\sigma_0(V)](i-j)}, & i>j\geqslant 1,\\ \Delta^{[U,V](i+j)} + \Delta^{[\sigma_0(U),V](j-i)}, & j>i\geqslant 1,\\ \Delta^{[U,V](2i)} + \Delta^{[U,\sigma_0(V)](0)} + \Delta^{[\sigma_0(U),V](0)}, & i=j\geqslant 1,\\ \Delta^{[U,V](i+j)}, & i=0 \text{ or } j=0.\end{cases} \qquad (6.13)$$

C. Two remarkable graded subalgebras of $\mathfrak{k}$

In order to identify the structure of the algebra $\mathfrak{k}$ spanned by the set (6.9), it is natural to look for a gradation on $\mathfrak{k}$. We have not been able to find such gradations over $\mathbb{N}$ or $\mathbb{Z}$. (Gradations over $\mathbb{Z}_2$ are easy to find but not very informative.) However, we have found two remarkable infinite-dimensional graded subalgebras in $\mathfrak{k}$. In neither case is the gradation extendable to the whole algebra $\mathfrak{k}$ or, at least, not through a change of basis whose elements are *finite* linear combinations of the elements of the other.

Let $\mathfrak{h}_0$ and $\mathfrak{m}_0$ be the eigenspaces of σ_0 associated with $+1$ and -1, respectively. The first graded subalgebra $\mathfrak{s}_1\subset\mathfrak{k}$ is spanned by the generators $\{\Delta^{U(i)}, U\in\mathfrak{h}_0, i\in\mathbb{N}\}$. For this case, the involution σ_0 can be simply dropped out of the generating functions in the commutation rules (6.12). Let us look for a change of basis defined by a generating function

$$\hat{\delta}^U(\lambda) = f(\lambda)\Delta^U(g(\lambda)) \equiv \sum_{i=0}^{\infty}\lambda^{-i}\hat{\delta}^{U(i)}, \qquad (6.14)$$

where $f(\lambda)$ and $g(\lambda)$ are expandable around $\lambda=\infty$ and

such that the subalgebra $\mathcal{S}_1$ will be graded over N,

$$[\hat{\delta}^{U(i)},\hat{\delta}^{V(j)}]=\hat{\delta}^{[U,V](i+j)},\quad \text{for } U,V\in\mathfrak{h}_0,\ i,j\in\mathbf{N}. \tag{6.15}$$

Such commutation rules are summarized in the following relation [see Eq. (4.7)]:

$$[\hat{\delta}^{U}(\lambda),\hat{\delta}^{V}(\lambda')]$$
$$=[1/(\lambda'-\lambda)]\{\lambda'\hat{\delta}^{[U,V]}(\lambda)-\lambda\hat{\delta}^{[U,V]}(\lambda')\}. \tag{6.16}$$

This relation leads to functional equations for $f(\lambda)$ and $g(\lambda)$. Indeed,

$$\begin{aligned}&[\hat{\delta}^{U}(\lambda),\hat{\delta}^{V}(\lambda')]\\&\quad=ff'[\Delta^{U}(g),\Delta^{V}(g')]\\&\quad=\frac{ff'}{(g'-g)(gg'-1)}\\&\qquad\times\{g(g'^2-1)\Delta^{[U,V]}(g)-g'(g^2-1)\Delta^{[U,V]}(g')\}\\&\quad=\frac{1}{(g'-g)(gg'-1)}\{f'g(g'^2-1)\hat{\delta}^{[U,V]}(\lambda)\\&\qquad-g'f(g^2-1)\hat{\delta}^{[U,V]}(\lambda')\},\end{aligned} \tag{6.17}$$

where the following shorthand notation has been used: $f\equiv f(\lambda)$, $f'\equiv f(\lambda')$, $g\equiv g(\lambda)$, and $g'\equiv g(\lambda')$. Comparing (6.16) and (6.17), one gets

$$\frac{1}{\lambda'-\lambda}=\frac{f'g(g'^2-1)}{\lambda'(g'-g)(gg'-1)}=\frac{fg'(g^2-1)}{\lambda(g'-g)(gg'-1)}. \tag{6.18}$$

The most general solutions $f(\lambda)$ and $g(\lambda)$ depend on two constants α and β and read

$$\begin{aligned}f_{\pm}(\lambda)&=\pm\alpha\lambda/\sqrt{(\alpha\lambda+\beta)^2-4},\\g_{\pm}(\lambda)&=(\alpha\lambda+\beta)/2\pm\sqrt{(\alpha\lambda+\beta)^2/4-1}.\end{aligned} \tag{6.19}$$

Since the generating function $\hat{\delta}^{U}(\lambda)$ is to be expanded around $\lambda=\infty$ [see Eq. (6.14)], this requires that the behavior of $g(\lambda)$ as $\lambda\to\infty$ should be linear in λ. However,

$$g_{\pm}(\lambda)\underset{\lambda\to\infty}{\to}(\alpha\lambda)^{\pm 1}, \tag{6.20}$$

which forces the set of solutions $(f_-(\lambda),g_-(\lambda))$ to be discarded. Hence, the solutions $f(\lambda)$ and $g(\lambda)$ are

$$\begin{aligned}f(\lambda)&=\alpha\lambda/\sqrt{(\alpha\lambda+\beta)^2-4},\\g(\lambda)&=(\alpha\lambda+\beta)/2+\sqrt{(\alpha\lambda+\beta)^2/4-1}.\end{aligned} \tag{6.21}$$

The first $\hat{\delta}^{U(i)}$ are, for $\alpha=1$ and $\beta=0$,

$$\begin{aligned}\hat{\delta}^{U(0)}&=\Delta^{U(0)},\\\hat{\delta}^{U(1)}&=\Delta^{U(1)},\\\hat{\delta}^{U(2)}&=\Delta^{U(2)}+\Delta^{U(0)},\\\hat{\delta}^{U(3)}&=\Delta^{U(3)}+3\Delta^{U(1)},\\\hat{\delta}^{U(4)}&=\Delta^{U(4)}+4\Delta^{U(2)}+3\Delta^{U(0)},\\\hat{\delta}^{U(5)}&=\Delta^{U(5)}+5\Delta^{U(3)}+10\Delta^{U(1)}.\end{aligned} \tag{6.22}$$

The inverse of this change of basis can also be written in terms of a generating function

$$\Delta^{U}(\lambda)=F(\lambda)\hat{\delta}^{U}(G(\lambda)) \tag{6.23}$$

again to be expanded around $\lambda=\infty$. The functions $F(\lambda)$ and $G(\lambda)$ are

$$\begin{aligned}F(\lambda)&=(\lambda^2-1)/(\lambda^2-\beta\lambda+1),\\G(\lambda)&=(\lambda^2-\beta\lambda+1)/\alpha\lambda.\end{aligned} \tag{6.24}$$

The generating functions (6.14) and (6.23) have a simple interpretation. Indeed replacing $\hat{\delta}^{U(i)}$ simply by x^i in (6.23) allows us to identify which polynomials have as coefficients those of the change of basis,

$$\begin{aligned}\mathbf{X}(\lambda,x)&=F(\lambda)\sum_{i=0}^{\infty}(G(\lambda))^{-i}x^i\\&=\frac{1-1/\lambda^2}{1/\lambda^2-\beta/\lambda+1}\\&\quad\times\left[1-\frac{\alpha x}{\lambda(1/\lambda^2-\beta/\lambda+1)}\right]^{-1}\\&=\frac{1-1/\lambda^2}{1/\lambda^2-\beta/\lambda-\alpha x/\lambda+1}.\end{aligned} \tag{6.25}$$

Setting $z\equiv 1/\lambda$, $\alpha=1$, and $\beta=0$, this is nothing but the generating function for the Chebyshev polynominals of the first kind $C_n(x)$ (Ref. 24):

$$\sum_{n=0}^{\infty}z^nC_n(x)=\frac{1-z^2}{1-xz+z^2}+1,$$
$$\text{for } -1<x<1 \text{ and } |z|<1. \tag{6.26}$$

[Only the zeroth polynomial $C_0(x)=2$ is different (by a factor 2) from the zeroth polynomials defined by (6.25).] The reason why the Chebyshev polynomials arise in this context is that their recurrence relations have precisely the structure of the commutation rules (6.13) restricted to $\mathcal{S}_1$. Indeed

$$C_i(x)C_j(x)=C_{i+j}(x)+C_{i-j}(x),\quad i\geqslant j. \tag{6.27}$$

Hence the coefficients in (6.22) are the coefficients in the expression of the powers x^i of x in terms of Chebyshev polynomials C_n. (A similar graded subalgebra has been identified for the self-dual Yang–Mills system by Chau and Wu.[25] However, its existence has been proved recursively and, hence, the relationship with the Chebyshev polynomials missed.) Attempts to find a change of basis for the elements of the form $\Delta^{U(i)}$, $U\in\mathfrak{m}_0$, consistent with the gradation (6.15) have failed. It does not seem possible to find a gradation on $\mathfrak{k}$ whose restriction to $\mathcal{S}_1$ is (6.15).

The second graded algebra $\mathcal{S}_2\subset\mathfrak{k}$ is spanned by $\{\Delta^{U(i)}$, $U\in\mathfrak{h}_0$, i even $\in\mathbf{N}$ and $\Delta^{U(j)}$, $U\in\mathfrak{m}_0$, j odd $\in\mathbf{N}\}$. It is also graded over N:

$$[\tilde{\delta}^{U(i)},\tilde{\delta}^{V(j)}]=\tilde{\delta}^{[U,V](i+j)}, \tag{6.28}$$

for $i,j\in\mathbf{N}$ and U and V in $\mathfrak{h}_0$ or $\mathfrak{m}_0$ according to whether i and j are even or odd, respectively. Since the two subspaces $\mathfrak{h}_0$ or $\mathfrak{m}_0$ of $\mathfrak{g}$ alternate in the graded structure, it is harder here to take advantage of the compactness of the formalism of generating functions. However it is straightfoward to transform the results obtained above for $\mathcal{S}_1$. Indeed, let us first introduce tilded Chebyshev polynomials $\tilde{C}_n(x)$ by

$$\tilde{C}_n(x)=i^{-n}C_n(ix). \tag{6.29}$$

Since the $C_n(x)$ are of the same parity as the integer n, the $\tilde{C}_n(x)$'s have real coefficients. Their generating function is

$$(1+z^2)/(1-xz-z^2)+1 \tag{6.30}$$

and their recurrence formula is

$$\tilde{C}_i(x)\tilde{C}_j(x) = \tilde{C}_{i+j}(x) + (-1)^j\tilde{C}_{i-j}(x), \quad i \geqslant j, \tag{6.31}$$

which has exactly the form of the commutation rules (6.13) restricted to $\mathcal{S}_2$:

$$[\Delta^{U(i)},\Delta^{V(j)}] = \Delta^{[U,V](i+j)} + (-1)^j\Delta^{[U,V](i-j)}, \quad i>j\geqslant 1. \tag{6.32}$$

Again, the coefficients α_j^i in the change of basis

$$\tilde{\delta}^{U(i)} = \sum_{j=0}^{i} \alpha_j^i \Delta^{U(j)} \tag{6.33}$$

will be the coefficients in the expression of x^i in terms of the tilded polynomials $\tilde{C}_j(x)$. As before, we did not succeed in defining a change of basis on a complementary subspace of $\mathcal{S}_2$ compatible with the gradation (6.28) on $\mathcal{S}_2$.

Let us recapitulate the results obtained in this section. *A Lie algebra $\mathfrak{k}$ of infinitesimal transformations for the σ model with values in a RSS G/H has been constructed. Even though $\mathfrak{k}$ does not seem to bear any $\mathbb{N}$ or $\mathbb{Z}$ gradation, two graded subalgebras $\mathcal{S}_1 \subset \mathfrak{k}$ and $\mathcal{S}_2 \subset \mathfrak{k}$ have been identified. The first has structure $\mathcal{S}_1 \simeq \mathfrak{h} \otimes \mathbb{R}[t]$ and the second is $\mathcal{S}_2 \simeq (\oplus_{i\in\mathbb{N}} \mathfrak{h} \otimes t^{2i}) \oplus (\oplus_{i\in\mathbb{N}} \mathfrak{m} \otimes t^{2i+1})$, where $\mathfrak{h}$ and $\mathfrak{m}$ are the eigenspaces of the involution σ defining the Riemannian symmetric space G/H with eigenvalues $+1$ and -1, respectively.*

VII. CONCLUSION

One of the hopes triggered by the introduction of infinitely many symmetry transformations by Dolan was that they might give rise to an infinite set of conserved quantities through the Noether theorem and hence be used to solve both the classical and (if there is no anomaly) the quantum theories. A closer analysis shows, however, that this is not quite so. Davies, Houston, Leinaas, and Macfarlane[26] observed indeed that the Noether theorem cannot be naively used because the symmetries are nonlocal. Using a generalized Noether theorem, they concluded that the infinitesimal transformations of $\hat{\mathfrak{g}}_1$ (the subspace $i=1$ in the gradation) are not canonical transformations. Setting these transformations of $\hat{\mathfrak{g}}_1$ in Hamiltonian formalism changes the structure of the loop algebra. We refer the reader to de Vega, Eichenherr, and Maillet[27] for a discussion of the algebraic structure spanned by these transformations.

These infinite-dimensional Lie algebras are more promising when viewed as in Takasaki's approach. There, they might lead to a better understanding of the structure of the solution space of nonlinear σ models. There exist remarkable similarities between this formulation of the SDYM system and the σ models (Sec. III) on the first hand and that of the Kadomtsev–Petviashvili and Korteweg–de Vries equations in the framework of the Kyoto school on the other hand. If these similarities are more than formal, the techniques developed by the Kyoto school could become available and new types of solutions for the σ models be uncovered. But there are many unresolved problems.

First, there is the problem of the formal integration of the σ models. Takasaki[8] had succeeded in giving all formal power series solutions of the SDYM equations. Harnad and one of the authors[19] have extended this result to the supersymmetric ($N=3$) Yang–Mills equations. However, for the σ models, we have not been able to give such a construction.

Second, there is no explicit form of the action for the group associated to these infinitesimal symmetries. According to the work of Ueno and Nakamura[4] (see also Ref. 8), it is natural to think that the action of an element in this group is equivalent to solving a (ξ,η)-dependent Riemann–Hilbert problem. But this is a notoriously difficult problem. To our knowledge, there is no example of solution of a (regular) Riemann–Hilbert problem for either the SDYM equations or for the σ models.

As a third and last point, we would like to underline the possibility of the existence of finite dimensional orbits under the action of this infinite-dimensional group. In the case of the Kadomtsev–Petviashvili equation, these finite-dimensional orbits are known to exist and can be characterized, at least formally. It is fair to think that, for the SDYM equations and/or the σ models on Euclidean space, such orbits exist. On these orbits, only the few first infinitesimal transformations would be linearly independent. Further investigations are necessary to clarify these questions.

ACKNOWLEDGMENTS

The authors wish to thank Guy Arsenault for fruitful comments and are grateful to Michael Forger for useful discussions concerning the Riemann–Hilbert approach to Sec. VI.

One of the authors (Y. S.-A.) was supported in part by the Natural Sciences and Engineering Research Council of Canada and the "Fonds FCAR pour l'aide et le soutien à la recherche."

[1]L. Dolan, "Kac–Moody algebra is hidden symmetry of chiral models," Phys. Rev. Lett. **47**, 1371 (1981).
[2]Y.-S. Wu, "Extension of the hidden symmetry algebra in classical principal chiral models," Nucl. Phys. B **211**, 160 (1983); "The group theoretical aspects of infinitesimal Riemann–Hilbert transform and hidden symmetry," Commun. Math. Phys. **90**, 461 (1983).
[3]M.-L. Ge and Y.-S. Wu, "An explicit approach to the group structure of hidden symmetry," Phys. Lett. B **108**, 411 (1982); C. Devchand and D. B. Fairlie, "A generating function for hidden symmetries of chiral models," Nucl. Phys. B **194**, 232 (1982).
[4]K. Ueno and Y. Nakamura, "The hidden symmetry of chiral fields and the Riemann–Hilbert problem," Phys. Lett. B **117**, 208 (1982).
[5]K. Ueno, "Infinite dimensional Lie algebras acting on chiral fields and the Riemann–Hilbert problem," Publ. RIMS, Kyoto Univ. **19**, 59 (1983).
[6]J. Harnad, Y. Saint-Aubin, and S. Shnider, "Bäcklund transformations for nonlinear sigma models with values in Riemannian symmetric spaces," Commun. Math. Phys. **92**, 329 (1984).
[7]P. Winternitz, "Lie groups and solutions of nonlinear partial differential equations." in *Lecture Notes in Physics*, Vol. 189 (Springer, Berlin, 1983).
[8]K. Takasaki, "A new approach to the self-dual Yang–Mills equations," Commun. Math. Phys. **94**, 35 (1984).
[9]M. Sato, "Soliton equations as dynamical systems on an infinite dimensional Grassmann manifold," RIMS Kyoto Univ. **439**, 30 (1981).
[10]M. Jimbo and T. Miwa, "Solitons and infinite dimensional Lie algebras," Publ. RIMS, Kyoto Univ. **19**, 943 (1983).
[11]A. V. Mikhaïlov, "The reduction problem and the inverse scattering method," Physica D **3**, 73 (1981).
[12]E. Witten, "Non-Abelian bosonization in two dimensions," Commun. Math. Phys. **92**, 455 (1984).
[13]S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces* (Academic, New York, 1978).
[14]H. Eichenherr and M. Forger, "On the dual symmetry of the non-linear

sigma models," Nucl. Phys. B **155**, 381 (1979); "More about non-linear sigma models on symmetric spaces," Nucl. Phys. B **164**, 528 (1980).
[15]J.-P. Antoine and B. Piette, "Classical nonlinear σ models on compact and non-compact Grassman manifolds," preprint, Université Catholique de Louvain, 1986.
[16]V. E. Zakharov and A. V. Mikhaïlov, "Relativistically invariant two-dimensional models of field theory which are integrable by means of the inverse scattering problem method," Zh. Eksp. Teor. Fiz. **74**, 1953 (1978) [Sov. Phys. JETP **47**, 1017 (1978)]; "On the integrability of classical spinor models in two-dimensional space-time," Commun. Math. Phys. **74**, 21 (1980).
[17]H. J. de Vega, "Field theories with an infinite number of conservation laws and Bäcklund transformations in two dimensions," Phys. Lett. B **87**, 233 (1979); "Current Algebras in Sigma Models Including Wess–Zumino Terms," preprint PAR-LPTHE 85-23, 1985.
[18]K. Pohlmeyer, "Integrable Hamiltonian systems and interactions through quadratic constraints," Commun. Math. Phys. **46**, 207 (1976).
[19]J. Harnad and M. Jacques, "Formal power series solutions of supersymmetric ($N = 3$) Yang–Mills equations," J. Math. Phys. **27**, 2394 (1986).
[20]J. Harnad, Y. Saint-Aubin, and S. Shnider, " The soliton correlation matrix and the reduction problem for integrable systems," Commun. Math. Phys. **93**, 33 (1984).
[21]M. Forger and P. Zizzi, "Twisted chiral models with Wess–Zumino term, and strings," Nucl. Phys. B **287**, 131 (1987).
[22]K. Uhlenbeck, "Harmonic maps into Lie groups (classical solutions of the chiral model)," University of Chicago preprint, 1985.
[23]M. Jacques and Y. Saint-Aubin, " Construction of an infinite set of symmetry transformations acting on the solution space of the $\mathbf{CP}^{n-1}$ σ models," in *Proceedings of the XXIInd Karpacz Winter School*, edited by A. Jadczyk (World Scientific, Singapore, 1986).
[24]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[25]L.-L. Chau and Y.-S. Wu, "More about hidden-symmetry algebra for the self-dual Yang–Mills system," Phys. Rev. D **26**, 3581 (1982).
[26]M. C. Davies, P. J. Houston, J. M. Leinaas, and A. J. Macfarlane, "Hidden symmetries as canonical transformations for the chiral model," Phys. Lett. B **119**, 187 (1982).
[27]H. J. de Vega, H. Eichenherr, and J. M. Maillet, "Classical and quantum algebras of non-local charges in σ models," Commun. Math. Phys. **92**, 507 (1984).

Baker–Campbell–Hausdorff relations for the super-Poincaré group

V. Alan Kostelecký
Department of Physics, Indiana University, Bloomington, Indiana 47405 and Theory Division, CERN, 1211 Genève 23, Switzerland

D. Rodney Truax
Department of Chemistry, University of Calgary, Calgary, Alberta, T2N 1N4 Canada

(Received 27 June 1986; accepted for publication 17 June 1987)

Baker–Campbell–Hausdorff relations are obtained for the connected supergroup associated with the super-Poincaré algebra iosp(1|4).

I. INTRODUCTION

The super-Poincaré algebra in its simplest form is an extension of the Poincaré algebra that includes four anticommuting generators, called supersymmetry generators, in addition to the four translation, three rotation, and three boost generators. This superalgebra has been the basis for much research in the past 15 years, especially in high-energy physics, where its representations are involved in the construction of globally supersymmetric field theories, supergravities, and superstring theories. A description of many of these applications may be found in Ref. 1.

Although the basic properties of superalgebras are well established, the mathematically rigorous definition of supergroups as abstract groups and as superanalytic supermanifolds[2] is relatively recent.[3] Since the original papers of Rogers, much progress has been made in elucidating the properties of these supermanifolds[4,5] and supergroups.[6,7] Explicit results for practical calculations exist, however, only for a few of the simpler supergroups. Thus, for example, the supergroup based on the super-Poincaré algebra with four anticommuting generators, called the super-Poincaré group, was first presented within Rogers' formalism in Ref. 3. In this formulation, the supergroup has not been studied extensively, although there are applications to physics making use of Rogers' formalism.[8]

The importance of Roger's supermanifolds stems from their generality, in that they incorporate various earlier supermanifold theories.[9] Working with Rogers' supergroups is advantageous because group elements may be assigned coordinates in ways similar to Lie groups. In Refs. 6 and 7, we have defined and explored three canonical coordinate schemes and related them to matrix supergroups. Although matrix techniques for the super-Poincaré group have also been used,[10] these typically do not take into account Rogers' supermanifold structure.

In Ref. 7, we examined Baker–Campbell–Hausdorff (BCH) relations for simple supergroups.[7,11,12] These formulas link different coordinate schemes for a given Lie group or supergroup. In applications of Lie groups, as, for example, in the theory of coherent states, it is often convenient, for physical reasons, to define the group action in one coordinate scheme but to carry out the group action in a different system that is more accessible computationally. Significant applications of BCH relations to supergroups remain to be made in the study of supercoherent states.

In this paper, we present BCH relations for the super-Poincaré group. The basic forms of the relations are derived as solutions of a set of differential equations obtained by an appropriate method.[7,13] The method was applied previously[7] to two other supergroups, CSQM(2) and CIOSP(1|2); however, the super-Poincaré group has significantly more generators (14) than either of these cases (3 and 5, respectively). Thus, in addition to their intrinsic interest and practical application, the calculations presented here provide further evidence of the viability of the differential equation approach[13] to BCH relations. The differential equation technique lends itself quite naturally to treating infinite-dimensional unitary representations of noncompact supergroups. In principle, finite-dimensional matrix methods could encounter difficulties in correctly defining BCH formulas for such cases.

In Sec. II, we present our notation and conventions both for the super-Poincaré algebra and for the various canonical and noncanonical forms of the supergroup elements that we use. The BCH relations for canonical coordinates in normal sequence are obtained in Sec. III, which also contains a description of the various techniques involved in their derivation. In Sec. IV, we extend these results to BCH relations for canonical coordinates in non-normal sequence. The results are further extended to noncanonical coordinates in Sec. V.

Note that our conventions are those of Refs. 6 and 7, which were based on those of Rogers.[2,3] We remark, in particular, that we do *not* use the summation convention. Also, we do not repeat here basic results for supergroups developed elsewhere. To follow the calculations in detail, the reader will need some familiarity with Ref. 7. However, the BCH relations presented should be readily accessible for practical use.

II. THE SUPER-POINCARÉ ALGEBRA AND THE SUPER-POINCARÉ GROUP

A. The super-Poincaré algebra

The super-Poincaré algebra with four supersymmetry generators may be derived from the simple superalgebra osp(1|4) by an Inönü–Wigner contraction[14] that leaves the usual Poincaré algebra as a subalgebra. The procedure is analogous to that presented in Ref. 7 for obtaining iosp(1|2) from osp(1|2). There exist also other inhomogeneous superalgebras, generically referred to as N-extended super-Poincaré algebras, that may be obtained by contraction from the simple superalgebra osp(N|4). In the remainder of this pa-

 0022-2488/87/102480-08$02.50 © 1987 American Institute of Physics

per, "super-Poincaré algebra" is taken to mean exclusively the case $N = 1$.

In our notation, the nonvanishing graded commutation relations of the super-Poincaré algebra are

$$\{Q_1,Q_3\} = P_1, \quad \{Q_2,Q_3\} = P_3, \tag{2.1}$$
$$\{Q_1,Q_4\} = P_2, \quad \{Q_2,Q_4\} = P_4;$$

$$[J_1,Q_2] = -Q_1, \quad [J_2,Q_1] = \tfrac{1}{2}Q_1, \tag{2.2}$$
$$[J_2,Q_2] = -\tfrac{1}{2}Q_2, \quad [J_3,Q_1] = Q_2;$$

$$[J_4,Q_4] = Q_3, \quad [J_5,Q_3] = -\tfrac{1}{2}Q_3, \tag{2.3}$$
$$[J_5,Q_4] = \tfrac{1}{2}Q_4, \quad [J_6,Q_3] = -Q_4;$$

$$[J_1,P_3] = -P_1, \quad [J_1,P_4] = -P_2, \tag{2.4}$$
$$[J_2,P_1] = \tfrac{1}{2}P_1, \quad [J_2,P_3] = -\tfrac{1}{2}P_3,$$
$$[J_2,P_2] = \tfrac{1}{2}P_2, \quad [J_2,P_4] = -\tfrac{1}{2}P_4,$$
$$[P_3,P_1] = P_3, \quad [P_3,P_2] = P_4;$$

$$[J_4,P_2] = P_1, \quad [J_4,P_4] = P_3, \tag{2.5}$$
$$[J_5,P_1] = -\tfrac{1}{2}P_1, \quad [J_5,P_2] = \tfrac{1}{2}P_2,$$
$$[J_5,P_3] = -\tfrac{1}{2}P_3, \quad [J_5,P_4] = \tfrac{1}{2}P_4,$$
$$[J_6,P_1] = -P_2, \quad [J_6,P_3] = -P_4;$$

$$[J_1,J_2] = -J_1, \quad [J_4,J_5] = J_4, \tag{2.6}$$
$$[J_1,J_3] = -2J_2, \quad [J_4,J_6] = 2J_5,$$
$$[J_2,J_3] = -J_3, \quad [J_5,J_6] = J_6.$$

The Q_m, $m = 1,...,4$, generate supersymmetry transformations, the P_l, $l = 1,...,4$, generate translations, and the J_n, $n = 1,...,6$, generate rotations and boosts. These generators satisfy the following adjoint relations[1]:

$$Q_1^\dagger = -Q_3, \quad Q_2^\dagger = -Q_4, \quad Q_3^\dagger = -Q_1, \quad Q_4^\dagger = -Q_2,$$
$$P_1^\dagger = P_1, \quad P_2^\dagger = P_3, \quad P_3^\dagger = P_2, \quad P_4^\dagger = P_4, \tag{2.7}$$
$$J_1^\dagger = J_4, \quad J_2^\dagger = J_5, \quad J_3^\dagger = J_6,$$
$$J_4^\dagger = J_1, \quad J_5^\dagger = J_2, \quad J_6^\dagger = J_3,$$

which are essential for computing unitary supergroup operators.

Several different conventions and notations for the super-Poincaré algebra exist in the literature. In Eqs. (2.1)–(2.7), we have adopted the *conventions* of Ref. 1, Chap. 3. To make formulas and calculations less cumbersome, however, our *notation* is different. The notation of Ref. 1 may be recovered by effectuating the following replacements:

$$Q_1 \leftrightarrow Q_+, \quad P_1 \leftrightarrow P_{+\dot{+}}, \quad J_1 \leftrightarrow J_{++}, \quad J_4 \leftrightarrow J_{\dot{+}\dot{+}},$$
$$Q_2 \leftrightarrow Q_-, \quad P_2 \leftrightarrow P_{+\dot{-}}, \quad J_2 \leftrightarrow J_{+-}, \quad J_5 \leftrightarrow J_{\dot{+}\dot{-}}, \tag{2.8}$$
$$Q_3 \leftrightarrow Q_{\dot{+}}, \quad P_3 \leftrightarrow P_{-\dot{+}}, \quad J_3 \leftrightarrow J_{--}, \quad J_6 \leftrightarrow J_{\dot{-}\dot{-}},$$
$$Q_4 \leftrightarrow Q_{\dot{-}}, \quad P_4 \leftrightarrow P_{-\dot{-}}.$$

The conventions of Ref. 1 are especially convenient for calculational purposes. Note, however, that the operators P_l and J_n are *not* the usual momentum and angular-momentum physical observables. The P_l are light-cone-type variables defined[1] as linear combinations of the standard four-momenta. Furthermore, as is usual practice, the Hermitian generators of the Lorentz algebra so(3,1) have been linearly combined to form the J_n. We remark that any BCH relation presented in this paper may readily be converted to an expression in terms of the standard four-momenta, using Lemma 2 of Ref. 7. This is possible because the P_l form an Abelian subalgebra of iosp(1|4).

B. The super-Poincaré group

Associated with the super-Poincaré algebra is a connected supergroup whose elements may be obtained (see Theorem 2 of Ref. 6) by exponentiation of the superalgebra generators with Grassmann-valued variables. We refer to this supergroup as the super-Poincaré group.

The precise form of the supergroup elements depends upon the parametrization scheme adopted for the exponentiation. For supergroups, there exist three kinds of canonical coordinates.[6,7] For elements $\mathbf{g}$ of the super-Poincaré group, canonical coordinates of the first kind take the form

$$\mathbf{g}_{\mathrm{I}} = \exp\left(\sum_{l=1}^{4} p^l P_l + \sum_{m=1}^{4} q^m Q_m + \sum_{n=1}^{6} j^n J_n\right), \tag{2.9}$$

where the p^l and j^n are variables taking values in the even part ${}^0B_L$ of a Grassmann algebra B_L over $\mathbb{C}^L$, and where the q^m are variables taking values in the odd part ${}^1B_L$. For a complete description of this construction and the discussion that follows for the case of a general supergroup, the reader should consult Refs. 6 and 7. Note that a Hermitian basis for the Grassmann algebra is established in Kostelecký and Rabin.[8]

Canonical coordinates of the second kind for a supergroup are constructed[6] in terms of exponentials of the individual basis elements of the Lie algebra that is associated with the superalgebra in question. However, because each Q_m anticommutes with itself, it turns out that these canonical coordinates are identical to the canonical coordinates of the third kind (corollary to Theorem 1 of Ref. 7). Therefore, in this paper we proceed directly to consideration of canonical coordinates of the third kind, which for elements $\mathbf{g}$ of the super-Poincaré group take the form

$$\mathbf{g}_{\mathrm{III}} = \prod_{l=1}^{4} \exp(\alpha^l P_l) \prod_{m=l}^{4} \exp(\beta^m Q_m) \prod_{n=1}^{6} \exp(\gamma^n J_n). \tag{2.10}$$

Here, $\alpha^l, \gamma^n \in {}^0B_L$ and $\beta^m \in {}^1B_L$. Note that the order of appearance of the 14 exponentials in Eq. (2.10) is important; by definition, we say that this sequence is "normal."[7] Other non-normal sequences are possible and, indeed, will appear in the subsequent sections.

In addition to the canonical coordinate schemes, various noncanonical parametrizations may also be defined. In Sec. V, we shall consider extensions of the BCH relations to certain noncanonical coordinate schemes. For elements $\mathbf{g}$ of the super-Poincaré group, these noncanonical coordinates take the form

$$\mathbf{g}_{\mathrm{NC}} = \exp\left(\sum_{l=1}^{4} a^l P_l\right) \exp\left(\sum_{m=1}^{4} b^m Q_m\right) \exp\left(\sum_{n=1}^{6} c^n J_n\right), \tag{2.11}$$

where $a^l, c^n \in {}^0B_L$ and $b^m \in {}^1B_L$. Again, this is defined as the normal sequence; several non-normal sequences may also be introduced.

To obtain unitary supergroup operators in terms of canonical coordinates of the first kind, Eq. (2.9), we require the condition

$$\mathbf{g}_I^\dagger = \mathbf{g}_I^{-1}. \tag{2.12}$$

Computing the adjoint and the inverse of (2.9) and using the properties (2.7) of the generators, we obtain the conditions on the Grassmann parameters,

$$\begin{aligned} &q^3 = -(q^1)^*, \quad q^4 = -(q^2)^*, \\ &p^1 = -(p^1)^*, \quad p^3 = -(p^2)^*, \quad p^4 = -(p^4)^*, \\ &j^4 = -(j^1)^*, \quad j^5 = -(j^2)^*, \quad j^6 = -(j^3)^*. \end{aligned} \tag{2.13}$$

Note that p^1 and p^4 are pure imaginary as a consequence of the fact that P_1 and P_4 are self-adjoint.

Thus requiring that the parameters p^l, q^m, j^n of Eq. (2.9) satisfy Eq. (2.13) ensures that the supergroup elements are unitary. In the remainder of the paper, the constraints (2.13) are *not* assumed. However, they may be imposed if desired.

III. BCH RELATIONS FOR CANONICAL COORDINATES IN NORMAL SEQUENCE

Here, we construct explicitly the BCH relations between canonical coordinates of the first and third kinds using the differential equation method expounded in Ref. 7. Only canonical coordinates of the third kind in normal sequence are considered in this section.

Applying the general method of Ref. 7, we introduce a real parameter t and write

$$\begin{aligned} &\exp\left[t\left(\sum_{l=1}^{4} p^l P_l + \sum_{m=1}^{4} q^m Q_m + \sum_{n=1}^{6} j^n J_n\right)\right] \\ &\quad = \prod_{l'=1}^{4} \exp(\alpha^{l'} P_{l'}) \prod_{m'=1}^{4} \exp(\beta^{m'} Q_{m'}) \\ &\qquad \times \prod_{n'=1}^{6} \exp(\gamma^{n'} J_{n'}), \end{aligned} \tag{3.1}$$

where $\alpha^{l'}, \beta^{m'}, \gamma^{n'}$ are taken as unknown functions of p^l, q^m, j^n, and t:

$$\begin{aligned} &\alpha^{l'} = \alpha^{l'}(p^l, q^m, j^n, t) \in {}^0B_L, \\ &\beta^{m'} = \alpha^{m'}(p^l, q^m, j^n, t) \in {}^1B_L, \\ &\gamma^{n'} = \gamma^{n'}(p^l, q^m, j^n, t) \in {}^0B_L, \\ &\qquad p^l, j^n \in {}^0B_L, \quad q^m \in {}^1B_L, \quad t \in \mathbb{R}. \end{aligned} \tag{3.2}$$

For $t = 1$, Eq. (3.1) and knowledge of the functions in Eq. (3.2) together form the desired BCH relation.

We can obtain the explicit form of Eqs. (3.2) by solving a set of 14 coupled linear first-order differential equations in the real parameter t. These equations are obtained by differentiating Eq. (3.1), simplifying, and equating coefficients of the superalgebra generators. The appropriate boundary conditions[7] are $\alpha^l(0) = \beta^m(0) = \gamma^n(0) = 0$. The details of the process for obtaining the differential equations have been discussed in Ref. 7; the present case is a straightforward application of those techniques.

The methods yield the following 14 differential equations, where a dot above a symbol means differentiation with respect to t:

$$\dot\gamma^6 e^{\gamma^5} = j^6, \tag{3.3}$$

$$\dot\gamma^5 + 2\gamma^4 \dot\gamma^6 e^{\gamma^5} = j^5, \tag{3.4}$$

$$\dot\gamma^4 + \gamma^4 \dot\gamma^5 + (\gamma^4)^2 \dot\gamma^6 e^{\gamma^5} = j^4, \tag{3.5}$$

$$\dot\gamma^3 e^{-\gamma^2} = j^3, \tag{3.6}$$

$$\dot\gamma^2 - 2\gamma^1 \dot\gamma^3 e^{-\gamma^2} = j^2, \tag{3.7}$$

$$\dot\gamma^1 - \gamma^1 \dot\gamma^2 + (\gamma^1)^2 \dot\gamma^3 e^{-\gamma^2} = j^1, \tag{3.8}$$

$$\dot\beta^4 - \tfrac{1}{2}\beta^4 \dot\gamma^5 + \dot\gamma^6 e^{\gamma^5}(\beta^3 - \gamma^4 \beta^4) = q^4, \tag{3.9}$$

$$\begin{aligned} &\dot\beta^3 + \dot\gamma^5(\tfrac{1}{2}\beta^3 - \gamma^4\beta^4) \\ &\quad + \dot\gamma^6 e^{\gamma^5}\gamma^4(\beta^3 - \gamma^4\beta^4) - \beta^4 \dot\gamma^4 = q^3, \end{aligned} \tag{3.10}$$

$$\dot\beta^2 + \tfrac{1}{2}\dot\gamma^2 \beta^2 - \dot\gamma^3 e^{-\gamma^2}(\gamma^1\beta^2 + \beta^1) = q^2, \tag{3.11}$$

$$\begin{aligned} &\dot\beta^1 + \dot\gamma^1\beta^2 - \dot\gamma^2(\gamma^1\beta^2 + \tfrac{1}{2}\beta_1) \\ &\quad + \dot\gamma^3 e^{-\gamma^2}\gamma^1(\beta_1 + \gamma^1\beta_2) = q^1, \end{aligned} \tag{3.12}$$

$$\begin{aligned} &\dot\alpha^4 - \beta^2\dot\beta^4 + \tfrac{1}{2}\dot\gamma^2\alpha^4 \\ &\quad - \dot\gamma^3 e^{-\gamma^2}(\alpha^2 + \gamma^1\alpha^4) + \tfrac{1}{2}\dot\gamma^5(\beta^2\beta^4 - \alpha^4) \\ &\quad + \dot\gamma^6 e^{\gamma^5}[\alpha^3 - \beta^2\beta^3 + \gamma^4(\beta^2\beta^4 - \alpha^4)] = p^4, \end{aligned} \tag{3.13}$$

$$\begin{aligned} &\dot\alpha^3 - \beta^2\dot\beta^3 + \tfrac{1}{2}\dot\gamma^2\alpha^3 \\ &\quad - \dot\gamma^3 e^{-\gamma^2}(\alpha^1 + \gamma^1\alpha^3) + \dot\gamma^4(\beta^2\beta^4 - \alpha^4) + \tfrac{1}{2}\dot\gamma^5 \\ &\quad \times [\alpha^3 - \beta^2\beta^3 + 2\gamma^4(\beta^2\beta^4 - \alpha^4)] + \dot\gamma^6 e^{\gamma^5}\gamma^4 \\ &\quad \times [\alpha^3 - \beta^2\beta^3 + \gamma^4(\beta^2\beta^4 - \alpha^4)] = p^3, \end{aligned} \tag{3.14}$$

$$\begin{aligned} &\dot\alpha^2 - \beta^1\dot\beta^4 + \alpha^4\dot\gamma^1 \\ &\quad - \dot\gamma^2(\tfrac{1}{2}\alpha^2 + \gamma^1\alpha^4) + \dot\gamma^3\gamma^1 e^{-\gamma^2}(\alpha^2 + \gamma^1\alpha^4) \\ &\quad + \tfrac{1}{2}\dot\gamma^5(\beta^1\beta^4 - \alpha^2) + \dot\gamma^6 e^{\gamma^5} \\ &\quad \times [\alpha^1 - \beta^1\beta^3 + \gamma^4(\beta^1\beta^4 - \alpha^2)] = p^2, \end{aligned} \tag{3.15}$$

$$\begin{aligned} &\dot\alpha^1 - \beta^1\dot\beta^3 + \alpha^3\dot\gamma^1 - \dot\gamma^2(\tfrac{1}{2}\alpha^1 + \gamma^1\alpha^3) \\ &\quad + \dot\gamma^3\gamma^1 e^{-\gamma^2}(\alpha^1 + \gamma^1\alpha^3) + \dot\gamma^4(\beta^1\beta^4 - \alpha^2) + \tfrac{1}{2}\dot\gamma^5 \\ &\quad \times [\alpha^1 - \beta^1\beta^3 + 2\gamma^4(\beta^1\beta^4 - \alpha^2)] + \dot\gamma^6 e^{\gamma^5}\gamma^4 \\ &\quad \times [\alpha^1 - \beta^1\beta^3 + \gamma^4(\beta^1\beta^4 - \alpha^2)] = p^1. \end{aligned} \tag{3.16}$$

The remainder of this section is concerned with solving these equations.

In fact, each of the ordered sets of Eqs. {(3.3)–(3.5)}, {(3.6)–(3.8)}, {(3.9)–(3.10)}, and {(3.11)–(3.12)} is self-coupled and can be solved using only the solutions to the sets occurring earlier in the ordered sequence, as we show below. This occurs because the J_n form a direct-product Lie subalgebra, because the Q_m commute with the P_l, and because Q_1, Q_3 anticommute with Q_2, Q_4, respectively.

The differential equations thus admit a natural sequence for solution. We outline first the solution of Eqs. (3.3)–(3.5). Substitution of (3.3) and (3.4) into (3.5) yields a Riccati equation[15] for γ^4:

$$\dot\gamma^4 + j^5\gamma^4 - j^6(\gamma^4)^2 = j^4. \tag{3.17}$$

The solution may be shown by standard methods[13] to be

$$\gamma^4 = j^4 \sinh(K_2 t)/[K_2 \cosh(K_2 t) + \tfrac{1}{2} j^5 \sinh(K_2 t)], \tag{3.18}$$

where

$$\Delta_2^2 = |\tfrac{1}{4}(j^5)^2 - j^4 j^6|,$$
$$\sigma_2 = \mathrm{sgn}[\tfrac{1}{4}(j^5)^2 - j^4 j^6],\ \hat{\sigma}_2 = +\sqrt{\sigma_2}, \tag{3.19}$$
$$K_2 = \hat{\sigma}_2 \Delta_2,$$

and where we used the initial condition $\gamma^4(0) = 0$. Then, the solutions for γ^5 and γ^6 can be obtained by straightforward integration of Eq. (3.4) followed by (3.3), again using the initial conditions:

$$\gamma^5 = 2\ln[\cosh(K_2 t) + \tfrac{1}{2} j^5 K_2^{-1} \sinh(K_2 t)], \tag{3.20}$$
$$\gamma^6 = j^6 \sinh(K_2 t)/[K_2 \cosh(K_2 t) + \tfrac{1}{2} j^5 \sinh(K_2 t)]. \tag{3.21}$$

By inspection, the corresponding solutions γ^1, γ^2, γ^3 to Eqs. (3.6)–(3.8) are obtained by substituting $j^6 \leftrightarrow j^3$, $j^5 \leftrightarrow -j^2$, $j^4 \leftrightarrow j^1$, $\gamma^6 \leftrightarrow \gamma^3$, $\gamma^5 \leftrightarrow -\gamma^2$, $\gamma^4 \leftrightarrow \gamma^1$. Explicitly, this yields

$$\gamma^1 = j^1 \sinh(K_1 t)/[K_1 \cosh(K_1 t) - \tfrac{1}{2} j^2 \sinh(K_1 t)], \tag{3.22}$$
$$\gamma^2 = -2\ln[\cosh(K_1 t) - \tfrac{1}{2} j^2 K_1^{-1} \sinh(K_1 t)], \tag{3.23}$$
$$\gamma^3 = j^3 \sinh(K_1 t)/[K_1 \cosh(K_1 t) - \tfrac{1}{2} j^2 \sinh(K_1 t)], \tag{3.24}$$

where

$$\Delta_1^2 = |\tfrac{1}{4}(j^2)^2 - j^1 j^3|$$
$$\sigma_1 = \mathrm{sgn}[\tfrac{1}{4}(j^2)^2 - j^1 j^3],\quad \hat{\sigma}_1 = +\sqrt{\sigma_1}, \tag{3.25}$$
$$K_1 = \hat{\sigma}_1 \Delta_1.$$

The results (3.18) and (3.20)–(3.24) are the same as those yielded for a BCH relation of the direct-product Lie algebra su(1,1) ⊗ su(1,1) (see, for example, Ref. 13). This occurs because the J_m generate this Lie subalgebra of the super-Poincaré algebra.

Next, we turn to the solution of Eqs. (3.9) and (3.10). By substitution for $\dot{\gamma}^4$, $\dot{\gamma}^5$, $\dot{\gamma}^6$ from Eqs. (3.3)–(3.5), we obtain

$$\dot{\beta}^4 - \tfrac{1}{2} j^5 \beta^4 + j^6 \beta^3 = q^4, \tag{3.26}$$
$$\dot{\beta}^3 + \tfrac{1}{2} j^5 \beta^3 - j^4 \beta^4 = q^3. \tag{3.27}$$

Solving Eq. (3.26) for β^3 and using the result together with its t derivative in Eq. (3.27) yields a second-order equation for $\ddot{\beta}^4$. The solution is straightforward, as is the subsequent solution of (3.26). Fixing the four integration constants by using the initial conditions and by requiring consistency of the solutions with Eqs. (3.26) and (3.27), we find

$$\beta^4 = q^4 K_2^{-1} \sinh(K_2 t) + K_2^{-2}(\tfrac{1}{2} j^5 q^4 - j^6 q^3)(\cosh(K_2 t) - 1), \tag{3.28}$$
$$\beta^3 = q^3 K_2^{-1} \sinh(K_2 t) - K_2^{-2}(\tfrac{1}{2} j^5 q^3 - j^4 q^4)(\cosh(K_2 t) - 1). \tag{3.29}$$

In a similar manner, Eqs. (3.11) and (3.12) may be reexpressed as

$$\dot{\beta}^2 + \tfrac{1}{2} j^2 \beta^2 - j^3 \beta^1 = q^2, \tag{3.30}$$
$$\dot{\beta}^1 - \tfrac{1}{2} j^2 \beta^1 + j^1 \beta^2 = q^1. \tag{3.31}$$

Making the substitutions $j^6 \leftrightarrow j^3$, $j^5 \leftrightarrow -j^2$, $j^4 \leftrightarrow j^1$, $\beta^4 \leftrightarrow \beta^2$, $\beta^3 \leftrightarrow -\beta^1$, $q^4 \leftrightarrow q^2$, $q^3 \leftrightarrow -q^1$ enables us to write the solutions by inspection:

$$\beta^2 = q^2 K_1^{-1} \sinh(K_1 t) - K_1^{-2}(\tfrac{1}{2} j^2 q^2 - j^3 q^1)(\cosh(K_1 t) - 1), \tag{3.32}$$
$$\beta^1 = q^1 K_1^{-1} \sinh(K_1 t) + K_1^{-2}(\tfrac{1}{2} j^2 q^1 - j^1 q^2)(\cosh(K_1 t) - 1). \tag{3.33}$$

We are left with the four coupled differential Eqs. (3.13)–(3.16) for $\dot{\alpha}^1$, $\dot{\alpha}^2$, $\dot{\alpha}^3$, $\dot{\alpha}^4$. As for the other variables, these equations may be simplified by substitution of Eqs. (3.3)–(3.12) for $\dot{\gamma}^n$ and $\dot{\beta}^m$. Introducing the four-component column vector

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}, \tag{3.34}$$

these four equations may be written in abbreviated form as

$$\dot{\boldsymbol{\alpha}} = \mathbb{P}\boldsymbol{\alpha} + \mathbf{r}(t). \tag{3.35}$$

Here, $\mathbb{P}$ is the constant 4×4 matrix

$$\mathbb{P} = \begin{pmatrix} j^- & j^4 & -j^1 & 0 \\ -j^6 & j^+ & 0 & -j^1 \\ j^3 & 0 & -j^+ & j^4 \\ 0 & j^3 & -j^6 & -j^- \end{pmatrix}, \tag{3.36}$$

where

$$j^\pm = \tfrac{1}{2}(j^2 \pm j^5), \tag{3.37}$$

and where $\mathbf{r}(t)$ is the four-component vector given by

$$r^1 = \beta^1 q^3 + p^1,\quad r^2 = \beta^1 q^4 + p^2,$$
$$r^3 = \beta^2 q^3 + p^3,\quad r^4 = \beta^2 q^4 + p^4. \tag{3.38}$$

Equation (3.35) has as general solution[16]

$$\boldsymbol{\alpha} = \Psi(t)\mathbf{c} + \Psi(t)\int_0^t \Psi^{-1}(s)\mathbf{r}(s)\,ds. \tag{3.39}$$

The constant vector $\mathbf{c}$ is to be determined by the initial conditions. Also, $\Psi(t)$ is a fundamental 4×4 matrix solution,[16] i.e., $\Psi(t)$ is a 4×4 matrix whose columns are four linearly independent solutions to the homogeneous equations $\dot{\boldsymbol{\alpha}} = \mathbb{P}\boldsymbol{\alpha}$. Since $\Psi(0) \neq 0$, as we argue below, the initial condition $\boldsymbol{\alpha}(0) = 0$ implies $\mathbf{c} = 0$.

A fundamental matrix solution to the homogeneous equation $\dot{\boldsymbol{\alpha}} = \mathbb{P}\boldsymbol{\alpha}$ may be constructed by applying the trial solution $\boldsymbol{\alpha} = \mathbf{A}e^{\lambda t}$ and diagonalizing the resulting matrix of coefficients for $\mathbf{A}$. The method is described in detail in Ref. 16. We obtain

$$\Psi(t) = \begin{pmatrix} j^1 j^4 e^{\lambda_1 t} & j^1 j^4 e^{\lambda_2 t} & j^1 j^4 e^{\lambda_3 t} & j^1 j^4 e^{\lambda_4 t} \\ j^1 \phi^5_+ e^{\lambda_1 t} & -j^1 \phi^5_- e^{\lambda_2 t} & -j^1 \phi^5_- e^{\lambda_3 t} & j^1 \phi^5_+ e^{\lambda_4 t} \\ -j^4 \phi^2_- e^{\lambda_1 t} & j^4 \phi^2_+ e^{\lambda_2 t} & -j^4 \phi^2_- e^{\lambda_3 t} & j^4 \phi^2_+ e^{\lambda_4 t} \\ -\phi^2_- \phi^5_+ e^{\lambda_1 t} & -\phi^2_+ \phi^5_- e^{\lambda_2 t} & \phi^2_- \phi^5_- e^{\lambda_3 t} & \phi^2_+ \phi^5_+ e^{\lambda_4 t} \end{pmatrix}, \tag{3.40}$$

where

$$\lambda_1 = -\lambda_2 = K_1 + K_2, \quad \lambda_3 = -\lambda_4 = K_1 - K_2, \tag{3.41}$$

and

$$\phi^2_{\pm} = K_1 \pm \tfrac{1}{2} j^2, \quad \phi^5_{\pm} = K_2 \pm \tfrac{1}{2} j^5. \tag{3.42}$$

The inverse fundamental matrix may be calculated as

$$\Psi^{-1}(t) = (4K_1K_2 j^1 j^4)^{-1} \begin{pmatrix} \phi^2_+ \phi^5_- e^{-\lambda_1 t} & j^4 \phi^2_+ e^{-\lambda_1 t} & -j^1 \phi^5_- e^{-\lambda_1 t} & -j^1 j^4 e^{-\lambda_1 t} \\ \phi^2_- \phi^5_+ e^{-\lambda_2 t} & -j^4 \phi^2_- e^{-\lambda_2 t} & j^1 \phi^5_+ e^{-\lambda_2 t} & -j^1 j^4 e^{-\lambda_2 t} \\ \phi^2_+ \phi^5_+ e^{-\lambda_3 t} & -j^4 \phi^2_+ e^{-\lambda_3 t} & -j^1 \phi^5_+ e^{-\lambda_3 t} & j^1 j^4 e^{-\lambda_3 t} \\ \phi^2_- \phi^5_- e^{-\lambda_4 t} & j^4 \phi^2_- e^{-\lambda_4 t} & j^1 \phi^5_- e^{-\lambda_4 t} & j^1 j^4 e^{-\lambda_4 t} \end{pmatrix}. \tag{3.43}$$

With expressions (3.38), (3.40), and (3.43), the general solution for α may now be obtained from Eq. (3.39). The integrations are straightforward. After some algebra, the expression for α^1 takes the form

$$\begin{aligned} \alpha^1 = {} & 2K_1 k^{-2} q^1 C^2 \sinh(K_1 t) + k^{-2} C^1 C^2 \cosh(K_1 t) + k^{-2} K_+^{-1} [K_1 \cosh(K_+ t) + K_2][q^1 q^3 k - \tfrac{1}{2} C^1 C^2] \\ & + k^{-2} K_-^{-1} [K_1 \cosh(K_- t) + K_2][-q^1 q^3 k - \tfrac{1}{2} C^1 C^2] + k^{-2} K_1 K_+^{-1} \sinh(K_+ t)[K_2 C^1 q^3 - K_1 q^1 C^2] \\ & - k^{-2} K_1 K_-^{-1} \sinh(K_- t)[K_2 C^1 q^3 + K_1 q^1 C^2] + \tfrac{1}{2} k^{-1} K_+^{-1} [1 - \cosh(K_+ t)][K_1 D^1 - K_2 D^2] \\ & - \tfrac{1}{2} k^{-1} K_-^{-1} [1 - \cosh(K_- t)][K_1 D^1 + K_2 D^2] + \tfrac{1}{2} k^{-1} K_+^{-1} \sinh(K_+ t)[p^1 k - \tfrac{1}{2}(D^3 j^5 - 2D^4 j^4)] \\ & + \tfrac{1}{2} k^{-1} K_-^{-1} \sinh(K_- t)[p^1 k + \tfrac{1}{2}(D^3 j^5 - 2D^4 j^4)]. \end{aligned} \tag{3.44}$$

In this expression, we have defined

$$k = 2K_1K_2, \quad K_{\pm} = K_1 \pm K_2, \tag{3.45}$$

and

$$\begin{aligned} C^1 &= q^1 j^2 - 2q^2 j^1, \quad C^2 = q^3 j^5 - 2q^4 j^4, \\ D^1 &= p^1 j^5 - 2p^2 j^4, \quad D^2 = p^1 j^2 - 2p^3 j^1, \\ D^3 &= p^1 j^2 - 2p^3 j^1, \quad D^4 = p^2 j^2 - 2p^4 j^1. \end{aligned} \tag{3.46}$$

The reader should note that C^1 and C^2, like the q^m, are odd Grassmann variables.

Similarly, the solutions for α^2, α^3, and α^4 may be obtained. Due to the structure of the differential equations (3.35), these solutions can be expressed in the form of Eq. (3.44) with simple parameter substitutions. If we write explicitly the functional dependence of α^1 in Eq. (3.44) as

$$\alpha^1 = \alpha^1(p^1,p^2,p^3,p^4;q^1,q^2,q^3,q^4;j^1,j^2,j^3,j^4,j^5,j^6), \tag{3.47}$$

then the solutions α^2, α^3, α^4 may be written

$$\begin{aligned} \alpha^2 = {} & \alpha^1(p^2,p^1,p^4,p^3;q^1,q^2,q^4,q^3; \\ & j^1,j^2,j^3,-j^6,-j^5,-j^4), \end{aligned} \tag{3.48}$$

$$\begin{aligned} \alpha^3 = {} & \alpha^1(p^3,p^4,p^1,p^2;q^2,q^1,q^3,q^4; \\ & -j^3,-j^2,-j^1,j^4,j^5,j^6), \end{aligned} \tag{3.49}$$

$$\begin{aligned} \alpha^4 = {} & \alpha^1(p^4,p^3,p^2,p^1;q^2,q^1,q^4,q^3; \\ & -j^3,-j^2,-j^1,-j^6,-j^5,-j^4), \end{aligned} \tag{3.50}$$

Note in particular that under any of these parameter changes K_1 and K_2 remain unaffected.

In summary, we have obtained in this section the BCH relation between canonical coordinates of the first and third kinds, in normal sequence. The relation is of the form of Eqs. (3.1) and (3.2) with $t = 1$, where the 14 equations in (3.2) are given explicitly by Eqs. (3.18), (3.20)–(3.24), (3.28), (3.29), (3.32), (3.33), (3.44), and (3.48)–(3.50). We remark that the results may also be applied in the limit in which any of the parameters p^l, q^m, j^n are taken to zero.

IV. BCH RELATIONS FOR CANONICAL COORDINATES IN NON-NORMAL SEQUENCE

In this section, we construct BCH relations between canonical coordinates of the first and third kinds for certain non-normal sequences. Although we could proceed as in Sec. III, obtaining and solving sets of 14 differential equations for each BCH relation, it is simpler to proceed by relating the various non-normal sequences for canonical coordinates of the third kind to the normal sequence. Combined with the BCH relation already obtained, these results will yield BCH relations of the desired type.

We begin by recalling that the P_l commute among themselves and with the Q_m. Furthermore, Q_1, Q_3 anticommute with Q_2,Q_4, respectively. Therefore, by Lemma 2 of Ref. 7, the explicit solutions of the form (3.2) that we have obtained in Sec. III will be valid for a BCH relation of the form (3.1) but with the exponentials involving $P_{l'}$ and $Q_{m'}$ taken in any order, *provided* all exponentials with $P_{l'}$ and $Q_{m'}$ are to the left of those with $J_{n'}$ and *provided* the exponentials with Q_1, Q_2 appear to the left of those with Q_3 and Q_4. In particular, this means that the BCH relation

$$\begin{aligned} & \exp\left(\sum_{l=1}^{4} p^l P_l + \sum_{m=1}^{4} q^m Q_m + \sum_{n=1}^{6} j^n J_n\right) \\ & \quad = \prod_{m'=1}^{4} \exp(\beta^{m'} Q_{m'}) \prod_{l'=1}^{4} \exp(\alpha^{l'} P_{l'}) \prod_{n'=1}^{6} \exp(\gamma^{n'} J_{n'}) \end{aligned} \tag{4.1}$$

is obtained with the same explicit solutions of the form (3.2) as we found in Sec. III.

Next, we establish a link of the following kind:

$$\prod_{l=1}^{4}\exp(\alpha^l P_l)\prod_{m=1}^{4}\exp(\beta^m Q_m)\prod_{n=1}^{6}\exp(\gamma^n J_n)$$
$$=\prod_{l'=1}^{4}\exp(\alpha^{l'}P_{l'})\prod_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})\prod_{m'=1}^{4}\exp(\sigma^{m'}Q_{m'}), \tag{4.2}$$

where the quantities $\sigma^{m'}$ are functions of β^m and γ^n. We proceed by inserting the identity I, in the form

$$I=\prod_{k=1}^{6}\exp(\gamma^k J_k)\prod_{k'=6}^{1}\exp(-\gamma^{k'}J_{k'}), \tag{4.3}$$

between the P_l and Q_m exponentiations on the left-hand side of Eq. (4.2). Then, the piece

$$\prod_{k'=6}^{1}\exp(-\gamma^{k'}J_{k'})\prod_{m=1}^{4}\exp(\beta^m Q_m)\prod_{n=1}^{6}\exp(\gamma^n J_n) \tag{4.4}$$

of the resulting expression can be simplified by repeated use of Theorem 2, Theorem 3, and Lemma 7 of Ref. 7, in a fashion analogous to their use in obtaining the differential equations (3.3)–(3.16).

Thus, for example, we find

$$\begin{aligned}&\exp(-\gamma^1 J_1)\prod_{m=1}^{4}\exp(\beta^m Q_m)\exp(\gamma^1 J_1)\\&=e^{-\gamma^1 J_1}e^{\beta^1 Q_1}e^{\gamma^1 J_1}\cdot e^{-\gamma^1 J_1}e^{\beta^2 Q_2}e^{\gamma^1 J_1}\cdot e^{-\gamma^1 J_1}\\&\quad\times e^{\beta^3 Q_3}e^{\gamma^1 J_1}\cdot e^{-\gamma^1 J_1}e^{\beta^4 Q_4}e^{\gamma^1 J_1}\\&=(1+\beta^1 Q_1)[1+\beta^2(Q_2+\gamma^1 Q_1)]\\&\quad\times(1+\beta^3 Q_3)(1+\beta^4 Q_4)\\&=e^{(\beta^1+\beta^2\gamma^1)Q_1}e^{\beta^2 Q_2}e^{\beta^3 Q_3}e^{\beta^4 Q_4}.\end{aligned} \tag{4.5}$$

Proceeding in this manner ultimately yields expressions for the $\sigma^{m'}$ of Eq. (4.2) in terms of β^m and γ^n:

$$\begin{aligned}\sigma^1&=(\beta^1+\beta^2\gamma^1)\,e^{-(1/2)\gamma^2},\\\sigma^2&=\beta^2 e^{(1/2)\gamma^2}-(\beta^1+\beta^2\gamma^4)\gamma^3 e^{-(1/2)\gamma_2},\\\sigma^3&=(\beta^3-\beta^4\gamma^4)\,e^{(1/2)\gamma^5},\\\sigma^4&=\beta^4 e^{-(1/2)\gamma^5}+(\beta^3-\beta^4\gamma^4)\gamma^6 e^{(1/2)\gamma^5}.\end{aligned} \tag{4.6}$$

Substitution of the solutions for β^m, γ^n obtained in Sec. III and using the resulting form of Eq. (4.2) yields the BCH relation of the form

$$\exp\left(\sum_{l=1}^{4}p^l P_l+\sum_{m=1}^{4}q^m Q_m+\sum_{n=1}^{6}j^n J_n\right)$$
$$=\prod_{l'=1}^{4}\exp(\alpha^{l'}P_{l'})\prod_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})\prod_{m'=1}^{4}\exp(\sigma^{m'}Q_{m'}). \tag{4.7}$$

By an analogous method, we can establish the BCH relation of the type

$$\exp\left(\sum_{l=1}^{4}p^l P_l+\sum_{m=1}^{4}q^m Q_m+\sum_{n=1}^{6}j^n J_n\right)$$
$$=\prod_{m'=1}^{4}\exp(\beta^{m'}Q_{m'})\sum_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})\prod_{l'=1}^{4}\exp(\rho^{l'}P_{l'}). \tag{4.8}$$

In this case, the $\rho^{l'}$ are functions of the α^l and γ^n of Sec. III, given by

$$\begin{aligned}\rho^1&=(\alpha^1-\alpha^2\gamma^4-\alpha^3\gamma^1-\alpha^4\gamma^1\gamma^4)e^{-(1/2)\gamma^-},\\\rho^2&=(\alpha^1-\alpha^2\gamma^4+\alpha^3\gamma^1-\alpha^4\gamma^1\gamma^4)\gamma^6 e^{-(1/2)\gamma^-}\\&\quad+(\alpha^2+\alpha^4\gamma^1)e^{-(1/2)\gamma^+},\\\rho^3&=-(\alpha^1-\alpha^2\gamma^4+\alpha^3\gamma^1-\alpha^4\gamma^1\gamma^4)\gamma^3 e^{-(1/2)\gamma^-}\\&\quad+(\alpha^3-\alpha^4\gamma^4)e^{(1/2)\gamma^+},\\\rho^4&=-(\alpha^1-\alpha^2\gamma^4+\alpha^3\gamma^1-\alpha^4\gamma^1\gamma^4)\gamma^3\gamma^6 e^{-(1/2)\gamma^-}\\&\quad-(\alpha^2+\alpha^4\gamma^1)\gamma^3 e^{-(1/2)\gamma^+}\\&\quad+(\alpha^3-\alpha^4\gamma^4)\gamma^6 e^{(1/2)\gamma^+}+\alpha^4 e^{(1/2)\gamma^-}.\end{aligned} \tag{4.9}$$

In these equations, $\gamma^\pm=\gamma^2\pm\gamma^5$. Again, substitution of the solutions of Sec. III for α^l and γ^n yields the desired BCH relation.

Finally, consider the BCH relations of the type

$$\exp\left(\sum_{l=1}^{4}p^l P_l+\sum_{m=1}^{4}q^m Q_m+\sum_{n=1}^{6}j^n J_n\right)$$
$$=\prod_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})\prod_{l'=1}^{4}\exp(\rho^{l'}P_{l'})\prod_{m'=1}^{4}\exp(\sigma^{m'}Q_{m'}). \tag{4.10}$$

This is, in fact, a class of relations of the same size as that of the coordinates in normal sequence, for the same reasons [see Eq. (4.1) and the associated discussion]. Furthermore, the solution is already known, since Eq. (4.10) can be obtained from Eq. (4.7) by applying a similar derivation to that of Eq. (4.8). Thus, the solutions for $\rho^{l'}$ and $\sigma^{m'}$ in Eq. (4.10) are just those of Eqs. (4.6) and (4.9). As before, substitution of the solutions for α^l, β^m, and γ^n from Sec. III then yields the desired BCH relation.

In summary, this section contains BCH relations for canonical coordinates in non-normal sequence. Schematically, the expressions obtained relate $e^{\Sigma(P+Q+J)}$ to $\Pi e^Q\Pi e^P\Pi e^J$ and certain permutations [Eq. (4.1)], to $\Pi e^P\Pi e^J\Pi e^Q$ [Eqs. (4.6) and (4.7)], to $\Pi e^Q\Pi e^J\Pi e^P$ [Eqs. (4.8) and (4.9)], and to $\Pi e^J\Pi e^P\Pi e^Q$ and certain permutations [Eq. (4.10)]. Note that all expressions remain valid in the limit in which any of the parameters p^l, q^m, j^n are taken to zero.

V. BCH RELATIONS FOR NONCANONICAL COORDINATES

It is also possible to obtain BCH relations between canonical coordinates of the first kind and certain noncanonical coordinate schemes. We present in this section some of the methods for obtaining such relations.

First, consider noncanonical coordinates as defined in Eq. (2.11). A relationship is readily found between canonical coordinates of the third kind in normal sequence, Eq. (2.10), and the noncanonical coordinates of Eq. (2.11). From this, BCH relation between Eqs. (2.9) and (2.11) can be established.

To begin, we note that since the J_n form a subalgebra of the Poincaré subalgebra, there exists a BCH relation of the

form

$$\prod_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})=\exp\left(\sum_{n=1}^{6}j^{n}J_{n}\right), \tag{5.1}$$

where the $\gamma^{n'}$ are given as functions of the j^{n} by Eqs. (3.18) and (3.20)–(3.24). Next, recalling that the P_l commute among themselves, we have

$$\prod_{l'=1}^{4}\exp(\alpha^{l'}P_{l'})=\exp\left(\sum_{l'=1}^{4}\alpha^{l'}P_{l'}\right). \tag{5.2}$$

Finally, using the anticommutators for the Q_m in Eq. (2.1) and also Lemma 2 of Ref. 7, we have

$$\begin{aligned}\prod_{m=1}^{4}&\exp(\beta^{m}Q_{m})\\ &=\exp(\beta^{1}Q_{1}+\beta^{2}Q_{2})\exp(\beta^{3}Q_{3}+\beta^{4}Q_{4})\\ &=\exp\Big(\sum_{m=1}^{4}\beta^{m}Q_{m}\\ &\quad+\tfrac{1}{2}[\beta^{1}Q_{1}+\beta^{2}Q_{2},\beta^{3}Q_{3}+\beta^{4}Q_{4}]\Big)\\ &=\exp\Big(\sum_{m=1}^{4}\beta^{m}Q_{m}\Big)\exp(-\tfrac{1}{2}[\beta^{1}\beta^{3}P_{1}\\ &\quad+\beta^{1}\beta^{4}P_{2}+\beta^{2}\beta^{3}P_{3}+\beta^{2}\beta^{4}P_{4}]).\end{aligned} \tag{5.3}$$

Combining Eqs. (5.1)–(5.3) and the results of Sec. III yields the desired BCH relation as

$$\begin{aligned}\exp&\Big(\sum_{l=1}^{4}p^{l}P_{l}+\sum_{m=1}^{4}q^{m}Q_{m}+\sum_{n=1}^{6}j^{n}J_{n}\Big)\\ &=\exp\Big(\sum_{l'=1}^{4}a^{l'}P_{l'}\Big)\exp\Big(\sum_{m'=1}^{4}\beta^{m'}Q_{m'}\Big)\exp\Big(\sum_{n'=1}^{6}j^{n'}J_{n'}\Big),\end{aligned} \tag{5.4}$$

where

$$\begin{aligned}&a^{1}=\alpha^{1}-\tfrac{1}{2}\beta^{1}\beta^{3},\quad a^{2}=\alpha^{2}-\tfrac{1}{2}\beta^{1}\beta^{4},\\ &a^{3}=\alpha^{3}-\tfrac{1}{2}\beta^{2}\beta^{3},\quad a^{4}=\alpha^{4}-\tfrac{1}{2}\beta^{2}\beta^{4}.\end{aligned} \tag{5.5}$$

Since the P_l commute with the Q_m, we can interchange the order of the $P_{l'}$ and $Q_{m'}$ exponentials on the right-hand side of Eq. (5.4) without changing the solutions (5.5). This immediately gives another BCH relation with the noncanonical coordinates in non-normal sequence:

$$\begin{aligned}\exp&\Big(\sum_{l=1}^{4}p^{l}P_{l}+\sum_{m=1}^{4}q^{m}Q_{m}+\sum_{n=1}^{6}j^{n}J_{n}\Big)\\ &=\exp\Big(\sum_{m'=1}^{4}q^{m'}Q_{m'}\Big)\exp\Big(\sum_{l'=1}^{4}a^{l'}P_{l'}\Big)\exp\Big(\sum_{n'=1}^{6}j^{n'}J_{n'}\Big).\end{aligned} \tag{5.6}$$

Other non-normal sequences are possible for this type of noncanonical coordinates. For instance, a BCH relation of the form

$$\begin{aligned}\exp&\Big(\sum_{l=1}^{4}p^{l}P_{l}+\sum_{m=1}^{4}q^{m}Q_{m}+\sum_{n=1}^{6}j^{n}J_{n}\Big)\\ &=\exp\Big(\sum_{m'=1}^{4}q^{m'}Q_{m'}\Big)\exp\Big(\sum_{n'=1}^{6}j^{n'}J_{n'}\Big)\exp\Big(\sum_{l'=1}^{4}x^{l'}P_{l'}\Big)\end{aligned} \tag{5.7}$$

may be obtained by insertion of the identity I,

$$\begin{aligned}I&=\exp\Big(\sum_{k=1}^{6}j^{k}J_{k}\Big)\exp\Big(-\sum_{k'=1}^{6}j^{k'}J_{k'}\Big)\\ &\equiv\prod_{k=1}^{6}\exp(\gamma^{k}J_{k})\prod_{k'=6}^{1}\exp(-\gamma^{k'}J_{k'}),\end{aligned} \tag{5.8}$$

between the $Q_{m'}$ and $P_{l'}$ exponentials on the right-hand side of Eq. (5.6). From the similarity to the analysis leading to Eq. (4.8), we find immediately

$$\mathbf{x}=\boldsymbol{\rho}(\alpha^{l},\gamma^{n}), \tag{5.9}$$

where $\boldsymbol{\rho}(\alpha^{l},\gamma^{n})$ are given by Eqs. (4.9). This establishes the explicit form of the BCH relation (5.7).

Next, consider the BCH relation of the type

$$\begin{aligned}\exp&\Big(\sum_{l=1}^{4}p^{l}P_{l}+\sum_{m=1}^{4}q^{m}Q_{m}+\sum_{n=1}^{6}j^{n}J_{n}\Big)\\ &=\exp\Big(\sum_{l'=1}^{4}a^{l'}P_{l'}\Big)\exp\Big(\sum_{n'=1}^{6}j^{n'}J_{n'}\Big)\exp\Big(\sum_{m'=1}^{4}b^{m'}Q_{m'}\Big).\end{aligned} \tag{5.10}$$

We obtain this form by inserting the identity I of Eq. (5.8) between the $P_{l'}$ and $Q_{m'}$ exponentials of Eq. (5.4) and using

$$\begin{aligned}\prod_{k'=6}^{1}&\exp(-\gamma^{k'}J_{k'})\exp\Big(\sum_{m'=1}^{4}q^{m'}Q_{m'}\Big)\prod_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})\\ &=\exp\Big[\prod_{k'=6}^{1}\exp(-\gamma^{k'}J_{k'})\Big(\sum_{m'=1}^{4}q^{m'}Q_{m'}\Big)\\ &\quad\times\prod_{n'=1}^{6}\exp(\gamma^{n'}J_{n'})\Big],\end{aligned} \tag{5.11}$$

which follows from Theorem 3 of Ref. 7. Thus the $b^{m'}$ of Eq. (5.10) must be functions of q^{m} and γ^{n}. By comparison with the calculation leading to Eq. (4.7), we find

$$\mathbf{b}=\boldsymbol{\sigma}(q^{m},\gamma^{n}), \tag{5.12}$$

where $\boldsymbol{\sigma}(\beta^{m},\gamma^{n})$ are the solutions (4.6). Thus the BCH relation (5.10) is also explicitly obtained.

Finally, we may readily determine the BCH relations

$$\begin{aligned}\exp&\Big[\sum_{l=1}^{4}p^{l}P_{l}+\sum_{m=1}^{4}q^{m}Q_{m}+\sum_{n=1}^{6}j^{n}J_{n}\Big]\\ &=\exp\Big(\sum_{n'=1}^{6}j^{n'}J_{n'}\Big)\exp\Big(\sum_{l'=1}^{4}x^{l'}P_{l'}\Big)\exp\Big(\sum_{m'=1}^{4}b^{m'}Q_{m'}\Big)\\ &\equiv\exp\Big(\sum_{n'=1}^{6}j^{n'}J_{n'}\Big)\exp\Big(\sum_{m'=1}^{4}b^{m'}Q_{m'}\Big)\exp\Big(\sum_{l'=1}^{4}x^{l'}P_{l'}\Big).\end{aligned} \tag{5.13}$$

The identity (5.8) is inserted in front of the $P_{l'}$ exponential of Eq. (5.10), and the steps leading to Eq. (5.9) are repeated. Thus the expressions for $b^{m'}$ and $x^{l'}$ in Eq. (5.13) are precisely those of Eqs. (5.9) and (5.12).

In this section, we have obtained BCH relations for certain noncanonical coordinates. In schematic form, the results may be summarized as relating $e^{\Sigma(P+Q+J)}$ to $e^{\Sigma P}e^{\Sigma Q}e^{\Sigma J}$ [Eqs. (5.4) and (5.5)], to $e^{\Sigma Q}e^{\Sigma P}e^{\Sigma J}$ [Eq. (5.6)], to $e^{\Sigma Q}e^{\Sigma J}e^{\Sigma P}$ [Eqs. (5.7) and (5.9)], to $e^{\Sigma P}e^{\Sigma J}e^{\Sigma Q}$ [Eqs. (5.10) and (5.12)], and to $e^{\Sigma J}e^{\Sigma P}e^{\Sigma Q}$ or $e^{\Sigma J}e^{\Sigma Q}e^{\Sigma P}$ [Eq. (5.13)]. As for other BCH relations in this paper, the results remain valid in the limit in which any of the parameters p^{l}, q^{m}, j^{n} are taken to zero.

ACKNOWLEDGMENTS

One of the authors (D.R.T.) thanks the members of the Department of Physics in Indiana University for their kind hospitality. The other author (V.A.K.) thanks André Neveu for a careful reading of the manuscript.

This work was supported in part by the United States Department of Energy under Contract No. DE-AC02-84ER40125, Task B; and by the Natural Sciences and Engineering Research Council of Canada.

[1]S. J. Gates, Jr., M. T. Grisaru, M. Rocek, and W. Siegel, *Superspace* (Benjamin–Cummings, Reading, MA, 1983).
[2]A. Rogers, J. Math. Phys. **21**, 1352 (1980).
[3]A. Rogers, J. Math. Phys. **22**, 939 (1981).
[4]A. Rogers, J. Math. Phys. **22**, 443 (1981); **26**, 385, 2749 (1985).
[5]C. P. Boyer and J. Gitler, Trans. Am. Math. Soc. **258**, 241 (1984); J. Hoyos, M. Quiros, J. Ramirez Mittelbrunn, and F. J. de Urries, J. Math. Phys. **25**, 833, 841, 847 (1984); J. M. Rabin and L. Crane, Commun. Math. Phys. **100**, 141 (1985); **102**, 123 (1985).
[6]D. R. Truax, V. A. Kostelecký, and M. M. Nieto, J. Math. Phys. **27**, 354 (1986).
[7]V. A. Kostelecký, M. M. Nieto, and D. R. Truax, J. Math. Phys. **27**, 1419 (1986).
[8]V. A. Kostelecký, Nucl. Phys. B **219**, 167 (1983); V. A. Kostelecký and J. M. Rabin, J. Math. Phys. **25**, 2744 (1984); in *Supersymmetry in Physics*, edited by V. A. Kostelecký and D. K. Campbell (North-Holland, Amsterdam, 1985), p. 213.
[9]B. Kostant, *Lecture Notes in Mathematics*, Vol. 570 (Springer, New York, 1977), p. 617; F. A. Berezin and D. A. Leites, Sov. Math. Dokl. **16**, 1218 (1975); M. Batchelor, Trans. Am. Math. Soc. **253**, 329 (1979); 258, 257 (1980); B. DeWitt, *Supermanifolds* (Cambridge U.P., Cambridge, 1984); see also Ref. 2.
[10]See, for example, R. G. Yates, Commun. Math. Phys. **76**, 255 (1980).
[11]J. E. Campbell, Proc. London Math. Soc. **28**, 381 (1897); H. F. Baker, *ibid.* **34**, 347 (1902); **35**, 333 (1903); **2**, 293 (1905); F. Hausdorff, Ber. Sachsischen Akad. Wiss. Math. Phys. Kl. Leipzig **58**, 19 (1906).
[12]For further developments, see, for example, R. M. Wilcox, J. Math. Phys. **8**, 962 (1967); B. Mielnik, Ann. Inst. H. Poincaré A **12**, 215 (1970); R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley, New York, 1974).
[13]D. R. Truax, Phys. Rev. D **31**, 1988 (1985).
[14]E. Inönü and E. P. Wigner, Proc. Natl. Acad. Sci. (USA) **39**, 510 (1953).
[15]See, for example, E. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).
[16]W. E. Boyce and R. C. DiPrima, *Elementary Differential Equations and Boundary Value Problems* (Wiley, New York, 1969).

Chiral symmetry breakdown. I. Gauge dependence in constant vertex approximation

D. Atkinson and P. W. Johnson[a)]
Institute for Theoretical Physics, P. O. Box 800, 9700 AV Groningen, The Netherlands

(Received 7 March 1987; accepted for publication 13 May 1987)

An approximate quark propagator equation in a Landau-like gauge is analyzed and it is shown that there is a critical value of the coupling constant, corresponding to the onset of dynamical chiral symmetry breakdown, provided that (a) there is an infrared cutoff, which can be supplied by an effective gluon mass, and (b) there is an ultraviolet cutoff, which may be engendered by a running coupling constant. Dynamical chiral symmetry breakdown is shown not to occur in other gauges under the same circumstances, thus casting doubt upon the approximations commonly used.

I. INTRODUCTION

The idea that quarks obtain effective (constituent) masses as a result of a dynamical breakdown of chiral symmetry has received a great deal of attention in recent years.[1-10] We propose to examine this attractive hypothesis by a detailed analysis of truncated Dyson–Schwinger equations for the quark propagator. In this paper, we will restrict attention to the approximation in which the gluon–quark vertex is replaced by the bare value, whereas the gluon acquires an effective mass, while its propagator retains the tensor structure of the bare propagator. This resembles the first Johnson–Baker–Willey (JBW) approximation for the electron propagator of QED.[11-13]

In pioneering work over a decade ago, Maskawa and Nakajima[2] studied the truncated Dyson–Schwinger equation in a JBW-like approximation. Their analysis in a "Landau-like" gauge showed that spontaneous chiral symmetry breakdown occurs when a Pauli–Villars ultraviolet cutoff parameter Λ is introduced, *and* that spontaneous breakdown survives in the continuum limit $\Lambda \to \infty$. We obtain similar conclusions in that gauge, but using a smooth ultraviolet cutoff function, the choice being motivated by QCD. Like Maskawa and Nakajima in Ref. 2, and unlike several recent authors,[3-9] we have gone to some care in analyzing coupled Dyson–Schwinger equations for the two functions appearing in the quark propagator. The formalism is described in Sec. II, and the Landau-like gauge is analyzed in Sec. III.

The Landau-like gauge of Ref. 2 leads to Dyson–Schwinger equations which are relatively well behaved in the ultraviolet, whereas in other covariant gauges they become more singular. The case of Feynman gauge with *finite* momentum cutoff parameter Λ has also been analyzed in Ref. 2. We show in Sec. IV that, because of ultraviolet singularities in the continuum limit, $\Lambda \to \infty$, in Feynman gauge the regularized quark propagator corresponds to massless, free quarks. The Dyson–Schwinger equations exhibit spontaneous chiral symmetry breaking at finite cutoff Λ, because the quark mass operator satisfies a homogeneous Fredholm integral equation in that case, but the solution becomes "trivialized" upon renormalization in the continuum limit. Our logarithmic ultraviolet cutoff function reduces the degree of the divergence in the continuum limit, before renormalization, from $\log \Lambda$ (Ref. 2) to $\log\log \Lambda$; but it does not eliminate the need for regularization.

We have established that, in the JBW-like approximation, the quark propagator exhibits a sensitivity to the choice of gauge. This apparent "gauge dependence" of spontaneous chiral symmetry breaking is a consequence of the fact that truncated Dyson–Schwinger equations in the JBW-like scheme have ultraviolet singularities in most gauges. It is our conclusion that such a truncation is inadequate for studying spontaneous chiral symmetry breaking, and we intend in the future to study the problem for truncation schemes in which our choice of vertex function is motivated by the Slavnov–Taylor identities. In addition, asymptotic freedom imposes constraints upon the ultraviolet behavior of the propagator and vertex function.

II. DYSON–SCHWINGER EQUATION

The quark propagator satisfies the integral equation

$$S_F'^{-1}(p) = \not{p} - \frac{i\lambda}{(2\pi)^4}\int d^4p'\,\gamma_\mu S_F'(p')\gamma_\nu D_F'^{\mu\nu}(p'-p), \tag{2.1}$$

where we have approximated the full by the bare vertex. Here λ is the square of the QCD coupling constant, times a color factor, and the bare quark mass is zero. We suppose that the gluon has an effective mass, generated by self-interaction. The correct form for a massive vector propagator in a gauge theory is

$$\frac{1}{k^2 - m^2 + i\epsilon}\left[-g^{\mu\nu} + (1-a)\frac{k^\mu k^\nu}{k^2 - am^2 + i\epsilon}\right]. \tag{2.2}$$

We multiply this by a factor $\omega(-k^2)$ that satisfies $\omega(0) = 1$ and $\omega(-k^2) \sim [\log(-k^2)]^{-1}$ as $-k^2 \to \infty$, in order to allow for the decrease of the running coupling constant in a non-Abelian gauge theory. Thus

[a)] Permanent address: Physics Department, Illinois Institute of Technology, Chicago, Illinois 60616.

 0022-2488/87/102488-06$02.50 © 1987 American Institute of Physics

$$D'^{\mu\nu}_F(k) = \omega(-k^2)\left[\frac{-g^{\mu\nu}}{k^2-m^2+i\epsilon}+\frac{k^\mu k^\nu}{m^2}\times\left(\frac{1}{k^2-m^2+i\epsilon}-\frac{1}{k^2-am^2+i\epsilon}\right)\right], \tag{2.3}$$

where we have rearranged the tensor for calculational convenience and where a is the gauge parameter. A suitable form for ω is

$$\omega(-k^2) = \frac{k^2}{k^2-m^2+i\epsilon}\left[\log\left(1-\frac{k^2}{m^2}-i\epsilon\right)\right]^{-1}, \tag{2.4}$$

although the results do not depend on the detailed expression.

The inverse quark propagator has the form

$$S'^{-1}_F(p) = \alpha(-p^2)+\not{p}\beta(-p^2), \tag{2.5}$$

where α and β are scalar functions, so that

$$S'_F(p) = \frac{\alpha(-p^2)-\not{p}\beta(-p^2)}{\alpha^2(-p^2)-p^2\beta^2(-p^2)}. \tag{2.6}$$

Upon insertion of these formulas into (2.1), two coupled equations for α and β can be obtained; and, after Wick rotation, one gets (with $x=p^2$, $y=p'^2$),

$$\alpha(x) = \frac{\lambda}{\pi^2}\int_0^\infty dy\,K(x,y)\frac{y\alpha(y)}{\alpha^2(y)+y\beta^2(y)}, \tag{2.7}$$

$$\beta(x) = 1+\frac{\lambda}{\pi^2}\int_0^\infty dy\,L(x,y)\frac{y\beta(y)}{\alpha^2(y)+y\beta^2(y)}, \tag{2.8}$$

where

$$K(x,y) = \int_0^\pi d\theta\sin^2\theta\cdot\omega(k^2)\left[\frac{3}{k^2+m^2}+\frac{a}{k^2+am^2}\right] \tag{2.9}$$

and

$$L(x,y) = \int_0^\pi d\theta\frac{\omega(k^2)}{k^2+m^2}\left[2(y/x)^{1/2}\cos\theta+(1-a)\times\frac{(y-x)k^2[(y/x)^{1/2}\cos\theta-1]}{k^2+am^2}\right]\sin^2\theta, \tag{2.10}$$

with

$$k^2 = (p'-p)^2 = x+y-2(xy)^{1/2}\cos\theta. \tag{2.11}$$

We require that $\omega(x)$ be a monotonically decreasing function of x for Euclidean momenta. With the Euclidean version of (2.4), namely,

$$\omega(x) = \frac{x}{x+m^2}\left[\log\left(1+\frac{x}{m^2}\right)\right]^{-1}, \tag{2.12}$$

it is not possible to evaluate the kernels K and L in terms of elementary functions.

A simplification is to replace $\omega(k^2)$ by unity, i.e., the coupling does not run. This has a profound (and nonphysical) effect on the behavior of the equation. The kernels can now be evaluated,

$$K(x,y) = \tfrac{3}{8}k(x,y,m^2)+\tfrac{1}{8}ak(x,y,am^2), \tag{2.13}$$

$$L(x,y) = \tfrac{1}{4}yk^2(x,y,m^2)-(1/16m^2x)\times[[(y-x)^2+m^2(y+x)]k(x,y,m^2)-[(y-x)^2+am^2(y+x)]k(x,y,am^2)] \tag{2.14}$$

with

$$k(x,y,m^2) = [x+y+m^2+[(x+y+m^2)^2-4xy]^{1/2}]^{-1}. \tag{2.15}$$

This was essentially the case considered by Maskawa and Nakajima,[2] together with the Pauli–Villars cutoff version.

In the limit that the gluon mass m tends to zero, we find

$$k(x,y,0) = \frac{\theta(x-y)}{2x}+\frac{\theta(y-x)}{2y}. \tag{2.16}$$

For $m\neq 0$, the approximation

$$k(x,y,m^2)\approx\frac{\theta(x-y)}{2(x+m^2)}+\frac{\theta(y-x)}{2(y+m^2)} \tag{2.17}$$

is exact in the limits $x\to 0$ and $x\to\infty$, $y\to 0$ and $y\to\infty$, and it is a strict upper bound for all positive x and y. In this paper, we shall use the form (2.17) exclusively, although we propose to consider the exact expression in a future publication.

The above approximation is improved by reinstating the running coupling constant. Unfortunately, the dependence of k^2 on the angle θ in (2.9) and (2.10) makes it impossible to evaluate the integrals in closed form when the ω of (2.12) is present. However, if one sets

$$\omega(k^2)\approx\omega(p^2)\theta(p^2-p'^2)+\omega(p'^2)\theta(p'^2-p^2), \tag{2.18}$$

one obtains, instead of the kernels K and L of Eqs. (2.13) and (2.14), respectively,

$$[\omega(x)\theta(x-y)+\omega(y)\theta(y-x)]K(x,y), \tag{2.19}$$

$$[\omega(x)\theta(x-y)+\omega(y)\theta(y-x)]L(x,y). \tag{2.20}$$

The approximation (2.18) for the smooth, monotonic function $\omega(k^2)$ is good when $p^2\gg p'^2$ or $p^2\ll p'^2$, but not when p^2 and p'^2 are comparable in magnitude. However, the approximation is not expected to affect either the infrared or the ultraviolet behaviors of the solution.

With the approximations (2.17) and (2.18), the kernels read

$$K(x,y) = \tfrac{1}{16}[\omega(x)\mu(x)\theta(x-y)+\omega(y)\mu(y)\theta(y-x)], \tag{2.21}$$

$$L(x,y) = (y/32)[\omega(x)\nu(x,y)\theta(x-y)+\omega(y)\nu(y,x)\theta(y-x)], \tag{2.22}$$

where

$$\mu(x) = 3/(x+m^2)+a/(x+am^2), \tag{2.23}$$

$$\nu(x,y) = \frac{2}{(x+m^2)^2}-\frac{1}{m^2x}\left[\frac{(y-x)^2+m^2(y+x)}{x+m^2}-\frac{(y-x)^2+am^2(y+x)}{x+am^2}\right]. \tag{2.24}$$

The Feynman gauge ($a=1$) is especially simple,

$$D'^{\mu\nu}_F(k) = [-g^{\mu\nu}/(k^2-m^2+i\epsilon)]\omega(-k^2), \tag{2.25}$$

$$\mu(x) = 4/(x+m^2), \tag{2.26}$$

$$\nu(x,y) = 2/(x+m^2)^2. \tag{2.27}$$

It turns out that in this gauge (and others), an ultraviolet cutoff is necessary (see Sec. IV). On the other hand, no such cutoff is required in the Landau gauge ($a=0$),

$$D_F'^{\mu\nu}(k) = \frac{-g^{\mu\nu} + k^\mu k^\nu/(k^2+i\epsilon)}{k^2 - m^2 + i\epsilon}\omega(-k^2), \tag{2.28}$$

$$\mu(x) = 3/(x+m^2), \tag{2.29}$$

$$\nu(x,y) = \frac{2}{(x+m^2)^2} + \frac{y-3x}{x^2(x+m^2)}. \tag{2.30}$$

As can be seen from the denominator in (2.30), an infrared singularity has been introduced, a gauge artifact, and this turns out to be a nuisance. To avoid this difficulty, Maskawa and Nakajima[2] introduce what they called the Landau-like gauge, with the gluon propagator

$$D_F'^{\mu\nu}(k) = \left[\frac{-g^{\mu\nu}}{k^2 - m^2 + i\epsilon} + \frac{k^\mu k^\nu}{(k^2 - m^2 + i\epsilon)^2}\right]\omega(-k^2), \tag{2.31}$$

for which the kernels K and L, with the approximations (2.17) and (2.18), have the form (2.21) and (2.22), with

$$\mu(x) = 3/(x+m^2) + m^2/(x+m^2)^2, \tag{2.32}$$

$$\nu(x,y) = 2m^2/(x+m^2)^3. \tag{2.33}$$

Here the good ultraviolet properties have been retained, while the artificial infrared divergence has been removed.

In Sec. III, we consider this Landau-like gauge, without ultraviolet cutoff; while the Feynman gauge is treated in Sec. IV. In the latter case, a Pauli–Villars cutoff has to be introduced.

The major purpose is to find out conditions under which there is a critical $\lambda_c > 0$, such that, for $0 < \lambda < \lambda_c$, Eqs. (2.7) and (2.8) only have the chiral solution $\alpha \equiv 0$, while for $\lambda > \lambda_c$, there is also a nontrivial solution, $\alpha \neq 0$. To investigate such a bifurcation point λ_c we differentiate the equations functionally w.r.t. α, and set $\alpha = 0$,

$$\delta\alpha(x) = \frac{\lambda}{\pi^2}\int_0^\infty dy\, K(x,y)\,\frac{\delta\alpha(y)}{\beta^2(y)}, \tag{2.34}$$

$$\beta(x) = 1 + \frac{\lambda}{\pi^2}\int_0^\infty dy\, L(x,y)\,\frac{1}{\beta(y)}. \tag{2.35}$$

III. LANDAU-LIKE GAUGE

In the case (2.31)–(2.33), we can write the bifurcation equations (2.34) and (2.35) in the form

$$\delta\alpha(x) = \frac{\lambda}{16\pi^2}\left\{\int_0^x dy\,\rho(x) + \int_x^\infty dy\,\rho(y)\right\}\frac{\delta\alpha(y)}{\beta^2(y)}, \tag{3.1}$$

$$\beta(x) = 1 + \frac{\lambda}{16\pi^2}\left\{\int_0^x dy\,\sigma(x) + \int_x^\infty dy\,\sigma(y)\right\}\frac{y}{\beta(y)}, \tag{3.2}$$

where

$$\rho(x) = [3/(x+m^2) + m^2/(x+m^2)^2]\omega(x), \tag{3.3}$$

$$\sigma(x) = [m^2/(x+m^2)^3]\,\omega(x). \tag{3.4}$$

We study first Eq. (3.2). To this end, consider the mapping

$$\bar\beta(x) = P(\beta;x), \tag{3.5}$$

where $P(\beta;x)$ is just the right-hand side of (3.2). Let B be the Banach space of real, continuous functions $f(x)$ with supremum norm, for which the following inequalities hold:

$$0 < \beta_m \leqslant f(x) \leqslant \beta_M < \infty. \tag{3.6}$$

We shall specify β_m and β_M in a moment.

Next, define the function

$$P(x) = \frac{1}{16\pi^2}\left\{\int_0^x y\,dy\,\sigma(x) + \int_x^\infty y\,dy\,\sigma(y)\right\}. \tag{3.7}$$

It is easy to see that $P(x)$ is non-negative and monotonically decreasing in $0 < x < \infty$. Thus

$$0 \leqslant P(x) \leqslant P(0) < \infty. \tag{3.8}$$

A computer estimate gives

$$P(0) = \frac{1}{16\pi^2}\int_0^\infty d\omega\,\frac{\omega^2}{(1+\omega)^4\log(1+\omega)} \approx 0.001\,82. \tag{3.9}$$

Because of the positivity of $\sigma(x)$, we see from (3.2) and (3.5) that

$$\bar\beta(x) \geqslant 1 \tag{3.10}$$

and

$$\bar\beta(x) \leqslant 1 + \lambda P(0), \tag{3.11}$$

so that, if we define

$$\beta_m = 1 \tag{3.12}$$

and

$$\beta_M = 1 + \lambda P(0), \tag{3.13}$$

we see that the space B is mapped into itself by the nonlinear operator, P. Indeed, the image of B is actually compact in norm, since

$$\frac{d}{dx}\bar\beta(x) = \frac{\lambda}{16\pi^2}\left[\frac{d}{dx}\sigma(x)\right]\int_0^x \frac{y\,dy}{\beta(y)}. \tag{3.14}$$

Now $d\sigma/dx$ is negative, so $d\bar\beta/dx$ is also negative, and

$$-\frac{d}{dx}\bar\beta(x) \leqslant -\frac{\lambda}{32\pi^2}x^2\left[\frac{d}{dx}\sigma(x)\right] \leqslant \text{const}, \tag{3.15}$$

i.e., $|d\bar\beta/dx|$ has a bound that is independent of β.

Since P is a completely continuous operator that maps B into itself, we can use the Schauder theorem to assert that there is at least one fixed point, $\bar\beta = \beta$, in B, i.e., at least one solution of (3.2). To show that the solution is unique in B, we subtract $\beta(0)$,

$$\beta(x) = \beta(0) + \frac{\lambda}{16\pi^2}\int_0^x dy[\sigma(x) - \sigma(y)]\,\frac{y}{\beta(y)}. \tag{3.16}$$

Any solution of (3.2) is also a solution of (3.16), on the condition that $\beta(0)$ has the correct value. We first show that no two different solutions in B can have the same $\beta(0)$. For suppose that β_1 and β_2 both satisfy (3.16), and that $\beta_1(0) = \beta_2(0)$. Then

$$\beta_1(x)-\beta_2(x)=\frac{\lambda}{16\pi^2}\int_0^x y\,dy[\sigma(y)-\sigma(x)]\times\frac{\beta_1(y)-\beta_2(y)}{\beta_1(y)\beta_2(y)}. \tag{3.17}$$

Hence

$$|\beta_1(x)-\beta_2(x)|\leqslant\frac{\lambda}{16\pi^2\beta_m^2}\sum(x)\sup_{0<y<x}|\beta_1(y)-\beta_2(y)|, \tag{3.18}$$

where

$$\sum(x)=\int_0^x y\,dy[\sigma(y)-\sigma(x)]. \tag{3.19}$$

Let us take $X>0$ to be so small that, for any $x\in[0,X]$,

$$\sum(x)\leqslant\frac{16\pi^2\beta_m^2}{\lambda}K \tag{3.20}$$

with $K<1$. Then

$$\sup_{0<x<X}|\beta_1(x)-\beta_2(x)|\leqslant K\sup_{0<x<X}|\beta_1(x)-\beta_2(x)|, \tag{3.21}$$

which is only possible if $\beta_1(x)=\beta_2(x)$ for $0\leqslant x\leqslant X$. Since $\beta(x)$ satisfies the differential equation

$$\left[\frac{\beta'(x)}{\sigma'(x)}\right]'=\frac{\lambda}{16\pi^2}\frac{x}{\beta(x)}, \tag{3.22}$$

it is easy to extend this identity to all x values.

Next consider the case that $\beta_1(0)\neq\beta_2(0)$. For definiteness, we set $\beta_1(0)>\beta_2(0)$. Instead of (3.17) we have

$$\beta_1(x)-\beta_2(x)=\beta_1(0)-\beta_2(0)+\frac{\lambda}{16\pi^2}\int_0^x y\,dy[\sigma(y)-\sigma(x)]\times\frac{\beta_1(y)-\beta_2(y)}{\beta_1(y)\beta_2(y)}. \tag{3.23}$$

This has the structure of a linear Volterra equation for $\beta_1-\beta_2$, given $\beta_1\beta_2$; and the Neumann series is guaranteed to have an infinite radius of convergence. Since $\sigma(y)\geqslant\sigma(x)$ for $y\leqslant x$, each term in the series is non-negative, and so, for all x

$$\beta_1(x)-\beta_2(x)\geqslant\beta_1(0)-\beta_2(0). \tag{3.24}$$

Now it follows from (3.2) that $\beta(\infty)=1$, so by taking $x=\infty$ in (3.24) we find $\beta_2(0)\geqslant\beta_1(0)$, which contradicts $\beta_1(0)>\beta_2(0)$. Hence $\beta_1(0)=\beta_2(0)$ and, as we have seen this implies $\beta_1(x)\equiv\beta_2(x)$.

Having shown that (3.2) has a unique solution in B, we turn to (3.1). Let us write it in the form

$$\delta\alpha(x)=\frac{\lambda}{16\pi^2}\int_0^\infty dy\,H(x,y)\delta\alpha(y), \tag{3.25}$$

where

$$H(x,y)=[\rho(x)\theta(x-y)+\rho(y)\theta(y-x)]\beta^{-2}(y). \tag{3.26}$$

Thanks to the fact that $\beta(y)$ is bounded from below, we can show that H is a positive L^2 kernel,

$$\begin{aligned}\int_0^\infty\int_0^\infty dx\,dy\,H^2(x,y)&=\int_0^\infty dx\,\rho^2(x)\int_0^x\frac{dy}{\beta^4(y)}+\int_0^\infty dy\,\frac{\rho^2(y)}{\beta^4(y)}dx\\&\leqslant\frac{2}{\beta_m^4}\int_0^\infty\frac{d\omega}{\log^2(1+\omega)}\frac{\omega^3}{(1+\omega)^4}\times\left(3+\frac{1}{1+\omega}\right)^2<\infty.\end{aligned} \tag{3.27}$$

Hence (3.25) is a classic Fredholm equation, and thus, if we require $\delta\alpha(x)$ to belong to L^2, the spectrum is discrete; in particular, there is a smallest value $\lambda_c>0$ such that (3.25) has only the trivial solution, $\delta\alpha(x)\equiv0$, if $0<\lambda<\lambda_c$, while it has a nontrivial solution if $\lambda=\lambda_c$.

The existence of a critical point λ_c is crucially dependent on limiting $\delta\alpha$ to L^2. Equation (3.1) is equivalent to the differential equation

$$\frac{d}{dx}\left[\frac{(d/dx)\delta\alpha(x)}{(d/dx)\rho(x)}\right]=\frac{\lambda}{16\pi^2}\frac{\delta\alpha(x)}{\beta^2(x)}, \tag{3.28}$$

with the boundary condition

$$\frac{d}{dx}\delta\alpha(x)\underset{x\to0}{\to}0. \tag{3.29}$$

The asymptotic behavior of (3.28) for large x is

$$\frac{d}{dx}\left[x^2\log\frac{x}{m^2}\frac{d}{dx}\delta\alpha(x)\right]+\frac{3\lambda}{16\pi^2}\delta\alpha(x)\sim0, \tag{3.30}$$

where we have used the fact that $\beta(\infty)=1$. This admits two solutions, which have the asymptotic behaviors

$$f_R(x)\sim(1/x)[\log(x/m^2)]^{-1+3\lambda/16\pi^2}, \tag{3.31}$$

$$f_I(x)\sim[\log(x/m^2)]^{-3\lambda/16\pi^2}. \tag{3.32}$$

The general solution of (3.28) is

$$\delta\alpha(x)=Af_R(x)+Bf_I(x), \tag{3.33}$$

but in order for this to solve the integral equation (3.1), the boundary condition (3.29) needs to be imposed. This fixes the ratio B/A, the remaining constant being a trivial normalization. It should be noted that a solution of the form (3.33) exists for any λ, but that it is not in L^2 in general. The smallest value of λ for which $B=0$ is precisely λ_c, and $\delta\alpha$ is then the regular solution f_R, which is in L^2.

In conclusion, we have seen that the bifurcation equations, in the Landau-like gauge, yield a critical point λ_c only if some information external to the Dyson–Schwinger system is used, in order to exclude the irregular solution $f_I(x)$.[3,8,9]

IV. FEYNMAN GAUGE

In Feynman gauge the bifurcation equation (2.35) for β has the specific form

$$\beta(x)=1+\frac{\lambda}{16\pi^2}\int_0^\infty dy\,\sigma(x_{\max})\frac{y}{\beta(y)}, \tag{4.1}$$

where $x_{\max}=\max(x,y)$, and where by hypothesis the function

$$\sigma(x)=\omega(x)/(x+m^2)^2 \tag{4.2}$$

is positive and monotonically decreasing. Notice that the function σ has one inverse power of $(x+m^2)$ less than that of Sec. III. We shall in fact show that (4.1) has *no* solutions.

Equation (4.1) is equivalent to the integrodifferential equation

$$\beta'(x) = \frac{\lambda}{16\pi^2}\sigma'(x)\int_0^x dy\,\frac{y}{\beta(y)}, \tag{4.3}$$

along with the boundary condition

$$\beta(\infty) = 1. \tag{4.4}$$

Let us consider the case in which $\beta(0) > 0$, and define the domain $\mathfrak{D}$ on which β remains positive,

$$\mathfrak{D} = \{x | \beta(y) > 0 \text{ for } y \in [0,x]\}. \tag{4.5}$$

It follows from (4.3) that β is monotonically decreasing on $\mathfrak{D}$. As a consequence

$$\beta'(x) \leqslant [\lambda/32\pi^2\beta(0)]x^2\sigma'(x) \tag{4.6}$$

for $x \in \mathfrak{D}$. Integrating, we obtain

$$\beta(0) - \beta(x) \geqslant \frac{\lambda}{16\pi^2\beta(0)}\int_0^x dy\, y\sigma(y). \tag{4.7}$$

It follows from (4.2) and the definition (2.4) of ω that, at large y,

$$\sigma(y) \sim 1/y^2 \log y, \tag{4.8}$$

so that the integral in (4.7) approaches $\log\log x$ asymptotically at large x. Because of this divergence, the function $\beta(x)$ must approach zero at a finite point x_0 on the positive real x axis. In the vicinity of such a point, the solution to the differential equation (4.3) has the behavior

$$\beta(x) \sim (x_0 - x)[(\lambda x_0/8\pi^2)\sigma'(x_0)\log(x_0 - x)]^{1/2}. \tag{4.9}$$

The solution to Eq. (4.3) consequently has a branch point at $x = x_0$, with the real-analytic continuation having a branch cut for $x > x_0$. Furthermore, this solution of (4.3) has the asymptotic form

$$\beta(x) \sim \pm[(-\lambda/8\pi^2)\log\log x]^{1/2} \tag{4.10}$$

as x becomes large within the cut plane. Such solutions are not consistent with the boundary condition (4.4), so that they do not satisfy the integral equation (4.1), even if x is allowed to be complex.

We have shown that there are no solutions of (4.1) for $\beta(x)$ positive. Since $-\beta(x)$ satisfies Eq. (4.3) if $\beta(x)$ is a solution, there are also no solutions of (4.1) for $\beta(0)$ negative. For $\beta(0) = 0$, the solution to (4.3) has the following asymptotic behavior at small x:

$$\beta(x) \sim \pm[(\lambda/12\pi^2)\sigma'(0)x^3]^{1/2}, \tag{4.11}$$

where $\sigma'(0) < 0$. In this case the real-analytic solution has a branch cut for $x > 0$, and it also has asymptotic behavior (4.10) at large x. Therefore there are no solutions to the integral equation for this case either.

The integral equation (4.1), considered for any positive, strictly decreasing weight functions $\omega(x)$, has no solutions whenever

$$\lim_{x\to\infty}\int_0^x dy\, y\sigma(y) = \infty. \tag{4.12}$$

If the weight function $\omega(x)$ were chosen to decrease slightly faster—say, $O[(\log x)^{-1-\epsilon}]$ for $\epsilon > 0$—the integral (4.12) would converge and the integral equation would have a solution. This might well be affected by modifying the approximation for the quark–gluon vertex function—a matter we propose to take up in the future—but for the present we shall discuss the more standard Pauli–Villars cutoff procedure.

In the Pauli–Villars approach, we replace the function $\sigma(x_{\max})$ in the nonlinear integral equation (4.1) by the function $\tau(x_{\max})$,

$$\tau(x) = \omega(x)\,[1/(x+m^2)^2 - 1/(x+\Lambda^2)^2]; \tag{4.13}$$

with $\Lambda \gg m$. Equivalently, the function $\beta(x)$ will satisfy the nonlinear Volterra equation

$$\beta(x) = \beta(0) + \frac{\lambda}{16\pi^2}\int_0^x dy\,\frac{y}{\beta(y)}[\tau(x) - \tau(y)], \tag{4.14}$$

along with the boundary condition

$$\beta(\infty) = 1. \tag{4.15}$$

Let us consider the solution of Eq. (4.14), starting from a given initial value $\beta(0) > 0$. We define $\mathfrak{D}$ as the domain over which β remains positive; vide Eq. (4.5). For x in $\mathfrak{D}$, the Volterra equation has a unique monotonically decreasing solution $\beta(x)$. Furthermore, the value of β at fixed x is monotonically increasing as a function of the initial value $\beta(0)$. On the domain $\mathfrak{D}$, $\beta(x)$ satisfies the bound

$$\beta(x) \geqslant \beta(0) - I/\beta(x), \tag{4.16}$$

where

$$I = \frac{\lambda}{16\pi^2}\int_0^\infty dy\, y\tau(y). \tag{4.17}$$

If we choose

$$\beta(0) > [4I]^{1/2}, \tag{4.18}$$

it follows from (4.16) that $\beta(x)$ is positive for all $x \geqslant 0$.

We have shown that, for $\beta(0)$ chosen sufficiently large, the nonlinear integral equation (4.14) has a unique positive solution for $x > 0$. For a particular choice of $\beta(0)$, one satisfies condition (4.15). One can show directly from the integral equation that, to meet (4.15), the initial value $\beta(0)$ lies somewhere between the limits

$$I + [I^2+4]^{1/2}/2 \leqslant \beta(0) \leqslant I + 1. \tag{4.19}$$

Consequently, there is a unique solution to the integral equation corresponding to (4.1), with a Pauli–Villars cutoff inserted.

We have shown the existence of a unique positive solution of the cutoff integral equation, but the question remains as to the limit in which the cutoff parameter Λ becomes large. For our case the integral $I(\Lambda)$, defined in (4.17), has the form

$$I(\Lambda) = \frac{\Lambda}{16\pi^2}\int_0^\infty dy\, y\omega(y)\left[\frac{1}{(y+m^2)^2} - \frac{1}{(y+\Lambda^2)^2}\right]. \tag{4.20}$$

Because of (4.12), the integral $I(\Lambda)$ must diverge in the limit $\Lambda \to \infty$. In fact, one can show that

$$I(\Lambda) \sim (\lambda/16\pi^2)\log\log\Lambda \tag{4.21}$$

at large Λ. Because of this asymptotic behavior, along with the bounds (4.19), it follows that

$$\beta(0,\Lambda) \sim (\lambda/16\pi^2)\log\log\Lambda \tag{4.22}$$

as $\Lambda \to \infty$. In fact, one may show that, for fixed x, the renormalized function

$$\tilde{\beta}(x) \equiv \lim_{\Lambda\to\infty} [\beta(x,\Lambda)/\beta(0,\Lambda)] = 1. \tag{4.23}$$

The integral equation (2.34) for $\delta\alpha(x)$, with $\beta(x,\Lambda)$ inserted, exhibits chiral symmetry breaking, in that for λ less than some critical value $\lambda_c > 0$, the only solution is $\delta\alpha = 0$. The analysis in Feynman gauge is similar to that of Sec. III in Landau gauge. The critical coupling λ_c depends upon Λ, and in fact

$$\lambda_c \sim [\beta(0,\Lambda)]^2. \tag{4.24}$$

In other words, the only consistent solution of (2.34) for fixed coupling λ in the limit as the cutoff Λ becomes large is

$$\delta\alpha(x,\Lambda) = 0. \tag{4.25}$$

The renormalized function $\delta\tilde{\alpha}(x)$ is also zero,

$$\delta\tilde{\alpha}(x) \equiv \lim_{\Lambda\to\infty} \frac{\delta\alpha(x,\Lambda)}{\beta(0,\Lambda)} = 0.$$

We therefore find that in Feynman gauge, the normalized inverse quark propagator $\tilde{S}^{-1}(p)$ corresponds to a massless, free quark,

$$\tilde{S}^{-1}(p) = \lim_{\Lambda\to\infty} \frac{\alpha(x,\Lambda) + \not{p}\beta(x,\Lambda)}{\beta(0,\Lambda)} = \not{p}. \tag{4.26}$$

In summary, we have shown that there is no solution of the bifurcation equation (4.1) in Feynman gauge, because of problems in the ultraviolet. There is a solution to the Dyson–Schwinger equations when a Pauli–Villars cutoff parameter Λ is introduced, but the renormalized propagator corresponds to free, massless quarks in the limits as $\Lambda \to \infty$. One would expect the phenomenon of chiral symmetry breaking to gauge invariant, but our algorithm for truncation of the Dyson–Schwinger equation is explicitly gauge dependent. The difficulty can be plausibly traced to the naive JBW treatment of the full vertex function.

ACKNOWLEDGMENTS

We would like to thank M. Koopmans for useful discussions.

This work has been partially supported by the National Science Foundation. One of us (PWJ) would like to thank the Stichting FOM (Fundamenteel Onderzoek der Materie), financially supported by the Nederlandse Organisatie voor Zuiver Wetenschappelijk Onderzoek, for its support.

[1] Y. Nambu and G. Jona-Lasinio, Phys. Rev. **122**, 345 (1961).
[2] T. Maskawa and H. Nakajima, Prog. Theor. Phys. **52**, 1326 (1974);**54**, 860 (1975).
[3] P. Fomin, V. Gusynin, V. Miranskii, and Y. Sitenko, Rev. Nuovo Cimento **6**, 1 (1983); V. Miranskii, Phys. Lett. B **165**, 401 (1985);V. Miranskii and P. Fomin, Sov. J. Part. Nucl. **16**, 203 (1985).
[4] K. Higashijima, Phys. Rev. D **29**, 1228 (1984).
[5] Y. Okumura, Phys. Lett. B **156**, 259 (1985).
[6] T. Akiba and T. Yamagida, "Hierarchic chiral condensate," Tohoku University preprint TU/85/292.
[7] D. Atkinson, P. Johnson, and M. Koopmans, Z. Phys. C **34**, 99 (1987).
[8] D. Atkinson and P. Johnson, Phys. Rev. D **39**, 1943 (1987).
[9] W. Bardeen, C. Leung, and S. Love, "The dilaton and chiral symmetry breaking," Fermilab-Pub-85/172-T; "Spontaneous symmetry breaking in scale invariant quantum electrodynamics," Fermilab-Pub-85/173-T.
[10] R. Delbourgo and B. Keck, J. Phys. G: Nucl. Phys. **6**, 275 (1980).
[11] L. Landau, A. Abrikosov, and I. Khalatnikov, Nuovo Cimento Suppl. **3**, 80 (1956).
[12] K. Johnson, M. Baker, and R. Willey, Phys. Rev. **136**, B 1111 (1964).
[13] R. Fukuda and T. Kugo, Nucl. Phys. B **117**, 250 (1976); D. Atkinson and D. Blatt, Nucl. Phys. B **151**, 342 (1979).

Chiral symmetry breakdown. II. Ansatz for vertex in ultraviolet

D. Atkinson
Institute for Theoretical Physics, P. O. Box 800, 9700 AV Groningen, The Netherlands

(Received 17 March 1987; accepted for publication 13 May 1987)

An earlier analysis of the Dyson–Schwinger equation for the quark propagator is improved by taking the Slavnov–Taylor identity into account in the ultraviolet. It is found that chiral symmetry breaking occurs above a critical coupling in the Landau gauge; but that this result does not hold in other gauges.

I. INTRODUCTION

Much study has been devoted to the tantalizing possibility that the (constituent) masses of quarks arise from the nonperturbative breakdown of chiral symmetry.[1–5] More specifically, it is supposed that the bare quark mass vanishes; and the Dyson–Schwinger equation for the quark propagator is then analyzed for signs of chiral symmetry breaking. The most popular scenario is that in which this breakdown occurs only if the QCD coupling λ is greater than a certain critical value λ_C: this point constitutes then a bifurcation of the mass function from the trivial to a nontrivial solution of the equation.

Some authors confine themseleves to the Landau gauge and assume that the trace of $\not{p}$ times the inverse quark propagator is p^2. This is only correct if the gluon remains massless. If the gluon acquires an effective mass, as a result of self-interaction, this trace is $p^2\beta(p^2)$, where β is a function that has to be obtained from the Dyson–Schwinger equation. In Refs. 6 and 7, we showed that, in the approximation $\beta(p^2)\equiv 1$, a positive bifurcation point λ_c exists only if both infrared and ultraviolet cutoffs are introduced. In Ref. 8 we elaborated the analysis by treating β properly: in the presence of an infrared cutoff, in the form of an effective gluon mass, and an ultraviolet cutoff, provided naturally by the logarithmic decrease of the running coupling constant, we found again that $\lambda_c > 0$ in the Landau gauge. However, in the Feynman gauge (and in other gauges), it turned out that there is no solution of the equation for $\beta(p^2)$, unless a Pauli–Villars cutoff Λ is introduced. As $\Lambda\to\infty$, so $\lambda_c\to 0$, thus indicating an extreme gauge dependence that casts doubt on the credibility of the approach.

The most questionable approximation made in Ref. 8 is the replacement of the full quark–gluon vertex $\Gamma_\nu(p',p)$ by its bare value γ_ν. Since the difficulties in the Feynman gauge are associated with ultraviolet divergences, and since the inverse quark propagator behaves like $\not{p}\beta(p^2)$ as $p^2\to\infty$, a better approximation for Γ_ν should be γ_ν, multiplied by β, since this is consistent with the Ward–Takahashi identity in the ultraviolet regime. It is true that the correct Slavnov–Taylor identity of a non-Abelian theory contains matrix elements of ghost fields, as Miransky has pointed out[9]; but it might reasonably be hoped that these do not alter the ultraviolet behavior of the quark propagator.

In this paper we undertake a treatment of the quark propagator, with the above-mentioned improvement in the approximation for Γ_ν. We find that the analysis is much easier than that of Ref. 8; but the fundamental conclusions remain unchanged: λ_c is positive in the Landau gauge, and $\lambda_c\to 0$ as $\Lambda\to\infty$ in the Feynman gauge.

Ten years ago, Weinberg[10] suggested that a positive bifurcation point λ_c is not to be expected, since, if it were to exist, it would surely be gauge dependent; and the onset of a phenomenon such as chiral symmetry breaking presumably ought not to depend on the gauge that one chooses. Our conclusions support this conjecture; and, in this connection, a parallel analysis that employs Delbourgo's gauge technique,[11] in which the Ward–Takahashi identity is respected at all momenta, similarly yields a gauge dependence of λ_c.

In Sec. II, we briefly recall the formalism, while the analysis is carried out in Sec. III. An Appendix is devoted to the bifurcation theory that is required in the body of the paper.

In conclusion, although the general result of this work suggests that the existence of a gauge-independent bifurcation point $\lambda_c > 0$ is untenable, the hope might reasonably be entertained that our general methods will yield more positive results in other situations. In particular, in finite-temperature field theory, one expects a phase transition to the plasma state above a critical temperature T_c and bifurcation theory should prove a useful tool.

II. DYSON–SCHWINGER EQUATION AND SLAVNOV–TAYLOR IDENTITY

The Dyson–Schwinger equation for the quark propagator may be written in Euclidean space in the form

$$S_F'^{-1}(p) = \not{p} + \frac{\lambda}{(2\pi)^4}\int d^4p'\,\gamma_\mu S_F'(p')\Gamma_\nu(p',p) \times D'_{F\mu\nu}(p'-p), \tag{2.1}$$

where λ is the square of the QCD coupling constant, times a color factor. Here D'_F is the gluon propagator, and we shall equip it with a mass and a running coupling,

$$D'_{F\mu\nu}(k) = \omega(k^2)D_{F\mu\nu}(k). \tag{2.2}$$

Here D_F is the bare propagator for a massive vector field, and $\omega(x)$ is a given function with the following properties:

$$\omega(0)=1,\quad \omega(x)\sim(\log x)^{-1}\quad \text{as } x\to\infty,\quad \frac{d\omega(x)}{dx}\leqslant 0. \tag{2.3}$$

The Slavnov–Taylor identity, with ghosts neglected, is

$$(p'-p)_\nu\Gamma_\nu(p'-p) = S_F'^{-1}(p) - S_F'^{-1}(p'), \tag{2.4}$$

and this relates the longitudinal part of the quark–gluon vertex to the inverse of the quark propagator. If we set

 0022-2488/87/102494-04$02.50 © 1987 American Institute of Physics

$$S_F'^{-1}(p) = \alpha(p^2) - \not{p}\beta(p^2), \tag{2.5}$$

then we expect that, as $p \to \infty$ at fixed p', (2.4) will read asymptotically

$$p_\nu \Gamma_\nu(p',p) \sim \not{p}\beta(p^2) \tag{2.6}$$

and similarly for $p' \to \infty$ at fixed p. This motivates the ansatz

$$\Gamma_\nu(p',p) \approx \gamma_\nu \beta(p_>^2), \tag{2.7}$$

where $p_>^2 = \max(p^2,p'^2)$, which should respect the ultraviolet behavior of the theory better than does the constant vertex approximation of Ref. 8.

As in Ref. 8, we approximate the running coupling function ω by

$$\omega(k^2) = \omega[(p'-p)^2] \approx \omega(p_>^2), \tag{2.8}$$

and we evaluate the angular integrals in (2.1). This results in the coupled integral equations

$$\alpha(x) = \frac{\lambda}{\pi^2}\int_0^\infty dy\, K(x,y)\,\frac{y\alpha(y)\beta(x_>)}{\alpha^2(y)+y\beta^2(y)}, \tag{2.9}$$

$$\beta(x) = 1 + \frac{\lambda}{\pi^2}\int_0^\infty dy\, L(x,y)\,\frac{y\beta(y)\beta(x_>)}{\alpha^2(y)+y\beta^2(y)}. \tag{2.10}$$

The kernels K and L were given explicitly in Ref. 8, and we do not reproduce them here, nor shall we repeat the discussion of their further approximation.

III. BIFURCATION EQUATIONS

As in I, we shall consider the Feynman gauge, and a modification of the Landau gauge, the so-called Landau-like gauge of Maskawa and Nakajima,[1] for technical convenience. Upon differentiating (2.9) functionally with respect to α, and setting $\alpha = 0$, we obtain the following equations:

$$\delta\alpha(x) = \frac{\lambda}{16\pi^2}\int_0^\infty dy\,\rho(x_>)\beta(x_>)\frac{\delta\alpha(y)}{\beta^2(y)}, \tag{3.1}$$

and

$$\beta(x) = 1 + \frac{\lambda}{16\pi^2}\int_0^\infty dy\,\sigma(x_>)\beta(x_>)\frac{y}{\beta(y)}, \tag{3.2}$$

where $x_> = \max(x,y)$, and where

$$\rho(x) = [4/(x+m^2)]\omega(x), \tag{3.3}$$

$$\sigma(x) = [1/(x+m^2)^2]\omega(x), \tag{3.4}$$

in the Feynman gauge, and

$$\rho(x) = [3/(x+m^2) + m^2/(x+m^2)^2]\omega(x), \tag{3.5}$$

$$\sigma(x) = [m^2/(x+m^2)^3]\omega(x), \tag{3.6}$$

in the Landau-like gauge. Here m is the effective gluon mass, which is assumed to arise from gluon–gluon interaction.

Consider first Eq. (3.2), which can be written

$$\beta(x) = 1 + \frac{\lambda}{16\pi^2}\int_0^x y\,dy\,\sigma(x)\frac{\beta(x)}{\beta(y)} + \frac{\lambda}{16\pi^2}\int_x^\infty y\,dy\,\sigma(y). \tag{3.7}$$

The last integral here is convergent in the Landau-like gauge; but it is log log divergent in the Feynman gauge. Convergence can be achieved in this case by the imposition of a Pauli–Villars cutoff, which has the effect of replacing (3.4) by

$$\sigma(x) = [1/(x+m^2)^2 - 1/(x+\Lambda^2)^2]\omega(x). \tag{3.8}$$

Divide (3.7) throughout by $\beta(x)$ and define $\gamma(x) = [\beta(x)]^{-1}$, thus obtaining

$$\frac{\gamma(x)}{f(x)} = 1 - \frac{\lambda}{16\pi^2}\sigma(x)\int_0^x y\,dy\,\gamma(y), \tag{3.9}$$

where

$$[f(x)]^{-1} = 1 + \frac{\lambda}{16\pi^2}\int_x^\infty y\,dy\,\sigma(y). \tag{3.10}$$

Note that (3.9) is a linear Volterra equation that can be converted into a linear differential equation for $\gamma(x)$. The unique solution of the Volterra equation is

$$\gamma(x) = f(x) - \frac{\lambda}{16\pi^2}\sigma(x)\int_0^x y\,dy\,f^2(y). \tag{3.11}$$

From (3.10) we see that $f(x) \to 1$ as $x \to \infty$, whether we take σ to be given by (3.6), the Landau-like gauge, or by (3.8), the Feynman gauge with Pauli–Villars cutoff. Hence, from (3.11), $\gamma(x) \to 1$ as $x \to \infty$.

Further,

$$\gamma(0) = f(0) = \left[1 + \frac{\lambda}{16\pi^2}\int_0^\infty y\,dy\,\sigma(y)\right]^{-1} > 0; \tag{3.12}$$

and moreover, it is easy to check from (3.11) that

$$\gamma'(x) = -\frac{\lambda}{16\pi^2}\sigma'(x)\int_0^x dy\,f^2(y), \tag{3.13}$$

which is positive, since $\sigma'(x)$ is negative. Hence, as x increases from zero to infinity, $\gamma(x)$ increases monotonically from $f(0)$ to unity, and $\beta(x)$ decreases monotonically from $[f(0)]^{-1}$ to unity.

We turn now to (3.1), which we rewrite

$$\delta\alpha(x) = \frac{\lambda}{16\pi^2}\int_0^\infty dy\,F(x,y)\delta\alpha(y), \tag{3.14}$$

where

$$F(x,y) = \frac{\rho(x)\beta(x)}{\beta^2(y)}\theta(x-y) + \frac{\rho(y)}{\beta(y)}\theta(y-x). \tag{3.15}$$

The kernel F is square integrable,

$$\begin{aligned}\|F\|^2 &= \int_0^\infty dx\int_0^x dy\,\frac{\rho^2(x)\beta^2(x)}{\beta^4(y)} + \int_0^\infty dx\int_x^\infty dy\,\frac{\rho^2(y)}{\beta^2(y)} \\ &\leqslant \left[\frac{1}{[f(0)]^2}+1\right]\int_0^\infty dx\,x\rho^2(x).\end{aligned} \tag{3.16}$$

In the Landau-like gauge,

$$\begin{aligned}\rho(x) &= [3/(x+m^2) + m^2/(x+m^2)^2]\omega(x) \\ &\leqslant [4/(x+m^2)]\omega(x),\end{aligned} \tag{3.17}$$

and the last expression is just ρ in the Feynman gauge. So in both gauges

$$\|F\|^2 \leqslant 16\left[\frac{1}{[f(0)]^2}+1\right]\int_0^\infty dx\,\frac{x}{(x+m^2)^2}\omega^2(x), \tag{3.18}$$

which is convergent, since $\omega^2(x) \sim (\log x)^{-2}$ as $x \to \infty$. Notice that the running coupling function ω is essential for this

convergence. Since (3.14) is a homogeneous Fredholm equation, it only has a nontrivial L^2 solution $\delta\alpha$ for λ on a point set. The smallest positive point in this set, say λ_c, which necessarily satisfies

$$\lambda_c \geqslant 16\pi^2/\|F\|, \tag{3.19}$$

corresponds to the bifurcation of a nontrivial L^2 solution $\alpha(x)$ of Eqs. (2.9) and (2.10) away from the trivial solution (see the Appendix).

Equation (3.14) is equivalent to the differential equation,

$$\frac{d}{dx}\left[\frac{(d/dx)\delta\alpha(x)}{(d/dx)[\rho(x)\beta(x)]}\right] = \frac{\lambda}{16\pi^2}\frac{\delta\alpha(x)}{\beta^2(x)} \tag{3.20}$$

with the boundary condition,

$$\frac{d}{dx}\delta\alpha(x) \underset{x\to 0}{\to} 0. \tag{3.21}$$

According to the general theory of linear, second-order, ordinary differential equations, Eq. (3.20) has two independent solutions, say f_R and f_I, and the general solution of (3.14) is

$$\delta\alpha(x) = Af_R(x) + Bf_I(x); \tag{3.22}$$

and the ratio of A to B is determined by the boundary condition (3.21). The solution is thus unique, up to a normalization.

The ultraviolet behaviors of the regular and irregular solutions follow from the fact that $\beta(x)$ tends to unity as $x\to\infty$, that $\rho(x)$ is given by (3.3) or (3.5), and that $\omega(x)$ satisfies (2.3). We find

$$f_R(x) \sim x^{-1}(\log x)^{-1+b}, \tag{3.23}$$

$$f_I(x) \sim (\log x)^{-b}, \tag{3.24}$$

as $x\to\infty$, where $b = \lambda/4\pi^2$ in the Feynman gauge and $b = 3\lambda/16\pi^2$ in the Landau-like gauge. The solution (3.22) is square integrable only if $B = 0$, and the smallest value of λ for which this happens is precisely λ_c, the bifurcation point.

The whole analysis is applicable to the Landau-like gauge without cutoff, or the Feynman gauge with cutoff. As $\Lambda\to\infty$ in the latter case, however, $\beta(0)\sim\log\log\Lambda$. Subtract $\beta(0)$ from (3.7),

$$\beta(x) = \beta(0) + \frac{\lambda}{16\pi^2}\int_0^x y\,dy\left[\sigma(x)\frac{\beta(x)}{\beta(y)} - \sigma(y)\right], \tag{3.25}$$

and define a renormalized $\tilde\beta(x) = Z_2\beta(x)$, where $Z_2 = [\beta(0)]^{-1}$. The renormalized version of (3.25) is

$$\tilde\beta(x) = 1 + \frac{\lambda Z_2}{16\pi^2}\int_0^x y\,dy\left[\sigma(x)\frac{\tilde\beta(x)}{\tilde\beta(y)} - \sigma(y)\right]. \tag{3.26}$$

As $\Lambda\to\infty$, $Z_2\to 0$ and $\tilde\beta(x)\to 1$. The renormalization constant Z_2 may *not* be absorbed into a redefinition of the coupling $\tilde\lambda = \lambda Z_2$ for the coupling should be renormalized by the gluon renormalization constant Z_3 which we have effectively approximated by unity. In the usual perturbative renormalization, one would expand the integral in (3.26) to order λ^n, and Z_2 to order λ^{n-1}, allowing the infinities to cancel in the usual way. However, the present nonperturbative approach, if it is to be viable, must deal with all divergences in one fell swoop. The renormalized $\delta\tilde\alpha(x) = Z_2\delta\alpha(x)$ satisfies

$$\delta\tilde\alpha(x) = \frac{\lambda Z_2}{16\pi^2}\int_0^\infty dy\,\rho(x_>)\tilde\beta(x_>)\frac{\delta\tilde\alpha(y)}{\tilde\beta^2(y)}, \tag{3.27}$$

from which we see that $\delta\tilde\alpha(x)\to 0$ as $\Lambda\to\infty$. Hence, as the cutoff is removed, the quark propagator tends to the bare form, $(\not p)^{-1}$. Thus we have demonstrated a gauge dependence of a most extreme kind: chiral symmetry breakdown in the Landau-like gauge and none in the Feynman gauge—a most absurd result.

APPENDIX: BIFURCATION THEORY

We present a theorem in bifurcation theory and apply it to the coupled equations (2.9) and (2.10), in a neighborhood of the trivial solution, $\alpha(x) = 0$.

Theorem: Suppose that

$$\alpha(x) = \lambda T(\alpha;x), \tag{A1}$$

where α belongs to some real Hilbert space H, T is a nonlinear operator on H, and λ is a real number. Suppose further that T is thrice Fréchet differentiable, and that

$$T(-\alpha;x) = -T(\alpha,x), \tag{A2}$$

so that $T(0;x) = 0$, which implies that (A1) possesses the trivial solution. Let the first Fréchet derivative at the trivial solution $T'(0;x)$ be compact on H, and suppose that λ_c is such that the linear equation

$$\delta\alpha(x) = \lambda_c[T'(0;\cdot)\delta\alpha](x) \tag{A3}$$

has precisely one nontrivial, linearly independent solution [i.e., λ_c^{-1} belongs to the (point) spectrum of $T'(0;\cdot)$, the corresponding null space of $1-\lambda_c T'(0;\cdot)$ being one dimensional].

Then there exist precisely two nontrivial solutions of (A1), differing only in sign, for λ in a half-neighborhood of λ_c (i.e., $\lambda > \lambda_c$ or $\lambda < \lambda_c$). A proof can be found in Pimbley's book.[12]

In Eqs. (2.9) and (2.10), there is the complication that α and β satisfy coupled equations, the trivial solution corresponding to $\alpha(x)\equiv 0$ and

$$\beta(x) = 1 + \frac{\lambda}{\pi^2}\int_0^\infty dy\,L(x,y)\frac{\beta(x_>)}{\beta(y)}. \tag{A4}$$

However, we can treat β as an implicit function of α. On differentiating (2.9) and (2.10) functionally with respect to α, we find

$$\delta\alpha(x) = \frac{\lambda}{\pi^2}\int_0^\infty y\,dy\,K(x,y) \times\left[\frac{\delta\alpha(y)\beta(x_>) + \alpha(y)\delta\beta(x_>)}{\alpha^2(y) + y\beta^2(y)} - \frac{2\alpha(y)\beta(x_>)[\alpha(y)\delta\alpha(y) + y\beta(y)\delta\beta(y)]}{[\alpha^2(y) + y\beta^2(y)]^2}\right], \tag{A5}$$

$$\delta\beta(x) = \frac{\lambda}{\pi^2}\int_0^\infty y\,dy\,L(x,y) \times\left[\frac{\delta\beta(y)\beta(x_>) + \beta(y)\delta\beta(x_>)}{\alpha^2(y) + y\beta^2(y)} - \frac{2\beta(y)\beta(x_>)[\alpha(y)\delta\alpha(y) + y\beta(y)\delta\beta(y)]}{[\alpha^2(y) + y\beta^2(y)]^2}\right]. \tag{A6}$$

These equations reduce, at $\alpha(x) = 0$, to

$$\delta\alpha(x) = \frac{\lambda}{\pi^2}\int_0^\infty dy\,K(x,y)\frac{\beta(x_>)}{\beta^2(y)}\delta\alpha(y), \tag{A7}$$

$$\delta\beta(x) = \frac{\lambda}{\pi^2}\int_0^\infty dy\,L(x,y)\left[\frac{\delta\beta(x_>)}{\beta(y)} - \frac{\beta(x_>)\delta\beta(y)}{\beta^2(y)}\right]. \tag{A8}$$

The bifurcation equations (A7) and (A4) are, respectively, equivalent to Eqs. (3.1) and (3.2). The possible existence of a nontrivial solution of (A8) is irrelevent to the applicability of the theorem, since Eqs. (A7) and (A8) are decoupled from one another.

We must now check the conditions of the theorem. The space is L^2, and the nonlinear operator T is given in implicit form by Eqs. (2.9) and (2.10). The oddness condition (A2) is clearly satisfied, and it is easy to check that T is thrice Fréchet differentiable. In Sec. III it is shown that (A7) is a classic Fredholm equation, which means that $T'(0;\cdot)$ is compact on L^2. The fact that the null space of $1 - \lambda_c T'(0;\cdot)$ is one dimensional is implied by the analysis of Eq. (3.14) in Sec. III, in which it is shown that the solution is unique, up to a normalization.

[1] T. Maskawa and H. Nakajima, Prog. Theor. Phys. **52**, 1326 (1974); **54**, 860 (1975).
[2] P. I. Fomin, V. P. Gusynin, V. A. Miransky, and Yu. A. Sitenko, Rev. Nuovo Cimento **6**, 1 (1983).
[3] K. Higashijima, Phys. Rev. D **29**, 1228 (1984).
[4] Y. Okumura, Phys. Lett. B **156**, 259 (1985).
[5] R. Acharya and P. Narayana-Swamy, Phys. Rev. D **26**, 2797 (1982); Nuovo Cimento A **86**, 157 (1985); Z. Phys. C **28**, 463 (1985).
[6] D. Atkinson and P. W. Johnson, Phys. Rev. D **35**, 1943 (1987).
[7] D. Atkinson, P. W. Johnson, and M. Koopmans, Z. Phys. C **34**, 99 (1987).
[8] D. Atkinson and P. W. Johnson, J. Math. Phys. **28**, 2488 (1987).
[9] V. A. Miransky, Phys. Lett. B **165**, 401 (1985).
[10] S. Weinberg, Phys. Rev. D **13**, 974 (1976).
[11] D. Atkinson, A. Hulsebos, and P. W. Johnson, "Chiral symmetry breakdown. III. Delbourgo's gauge technique," Groningen preprint, 1986.
[12] G. H. Pimbley, "Eigenfunction branches of nonlinear operators, and their bifurcations," *Lecture Notes in Mathematics*, Vol. 104 (Springer, Berlin, 1969).

The group-theoretical treatment of aberrating systems. III. The classification of asymmetric aberrations

Kurt Bernardo Wolf[a)]
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas/Cuernavaca, Universidad Nacional Autónoma de México, Apdo. Postal 20-726, 01000 México D. F., Mexico

(Received 16 October 1986; accepted for publication 13 May 1987)

A Lie-theoretical classification of the aberrations of systems modeled by asymmetric optical devices is given. The classification is done on the basis of aberration order and axial symmetry of the first-order part. This leads to finite-dimensional (nonunitary) representations of $sp(4,R)$ reduced with respect to its $sp(2,R)$ subalgebra, with helicity and "symplectic spin" labels. Based on pure-magnifier systems, a weight label reproduces and completes Seidel's traditional classification of axis-symmetric aberrations. Based on other first-order systems such as optical fibers, other classification schemes are indicated.

I. INTRODUCTION

In this series of articles,[1,2] we study the Lie theoretical aspects of *aberrating systems*, i.e., systems whose action on phase space is nonlinear and amenable to expansion by aberration *order*.

First order corresponds to linear transformations of an ideal "design" system. Departures from linearity are termed *aberrations*. The model we regard here is that of geometrical optics, but applications and problems also lie in ring and linear accelerator design, wave optics, and radar detection. Reference 3 contains several basic accounts of these directions of inquiry.

In the first two parts of this series, we considered *aligned* lens[1] and fiber[2] systems in detail to third aberration order. Optical alignment means that the elements of such a system are all invariant under rotations around that common optical axis. For these systems, the aberrations have been given names such as spherical and oblique spherical aberration, circular and elliptic coma, astigmatism, curvature, and distortion.[4] These designations have been attributed to Seidel, whose generally quoted paper[5] in fact does not establish the full nomenclature and treats only meridional rays to third aberration order. No visible, systematic classification is known to the author for asymmetric optical systems.

When no symmetry axis is present, the three-dimensional optics of two-dimensional screens requires the *four*-dimensional symplectic algebra $sp(4,R)$. In accelerator design,[6] the program MARYLIE treats $sp(4,R)$-asymmetric magnetic elements; chromatic dispersion requires $sp(6,R)$. Aberrations are handled in a Cartesian basis, by lexicographical order of monomials.

Our purpose here is to enlarge the $sp(2,R)$ symplectic classification[1] to $sp(4,R)$ and to present some specific results in aberration order 4. Section II recapitulates the concepts of optical phase space,[7-9] Lie operators and Lie transformations,[10] needed for the Dragt–Finn factorization[11] of symplectic maps by aberration order. We then present the problem of classification of aberrations in the basis provided by axis-symmetric optical systems of pure magnification. Section III proceeds to construct an appropriate basis for $sp(4,R)$ to accommodate asymmetric aberrations of any order labeled by symplectic spin, Seidel weight, and helicity.

The explicit expressions of the zero-helicity aberration generators has been given in Ref. 12 and appears developed in Ref. 9. They lead, in fact, a close parallel with the states of a symmetric-quantum harmonic oscillator with angular momentum classification of orbitals.[13] The Seidal third- and fifth-order aberrations coincide with the orbitals of the $2s$–$1d$ and $2p$–$1f$ shells in the nuclear model.

The introduction of helicity lifts aberrations out of $sp(2,R)$. Section IV shows how the $sp(4,R)$ multiplets build: each spin-j aberration multiplet unfolds into $2j+1$ copies of different helicities up to j. For aberration orders 2, 3, and 4, there are, respectively, 20, 35, and 56 independent aberrations. Section V contains the analytical formulas for the general case and tables for orders 2 and 3.

The simple model of multipole kicks[14] and quasiflat refracting surfaces is given in Sec. VI. The concluding discussion in Sec. VII gives some pros and cons of the Lie-theoretical classification of aberrations. Selection rules for aberration coefficients in refracting surfaces[9,12] as well as computational simplicity favor the Cartesian monomial classification of aberrations in optical elements. Full optical systems designed on pure magnifier and fiber properties, we contend, may profit from the insight of Lie methods.

II. OPTICAL PHASE SPACE AND sp(4,R)

The phase space of geometrical light rays that cross a *reference* plane $z=0$, is parametrized by a *position* two-vector $\mathbf{q}\in R^2$ (the intersection of the ray with the plane), and a *momentum* two-vector $\mathbf{p}$. The latter is the projection of a three-vector $\vec{n}$ along the ray *on* the reference plane. The length of the three-vector is $n(\mathbf{q})$, the refractive index of the medium at $\mathbf{q}$. We introduce Cartesian coordinates on the plane and write

$$\mathbf{w}=\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix},\quad \mathbf{p}=\begin{pmatrix}p_x\\ p_y\end{pmatrix},\quad \mathbf{q}=\begin{pmatrix}q_x\\ q_y\end{pmatrix}. \tag{2.1}$$

We should distinguish by a sign s, rays in the $+z$ direction ($s=+1$) that we regard as "forward," and "backward" rays in the $-z$ direction ($s=-1$). Optical systems act through *canonical* transformations (i.e., symplectomorphisms) of phase space, preserving the Poisson brackets of

[a)] Member of Centro Internacional de Física y Matemáticas Aplicadas, AC (Mexico).

 0022-2488/87/102498-10$02.50 © 1987 American Institute of Physics

functions f, g thereof,

$$\{f,g\}(\mathbf{w}) = \sum_{j=x,y}\left(\frac{\partial f}{\partial q_j}\frac{\partial g}{\partial p_j} - \frac{\partial f}{p_j}\frac{\partial g}{\partial q_j}\right) = (\hat{fg})(\mathbf{w}), \quad (2.2)$$

at least locally. In the last expression we define the *Lie operator* $\hat{f}$ associated to the function $f(\mathbf{w})$. We use the circumflex notation as in Refs. 1 and 2. Other common notations are $:f:$,[7] $\{f,\circ\}$,[10] f_{0p}.[15] A fundamental property of the association $f \to \hat{f}$ is that it carries Poisson brackets into commutators,

$$(\{f,g\})\hat{} = [\hat{f},\hat{g}]. \quad (2.3)$$

Lie operators generate *Lie transformations*[10] $\mathbb{G}$ through the exponential map

$$\mathbb{G}_f = \exp\hat{f} = \sum_{n=0}^{\infty}\frac{1}{n!}(\hat{f})^n. \quad (2.4)$$

Lie transformations are locally canonical; those having the origin of phase space invariant may be written as a factorized product[11]

$$\mathbb{G}_f = \cdots \exp\hat{f}_5 \exp\hat{f}_4 \exp\hat{f}_3 \exp\hat{f}_2, \quad (2.5)$$

where $f_N(\mathbf{w})$ is a polynomial homogeneous of degree N in the components of $\mathbf{w}$. This is a formal expansion and we cannot at present say much about its *global* properties except in the framework that follows.[7] We simply replace the optical system by a system that exhibits a mechanical-type momentum. We disregard the bound $|\mathbf{p}| \leqslant n(\mathbf{q})$, i.e, that the space of directions is a sphere,[16] in favor of $\mathbf{p} \in R^2$. This allows us to suppress the backward rays to treat metaxial rays in regions still "far" from rays perpendicular to the optical axis.

For $N = \operatorname{degr} f_N$, we note that

$$\operatorname{degr}\{f,g\} = \operatorname{degr} f + \operatorname{degr} g - 2. \quad (2.6)$$

From (2.4) and (2.6) we see thus that $\hat{f}_2$ returns the degree of g and so generates linear transformations of the phase space vector $\mathbf{w}$ corresponding to *paraxial* optical systems (i.e., Gaussian thin lenses, small angles). Such systems are well known to be amenable to 4×4 matrix algebra[6] (2×2 matrices for axis-symmetric systems) that indeed necessitate an R^4 phase space with $\mathbf{p} \in R^2$.

Due to (2.6), the general factor $\exp\hat{f}_N$, $N > 2$, generates *aberrations*, i.e., nonlinear transformations of phase space as

$$\exp\hat{f}_N\mathbf{w} = \mathbf{w} + \{f_N,\mathbf{w}\} + (1/2!)\{f_N,\{f_N,\mathbf{w}\}\} + \cdots$$
$$= \mathbf{w} + \mathbf{w}_{N-1} + \mathbf{w}_{2N-1} + \cdots, \quad (2.7)$$

where $\mathbf{w}_{N-1}(\mathbf{w})$ is a function of degree $A = N-1$ of $\mathbf{w}$, defined as $A = N-1$ aberration *order* of the Lie transformation $\exp\hat{p}_N$. On f_N, the number d_N of independent monomials ${p_x}^{m_x}{p_y}^{m_y}{q_x}^{n_x}{q_y}^{n_y}$, $m_x + m_y + n_x + n_y = N$, is $d_1 = 4$, $d_2 = 10$, $d_3 = 20$, $d_4 = 35$,..., $d_N = \frac{1}{6}(N+1)(N^2+5N+6)$. For $N > 2$, d_N gives the number of independent aberrations of order $A = N-1$.

The problem of *classification* of aberrations is that of labeling them in accord with some clear criterion. The first label, aberration order A, has been given. This is tailored to the Dragt–Finn factorization (2.5) and seems to correspond closely with what is needed in practice. The second criterion we introduce[13] is that our interest lies around *image-forming* devices. Ideally, these perform linear maps of *pure magnification*

$$\exp(\tau\mathbf{p}\cdot\mathbf{q})\hat{}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix} = \begin{pmatrix}e^{\tau} & 0\\ 0 & e^{-\tau}\end{pmatrix}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix}. \quad (2.8)$$

Unfortunately, aberration is unavoidable. This is due to the basic geometric fact that even free propagation by z in a homogeneous medium is already nonlinear: $\mathbf{p} \to \mathbf{p}$ but $\mathbf{q} \to \mathbf{q} + z\mathbf{p}/\sqrt{(n^2-p^2)}$ the last summand is of size $z\tan\theta$, where θ is the angle between the ray and the optical axis. This transformation is generated by the optical Hamiltonian[7] $H(\mathbf{p}) = -\sqrt{(n^2-p^2)}$ (times the sign factor s if we were to include backward rays). Free propagation is the Lie transformation

$$\mathbb{G}_{-zH} = \exp(z\sqrt{n^2-p^2})\hat{}$$
$$= \prod_{k=\infty}^{1}\exp\left(z\frac{(2k-3)!!}{(2k)!!n^{2k-1}}(p^2)^k\right)\hat{}. \quad (2.9)$$

The k th factor above is f_{2k}, function of only the ray direction $\mathbf{p}$, in the form of powers of $p^2 = p_x^2 + p_y^2$. (This is *spherical aberration* of order $A = 2k-1$, excluding $k = 1$, i.e., $N = 2$, $A = 1$, the linear term.) Also, optical refraction surfaces inevitably aberrate.[9]

The ideal magnifier (2.8) has two invariants, $\mathbf{p}\cdot\mathbf{q} = p_xq_x + p_yq_y$ and $\mathbf{p}\times\mathbf{q} = p_xq_y - p_yq_x$. The first is the generator of the transformation, while the second merits some attention, since its square is called the *Petzval* (or skewness) invariant in optics.[7,9] If $\mathbf{p}\times\mathbf{q} = 0$, the ray is *meridional* and contained in a plane with the optical axis; rays for which $\mathbf{p}\times\mathbf{q} \neq 0$ are *skew* rays of the system, and those for which $\mathbf{p}\times\mathbf{q} = |\mathbf{p}|\,|\mathbf{q}|$ are *saggital*. The set of linear transformations for which $\mathbf{p}\times\mathbf{q}$ is an invariant is the group $\mathrm{Sp}(2,R)\times\mathrm{SO}(2)$. The first factor is the set of all axis-symmetric optical systems generated by degree-2 polynomials in the rotation-invariant variables

$$\exp(\alpha p^2 + \beta\mathbf{p}\cdot\mathbf{q} + \gamma q^2)\hat{}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix}$$
$$= \begin{pmatrix}\cos\omega + \beta\,\mathrm{sinc}\,\omega & 2\gamma\,\mathrm{sinc}\,\omega\\ -2\alpha\,\mathrm{sinc}\,\omega & \cos\omega - \beta\,\mathrm{sinc}\,\omega\end{pmatrix}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix}, \quad (2.10a)$$

$$\omega = \pm\sqrt{4\alpha\gamma - \beta^2}, \quad \mathrm{sinc}\,\omega = \omega^{-1}\sin\omega, \quad (2.10b)$$

and the second factor that of pure rotations around the optical axis,

$$\exp(\phi\mathbf{p}\times\mathbf{q})\hat{}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix} = \begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix},$$
$$\mathbf{R} = \begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix}. \quad (2.11)$$

Canonical transformations of phase space preserve areas,[17] and pure-magnification devices, even in geometrical optics, contain a germ of the uncertainty principle in that they allow a reduction of image size $\mathbf{q} \to e^{-\delta}\mathbf{q}$ only at the cost of a spread in directions: $\mathbf{p} \to e^{+\delta}\mathbf{p}$ (in the $\mathbf{p} \in R^2$ model). Pure rotators (2.11), on the other hand, do not exist in axially symmetric light optics.[18] These systems exhibit in addition the discrete space reflection symmetry $p_x \mapsto -p_x$, $q_x \mapsto -q_x$ but $p_y \mapsto p_y$, $q_y \mapsto q_y$ (across any meridional plane). Under reflection, p^2, $\mathbf{p}\cdot\mathbf{q}$, and q^2 do not change sign, but $\mathbf{p}\times\mathbf{q}$ does. Only the former three variables thus may

appear in the aberration functions of axially symmetric optical systems.

The discussion in this section points to our use of the Lie operators of $\mathbf{p}\cdot\mathbf{q}$ and $\mathbf{p}\times\mathbf{q}$ to classify aberrations as, we should note, the associated Lie operators commute since $\{\mathbf{p}\cdot\mathbf{q}, \mathbf{p}\times\mathbf{q}\} = 0$. Through their effect on phase space we see that the former is noncompact and the latter compact. This classification is not *complete*, however, as in the example[7] of the degeneracy of astigmatism [generated by $(\mathbf{p}\cdot\mathbf{q})^2$] and curvature of field (generated by p^2q^2), where both have zero eigenvalue under $(\mathbf{p}\cdot\mathbf{q})\hat{}$ and $(\mathbf{p}\times\mathbf{q})\hat{}$.

We shall now use the known structure of sp(4,R) to accommodate the linearly independent aberrations into multiplets where $(\mathbf{p}\cdot\mathbf{q})\hat{}$ and $(\mathbf{p}\times\mathbf{q})\hat{}$ will be *weight* operators, embedding the extant results[1,12] on axis-symmetric sp(2,R) systems.

III. sp(4,R) AND AXIS-SYMMETRIC ABERRATIONS

The pure-magnifier matrix (2.8) is diagonal, but the pure rotator (2.11) is not. We thus begin by introducing the helicity *basis* of phase space

$$p_\pm = (1/\sqrt{2})(p_x \pm ip_y), \quad q_\pm = (1/\sqrt{2})(q_x \pm iq_y). \tag{3.1}$$

The following expressions and brackets may be seen to hold:

$$p_+p_- = \tfrac{1}{2}p^2, \quad q_+q_- = \tfrac{1}{2}q^2, \tag{3.2a}$$

$$p_+q_- + p_-q_+ = \mathbf{p}\cdot\mathbf{q}, \quad i(p_+q_- - p_-q_+) = \mathbf{p}\times\mathbf{q}. \tag{3.2b}$$

Lie operators are

$$\hat{f} = \frac{\partial f}{\partial q_-}\frac{\partial}{\partial p_+} + \frac{\partial f}{\partial q_+}\frac{\partial}{\partial p_-} - \frac{\partial f}{\partial p_-}\frac{\partial}{\partial q_+} - \frac{\partial f}{\partial p_+}\frac{\partial}{\partial q_-}. \tag{3.3}$$

In particular, the basic Poisson brackets are now

$$\{q_\pm, p_\pm\} = 0, \quad \{q_\pm, p_\mp\} = 1; \quad \{q_\sigma, q_\tau\} = \{p_\sigma, p_\tau\} = 0. \tag{3.4}$$

Magnification distinguishes between $\mathbf{p}$ and $\mathbf{q}$ while rotation classifies $\mathbf{w}_+$ and $\mathbf{w}_-$. In terms of the coordinates p_+, p_-, q_+, q_- we may write the functions corresponding to the generators[18] of sp(4,R) in the Weyl–Cartan basis (see Fig. 1); they are

$$K^+_- = q_+^2, \quad K^+_0 = p_+q_+, \quad K^+_+ = p_+^2, \tag{3.5a}$$

$$K^0_- = q_+q_-, \quad K^0_0 = \tfrac{1}{2}(p_+q_- + p_-q_+),$$

$$K^0_+ = p_+p_-, \tag{3.5b}$$

$$K^-_- = q_-^2, \quad K^-_0 = p_-q_-, \quad K^-_+ = p_-^2; \tag{3.5c}$$

$$L = \tfrac{1}{2}(p_+q_- - p_-q_+). \tag{3.6}$$

The two weight operators are described by

$$K^0_0 = \tfrac{1}{2}\mathbf{p}\cdot\mathbf{q}, \quad L = -(i/2)\mathbf{p}\times\mathbf{q}; \tag{3.7a}$$

$$\hat{K}^0_0 = \frac{1}{2}\left(p_+\frac{\partial}{\partial p_+} + p_-\frac{\partial}{\partial p_-} - q_+\frac{\partial}{\partial q_+} - q_-\frac{\partial}{\partial q_-}\right),$$

$$\hat{L} = \frac{1}{2}\left(p_+\frac{\partial}{\partial p_+} - p_-\frac{\partial}{\partial p_-} + q_+\frac{\partial}{\partial q_+} - q_-\frac{\partial}{\partial q_-}\right), \tag{3.7b}$$

$$\{K^0_0, K^\lambda_\mu\} = \mu K^\lambda_\mu, \quad \{L, K^\lambda_\mu\} = \lambda K^\lambda_\mu, \tag{3.7c}$$

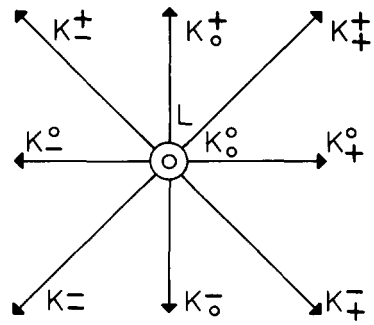


FIG. 1. Root diagram for sp(4,R). (For convenience we tilt it 45° with respect to the usual presentation of the C_2 Cartan root diagram.)

with $\mu,\lambda = -,0,+$. The coordinates of the root vectors in the diagram will be $\alpha = (\mu,\lambda)$.

We have several subalgebras worth noting. Our former work[1,2,9,12] regarded the "horizontal" monomials relevant to axis-symmetric systems, p^2, $\mathbf{p}\cdot\mathbf{q}$, and q^2. These are the generators of the horizontal sp(2,R) subalgebra in Fig. 1:

$$K^0_- = \tfrac{1}{2}q^2, \quad K^0_0 = \tfrac{1}{2}\mathbf{p}\cdot\mathbf{q}, \quad K^0_+ = \tfrac{1}{2}p^2; \tag{3.8a}$$

$$\hat{K}^0_- = q_+\frac{\partial}{\partial p_+} + q_-\frac{\partial}{\partial p_-},$$

$$\hat{K}^0_+ = -p_+\frac{\partial}{\partial q_+} - p_-\frac{\partial}{\partial q_-}; \tag{3.8b}$$

$$\{K^0_+, K^0_-\} = -2K^0_0. \tag{3.9}$$

In Ref. 9, the K^0_σ were denoted simply as K_σ.

A second subalgebra is the "vertical" one of Fig. 1, generated by K^+_0, L, and K^-_0. Since $\{K^+_0, K^-_0\} = 2L$ it is the compact su(2) subalgebra. The two "diagonal" subalgebras are $\{\frac{1}{2}K^+_+, \frac{1}{2}K^-_-\} - 2U$ and $\{\frac{1}{2}K^-_+, \frac{1}{2}K^+_-\} = -2V$ with $U = \frac{1}{2}(K^0_0 + L)$ and $V = \frac{1}{2}(K^0_0 - L)$. The factor $\frac{1}{2}$ in the Poisson brackets come from the roots being the "long" ones of the algebra. Poisson brackets between functions corresponding to root vectors follow the standard form[19] $[\hat{X}_\alpha, \hat{X}_\beta] = N_{\alpha\beta}\hat{X}_{\alpha+\beta}$, $N_{\alpha\beta} = N_{-\alpha,-\beta} = -N_{\beta\alpha}$. The root vectors at 90°, if long, commute; if short, $N_{(0,+),(+,0)} = N_{(0,+),(-,0)} = -1$. At 135° we have $N_{(0,+),(-,+)} = N_{(0,+),(-,-)} = -2$ and $N_{(+,+),(0,-)} = -N_{(-,+),(0,-)} = -2$. [The signs for the compact algebra usp(4) are simpler and may be put as $N_{\alpha\beta} = |\alpha+\beta|^2$.]

For axis-symmetric systems involving only the horizontal sp(2,R) of Fig. 1, the classification of aberration generating functions f_N, $N = 2k$ even, is performed in the following way. We define the coordinates[12]

$$-(1/\sqrt{2})(\xi_1 + i\xi_2) = \xi_+ = (1/\sqrt{2})p^2 = \sqrt{2}p_+p_-, \tag{3.10a}$$

$$\xi_3 = \xi_0 = \mathbf{p}\cdot\mathbf{q} = p_+q_- + p_-q_+, \tag{3.10b}$$

$$(1/\sqrt{2})(\xi_1 - i\xi_2) = \xi_- = (1/\sqrt{2})q^2 = \sqrt{2}q_+q_-, \tag{3.10c}$$

and

$$\eta = \mathbf{p}\times\mathbf{q} = i(p_+q_- - p_-q_+). \tag{3.11}$$

For $\mathbf{p}$ and $\mathbf{q}$ real, the $\xi_\pm$, ξ_0, ξ_1, and η are real while ξ_2 is pure imaginary. In the R^3 ξ space, the Sp(2,R) action (2.10)

leaves invariant the spheres

$$\xi^2 = \xi_1^2 + \xi_2^2 + \xi_3^2 = \xi_0^2 - 2\xi_+\xi_- = (\mathbf{p}\cdot\mathbf{q})^2 - p^2q^2 = -\eta^2, \tag{3.12}$$

where η^2 is the Petzval (skewness) invariant. In terms of these coordinates, we build[12] the solid spherical harmonics [Ref. 20, Eq. (3.153)]

$$\mathscr{Y}^j_m(\xi) = [(2j+1)(j+m)!(j-m)!/4\pi]^{1/2} \times \sum_n \frac{1}{2^{m+2n}} \frac{(p^2)^{m+n}}{(m+n)!} \frac{(\mathbf{p}\cdot\mathbf{q})^{j-m-2n}}{(j-m-2n)!} \frac{(q^2)^n}{n!}, \tag{3.13a}$$

with $|m| \leqslant j$ integers, and finally we build

$$^k\mathscr{Y}^j_m(\mathbf{p},\mathbf{q}) = \eta^{k-j}\mathscr{Y}^j_m(\xi). \tag{3.13b}$$

The latter is an eigenfunction of $\hat{K}^0_0$ with eigenvalue m, subject to raising by $\hat{K}^0_+$ up to $m=j$ and lowering by $\hat{K}^0_-$ down to $m=-j$, and of $\hat{L}$ with *helicity* eigenvalue $\lambda = 0$. These functions provide the basis for the horizontal sp(2,R) aberrations on the zero-helicity plane; $k-j = 0,2,4,...$, $k-1$ or k, i.e., $j = k, k-2, ..., 1$ or 0.

In axial systems we may have only even powers of η; odd powers of η are not allowed by reflection symmetry across meridional planes. In Ref. 13 we remarked that this *symplectic Seidel* classification of aberrations placed them in one-to-one correspondence with the states of a quantum harmonic oscillator, Ψ_{kjm}, with k energy quanta, angular momentum j, and "magnetic" projection m along the axis $\xi_0 = \mathbf{p}\cdot\mathbf{q}$, i.e., pure-magnifier systems. The magnetic classification axis may be chosen to conform to other systems such as fibers, where $\frac{1}{2}(p^2+q^2) = -i\xi_2$ is more convenient. An su(3) algebra of operators $\xi_i/\partial\xi_j$, $i,j = 1,2,3$, may be formally set up to accommodate the aberrations of a given order $A = 2k-1$ into completely symmetric su(3) multiplets reduced with respect to so(3), characterized by eigenvalues under the number operator

$$\hat{N} = p_+\frac{\partial}{\partial p_+} + p_-\frac{\partial}{\partial p_-} + q_+\frac{\partial}{\partial q_+} + q_-\frac{\partial}{\partial q_-}, \tag{3.14a}$$

$$\hat{N}\,^k\mathscr{Y}^j_m(\mathbf{p},\mathbf{q}) = 2k\,^k\mathscr{Y}^j_m(\mathbf{p},\mathbf{q}). \tag{3.14b}$$

Note that $\hat{N}$ commutes with the sp(2,R) algebra (3.5) and (3.6), and is *extraneous* to sp(4,R) in the sense that there is *no* function $N(\mathbf{p},\mathbf{q})$ whose Lie operator is $\hat{N}$. The same remark holds for the other su(3) generators except for the three obtained from (3.10), basically the so(3) subalgebra of su(3).

It is not clear to us whether the algebra su(3) can be used beyond its role as a suggestive classification scheme. Recall that three-dimensional classical systems always admit an su(3) algebra[21] whose generators may be quite complicated functions of phase space.

The sp(4,R) Casimir operator is a function of the number operator

$$C = (\hat{K}^0_0)^2 - \tfrac{1}{2}\{\hat{K}^0_+,\hat{K}^0_-\}_+ + (\hat{L})^2 + \tfrac{1}{2}\{\hat{K}^+_0,\hat{K}^-_0\}_+ - \tfrac{1}{4}\{\hat{K}^+_+,\hat{K}^-_-\}_+ - \tfrac{1}{4}\{\hat{K}^+_-,\hat{K}^-_+\}_+ = \tfrac{1}{2}\hat{N}(\hat{N}+4), \tag{3.15}$$

where $\{\cdot,\cdot\}_+$ is the anticommutator. The corresponding classical function built out of the K^λ_μ's is identically zero. The eigenvalues of (3.15) are $2k(k+2)$. Since the representations are built out of the symmetric product of basis functions w_iw_j, etc., it is the totally symmetric one, with null fourth-degree Casimir operator. Also, from (3.13b) it is clear that the representations of the horizontal sp(2,R) contained therein are in general

$$j = N/2 = \tfrac{1}{2}(A+1) = k, k-1, k-2, ..., \tfrac{1}{2} \text{ or } 0. \tag{3.16}$$

We are dealing thus with the finite-dimensional (nonunitary) irreducible representations of sp(4,R) and sp(2,R); they will be in correspondence with the compact irreps of usp(4) = so(6) and usp(2) = su(2) = so(3).

IV. THE HELICITY OF ABERRATION MULTIPLETS

The basic dynamical observables of geometric optics (in a reference plane) are the two position components q_+, q_- and the two momentum components p_+, p_- given in Eq. (3.1). These are arranged in the lowest sp(4,R) multiplet shown in Fig. 2. This is not an *aberration* multiplet since these functions do not appear except in products for higher aberration multiplets. There are two horizontal sp(2,R) doublets q_+, p_+ and q_-, p_-.

Simple (symmetric) product of two basic sp(4,R) multiplets yields the adjoint representation mulltiplet shown in Fig. 3, basically a reproduction of the root diagram in Fig. 1. These functions also do not generate aberrations since their degree is 2 and they belong to f_2 functions. They allow us, however, to see the role that the vertical su(2) subalgebra plays in changing the *helicity* [i.e., the eigenvalue of $\hat{L} = (i/2)(\mathbf{p}\times\mathbf{q})$] of the extreme $-m$ sp(2,R) states. The subalgebra generators are

$$\hat{K}^+_0 = p_+\frac{\partial}{\partial p_-} - q_+\frac{\partial}{\partial q_-}, \quad \hat{K}^-_0 = p_-\frac{\partial}{\partial p_+} - q_-\frac{\partial}{\partial q_+} \tag{4.1}$$

and $\hat{L}$ is given in (3.7b).

For $m=j$, the vertical su(2) multplets will have the form

$$\mathscr{U}^j_\lambda = 2^j p_+^{j+\lambda} p_-^{j-\lambda} = \begin{cases} (p^2)^{j-\lambda}(\sqrt{2}p_+)^{2\lambda}, & j \geqslant \lambda \geqslant 0, \\ (p^2)^{j+\lambda}(\sqrt{2}p_-)^{-2\lambda}, & -j \leqslant \lambda \leqslant 0, \end{cases} \tag{4.2}$$

with the same spin-j integer or half-integer and helicity eigenvalue λ under $\hat{L}$. For $\lambda = 0$ we have equal amounts of p_+'s and p_-'s; we are then in one of the axis-symmetric aberrations (3.13).

Now, we follow the highest $\lambda = j$ (or lowest $\lambda = -j$) helicity sp(2,R) multiplet, starting from $\mathscr{U}^j_{\pm j} = 2^j p^j_\pm$, and

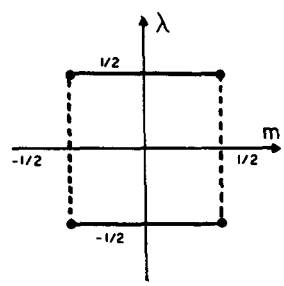


FIG. 2. The basic sp(4,R) multiplet. Continuous lines join the two "horizontal" sp(2,R) doublets.

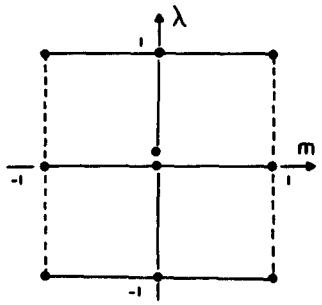


FIG. 3. The adjoint sp(4,R) multiplet. Dotted lines join the "vertical" su(2) multiplets.

move down sp(2,R) with $\hat{K}^0_-$. This yields

$$S_m^{\pm j} = 2^j p_\pm^{j+m} q_\pm^{j-m}, \quad m = j, j-1, \ldots, -j, \quad j \geqslant 0. \qquad (4.3)$$

Both the axis-symmetric aberration functions (3.13) and the highest-helicity aberration functions (4.3) transform in the same way under the group Sp(2,R) of paraxial, axis-symmetric optical transformations.

The simple product of two highest-weight states, $(p^2)^{j_1}$ and $(p_\pm)^{2j_2}$ constitutes a highest-weight state of sp(2,R) with Seidel weight $m = j = j_1 + j_2$ and helicity $\lambda = \pm j_2$. In particular, (4.2) shows that the sp(2,R) highest states $\mathscr{Y}^j_\lambda$ of spin $m = j$ and helicity $\lambda = \Lambda|\lambda|, \Lambda = \operatorname{sgn}\lambda$, are products between $\mathscr{Y}^{j-|\lambda|}_{j-|\lambda|}(p^2)$ and $S^\lambda_{|\lambda|}(p_\Lambda)$.

When $\mathscr{Y}^j_\lambda$ is subject to repeated action of the sp(2,R) lowering operator $\hat{K}^0_-$, it yields the m partners of the sp(2,R) multiplet of spin j, through $m = j, j-1, \ldots, -j$. These will be denoted ${}^k\mathscr{Z}^{j,\lambda}_m(\mathbf{p},\mathbf{q})$ and given below. So, every multiplet of axis-symmetric aberrations of integer spin j gets j positive-helicity partners $\lambda = 1,2,\ldots, j$ and j negative-helicity partners $\lambda = -1, -2, \ldots, -j$; their highest $m = j$ states belong to a $(2j+1)$-dimensional vertical su(2) multiplet. This is shown in Figs. 4 and 5 for aberration orders 2 and 3.

The coupling of $\mathscr{Y}^{j-|\lambda|}$ and S^λ to aberrations $\mathscr{Z}^{j,\lambda}$ of total symplectic spin j and helicity λ thus takes place through the "completely stretched" Wigner coefficients $C^{j-|\lambda|\,|\lambda|\,j}_{m_1,m_2,m_1+m_2}$ that will be detailed in the next section.

What has been said for the sp(4,R) multiplet that contains $(p^2)^j$ [the five sp(2,R) quintuplets of Fig. 5] is valid for the three triplets in the same figure, except that we need not resort to su(2)-lowering arguments [which mix sp(2,R)

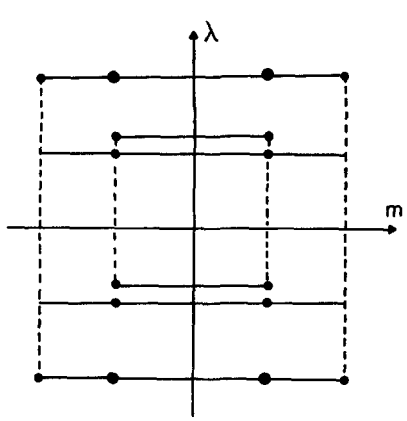


FIG. 4. The sp(4,R) multiplet corresponding to second-order aberrations.

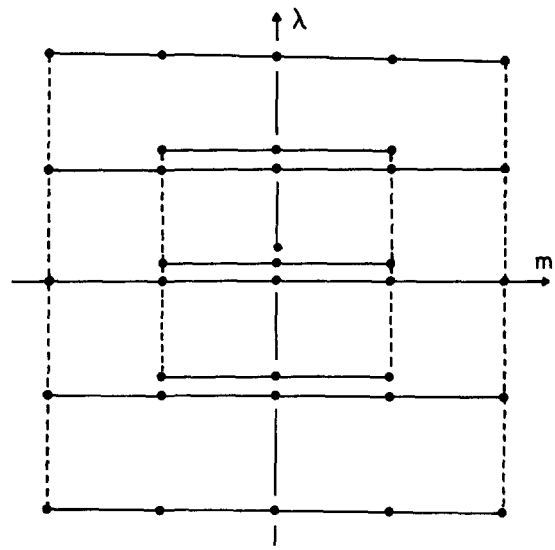


FIG. 5. The sp(4,R) multiplet corresponding to third-order aberrations.

multiplets]. It is clear that if we have an sp(4,R) multiplet of aberrations labeled ${}^k\mathscr{Z}^{j,\lambda}_m$ and we multiply by the sp(2,R) singlet $\eta = \mathbf{p}\times\mathbf{q}$ of zero weight and degree 2, we obtain aberrations ${}^{k+1}\mathscr{Z}^{j,\lambda}_m$ in an sp(4,R) multiplet corresponding to aberrations of order A increased by 2. In this way, the three sp(2,R) triplets and the singlet in Fig. 5 for third aberration order, are the higher repeaters of the three triplets and singlet of first order in Fig. 3.

A similar compounding between spin and helicity applies to half-integer symplectic spin multiplets corresponding to even aberration order. The latter contain no axis-symmetric aberration multiplets. Thus Fig. 4 contains the highest $j = k = \frac{3}{2}$ $(4\times4 = 16)$-plet, none of whose members have zero helicity; and a $j = k - 1 = \frac{1}{2}$ $(2\times2 = 4)$-plet, repeater by one power of η of the basic representation of sp(4,R) in Fig. 2.

In abstract, thus the sp(4,R)$\supset$sp(2,R) classification scheme we propose here yields the aberration functions ${}^k\mathscr{Z}^{j,\lambda}_m(\mathbf{p},\mathbf{q})$ labeled by $(k,j;\lambda,m)$, where we have the following.

k: labels the aberration *order* $A = 2k - 1 = 2,3,4,\ldots$, by $k = \frac{3}{2}, 2, \frac{5}{2}, \ldots$; it is the eigenvalue of $\hat{N}$, the number operator (3.14a), $N = 2k = A + 1$.

j: symplectic *spin*, $j = k, k-1, \ldots \frac{1}{2}$ or 0. The power of the skewness variable $\eta = \mathbf{p}\times\mathbf{q}$ in the aberration function is $k - j$.

λ: *helicity*, $\lambda = j, j-1, \ldots, -j$, eigenvalue of $\hat{L}$ in (3.7b); gives the excess of $\mathbf{w}_+$'s over $\mathbf{w}_-$'s.

m: *Seidel weight*, $m = j, j-1, \ldots, -j$, eigenvalue of $\hat{K}^0_0$ in (3.7a); gives the excess of $\mathbf{p}$'s over $\mathbf{q}$'s.

Axis-symmetric paraxial transformations of aberrating systems will only mix m's, respecting k, j, and λ.

We may write the generating polynomial of an asymmetric aberrating system of order $A = 2k - 1$ as

$$f_{2k}(\mathbf{p},\mathbf{q}) = \sum_{j=0\text{ or }\frac{1}{2}}^{k} \sum_{\lambda=-j}^{j} \sum_{m=-j}^{j} {}^k\sigma^{j,\lambda}_m \, {}^k\mathscr{Z}^{j,\lambda}_m(\mathbf{p},\mathbf{q}), \qquad (4.4)$$

where the ${}^k\sigma^{j,\lambda}_m$ are the sp(4,R)$\supset$sp(2,R)-Seidel aberration

coefficients. The generic association between the Seidel m-label and the christened axis-symmetric aberrations[4] was given in Ref. 13. Enclosing in parentheses aberrations that are present only for order 5 or higher, they are

$m=j$	spherical aberration,
$m=j-1$	(circular) coma,
$m=j-2$	(oblique spherical aberration),
$m=j-3$	(nameless),
⋮	⋮
$m=3-j$	(elliptical coma),
$m=2-j$	curvature of field/astigmatism,
$m=1-j$	distortion,
$m=-j$	pocus.[1]

The traditional classification is neither complete (pocus has not openly appeared), nondegenerate (by j), nor is it easy to see departure directions for asymmetries. We offer (4.4) as an attractive alternative.

We end this section with a word on reality: p_x, p_y, q_x, and q_y are real, so $p_+^* = p_-$ and $q_+^* = q_-$ are complex conjugates; η is real. It follows that $(\mathscr{Z}^\lambda)^* = \mathscr{Z}^{-\lambda}$, as may be seen on the sp(2,R)-highest weight states. If $f_{2k}(\mathbf{p},\mathbf{q})$ is to be a real polynomial, then the complex aberration coefficients in (4.4) must relate as

$$({}^k\mathscr{C}_m^{j,\lambda})^* = {}^k\mathscr{C}_m^{j,-\lambda}. \tag{4.5}$$

The operation of complex conjugation $p_+ \mapsto p_-$, $q_+ \mapsto q_-$ is equivalent to the reflection $p_y \mapsto -p_y$, $q_y \mapsto -q_y$. If the asymmetric system is even (or odd) under the latter transformation, then ${}^k\mathscr{C}_m^{j,\lambda} = +{}^k\mathscr{C}_m^{j,-\lambda}$ (or $-$) and the ${}^k\mathscr{C}_m^{j,\lambda}$ are real (or pure imaginary). Reflecting across the orthogonal line $p_x \mapsto p - p_x$, $q_x \mapsto -q_x$ effects $p_+ \mapsto -p_-$ and $q_+ \mapsto -q_-$, placing a factor of $(-1)^{2k}$ to what was said above for aberrations of orders 2,4,... ($k = \frac{3}{2}, \frac{5}{2}, \ldots$). If the asymmetric system is even (or odd) under the last reflection, the ${}^k\mathscr{C}_m^{j\lambda}$ are pure imaginary (or real). If both reflection symmetries are present and even ($\mathbf{p} \to -\mathbf{p}$, $\mathbf{q} \to -\mathbf{q}$) then all even-ordered aberrations will be zero.

V. THE ASYMMETRIC ABERRATION FUNCTIONS

There is an evident advantage in uniform notation for the identification of the aberration functions with standard special functions such as the solid spherical harmonics $\mathscr{Y}_m^j$, normalized by integration over the unit sphere (Ref. 20, Sec. 2.10). These functions involve numerical square roots, as (3.13) shows, for factors of the functions themselves, under the action of $\hat{K}^0_{\pm}$, and in the representation matrices carrying the Sp(2,R) paraxial transformations.

In geometric optics we have not yet required the integral over the sphere, i.e., the subspace of rays with fixed skewness. Moreover, square-root factors are not included in the traditional scale of the known axis-symmetric aberration coefficients.[7] Finally, symbolic and numeric computer programs[22] run faster when no square roots are present. For this reason we introduced in Ref. 1, and kept in Ref. 9, the *unnormalized* symplectic aberration polynomials[23]

$$\begin{aligned}
\mathscr{C}_m^j(\mathbf{p},\mathbf{q}) &= {}^j\mathscr{C}_m^j(p^2,\mathbf{p}\cdot\mathbf{q},q^2) \\
&= \sqrt{\frac{4\pi(2j+1)(j+m)!(j-m)!}{(2j-1)!!}}\,\mathscr{Y}_m^j(\xi) \\
&= \frac{j!(j+m)!(j-m)!}{(2j)!}\sum_n \frac{(p^2)^{m+n}}{(m+n)!} \\
&\quad\times\frac{(2\mathbf{p}\cdot\mathbf{q})^{j-m-2n}}{(j-m-2n)!}\,\frac{(q^2)^n}{n!},
\end{aligned} \tag{5.1a}$$

$${}^k\mathscr{C}_m^j = (\mathbf{p}\times\mathbf{q})^{k-j}\mathscr{C}_m^j. \tag{5.1b}$$

As a check, the coefficients of the terms $(p^2)^a(p\cdot q)^b(q^2)^c$ are all positive and sum to unity.

The aberration polynomials (5.1) are such that

$${}^j\mathscr{C}_j^j = (p^2)^j = (2p_+p_-)^j, \quad {}^j\mathscr{C}_{-j}^j = (q^2)^j = (2q_+q_-)^j. \tag{5.2}$$

They raise and lower through

$$\begin{aligned}
\hat{K}^0_+\,{}^k\mathscr{C}_m^j &= (m-j)\,{}^k\mathscr{C}_{m+1}^j, \\
\hat{K}^0_-\,{}^k\mathscr{C}_m^j &= (m+j)\,{}^k\mathscr{C}_{m-1}^j.
\end{aligned} \tag{5.3a}$$

Indeed, our choice of the extreme-helicity states in (4.3) was made so that also

$$\begin{aligned}
\hat{K}^0_+ S_m^\lambda &= (m-|\lambda|)S_{m+1}^\lambda, \\
\hat{K}^0_- S_m^\lambda &= (m+|\lambda|)S_{m-1}^\lambda,
\end{aligned} \tag{5.3b}$$

with the same form for the coefficients.

Now we build the sp(4,R) symplectic harmonics labeled in the last section through defining first the highest-weight state ($\Lambda = \operatorname{sgn}\lambda$),

$$\begin{aligned}
{}^k\mathscr{Z}_j^{j,\lambda} &= \eta^{k-j}\mathscr{C}_{j-|\lambda|}^j S_{|\lambda|}^\lambda \\
&= (\mathbf{p}\times\mathbf{q})^{k-j}(p^2)^{j-|\lambda|}(\sqrt{2}p_\Lambda)^{2|\lambda|}.
\end{aligned} \tag{5.4}$$

We construct the rest of the j multiplet through demanding that (5.3) hold for the ${}^k\mathscr{Z}_m^{j,\lambda}$ as well. Applying $\hat{K}^0_-$ to (5.4) we see that the general weight $-m$ aberration polynomial is given by the linear combination

$${}^k\mathscr{Z}_m^{j,\lambda} = \eta^{k-j}\sum_\mu C_{m,\mu}^{j,|\lambda|}\,\mathscr{C}_{m-\mu}^{j-|\lambda|}S_\mu^\lambda, \tag{5.5a}$$

where the coefficient $C_{m,\mu}^{j,|\lambda|}$ is related by root factors to the stretched su(2) Wigner coefficient [Ref. 20, Eq. (6.177)]. Recursion relations found from applying $\hat{K}^0_-$ and the base case (5.4) yield a coefficient in terms of binomial coefficients with no square roots,

$$C_{m,\mu}^{j,|\lambda|} = \binom{j+m}{|\lambda|+\mu}\binom{j-m}{|\lambda|-\mu}\binom{2j}{2|\lambda|}^{-1}. \tag{5.5b}$$

We note that the coefficients sum to unity.

The closed analytic form for the asymmetric aberration polynomials is thus

$$\begin{aligned}
&{}^k\mathscr{Z}_m^{j,\lambda}(\mathbf{p},\mathbf{q}) \\
&= [(2j)!]^{-1}(2|\lambda|)!2^{|\lambda|}(j-|\lambda|)!(j+m)!(j-m)! \\
&\quad\times(\mathbf{p}\times\mathbf{q})^{k-j}\sum_\mu \frac{p_\Lambda^{|\lambda|+\mu}}{(|\lambda|+\mu)!}\,\frac{q_\Lambda^{|\lambda|-\mu}}{(|\lambda|-\mu)!} \\
&\quad\times\sum_n\frac{(p^2)^{m-\mu+n}}{(m-\mu+n)!} \\
&\quad\times\frac{(2\mathbf{p}\cdot\mathbf{q})^{j-|\lambda|-m+\mu-2n}}{(j-|\lambda|-m+\mu-2n)!}\,\frac{(q^2)^n}{n!},
\end{aligned} \tag{5.6}$$

TABLE I. Second-order aberration polynomials.

$$^{3/2}\mathscr{L}^{3/2,3/2}_{3/2} = S^{3/2}_{3/2} = 2^{3/2}p^3_+$$

$$^{3/2}\mathscr{L}^{3/2,3/2}_{1/2} = S^{3/2}_{1/2} = 2^{3/2}p^2_+\, q_+$$

$$^{3/2}\mathscr{L}^{3/2,1/2}_{3/2} = \mathscr{L}^1_1 S^{1/2}_{1/2} = 2^{1/2}p^2 p_+ = 2^{3/2}p^2_+\, p_-$$

$$^{3/2}\mathscr{L}^{3/2,1/2}_{1/2} = \tfrac{1}{3}\mathscr{L}^1_1 S^{1/2}_{-1/2} + \tfrac{2}{3}\mathscr{L}^1_0 S^{1/2}_{1/2} = 2^{1/2}\tfrac{1}{3}(p^2 q_+ + 2\mathbf{p}\cdot\mathbf{q}\, p_+)$$

$$^{3/2}\mathscr{L}^{1/2,1/2}_{1/2} = \eta S^{1/2}_{1/2} = 2^{1/2}\mathbf{p}\times\mathbf{q}\, p_+$$

where the coefficients of $p^{m_+}_+ p^{m_-}_- q^{n_+}_+ q^{n_-}_-$ sum to 2^j. In the multiplet diagrams, reflection in helicity is

$$^k\mathscr{L}^{j,-\lambda}_m(\mathbf{p},\mathbf{q}) = (^k\mathscr{L}^{j,\lambda}_m(\mathbf{p},\mathbf{q}))^* = {}^k\mathscr{L}^{j\lambda}_m(\mathbf{w}_+ \leftrightarrow \mathbf{w}_-). \tag{5.7a}$$

Similarly, reflection in Seidel weight is obtained as

$$^k\mathscr{L}^{j,\lambda}_{-m}(\mathbf{p},\mathbf{q}) = (-1)^{k-j}\,{}^k\mathscr{L}^{j,\lambda}_m(\mathbf{q},\mathbf{p}). \tag{5.7b}$$

The last operation exchanges $\mathbf{p}$'s and $\mathbf{q}$'s, including the η factor; hence the sign.

In Tables I and II we give the symplectic aberration polynomials of orders 2 and 3 (Figs. 4 and 5) in the normalization (5.6) for the upper-right quadrant of the multiplet. Reflections in λ and m through (5.7) yield the full multiplet.

As we said in Sec. II, for a given aberration order there are $d_{2k} = \frac{1}{3}(2k+1)(2k^2+5k+3)$ independent $2k$ th-order monomials in $p^{m_x}_x p^{m_y}_y q_x{}^{n_x} q_y{}^{n_y}$; we can count the same number of $^k\mathscr{L}$'s. The coefficients to pass from the latter to the former are present in (5.6); the inverse transformation coefficients have a recursion relation reported in Ref. 12 for zero helicity.

The role of the aberration polynomials $^k\mathscr{L}^{j,\lambda}_m$ is not only to generate aberrations, but to serve as homogeneous space for the optical group action linearized thereby. If we return to (2.7) with f_{2k} written in the symplectic basis as (4.4) and write the basic quadruplet $^{1/2}\mathscr{L}^{1/2\,\lambda}_m$ for $\mathbf{w}$, then clearly we are in need of expressions of the Poisson brackets between the $\mathscr{L}$'s. To our knowledge, this operation has not been studied within representation theory. We may state, nevertheless, the following structure with selection rules for the indices:

$$\{{}^k\mathscr{L}^{j,\lambda}_m,\ {}^{k'}\mathscr{L}^{j',\lambda'}_{m'}\} = \sum_{j''=|m+m'|}^{j+j'-1} P^{jj'j''}_{m,m'}\ {}^{k+k'-1}\mathscr{L}^{j'',\lambda+\lambda'}_{m+m'}. \tag{5.8}$$

This also defines the Lie structures of the universal covering algebra of the one generated by the basic quadruplet, and those of its aberration algebras of order A briefly constructed in Ref. 1. We note in (5.8) that the weight labels λ and m compose additively, k is diminished by 1 [cf. (2.6)], and the sum ranges only over the j'' values present at the point $(m+m',\lambda+\lambda')$ on the multiplet diagram. The maximal multiplicity occurs nearest to (0,0) and is the integer part of $k+1$ for asymmetric aberrations. For zero helicity, the result (5.8) involves only sp(2,R) and the coefficients are given in terms of so(3) Wigner coefficients in Ref. 9.

TABLE. II. Third-order aberration polynomials.

$$^2\mathscr{L}^{2,2}_2 = S^2_2 = 4p^4_+$$

$$^2\mathscr{L}^{2,2}_1 = S^2_1 = 4p^3_+\, q_+$$

$$^2\mathscr{L}^{2,2}_0 = S^2_0 = 4p^2_+\, q^2_+$$

$$^2\mathscr{L}^{2,1}_2 = {}^1\mathscr{L}^1_1 S^1_1 = 2p^2 p^2_+ = 4p^3_+\, p_-$$

$$^2\mathscr{L}^{2,1}_1 = \tfrac{1}{2}\,{}^1\mathscr{L}^1_1 S^1_0 + \tfrac{1}{2}\,{}^1\mathscr{L}^1_0 S^1_1 = p^2 p_+ q_+ + \mathbf{p}\cdot\mathbf{q}\, p^2_+$$

$$^2\mathscr{L}^{2,1}_0 = \tfrac{1}{6}\,{}^1\mathscr{L}^1_1 S^1_{-1} + \tfrac{2}{3}\,{}^1\mathscr{L}^1_0 S^1_0 + \tfrac{1}{6}\,{}^1\mathscr{L}^1_{-1} S^1_1$$
$$= \tfrac{1}{3}(p^2 q^2_+ + 4\mathbf{p}\cdot\mathbf{q}\, p_+ q_+ + q^2 p^2_+) = 2\mathbf{p}\cdot\mathbf{q}\, p_+ q_+$$

$$^2\mathscr{L}^{2,0}_2 = {}^1\mathscr{L}^2_2 = (p^2)^2 = 4p^2_+\, p^2_-$$

$$^2\mathscr{L}^{2,0}_1 = {}^1\mathscr{L}^2_1 = p^2\mathbf{p}\cdot\mathbf{q}$$

$$^2\mathscr{L}^{2,0}_0 = {}^1\mathscr{L}^2_0 = \tfrac{1}{3}[p^2 q^2 + 2(\mathbf{p}\cdot\mathbf{q})^2]$$

$$^2\mathscr{L}^{1,1}_m = \eta S^1_m$$

$$^2\mathscr{L}^{1,0}_m = \eta\, {}^1\mathscr{L}^1_m$$

$$^2\mathscr{L}^{0,0}_0 = \eta^2$$

VI. EXAMPLES: FLAT AND QUASIFLAT REFRACTING SURFACES TO FOURTH ABERRATION ORDER

Let us consider two optical model elements to visualize the effect of asymmetric aberration on light rays: flat and quasiflat refracting surfaces.

Flat refracting surfaces exist as Fresnel lenses, such as may be seen in the back windows of some vans, with groves that may not be circular and/or of varying depth. The action of a flat Fresnel lens on optical phase space is to change the direction of all rays at the surface, namely $\mathbf{p}\mapsto\mathbf{p}'(\mathbf{p},\mathbf{q})$ and $\mathbf{q}\mapsto\mathbf{q}'=\mathbf{q}$. Since the transformation is canonical, $\mathbf{p}'$ must be of the form $\mathbf{p}+\boldsymbol{\kappa}'(\mathbf{q})$ and this is produced by the operator

$$\mathbb{G}_\kappa = \exp(\kappa(\mathbf{q})\hat{}). \tag{6.1}$$

We shall call $\kappa(\mathbf{q})$ the *kick* function since the model also applies to a potential kick in mechanical systems where the z axis is time, and we shall speak of *multipoles*, as in the thin-lens approximation to magnetic optics.

The expansion of the general kick function $\kappa(\mathbf{q})$ into aberration polynomials is

$$\kappa(\mathbf{q}) = \sum_{n_+,n_-=0}^{\infty} \kappa_{n_+n_-} q^{n_+}_+ q^{n_-}_-$$
$$= \sum_{k=0,1/2,\ldots}^{\infty} \sum_{\lambda=-k}^{k} \kappa^\lambda_k\, {}^k\mathscr{L}^{k,\lambda}_{-k}(\mathbf{q})$$
$$= \sum_{k=0,1/2,\ldots}^{\infty} \sum_{\lambda=0\text{ or }1/2}^{k} q^{2k}(C^\lambda_k \cos 2\lambda\phi + S^\lambda_k \sin 2\lambda\phi), \tag{6.2a}$$

$$\kappa_{n_+n_-} = \frac{1}{n_+!n_-!}\frac{\partial^{n_++n_-}}{\partial q^{n_+}_+\,\partial q^{n_-}_-}\kappa(\mathbf{q})\bigg|_{\mathbf{q}=0} = 2^k \kappa^{\lambda=(n_+-n_-)/2}_{k=(n_++n_-)/2}. \tag{6.2b}$$

In the last expressions we have written $q_x = q\cos\phi$ and $q_y = q\sin\phi$, so $^k\mathscr{L}^{k,\lambda}_{-k}(\mathbf{q}) = q^{2k}e^{2i\lambda\phi}$ and the coefficients are $C^\lambda_k = 2\,\mathrm{Re}\,\kappa^\lambda_k$ and $S^\lambda_k = -2i\,\mathrm{Im}\,\kappa^\lambda_k$. All other coefficients of $^k\mathscr{L}^{j,\lambda}_m$, $m\neq -j$, are zero. The $\mathscr{L}$'s present are only those of the leftmost vertical su(2) multiplet in Figs. 2–5.

The $k=0$ term is a constant and of no import (the Lie

operator of a constant is zero). The $k=\frac{1}{2}$ terms correspond to a linear (free-fall) potential kick, a thin prism (a Fresnel lens with straight groves), or a "thin" magnetic dipole across the beam axis. Being of first order, they are excluded from our treatment here.

The $k=1$ terms are the generators of linear transformations of phase space; they are produced by harmonic oscillator potential kicks, thin (Gaussian) lenses, and thin magnetic dipoles along the beam axis modeled by the ${}^{1}\mathscr{L}^{1,0}_{-1}=q^2$ term. The other two ${}^{1}\mathscr{L}^{1,\lambda}_{-1}$'s yield the independent quadrupole kicks $q_x q_y$ and $q_x^2-q_y^2$ that are in sp(4,R), linear but not axis symmetric. Since they produce linear transformations, they do not count among the aberrations.

Kick functions of angular dependence $\sim\sin(M\phi)$ may be called $2M$ poles; this may be used to model magnetic $2M$-pole arrangements of $2M$ alternating magnetic poles in a plane normal to the optical axis, and thin. Since $2M$ poles are invariant under rotations around the optical axis by $2\pi/M$, it follows that a pure $2M$-pole kick function may expand in (6.2) only into helicity components $\lambda=0,\ \pm M/2,\ \pm M,\ldots$, and these may appear only for $k\geqslant|\lambda|=M/2$, i.e., for aberration order $A\geqslant M-1$. Thus sextupoles ($M=3$, $k=\frac{3}{2}$) require at least second aberration order, octuples ($M=4$, $k=2$) require at least third order, etc.

Since each exponential factor terminates after the first term, the required Poisson brackets are, from (3.3),

$$\{{}^{k}\mathscr{L}^{k,\lambda}_{-k},p_{\pm}\}=\frac{\partial}{\partial q_{\mp}}\,{}^{k}\mathscr{L}^{k,\lambda}_{-k}(\mathbf{q})$$
$$=\sqrt{2}(k\mp\lambda)\,{}^{k-1/2}\mathscr{L}^{k-1/2,\lambda\pm1/2}_{-(k-1/2)}. \tag{6.3a}$$

The series may be summed to

$$\exp\kappa(\mathbf{q})\hat{}\,p_{\pm}=p_{\pm}+\frac{\partial\kappa(\mathbf{q})}{\partial q_{\mp}}, \tag{6.3b}$$

as is evident from (3.3).

A *quasiflat* refracting surface is an interface $z=\zeta(\mathbf{q})$ between two different optical media with refractive indices n, n', that coincides with the reference plane up to second derivatives at the chosen optical center. This means $\zeta(\mathbf{0})=0$, $\partial\zeta/\partial\mathbf{q}|_{\mathbf{q}=\mathbf{0}}=\mathbf{0}$, and $\partial^2\zeta/\partial q_\sigma\,\partial q_\xi|_{\mathbf{q}=\mathbf{0}}=0$. We may expand the quasiflat surface $\zeta(\mathbf{q})$ in ${}^{k}\mathscr{L}^{k,\lambda}_{-k}(\mathbf{q})$'s as in Eqs. (6.1) with coefficients where $\zeta^0_0=\zeta^\lambda_1=0$. We exclude here the axis-symmetric Gaussian thin-lens coefficient ζ^0_1 since it produces linear transformations ($\mathbf{q}\mapsto\mathbf{q}$, $\mathbf{p}\mapsto\mathbf{p}+2\zeta^0_1\mathbf{q}$) that would take us beyond the purpose of simple illustration. We exclude also the Gaussian thin saddle lenses $\zeta_1{}^{\pm1}$ since they lie in sp(4,R) outside sp(2,R).

Unlike multipole kicks, quasiflat surfaces are not *quite* flat. See Fig. 6. A ray crossing the reference plane at $\mathbf{q}$ in medium n, strikes the interface ζ at $\bar{\mathbf{q}}$ after free flight by a distance $z=\zeta(\bar{\mathbf{q}})$. This is described *at* the reference plane by virtual free flight *back*, in medium n', by $-z$. The intersection with the plane is $\mathbf{q}'$. The effect of an arbitrary refracting surface on optical phase space was introduced in Ref. 24 and described in the articles in this series[1,2] and in Ref. 9, so we need not repeat the derivation. It is shown that the refracting-surface transformation is a canonical transformation that *factorizes* in the manner described in Fig. 6 into two *root*

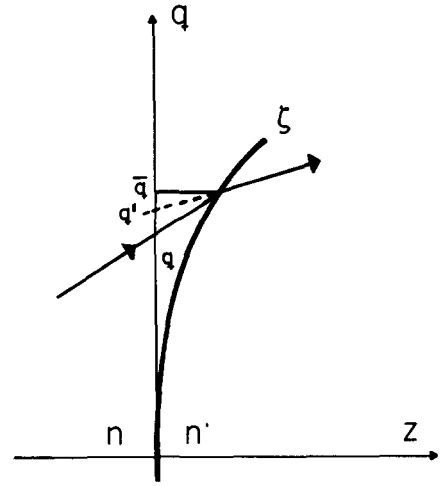


FIG. 6. Refraction at the interface between two media interpreted as a transformation at the reference plane.

transformations

$$\mathbb{S}_{n,n';\zeta}=\mathbb{R}_{n,\zeta}\mathbb{R}^{-1}_{n',\zeta}, \tag{6.4}$$

each of which is canonical and, written in the helicity basis, is

$$\mathbb{R}_{n,s}\colon\ p_{\pm}\mapsto\bar{p}_{\pm}=p_{\pm}+\sqrt{n^2-p^2}\,\frac{\partial\zeta(\bar{\mathbf{q}})}{\partial\bar{q}_{\mp}}, \tag{6.5a}$$

$$\mathbb{R}_{n,s}\colon\ q_{\pm}\mapsto\bar{q}_{\pm}=q_{\pm}+\zeta(\bar{\mathbf{q}})(p_{\pm}/\sqrt{n^2-p^2}). \tag{6.5b}$$

Observations that have been made before in this regard are that this set of equations solve implicitly for $\bar{\mathbf{q}}$ out of (6.5b). This process is amenable to expansion by aberration order; using symbolic computation programs, we have found explicit expressions to aberration order 9 for arbitrary axis-symmetric surfaces.[22,25] Here we proceed by hand through fourth order for asymmetric quasiflat surface involving $k=\frac{3}{2}$, 2, and $\frac{5}{2}$. We abbreviate the surface shape in (6.2) as

$$\zeta(\mathbf{q})=\zeta_3(\mathbf{q})+\zeta_4(\mathbf{q})+\zeta_5(\mathbf{q}),$$
$$\zeta_{2k}(\mathbf{q})=q^{2k}\sum_{\lambda=-k}^{k}\zeta^\lambda_k e^{2i\lambda\phi}. \tag{6.6}$$

Keeping terms in phase space to the aberration order, and this plus one in (6.6), we expand the inverse root function in (6.5b) and find

$$\bar{q}_{\pm}=q_{\pm}+(\zeta_3+\zeta_4+\zeta_5)(\bar{\mathbf{q}})$$
$$\times[(1/n)p_{\pm}(p^2/2n^3)p_{\pm}+\cdots]$$
$$=q_{\pm}+(1/n)\zeta_3(\bar{\mathbf{q}})p_{\pm}+o_5$$
$$=q_{\pm}+(1/n)\zeta_3(\mathbf{q})p_{\pm}+o_5. \tag{6.7}$$

In the last step we have replaced the left-hand side, $\bar{q}$, into $\zeta_3(\bar{\mathbf{q}})$ obtaining $\zeta_3(\mathbf{q})$ plus terms beyond aberration order. The fact that $\bar{\mathbf{q}}\neq\mathbf{q}$ shows that the quasiflat surface is not simply a flat kick; the summand $\zeta_3(\mathbf{q})\,p_{\pm}$ is of degree 4 and the distinction with kicks lies thus beyond aberration order 3.

We now replace the result (6.7) into (6.5a) expanding

the square root in the same manner,

$$\bar{p}_\pm = p_\pm + \left(n - \frac{p^2}{2n} - \cdots\right)\frac{\partial}{\partial \bar{q}_\mp}(\zeta_3+\zeta_4+\zeta_5)(\bar{\mathbf{q}})$$

$$= p_\pm + n\frac{\partial}{\partial q_\mp}(\zeta_3+\zeta_4+\zeta_5)(\mathbf{q})$$

$$-\frac{p_+p_-}{n}\frac{\partial}{\partial q_\mp}\zeta_3(\mathbf{q}) + o_5. \tag{6.8}$$

In the last step we have noted that $\partial\zeta_3/\partial\bar{\mathbf{q}}$ is of degree 2 in $\mathbf{q}$ and, since the lowest cross term in (6.7) would be already of degree 5, we wrote $\partial\zeta_3/\partial q + o_5$. The $q_\mp$ derivative of the aberration polynomials are found from (6.3).

Now we want to write the aberration polynomials $r_{2k}(\mathbf{p},\mathbf{q})$ that generate the root transformation (6.7) and (6.8) as a pure-aberration Lie transformation (2.5) with $r_2 = 0$. We recall the results for pure kicks and proceed through aberration orders 2 and 3,

$$r_3(\mathbf{q}) = n\zeta_3(\mathbf{q}) = n\sum_{\lambda=-3/2}^{3/2}\zeta_{3/2}^{\lambda\ 3/2}\mathscr{L}_{-3/2}^{3/2,\lambda}(\mathbf{q}), \tag{6.9a}$$

$$r_4(\mathbf{q}) = n\zeta_4(\mathbf{q}) = n\sum_{\lambda=-2}^{2}\zeta_2^{\lambda\ 2}\mathscr{L}_{-2}^{2,\lambda}(\mathbf{q}). \tag{6.9b}$$

The corresponding exponential series of the operators $\hat{r}_3$ and $\hat{r}_4$ acting on the position observables $\mathbf{q}$ do nothing, while on $\mathbf{p}$ they stop after the first term and account for the summands $n\,\partial\zeta_3/\partial q_\mp$ and $n\,\partial\zeta_4/\partial q_\mp$ in (6.8).

The polynomial r_5 responsible for fourth aberration order in the exponential series yields the fourth-order term of $\bar{\mathbf{q}}$ and $\bar{\mathbf{p}}$ through its first Poisson bracket with $\mathbf{q}$ and $\mathbf{p}$,

$$-\frac{\partial r_5}{\partial p_\mp} = \{r_5, q_\pm\} = \frac{1}{n}\zeta_3(\mathbf{q})p_\pm, \tag{6.10a}$$

$$\frac{\partial r_5}{\partial q_\mp} = \{r_5, p_\pm\} = n\frac{\partial}{\partial q_\mp}\zeta_5(\mathbf{q}) - \frac{p_+p_-}{n}\frac{\partial}{\partial q_\mp}\zeta_3(\mathbf{q}). \tag{6.10b}$$

The system is integrable because of the symplectic condition.[26] Its solution is

$$r_5(\mathbf{p},\mathbf{q}) = n\zeta_5(\mathbf{q}) - (1/2n)p^2\zeta_3(\mathbf{q}). \tag{6.11}$$

Thus the root transformation (6.5) to fourth order is

$$\mathbb{R}_{n,s} = \cdots\exp[n\zeta_5(\mathbf{q}) - (1/2n)p^2\zeta_3(\mathbf{q})]\hat{}$$
$$\times\exp n\zeta_4(\mathbf{q})\hat{}\ \exp n\zeta_3(\mathbf{q})\hat{}. \tag{6.12}$$

The quasiflat surface transformation (6.4) may be obtained directly from the two root ones, since[10] $e^{\hat{A}}e^{\hat{B}} = e^{\hat{B}}e^{\hat{A}}e^{[\hat{A},\hat{B}]+\cdots}$, where $A = r_3 + r_4 + r_5$ with n, and B the same with $-n'$; $\{A,B\}\sim\{r_3,r_5\}$ is of degree $3+5-2 = 6 > 5$ and lies beyond the aberration order. Hence the Lie transformation is

$$\mathbb{S}_{n,n';\zeta}$$
$$= \cdots\exp\left[(n-n')\zeta_5(\mathbf{q}) - \left(\frac{1}{2n} - \frac{1}{2n'}\right)p^2\zeta_3(\mathbf{q})\right]\hat{}$$
$$\times\exp(n-n')\zeta_4(\mathbf{q})\hat{}\ \exp(n-n')\zeta_3(\mathbf{q})\hat{}\ . \tag{6.13}$$

The phase-space transformations $\mathbf{w}' = \mathbb{S}_{n,n';\zeta}\mathbf{w} = \mathbb{R}_{n',\zeta}^{-1}\bar{\mathbf{w}}$ are also given by (6.7) and (6.8) through $n\mapsto n-n'$ and $1/n\mapsto 1/n - 1/n'$.

In the above expressions we may distinguish the kick terms present in (6.8) by the factor $\sim(n-n')$. They are functions only of $\mathbf{q}$ and belong to the "left edge" of each sp(4,R) mutliplet. The nonflatness term has a factor $\sim(1/n - 1/n')$ in both $q'_\pm$ and $p'_\pm$. The latter are the more interesting ones: for $q'_\pm$, following (6.7) it is $\sim p_\pm\zeta_3(\mathbf{q})$; for $p'_\pm$, following (6.8) it is $\sim p^2\,\partial\zeta_3(\mathbf{q})/\partial q_\mp$. A ζ_3 term at the (k,j,m,λ) position $(\frac{3}{2},\frac{3}{2};-\frac{3}{2},[\lambda=\frac{3}{2},-\frac{1}{2},\frac{1}{2}$, or $\frac{3}{2}])$ is shifted in the former to $(2,[j=2,1];-1,\lambda\pm\frac{1}{2})$ and in the latter to $(2,[j=2,1,0];0,|\lambda\pm\frac{1}{2}|\leqslant 1)$, with a mixture of two or three values of the symplectic spin j. We shall not go into details beyond this point; the explicit results in terms of the sp(4,R) aberration polynomials may be found using (6.3) and the two tables of Sec. V.

VII. SOME FURTHER ISSUES AND CONCLUDING REMARKS

It seems to us that Lie methods in aberrating systems require a balance between computational ease and mathematical thoroughness. The examples in the last section would also be quite tractable using the "unclassifiable"[27] monomials $C_{m_+m_-n_+n_-}p_+^{m_+}p_-^{m_-}q_+^{n_+}q_-^{n_-}$ for both the aberration polynomials r_k $(\mathbf{p},\mathbf{q})$ and the nonlinear map of phase space $\mathbf{p}\mapsto\bar{\mathbf{p}}(\mathbf{p},\mathbf{q}),\mathbf{q}\mapsto\bar{\mathbf{q}}(\mathbf{p},\mathbf{q})$. Indeed, if we ask only that the refracting surface be *tangent* to the reference plane at the optical center, we derived in Ref. 12 a set of *selection rules*[9] obeyed by the monomial aberration coefficients of axis-symmetric systems. Concretely, spherical aberration ($[p^2]^k$), circular coma ($[p^2]^{k-1}\,\mathbf{p}\cdot\mathbf{q}$), and all aberrations generated by $[p^3]^{k-\kappa}(p\cdot q)^\kappa$, $\kappa = 0,1,\ldots,k$, are *zero*.

We may apply exactly the same reasoning to the asymmetric surface $\zeta(\mathbf{q}) = \zeta_{\sigma\tau}q_\sigma q_\tau + \zeta_{\sigma\tau\phi}q_\sigma q_\tau q_\phi + \cdots(\sigma,\tau,\phi = +,-)$ and obtain selection rules for the coefficients r of $\mathbb{R}_{n,\zeta}$,

$$r_{2k}(\mathbf{q}) = \sum_{m_+ + m_- + n_+ + n_- = 2k} r_{m_+m_-n_+n_-}p_+^{m_+}p_-^{m_-}q_+^{n_+}q_-^{n_-}. \tag{7.1}$$

We shall not repeat the details since they follow closely the arguments presented in Ref. 9. We obtain

$$r_{m_+m_-n_+n_-} = 0 \quad \text{for } n_+ + n_- \leqslant 1, \tag{7.2a}$$

i.e., the rightmost two columns of every symplectic aberration multiplet are absent: spherical aberration and circular coma, with all their helicity versions. For the third column from the right, we find

$$r_{m_+m_-n_+n_-} = 0 \quad \text{for } n_+ + n_- = 2 \text{ and } m_+ \neq m_-, \tag{7.2b}$$

i.e., only the monomial aberrations $p^2q_+^{n_+}q_-^{n_-}$, $n_+ + n_- = 2$ (so $\lambda = 0, \pm 1$) are nonzero.

The point we want to emphasize here is that the selection rules are imposed by nature on the coefficients of the *monomials* $p_+^{m_+}p_-^{m_-}q_+^{n_+}q_-^{n_-}$, *not* on the coefficients of the symplectic polynomials ${}^k\mathscr{L}_m^{j,\lambda}$. By itself, this result would argue against the usefulness of our classification. This, we saw in Sec. II is based on pure magnifiers; it represents the "best choice of balance" between refracting surface transformations and free propagation (2.8), where only $f_{kk00}(z)$ is different from zero. Principally, it is the symplectic spin j

that may be questioned for convenience. Let us therefore present the basics of another development that argues for the economy of Lie theory in aberration optics.[13]

Optical fibers with bends or other defects may be expected to suffer from asymmetric aberrations. Moreover, the Seidel aberrations of a fiber[2] are bound to a paraxial harmonic oscillator motion and describe epicycles in the complex plane. A simpler description of their behavior becomes evident already for axis-symmetric systems[13] when we refer their weight (m) classification to the oscillator axis

$$H^{\text{osc}} = \tfrac{1}{2}(p^2 + q^2) = \sqrt{2}(\xi_+ + \xi_-) = -i\xi_2. \qquad (7.3)$$

The transformation from the Seidel axis for magnifiers to the coherent-state axis for fibers is through a (complex) rotation of $\pi/2$ around the $\xi_1 = -\tfrac{1}{2}(p^2 - q^2)$ axis. This is Bargmann's transformation[28]

$$\exp\left[-\frac{1}{8}i\pi(p^2 - q^2)\hat{}\right]\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & i\\ i & 1\end{pmatrix}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix}. \qquad (7.4)$$

Under this transformation, the components of each sp(2,R) spin multiplet j mix only among themselves. Instead of a Seidel weight m, we shall have a "coherent state" weight m'; aberration order, sp(2,R) spin, and helicity are the same. The coherent state basis aberration coefficients now follow multiply periodic circular motion (with z) in the complex plane that does not surround the origin. The adaptability of the Lie classification scheme to the paraxial system under consideration will extend to the general asymmetric situation as well.

In this article we have presented as the main result the classification of asymmetric aberrations; we have also brushed several issues we left aside as lateral, and questions remain to be answered. There is a need for more realistic examples of optical systems, analyzed and computed in greater depth. This we propose to do in future work.

[1]M. Navarro Saad and K. B. Wolf, "The group-theoretical treatment of aberrating systems. I. Aligned lens systems in third aberration order," J. Math. Phys. **27**, 1449 (1986).

[2]K. B. Wolf, "The group-theoretical treatment of aberrating systems. II. Axis-symmetric inhomogeneous systems and fiber optics in third aberration order," J. Math. Phys. **27**, 1458 (1986).

[3]*Lie Methods in Optics*, Proceedings of the CIFMO CIO Workshop, edited by J. Sánchez Mondragón and K. B. Wolf, held at León, México, 7–10 January 1985, *Lecture Notes in Physics*, Vol. 250 (Springer, Berlin, 1986).

[4]H. Buchdahl, *Optical Aberration Coefficients* (Dover, New York, 1968); *An Introduction to Hamiltonian Optics* (Cambridge U. P., Cambridge, 1970).

[5]L. Seidel, "Zur Dioptrik," Astron. Nachr. **871**, 105 (1853).

[6]A. J. Dragt, *Lectures in Nonlinear Orbit Analysis*, AIP Conference Proceedings, Vol. 87 (AIP, New York, 1982).

[7]A. J. Dragt, "Lie-algebraic theory of geometrical optics and optical aberrations," J. Opt. Soc. Am. **72**, 372 (1982).

[8]T. Sekiguchi and K. B. Wolf, "The Hamiltonian formulation of optics," to be published in Am. J. Phys.

[9]A. J. Dragt, E. Forest, and K. B. Wolf, "Foundations of Lie algebraic theory of geometrical optics," in Ref. 3.

[10]S. Steinberg, "Lie series, Lie transformations, and their applications," in Ref. 3.

[11]A. J. Dragt and J. Finn, "Lie series and invariant functions for analytic symplectic maps," J. Math. Phys. **17**, 2215 (1976).

[12]K. B. Wolf, "Symmetry in Lie optics," Ann. Phys. **172**, 1 (1986).

[13]K. B. Wolf, "Symmetry adapted classification of aberrations," Comunicaciones Técnicas IIMAS preprint No. 440, 1986.

[14]A. J. Dragt and E. Forest, "Computation of nonlinear behavior of Hamiltonian systems using Lie algebraic methods," J. Math. Phys. **24**, 2734 (1983); E. Forest, "Lie algebraic maps and invariants produced by tracking codes," SSC report 78, July 1986.

[15]A. Katz, *Classical Mechanics, Quantum Mechanics, Field Theory* (Academic, New York, 1965).

[16]"A Euclidean algebra of Hamiltonian observables in Lie optics," Kinam **6**, 141 (1985).

[17]H. Goldstein, *Classical Mechanics* (Addison–Wesley, Reading, MA, 1959).

[18]M. Havlíček and W. Lassner, "Canonical realizations of the Lie algebra sp(2n,R)," Int. J. Theor. Phys. **15**, 867 (1976); "On the 'near to minimal' canonical realizations of the Lie algebra C_n," Int. J. Theor. Phys. **15**, 877 (1976).

[19]R. Gilmore, *Lie groups, Lie algebras, and Some of Their Applications* (Wiley, New York, 1974).

[20]L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics*, Encyclopedia of Mathematics, Vol. 8, edited by G.-C. Rota (Addison–Wesley, Reading, MA, 1981), Sec. 3.10.

[21]P. Chand, C. L. Mehta, N. Mukunda, and E. C. G. Sudarshan, "Realizations of Lie algebras by analytic functions of generators of a given Lie algebra," J. Math. Phys. **8**, 2048 (1967).

[22]M. Navarro Saad and K. B. Wolf, "Applications of a factorisation theorem for ninth-order aberration optics," J. Symb. Comp. **1**, 235 (1985).

[23]In Refs. 1, 2, and 12 we built the aberration polynomials using factors of $\eta/\sqrt{2}$, instead of η as here; the polynomials are denoted there as ${}^k\chi^j_m$ using the Greek letter χ (chi), while here they are denoted $\mathscr{X}$ (script X). See Ref. 9, p. 143.

[24]M. Navarro Saad and K. B. Wolf, "Factorization of the phase-space transformation produced by an arbitrary refracting surface," J. Opt. Soc. Am. A **3**, 340 (1986).

[25]M. Navarro Saad, Ph.D. thesis, Universidad Nacional Autónoma de México, 1985.

[26]E. Forest, thesis, University of Maryland, 1984.

[27]We call the monomials "unclassifiable" since the four number operators $p_+\,\partial/\partial p_+$, etc. are *not* Lie operators $\hat{f}$ obtained from a phase-space function f in the manner (3.3).

[28]V. Bargmann, "On a Hilbert space of analytic functions and an associated integral transform. Part 1," Comm. Pure Appl. Math. **14**, 187 (1961); see also in K. B. Wolf, *Integral Transforms in Science and Engineering* (Plenum, New York, 1979), Sec. 9.1.2.

Stability and bifurcation of a rotating planar liquid drop

D. Lewis and J. Marsden
Department of Mathematics, University of California, Berkeley, California 94720

T. Ratiu
Department of Mathematics, University of Arizona, Tucson, Arizona 85721

(Received 23 January 1987; accepted for publication 17 June 1987)

The stability and symmetry breaking bifurcation of a planar liquid drop is studied using the energy-Casimir method and singularity theory. It is shown that a rigidly rotating circular drop of radius r with surface tension coefficient τ and angular velocity $\Omega/2$ is stable if $(\Omega/2)^2 < 3\tau/r^3$. A new branch of stable rigidly rotating relative equilibria invariant under rotation through π and reflection across two axes bifurcates from the branch of circular solutions when $(\Omega/2)^2 = 3\tau/r^3$.

I. INTRODUCTION

Bifurcation of systems with symmetry has been a subject of much interest in recent years. Symmetric systems are common in nature and even more common in the literature, as multidimensional bifurcation problems possessing symmetry are typically more tractable than asymmetric problems of comparable dimensions. The requirement that the bifurcation equation be equivariant under the action of a given group G, i.e., that $f(g\cdot\mathbf{x},\lambda) = g\cdot f(\mathbf{x},\lambda)$ for all $g\in G$, can force the bifurcation equation to take on a relatively simple form. For example, if one considers a function f on $\mathbb{R}$ which is equivariant with respect to the $\mathbf{Z}_2$ action $x \to -x$ it is clear that f can be written as $\tilde{f}(x^2)x$ for some function $\tilde{f}$. [See Golubitsky and Schaeffer[1] for a thorough presentation of the singularity theory approach to bifurcations with (and without) symmetry.]

The class of bifurcation equations with which we are particularly concerned here arise in Hamiltonian systems with symmetry. Using the energy-Casimir method (cf. Holm *et al.*[2]), one can typically find a combination C of conserved quantities such that a given (relative) equilibrium of a Hamiltonian system is a critical point of $H + C$, where H is the usual Hamiltonian of the system. The bifurcation parameter may appear in either the Hamiltonian itself or in the added conserved quantities; if we denote the parameter-dependent modified Hamiltonian by $(H + C)_\lambda$, then the appropriate bifurcation equation is $D_{\mathbf{x}}(H + C)_\lambda(\mathbf{x}) = \mathbf{0}$.

Invariance of the Hamiltonian under a given group action usually induces constraints on the form of its differential. In the analysis of a symmetric bifurcation problem it is important to exploit these constraints as fully as possible; behavior exceptional in an asymmetric context may be typical or even necessary if all existing symmetry is taken into account. Several important generic properties of bifurcations of Hamiltonian systems are presented in Golubitsky and Stewart.[3] The present paper is largely the result of discussions with Golubitsky and Stewart; the lemma presented here is a variation on results due to Cicogna[4] and Golubitsky *et al.*[5]

There are a number of well known, but as yet incompletely understood, examples of bifurcation with symmetry breaking in hydrodynamics, including Taylor–Couette flow and the vortex breakdown. The energy-Casimir method has been applied to a wide variety of hydrodynamic problems with a great deal of success in recent years (see Holm *et al.*[2] for a generous selection of applications of the energy-Casimir method). In earlier works we have determined the Hamiltonian structure for free boundary fluid problems (see Lewis *et al.*[6]) and formal stability for the two-dimensional circular liquid drop (see Lewis *et al.*[7]); in Lewis,[8] conditional nonlinear stability under the same hypotheses is established. The method is readily applicable to analytic solutions (e.g., the Kelvin–Stuart cat's eye, cf. Holm *et al.*[9]) and should be implementable for approximate numerical solutions.

Our basic approach is to determine the stability of a relatively simple equilibrium flow by applying the energy-Casimir method and then, at the point at which this flow loses formal stability, apply the techniques of symmetric bifurcation theory to gain information about the new, typically more complicated, solution branch. The techniques and general results discussed here are not, however, restricted to problems in fluid dynamics; another class of examples currently being studied is the stability of coupled rigid bodies and spacecraft with flexible attachments; see Krishnaprasad and Marsden.[10]

The paper consists of three sections. Section II gives a brief (and incomplete) summary of existing results in this area. Section III contains a lemma outlining conditions under which bifurcation of the critical manifold of an $SO(2)$ invariant function on $\mathbb{R}^2$ can be shown to occur. Section IV discusses, as an application of the lemma, the bifurcation of a two-dimensional rotating liquid drop with surface tension from a rigidly rotating circular configuration. In future publications we hope to present some numerical studies of the drop configurations and possibly search for boundary bifurcations from the "flip" symmetric two-lobed branch.

II. BACKGROUND

Rotating liquid drops have been the object of intense study, both in the nineteenth century and in the last twenty years. While the original research was necessarily restricted

 0022-2488/87/102508-08$02.50 © 1987 American Institute of Physics

to the study of approximate theoretical and experimental models, recent work has benefitted greatly from the availability of computer simulation and elaborate and accurate experimental configurations. Swiatecki[11] provides a thorough review of research in this area up to the early seventies.

The principal analytic approach to the study of the equilibrium configurations and their stability has been to analyze linearized models and low-order approximations of the actual drop shapes. Analytic linear stability results for axisymmetric drops held together by surface tension have been found by Chandrasekhar[12] using the method of virials. Second-order expansions for the evolution of a perturbed spherical drop have been developed by Tsamopoulos and Brown.[13]

Several thorough numerical studies of rotating liquid drops have been made. Brown and Scriven[14] use a finite element code to trace the bifurcations of an initially spherical rotating drop held together by surface tension; they analyze the linear stability of the solution branches and show general agreement with Chandrasekhar's analytic results. Benner[15] has performed numerical studies of cylindrical (i.e., planar) drops under the effect of surface tension and traced the evolution of small potential flow perturbations of the stationary circular solution. The results of his simulations indicate that these perturbations remain bounded for at least a short period of time. Both the calculations of Brown and Scriven and Benner assume that the drop possesses reflectional symmetry across some axis; equilibria lacking this symmetry could conceivably appear through subsequent secondary bifurcations.

Experimental research regarding rotating liquid drops with surface tension dates back to Plateau's study of fat globules suspended in a liquid of nearly equal density. The most dramatic recent research is that of Wang *et al.*[16]; these experiments, which involved free floating, acoustically accelerated droplets, were conducted in near zero gravity in Spacelab. The observed bifurcation of a family of two-lobed drops from a family of oblate, axisymmetric drops agrees qualitatively with both the analytic and numerical predictions, although there are some unresolved quantitative discrepancies. (In particular, the bifurcation from the axisymmetric to the two-lobed branch appears to have occurred somewhat earlier than predicted.)

III. BIFURCATION LEMMA

The initial step in the analysis of a given bifurcation is to establish that a bifurcation has, in fact, taken place. It is typically the case that if a known solution loses stability as a given parameter is varied, then a "transfer of stability" occurs and another stable solution exists for nearby parameter values. This supposition must, however, be checked in each case. In complicated examples, e.g., those obtained from large or even infinite-dimensional systems by Liapunov–Schmidt reduction, the task of determining points of bifurcation need not be trivial.

At a point of bifurcation one typically expects to see a new one-dimensional solution branch emerge; a typical nondegeneracy condition for bifurcation results is that only one eigenvalue of the system pass through zero at the point of bifurcation. In problems without symmetry, or with discrete symmetry, this is an entirely reasonable assumption, but if the symmetry group is continuous, it may be impossible to satisfy. If a map f is equivariant under the linear action of a group G, then the following situation occurs. If $\mathbf{x}$ is a zero of f then, for any $g\in G$, $g\cdot\mathbf{x}$ must be a zero as well, since

$$f(g\cdot\mathbf{x}) = g\cdot f(\mathbf{x}) = \mathbf{0}$$

if $f(\mathbf{x}) = \mathbf{0}$. Thus the solution branches are made up of orbits of the group action. If $\mathcal{G}$ acts freely on a given solution branch, then the dimension of that branch cannot be less than the dimension of G. Even if the action is not free, it may still force the solution branch to be multidimensional, implying that at the point of bifurcation multiple eigenvalues pass through zero simultaneously. In this case many standard bifurcation theorems may not be applicable.

If analyzed strictly with regard to dimension, the study of bifurcation problems with continuous symmetry groups may appear to be extremely difficult. In fact, the multidimensional solution branches are usually redundant; all essential information about the bifurcation may be obtained by studying a representative point in the orbit swept out by the group action. In some cases it is feasible to explicitly reduce the original manifold by the group action, but there are circumstances under which this reduction can be somewhat complicated. For example, if one considers a linear group action on a vector space, the action at the origin is not free and the reduced space may fail to be a manifold at that point. Thus, if one is considering a bifurcation from the "trivial" solution $(\mathbf{0},\lambda)$, analytic difficulties arise exactly at the point of interest. In such cases it seems preferable to leave the state space unaltered and instead generalize the usual criteria for bifurcation to account for the redundancy induced by the group action. The central result of this section is a simple generalization to the case of the group $SO(2)$ acting on $\mathbb{R}^2$. (In this case both eigenvalues pass through zero simultaneously at a point of bifurcation never leaving the imaginary axis.)

The following lemma is a modification of results of Cicogna[4] and Golubitsky *et al.*[5] The idea behind the lemma is to split the bifurcation map into a scalar function that depends on the bifurcation parameter and a multidimensional map that is independent of the parameter and equal to zero at the bifurcation point; one then applies the implicit function theorem to the scalar equation to establish the existence of a new solution branch. The second result in this section is an application of the lemma to the differential of an $SO(2)$ invariant function on $\mathbb{R}^2$, where the restrictions imposed on the function by $SO(2)$ invariance guarantee that the decomposition of the differential into scalar and vector-valued components is possible.

Lemma 1: Let V be a vector bundle over a manifold M and $\lambda\in\mathbb{R}$. Let F be a λ-dependent section of V. Assume $F(\mathbf{x},\lambda) = g(\mathbf{x},\lambda)\cdot\mathbf{h}(\mathbf{x})$ for some (smooth) maps $g: M\times\mathbb{R}\to\mathbb{R}$ and $\mathbf{h}: M\to V$. Let $S_0 = \{\mathbf{x}: \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$. If for some point $(\mathbf{x}_0,\lambda_0)$ with $\mathbf{x}_0\in S_0$ we have

(i) $D_{\mathbf{x}}F(\mathbf{x}_0,\lambda_0) = \mathbf{0}$;

(ii) $D_{\mathbf{x}}\mathbf{h}(\mathbf{x}_0)\neq\mathbf{0}$;

(iii) $D_{\mathbf{x}\lambda}F(\mathbf{x}_0,\lambda_0)\neq\mathbf{0}$,

then a branch (or possibly family) of solutions (i.e., points mapped into 0) bifurcates from the trivial solution manifold S_0 at $(\mathbf{x}_0,\lambda_0)$.

Proof:

$$\mathbf{0} = D_{\mathbf{x}}F(\mathbf{x}_0,\lambda_0)$$

$$= D_{\mathbf{x}}\, g(\mathbf{x}_0,\lambda_0)\mathbf{h}(\mathbf{x}_0) + g(\mathbf{x}_0,\lambda_0)D_{\mathbf{x}}\mathbf{h}(\mathbf{x}_0)$$

implies $g(\mathbf{x}_0,\lambda_0) = 0$, since $\mathbf{x}\in S_0$ implies $\mathbf{h}(\mathbf{x}_0) = \mathbf{0}$ and, by (ii), $D_{\mathbf{x}}\mathbf{h}(\mathbf{x}_0) \neq \mathbf{0}$. Similarly,

$$\mathbf{0} \neq D_{\mathbf{x}\lambda}F(\mathbf{x}_0,\lambda_0)$$

$$= D_{\lambda}g(\mathbf{x}_0,\lambda_0)D_{\mathbf{x}}\mathbf{h}(\mathbf{x}_0)$$

implies $D_{\lambda}g(\mathbf{x}_0,\lambda_0) \neq 0$. Thus we can apply the implicit function theorem to g and find a function $\Lambda: M \to \mathbb{R}$ such that $g(\mathbf{x},\Lambda(\mathbf{x})) = 0$ for all $\mathbf{x}$ in a neighborhood of $\mathbf{x}_0$. It follows that there must be a set of solutions of $F = \mathbf{0}$ passing through S_0 at $(\mathbf{x}_0,\lambda_0)$. ■

We now specialize the above result to the study of critical points of an SO(2) invariant function on $\mathbb{R}^2$.

Corollary 1: If

(i) $f: \mathbb{R}^2\times\mathbb{R}\to\mathbb{R}$ is (smooth and) invariant under the standard SO(2) action on $\mathbb{R}^2$;

(ii) $D_{\mathbf{xx}}f(0,0,\lambda_0) = \mathbf{0}$ for some λ_0;

(iii) $D_{\mathbf{xx}\lambda}f(0,0,\lambda_0) \neq \mathbf{0}$,

then a branch of critical points of f emanates from the trivial critical point branch $(0,0,\lambda)$ at λ_0.

Proof: The invariance of $f: \mathbb{R}^2\times\mathbb{R}\to\mathbb{R}$ under the SO(2) action implies the existence of a function $\tilde{f}: \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ such that $f(x,y,\lambda) = \tilde{f}(x^2+y^2,\lambda)$. (For smoothness of $\tilde{f}$, see Golubitsky and Shaeffer.[1]) Identifying $T^*\mathbb{R}^2$ with $\mathbb{R}^2\times\mathbb{R}^2$, it follows that

$$D_{\mathbf{x}}f(x,y,\lambda) = \frac{\partial\tilde{f}}{\partial r}(x^2+y^2,\lambda)(2x,2y).$$

Thus, letting $g(x,y,\lambda) = (\partial\tilde{f}/\partial r)(x^2+y^2,\lambda)$ and $\mathbf{h}(x,y) = (2x,2y)$, we have

$$F(x,y,\lambda) = D_{\mathbf{x}}f(x,y,\lambda)$$

$$= g(x,y,\lambda)\cdot\mathbf{h}(x,y).$$

Conditions (ii) and (iii) imply that

$$D_{\mathbf{x}}F(0,0,\lambda_0) = D_{\mathbf{xx}}f(0,0,\lambda_0)$$

$$= \mathbf{0}$$

and

$$D_{\mathbf{x}\lambda}F(0,0,\lambda_0) = D_{\mathbf{xx}\lambda}f(0,0,\lambda_0)$$

$$\neq \mathbf{0}.$$

Differentiating the linear map $\mathbf{h}$ gives

$$D_{\mathbf{x}}\mathbf{h}(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Thus the conditions of the lemma are satisfied and a branch of nonzero solutions of $F = \mathbf{0}$, i.e., critical points of f, must branch from $(0,0,\lambda_0)$. ■

Remark: The above result for SO(2) acting on $\mathbb{R}^2$ can be generalized to the case of an n dimensional Lie group G acting on an $n+1$ dimensional manifold $\mathcal{M}$. If a function $f: \mathcal{M}\times\mathbb{R}\to\mathbb{R}$ is G invariant, then Df typically lies in a one-dimensional subspace of the cotangent bundle of $\mathcal{M}$; thus, if the appropriate nondegeneracy conditions are satisfied, the lemma can be applied. More precisely, let $f: \mathcal{M}\to\mathbb{R}$ be a G invariant function, $x\in\mathcal{M}$ and $\Theta(s)$ be a curve in G tangent to a vector $\xi\in\mathcal{G}$, the Lie algebra of G, at $s = 0$. Differentiating the equality $f(\Theta(s)\cdot\mathbf{x}) = f(\mathbf{x})$, one sees that $D_{\mathbf{x}}f(\mathbf{x})\cdot\xi_{\mathcal{M}}(x) = 0$. Here $\xi_{\mathcal{M}}(x)$ denotes the infinitesimal generator of ξ, defined by $\xi_{\mathcal{M}}(x) = (d/ds)|_{s=0}\,\Theta(s)\cdot\mathbf{x}$. [For example, in the case of $\mathbb{R}^2$ with the usual SO(2) action, $1_{\mathcal{M}}(\mathbf{x}) = \hat{\mathbf{z}}\times\mathbf{x}$.] Let $\xi^1,\ldots,\xi^n$ be a basis of $\mathcal{G}$. At any point $\mathbf{x}$ in $\mathcal{M}$ at which G acts freely, $\xi^1_{\mathcal{M}}(\mathbf{x}),\ldots,\xi^n_{\mathcal{M}}(\mathbf{x})$ span an n-dimensional subspace $\Xi_{\mathbf{x}}$ of $T_{\mathbf{x}}\mathcal{M}$. Then $Df(\mathbf{x})$ must lie in the one-dimensional subspace $\Xi^{\perp}_{\mathbf{x}}$ of $T^*_{\mathbf{x}}\mathcal{M}$ consisting of one forms annihilating $\Xi_{\mathbf{x}}$. Any nondegenerate local section of $\Xi^{\perp}_{\mathbf{x}}$ will serve as $\mathbf{h}$, so that the lemma may be applied.

IV. ROTATING PLANAR LIQUID DROP

As an application of the preceding results, we consider a planar liquid drop consisting of an incompressible, inviscid fluid with a free boundary and forces of surface tension on the boundary. The dynamic variables are the free boundary Σ and the spatial velocity field $\mathbf{v}$, a divergence-free vector field on the region D_{Σ} bounded by Σ. The surface Σ is an element of the set $\mathcal{S}$ of closed curves in $\mathbb{R}^2$ diffeomorphic to the boundary of a reference region D and enclosing the same area as D. We let $\mathcal{N}$ denote the space of all such pairs $(\Sigma,\mathbf{v})$. The Hamiltonian approach to hydrodynamic problems was introduced in the fixed boundary case by Arnold[17] and developed by Marsden and Weinstein.[18] The free boundary case has also been studied by Sedenko and Iudovich.[19]

The equations of motion for an ideal fluid with a free boundary Σ with surface tension τ are

$$\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla p, \quad \frac{\partial\Sigma}{\partial t} = \langle\mathbf{v},\nu\rangle, \qquad (1)$$

$$\operatorname{div}\mathbf{v} = 0 \quad \text{and} \quad p|\Sigma = \tau\kappa,$$

where ν is the unit normal to the surface, Σ, κ is the mean curvature of Σ, and τ is the surface tension coefficient, a numerical constant.

The Poisson bracket will be defined for functions F,G: $\mathcal{N}\to\mathbb{R}$, which possess *functional derivatives* defined as follows.

(i) $\delta F/\delta\mathbf{v}$ is a divergence-free vector field on D_{Σ} such that

$$D_{\mathbf{v}}F(\Sigma,\mathbf{v})\cdot\delta\mathbf{v} = \int_{D_{\Sigma}}\left\langle\frac{\delta F}{\delta\mathbf{v}},\delta\mathbf{v}\right\rangle dA,$$

where the partial (Fréchet) derivative $D_{\mathbf{v}}F$ is computed with Σ fixed.

(ii) $\delta F/\delta\varphi$ is the function on Σ with integral zero given by

$$\frac{\delta F}{\delta\varphi} = \left\langle\frac{\delta F}{\delta\mathbf{v}},\nu\right\rangle.$$

(The symbol φ represents the potential for the gradient part of $\mathbf{v}$ in the Helmholtz, or Hodge, decomposition.)

(iii) $\delta F/\delta\Sigma$ is a function on Σ determined up to an additive constant as follows. A variation $\delta\Sigma$ of Σ is identified with a function on Σ representing the infinitesimal variation of Σ in its normal direction. It follows from the incompressibility assumption that $\delta\Sigma$ has integral zero. Let $\delta F/\delta\Sigma$ be

the function determined up to an additive constant by

$$\int_\Sigma \frac{\delta F}{\delta \Sigma}\,\delta\Sigma\,ds = D_\Sigma F(\Sigma,\mathbf{v})\cdot\delta\Sigma.$$

We now define a Poisson bracket on $\mathcal{N}$ as follows. For functions F and G mapping $\mathcal{N}$ to $\mathbb{R}$ and possessing functional derivatives as defined above, set

$$\{F,G\} = \int_{D_\Sigma} \left\langle \omega, \frac{\delta F}{\delta \mathbf{v}} \times \frac{\delta G}{\delta \mathbf{v}} \right\rangle dA + \int_\Sigma \left(\frac{\delta F}{\delta \Sigma}\frac{\delta G}{\delta \varphi} - \frac{\delta G}{\delta \Sigma}\frac{\delta F}{\delta \varphi} \right) ds, \tag{2}$$

where $\omega = \operatorname{curl} \mathbf{v}$. This Poisson bracket on $\mathcal{N}$ is derived from the canonical cotangent bracket on $T^*\mathscr{C}$, where, in the two-dimensional case, $\mathscr{C} = \mathrm{Emb}_{\mathrm{vol}}(D,\mathbb{R}^2)$ is the manifold of volume-preserving embeddings of a two-dimensional reference manifold D into $\mathbb{R}^2$, by reduction by the group $G = \mathrm{Diff}_{\mathrm{vol}}(D)$, the group of volume-preserving diffeomorphisms of D (i.e., the group of particle relabeling transformations). (See Lewis *et al.*[6] for details.)

We take our Hamiltonian to be

$$H(\Sigma,\mathbf{v}) = \int_{D_\Sigma} \frac{1}{2}|\mathbf{v}|^2\,dA + \tau\int_\Sigma ds. \tag{3}$$

The functional derivatives of H are computed to be

$$\frac{\delta H}{\delta \mathbf{v}} = \mathbf{v}, \quad \frac{\delta H}{\delta \varphi} = \left\langle \frac{\delta H}{\delta \mathbf{v}}, \nu \right\rangle = \langle \mathbf{v},\nu\rangle,$$

$$\frac{\delta H}{\delta \Sigma} = \frac{1}{2}|\mathbf{v}|^2 + \tau\kappa,$$

where $\delta H/\delta\Sigma$ is taken modulo constants. For this H and the Poisson bracket (2), the equations of motion (1) for the free boundary fluid with surface tension are equivalent to the relation $\partial F/\partial t = \{F,H\}$ for all functions F on $\mathcal{N}$ possessing functional derivatives.

We consider the stability of the planar incompressible fluid flow such that the boundary Σ_e is a circle of radius r and the fluid is rigidly rotating with angular velocity Ω. We shall apply the energy-Casimir method as follows. For the circular equilibrium solution of the equations of motion, we shall find a conserved quantity C such that $H_C = H + C$ has a critical point at the equilibrium. We shall then test for definiteness of the second variation of H_C at the equilibrium point. If it is definite, then the equilibrium is said to be formally stable. (See Holm *et al.*[2] for a thorough description and applications of the energy-Casimir method. For details of the following stability analysis, see Lewis *et al.*[7])

One class of conserved quantities consists of the Casimirs of the Poisson manifold $\mathcal{N}$, i.e., functions C on $\mathcal{N}$ satisfying $\{C,F\} = 0$ for all functions F for which the bracket is defined. We will make use of Casimirs of the form

$$C_1(\Sigma,\mathbf{v}) = \int_{D_\Sigma} \Phi(\omega)\,dA,$$

where Φ is a C^2 function on $\mathbb{R}^2$ and $\omega = \langle \operatorname{curl}\mathbf{v},\hat{\mathbf{z}}\rangle$. We will also include the angular momentum

$$J(\Sigma,v) = \int_{D_\Sigma} \langle \mathbf{x}\times\mathbf{v},\hat{\mathbf{z}}\rangle\,dA.$$

Here J is the momentum map associated to the left action of the group $O(2)$ on $\mathcal{N}$. The conservation of J is a consequence of the invariance of the Hamiltonian H under the $O(2)$ action, which implies $\partial J/\partial t = \{J,H\} = 0$. The inclusion of J in the modified Hamiltonian H_C allows us, roughly speaking, to view the fluid from a rotating frame with arbitrary angular velocity.

We take our total conserved quantity to be

$$H_C(\Sigma,v) = \int_{D_\Sigma} \left(\frac{1}{2}|\mathbf{v}|^2 - \mu\langle \mathbf{x}\times\mathbf{v},\hat{\mathbf{z}}\rangle + \Phi(\omega) \right) dA + \tau\int_\Sigma ds,$$

where μ is a constant, as yet undetermined. Using elementary vector identities, we can rewrite H_C as

$$H_C(\Sigma,\mathbf{v}) = \int_{D_\Sigma} \left(\frac{1}{2}|\tilde{\mathbf{v}}|^2 - \frac{1}{2}\mu^2|\mathbf{x}|^2 + \Phi(\omega) \right) dA + \tau\int_\Sigma ds,$$

where $\tilde{\mathbf{v}} = \mathbf{v} - \mu\hat{z}\times\mathbf{x}$. This rephrasing corresponds to viewing the fluid from a flame rotating with constant angular velocity μ; $\tilde{\mathbf{v}}$ is the fluid velocity in the rotating flame.

The first variation of H_C is computed to be

$$DH_C(\Sigma,\mathbf{v})\cdot(\delta\Sigma,\delta\mathbf{v}) \tag{4}$$

$$= \int_{D_\Sigma} (\langle \tilde{\mathbf{v}},\delta\mathbf{v}\rangle + \Phi'(\omega)\cdot\langle \operatorname{curl}\delta\mathbf{v},\hat{\mathbf{z}}\rangle)dA \tag{5}$$

$$+ \int_\Sigma \left(\frac{1}{2}|\tilde{\mathbf{v}}|^2 - \frac{1}{2}\mu^2|\mathbf{x}|^2 + \tau\kappa + \Phi(\omega) \right)\delta\Sigma\,ds. \tag{6}$$

We now consider the case where Σ_e is a circle of radius r and $v_e = (\Omega/2)\hat{\mathbf{z}}\times\mathbf{x}$ for some constant Ω, i.e., the equilibrium flow is rigid rotation with angular velocity Ω. The circle Σ_e has constant mean curvature $\kappa = 1/r$. We require DH_C to vanish at this equilibrium. Since $\omega_e = \langle \operatorname{curl}\mathbf{v}_e,\hat{\mathbf{z}}\rangle = \Omega$, DH_C depends on Φ only through the constants $\Phi(\Omega)$ and $\Phi'(\Omega)$. If we set $\mu = \Omega/2$, corresponding to choosing a frame moving with the rigidly rotating fluid, then $\tilde{\mathbf{v}}_e = 0$, so

$$DH_C(\Sigma_e,\mathbf{v}_e)\cdot(\delta\Sigma,\delta\mathbf{v})$$

$$= \int_{D_\Sigma} \Phi'(\Omega)\cdot\langle \operatorname{curl}\delta\mathbf{v},\hat{\mathbf{z}}\rangle dA + \left(-\frac{1}{2}\left(\frac{\Omega}{2}\right)^2 r^2 + \frac{\tau}{r} + \Phi(\Omega) \right)\int_\Sigma \delta\Sigma\,ds$$

$$= \int_{D_\Sigma} \Phi'(\Omega)\cdot\langle \operatorname{curl}\delta\mathbf{v},\hat{\mathbf{z}}\rangle dA,$$

since $\delta\Sigma$ satisfies $\int_\Sigma \delta\Sigma\,ds = 0$. Thus $DH_C(\Sigma_e,\mathbf{v}_e) = 0$ iff $\Phi'(\Omega) = 0$. For convenience we choose $\Phi = 0$. (Other choices of Φ will give better stability estimates.)

The second variation of H_C at a general point $(\Sigma,\mathbf{v})$ is calculated to be

$$D^2H_C(\Sigma,\mathbf{v})\cdot(\delta\Sigma,\delta\mathbf{v})^2$$
$$= \int_{D_\Sigma} (|\delta\mathbf{v}|^2 + \Phi''(\omega)\cdot|\operatorname{curl}\delta\mathbf{v}|^2)dA$$
$$+ \int_\Sigma \Big[2(\langle\tilde{\mathbf{v}},\delta\mathbf{v}\rangle + \Phi'(\omega)\cdot\langle\operatorname{curl}\delta\mathbf{v},\hat{\mathbf{z}}\rangle)\delta\Sigma$$
$$+ (\tfrac{1}{2}|\tilde{\mathbf{v}}|^2 - \tfrac{1}{2}\mu^2|\mathbf{x}|^2 + \tau\kappa + \Phi(\omega))(\delta^2\Sigma + \kappa\delta\Sigma^2)$$
$$+ \frac{\partial}{\partial\nu}\left(\frac{1}{2}|\tilde{\mathbf{v}}|^2 - \frac{1}{2}\mu^2|\mathbf{x}|^2 + \Phi(\omega)\right)\delta\Sigma^2$$
$$- \tau(\Delta\delta\Sigma)\delta\Sigma - \tau\kappa^2\delta\Sigma^2\Big]\,ds,$$

where Δ is the Laplacian on Σ and $\delta^2\Sigma$ is the variation of $\delta\Sigma$ with respect to Σ. (The presence of the terms involving $\delta^2\Sigma$ is due to the constraints on the variations of Σ arising from the fact that the manifold $\mathcal{S}$ of boundary curves is not a linear space; for fixed Σ the space of $\mathbf{v}$'s on Σ is linear, so no such $\delta^2\mathbf{v}$ term arises.)

For the circular flow described above the second variation reduces to

$$D^2H_C(\Sigma_e,\mathbf{v}_e)\cdot(\delta\Sigma,\delta\mathbf{v})^2$$
$$= \int_{D_\Sigma} |\delta\mathbf{v}|^2\,dA$$
$$- \int_\Sigma \left[\left(\frac{\Omega}{2}\right)^2 r\delta\Sigma^2 + \tau(\Delta\delta\Sigma)\delta\Sigma + \frac{\tau}{r^2}\delta\Sigma^2\right] ds.$$

It follows that $D^2H_C(\Sigma_e,\mathbf{v}_e)$ is positive definite iff

$$\tau\int_\Sigma \left(-\frac{1}{r^2}\delta\Sigma^2 - (\Delta\delta\Sigma)\delta\Sigma\right) ds > \left(\frac{\Omega}{2}\right)^2 r\int_\Sigma \delta\Sigma^2\,ds \tag{7}$$

for all area preserving variations $\delta\Sigma$.

We simplify the expression of this condition by estimating $-(\Delta\delta\Sigma)\delta\Sigma$ using eigenvalues of the negative of the Laplacian on the circle of radius r. The eigenfunctions are $\delta\Sigma_{k,\phi}(\theta) = \cos k(\theta-\phi)$ with eigenvalues $\lambda_{k,\phi} = (k/r)^2$ for all positive integers k. The eigenfunction $\delta\Sigma_{1,\phi} = \cos(\theta-\phi)$ corresponds to an infinitesimal translation in the ϕ direction. If we wish to consider our system modulo position, regarding two configurations as equivalent if one can be obtained from the other by a Euclidean motion, then we can simply ignore the perturbations generated by the lowest eigenfunctions $\delta\Sigma_{1,\phi}$ and test for the definiteness of D^2H_C only with respect to perturbations which actually distort the drop shape. In this case, taking $\lambda_{2,\phi} = 4/r^2$ as the lowest admissible eigenvalue, D^2H_C is positive definite iff

$$3\tau/r^3 > (\Omega/2)^2. \tag{8}$$

It follows from the stability analysis above that the rigidly rotating circular drop $(\Sigma_e,\mathbf{v}_e)$ is formally stable iff (8) holds. If we fix values for τ and r and consider the rotation rate Ω as a variable parameter, then the above statement may be interpreted as saying that the circular solution loses (formal) stability as the parameter Ω increases through the critical value $\Omega_2 = \sqrt{12\tau/r^3}$. Typically, one expects that at a point where a known curve of solutions loses stability (in this case, when the second variation of the Hamiltonian loses definiteness) a "new" branch of solutions bifurcates from the known curve. Thus we look for a bifurcation of critical points of $H - (\Omega/2)J$ at $(\Sigma_e,\mathbf{v}_e)$ when $\Omega = \Omega_2$.

We now consider the O(2) action on the manifold $\mathcal{M}$. This action is induced by the O(2) action on $\mathbb{R}^2$ as follows: Let $R_\gamma\colon \mathbb{R}^2\to\mathbb{R}^2$ denote the action of $\gamma\in$O(2) on $\mathbb{R}^2$. Then $\gamma\cdot\Sigma = \{R_\gamma(\mathbf{x})\colon \mathbf{x}\in\Sigma\}$ and $\gamma\cdot(\Sigma,\mathbf{v}) = (\gamma\cdot\Sigma, R_{\gamma_*}\mathbf{v})$. We are concerned here primarily with relative equilibria; in particular, we are seeking equilibria whose motion is given by the action of some curve in the group O(2). Since the motion of our configurations must be continuous, we do not allow a sudden flip; hence the motion must be given by a smooth rotation. We choose to work with the group O(2) so as to be able to capture any reflectional symmetries of the equilibrium configurations, although this is not the appropriate group for a study of the dynamics of the problem. While the Hamiltonian is invariant under the O(2) action, the dynamics are not invariant under reflection; hence, if one wishes to consider the time-dependent behavior of solutions near the bifurcating equilibria, it is necessary to take SO(2), rather than O(2), as the appropriate symmetry group. The SO(2) action preserves both the bracket and the Hamiltonian; thus the theory of bifurcations of Hamiltonian systems with symmetry may be applied in this case.

When discussing the symmetries of a given configuration it is convenient to do so within a given rotating frame. This is motivated as follows: consider a drop moving in rigid rotation with angular velocity Ω; if the drop shape is fixed at some time t_0 by a reflection across an axis $\bar{\mathbf{x}}$, then at time t it must be fixed by reflection across $R_{\Omega(t-t_0)/2}\bar{x}$, where $R_{\Omega(t-t_0)/2}$ denotes rotation through the angle $\Omega(t-t_0)/2$, while in general it will not continue to be fixed by reflection across $\bar{\mathbf{x}}$. Thus, while the conjugacy class of the isotropy subgroup of the drop is fixed, the actual axes of symmetry of the drop vary in time. Shifting the problem to a rotating frame eliminates this complication; a rigidly rotating drop is stationary in the appropriately chosen frame and hence has a constant isotropy subgroup.

Another advantage of viewing drop symmetries from within a rotating frame is that in this context one can have nontrivial velocity fields which are fixed by orientation reversing actions. More specifically, if one considers rigidly rotating equilibrium configurations, then such drops are fixed points of some subgroup of the O(2) action in the sense that the drop shape is preserved by the subgroup, although the velocity field is reversed. [If one incorporates a time reversal as part of the flip action, then rigid rotation is fixed by the O(2) action.] Within an appropriately chosen rotating frame the velocity field of a rigidly rotating drop is equal to zero; thus, if we consider the action of O(2) within this frame, the drops described above are actual fixed points under the action. For these reasons we shall now shift the problem to a rotating frame and work with triples $(\Sigma,\tilde{\mathbf{v}},\Omega)$, where Σ denotes as usual the drop boundary, Ω is the rotation rate of the rotating frame, and $\tilde{\mathbf{v}}$ is the velocity field in the rotating frame. For an arbitrary pair $(\Sigma,\mathbf{v})$, we take Ω to be the average angular velocity of the velocity field, i.e.,

$$\Omega = \frac{1}{\text{volume } D_\Sigma}\int_{D_\Sigma}\langle\operatorname{curl}\mathbf{v},\hat{\mathbf{z}}\rangle dA;$$

for a rigidly rotating drop, this sets the frame rotation rate

equal to the rotation rate of the drop. For example, the configuration $(\Sigma,(\Omega/2)\hat{\mathbf{z}}\times\mathbf{x})$ is identified with the triple $(\Sigma,0,\Omega)$. The dynamics in the rotating frame are determined by the bracket

$$\{F,G\}=\int_{D_\Sigma}\left\langle\tilde{\omega}+\Omega\hat{\mathbf{z}},\frac{\delta F}{\delta\tilde{\mathbf{v}}}\times\frac{\delta G}{\delta\tilde{\mathbf{v}}}\right\rangle dA + \int_\Sigma\left(\frac{\delta F}{\delta\Sigma}\frac{\delta G}{\delta\tilde{\varphi}}-\frac{\delta G}{\delta\Sigma}\frac{\delta F}{\delta\tilde{\varphi}}\right)ds,$$

where $\tilde{\omega}$ (respectively, $\delta F/\delta\tilde{\mathbf{v}}$ and $\delta F/\delta\tilde{\varphi}$) is the vorticity (respectively, functional derivatives of F with respect to $\tilde{\mathbf{v}}$ and $\tilde{\varphi}$), and the Hamiltonian $\tilde{H}:\mathcal{N}\times\mathbb{R}\to\mathbb{R}$ is given by

$$\tilde{H}(\Sigma,\tilde{\mathbf{v}},\Omega)=\frac{1}{2}\int_{D_\Sigma}\left(|\tilde{\mathbf{v}}|^2-\left(\frac{\Omega}{2}\right)^2|\mathbf{x}|^2\right)dA+\tau\int_\Sigma ds = \left(H-\frac{\Omega}{2}J\right)\left(\Sigma,\tilde{\mathbf{v}}+\frac{\Omega}{2}\hat{\mathbf{z}}\times\mathbf{x}\right).$$

The trivial solution $(\Sigma_e,\mathbf{v}_e)=(\Sigma_e,\mathbf{0},\Omega)$ is a fixed point of the O(2) action in the rotating frame; we expect that the new solution branch bifurcating from $(\Sigma_e,\mathbf{v}_e)$ should be fixed by some subgroup of O(2). We find, in fact, that the new solutions have isotropy subgroup conjugate to the subgroup $\mathbf{Z}_2\times\mathbf{Z}_2$ of O(2) generated by rotation through π and reflection across the x axis. (For a discussion of the theory of bifurcation with symmetry relevant here, see Ihrig and Golubitsky[20] or Golubitsky *et al.*[5])

As we are concerned only with the immediate neighborhood of the point $(\Sigma_e,\mathbf{v}_e)$, it is convenient to work in normal coordinates centered at $(\Sigma_e,\mathbf{v}_e)$. We endow $\mathcal{N}$ with the O(2) invariant metric

$$\langle\langle(\delta\Sigma,\delta\mathbf{v}),(\tilde{\delta\Sigma},\tilde{\delta\mathbf{v}})\rangle\rangle=\int_\Sigma\delta\Sigma\tilde{\delta\Sigma}\,ds+\int_{D_\Sigma}\langle\delta\mathbf{v},\tilde{\delta\mathbf{v}}\rangle dA$$

and use the exponential map exp associated to the metric given above to map a neighborhood V of (0,0) in $T_{(\Sigma_e,\mathbf{v}_e)}\mathcal{N}$ diffeomorphically onto a neighborhood U of $(\Sigma_e,\mathbf{v}_e)$ in $\mathcal{N}$. We define the function $\hat{H}$ on $V\times\mathbb{R}$ to be the pullback of the Hamiltonian plus conserved quantity;

$$\hat{H}((\delta\Sigma,\delta\mathbf{v}),\Omega)=\tilde{H}(\exp(\delta\Sigma,\delta\mathbf{v})).$$

It follows from the invariance of $\tilde{H}$ and the equivariance of exp that $\hat{H}$ is O(2) invariant.

We construct the bifurcation equation using the Liapunov–Schmidt procedure. First we construct the splitting $V=V_1\oplus V_2$, where $V_1=\operatorname{Ker}D^2\hat{H}(0,0,\Omega_2)$ and V_2 is the $\langle\langle\ ,\ \rangle\rangle$ orthogonal complement to V_1. We have

$$D^2\hat{H}(0,0,\Omega_2)=D^2(H-(\Omega_2/2)J)(\Sigma_e,\mathbf{v}_e).$$

Thus

$$V_1=\operatorname{Ker}D^2(H-(\Omega_2/2)J)(\Sigma_e,\mathbf{v}_e) =\{(\cos 2\theta,0),(\sin 2\theta,0)\}.$$

The pure rotation elements of O(2) act on the $\delta\Sigma$ component of $(\delta\Sigma,\delta\mathbf{v})$ by a negative phase shift, i.e.,

$$R^r_\varphi\cdot\delta\Sigma(\theta)=\delta\Sigma(\theta-\varphi);$$

a reflection across the axis at an angle φ to the x axis is given by

$$R^f_\varphi\cdot\delta\Sigma(\theta)=\delta\Sigma(2\varphi-\theta).$$

Let

$$F:V\times\mathbb{R}\to V,$$
$$(\delta\Sigma,\delta\mathbf{v},\Omega)\to\left(\frac{\delta\hat{H}}{\delta\Sigma}(\delta\Sigma,\delta\mathbf{v},\Omega),\frac{\delta\hat{H}}{\delta\mathbf{v}}(\delta\Sigma,\delta\mathbf{v},\Omega)\right)$$

denote the map determined by

$$\int_{D_\Sigma}\left\langle\frac{\delta\hat{H}}{\delta\mathbf{v}}(\delta\Sigma,\delta\mathbf{v},\Omega),\delta\mathbf{v}\right\rangle dA+\int_\Sigma\frac{\delta\hat{H}}{\delta\Sigma}(\delta\Sigma,\delta\mathbf{v},\Omega)\delta\Sigma\,ds = D\hat{H}(\delta\Sigma,\delta\mathbf{v},\Omega)\cdot(\delta\Sigma,\delta\mathbf{v}),$$

where Σ denotes the first component of $\exp(\delta\Sigma,\delta\mathbf{v})$, for all $(\delta\Sigma,\delta\mathbf{v})\in V$. Let $\mathbf{P}$ denote the orthogonal projection $\mathbf{P}:V\to V_2$. The mapping

$$\mathbf{P}\circ F:V_1\times V_2\times\mathbb{R}\to V_2$$

is nonsingular at $(0,0,\Omega_2)$; hence, by the implicit function theorem, there exists an O(2) equivariant mapping $\mathbf{u}:V_1\times\mathbb{R}\to V_2$ such that

$$(\mathbf{P}\circ F)((\delta\Sigma,0)+\mathbf{u}((\delta\Sigma,0),\Omega),\Omega)=0$$

for all $(\delta\Sigma,0)\in V_1$. The bifurcation equation is then given by

$$(\mathrm{Id}-\mathbf{P})\circ F((\delta\Sigma,0)+\mathbf{u}((\delta\Sigma,0),\Omega),\Omega)=0.$$

We introduce the coordinate chart Ψ on a neighborhood $W\times Y$ in $\mathbb{R}^2\times\mathbb{R}$, given by

$$\Psi:W\times Y\to V_1\times Y,$$
$$(x,y,\Omega)\to((x\cos 2\theta+y\sin 2\theta,0)+\mathbf{u}((x\cos 2\theta+y\sin 2\theta,0),\Omega),\Omega).$$

We pull back $\hat{H}$ by Ψ to obtain the bifurcation Hamiltonian $\check{H}:W\times Y\to\mathbb{R}$ given by $\check{H}=\hat{H}\circ\Psi$. In summary, we have reduced the original problem to that of finding critical points of an O(2) invariant function on a two-dimensional space with an O(2) invariant metric.

The bifurcation space W possesses nontrivial symmetry. This symmetry is not artifically imposed on the system; it is a natural property of $\operatorname{Ker}D^2\tilde{H}(\Sigma_e,\mathbf{0})$ which is inherited by the bifurcation space. The O(2) action on W induced by that on V_1 is simply twice the standard O(2) action on $\mathbb{R}^2$; i.e., for $\mathbf{x}=(x,y),\theta\cdot\mathbf{x}=R_{2\theta}(\mathbf{x})$. In this action, rotation through π is equivalent to the identity action, thus the entire space W is fixed by the subgroup $\mathbf{Z}_2$ generated by rotation through π. We also note that any element (x,y) of W is fixed by reflection across the line through the angles $\arctan(x/y)$ and $\arctan(x/y)+\pi/2$. Thus any element of W has isotropy subgroup $\mathrm{O}(2)_{\mathbf{x}}$ conjugate to $\mathbf{Z}_2\times\mathbf{Z}_2$. Since the mappings $\mathbf{u}$ and exp are equivariant, it follows that any solution $\mathbf{x}\in W\times Y$ of the bifurcation equation must be mapped to an $\mathrm{O}(2)_{\mathbf{x}}$ invariant solution in $\mathcal{N}\times\mathbb{R}$ (the "rotating frame space") under $\exp\circ\Psi$.

There are two possible methods for demonstrating that a bifurcation does, in fact, occur. If we consider the group O(2) acting on the space W, then each isotropy subgroup $\mathrm{O}(2)_{\mathbf{x}}$, for some nonzero element $\mathbf{x}$ of W, has a one-dimensional fixed point space consisting of the line spanned by $\mathbf{x}$. Thus, we can apply the equivariant branching lemma to show that there is a branch of relative equilibria with isotropy subgroup $\mathrm{O}(2)_{\mathbf{x}}$ branching from the trivial solution branch at $\Omega=\Omega_2$. The equivariant branching lemma states

that, given a Lie group G acting on a vector space V such that

(i) $\mathrm{Fix}(G) = \{\mathbf{0}\}$;

(ii) $\Gamma \subset G$ is an isotropy subgroup satisfying $\dim(\mathrm{Fix}(\Gamma)) = 1$;

(iii) $g: V \times \mathbb{R} \to V$ is a G-equivariant bifurcation problem satisfying $D_\lambda Dg(\mathbf{0},\lambda_0)\cdot \mathbf{v}_0 \neq 0$ for some λ_0 and some nonzero $\mathbf{v}_0 \in \mathrm{Fix}(\Gamma)$,

then there exists a branch of solutions $(t\mathbf{v}_0, \lambda(t))$ to the equation $g(\mathbf{v},\lambda) = \mathbf{0}$. (See Cicogna[4] or Golubitsky *et al.*[5] for a proof of the equivariant branching lemma.) The first two conditions are clearly satisfied for the $O(2)$ action on W; we take, for example, the subgroup $\mathbf{Z}_2 \times \mathbf{Z}_2$ corresponding to reflection across the x axis and rotation through π as our isotropy subgroup and let $\mathbf{v}_0 = (1,0)$. The equivariance of the map $F = D\breve{H}$ follows from the $O(2)$ invariance of $\breve{H}$; the fact that $DF(0,0,\Omega_2) = D^2\breve{H}(0,0,\Omega_2) = \mathbf{0}$ implies that the map F and the point $(0,0,\Omega_2)$ form a "bifurcation problem." Finally, we compute that

$$\begin{aligned} D_\Omega DF(0,0,\Omega_2)\cdot(1,0) &= D_\Omega\, D^2\breve{H}(0,0,\Omega_2)\cdot(1,0) \\ &= -\Omega_2 \pi r^2/2 \\ &\neq 0, \end{aligned}$$

thus the conditions of the equivariant branching lemma are fulfilled and a branch of solutions of $F(x,0,\Omega) = \mathbf{0}$ must exist. It follows from the equivariance of the equations that the existence of one solution branch implies the existence of an entire circle of solution branches swept out by the group action.

If we wish to consider only symplectic group actions, then we must restrict our attention to the group $SO(2)$, which preserves the symplectic two-form on the space W. In this case, there are no one-dimensional fixed point spaces, so the equivariant branching lemma is not applicable. We can, however, apply the corollary given above to show that a bifurcation occurs. [The fact that the $SO(2)$ action on W is twice the usual $SO(2)$ action does not effect the applicability of the corollary.] The space W and function $\breve{H}$ clearly satisfy condition (i) of the corollary; we shall show that the point $(0,0,\Omega_2)$ satisfies conditions (ii) and (iii):

(ii) $$\begin{aligned} D^2\breve{H}(0,0,\Omega_2) &= \begin{pmatrix} (3\tau/r^2 - (\Omega_2/2)^2 r)\pi r & 0 \\ 0 & (3\tau/r^2 - (\Omega_2/2)^2 r)\pi r \end{pmatrix} \\ &= 0; \end{aligned}$$

(iii) $$\begin{aligned} D_\Omega\, D^2\breve{H}(0,0,\Omega_2) &= \begin{pmatrix} -\Omega_2 \pi r^2/2 & 0 \\ 0 & -\Omega_2 \pi r^2/2 \end{pmatrix} \\ &\neq 0; \end{aligned}$$

provided that $\Omega_2 = \sqrt{12\tau/r^3} \neq 0$ (e.g., that the surface tension coefficient τ is nonzero).

Thus the corollary applies to W and $\breve{H}$ and so there is a branch of critical points of $\breve{H}$ bifurcating from $(0,0,\Omega)$ at $\Omega = \Omega_2$. Note: The matrices computed above are simply scalar multiples of the identity matrix; these scalars are the relevant quantities which must be computed when checking the conditions of the equivariant branching lemma in the $O(2)$ case. Taking the image of the solution branch under the map $\exp \circ \Psi$, we obtain a curve in $\mathcal{N}$ of critical points of the original function $H + \mu J$. The elements in $\mathcal{N}$ thus obtained have the same isotropy subgroups as their preimages in W; in particular, the isotropy subgroups of elements along the new branch near the bifurcation point contain a subgroup conjugate to $\mathbf{Z}_2 \times \mathbf{Z}_2$.

By computing higher-order derivatives of the bifurcation equation, it may be seen that the bifurcation equation has normal form $\mathbf{0} = -\nabla((x^2+y^2)(x^2+y^2+\Omega-\Omega_2))$ (see Lewis[8] for details). Thus, the bifurcation at Ω_2 is subcritical with respect to the bifurcation parameter Ω (i.e., locally the nontrivial solutions exist only for values of Ω less than Ω_2).

Remark 1: The bifurcation is supercritical with respect to angular momentum. Angular momentum is the "physically appropriate" bifurcation parameter in the sense that angular momentum is a physically meaningful conserved quantity for all isolated flows (whereas the bifurcation parameter Ω, which functions mathematically as a Lagrange multiplier, is related to angular velocity, a physical parameter which is only defined for rigidly rotating flows). In this case, the bifurcation equation has normal form $\mathbf{0} = \nabla((x^2+y^2)(x^2+y^2+\mu^2-\mu))$, where μ is the bifurcation parameter and μ_2 is the angular momentum at the bifurcation point; the energy-Casimir method shows the new branch is formally stable near the bifurcation point, which agrees with the general notion of transfer of stability if one views the bifurcation as supercritical. Despite the greater physical relevance of angular momentum, we have chosen the Lagrange multiplier Ω as the bifurcation parameter, since the necessary computations are straightforward in this context and it is easy to interpret the results with respect to angular momentum once the bifurcation branches have been determined.

Remark 2: The symplectic form induced on the reduced space W is a multiple of the standard symplectic two-form on $\mathbb{R}^2$, given by $\omega((x,y),(\tilde{x},\tilde{y})) = y\tilde{x} - \tilde{y}x$, which changes sign under the action of reflections; hence, as remarked above, the symplectic structure on the reduce space W is *not* preserved by the action of the orientation reversing elements of $O(2)$. The symplectic form is, however, preserved under the action of S^1; hence the analysis of Golubitsky and Stewart[4] may be applied, viewing the drop as an S^1 invariant Hamiltonian system. We see that in this context the behavior of the drops near the point of bifurcation is generic.

Remark 3: It can be seen from the second variation of $H + \mu J$ (or $H + C$) that the variation will be indefinite in the direction of $(\delta\Sigma_{k,\phi}, 0) = (\cos k(\theta - \phi), 0)$ when $\mu^2 = (k^2 - 1)\tau/r^3$. It may be shown as above that a subcritical bifurcation occurs at $\Omega_k = \sqrt{4(k^2-1)\tau/r^3}$. The solution branches intersecting the trivial solution branch are invariant under rotation through $2\pi/k$ and flips across lines conjugate to $n\pi/k$; thus their isotropy subgroups are conjugate to D_k, the dihedral group of symmetries of a k-gon. Note: D_k is the semidirect product $\mathbf{Z}_2 \circledS \mathbf{Z}_k$, where $\mathbf{Z}_2$ acts on $\mathbf{Z}_k$ by negation, i.e., by reversing the rotation associated with the elements of $\mathbf{Z}_k$.

Remark 4: The remark in Lewis *et al.*[7] regarding three-

dimensional equilibria is incorrect; it will be corrected elsewhere.

ACKNOWLEDGMENTS

We thank H. Abarbanel, D. Holm, R. Montgomery, and C. Rosenkilde for useful conversations and helpful remarks.

D. L. and J. M. were partially supported by DOE Contract No. DE-AT03-85ER 12097. T. R. was partially supported by a NSF postdoctoral fellowship and a Sloan Foundation fellowship.

[1]M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory* (Springer, Berlin, 1985), Vol. 1.

[2]D. Holm, J. Marsden, T. Ratiu, and A. Weinstein, "Nonlinear stability of fluid and plasma equilibria," Phys. Rep. **123**, 1 (1985). See also Phys. Lett. A **98**, 15 (1983).

[3]M. Golubitsky and I. Stewart, "Generic bifurcation of Hamiltonian systems with symmetry," Physica D **24**, 391 (1987).

[4]G. Cicogna, "Symmetry breakdown from bifurcation," Lett. Nuovo Cimento **31**, 600 (1981).

[5]M. Golubitsky, D. G. Schaeffer, and I. Stewart, *Singularities and Groups in Bifurcation Theory* (Springer, Berlin, to be published), Vol. 2.

[6]D. Lewis, J. Marsden, R. Montgomery, and T. Ratiu, "The Hamiltonian structure for dynamic free boundary problems," Physica D **18**, 391 (1986).

[7]D. Lewis, J. Marsden, and T. Ratiu, "Formal stability of liquid drops with surface tension," *Proceedings of Conference on New Perspectives in Nonlinear Dynamics*, edited by M. Schlesinger, R. Cawley, A. Saenz, and W. Zachary (World Scientific, Singapore, 1986).

[8]D. Lewis, Thesis, University of California at Berkeley, 1987.

[9]D. Holm, J. Marsden, and T. Ratiu, "Nonlinear stability of the Kelvin–Stuart cat's eyes flow," Lect. Appl. Math. **23**, 171 (1986).

[10]P. S. Krishnaprasad and J. E. Marsden, "Hamiltonian structures and stability for rigid bodies with flexible attachments," Arch. Rational Mech. Anal. **98**, 71 (1987).

[11]W. J. Swiatecki, "The rotating, charged or gravitating liquid drop and problems in nuclear physics and astronomy," in *Proceedings of the International Colloquium on Drops and Bubbles*, edited by D. J. Collins, M. S. Plesset, and M. M. Saffren, pp. 52–78.

[12]S. Chandrasekhar, "The stability of a rotating liquid drop," Proc. R. Soc. London Ser. A **284**, 1 (1965).

[13]J. A. Tsamopoulos and R. A. Brown, "Nonlinear oscillations of inviscid drops and bubbles," J. Fluid Mech. **127**, 519 (1983).

[14]R. A. Brown and L. E. Scriven, "The shape and stability of rotating liquid drops," Proc. R. Soc. London Ser. A **371**, 331 (1980).

[15]R. Benner, preprint, 1986.

[16]T. G. Wang, E. H. Trinh, A. P. Croonquist, and D. D. Elleman, "Shapes of rotating free drops: Spacelab experimental results," Phys. Rev. Lett. **56**, 452 (1986).

[17]V. I. Arnold, "Variational principle for three-dimensional steady-state flows of an ideal fluid," Prikl. Mat. Mekh. **29**, 846 (1965) [J. Appl. Math. Mech. **29**, 1002 (1965)]; "Sur la géometrie differentielle des groupes de Lie de dimension infinie et ses applications a l'hydrodynamique des fluides parfaits," Ann. Inst. Fourier (Grenoble) **16**, 319 (1966); "Sur un principe variationnel pour les écoulements stationaires des liquides parfaits et ses applications aux problèmes de stabilité non linéairs." J. Mec. **5**, 29 (1966); "An *a priori* estimate in the theory of hydrodynamic stability," Izv. Vyssh. Uchebn. Z. Math. **54**, 3 (1966) [Transl. Am. Math. Soc. **19**, 267 (1969)].

[18]J. E. Marsden and A. Weinstein, "The Hamiltonian structure of the Maxwell–Vlasov equations," Physica D **4**, 394 (1982); "Coadjoint orbits, vortices and Clebsch variables for incompressible fluids," *ibid.* **7**, 305 (1983).

[19]V. I. Sedenko and V. I. Iudovich, "Stability of steady flows of ideal incompressible fluid with free boundary," Prikl. Mat. Mekh. **42**, 1049 (1978) [J. Appl. Math. Mech. **42**, 1148 (1978)].

[20]E. Ihrig and M. Golubitsky, "Pattern selection with O(3) symmetry," Physica D **13**, 1 (1984).

Effective conductivity of periodic composites composed of two very unequal conductors

Joseph B. Keller
Departments of Mathematics and Mechanical Engineering, Stanford University, Stanford, California 94305

(Received 24 February 1987; accepted for publication 27 May 1987)

The effective conductivity tensor is calculated for a periodic composite composed of alternating rectangular blocks of two very unequal conductors. The two-dimensional case of a checkerboard pattern of rectangles is also treated, and Gautesen's result for it is obtained. The checkerboard of parallelograms is treated, too. The method can be applied to alternating parallelepipeds and to certain other configurations.

I. INTRODUCTION

We consider the effective conductivity tensor $\Sigma(\sigma_a,\sigma_b)$ of certain two- and three-dimensional periodic composites composed of two materials with scalar conductivities σ_a and σ_b. Examples are the "checkerboard" patterns of rectangles or parallelograms in two dimensions (Fig. 1) and the analogous arrangement of rectangular blocks or parallelepipeds in three dimensions (Fig. 2). We shall show how to calculate Σ asymptotically as σ_a/σ_b tends to zero or to infinity. This work grew out of an attempt to obtain a simpler deriviation of one of Gautesen's[1] recent results for a rectangular "checkerboard" in two dimensions.

First we shall present our result for the three-dimensional alternating arrangment of rectangular blocks shown in Fig. 2. Let the edges of the blocks be parallel to the axes, and let h_i be the length of the edge parallel to the x_i axis. Clearly the axes are the principal directions of Σ. Our result for Σ_{11} is

$$\Sigma_{11}(\sigma_a,\sigma_b)\sim[h_1(h_2+h_3)/h_2h_3](\sigma_a\sigma_b)^{1/2} \quad \text{as } \sigma_a/\sigma_b\to 0 \text{ or } \infty. \tag{1.1}$$

For cubes this yields $2(\sigma_a\sigma_b)^{1/2}$, which was obtained before by Milton[2] and by Söderberg and Grimvall,[3] while when h_3 tends to infinity it yields Gautesen's two-dimensional result $(h_1/h_2)(\sigma_a,\sigma_b)^{1/2}$. Cyclic permutation of indices in (1.1) yields Σ_{22} and Σ_{33}.

In Sec. II we derive the result for a two-dimensional rectangular checkerboard and in Sec. III we derive the three-dimensional result (1.1). In Sec. IV we calculate the conductance σ between two highly conducting parallelograms that meet at a corner. Then in Sec. V we use σ to determine Σ for a checkerboard of parallelograms. The result (1.1) is not uniform in the h_i, so an appropriate modification of it is discussed in Sec. VI. Finally in Sec. VII we discuss these results and indicate some generalizations of them.

II. RECTANGULAR CHECKERBOARD PATTERN

We begin with the two-dimensional checkerboard of rectangles with conductivities σ_a and σ_b shown in Fig. 1(a). The principal axes of the effective conductivity tensor Σ are the x_1 and x_2 axes, so $\Sigma_{12}=\Sigma_{21}=0$. By definition Σ_{11} is just the average current density in the x_1 direction resulting from an electric field of unit strength in the x_1 direction. We suppose that $\sigma_a \gg \sigma_b$. Then the current will flow through the highly conducting regions as much as possible, and it will traverse the poorly conducting regions only at the corners where it goes from one highly conducting rectangle to a diagonally adjacent one. In Sec. IV we shall show that there is a well-defined conductance σ associated with such a corner.

We now use σ to find the current density due to a unit electric field along the x_1 axis. This field produces a voltage difference h_1 between the planes $x_1=-h_1/2$ and $x_1=h_1/2$. As a result a current $h_1\sigma$ flows across each corner, and the resulting current density Σ_{11} is this current divided by the vertical spacing h_2 between corners. Thus $\Sigma_{11}\sim h_1\sigma/h_2$, and similarly $\Sigma_{22}\sim h_2\sigma/h_1$. These results are asymptotic as $\sigma_a/\sigma_b\to\infty$ because only then can we associate all the conductance with the corners.

For a square checkerboard we have shown[4,5] that $\Sigma_{11}=\Sigma_{22}=(\sigma_a\sigma_b)^{1/2}$. Therefore by applying our asymptotic result to this case, for which $h_1=h_2$, we find that $\sigma\sim(\sigma_a\sigma_b)^{1/2}$. By using this value of σ in the preceding for-

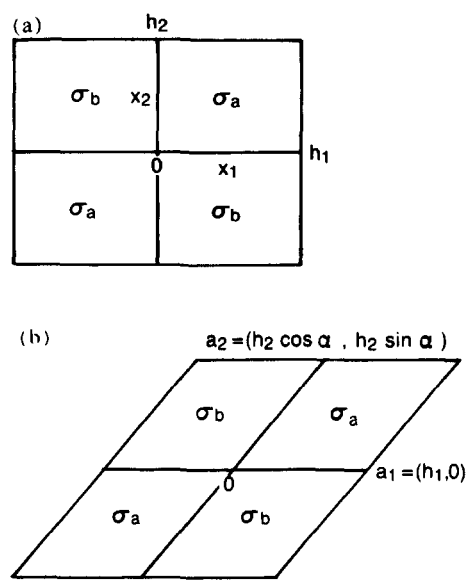


FIG. 1. (a) Part of a checkerboard pattern of rectangles with conductivities σ_a and σ_b. Edges parallel to the x_1 axis are of length h_1 and those parallel to the x_2 axis are of length h_2. (b) Part of an alternating pattern of parallelograms with conductivities σ_a and σ_b. The vertices are generated by the vectors $a_1=(h_1,0)$ and $a_2=(h_2\cos\alpha, h_2\sin\alpha)$.

 0022-2488/87/102516-05$02.50 © 1987 American Institute of Physics

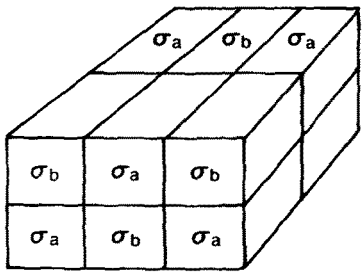


FIG. 2. Part of an alternating arrangment of rectangular parallelipipeds of conductivities σ_a and σ_b.

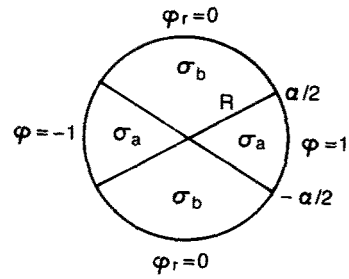


FIG. 3. A corner of the pattern in Fig. 1(b), rotated to be symmetric about the coordinate axes. The value of the potential $\varphi = \pm 1$ and its derivative $\varphi_r = 0$ are indicated on a circle of radius R centered at the vertex.

mulas, we obtain for the rectangular checkerboard

$$\Sigma_{11} \sim (h_1/h_2)(\sigma_a\sigma_b)^{1/2}, \qquad \Sigma_{22} \sim (h_2/h_1)(\sigma_a\sigma_b)^{1/2} \quad \text{as } \sigma_a/\sigma_b \to \infty. \tag{2.1}$$

This is just Gautesen's result,[1] which he derived in a different way that proves it to be asymptotically correct. In Sec. IV we shall calculate σ directly for general corners, including rectangular ones, and again obtain the value $(\sigma_a\sigma_b)^{1/2}$ for the present case.

III. RECTANGULAR BLOCK PATTERN IN THREE DIMENSIONS

We shall now obtain the result (1.1) for the medium of alternating rectangular blocks shown in Fig. 2. The diagonal element Σ_{11} is, as before, the average current density in the x_1 direction due to a unit electric field in the x_1 direction. When $\sigma_a \gg \sigma_b$ the current will flow through the highly conducting blocks as much as possible. It will pass through the poorer conductors only along the edges where it goes from one highly conducting block to another. The conductance per unit length of such an edge is just σ, where σ is the two-dimensional conductance introduced in the preceding section. The voltage between the planes $x_1 = \pm h_1/2$ is just h_1. Therefore the current through each highly conducting block is $(2h_2 + 2h_3)h_1\sigma$ because $2h_2 + 2h_3$ is the length of edge between a highly conducting block and its highly conducting neighbors in the direction of increasing x_1. The current density is obtained by dividing this current by the area $2h_1h_2$, which is the cross-sectional area normal to the x_1 axis of a highly conducting block and a poorer conducting neighbor. In this way we get $\Sigma_{11} \sim h_1(h_2 + h_3)\sigma/h_2h_3$. When we use the value $\sigma \sim (\sigma_a\sigma_b)^{1/2}$ in this formula, we obtain our result (1.1).

IV. RESISTANCE OF A CORNER

In order to treat the two-dimensional medium of alternating parallelograms shown in Fig. 1(b), we shall first determine the conductance $\sigma(\alpha)$ of the corner shown in Fig. 3. The medium with the high conductivity σ_a occupies the sector $-\alpha/2 < \theta < \alpha/2$ and the opposite sector, while the other two sectors contain the medium of conductivity σ_b. The corner is surrounded by a circle of radius R which is an insulator in the σ_b regions and a perfect conductor in the σ_a regions. Its potential is $+1$ in the interval $-\alpha/2 < \theta < \alpha/2$ and -1 in the opposite sector. Then the current between these two conductors is just the potential difference multiplied by the conductance, i.e., 2σ. We shall calculate the current and thus determine σ.

In terms of polar coordinates $\hat{\rho}, \theta$ the potential φ must be a function of $\hat{\rho}/R$ and θ, by dimensional analysis: $\varphi = \varphi(\hat{\rho}/R, \theta)$. Then the current, which is equal to 2σ, is given by

$$2\sigma(\alpha) = \int_{-\alpha/2}^{\alpha/2} \sigma_a \frac{\partial\varphi}{\partial\hat{\rho}}\left(\frac{\hat{\rho}}{R}, \theta\right)\Big|_{\hat{\rho}=R} R\, d\theta = \sigma_a \int_{-\alpha/2}^{\alpha/2} \varphi_\rho(1,\theta)\, d\theta. \tag{4.1}$$

Here φ_ρ is the derivative of φ with respect to its first argument $\rho = \hat{\rho}/R$. From (4.1) we see that σ is independent of R, the radius of the circular conductors and insulators, so it can be interpreted as a property of the corner.

To simplify (4.1) we use the symmetry of φ about $\theta = 0$ to write the integral as twice the integral from 0 to $\alpha/2$:

$$\alpha = \sigma_a \int_0^{\alpha/2} \varphi_\rho(1,\theta)\, d\theta. \tag{4.2}$$

Now φ must be a harmonic function satisfying the following conditions:

$$\varphi(1,\theta) = 1, \quad 0 < \theta < \alpha/2, \tag{4.3}$$

$$\varphi_\rho(1,\theta) = 0, \quad \alpha/2 < \theta < \pi/2, \tag{4.4}$$

$$\varphi_\theta(\rho,0) = \varphi(\rho,\pi/2) = 0, \quad 0 < \rho < 1, \tag{4.5}$$

$$\varphi\left(\rho, \frac{\alpha}{2} -\right) = \varphi\left(\rho, \frac{\alpha}{2} +\right), \qquad \sigma_a\varphi_\theta\left(\rho, \frac{\alpha}{2} -\right) = \sigma_b\varphi_\theta\left(\rho, \frac{\alpha}{2} +\right), \quad 0 < \rho < 1. \tag{4.6}$$

Equation (4.3) follows from the specification of the potential on the conductor, (4.4) is the condition of no current flow into the insulator, (4.5) expresses the evenness of φ about $\theta = 0$ and its oddness about $\theta = \pi/2$, while (4.6) states that φ and the normal component of current are continuous at $\theta = \alpha/2$.

To solve for φ we write

$$\varphi = A_a\rho^\nu \cos\nu\theta, \quad 0 \leqslant \theta \leqslant \alpha/2, \tag{4.7}$$

$$\varphi = A_b\rho^\nu \sin\nu(\pi/2 - \theta), \quad \alpha/2 \leqslant \theta \leqslant \pi/2. \tag{4.8}$$

These functions are harmonic for any ν and they satisfy (4.5). Upon imposing (4.6) we get

$$A_a \cos\frac{\nu\alpha}{2} = A_b \sin\nu\left(\frac{\pi}{2}-\frac{\alpha}{2}\right),$$
$$A_a\sigma_a \sin\frac{\nu\alpha}{2} = A_b\sigma_b \cos\nu\left(\frac{\pi}{2}-\frac{\alpha}{2}\right). \tag{4.9}$$

Dividing the second equation in (4.9) by the first yields

$$\sigma_a \tan\frac{\nu\alpha}{2} = \sigma_b \cot\nu\left(\frac{\pi}{2}-\frac{\alpha}{2}\right). \tag{4.10}$$

When $\sigma_a/\sigma_b \gg 1$, it follows from (4.10) that the first positive root for ν is small. Therefore we expand tan and cot and solve for ν to obtain

$$\nu \sim 2(\sigma_b/\alpha(\pi-\alpha)\sigma_a)^{1/2}, \quad \text{for } \sigma_a/\sigma_b \gg 1. \tag{4.11}$$

Now (4.7) and (4.8) become

$$\varphi \sim A_a\rho^{\nu}, \quad 0\leqslant\theta\leqslant\alpha/2, \tag{4.12}$$

$$\varphi \sim A_b\rho^{\nu}\nu(\pi/2-\theta), \quad \alpha/2\leqslant\theta\leqslant\pi/2. \tag{4.13}$$

By using (4.12) in (4.3) we find that $A_a \sim 1$ and then the first of Eqs. (4.9) yields $A_b \sim 2/\nu(\pi-\alpha)$. We also see from (4.13) that (4.4) is satisfied to order ν. Finally we use (4.12) for φ in (4.2) with $A_a \sim 1$ to get

$$\sigma(\alpha) \sim \sigma_a\nu\alpha/2. \tag{4.14}$$

Then by substituting (4.11) for ν into (4.14) we obtain the final result

$$\sigma(\alpha) \sim (\alpha\sigma_a\sigma_b/(\pi-\alpha))^{1/2}, \quad \text{for } \sigma_a/\sigma_b \gg 1. \tag{4.15}$$

When $\alpha = \pi/2$ this reduces to the result $\sigma(\pi/2) \sim (\sigma_a\sigma_b)^{1/2}$, which we obtained in Sec. II.

V. PARALLELOGRAMS IN A CHECKERBOARD PATTERN

We shall use the result (4.15) to calculate Σ for the two-dimensional checkerboard of parallelograms shown in Fig. 1(b). First we note that the average current density I is related to the average applied field E by $I = \Sigma E$, and therefore the component of I parallel to the unit vector n is

$$n\cdot I = n\cdot\Sigma E. \tag{5.1}$$

By using this relation for three pairs of values of n and E, we shall obtain three equations from which to determine the three independent components of Σ.

First we introduce the two vectors a_1 and a_2, which generate the lattice of vertices, defined by $a_1 = h_1(1,0)$ and $a_2 = h_2(\cos\alpha,\sin\alpha)$. Here h_1 and h_2 are the lengths of the two sides of a parallelogram, and α is the angle between them. The unit normals to these sides are $b_1 = (0,1)$ and $b_2 = (\sin\alpha, -\cos\alpha)$. Now we choose $n = E = b_1$ in (5.1) to obtain

$$b_1\cdot I = \Sigma_{22}. \tag{5.2}$$

To compute the current density on the left side of (5.2) we note that the voltage across a parallelogram in the vertical direction is $h_2\sin\alpha$. The vertical current through one highly conducting parallelogram is the sum of currents across two corners with angles α and $\pi-\alpha$. Thus the current is $h_2\sin\alpha[\sigma(\alpha)+\sigma(\pi-\alpha)]$. The current density is obtained by dividing this current by $2h_1$, the horizontal extent of two parallelograms. Thus (5.2) yields

$$\Sigma_{22} \sim [(h_2\sin\alpha)/2h_1][\sigma(\alpha)+\sigma(\pi-\alpha)]. \tag{5.3}$$

Finally (4.15) and (5.3) give

$$\Sigma_{22} \sim \frac{h_2}{2h_1}(\sigma_a\sigma_b)^{1/2} \times\sin\alpha\left[\left(\frac{\alpha}{\pi-\alpha}\right)^{1/2}+\left(\frac{\pi-\alpha}{\alpha}\right)^{1/2}\right]. \tag{5.4}$$

Next we take $n = b_2$ and $E = b_1$ in (5.1) to get

$$b_2\cdot I = \sin\alpha\Sigma_{12} - \cos\alpha\Sigma_{22}. \tag{5.5}$$

The vertical voltage across a parallelogram is still $h_2\sin\alpha$. The net current through a parallelogram in the b_2 direction is the difference between the current in at the corner of angle α and the current out at the corner of angle $\pi-\alpha$. Thus it is $h_2\sin\alpha[\sigma(\alpha)-\sigma(\pi-\alpha)]$, and it must be divided by the width $2h_2$ of two parallelograms in the a_2 direction. Thus (5.5) becomes

$$\sin\alpha\Sigma_{12} - \cos\alpha\Sigma_{22} \sim [(\sin\alpha)/2][\sigma(\alpha)-\sigma(\pi-\alpha)]. \tag{5.6}$$

Solving for Σ_{12} in (5.6) with the aid of (5.3) yields

$$\Sigma_{12} \sim [(h_2\cos\alpha)/2h_1][\alpha(\alpha)+\sigma(\pi-\alpha)] + \tfrac{1}{2}[\sigma(\alpha)-\sigma(\pi-\alpha)]. \tag{5.7}$$

This and (4.15) for σ gives

$$\Sigma_{12} \sim \frac{h_2}{2h_1}(\sigma_a\sigma_b)^{1/2}\cos\alpha\left[\left(\frac{\alpha}{\pi-\alpha}\right)^{1/2}+\left(\frac{\pi-\alpha}{\alpha}\right)^{1/2}\right] + \frac{(\sigma_a\sigma_b)^{1/2}}{2}\left[\left(\frac{\alpha}{\pi-\alpha}\right)^{1/2}-\left(\frac{\pi-\alpha}{\alpha}\right)^{1/2}\right]. \tag{5.8}$$

As a third choice we take $n = E = b_2$ in (5.1), which becomes

$$b_2\cdot I = \sin^2\alpha\Sigma_{11} - 2\sin\alpha\cos\alpha\Sigma_{12} + \cos^2\alpha\Sigma_{22}. \tag{5.9}$$

The voltage in the b_2 direction across one parallelogram is $b_2\cdot a_1 = h_1\sin\alpha$ and the current in the b_2 direction is $h_1\sin\alpha[\sigma(\alpha)+\sigma(\pi-\alpha)]$. Dividing this current by $2h_2$ and using it in (5.9) yields

$$\sin^2\alpha\Sigma_{11} - 2\sin\alpha\cos\alpha\Sigma_{12} + \cos^2\alpha\Sigma_{22} \sim [(h_1\sin\alpha/2h_2][\sigma(\alpha)+\sigma(\pi-\alpha)]. \tag{5.10}$$

Solving for Σ_{11} leads to

$$\Sigma_{11} \sim \left(\frac{h_1}{h_2}+\frac{h_2}{h_1}\cos^2\alpha\right)\frac{1}{2\sin\alpha}[\sigma(\alpha)+\sigma(\pi-\alpha)] + \frac{\cos\alpha}{\sin\alpha}[\sigma(\alpha)-\sigma(\pi-\alpha)]. \tag{5.11}$$

When (4.15) is used in (5.11) it becomes

$$\Sigma_{11} \sim \left(\frac{h_1}{h_2}+\frac{h_2\cos^2\alpha}{h_1}\right)\frac{(\sigma_a\sigma_b)^{1/2}}{2\sin\alpha} \times\left[\left(\frac{\alpha}{\pi-\alpha}\right)^{1/2}+\left(\frac{\pi-\alpha}{\alpha}\right)^{1/2}\right] + (\sigma_a\sigma_b)^{1/2} \times\frac{\cos\alpha}{\sin\alpha}\left[\left(\frac{\alpha}{\pi-\alpha}\right)^{1/2}-\left(\frac{\pi-\alpha}{\alpha}\right)^{1/2}\right]. \tag{5.12}$$

Equations (5.4), (5.8), and (5.12) determine Σ.

VI. NONUNIFORMITY

The result (2.1) is not valid when h_2/h_1 tends to zero or to infinity. To obtain a result that is uniformly valid we must take account of the conductivity of the material away from the corner. We can do this roughly by replacing σ in the expression $\Sigma_{11}\sim h_1\sigma/h_2$ by the series-parallel conductance

$$\frac{1}{\sigma^{-1}+2h_1/h_2\sigma_a}+\frac{h_2}{h_1/2\sigma_a+h_1/2\sigma_b}$$
$$\sim\frac{1}{\sigma^{-1}+2h_1/h_2\sigma_a}+\frac{2h_2\sigma_b}{h_1}. \tag{6.1}$$

The first term accounts for the fact that the corner is in series with the resistance of half the rectangle of material σ_a, and this resistance tends to $2h_1/h_2\sigma_a$ as h_1/h_2 becomes large. The second term represents the conductance directly across the rectangles, which tends to $2h_2\sigma_b/h_1$ as $\sigma_a/\sigma_b\to\infty$. Then Σ_{11} becomes, with $\sigma=(\sigma_a\sigma_b)^{1/2}$ in (6.1),

$$\Sigma_{11}\sim\frac{h_1}{h_2}(\sigma_a\sigma_b)^{1/2}\left[1+\frac{2h_1}{h_2}\left(\frac{\sigma_a}{\sigma_b}\right)^{1/2}\right]^{-1}+2\sigma_b$$
$$\text{as}\quad\frac{\sigma_a}{\sigma_b}\to\infty. \tag{6.2}$$

By interchanging h_1 and h_2 in (6.2) we get Σ_{22}.

From (6.2) we find that

$$\Sigma_{11}\sim2\sigma_b,\quad\text{for}\quad\frac{h_1}{h_2}\ll\left(\frac{\sigma_b}{\sigma_a}\right)^{1/2}, \tag{6.3}$$

$$\Sigma_{11}\sim\frac{\sigma_a}{2},\quad\text{for}\quad\frac{h_1}{h_2}\gg\left(\frac{\sigma_a}{\sigma_b}\right)^{1/2}, \tag{6.4}$$

$$\Sigma_{11}\sim\frac{h_1}{h_2}(\sigma_a\sigma_b)^{1/2},\quad\text{for}\quad\left(\frac{\sigma_b}{\sigma_a}\right)^{1/2}\ll\frac{h_1}{h_2}\ll\left(\frac{\sigma_a}{\sigma_b}\right)^{1/2}. \tag{6.5}$$

The conditions for validity of (2.1) are thus those in (6.5).

In the same way, we can modify (1.1) for rectangular blocks to obtain

$$\Sigma_{11}\sim\frac{h_1}{2h_2h_3}\left[\left(\frac{(\sigma_a\sigma_b)^{-1/2}}{2h_2+2h_3}+\frac{h_1}{h_2h_3\sigma_a}\right)^{-1}+2h_2h_3\left(\frac{h_1}{2\sigma_a}+\frac{h_1}{2\sigma_b}\right)^{-1}\right]$$
$$\sim\frac{h_1(h_2+h_3)}{h_2h_3}(\sigma_a\sigma_b)^{1/2}\times\left[1+\frac{2h_1(h_2+h_3)}{h_2h_3}\left(\frac{\sigma_b}{\sigma_a}\right)^{1/2}\right]^{-1}+2\sigma_b. \tag{6.6}$$

Thus

$$\Sigma_{11}\sim2\sigma_b,\quad\text{for}\quad\frac{h_1(h_2+h_3)}{h_2h_3}\ll\left(\frac{\sigma_b}{\sigma_a}\right)^{1/2}, \tag{6.7}$$

$$\Sigma_{11}\sim\frac{\sigma_a}{2},\quad\text{for}\quad\frac{h_1(h_2+h_3)}{h_2h_3}\gg\left(\frac{\sigma_a}{\sigma_b}\right)^{1/2}, \tag{6.8}$$

$$\Sigma_{11}\sim\frac{h_1(h_2+h_3)}{h_2h_3}(\sigma_a\sigma_b)^{1/2},$$
$$\text{for}\quad\left(\frac{\sigma_b}{\sigma_a}\right)^{1/2}\ll\frac{h_1(h_2+h_3)}{h_2h_3}\ll\left(\frac{\sigma_a}{\sigma_b}\right)^{1/2}. \tag{6.9}$$

VII. DISCUSSION

The method of Sec. V can be applied to a three-dimensional alternating configuration of parallelepipeds, using the value of σ given by (4.15). Furthermore all of our results remain valid if the squares, parallelograms, rectangular blocks, or parallelepipeds are distorted, provided that their shapes near the corners in two dimensions, and near the edges in three dimensions, are unchanged. In addition the method can be applied to three-dimensional periodic media with curved edges and a variable angle $\alpha(s)$ along each edge. Then we must integrate $\sigma[\alpha(s)]$ along each edge to find its conductance.

The concept of corner conductance can be extended to other kinds of "corners" besides those treated in Sec. IV. For example, suppose that the two highly conducting sectors in Fig. 3 did not meet, but were separated by a small gap filled with the low conductance material. Then the conductance between the two highly conducting sectors could still be defined, and the same method could employed. The results of Sec. II, III, V would still apply with the appropriate value of σ.

The possibility of analyzing a continuous problem by replacing it with a network of lumped elements is a consequence of the asymptotic behavior of the solution with respect to some parameter. In the present case the parameter is the conductivity ratio σ_a/σ_b, which tends to zero or to infinity. In other cases it is a geometrical ratio. The analytical basis for the procedure is provided by the method of matched asymptotic expansions. In the present case, for example, the construction in Sec. IV provides the leading term in the inner expansion valid near each corner of the rectangles or parallelograms. The leading term in the outer expansion within each highly conducting rectangle or parallelogram is a harmonic function. It has current sources at two vertices and curent sinks at the other two, and a vanishing normal derivative on the boundaries. The magnitudes of the currents are determined by matching the inner and outer expansions. By constructing these expansions we could obtain further terms in the asymptotic expansion of Σ.

We have used similar ideas before to treat periodic configurations of perfectly conducting cylinders or spheres, or nonconducting cylinders, in a finitely conducting matrix.[6] Batchelor and O'Brien[7] carried it over to highly conducting bodies, and Buchal and Keller[8] extended it to time harmonic problems.

ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research, the Air Force Office of Scientific Research, and the National Science Foundation.

[1]A. Gautesen, "The effective conductivity of a rectangular checkerboard composite," preprint, Ames Research Laboratory, Iowa State University.
[2]G. W. Milton, "Theoretical studies of the transport properties of inhomogeneous media," Report Physics IV, Department of Theoretical Physics, University of Sydney, 1979.
[3]M. Söderberg and G. Grimvall, "Current distributions for a two-phase material with chequer-board geometry," J. Phys. C: Solid State Phys. **16**, 1085 (1983).
[4]J. B. Keller, "A theorem on the conductivity of a composite medium," J. Math. Phys. **5**, 548 (1964).
[5]J. Nevard and J. B. Keller, "Reciprocal relations for effective conductivities of anisotropic media," J. Math. Phys. **26**, 2761 (1985).
[6]J. B. Keller, "Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders, or nonconducting cylinders," J. Appl. Phys. **34**, 991 (1963).
[7]G. K. Batchelor and R. W. O'Brien, "Thermal or electrical conduction through a granular material," Proc. R. Soc. London Ser. A **355**, 313 (1977).
[8]R. Buchal and J. B. Keller, "Impedance between perfect conductors in a finitely conducting medium with application to composite media," J. Appl. Phys. **34**, 3414 (1963).